SIAM Review
Vol. 57, No. 4, pp. 483–531

A Survey of Projection-Based Model Reduction Methods for Parametric Dynamical Systems*

Peter Benner†
Serkan Gugercin‡
Karen Willcox§

Abstract. Numerical simulation of large-scale dynamical systems plays a fundamental role in studying a wide range of complex physical phenomena; however, the inherent large-scale nature of the models often leads to unmanageable demands on computational resources. Model reduction aims to reduce this computational burden by generating reduced models that are faster and cheaper to simulate, yet accurately represent the original large-scale system behavior. Model reduction of linear, nonparametric dynamical systems has reached a considerable level of maturity, as reflected by several survey papers and books. However, parametric model reduction has emerged only more recently as an important and vibrant research area, with several recent advances making a survey paper timely. Thus, this paper aims to provide a resource that draws together recent contributions in different communities to survey the state of the art in parametric model reduction methods.

Parametric model reduction targets the broad class of problems for which the equations governing the system behavior depend on a set of parameters. Examples include parameterized partial differential equations and large-scale systems of parameterized ordinary differential equations. The goal of parametric model reduction is to generate low-cost but accurate models that characterize system response for different values of the parameters. This paper surveys state-of-the-art methods in projection-based parametric model reduction, describing the different approaches within each class of methods for handling parametric variation and providing a comparative discussion that lends insights to potential advantages and disadvantages in applying each of the methods. We highlight the important role played by parametric model reduction in design, control, optimization, and uncertainty quantification—settings that require repeated model evaluations over different parameter values.

Key words. dynamical systems, parameterized model reduction, (Petrov–)Galerkin projection, Krylov subspace method, moments, interpolation, proper orthogonal decomposition, balanced truncation, greedy algorithm

AMS subject classifications. 35B30, 37M99, 41A05, 65K99, 93A15, 93C05

DOI. 10.1137/130932715

*Received by the editors August 13, 2013; accepted for publication (in revised form) June 29, 2015; published electronically November 5, 2015.
http://www.siam.org/journals/sirev/57-4/93271.html

†Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1, D-39106 Magdeburg, Germany (benner@mpi-magdeburg.mpg.de). The work of this author was supported by DFG grant BE 2174/16-1 "Multivariate Interpolation Methods for Parametric Model Reduction."

‡Department of Mathematics, Virginia Tech, Blacksburg, VA 24061-0123 (gugercin@math.vt.edu). The work of this author was supported by NSF grant DMS-1217156 (Program Manager L.M. Jameson).

§Department of Aeronautics & Astronautics, Massachusetts Institute of Technology, Cambridge, MA 02139 (kwillcox@mit.edu). The work of this author was supported by AFOSR Computational Mathematics grant FA9550-12-1-0420 (Program Manager F. Fahroo) and the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research, Applied Mathematics program under awards DE-FG02-08ER2585 and DE-SC0009297 (Program Manager A. Landsberg).

1. Introduction. Dynamical systems are the basic framework for modeling and control of an enormous variety of complex systems of scientific interest or industrial value. Examples include heat transfer, fluid dynamics, chemically reacting flows, biological systems, signal propagation and interference in electronic circuits, wave propagation and vibration suppression in large structures, and design of micro-electro-mechanical systems (MEMS). Numerical simulation of the associated models has been one of the few available means for studying complex underlying physical phenomena. However, the growing need for improved accuracy requires the inclusion of more detail in the modeling stage, leading inevitably to larger-scale, more complex models of dynamical systems. The ever increasing push towards improving system performance leads to a need to simulate many different possible realizations of the system. Performing multiple simulations in such large-scale settings often presents unmanageably large demands on computational resources. Alleviation of this computational burden is the main motivation for deriving *reduced models*—low-dimensional, efficient models that are fast to solve but that approximate well the underlying high-resolution

simulations. The field of *model reduction* encompasses a broad set of mathematical methods to generate and evaluate these reduced models.

In this paper, we focus on the broad class of problems for which the equations representing the system dynamics depend on a set of parameters, and the goal is to characterize system response for different values of the parameters. These parameters may enter the models in many ways, representing, for example, material properties, system geometry, system configuration, initial conditions, and boundary conditions. This parametric dependence presents a unique set of challenges for model reduction—building a reduced model requires querying the expensive underlying full model, thus one cannot afford to create a new reduced model for every change in the parameter values. Hence, the desired approach is to generate a parametric reduced model that approximates the original full-order dynamical system with high fidelity over a range of parameters. This is the goal of *parametric model reduction*. This survey provides an overview of state-of-the-art methods in parametric model reduction for large-scale dynamical systems. Recent years have seen considerable progress in this field, with several classes of methods emerging. This paper highlights the different approaches within each class of methods for handling parametric variation. We provide a comparative discussion that lend insights to potential advantages and disadvantages in applying each of the methods.

1.1. Applications of Parametric Model Reduction. Why is parametric model reduction important and useful? It fills a critical need in design, control, optimization, and uncertainty quantification settings—settings that require repeated model evaluations over a potentially large range of parameter values. These are settings in which we are willing to forgo a large up-front cost, the so-called *offline* cost, in order to obtain a reduced model that allows rapid yet accurate simulation over the range of parameters, the so-called *online* phase.

The *design* setting may require evaluation of system performance over a range of parameter values representing critical design constraints. For example, parametric reduced models capturing the coupled fluid dynamic and structural dynamic behavior of an aircraft configuration provide rapid evaluation of aeroelastic performance over a range of aircraft operating conditions [8, 158, 159]. This enables rapid characterization of the aircraft's flight envelope, calculations that would otherwise require many weeks of computation time. Parametric reduced models have also shown to be an important enabling technology in the synthesis and design of interconnect [55, 71], semiconductor devices [130], and MEMS [31, 91] as well as in electrochemical [95] and electrothermal applications [96].

In *control* design we desire to drive the system dynamics into a desired configuration, while accounting for parametrically varying dynamics. Examples include design process control in Rayleigh–Bénard convection with varying Rayleigh number [163], and control for fluid flow with varying Reynolds number and/or shape parameters [127, 170]. Instead of designing a new controller for every new parameter, which would be a large computational burden and intractable for online controller design, one could design either a single reduced-order controller that performs effectively over the full parameter space or one that parametrically adapts to the changes. Model predictive control is another example for which a cheap and reliable model is required to control the system of interest. Model reduction has been applied to model predictive control in the nonparametric [133] and parametric [7] cases.

Both design and control can involve *optimization*, where the goal may be an optimal system configuration with respect to a certain performance objective (e.g.,

maximal throughput, minimal weight, minimal energy consumption, etc.) or an optimal controller. Most optimization algorithms require multiple evaluations of the forward model for varying parameter configurations; this is where parametric reduced models can play a role. Past work has used trust regions to manage the reduced model as the optimization proceeds [17, 88, 227]. Another approach for optimal control incorporates the optimality conditions into derivation of the reduced model [144]. A combination of domain decomposition and model reduction has also been developed for optimal control and shape optimization problems [12, 13]. Another recent framework employed parametric reduced models that ensure exact matching of the objective function and gradient evaluations for a subset of parameter values [29, 80]. Model reduction for optimization problems constrained by partial differential equations (PDEs) has recently been surveyed in [49].

Uncertainty quantification is another area that demands repeated model evaluations—often many thousands or even millions of evaluations are needed to sample the uncertainty space (e.g., using Monte Carlo sampling). For example, using parametric reduced models, the forward propagation of uncertainty through complex systems, such as those typically modeled using large-scale computational fluid dynamic (CFD) models can be achieved in turnaround times useful for design [62]. Parametric reduced basis models have also been combined with stochastic collocation to solve systems governed by PDEs with random coefficients [86]. Another example is large-scale statistical inverse problems for which Markov chain Monte Carlo methods are computationally intractable, either because they require excessive amounts of CPU time or because the parameter space is too large to be explored effectively by state-of-the-art sampling methods. In these cases, parametric model reduction over both state and parameter spaces can make tractable the solution of large-scale inverse problems that otherwise cannot be solved [80, 102, 156, 220, 72].

While all four of these areas—design, control, optimization, and uncertainty quantification—require repeated model evaluations over the parameter space, the nature of the parameter sampling in an optimization setting is generally different from that in the other settings. For design, control, and uncertainty quantification, we are often (although not always) interested in querying the model over a range of parameter values (e.g., considering a range of operating conditions in a control setting or considering a range of probable parameter values in an uncertainty quantification setting). In contrast, an optimization search takes a targeted path through the parameter space; thus, a reduced model built for optimization requires a balance between exploration and exploitation of the design space [139]. For this reason, building the reduced model in an offline phase, as discussed above, is likely inefficient in an optimization setting. Doing so may result in large regions of the parameter space being sampled in the offline phase but not exploited in the online phase, as these regions will not be explored by the optimizer. A more effective strategy in the optimization setting blends the offline and online phases by rebuilding or adapting the reduced model as the optimization search proceeds. Trust region methods are one way to achieve this [17, 88, 227].

1.2. Parametric Model Reduction and Surrogate Modeling. Parametric model reduction is but one approach within the more general area of surrogate modeling strategies for reducing computational burden in applications such as design, control, optimization, and uncertainty quantification. The paper [84] categorizes surrogate models into three different classes: data-fit models, projection-based reduced models, and hierarchical models. Data-fit models include response surface methods that use

interpolation or regression of simulation data to fit a model for the system output as a function of the parameters. In the statistical literature, Gaussian processes have been used extensively as data-fit surrogates for complex computational models [140], while data-fit surrogate approaches in optimization include polynomial response surfaces [85, 108, 137, 216], radial basis functions [224], and Kriging models [204]. Stochastic spectral approximations, commonly used in uncertainty quantification, are another form of response surface model. These methods are based on polynomial chaos representations of random variables and processes [65, 223] and exploit regularity in the dependence of an output (or solution field) on uncertain parameters. Stochastic spectral approximations have been used both in forward propagation of uncertainty [107, 226] and for statistical inverse problems [168, 169]. Hierarchical surrogate models include a range of physics-based models of reduced computational cost (and possibly lower accuracy). Hierarchical surrogates are derived from higher-fidelity models using approaches such as simplifying physics assumptions, coarser grids, alternative basis expansions, and looser residual tolerances. Examples in the literature include the use of simplified-physics models in design optimization [3, 167], multigrid approaches to optimization of systems governed by differential equations [124, 151], and mesh coarsening for solving a linear inverse problem [18].

Each of these different classes of surrogate models has different advantages and disadvantages. Simplified physics models are typically used in an opportunistic way (i.e., when they are naturally available, as is the case in a great deal of engineering practice), while data-fit and reduced models are derived mathematically from a higher-fidelity model. A significant advantage of data-fit approaches is that the offline process of deriving the surrogate model is nonintrusive. That is, the high-fidelity model can be run in "black-box" mode, where the only task is to specify a set of input parameters and generate the corresponding system output predictions. In contrast, the offline process of deriving a projection-based reduced model is fairly intrusive; as we will see in the next section, it requires projections of system operators onto a reduced subspace. One of the potential advantages of a projection-based reduced model is that it retains the underlying structure of the model. This is of particular importance for dynamical systems, where there are often benefits to retaining the notion of "system state"—meaning that our reduced model can be evolved dynamically in time (albeit in a reduced subspace, but with an explicit relationship to the original state space). Handling dynamical evolution with a data-fit model is more challenging and typically leads to a loss of flexibility in the surrogate model (e.g., a statistical data-fit model may be applicable only to aggregate outputs and/or to the specific conditions under which it was derived, while a reduced model preserves the ability to simulate different conditions within its range of applicability). Another advantage of projection-based approaches is that, using the underlying system structure, they permit a rigorous framework for deriving system-theoretic error bounds and error estimates.

The question of which surrogate modeling approach to use for a given problem does not have a definitive answer—much depends on the structure of the problem, the design/optimization/control/uncertainty quantification task at hand, the availability of alternate models, and level of access to system operators. Projection-based reduced models are known to be effective for systems where the input–output map is of low rank (or can be well approximated by a low-rank representation) [14]. In the case of a linear time-invariant (LTI) state-space system, this rank is described precisely by the *Hankel singular values* of the system; model reduction is effective for systems whose Hankel singular values exhibit rapid decay. For LTI systems that arise from discretization of PDEs, it is the physical nature of the underlying PDE operator as

well as the particular input and output quantities of interest that determine the rate of decay of the Hankel singular values of the discretized system. Diffusive processes (e.g., the heat equation) tend to be associated with rapid singular value decay, while convection-dominated problems are not. However, the particular inputs and outputs under consideration play a key role—even in the presence of strong convection, an output that corresponds to an integrated quantity (e.g., an average solution over the domain) or a highly localized quantity (e.g., the solution at a particular spatial location) can be characterized by a low-dimensional input–output map [155].

In the case of nonlinear and/or parameterized systems, the elegant and rigorous characterization of system rank is lost; however, we will see that the concept can be extended empirically in useful ways. For example, the calculation of singular values associated with an ensemble of representative system solutions ("snapshots") lends useful guidance in determining to what extent a system is amenable for reduction. Smoothness or regularity of the solution with respect to the parameters is often indicative of a problem that is reducible, although we emphasize again the important role played by the output quantity of interest—even if the state solution depends in an irregular way on the parameters, the observation equation (describing the state-to-output map) often acts as a smoothing filter. Further, in some cases, a change of coordinate system or other modification of the problem can expose low-dimensional structure, as in the projective integration method that exploits system symmetries such as translational invariance (associated with traveling solutions) and scale invariance [138].

1.3. Parametric Model Reduction in Action. A further question that relates to the amount of reduction that can be achieved is, When is model reduction worthwhile? That is, when does the upfront offline cost outweigh the benefit of having rapid online evaluations? The recent advances in parametric model reduction surveyed in this paper have made significant strides towards scalable methods that reduce the cost of the offline phase; however, once again, a definitive answer to this question remains problem and context dependent. There are some cases, such as onboard structural assessment of an aircraft wing to support real-time mission replanning [4, 166], where one is willing to tolerate an expensive offline phase requiring high-performance computing in order to have the capability to react to aircraft data acquired in real time. There are other cases, such as the design optimization setting discussed above, where it may be just as efficient (or even more so) to solve the original problem using the full model than to derive the reduced model. Throughout this paper we provide discussions of both the quantitative and qualitative aspects of parametric model reduction, as well as examples from the literature, all of which can help guide the practitioner in making this assessment for their particular problem.

We lend more insight to the question of what kinds of problems are amenable to parametric model reduction by citing some success stories from the literature.

- In [40], the authors consider thermal modeling of electric motors, specifically two models used by Robert Bosch GmbH (the world's largest supplier of automotive components[1]) for designing integrated motor generators for plug-in hybrid electric vehicles. The models discretize the heat equation on the rather complicated geometry of the engine and comprise 7 and 13 design parameters, respectively, describing engine geometry and material parameters such

[1] According to 2011 revenues; see http://en.wikipedia.org/wiki/Robert_Bosch_GmbH, accessed July 15, 2014.

as heat transfer coefficients and thermal resistances. The simulation times for these models using the parametric model reduction-via-bilinearization approach (described in section 4.1) can be accelerated by a factor of 500 and 300, respectively.

- Batch chromatography is one of the major processes used for separation problems in chemical engineering. Therefore, finding the optimal operation mode is of significant economical value. The mathematical model of batch chromatography consists of a system of coupled instationary nonlinear advection-diffusion-reaction equations. In [230], the optimization of such a process for the separation of two components is considered. The model is discretized in space using the finite volume method. Applying the reduced basis method (section 4.3) with a novel greedy snapshot selection (section 3.4) with respect to time and discrete empirical interpolation (section 2.3) to treat the nonlinearity, the optimization process could be accelerated by a factor of 54.
- In [159], the authors use proper orthogonal decomposition (POD; section 3.3) to create a reduced-order aeroelastic model of a complete F-16 configuration with clean wings. The full-order aeroelastic model comprises a finite element structural model with 168,799 degrees of freedom coupled with an Euler CFD model with more than 2 million degrees of freedom. Local POD models of dimension $r = 90$ computed at sampled Mach numbers are interpolated to yield a parametric reduced model that estimates aeroelastic damping ratio coefficients to within 10% accuracy.
- In [154], the authors consider a convection-diffusion model of contaminant transport on a complex three-dimensional domain, parameterized by the initial distribution of contaminant concentration. A POD model (section 3.3) with structure-exploiting parametric sampling reduces the dimension of the problem from more than one million states to 800 states, with no loss of accuracy in the parameter inversion estimates or in the subsequent predictions of future contaminant evolution. The speedups in computing time in this case are approximately 3,000.
- In [147], the authors use the reduced basis method (section 4.3) to create a reduced model of nonlinear viscous flows with varying Young's modulus and varying geometrical parameters, representing arterial wall shape in a fluid-structure interaction problem with a stenosis and a shape optimization process of an arterial bypass. They achieve a reduction from a full model of state dimension 35,000 to a reduced model of state dimension 20 while maintaining accuracy levels of $O(10^{-2})$ in the vorticity estimates for the shape optimization process, and a reduction from a full model of state dimension 16,000 to a reduced model of state dimension 8 while maintaining accuracy levels of $O(10^{-2})$ in the viscous energy dissipation estimates for the fluid-structure interaction simulation in presence of a stenosis in the arterial branch. The authors estimate speedups in computing time ranging from approximately 28 in the stenosis example to approximately 450 in the bypass example. These examples represents complexity in terms of dealing with multiphysics (fluid-structure interaction), shape optimization under uncertainty, nonlinear flows, and advanced geometrical parametrization.
- In [200], the authors use the reduced basis method (section 4.3) to model control rod movement for a nuclear reactor core. The neutron kinetics are described by a multigroup diffusion equation parametrized by the height of the control rod (i.e., how far the rod is inserted). The reduced model has

a computational speedup of a factor of 30,000 relative to a full model of dimension 133,810, while reproducing the neutron flux distribution with a relative accuracy of 10^{-4}.

In addition to these examples of parametric model reduction, several studies have compared various surrogate modeling approaches in the context of specific examples. For example, [32] shows that in a two-dimensional driven cavity flow example for a viscoelastic material, a projection-based reduced model with 60 degrees of freedom performs significantly better than a coarser discretization with 14,803 degrees of freedom. The paper [100] compares projection-based reduced models to stochastic spectral approximations in a statistical inverse problem setting and concludes that, for an elliptic problem with low parameter dimension, the reduced model requires fewer offline simulations to achieve a desired level of accuracy, while the polynomial chaos-based surrogate is cheaper to evaluate in the online phase. In [74] parametric reduced models are compared with Kriging models for a thermal fin design problem and for prediction of contaminant transport. For those examples, the Kriging models are found to be rapid to solve, to be implementable with a nonintrusive approach, and to provide an accurate approximation of the system output for conditions over which they were derived. The projection-based reduced models are found to be more flexible with regard to approximating system response for parameters and initial conditions other than those over which they were derived. A detailed comparison of Kriging(-type) methods and projection-based parametric model reduction methods is, to the best of our knowledge, not currently available, although [70] derives a connection between a residual-minimizing model interpolation approach and Kriging by noting a relationship between the linear system of equations solved to determine the Kriging model and the linear system of equations resulting from a constrained least squares formulation of the model reduction method. Given the common interpolatory perspective that underlies both Kriging and many parametric model reduction methods, a deeper investigation into their mathematical relationships and an extensive comparison of method performance would be a welcome addition to the literature.

1.4. Outline of the Paper. In this paper, our focus is on dynamical systems—a large class of problems for which projection-based reduced model approaches will often be a good choice, as highlighted by the examples in the previous subsection. The remainder of the paper is organized as follows: In section 2, we define the problem setup and introduce the general framework for projection-based model reduction of parameterized dynamical systems. We also discuss measures of reduced model error. Sections 3 and 4 discuss in detail the building blocks of parametric model reduction. In particular, section 3 presents three different methods for deriving the reduced-order basis: rational interpolation methods, balanced truncation, and proper orthogonal decomposition. We also discuss methods for sampling the parameter space. Section 4 covers methods for constructing the parameterized reduced model. These construction methods are divided into those that use global information over the parameter space, and those that use local information combined with an interpolation method. Section 5 provides a comparative discussion of the various projection-based model reduction approaches, along with their relative advantages and disadvantages. Finally, section 6 concludes the paper with a discussion of open challenges and future outlook.

2. General Problem Setup. In this section we define the parameterized dynamical systems of interest. We present the general projection-based model reduction framework for systems that are linear in state but have general nonlinear parametric dependence. The projection framework also applies to systems that are nonlinear in

state, but as discussed in section 6, the model reduction theory in this case is much less developed. Throughout the discussion, we indicate which aspects of the methods carry over to the nonlinear-in-state case. Section 2 concludes with a discussion of error measures to assess the quality of the reduced model.

2.1. Parameterized Dynamical Systems. We consider dynamical systems that are linear in state and parameterized with d parameters $\mathbf{p} = [p_1, \ldots, p_d]^T \in \Omega \subset \mathbb{R}^d$ (usually, Ω is a bounded domain) as

$$\begin{aligned} \mathbf{E}(\mathbf{p})\,\dot{\mathbf{x}}(t;\mathbf{p}) &= \mathbf{A}(\mathbf{p})\,\mathbf{x}(t;\mathbf{p}) + \mathbf{B}(\mathbf{p})\,\mathbf{u}(t), \\ \mathbf{y}(t;\mathbf{p}) &= \mathbf{C}(\mathbf{p})\,\mathbf{x}(t;\mathbf{p}) \end{aligned} \qquad \text{with} \quad \mathbf{x}(0;\mathbf{p}) = \mathbf{0}, \tag{2.1}$$

where $t \in [0, \infty)$. The state-vector is denoted by $\mathbf{x}(t;\mathbf{p}) \in \mathbb{R}^n$. $\mathbf{u}(t) \in \mathbb{R}^m$ and $\mathbf{y}(t;\mathbf{p}) \in \mathbb{R}^q$ denote, respectively, the inputs (excitations) and outputs (observations or measurements) of the underlying model. Hence, the model has m inputs and q outputs. The state-space matrices, then, have the dimensions $\mathbf{E}(\mathbf{p})$, $\mathbf{A}(\mathbf{p}) \in \mathbb{R}^{n\times n}$, $\mathbf{B}(\mathbf{p}) \in \mathbb{R}^{n\times m}$, and $\mathbf{C}(\mathbf{p}) \in \mathbb{R}^{q\times n}$. We focus on models that are linear in state, but we allow nonlinear parametric dependency in all system matrices. The length of the state-vector $\mathbf{x}(t;\mathbf{p})$ (i.e., n) is called the dimension of the parametric model (2.1). For ease of presentation, we will assume that for every $\mathbf{p} \in \Omega$, $\mathbf{E}(\mathbf{p})$ is nonsingular; however, most of the discussion can be extended to the singular $\mathbf{E}(\mathbf{p})$ case as discussed in section 5. We will further assume that the original model in (2.1) is asymptotically stable for every $\mathbf{p} \in \Omega$; i.e., the eigenvalues of the matrix pencil $\lambda\mathbf{E}(\mathbf{p}) - \mathbf{A}(\mathbf{p})$ have negative real parts; however, most of the methods can be applied with small modifications to unstable systems, as discussed further in section 5.

We are interested in cases where n is very large, typically exceeding hundreds of thousands. The goal is to replace the original large-scale model (2.1), sometimes called the "truth model" or the "full model," with a reduced model of the form

$$\begin{aligned} \mathbf{E}_r(\mathbf{p})\,\dot{\mathbf{x}}_r(t;\mathbf{p}) &= \mathbf{A}_r(\mathbf{p})\,\mathbf{x}_r(t;\mathbf{p}) + \mathbf{B}_r(\mathbf{p})\,\mathbf{u}(t), \\ \mathbf{y}_r(t;\mathbf{p}) &= \mathbf{C}_r(\mathbf{p})\,\mathbf{x}_r(t;\mathbf{p}), \end{aligned} \qquad \text{with} \quad \mathbf{x}_r(0;\mathbf{p}) = \mathbf{0}, \tag{2.2}$$

such that the reduced output $\mathbf{y}_r(t;\mathbf{p}) \in \mathbb{R}^q$ is a good approximation of $\mathbf{y}(t;\mathbf{p})$ with respect to an appropriate error measure. Note that the reduced state-vector $\mathbf{x}_r(t;\mathbf{p})$ has length $r \ll n$, and the reduced state-space matrices have dimensions $\mathbf{E}_r(\mathbf{p})$, $\mathbf{A}_r(\mathbf{p}) \in \mathbb{R}^{r\times r}$, $\mathbf{B}_r(\mathbf{p}) \in \mathbb{R}^{r\times m}$, and $\mathbf{C}_r(\mathbf{p}) \in \mathbb{R}^{q\times r}$; hence the dimension is reduced from n down to $r \ll n$. This reduction process is illustrated pictorially in Figure 2.1.

2.2. Projection-Based Model Reduction. Parametric model reduction can be approached from a variety of viewpoints. This paper focuses on projection-based approaches. To motivate the challenges associated with parameterized systems, first consider the general projection-based reduction approach for a system with no parametric dependence; i.e.,

$$\mathbf{E}\,\dot{\mathbf{x}}(t) = \mathbf{A}\,\mathbf{x}(t) + \mathbf{B}\,\mathbf{u}(t), \quad \mathbf{y}(t) = \mathbf{C}\,\mathbf{x}(t). \tag{2.3}$$

One approximates the unknown state quantities in a basis of reduced dimension and projects the governing equations onto a suitably defined low-dimensional subspace. In particular, choose r-dimensional test and trial subspaces, denoted by $\mathcal{V}$ and $\mathcal{W}$, respectively, where $r \ll n$. Define the associated basis matrices $\mathbf{V} \in \mathbb{R}^{n\times r}$ and $\mathbf{W} \in \mathbb{R}^{n\times r}$, where $\mathcal{V} = \mathrm{Ran}(\mathbf{V})$ and $\mathcal{W} = \mathrm{Ran}(\mathbf{W})$, with *Ran* denoting the range. Using the approximation that the full state $\mathbf{x}(t)$ evolves in the r-dimensional subspace

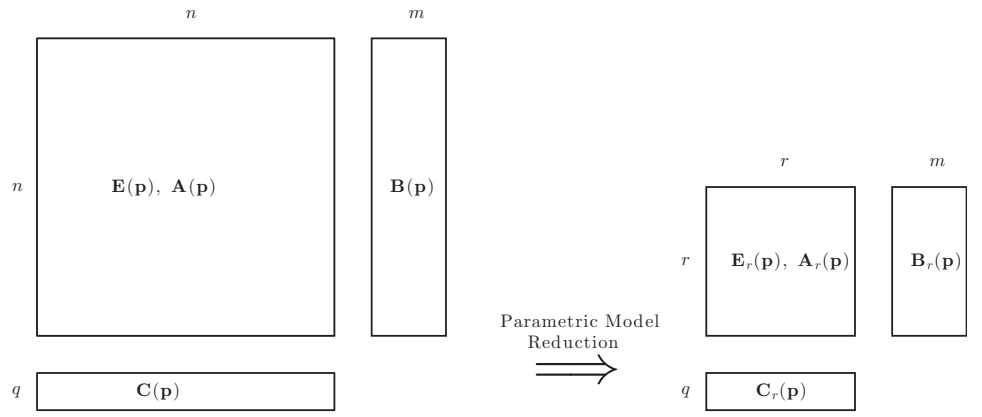

Fig. 2.1 *Parametric model reduction.*

$\mathcal{V}$, we write $\mathbf{x}(t) \approx \mathbf{V}\mathbf{x}_r(t)$, where $\mathbf{x}_r(t) \in \mathbb{R}^r$. Using this approximation in (2.3) defines a residual $(\mathbf{E}\mathbf{V}\dot{\mathbf{x}}_r(t) - \mathbf{A}\mathbf{V}\mathbf{x}_r(t) - \mathbf{B}\,\mathbf{u}(t))$ and the reduced model output $\mathbf{y}_r(t) = \mathbf{C}\mathbf{V}\mathbf{x}_r(t)$. Then enforcing the Petrov–Galerkin condition

$$\mathbf{W}^T\Big(\mathbf{E}\mathbf{V}\dot{\mathbf{x}}_r(t) - \mathbf{A}\mathbf{V}\mathbf{x}_r(t) - \mathbf{B}\,\mathbf{u}(t)\Big) = \mathbf{0} \tag{2.4}$$

leads to the reduced system

$$\mathbf{E}_r\dot{\mathbf{x}}_r(t) = \mathbf{A}_r\,\mathbf{x}_r(t) + \mathbf{B}_r\,\mathbf{u}(t), \quad \mathbf{y}_r(t) = \mathbf{C}_r\,\mathbf{x}_r(t), \tag{2.5}$$

where the reduced matrices are given by

$$\mathbf{E}_r = \mathbf{W}^T\mathbf{E}\mathbf{V}, \quad \mathbf{A}_r = \mathbf{W}^T\mathbf{A}\mathbf{V}, \quad \mathbf{B}_r = \mathbf{W}^T\mathbf{B}, \quad \text{and} \quad \mathbf{C}_r = \mathbf{C}\mathbf{V}. \tag{2.6}$$

2.3. Projection Framework for Parameterized Systems. In the case of a system with no parametric dependence, the reduced quantities in (2.6) are precomputed constant matrices, and the reduced model can be evaluated with no further reference to the full model. However, in the case of a parameterized system, the reduced model will also be parameter dependent. Several challenges arise in achieving efficient construction and evaluation of a reduced model as the parameters vary.

The first challenge is how to introduce parametric dependence into the basis matrices $\mathbf{V}$ and $\mathbf{W}$. One option is to construct "global" basis matrices over the parameter space, that is, a single matrix $\mathbf{V}$ and a single matrix $\mathbf{W}$, each of which captures parametric dependence by embedding information regarding the entire parameter space (e.g., information collected by sampling multiple parameter values). A second option is to construct "local" basis matrices. That is, consider K parameter sample points $\hat{\mathbf{p}}_1, \ldots, \hat{\mathbf{p}}_K$. For the realization of the dynamical system corresponding to $\hat{\mathbf{p}}_i$ (i.e., $\mathbf{E}(\hat{\mathbf{p}}_i)\,\dot{\mathbf{x}}(t) = \mathbf{A}(\hat{\mathbf{p}}_i)\,\mathbf{x}(t) + \mathbf{B}(\hat{\mathbf{p}}_i)\,\mathbf{u}(t)$, $\mathbf{y}(t) = \mathbf{C}(\hat{\mathbf{p}}_i)\,\mathbf{x}(t)$), the state-space matrices are constant, and one computes appropriate local basis matrices $\mathbf{V}_i$ and $\mathbf{W}_i$ (i.e., $\mathbf{V}_i$ and $\mathbf{W}_i$ denote the basis matrices corresponding to the parameter $\hat{\mathbf{p}}_i$). There are several ways one could then use these local basis matrices to construct the parametric reduced model. For example, one might interpolate the local basis matrices over the parameter space and construct a family of parameterized reduced models, or one might construct local reduced models and then interpolate the reduced models themselves. These global and local approaches are discussed in detail in section 4.

A second challenge is achieving efficient evaluations of the parametric reduced model in the case that a global basis is used. For example, for a given $\mathbf{V}$ and $\mathbf{W}$ consider evaluating $\mathbf{A}_r(\mathbf{p}) = \mathbf{W}^T\mathbf{A}(\mathbf{p})\mathbf{V}$. For general parametric dependence, $\mathbf{A}_r$ cannot be precomputed; instead, evaluating the reduced model for a new parameter value $\mathbf{p}$ requires computing $\mathbf{A}(\mathbf{p})$ and subsequent pre- and postmultiplication by $\mathbf{W}^T$ and $\mathbf{V}$, respectively. These operations all depend on the (large) dimension n of the original problem. Fortunately, in some cases the structure of the problem admits an efficient strategy. For example, consider the case of affine parameter dependence with $M+1$ terms:

$$\mathbf{A}(\mathbf{p}) = \mathbf{A}_0 + \sum_{i=1}^{M} f_i(\mathbf{p})\mathbf{A}_i, \tag{2.7}$$

where the scalar functions f_i determine the parametric dependency, which can be nonlinear, and $\mathbf{A}_i \in \mathbb{R}^{n\times n}$ for $i = 0, \ldots, M$ are parameter independent. Then the reduced matrix is given by

$$\mathbf{A}_r(\mathbf{p}) = \mathbf{W}^T\mathbf{A}(\mathbf{p})\mathbf{V} = \mathbf{W}^T\mathbf{A}_0\mathbf{V} + \sum_{i=1}^{M} f_i(\mathbf{p})\mathbf{W}^T\mathbf{A}_i\mathbf{V}. \tag{2.8}$$

For affine parametric dependence in the other matrices $\mathbf{E}(\mathbf{p})$, $\mathbf{B}(\mathbf{p})$, and $\mathbf{C}(\mathbf{p})$, analogous expressions can be derived for the reduced-order counterparts. The two most important advantages of an affine parameterization are clear from the structure of the reduced model: First, once the basis matrices $\mathbf{V}$ and $\mathbf{W}$ are chosen, the component reduced-order matrices (e.g., $\mathbf{W}^T\mathbf{A}_i\mathbf{V}$, $i = 0, \ldots, M$) can be precomputed in the offline phase. Hence, the reduced model for a given $\mathbf{p}$ can be constructed without referring back to the original system, thus having a small online cost. Second, the reduced model has the same parametric structure as the original one, which may be appealing to designers who work with these models. Note that the affine representation in (2.7) always holds for any $\mathbf{A}(\mathbf{p})$ by letting $M = n^2$ and choosing $\mathbf{A}_i$ as matrices with only one nonzero entry. However, for the affine representation to have the computational advantages discussed above, one needs $M \ll n^2$ and explicit expressions for $f_i(\mathbf{p})$, $i = 1, \ldots, M$.

For the more general case where the parametric dependence is nonaffine, typically one must introduce an approximation strategy in order to avoid costly $\mathcal{O}(n)$ evaluations in forming the reduced matrices for each different parameter value—though some parametric model reduction methods do not require affine parameter dependence; see section 4. In some cases, the low-order terms of a Taylor series expansion provide a suitable approximate affine decomposition of the system matrices [62, 112, 218]. A more general approach that has been used successfully for nonlinear model reduction is selective sampling of the nonlinear terms combined with interpolation among these samples to recover an approximate nonlinear evaluation. Among this class of methods, the missing point estimation [19] and Gauss Newton with approximated tensors (GNAT) [66] methods both build upon the gappy POD interpolation method [87]; the empirical interpolation method (EIM) [26] and its discrete variant, the discrete empirical interpolation method (DEIM) [67], conduct interpolation on a low-dimensional basis for the nonlinear term. The EIM has been recently extended to the case where $\mathbf{A}(\mathbf{p})$ represents a PDE operator [77].

In our finite-dimensional state-space framework as in (2.1), the fast evaluation of $\mathbf{W}^T\mathbf{A}(\mathbf{p})\mathbf{V}$ can be explained by exploiting how DEIM is used in handling nonlinearities for model reduction of finite-dimensional nonlinear dynamical systems $\dot{\mathbf{x}}(t) =$

$\mathbf{F}(\mathbf{x}(t))$, where $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n$ is a nonlinear mapping. Let $\mathbf{a}(\mathbf{p}) = \mathrm{vec}(\mathbf{A}(\mathbf{p})) \in \mathbb{R}^{n^2}$ denote the vector obtained by stacking the columns of $\mathbf{A}(\mathbf{p})$.[2] Similarly, define

$$\mathbf{a}_r(\mathbf{p}) = \mathrm{vec}(\mathbf{A}_r(\mathbf{p})) = \mathrm{vec}(\mathbf{W}^T\mathbf{A}(\mathbf{p})\mathbf{V}) = (\mathbf{V}^T \otimes \mathbf{W}^T)\mathbf{a}(\mathbf{p}) \in \mathbb{R}^{r^2}, \tag{2.9}$$

where $\otimes$ denotes the Kronecker product. The goal is to approximate $\mathbf{a}(\mathbf{p})$ as $\mathbf{a}(\mathbf{p}) \approx \tilde{\mathbf{a}}(\mathbf{p}) = \boldsymbol{\Phi}\boldsymbol{\alpha}(\mathbf{p})$, using M degrees of freedom, where $\boldsymbol{\Phi} \in \mathbb{R}^{n^2 \times M}$ is constant and $\boldsymbol{\alpha}(\mathbf{p}) \in \mathbb{R}^M$. In this way, $\mathbf{a}_r(\mathbf{p})$ can be approximately computed, independently of n, using

$$\mathbf{a}_r(\mathbf{p}) = (\mathbf{V}^T \otimes \mathbf{W}^T)\mathbf{a}(\mathbf{p}) \approx (\mathbf{V}^T \otimes \mathbf{W}^T)\tilde{\mathbf{a}}(\mathbf{p}) = (\mathbf{V}^T \otimes \mathbf{W}^T)\boldsymbol{\Phi}\boldsymbol{\alpha}(\mathbf{p}) = \tilde{\mathbf{a}}_r(\mathbf{p}). \tag{2.10}$$

In (2.10), $(\mathbf{V}^T \otimes \mathbf{W}^T)\boldsymbol{\Phi}$ can be precomputed, and only M evaluations appearing in $\boldsymbol{\alpha}(\mathbf{p})$ need computing. One can interpret $\tilde{\mathbf{a}}(\mathbf{p})$ as $\tilde{\mathbf{a}}(\mathbf{p}) = \mathrm{vec}(\tilde{\mathbf{A}}(\mathbf{p}))$ such that $\tilde{\mathbf{A}}(\mathbf{p})$ is an approximation to $\mathbf{A}(\mathbf{p})$ that will allow faster online computation of $\mathbf{A}_r(\mathbf{p})$. The DEIM basis $\boldsymbol{\Phi}$ is constructed by sampling the coefficient matrix $\mathbf{A}(\mathbf{p})$ at parameter values $\mathbf{p}_1, \ldots, \mathbf{p}_M$. Let $\boldsymbol{\phi}_i = \mathrm{vec}(\mathbf{A}(\mathbf{p}_i))$; then, for simplicity assuming linear independence of the $\boldsymbol{\phi}_i$, the basis $\boldsymbol{\Phi}$ is given by $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1,\ \boldsymbol{\phi}_2 \ldots, \boldsymbol{\phi}_M] \in \mathbb{R}^{n^2 \times M}$. Using the DEIM algorithm, choose $\boldsymbol{\alpha}(\mathbf{p})$ to enforce that selected entries of $\tilde{\mathbf{a}}(\mathbf{p}_i)$ interpolate the corresponding entries of $\mathbf{a}(\mathbf{p}_i)$. Once mapped back to $\mathbf{A}(\mathbf{p})$ (i.e., when the vec operation is reversed), this corresponds to a selected set of entries of $\tilde{\mathbf{A}}(\mathbf{p}_i)$ exactly matching the corresponding entries of $\mathbf{A}(\mathbf{p}_i)$.

Let $z_1, z_2, \ldots, z_M$ be the indices to be exactly matched. There are a variety of methods to select these interpolation indices [26, 67, 180]. Construct the permutation matrix $\mathbf{Z} = [\mathbf{e}_{z_1}, \mathbf{e}_{z_2}, \ldots, \mathbf{e}_{z_M}] \in \mathbb{R}^{n^2 \times M}$, where $\mathbf{e}_i$ is the ith canonical vector in $\mathbb{R}^{n^2}$. Then forcing interpolation at the selected rows implies

$$\mathbf{Z}^T\mathbf{a}(\mathbf{p}) = \mathbf{Z}^T\boldsymbol{\Phi}\boldsymbol{\alpha}(\mathbf{p}) \quad \Longrightarrow \quad \boldsymbol{\alpha}(\mathbf{p}) = (\mathbf{Z}^T\boldsymbol{\Phi})^{-1}\mathbf{Z}^T\mathbf{a}(\mathbf{p}). \tag{2.11}$$

Hence, the approximation is given by

$$\tilde{\mathbf{a}}(\mathbf{p}) = \boldsymbol{\Phi}(\mathbf{Z}^T\boldsymbol{\Phi})^{-1}\mathbf{Z}^T\mathbf{a}(\mathbf{p}). \tag{2.12}$$

Note that, as discussed in [67], $\boldsymbol{\Phi}(\mathbf{Z}^T\boldsymbol{\Phi})^{-1}\mathbf{Z}^T$ in an oblique projector onto the range of $\boldsymbol{\Phi}$ and

$$\mathbf{Z}^T\tilde{\mathbf{a}}(\mathbf{p}) = \mathbf{Z}^T\mathbf{a}(\mathbf{p}), \tag{2.13}$$

i.e., the interpolation conditions are met at the required entries. Then, using (2.12) in (2.10), one obtains

$$\tilde{\mathbf{a}}_r(\mathbf{p}) = \mathrm{vec}(\tilde{\mathbf{A}}_r(\mathbf{p})) = (\mathbf{V}^T \otimes \mathbf{W}^T)\boldsymbol{\Phi}(\mathbf{Z}^T\boldsymbol{\Phi})^{-1}\mathbf{Z}^T\mathbf{a}(\mathbf{p}). \tag{2.14}$$

After reversing the vec operation, one obtains the reduced parametric coefficient matrix

$$\tilde{\mathbf{A}}_r(\mathbf{p}) = \mathbf{W}^T\Big(\sum_{i=1}^{M} \boldsymbol{\alpha}_i(\mathbf{p})\mathbf{A}(\mathbf{p}_i)\Big)\mathbf{V} = \sum_{i=1}^{M} \boldsymbol{\alpha}_i(\mathbf{p})\Big(\mathbf{W}^T\mathbf{A}(\mathbf{p}_i)\mathbf{V}\Big). \tag{2.15}$$

Note that in (2.15), the matrices $\mathbf{W}^T\mathbf{A}(\mathbf{p}_i)\mathbf{V}$ can be precomputed. Hence, for a new parameter value $\mathbf{p}$, the reduced-order matrix $\tilde{\mathbf{A}}_r(\mathbf{p})$ can be computed without any $\mathcal{O}(n)$ operations by exploiting (2.15) and (2.11). Also note that this DEIM-based

[2] Note that we define the n^2-dimensional vector $\mathbf{a}$ for simplicity of presentation, but in practice, one would never actually compute the basis representing the full matrix $\mathbf{A}$. (Indeed, in many cases of interest $\mathbf{A}$ will be sparse with many fewer than n^2 entries.)

affine approximation approach will be cost-effective in cases when one can cheaply compute only a few entries of $\mathbf{Z}^T\mathbf{a}(\mathbf{p})$, such as when $\mathbf{A}(\mathbf{p})$ is known analytically or when it is possible to compute some entries of $\mathbf{A}(\mathbf{p})$ independently of the others. Even though cheap computation of only a few entries of $\mathbf{Z}^T\mathbf{a}(\mathbf{p})$ is possible in many cases, there are exceptions such as the case of a dynamical system originating from the linearization of a nonlinear system around a steady state. For details, see [6, 158].

2.4. Error Measures. In model reduction, parametric or nonparametric, one needs to quantify the error introduced by the underlying approximation scheme. Different communities have used different, but closely related, error measures. In the POD and reduced-basis methods, one usually measures the error in time domain using either

$$\|\mathbf{y}(\cdot\,;\hat{\mathbf{p}}) - \mathbf{y}_r(\cdot\,;\hat{\mathbf{p}})\|_{L_2} = \left(\int_0^\infty \left\|\mathbf{y}(t;\hat{\mathbf{p}}) - \mathbf{y}_r(t;\hat{\mathbf{p}})\right\|_2^2 \mathrm{d}t\right)^{1/2} \tag{2.16}$$

or

$$\|\mathbf{y}(\cdot\,;\hat{\mathbf{p}}) - \mathbf{y}_r(\cdot\,;\hat{\mathbf{p}})\|_{L_\infty} = \sup_{t\geq 0} \left\|\mathbf{y}(t;\hat{\mathbf{p}}) - \mathbf{y}_r(t;\hat{\mathbf{p}})\right\|_\infty \tag{2.17}$$

for a given $\hat{\mathbf{p}}$. In the systems and control theory community, on the other hand, the concept of *transfer function* is used to analyze the accuracy of the reduced model by measuring the error in the frequency domain. Recall that $\mathbf{x}(0;\mathbf{p}) = \mathbf{0}$, and let $\mathbf{Y}(s;\mathbf{p})$ and $\mathbf{U}(s)$ denote the Laplace transforms of $\mathbf{y}(t;\mathbf{p})$ and $\mathbf{u}(t)$, respectively, where $s \in \mathbb{C}$ is the Laplace variable. Then one can take Laplace transforms of (2.1) and (2.2) to obtain

$$\mathbf{Y}(s;\mathbf{p}) = \mathbf{H}(s;\mathbf{p})\,\mathbf{U}(s) \quad \text{and} \quad \mathbf{Y}_r(s;\mathbf{p}) = \mathbf{H}_r(s;\mathbf{p})\,\mathbf{U}(s), \tag{2.18}$$

where $\mathbf{H}(s;\mathbf{p})$ and $\mathbf{H}_r(s;\mathbf{p})$ are, respectively, the (parameterized) full- and reduced-order transfer functions defined by

$$\mathbf{H}(s;\mathbf{p}) = \mathbf{C}(\mathbf{p})\,(s\,\mathbf{E}(\mathbf{p}) - \mathbf{A}(\mathbf{p}))^{-1}\,\mathbf{B}(\mathbf{p}) \tag{2.19}$$

and

$$\mathbf{H}_r(s;\mathbf{p}) = \mathbf{C}_r(\mathbf{p})\,(s\,\mathbf{E}_r(\mathbf{p}) - \mathbf{A}_r(\mathbf{p}))^{-1}\,\mathbf{B}_r(\mathbf{p}). \tag{2.20}$$

From (2.18), one can see that the output error $\mathbf{Y}(s;\mathbf{p}) - \mathbf{Y}_r(s;\mathbf{p})$ in the frequency domain is directly related to how well the reduced-order transfer function $\mathbf{H}_r(s;\mathbf{p})$ approximates $\mathbf{H}(s;\mathbf{p})$. For a given $\hat{\mathbf{p}}$, the two most common error measures are the $\mathcal{H}_\infty$ error norm, defined as

$$\left\|\mathbf{H}(\cdot\,;\hat{\mathbf{p}}) - \mathbf{H}_r(\cdot\,;\hat{\mathbf{p}})\right\|_{\mathcal{H}_\infty} = \sup_{\omega\in\mathbb{R}} \left\|\mathbf{H}(\imath\omega;\hat{\mathbf{p}}) - \mathbf{H}_r(\imath\omega;\hat{\mathbf{p}})\right\|_2, \tag{2.21}$$

and the $\mathcal{H}_2$ error norm defined as

$$\left\|\mathbf{H}(\cdot\,;\hat{\mathbf{p}}) - \mathbf{H}_r(\cdot\,;\hat{\mathbf{p}})\right\|_{\mathcal{H}_2} = \left(\frac{1}{2\pi}\int_{-\infty}^{\infty} \left\|\,\mathbf{H}(\imath\omega;\hat{\mathbf{p}}) - \mathbf{H}_r(\imath\omega;\hat{\mathbf{p}})\right\|_{\mathbf{F}}^2 \mathrm{d}\omega\right)^{1/2}, \tag{2.22}$$

where $\|\cdot\|_{\mathbf{F}}$ denotes the Frobenius form.[3] Here, recall that we assumed (2.1) to be pointwise asymptotically stable so that (2.21) and (2.22) are well-defined.

[3]Note that while the time-domain error measures (2.16) and (2.17) can be used for nonlinear dynamical systems as well, the frequency-domain measures (2.21) and (2.22) need to be modified. This is usually achieved by using the concept of Volterra series expansion; see, e.g., [36].

There is a strong connection between the frequency-domain error measures (2.21)–(2.22) and the aforementioned time-domain error measures (2.16)–(2.17); namely for a given input $\mathbf{u}(t)$ with bounded L_2 norm,

$$\|\mathbf{y}(\cdot\,;\hat{\mathbf{p}}) - \mathbf{y}_r(\cdot\,;\hat{\mathbf{p}})\|_{L_2} \leq \left\|\mathbf{H}(\cdot\,;\hat{\mathbf{p}}) - \mathbf{H}_r(\cdot\,;\hat{\mathbf{p}})\right\|_{\mathcal{H}_\infty} \|\mathbf{u}\|_{L_2} \tag{2.23}$$

and

$$\|\mathbf{y}(\cdot\,;\hat{\mathbf{p}}) - \mathbf{y}_r(\cdot\,;\hat{\mathbf{p}})\|_{L_\infty} \leq \left\|\mathbf{H}(\cdot\,;\hat{\mathbf{p}}) - \mathbf{H}_r(\cdot\,;\hat{\mathbf{p}})\right\|_{\mathcal{H}_2} \|\mathbf{u}\|_{L_2}; \tag{2.24}$$

see [14, 231] for details. Hence, there is an equivalency in measuring the error in the time domain and the frequency domain. Even though these types of error measures, which are pointwise in the parameter $\mathbf{p}$, prove effective in practice, the ultimate goal is to minimize a joint error measure defined on a composite frequency and parameter space. The work [29] introduced such a composite error measure, $\mathcal{L}_2$ error in the parameter space and $\mathcal{H}_2$ error in the frequency domain, leading to

$$\|\mathbf{H} - \mathbf{H}_r\|^2_{\mathcal{H}_2\otimes\mathcal{L}_2(\Omega)} \stackrel{\text{def}}{=} \frac{1}{2\pi}\int_{-\infty}^{+\infty}\int_{\Omega} \left\|\mathbf{H}(\imath\omega,\,\mathbf{p}) - \mathbf{H}_r(\imath\omega,\,\mathbf{p})\right\|_F^2 \,\mathrm{d}p_1 \ldots \mathrm{d}p_d \,\mathrm{d}\omega. \tag{2.25}$$

Then, for a specific parameter dependency, [29] introduced a method to minimize this error measure, as explained in section 3.4. One can similarly define a composite error measure in uniform norm in both frequency and parameter domains, namely,

$$\|\mathbf{H} - \mathbf{H}_r\|_{\mathcal{H}_\infty\otimes\mathcal{L}_\infty(\Omega)} \stackrel{\text{def}}{=} \sup_{\omega\in\mathbb{R},\ \mathbf{p}\in\Omega} \|\mathbf{H}(\imath\omega,\,\mathbf{p}) - \mathbf{H}_r(\imath\omega,\,\mathbf{p})\|_2\,. \tag{2.26}$$

The ultimate goal is to construct parametric reduced models minimizing these composite measures; however, except for special cases as considered in [29], this goal remains an open challenge.

3. Basis Computation.

This section presents three different methods for deriving the reduced basis matrices $\mathbf{V}$ and $\mathbf{W}$: rational interpolation methods, balanced truncation, and proper orthogonal decomposition (POD). The section concludes with a discussion of strategies for sampling the parameter space.

3.1. Rational Interpolation Methods.

Over the last decade, numerous parametric model reduction methods based on a rational interpolation framework have been suggested. In accordance with the historical development, we first present the idea of multivariate Padé approximation, or "multimoment matching," and then discuss the more general tangential interpolation approach. The tangential interpolation setting proposed in [29] provides a unifying framework for interpolatory projection-based model reduction of parametric systems and also paves the way to produce optimal (at least locally optimal) parametric reduced models for the composite $\mathcal{H}_2\otimes\mathcal{L}_2$ error measure.

3.1.1. Moment-Matching.

Moment-matching and Padé(-type) approximation for linear nonparametric systems determine a reduced model that satisfies the Hermite interpolation conditions

$$\frac{d^k}{ds^k}\mathbf{H}(\hat{s}) = \frac{d^k}{ds^k}\mathbf{H}_r(\hat{s}) \quad \text{for} \quad k = 0, 1, \ldots, N, \tag{3.1}$$

up to a maximal order N for $\hat{s} \in \mathbb{C}$ not a pole of $\mathbf{H}$. This yields a reduced model whose transfer function $\mathbf{H}_r(s)$ coincides in as many coefficients of its Taylor expansion (also called "moments") about $\hat{s}$ as possible for a given order of the reduced model. See, e.g., [22, 101] for a review of this approach and its close connection to the

(nonsymmetric) Lanczos process. The case $\hat{s} = 0$ is generally referred to as moment-matching, while for $\hat{s} \neq 0$ we obtain shifted moments, and $\hat{s} = \infty$ leads to matching of the Markov parameters of the full system. One can also match moments (i.e., Taylor series coefficients) around multiple expansion (interpolation) points $\hat{s}_1, \ldots, \hat{s}_k$ as opposed to a single expansion point $\hat{s}$, leading to the concept of multipoint moment-matching, also called multipoint rational interpolation. Surveys on this class of model reduction methods can be found in [22, 101]; see also [14, 15, 30, 44, 202].

The moment-matching idea can easily be extended to parametric model reduction by using multivariate Taylor expansion of $\mathbf{H}(s, \mathbf{p})$ about $(\hat{s}, \hat{\mathbf{p}})$. This has been discussed in numerous publications in the past two decades, e.g., [93, 96, 119, 153, 222] for the single parameter case, [83] for a special two-parameter case arising in structural dynamics, [89, 91, 150] for linear and polynomial parametric dependence, and [71, 120, 164, 176] for more general parametric dependence but only in some of the state-space matrices. Moment-matching/interpolation properties can be proved (see, e.g., [43, 71, 93, 120, 222]) analogously as for standard moment-matching methods such as Padé-via-Lanczos [90, 103].

3.1.2. Tangential Interpolation. In the tangential interpolation setting, the model reduction task is posed as follows: given a frequency interpolation point $\hat{s} \in \mathbb{C}$, a parameter interpolation point $\hat{\mathbf{p}} \in \mathbb{R}^d$, and a nontrivial direction vector $\hat{\mathbf{r}} \in \mathbb{C}^m$, construct a reduced parametric model via projection as in (2.6) such that $\mathbf{H}_r(s, \mathbf{p})$ interpolates $\mathbf{H}(s, \mathbf{p})$ at $(s, \mathbf{p}) = (\hat{s}, \hat{\mathbf{p}})$ along the direction $\hat{\mathbf{r}}$, i.e., $\mathbf{H}_r(\hat{s}, \hat{\mathbf{p}})\hat{\mathbf{r}} = \mathbf{H}(\hat{s}, \hat{\mathbf{p}})\hat{\mathbf{r}}$. In this case, we say $\mathbf{H}_r(s, \mathbf{p})$ tangentially interpolates $\mathbf{H}(s, \mathbf{p})$ at $(s, \mathbf{p}) = (\hat{s}, \hat{\mathbf{p}})$ along the right direction vector $\hat{\mathbf{r}}$. Similarly, if $\hat{\boldsymbol{\ell}}^T \mathbf{H}_r(\hat{s}, \hat{\mathbf{p}}) = \hat{\boldsymbol{\ell}}^T \mathbf{H}(\hat{s}, \hat{\mathbf{p}})$ for a nontrivial vector $\hat{\boldsymbol{\ell}} \in \mathbb{C}^q$, we say $\mathbf{H}_r(s, \mathbf{p})$ tangentially interpolates $\mathbf{H}(s, \mathbf{p})$ along the left direction vector $\hat{\boldsymbol{\ell}}$. Finally, if $\hat{\boldsymbol{\ell}}^T \mathbf{H}_r'(\hat{s}, \hat{\mathbf{p}})\hat{\mathbf{r}} = \hat{\boldsymbol{\ell}}^T \mathbf{H}'(\hat{s}, \hat{\mathbf{p}})\hat{\mathbf{r}}$, where $'$ denotes differentiation with respect to the frequency variable s, we say $\mathbf{H}_r(s, \mathbf{p})$ is a bitangential Hermite interpolant to $\mathbf{H}(s, \mathbf{p})$. For a discussion of projection-based tangential interpolation for *nonparametric* systems, we refer the reader to [104]. Note that if $m > 1$ or $q > 1$, i.e., if the system is not single-input single-output (SISO), the tangential interpolation framework is different from the standard interpolation in moment-matching where one enforces matrix interpolation, i.e., $\mathbf{H}(\hat{s}, \hat{\mathbf{p}}) = \mathbf{H}_r(\hat{s}, \hat{\mathbf{p}})$. This difference can prove crucial, especially for systems with a large number of inputs and outputs, as discussed in more detail below in Remark 1. Another important rationale for tangential interpolation is that it forms the necessary conditions for optimal model reduction in the $\mathcal{H}_2$ norm as shown in (3.9) below.

3.1.3. Computing the Basis for Rational Interpolation Methods. The recent work in [29] provided a unifying projection-based framework for structure-preserving, tangential interpolatory parametric model reduction, which also permitted a robust implementation. In what follows, we present the main results using the framework of [29]. As stated earlier, optimal control and optimization are two of the common applications for parametric model reduction. In these settings, when the objective functions are approximated, [2] shows that to establish convergence of the underlying optimization technique it is sufficient that the approximate models are first-order accurate, meaning that the gradient with respect to the optimization variable is matched. In our setting, this first-order accuracy corresponds to matching the gradient of the transfer function $\mathbf{H}(s, \mathbf{p})$. Therefore, we present the results only up to matching the first derivatives with respect to $\mathbf{p}$ (although higher derivatives can be matched in this framework).

The following theorem from [29] describes the conditions on the matrices $\mathbf{V}$ and $\mathbf{W}$ such that $\mathbf{H}_r(s,\mathbf{p}) = \mathbf{C}_r(\mathbf{p})\left(s\,\mathbf{E}_r(\mathbf{p}) - \mathbf{A}_r(\mathbf{p})\right)^{-1}\mathbf{B}_r(\mathbf{p})$, obtained by projection as in (2.6), is a rational tangential interpolant to $\mathbf{H}(s,\mathbf{p})$. $\hat{s} \in \mathbb{C}$ denotes an interpolation point in the frequency domain and $\hat{\mathbf{p}}$ is an interpolation point in the parameter domain.

THEOREM 3.1. *Let $\hat{s} \in \mathbb{C}$ and $\hat{\mathbf{p}} \in \mathbb{C}^d$ be chosen such that both $\hat{s}\,\mathbf{E}(\hat{\mathbf{p}}) - \mathbf{A}(\hat{\mathbf{p}})$ and $\hat{s}\,\mathbf{E}_r(\hat{\mathbf{p}}) - \mathbf{A}_r(\hat{\mathbf{p}})$ are invertible. Suppose $\hat{\mathbf{r}} \in \mathbb{C}^m$ and $\hat{\boldsymbol{\ell}} \in \mathbb{C}^q$ are two nontrivial vectors.*

(a) *If $\left(\hat{s}\,\mathbf{E}(\hat{\mathbf{p}}) - \mathbf{A}(\hat{\mathbf{p}})\right)^{-1}\mathbf{B}(\hat{\mathbf{p}})\hat{\mathbf{r}} \in \mathrm{Ran}(\mathbf{V})$, then*

$$\mathbf{H}(\hat{s},\hat{\mathbf{p}})\hat{\mathbf{r}} = \mathbf{H}_r(\hat{s},\hat{\mathbf{p}})\hat{\mathbf{r}}. \tag{3.2}$$

(b) *If $\left(\hat{\boldsymbol{\ell}}^T\mathbf{C}(\hat{\mathbf{p}})\left(\hat{s}\,\mathbf{E}(\hat{\mathbf{p}}) - \mathbf{A}(\hat{\mathbf{p}})\right)^{-1}\right)^T \in \mathrm{Ran}(\mathbf{W})$, then*

$$\hat{\boldsymbol{\ell}}^T\mathbf{H}(\hat{s},\hat{\mathbf{p}}) = \hat{\boldsymbol{\ell}}^T\mathbf{H}_r(\hat{s},\hat{\mathbf{p}}). \tag{3.3}$$

(c) *Suppose, in addition, that $\mathbf{E}(\mathbf{p})$, $\mathbf{A}(\mathbf{p})$, $\mathbf{B}(\mathbf{p})$, and $\mathbf{C}(\mathbf{p})$ are continuously differentiable in a neighborhood of $\hat{\mathbf{p}}$. If $\left(\hat{s}\,\mathbf{E}(\hat{\mathbf{p}}) - \mathbf{A}(\hat{\mathbf{p}})\right)^{-1}\mathbf{B}(\hat{\mathbf{p}})\hat{\mathbf{r}} \in \mathrm{Ran}(\mathbf{V})$ and $\left(\hat{\boldsymbol{\ell}}^T\mathbf{C}(\hat{\mathbf{p}})\left(\hat{s}\,\mathbf{E}(\hat{\mathbf{p}}) - \mathbf{A}(\hat{\mathbf{p}})\right)^{-1}\right)^T \in \mathrm{Ran}(\mathbf{W})$, then*

$$\nabla_{\mathbf{p}}\left(\hat{\boldsymbol{\ell}}^T\mathbf{H}(\hat{s},\hat{\mathbf{p}})\hat{\mathbf{r}}\right) = \nabla_{\mathbf{p}}\left(\hat{\boldsymbol{\ell}}^T\mathbf{H}_r(\hat{s},\hat{\mathbf{p}})\hat{\mathbf{r}}\right) \tag{3.4}$$

and

$$\hat{\boldsymbol{\ell}}^T\mathbf{H}'(\hat{s},\hat{\mathbf{p}})\hat{\mathbf{r}} = \hat{\boldsymbol{\ell}}^T\mathbf{H}_r'(\hat{s},\hat{\mathbf{p}})\hat{\mathbf{r}}, \tag{3.5}$$

where $'$ denotes differentiation with respect to s and $\nabla_{\mathbf{p}}$ denotes differentiation with respect to $\mathbf{p}$.

This theorem states that given $\hat{s}$, $\hat{\mathbf{p}}$, and a right direction vector $\hat{\mathbf{r}}$, adding one vector to the basis matrix $\mathbf{V}$ will satisfy the required right tangential interpolation condition (and analogously for the left direction vector $\hat{\boldsymbol{\ell}}$). The cost in each case is simply solving a (sparse) linear system, namely, $\left(\hat{s}\,\mathbf{E}(\hat{\mathbf{p}}) - \mathbf{A}(\hat{\mathbf{p}})\right)\mathbf{v} = \mathbf{B}(\hat{\mathbf{p}})\hat{\mathbf{r}}$ in the case of right tangential interpolation. This simplicity, both conceptually and with respect to implementation, is an advantage of the rational interpolation methods for parametric model reduction. As shown in Theorem 3.1, they also require basically no assumptions on system properties. The reduced model is generated quickly, in most cases requiring less time than generating a trajectory of the full-order model, making the offline phase relatively cheap.

Theorem 3.1(c) reveals an important additional fact. By adding one vector each to the bases $\mathbf{V}$ and $\mathbf{W}$ to tangentially interpolate $\mathbf{H}(s,\mathbf{p})$, one would match additionally the gradient of the transfer function $\mathbf{H}(s,\mathbf{p})$ with respect to $\mathbf{p}$ and the derivative with respect to s as shown in (3.4) and (3.5), respectively. This derivative matching is obtained for free, i.e., without any additional computation related to the gradient, and these quantities are matched without being computed; in other words no gradient with respect to $\mathbf{p}$ and no derivative with respect to s is computed in the construction of $\mathbf{V}$ or $\mathbf{W}$, yet they are still matched. This differs from the earlier work in interpolatory parametric model reduction methods, such as [55, 56, 71, 119, 120], where one-sided projection is employed, i.e., $\mathbf{W} = \mathbf{V}$. In one-sided approaches, to satisfy (3.4), gradients of $s\mathbf{E}(\mathbf{p}) - \mathbf{A}(\mathbf{p})$ with respect to $\mathbf{p}$ need to be computed and added to the basis $\mathbf{V}$. Of course, we emphasize that in applications where $\mathbf{E}(\mathbf{p})$ and $\mathbf{A}(\mathbf{p})$ are symmetric and symmetry needs to be preserved in the reduced model, one might be

restricted to the one-sided projection due to the nature of the problem. However, for the general nonsymmetric case, by taking advantage of the flexibility in choosing $\mathbf{W}$, one can obtain greater accuracy.

To illustrate Theorem 3.1 for multiple points, assume that the frequency interpolation points $\{s_i\}_{i=1}^{L} \in \mathbb{C}$ and the parameter interpolation points $\{\mathbf{p}_j\}_{j=1}^{K} \in \mathbb{R}^d$ are given together with the right tangential directions $\{\mathbf{r}_{ij}\}_{i=1,j=1}^{K,L} \subset \mathbb{C}^m$ and the left tangential directions $\{\boldsymbol{\ell}_{ij}\}_{i=1,j=1}^{K,L} \subset \mathbb{C}^q$. One may, for example, choose the tangential directions $\boldsymbol{\ell}_{ij}$ and $\mathbf{r}_{ij}$, respectively, as the leading left and right singular vectors of $\mathbf{H}(s_i, \mathbf{p}_j) \in \mathbb{C}^{q\times m}$. Below in section 3.1.4, we discuss a particular choice that results in $\mathcal{H}_2$ optimality for the parameter sample $\mathbf{p}_j$. For $i = 1, \ldots, K$ and $j = 1, \ldots, L$, define the vectors

(3.6)
$$\mathbf{v}_{ij} = \big(s_i\mathbf{E}(\mathbf{p}_j) - \mathbf{A}(\mathbf{p}_j)\big)^{-1}\mathbf{B}(\mathbf{p}_j)\mathbf{r}_{ij} \quad \text{and} \quad \mathbf{w}_{ij} = \big(s_i\mathbf{E}(\mathbf{p}_j) - \mathbf{A}(\mathbf{p}_j)\big)^{-T}\mathbf{C}(\mathbf{p}_j)^T\boldsymbol{\ell}_{ij}$$

and construct the model reduction bases $\mathbf{V}$ and $\mathbf{W}$ as follows:

$$\mathbf{V} = [\mathbf{v}_{11}, \ldots, \mathbf{v}_{1L}, \mathbf{v}_{21}, \ldots, \mathbf{v}_{2L}, \ldots, \mathbf{v}_{K1}, \ldots, \mathbf{v}_{KL}] \in \mathbb{C}^{n\times(KL)}, \tag{3.7}$$

$$\mathbf{W} = [\mathbf{w}_{11}, \ldots, \mathbf{w}_{1L}, \mathbf{w}_{21}, \ldots, \mathbf{w}_{2L}, \ldots, \mathbf{w}_{K1}, \ldots, \mathbf{w}_{KL}] \in \mathbb{C}^{n\times(KL)}. \tag{3.8}$$

Then the resulting projection-based parametric reduced model satisfies the interpolation conditions of Theorem 3.1 for every pair $(\hat{s}, \hat{\mathbf{p}}) = (s_i, \mathbf{p}_j)$ where $i = 1, \ldots, L$ and $j = 1, \ldots, K$. One can go a step further and match the Hessian (curvature) information with respect to the parameters as well. This is done in a similar way by adding additional vectors to the reduction subspaces. For example, for a given vector $\mathbf{n} \in \mathbb{C}^d$, by adding one additional vector to $\mathbf{V}$ and $\mathbf{W}$ each, one can match $\nabla^2_{\hat{\mathbf{p}}}\big(\hat{\boldsymbol{\ell}}^T\mathbf{H}(\hat{s}, \hat{\mathbf{p}})\hat{\mathbf{r}}\big)\mathbf{n}$, where $\nabla^2_{\hat{\mathbf{p}}}$ denotes the Hessian with respect to $\mathbf{p}$. For details, see [29].

REMARK 1. *Given $\hat{s} \in \mathbb{C}$, $\hat{\mathbf{p}} \in \mathbb{C}^d$, and $\hat{\mathbf{r}} \in \mathbb{C}^m$, tangential interpolation requires $\big(\hat{s}\,\mathbf{E}(\hat{\mathbf{p}}) - \mathbf{A}(\hat{\mathbf{p}})\big)^{-1}\mathbf{B}(\hat{\mathbf{p}})\hat{\mathbf{r}} \in \mathrm{Ran}(\mathbf{V})$; i.e., a single vector is added to the subspace. On the other hand, full matrix interpolation requires $\big(\hat{s}\,\mathbf{E}(\hat{\mathbf{p}}) - \mathbf{A}(\hat{\mathbf{p}})\big)^{-1}\mathbf{B}(\hat{\mathbf{p}}) \in \mathrm{Ran}(\mathbf{V})$; i.e., m new vectors are added to the subspace. If we require interpolation at, for example, L frequency points and K parameter points, full matrix interpolation can lead to* $\mathrm{Ran}(\mathbf{V})$ *having dimension as large as mLK. For applications where the system input dimension m is large, this would lead to a rather large reduced model dimension. In comparison, tangential interpolation will at most lead to a reduced dimension of LK, thus making the reduced order independent of the input dimension. Note that in the full matrix interpolation case one can keep the dimension growth modest by truncating the linearly dependent components from the model reduction bases $\mathbf{V}$ and $\mathbf{W}$ while still obtaining accurate reduced models; see, for example,* [43, 68, 93]. *Even though the subspace dimension grows more slowly in the tangential interpolation case, if KL is large in* (3.7) *and* (3.8) *due to choosing several frequency and parameter interpolation points, one might still encounter linearly dependent columns and similarly truncate these components from tangential interpolation bases as well.*

REMARK 2. *As noted above, if the system is not SISO, the tangential interpolation is different than the full-matrix interpolation, i.e., full moment-matching. Different methods proposed in these references differ in the way moments are computed (implicitly vs. explicitly) and in the number of (mixed) moments that are matched. Approaches based on explicitly computed moments suffer from the same numerical instabilities as analogous methods for model reduction of nonparametric systems* [90, 103]. *Implicit approaches appear to provide a robust resolution of these difficulties* [43, 93].

3.1.4. Optimal-$\mathcal{H}_2$ Tangential Interpolation for Nonparametric Systems. In rational interpolation methods, the choice of expansion points and tangential directions determines the accuracy of the reduced model. Until recently, this choice was largely ad hoc as the original interpolation/moment-matching framework gives no guidance on where to put the interpolation points. However, recently an interpolation-based analysis of the $\mathcal{H}_2$ error norm revealed an optimal selection strategy for expansion points and tangential directions.

Consider the nonparametric case. Suppose $\mathbf{H}_r(s)$ is an $\mathcal{H}_2$-optimal approximation to $\mathbf{H}(s) = \mathbf{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}$. Let $\widetilde{\mathbf{H}}_r(s) = \sum_{i=1}^{r} \frac{1}{s-\widetilde{\lambda}_i}\tilde{\mathbf{c}}_i\tilde{\mathbf{b}}_i^T$ be a partial fraction expansion of $\mathbf{H}_r(s)$. Note that $\tilde{\mathbf{c}}_i\tilde{\mathbf{b}}_i^T$ is the residue of $\mathbf{H}_r(s)$ at $s = \widetilde{\lambda}_i$; $\tilde{\mathbf{c}}_i$ and $\tilde{\mathbf{b}}_i^T$ are called the residue directions. Then the $\mathcal{H}_2$ optimality of $\mathbf{H}_r(s)$ means that, for $i = 1, \ldots, r$,

$$\begin{aligned} &\mathbf{H}(-\widetilde{\lambda}_i)\tilde{\mathbf{b}}_i = \widetilde{\mathbf{H}}_r(-\widetilde{\lambda}_i)\tilde{\mathbf{b}}_i, \quad \tilde{\mathbf{c}}_i^T\mathbf{H}(-\widetilde{\lambda}_i) = \tilde{\mathbf{c}}_i^T\widetilde{\mathbf{H}}_r(-\widetilde{\lambda}_i), \\ &\text{and} \quad \tilde{\mathbf{c}}_i^T\mathbf{H}'(-\widetilde{\lambda}_i)\tilde{\mathbf{b}}_i = \tilde{\mathbf{c}}_i^T\widetilde{\mathbf{H}}_r'(-\widetilde{\lambda}_i)\tilde{\mathbf{b}}_i. \end{aligned} \tag{3.9}$$

Meier and Luenberger in [173] proved these conditions for SISO systems originally; generalizations to MIMO systems were given in [64, 116, 214]. The optimality conditions in (3.9) mean that an $\mathcal{H}_2$-optimal reduced model $\mathbf{H}_r(s)$ is a bitangential Hermite interpolant to $\mathbf{H}(s)$. The optimal interpolation points are the mirror images of the poles of $\mathbf{H}_r(s)$, and the optimal tangential directions are the corresponding residue directions. Thus, the optimal points and associated tangent directions depend on the reduced model and are not known a priori. Gugercin, Antoulas, and Beattie [116] introduced the iterative rational Krylov algorithm (IRKA), which, using successive substitution, iteratively corrects the interpolation points and tangential directions until the optimality conditions in (3.9) are satisfied, i.e., until optimal interpolation points and tangential directions are reached. For details of IRKA, we refer the reader to [15, 116]. IRKA can be used in the context of parametric systems for computing locally optimal reduced models for the given parameter sample. This local information can then be used to construct a parametric reduced model using the ideas of section 4, as done in [29, 80].

3.2. Balanced Truncation. In the systems and control theory community, balanced truncation [177, 175] is one of the most common techniques for approximating linear dynamical systems without parametric dependency. In the parametric setting, balanced truncation can be employed to construct local reduced models at given parameter values. These local models can be used in various ways to construct a parametric reduced model (see, e.g., [28]), as will be discussed in section 4. Here we describe the construction of a balanced truncation reduced model at a single parameter point, $\hat{\mathbf{p}}$. For ease of notation, for the remainder of this subsection we denote $\mathbf{E}(\hat{\mathbf{p}})$ by $\mathbf{E}$, and similarly for the other matrices, so that $\mathbf{H}(s) = \mathbf{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}$ is our dynamical system (2.1) evaluated at $\hat{\mathbf{p}}$.

The balanced truncation basis matrices $\mathbf{V}$ and $\mathbf{W}$ depend on the two system Gramians, which are defined by the integrals

$$\mathcal{P} = \int_0^\infty e^{\mathbf{E}^{-1}\mathbf{A}t}\,\mathbf{E}^{-1}\mathbf{B}\mathbf{B}^T\mathbf{E}^{-T}\,e^{\mathbf{A}^T\mathbf{E}^{-T}t}\,\mathrm{d}t \tag{3.10}$$

and

$$\mathcal{Q} = \int_0^\infty \mathbf{E}^{-T}e^{\mathbf{A}^T\mathbf{E}^{-T}t}\,\mathbf{C}^T\mathbf{C}\,e^{\mathbf{E}^{-1}\mathbf{A}t}\mathbf{E}^{-1}\,\mathrm{d}t. \tag{3.11}$$

Here, $\mathcal{P} \in \mathbb{R}^{n\times n}$ and $\mathcal{Q} \in \mathbb{R}^{n\times n}$ are called *the reachability Gramian* and *the observability Gramian*, respectively. The reachability of a state $\mathbf{x}$ is a measure of how easy it is to reach the state $\mathbf{x}$ from the zero state. On the other hand, the observability of a state $\mathbf{x}_0$ is a measure of how easy it is to distinguish the initial state $\mathbf{x}_0$ from the zero state by observing the output $\mathbf{y}(t)$ in the case of zero input.

To explain these concepts further, define

$$\mathcal{J}_r(\mathbf{x}) := \min_{\mathbf{x}(-\infty)=\mathbf{0},\ \mathbf{x}(0)=\mathbf{x}} \|\mathbf{u}(t)\|^2, \quad t \le 0, \tag{3.12}$$

$$\mathcal{J}_o(\mathbf{x}_0) := \|\mathbf{y}(t)\|^2, \quad \mathbf{x}(0) = \mathbf{x}_0,\ \mathbf{u}(t) = \mathbf{0},\ t \ge 0. \tag{3.13}$$

$\mathcal{J}_r(\mathbf{x})$ is the minimal energy required to drive the system from the zero state at $t = -\infty$ to the state $\mathbf{x}$ at $t = 0$. On the other hand, $\mathcal{J}_o(\mathbf{x}_0)$ is the energy observed at the output due to the initial state $\mathbf{x}_o$ with zero input. The smaller the reachability energy $\mathcal{J}_r(\mathbf{x})$, the easier it is to reach the state $\mathbf{x}$. The larger the observability energy $\mathcal{J}_o(\mathbf{x}_0)$, the easier it is to observe the state $\mathbf{x}_0$. These two energies are completely determined by the Gramians $\mathcal{P}$ and $\mathcal{Q}$:

$$\mathcal{J}_r(\mathbf{x}) = \mathbf{x}^T\mathcal{P}^+\mathbf{x} \qquad \text{and} \qquad \mathcal{J}_r(\mathbf{x}_0) = \mathbf{x}_0^T\mathcal{Q}\,x_0, \tag{3.14}$$

where $\mathcal{P}^+$ denotes the Moore–Penrose pseudoinverse of $\mathcal{P}$. Thus, $\mathcal{P}$ and $\mathcal{Q}$ explain how important a state $\mathbf{x}$ is for the input-to-state and the state-to-output mappings, respectively.

When $\mathbf{H}(s)$ is asymptotically stable, both $\mathcal{P}$ and $\mathcal{Q}$ are positive semidefinite matrices. Square roots of the eigenvalues of the product $\mathcal{P}\mathcal{Q}$ are called the Hankel singular values of $\mathbf{H}(s)$, denoted by $\eta_i(\mathbf{H})$, and they are the singular values of the Hankel operator associated with $\mathbf{H}(s)$. The states corresponding to the smallest Hankel singular values are the least important states in terms of the input-to-output map; that is, these are states that are hard to reach and hard to observe. Model reduction via balanced truncation corresponds to eliminating those states corresponding to small Hankel singular values.

In practice one does not need to evaluate the infinite integrals (3.10) and (3.11); instead one solves the corresponding Lyapunov equations that $\mathcal{P}$ and $\mathcal{Q}$ satisfy, namely,

$$\mathbf{A}\mathcal{P}\mathbf{E}^T + \mathbf{E}\mathcal{P}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T = \mathbf{0} \quad \text{and} \quad \mathbf{A}^T\mathcal{Q}\mathbf{E} + \mathbf{E}^T\mathcal{Q}\mathbf{A} + \mathbf{C}^T\mathbf{C} = \mathbf{0}. \tag{3.15}$$

Instead of forming the full Gramians $\mathcal{P}$ and $\mathcal{Q}$ explicitly, one computes $\mathcal{P} = \mathbf{U}\mathbf{U}^T$ and $\mathcal{Q} = \mathbf{L}\mathbf{L}^T$ in the factored form. For effective methods to solve large-scale Lyapunov equations, see, e.g., [46, 47, 152, 219] for ADI-type methods, [117, 191, 199, 206] for the Smith method and its variants, [78, 135, 203, 210] for Krylov-based methods, or the recent survey [48] and the references therein.

Given the factorizations $\mathcal{P} = \mathbf{U}\mathbf{U}^T$ and $\mathcal{Q} = \mathbf{L}\mathbf{L}^T$, let $\mathbf{U}^T\mathbf{E}\mathbf{L} = \mathbf{Z}\mathbf{S}\mathbf{Y}^T$ be the singular value decomposition with $\mathbf{S} = \operatorname{diag}(\eta_1, \eta_2, \ldots, \eta_n)$. Let $\mathbf{S}_r = \operatorname{diag}(\eta_1, \eta_2, \ldots, \eta_r)$ with $\eta_{r+1} < \eta_r$ and $r < n$. Balanced truncation chooses the basis matrices

$$\mathbf{V} = \mathbf{U}\mathbf{Z}_r\mathbf{S}_r^{-1/2} \qquad \text{and} \qquad \mathbf{W} = \mathbf{L}\mathbf{Y}_r\mathbf{S}_r^{-1/2}, \tag{3.16}$$

where $\mathbf{Z}_r$ and $\mathbf{Y}_r$ denote the leading r columns of left singular vectors, $\mathbf{Z}$, and right singular vectors, $\mathbf{Y}$, respectively. Then the reduced model is obtained by following the projection in (2.6). The reduced model $\mathbf{H}_r(s)$ obtained by balanced truncation is asymptotically stable, and the $\mathcal{H}_\infty$ norm of the error system satisfies $\|\mathbf{H} - \mathbf{H}_r\|_{\mathcal{H}_\infty} \le 2(\eta_{r+1} + \cdots + \eta_n)$. For more details on balanced truncation, see [14].

3.3. Proper Orthogonal Decomposition. Due to its broad applicability to linear and nonlinear systems, the proper orthogonal decomposition (POD) has become widely used in many different application domains as a method for computing the reduced basis. While the rational interpolation methods of section 3.1 formulate the basis computation task in the frequency domain, POD formulations typically use the time domain. In the case of LTI systems, duality between time and frequency domain formulations reveals the connections between POD and balanced truncation.

3.3.1. Time Domain POD. POD was introduced for the analysis of turbulent flows by Lumley [162] and is closely related to methods used in other fields such as Karhunen–Loève expansions in stochastic process modeling [160, 142], principal component analysis in statistical analysis [132, 136], and empirical orthogonal eigenfunctions in atmospheric modeling [181]. POD basis vectors are computed empirically using sampled data collected over a range of relevant system dynamics, typically using the method of snapshots, introduced by Sirovich [205].

Consider a set of snapshots, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_s}$: state solutions computed at different instants in time and/or different parameter values, where $\mathbf{x}_j \in \mathbb{R}^n$ denotes the jth snapshot and one collects a total of $n_s < n$ snapshots. More specifically, write $\mathbf{x}_j = \mathbf{x}(t_j; \mathbf{p}_j)$, where t_j and $\mathbf{p}_j$ are, respectively, the time and parameter values for the jth snapshot. Define the snapshot matrix $\mathbf{X} \in \mathbb{R}^{n \times n_s}$ whose jth column is the snapshot $\mathbf{x}_j$. The (thin) singular value decomposition of $\mathbf{X}$ is written

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{Y}^T, \tag{3.17}$$

where the columns of the matrices $\mathbf{U} \in \mathbb{R}^{n \times n_s}$ and $\mathbf{Y} \in \mathbb{R}^{n_s \times n_s}$ are the left and right singular vectors of $\mathbf{X}$, respectively. $\mathbf{\Sigma} \in \mathbb{R}^{n_s \times n_s} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_{n_s})$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{n_s} \geq 0$, are the singular values of $\mathbf{X}$. The POD basis, $\mathbf{V}$, is chosen as the r left singular vectors of $\mathbf{X}$ that correspond to the r largest singular values. This yields an orthonormal basis.[4]

The POD basis is "optimal" in the sense that, for an orthonormal basis of size r, it minimizes the least squares error of snapshot reconstruction,

$$\min_{\mathbf{V} \in \mathbb{R}^{n \times r}} ||\mathbf{X} - \mathbf{V}\mathbf{V}^T\mathbf{X}||_F^2 = \min_{\mathbf{V} \in \mathbb{R}^{n \times r}} \sum_{i=1}^{n_s} ||\mathbf{x}_i - \mathbf{V}\mathbf{V}^T\mathbf{x}_i||_2^2 = \sum_{i=r+1}^{n_s} \sigma_i^2. \tag{3.18}$$

As can be seen from (3.18), the square of the error in snapshot representation is given by the sum of the squares of the singular values corresponding to those left singular vectors not included in the POD basis. Thus, the singular values provide quantitative guidance for choosing the size of the POD basis. A typical approach is to choose r so that

$$\frac{\sum_{i=1}^{r} \sigma_i^2}{\sum_{i=1}^{n_s} \sigma_i^2} > \kappa, \tag{3.19}$$

where κ is a user-specified tolerance, often taken to be 99.9% or greater. The numerator of (3.19) is often referred to as the "energy" captured by the POD modes.

Since the POD basis is constructed from sampled solutions, the POD method makes no assumptions about the form of the full model; POD applies to both linear

[4] Here, orthonormality is defined with respect to the Euclidean inner product. Often, it is meaningful to use the L_2 inner product instead, i.e., using the vector norm induced by the weighted inner product related to the mass matrix E. This holds in particular if the system is derived from a semidiscretization of a PDE by a finite element or finite volume method [49].

and nonlinear systems, as well as to parametrically varying systems. One can also include sensitivity information in the snapshot set [127, 131]. It is important to note that the optimality of the POD basis applies only to error in reconstruction of the snapshots, not to the error in solution of the POD-based reduced model. See [194] for a detailed analysis of the error of a POD-based reduced model. Clearly, the choice of snapshots is critical to the quality of the reduced model, although the POD theory per se gives no guidance on how to select the snapshots. One can interpret the POD as an approximation by quadrature of the reachability Gramian (3.10), as described in [146] and in more detail in the discussion of the balanced POD method below. Using this interpretation, a good snapshot set selection is thus one that leads to a good approximation of (3.10). However, in many cases the POD is used to create a reduced model that targets a particular range of system behavior, rather than attempting to approximate all reachable states. In those cases, the snapshot set is chosen based on knowledge of the desired range of validity of the reduced model. Optimal snapshot selection for nonparametric POD is considered in [145], where the time locations of snapshots are chosen to minimize the error between the POD solution and the trajectory of the original dynamical system. General strategies for snapshot selection in the parametric case are discussed in section 3.4.

3.3.2. Frequency Domain POD. For linear systems, one can derive the POD in the frequency domain. We present a brief discussion here because it highlights the connection between POD and balanced truncation. The POD basis vectors are the left singular vectors of the snapshot matrix $\mathbf{X}$, and thus are the eigenvectors of the matrix

$$\mathbf{K}_t = \mathbf{X}\mathbf{X}^T = \sum_{i=1}^{n_s} \mathbf{x}_i \mathbf{x}_i^T. \tag{3.20}$$

In the original formulation of the POD method of snapshots, $\mathbf{K}_t$ is referred to as the kernel [205]. In some applications of POD, the snapshots are centered to have zero mean, in which case $\mathbf{K}_t$ is a scaled covariance matrix.

Kim [141] develops the frequency domain POD method by showing that through a simple application of Parseval's theorem, in the single-input case one can write the kernel as

$$\mathbf{K}_\omega = \frac{1}{2\pi} \sum_{i=1}^{n_s} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^* \Delta\omega_i, \tag{3.21}$$

where now $\mathbf{K}_\omega$ is the frequency domain POD kernel. Here, $\bar{\mathbf{x}}_i$ is the ith (complex) snapshot computed at sample frequency ω_i,

$$\bar{\mathbf{x}}_i = (j\omega_i \mathbf{E} - \mathbf{A})^{-1}\mathbf{B}, \tag{3.22}$$

where $\mathbf{B}$ is now a vector (due to the assumption of a single input), $\Delta\omega_i = \omega_i - \omega_{i-1}$, and $*$ denotes the complex conjugate transpose.

3.3.3. Balanced POD. The connection between POD and balanced truncation is described in [146]. For a fixed parameter $\mathbf{p}$ and single input, consider the snapshot set generated by computing the impulse response of (2.1). In this case, the snapshot at time t_i is

$$\mathbf{x}_i = e^{\mathbf{E}^{-1}\mathbf{A}t_i}\mathbf{E}^{-1}\mathbf{B}, \tag{3.23}$$

and the POD kernel (3.20) becomes

$$\mathbf{K}_t = \sum_{i=1}^{n_s} e^{\mathbf{E}^{-1}\mathbf{A}t_i}\mathbf{E}^{-1}\mathbf{B}\mathbf{B}^T\mathbf{E}^{-T}e^{\mathbf{A}^T\mathbf{E}^{-T}t_i}. \tag{3.24}$$

Comparing (3.24) and (3.10), it can be seen that the POD kernel (with suitable scaling) computed from the system impulse response can be interpreted as an approximation to the system reachability Gramian. The POD basis vectors are the dominant eigenvectors of the POD kernel and thus approximate the most reachable modes in the system.

As discussed in [225], the approximation introduced by the POD is further highlighted through the frequency domain formulation. Using (3.22) in (3.21), write

$$\mathbf{K}_\omega = \frac{1}{2\pi}\sum_{i=1}^{n_s}(j\omega_i\mathbf{E}-\mathbf{A})^{-1}\mathbf{B}\mathbf{B}^T(-j\omega_i\mathbf{E}^T-\mathbf{A}^T)^{-1}\Delta\omega_i, \tag{3.25}$$

which shows that the frequency domain POD kernel again approximates the integral defining the reachability Gramian,

$$\mathcal{P} = \frac{1}{2\pi}\int_{-\infty}^{\infty}(j\omega\mathbf{E}-\mathbf{A})^{-1}\mathbf{B}\mathbf{B}^T(-j\omega\mathbf{E}^T-\mathbf{A}^T)^{-1}\,d\omega. \tag{3.26}$$

Based on this observation, [225] proposed an approximate balanced truncation approach using the POD method of snapshots. Computing a set of dual POD modes from snapshots of the dual (adjoint) system leads to an approximation of the most observable modes in the system. For a system with a single output, the ith (complex) dual snapshot $\bar{\mathbf{z}}_i$ computed at sample frequency ω_i is given by

$$\bar{\mathbf{z}}_i = (j\omega_i\mathbf{E}^T-\mathbf{A}^T)^{-1}\mathbf{C}^T. \tag{3.27}$$

Appropriately combining the primal and dual snapshots leads to an approximate balanced truncation. A modified algorithm for computing this snapshot-based approximate balanced truncation was proposed in [196]. Despite the analogy between the two methods, there is no asymptotic stability guarantee associated with balanced POD, unlike balanced truncation. The relation between frequency-domain POD and approximate balanced truncation is further discussed in [30], where also relations to balanced truncation via low-rank Gramians computed by the ADI method and to moment-matching methods are studied.

3.4. Parameter Sampling. We conclude this section on basis computation with a discussion of parameter sampling. The choice of parameter sample points is a critical question that arises in all methods to compute the basis. In the rational interpolation methods, one must select parameter samples at which interpolation conditions are applied; for balanced truncation, one must select parameter samples to generate the local LTI systems at which balanced truncation is applied; and in the POD, one must select parameter samples at which snapshots are computed. For problems with a small number of parameters, a structured or random sampling method (e.g., grid-based sampling or Latin hypercube sampling) is the simplest approach and, with a sufficiently high number of samples, will generate a rich set of data that covers the parameter space. For a moderate number of parameters, full grid sampling quickly becomes expensive, since the number of points in the grid grows exponentially with the dimension d. Latin hypercube sampling remains tractable, although a large number of sample points may be needed to ensure sufficient coverage. For moderate-dimension

problems that exhibit structure (e.g., smoothness), a sparse grid sampling approach will likely be another effective strategy. However, when the dimension of the parameter space is large ($d > 10$), it becomes challenging to balance sampling cost with coverage of the parameter space. These cases require more sophisticated sampling approaches, such as a problem-aware adaptive search of the parameter space. In this section, we review adaptive sampling via greedy search and local sensitivity methods. We also discuss optimal interpolation points for a special class of systems.

3.4.1. Adaptive Parameter Sampling via Greedy Search. Greedy sampling methods for model reduction approach the task of choosing parameter sample points one by one in an adaptive manner. The general steps in the greedy sampling approach are as follows. First, given a current reduced model,

$$\mathbf{E}_r(\mathbf{p})\,\dot{\mathbf{x}}_r(t;\mathbf{p}) = \mathbf{A}_r(\mathbf{p})\,\mathbf{x}_r(t;\mathbf{p}) + \mathbf{B}_r(\mathbf{p})\,\mathbf{u}(t), \quad \mathbf{y}_r(t;\mathbf{p}) = \mathbf{C}_r(\mathbf{p})\,\mathbf{x}_r(t;\mathbf{p}), \tag{3.28}$$

find the parameter value for which the error between the reduced model and the full model is largest:

$$\hat{\mathbf{p}} = \arg\max_{\mathbf{p}} \|\mathbf{y}(\cdot\,;\mathbf{p}) - \mathbf{y}_r(\cdot\,;\mathbf{p})\|_{L_2}. \tag{3.29}$$

Second, solve the full model at $\hat{\mathbf{p}}$ to generate new information with which to update the reduced model. Then with the updated reduced model, repeat these two steps until the error is acceptable. The greedy sampling method was first introduced in the context of reduced basis methods in [193]. It was further developed in reduced basis models for elliptic PDEs in [218], the steady incompressible Navier–Stokes equations in [217], and parabolic PDEs in [111, 112]. It has since been applied in conjunction with POD methods [61, 122, 230] and rational interpolation methods [79]. The key advantage of the greedy approach is that the search over the parameter space embodies the structure of the problem, so that the underlying system dynamics guide the selection of appropriate parameter samples.

In the first step, the task of finding the worst-case parameter value can be approached in a variety of ways. In the general case, using the actual reduced model error as a metric leads to a computationally intractable algorithm, since it requires evaluating the full model solution at many parameter points. Instead, one can use a posteriori error estimators if they are available for the given system under consideration (e.g., error estimators are available for dynamical systems arising from discretization of particular parameterized parabolic PDEs [112, 111]) or a residual-based error indicator [61]. The process of locating the parameter of maximal error indicator or error estimate can be conducted by a simple grid search if the parameter dimension is low. For problems with more than a handful of parameters, the greedy sampling approach can be formulated as a sequence of adaptive model-constrained optimization problems [61]. In the special case of a linear map between parameters and outputs, these optimization problems are convex and have explicit solutions in the form of an eigenvalue problem [27]. In the general case, the optimization problems are nonconvex and may lead to only a local maximum being determined. Another potential pitfall is that error estimators may not be tight and error indicators may be poor representatives of the actual error; in such cases, solving an approximate optimization problem using the error estimator or indicator will not result in maximizing (3.29). However, even if the global maximum of (3.29) is not found, the greedy search will often yield a good result that is more informed than random sampling. In [54], exponential convergence rates of greedy sampling are shown in a reduced basis framework,

both for strong greedy (the optimal next sampling point is always found) as well as in computationally feasible variants. In [156], the optimization-based greedy sampling approach was extended to construct both a basis for the state and a basis for the parameter, leading to models that have both reduced state and reduced parameters. This approach was demonstrated on a subsurface model with a distributed parameter representing the hydraulic conductivity over the domain, and shown to scale up to the case of a discretized parameter vector of dimension $d = 494$.

3.4.2. Adaptive Parameter Sampling via Local Sensitivity Analysis. Another possible method of selecting the parameter values for sampling, proposed in [55], is to use local sensitivity analysis to estimate whether a change in parameter value will result in states that are not well represented in the current reduced basis. That work uses a first-order Taylor series expansion to approximate the state solution as a function of small changes in parameters about the current sample point. The sensitivities of the state variables with respect to the parameters are then obtained by solving a large sparse linear system. When used in conjunction with the trajectory piecewise linear approximation of [55], this auxiliary sensitivity system can be assembled with no additional computational cost, using information already computed to create the reduced model. This approach was demonstrated for a diode circuit, a MEMS switch, and a nonlinear distributed circuit example.

3.4.3. Optimal Interpolation Points for a Special Class of Systems. Even though the parameter selection strategies outlined above have led to high quality reduced parametric models, they are not optimal in the error measures defined in section 2.4. Such an optimal parameter selection strategy for the joint $\mathcal{H}_2 \otimes \mathcal{L}_2$ error measure (2.25) was recently introduced by Baur et al. [29] for the special case of dynamical systems with affine parameter dependency in $\mathbf{B}(\mathbf{p})$ and $\mathbf{C}(\mathbf{p})$ and *no parameter dependency* in either $\mathbf{A}$ or $\mathbf{E}$. This method, inspired by IRKA of [116] for nonparametric systems, converts the problem into an equivalent nonparametric $\mathcal{H}_2$ minimization problem and optimizes over the frequency and parameter samples jointly.

4. Parameterized Reduced Model Generation. In this section, we discuss different strategies for constructing the parameterized reduced model. Many of these construction options are broadly applicable in that they can be used with any of the basis computation methods discussed in section 3. Section 4.1 discusses approaches that use a single global basis, while section 4.2 presents methods that use multiple local basis matrices each computed at different parameter values. In the following discussion, the parameter $\mathbf{p}$ belongs to a single domain $\Omega \subset \mathbb{R}^d$. Recent work has proposed approaches to split Ω into multiple subdomains and construct reduced models in each subdomain [10, 76, 82, 121, 190, 221]. Any of the model generation strategies described in the following can be applied in a partitioning setting by replacing Ω with the corresponding subdomain.

4.1. A Global Basis over the Parameter Space. A global basis is defined by a single pair of basis matrices $\mathbf{V}$ and $\mathbf{W}$, which are built by sampling information over a range of parameters. These basis matrices could be computed using any one of the methods described in section 3. In this case, the parametric reduced model takes the form

$$\begin{aligned} \mathbf{W}^T\mathbf{E}(\mathbf{p})\mathbf{V}\,\dot{\mathbf{x}}_r(t;\mathbf{p}) &= \mathbf{W}^T\mathbf{A}(\mathbf{p})\mathbf{V}\,\mathbf{x}_r(t;\mathbf{p}) + \mathbf{W}^T\mathbf{B}(\mathbf{p})\,\mathbf{u}(t), \\ \mathbf{y}_r(t;\mathbf{p}) &= \mathbf{C}(\mathbf{p})\mathbf{V}\,\mathbf{x}_r(t;\mathbf{p}). \end{aligned} \tag{4.1}$$

Equation (4.1) shows that construction of the reduced model requires evaluation of terms such as $\mathbf{A}_r(\mathbf{p}) = \mathbf{W}^T\mathbf{A}(\mathbf{p})\mathbf{V}$, $\mathbf{E}_r(\mathbf{p}) = \mathbf{W}^T\mathbf{E}(\mathbf{p})\mathbf{V}$, etc. As already discussed in section 2, if one evaluates $\mathbf{A}_r$, $\mathbf{E}_r$, and the other reduced matrices in this way for every new parameter value $\mathbf{p}$, then the reduced model will be inefficient and computational savings over solving the original full system will be small. Section 2 showed that for system matrices that depend affinely on the parameter $\mathbf{p}$, computational efficiency can be recovered. In this case, the parameterized reduced model is decomposed into reduced-order matrices that do not depend on the parameter and hence can be precomputed. In the more general case, a method such as EIM or DEIM (also discussed in section 2) is used to approximate $\mathbf{A}_r(\mathbf{p})$ and the other reduced-order matrices in a way that admits efficient evaluation over the parameter space.

4.1.1. Concatenation of the Basis. One of the common approaches to obtain the global basis matrices $\mathbf{V}$ and $\mathbf{W}$ is to concatenate the local basis matrices obtained for several parameter samples $\mathbf{p}_1, \ldots, \mathbf{p}_K$. Suppose that $\mathbf{V}_1, \ldots, \mathbf{V}_K$ and $\mathbf{W}_1, \ldots, \mathbf{W}_K$ denote the local basis matrices corresponding to $\mathbf{p}_1, \ldots, \mathbf{p}_K$. Then one can construct the global basis matrices $\mathbf{V}$ and $\mathbf{W}$ using

$$\mathbf{V} = [\mathbf{V}_1, \ \mathbf{V}_2, \ \ldots, \ \mathbf{V}_K] \ \text{ and } \ \mathbf{W} = [\mathbf{W}_1, \ \mathbf{W}_2, \ \ldots, \ \mathbf{W}_K]. \tag{4.2}$$

However, it is quite possible that the local matrices have common components among each other, leading to potentially rank-deficient global basis matrices $\mathbf{V}$ and $\mathbf{W}$. To avoid this scenario, the concatenation step is usually followed by an SVD or a rank-revealing QR factorization to remove these rank-deficient components from $\mathbf{V}$ and $\mathbf{W}$, leading to global basis matrices with orthonormal columns. It is important to note that even though theoretically it would not matter whether the local matrices $\mathbf{V}_i$ and $\mathbf{W}_i$ are orthogonalized prior to the concatenation step (since the reduced model is determined by the range, not by a specific basis), numerically it might (e.g., as in the case of Krylov subspaces where the power basis $[\mathbf{b}, \mathbf{A}\mathbf{b}, \ldots, \mathbf{A}^{k-1}\mathbf{b}]$ is replaced by an orthogonal basis robustly computed via an Arnoldi process [157, 212]).

The local basis matrices can be obtained using any one of the methods described in section 3. However, the method of choice results in different properties in the reduced model. The concatenation approach is especially appropriate when the local basis matrices are obtained by rational interpolation methods of section 3.1. Due to Theorem 3.1, even after concatenation, the final reduced parameterized model obtained by the global basis matrices will still interpolate the original model at every frequency and parameter interpolation point combination used in constructing every local basis matrix, assuming that the SVD performed on the concatenated matrices removed only the zero singular values. If this SVD step removes some small nonzero singular values below a numerical tolerance, then the interpolation will be approximate. This is in contrast to, for example, the balanced truncation approach. If the local basis matrices are obtained via balanced truncation (i.e., $\mathbf{V}_i$ and $\mathbf{W}_i$ are the balanced truncation basis matrices at the parameter values $\mathbf{p}_i$), then once the concatenation is performed, the resulting reduced model is no longer guaranteed to be balanced even at the parameter value $\mathbf{p}_i$. Nevertheless, concatenation of local balanced truncation basis may yield a good reduced parametric model, since the reduction subspace carries the important balanced truncation basis information for the individual parameters.

4.1.2. A Global Basis via Bilinearization. In the special case that only the matrix $\mathbf{A}$ depends on the parameter, and that dependence is affine, another method

to construct a global basis is to use a bilinearization approach. In this case, the parametric system takes the form

$$\begin{aligned} \mathbf{E}\,\dot{\mathbf{x}}(t;\mathbf{p}) &= \mathbf{A}(\mathbf{p})\,\mathbf{x}(t;\mathbf{p}) + \mathbf{B}\,\mathbf{u}(t), \\ \mathbf{y}(t;\mathbf{p}) &= \mathbf{C}\,\mathbf{x}(t;\mathbf{p}), \end{aligned} \tag{4.3}$$

with $\mathbf{A}(\mathbf{p})$ as in (2.7), i.e., affine parametric dependency is allowed in $\mathbf{A}$ only. Defining now auxiliary input functions $u_{m+j}(t) := f_j(\mathbf{p})$, $j = 1, \ldots, M$, and augmenting $\mathbf{B}$ by M zero columns (denoting the resulting input function by $\tilde{\mathbf{u}}$ and the new input matrix by $\tilde{\mathbf{B}}$), the parametric system (4.3) is transformed into a nonparametric *bilinear system*

$$\begin{aligned} \mathbf{E}\,\dot{\mathbf{x}}(t) &= \mathbf{A}_0\mathbf{x}(t) + \sum_{j=1}^{M} u_{m+j}(t)\mathbf{A}_i\mathbf{x}(t) + \tilde{\mathbf{B}}\,\tilde{\mathbf{u}}(t), \\ \mathbf{y}(t) &= \mathbf{C}\,\mathbf{x}(t). \end{aligned} \tag{4.4}$$

Now any model reduction method for bilinear systems can be applied to (4.4). Using any two-sided projection method for this task yields basis matrices $\mathbf{V}$ and $\mathbf{W}$. The structure of the system has been exploited to recast it in a way that removes the explicit parametric dependence and instead captures this dependence through the introduction of the auxiliary inputs; thus, the resulting $\mathbf{V}$ and $\mathbf{W}$ are global basis matrices for the original parametric system (4.3). Applying these basis matrices directly to (4.3) as in (2.8) results in a parametric reduced-order model. The bilinearization approach for parametric model reduction was first suggested in [35] and extended in several ways in [40]. Possible choices for determining the global basis include balanced truncation for bilinear systems [41, 110], Krylov subspace techniques [24, 58, 192], and $\mathcal{H}_2$-optimal methods [36, 99, 229]—in principle, any of the methods discussed in section 3 can be generalized to apply to bilinear systems, and thus to yield a global basis for model reduction of parametric systems that have the form of (4.3).

4.2. Local Bases at Multiple Parameter Points. As opposed to constructing fixed global basis matrices $\mathbf{V}$ and $\mathbf{W}$, one might construct several local ones by sampling the parameter space at points $\mathbf{p}_1, \ldots, \mathbf{p}_K$. As before, denote these local bases by $\mathbf{V}_k$ and $\mathbf{W}_k$ for $k = 1, \ldots, K$. One can then proceed by interpolating the local bases, interpolating the local reduced model matrices, or interpolating the local reduced model transfer functions, as described in more detail below. It is important to recognize that an LTI system has a nonunique state-space representation; in particular, any reduced-order basis that spans a given subspace of $\mathbb{R}^n$ is an equally valid representation of that subspace. This is an important consideration for approaches that interpolate local bases or local reduced model matrices, and is dealt with in various ways in the methods described below.

4.2.1. Interpolating the Local Bases. Given a set of local basis matrices, $\{\mathbf{V}_1, \ldots, \mathbf{V}_K\}$ and $\{\mathbf{W}_1, \ldots, \mathbf{W}_K\}$, a basis matrix for a new parameter value can be obtained by interpolating the local reduced bases. Straightforward interpolation of the entries in the basis vectors can lead to an interpolated quantity that does not preserve desired properties. Consider the simple case $\mathbf{V}_2 = -\mathbf{V}_1$. Let $\hat{\mathbf{p}} = \frac{\mathbf{p}_1 + \mathbf{p}_2}{2}$. In this case, a straightforward linear interpolation leads to $\mathbf{V}(\hat{\mathbf{p}}) = 0.5(\mathbf{V}_1 - \mathbf{V}_2) = \mathbf{0}$, which is clearly not a basis. Thus, as [8] recognized, the quantity to be interpolated should not be the local basis, but rather the underlying subspace. Thus, a better method interpolates the subspaces corresponding to $\mathbf{V}_k$ and $\mathbf{W}_k$ for $k = 1, \ldots, K$ on a tangent space to a manifold of these subspaces [8]. The manifold is chosen so as to preserve

desired properties. It also results in the property that the subspace spanned by the interpolated basis depends only on the underlying local subspaces, and is independent from the coordinate representations of the local bases. Below we briefly explain this approach for the particular case of constructing an orthonormal $\mathbf{V}$, where the $\mathbf{V}_k, k = 1, \ldots, K$, each represent an orthonormal basis, and the interpolation is done on the tangent space to the *Grassmann manifold*.

The Grassmann manifold $\mathcal{G}_{n,r}$ is the set of all r-dimensional subspaces of $\mathbb{R}^n$. Then, in the model reduction setting, the ranges of the model reduction bases $\mathbf{V}_k$, i.e., $\mathrm{Ran}(\mathbf{V}_k)$, can be suitably considered elements of $\mathcal{G}_{n,r}$. Define $\mathcal{R}_k = \mathrm{Ran}(\mathbf{V}_k) \in \mathcal{G}_{n,r}$. The Stiefel manifold, $\mathcal{S}_{n,r}$, is the set of all r-dimensional orthonormal bases of $\mathbb{R}^n$ for $1 \leq r \leq n$. Thus, the orthonormal basis $\mathbf{V}_k \in \mathbb{R}^{n\times r}$ for $\mathcal{R}_k$ is a point on the Stiefel manifold $\mathcal{S}_{n,r}$. For more details regarding the Stiefel and Grassmann manifolds, see, e.g., [8, 81].

The first step in the approach proposed in [8] is to choose a reference point. For simplicity, take $\mathcal{R}_1 \in \mathcal{G}_{n,r}$ as the reference point. Let $\mathcal{T}_{\mathcal{R}_1}$ be the tangent space of $\mathcal{G}_{n,r}$ at $\mathcal{R}_1$. The next step is to map all other subspaces $\mathcal{R}_2, \mathcal{R}_3, \ldots, \mathcal{R}_K$ onto the tangent space defined by $\mathcal{R}_1$. A point $\mathcal{R}_k \in \mathcal{G}_{n,r}$ in a neighborhood of $\mathcal{R}_1$ can be mapped to $\mathcal{T}_{\mathcal{R}_1}$ by the logarithmic map

$$\mathcal{X}_k = \mathrm{Log}_{\mathcal{R}_1}(\mathcal{R}_k) \in \mathcal{T}_{\mathcal{R}_1}. \tag{4.5}$$

Let $\mathbf{T}_k$ be the matrix representing the point $\mathcal{X}_k \in \mathcal{T}_{\mathcal{R}_1}$. To compute $\mathbf{T}_k$, first compute a thin SVD:

$$(\mathbf{I} - \mathbf{V}_1\mathbf{V}_1^T)\mathbf{V}_k(\mathbf{V}_1^T\mathbf{V}_k)^{-1} = \mathbf{U}_k\Sigma_k\mathbf{Z}_k^T. \tag{4.6}$$

Then

$$\mathbf{T}_k = \mathbf{U}_k \arctan(\Sigma_k)\mathbf{Z}_k^T. \tag{4.7}$$

The map $\mathbf{T}_k$ defines a geodesic on the manifold from $\mathcal{R}_1$ to $\mathcal{R}_k$.

Now, given a new parameter point $\hat{\mathbf{p}}$, the method interpolates the local bases in their mapped representations. That is, interpolate $\{\mathbf{T}_1, \ldots, \mathbf{T}_K\}$ using the parameter interpolation points $\{\mathbf{p}_1, \ldots, \mathbf{p}_K\}$. For example,

$$\mathbf{T}(\hat{\mathbf{p}}) = \sum_{k=1}^{K} L_k(\hat{\mathbf{p}})\mathbf{T}_k, \tag{4.8}$$

where $L_k(\hat{\mathbf{p}})$ are the Lagrange basis functions. Let $\hat{\mathcal{X}} \in \mathcal{T}_{\mathcal{R}_1}$ be the point corresponding to the matrix $\mathbf{T}(\hat{\mathbf{p}})$. After $\hat{\mathcal{X}}$ is computed, the exponential map

$$\mathrm{Exp}_{\mathcal{R}_1}(\hat{\mathcal{X}}) \in \mathcal{G}_{n,r} \tag{4.9}$$

maps it back to the original manifold $\mathcal{G}_{n,r}$. Let $\mathbf{V} \in \mathcal{S}_{n,r}$ span the resulting subspace. This results in an interpolated basis $\mathbf{V}$ at the new parameter point $\hat{\mathbf{p}}$. Numerically, the step mapping back to the original manifold is achieved by computing a thin SVD,

$$\mathbf{T}(\hat{\mathbf{p}}) = \hat{\mathbf{U}}\,\hat{\Sigma}\,\hat{\mathbf{Z}}^T, \tag{4.10}$$

followed by

$$\mathbf{V}(\hat{\mathbf{p}}) = \mathbf{V}_1\hat{\mathbf{Z}}\cos(\hat{\Sigma}) + \hat{\mathbf{U}}\sin(\hat{\Sigma}). \tag{4.11}$$

4.2.2. Interpolating the Local Reduced Model Matrices. Interpolating the local basis matrices as outlined above has the disadvantage that when the new basis matrices are computed for a given $\hat{\mathbf{p}}$, the multiplications $\mathbf{W}^T\mathbf{E}(\hat{\mathbf{p}})\mathbf{V}$ and $\mathbf{W}^T\mathbf{A}(\hat{\mathbf{p}})\mathbf{V}$ need to be recomputed. In general, these multiplications depend on the original system dimension n and thus are expensive. Recent work has addressed this issue for affine parametric dependence (as in (2.7)) by precomputing those quantities that do not depend on the parameters [207]. An approach to overcome this problem in the general nonaffine case is to interpolate reduced state-space quantities as opposed to the basis matrices themselves. This idea was recently introduced in [5, 9, 74, 161, 188]. The methods proposed in [9, 188] first perform a *congruence transformation* of the local basis matrices in $\{\mathbf{V}_1, \ldots, \mathbf{V}_K\}$ (and similarly for $\{\mathbf{W}_1, \ldots, \mathbf{W}_K\}$), so that the reduced systems are expressed in the same generalized coordinate system—this is to provide a consistent reduced basis representation across all parameters. Then the reduced-order coefficient matrices constructed from these transformed projection matrices can be interpolated by using matrix manifold interpolation [9] and direct interpolation [188]. In [74], global basis vectors $\mathbf{V}_k = \overline{\mathbf{V}}$ and $\mathbf{W}_k = \overline{\mathbf{W}}$ are used for all $\mathbf{p}_k$, and therefore there is no need to perform the congruence transformation on the bases before interpolating the reduced models.

Suppose one has local reduced-order coefficient matrices $\mathbf{A}_r(\mathbf{p}_k) = \mathbf{W}_k^T\mathbf{A}(\mathbf{p}_k)\mathbf{V}_k$ corresponding to parameters $\mathbf{p}_1, \ldots, \mathbf{p}_K$. To perform the congruence transformation, first select a reference reduced system. For simplicity, take $k = 1$ as the reference system. Next, compute transformation matrices $\mathbf{Q}_k \in \mathbb{R}^{r\times r}$ and $\mathbf{P}_k \in \mathbb{R}^{r\times r}$ for all $k = 1, \ldots, K$ by solving

$$\mathbf{Q}_k = \arg\min_{\mathbf{Q}} \|\mathbf{V}_k\mathbf{Q} - \mathbf{V}_1\|_F \quad \text{subject to} \quad \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_r \tag{4.12}$$

and

$$\mathbf{P}_k = \arg\min_{\mathbf{P}} \|\mathbf{W}_k\mathbf{P} - \mathbf{W}_1\|_F \quad \text{subject to} \quad \mathbf{P}^T\mathbf{P} = \mathbf{I}_r, \tag{4.13}$$

where the orthogonality constraints on $\mathbf{Q}$ and $\mathbf{P}$ assume that the local bases are orthogonal. The optimization problems (4.12) and (4.13), known as the orthogonal Procrustes problem [109], are solved analytically by the SVD. Towards this goal, define the SVDs $\mathbf{V}_k^T\mathbf{V}_1 = \mathbf{U}_{\mathbf{v}_k}\Sigma_{\mathbf{v}_k}\mathbf{Z}_{\mathbf{v}_k}^T$ and $\mathbf{W}_k^T\mathbf{W}_1 = \mathbf{U}_{\mathbf{w}_k}\Sigma_{\mathbf{w}_k}\mathbf{Z}_{\mathbf{w}_k}^T$. Then the solutions to (4.12) and (4.13) are given by

$$\mathbf{Q}_k = \mathbf{U}_{\mathbf{v}_k}\mathbf{Z}_{\mathbf{v}_k}^T \quad \text{and} \quad \mathbf{P}_k = \mathbf{U}_{\mathbf{w}_k}\mathbf{Z}_{\mathbf{w}_k}^T \quad \text{for} \quad k = 1, \ldots, K. \tag{4.14}$$

For $k = 1$, the reference system $\mathbf{Q}_1 = \mathbf{P}_1 = \mathbf{I}_r$. For $k = 1, \ldots, K$, define the transformed local reduction matrices by $\widetilde{\mathbf{V}}_k := \mathbf{V}_k\mathbf{Q}_k$ and $\widetilde{\mathbf{W}}_k := \mathbf{W}_k\mathbf{P}_k$. Then the congruence-transformed local state-space matrices are given by

$$\widetilde{\mathbf{A}}_r(\mathbf{p}_k) := \mathbf{P}_k^T\mathbf{W}_k^T\mathbf{A}(\mathbf{p}_k)\mathbf{V}_k\mathbf{Q}_k = \mathbf{P}_k^T\mathbf{A}_r(\mathbf{p}_k)\mathbf{Q}_k \quad \text{for} \quad k = 1, \ldots, K. \tag{4.15}$$

Note that the transformed basis matrices $\widetilde{\mathbf{V}}_k$ and $\widetilde{\mathbf{W}}_k$ are applied throughout the original system associated with $\mathbf{p}_k$ and therefore not only transform $\mathbf{A}_r(\mathbf{p}_k)$ to $\widetilde{\mathbf{A}}_r(\mathbf{p}_k)$, but also transform other coefficient matrices. More specifically, the kth local reduced model

$$\mathbf{E}_r(\mathbf{p}_k)\dot{\mathbf{x}}_r = \mathbf{A}_r(\mathbf{p}_k)\mathbf{x}_r + \mathbf{B}_r(\mathbf{p}_k)\mathbf{u}, \qquad \mathbf{y}_r = \mathbf{C}_r(\mathbf{p}_k)\mathbf{x}_r, \tag{4.16}$$

is transformed into

$$\widetilde{\mathbf{E}}_r(\mathbf{p}_k)\dot{\widetilde{\mathbf{x}}}_r = \widetilde{\mathbf{A}}_r(\mathbf{p}_k)\widetilde{\mathbf{x}}_r + \widetilde{\mathbf{B}}_r(\mathbf{p}_k)\mathbf{u}, \qquad \widetilde{\mathbf{y}}_r = \widetilde{\mathbf{C}}_r(\mathbf{p}_k)\widetilde{\mathbf{x}}_r, \tag{4.17}$$

where $\widetilde{\mathbf{E}}_r(\mathbf{p}_k) = \mathbf{P}_k^T\mathbf{E}_r(\mathbf{p}_k)\mathbf{Q}_k$, $\widetilde{\mathbf{B}}_r(\mathbf{p}_k) = \mathbf{P}_k^T\mathbf{B}_r(\mathbf{p}_k)$, and $\widetilde{\mathbf{C}}_r(\mathbf{p}_k) = \widehat{\mathbf{C}}_r(\mathbf{p}_k)\mathbf{Q}_k$. In contrast to the local systems in (4.16), the transformed systems in (4.17) are expected to lie in the same generalized coordinate system.

Then the question becomes how to construct the reduced-order matrices $\mathbf{A}_r(\hat{\mathbf{p}})$, $\mathbf{E}_r(\hat{\mathbf{p}})$, $\mathbf{B}_r(\hat{\mathbf{p}})$, and $\mathbf{C}_r(\hat{\mathbf{p}})$ for a new parameter value $\hat{\mathbf{p}}$ using the local congruence-transformed matrices. The reduced matrices $\mathbf{A}_r(\hat{\mathbf{p}})$ and $\mathbf{E}_r(\hat{\mathbf{p}})$ are obtained by applying (similar) manifold interpolation ideas explained in the previous section to the matrices $\{\widetilde{\mathbf{A}}_r(\mathbf{p}_1), \ldots, \widetilde{\mathbf{A}}_r(\mathbf{p}_K)\}$ and $\{\widetilde{\mathbf{E}}_r(\mathbf{p}_1), \ldots, \widetilde{\mathbf{E}}_r(\mathbf{p}_K)\}$, respectively, assuming nonsingular transformed matrices. On the other hand, when the transformed coefficient matrices are not square as in the case of $\widetilde{\mathbf{B}}_r(\mathbf{p}_k)$ and $\widetilde{\mathbf{C}}_r(\mathbf{p}_k)$, or are singular, the reduced-order matrices can be obtained for the new parameter $\hat{\mathbf{p}}$ by using direct interpolation from [9].

A similar approach based on matrix interpolation is also proposed in [188]. In this approach, the local reduced-order matrices are also first transformed and then interpolated; however, the transformation and the interpolation techniques are different from the ones used in [9]. Let $\mathbf{U}_r$ denote the first r dominant left singular vectors of $[\mathbf{V}_1, \ldots, \mathbf{V}_K]$. Define $\mathbf{M}_k := (\mathbf{W}_k^T\mathbf{U}_r)^{-1}$ and $\mathbf{N}_k := (\mathbf{V}_k^T\mathbf{U}_r)^{-1}$. Then the reduced system (4.16) is transformed into

$$\bar{\mathbf{E}}_r(\mathbf{p}_k)\dot{\bar{\mathbf{x}}}_r = \bar{\mathbf{A}}_r(\mathbf{p}_k)\bar{\mathbf{x}}_r + \bar{\mathbf{B}}_r(\mathbf{p}_k)\mathbf{u}, \qquad \bar{\mathbf{y}}_r = \bar{\mathbf{C}}_r(\mathbf{p}_k)\bar{\mathbf{x}}_r, \tag{4.18}$$

where $\bar{\mathbf{E}}_r(\mathbf{p}_k) = \mathbf{M}_k\mathbf{E}_r(\mathbf{p}_k)\mathbf{N}_k$, $\bar{\mathbf{A}}_r(\mathbf{p}_k) = \mathbf{M}_k\mathbf{A}_r(\mathbf{p}_k)\mathbf{N}_k$, $\bar{\mathbf{B}}_r(\mathbf{p}_k) = \mathbf{M}_k\mathbf{B}_r(\mathbf{p}_k)$, and $\bar{\mathbf{C}}_r(\mathbf{p}_k) = \mathbf{C}_r(\mathbf{p}_k)\mathbf{N}_k$. After this transformation, for a given parameter $\hat{\mathbf{p}}$, the matrices $\{\bar{\mathbf{E}}_r(\mathbf{p}_1), \ldots, \bar{\mathbf{E}}_r(\mathbf{p}_K)\}$, $\{\bar{\mathbf{A}}_r(\mathbf{p}_1), \ldots, \bar{\mathbf{A}}_r(\mathbf{p}_K)\}$, $\{\bar{\mathbf{B}}_r(\mathbf{p}_1), \ldots, \bar{\mathbf{B}}_r(\mathbf{p}_K)\}$, and $\{\bar{\mathbf{C}}_r(\mathbf{p}_1), \ldots, \bar{\mathbf{C}}_r(\mathbf{p}_K)\}$ are directly interpolated using any appropriate interpolation method.

Both of these methods achieve the goal of interpolating the reduced state-space quantities while avoiding interpolation of the basis matrices themselves and avoiding the multiplications $\mathbf{W}^T\mathbf{E}(\hat{\mathbf{p}})\mathbf{V}$ and $\mathbf{W}^T\mathbf{A}(\hat{\mathbf{p}})\mathbf{V}$, and thus are likely to be computationally more efficient than the methods described in section 4.2.1. The second approach differs from the first in that it uses the SVD of the concatenation of the local bases, $[\mathbf{V}_1, \ldots, \mathbf{V}_K]$, to define the common generalized coordinate system rather than picking one of the local systems as a reference. The tradeoff is then computing the SVDs of $2(K-1)$ matrices of size $n \times r$ to determine $\mathbf{Q}_k$ and $\mathbf{P}_k$ versus computing the SVD of a single matrix of size $n \times (rK)$ to determine $\mathbf{M}_k$ and $\mathbf{N}_k$; thus the computational complexity of these two options is roughly the same.

4.2.3. Interpolating the Local Transfer Functions. Yet another option for interpolating local information is to interpolate the transfer functions of local reduced models. Given the sampled parameter points $\mathbf{p}_k, k = 1, \ldots, K$, the local reduced models have transfer functions

$$\mathbf{H}_r(s, \mathbf{p}_k) = \mathbf{C}_r(\mathbf{p}_k)(s\mathbf{E}_r(\mathbf{p}_k) - \mathbf{A}_r(\mathbf{p}_k))^{-1}\mathbf{B}_r(\mathbf{p}_k), \quad k = 1, \ldots, K, \tag{4.19}$$

where $\mathbf{H}_r(s, \mathbf{p}_k)$ denotes the transfer function of the reduced model constructed at $\mathbf{p} = \mathbf{p}_k$. The order of each local reduced model is permitted to differ. Denote the order of the reduced model $\mathbf{H}_r(s, \mathbf{p}_k)$ as r_k. These local reduced models can be constructed with any set of local bases.

The reduced-order transfer function at a new parameter point $\hat{\mathbf{p}}$ can then be obtained by interpolating the $\mathbf{H}_r(s, \mathbf{p}_k)$:

$$\widehat{\mathbf{H}}_r(s, \hat{\mathbf{p}}) = \sum_{k=1}^{K} L_k(\hat{\mathbf{p}}) \mathbf{H}_r(s, \mathbf{p}_k), \tag{4.20}$$

where $L_k(\mathbf{p}_j) = \delta_{kj}$, $k, j = 1, \ldots, K$. This yields the functional interpolation condition

$$\widehat{\mathbf{H}}_r(s, \mathbf{p}_k) = \mathbf{H}_r(s, \mathbf{p}_k) \text{ for } k = 1, \ldots, K. \tag{4.21}$$

Because each local transfer function is independent of the representation of its local basis, the nonuniqueness of the state-space representation is not an issue for this interpolation approach.

This strategy was studied in detail in [28, 31]. Generally speaking, any interpolation technique could be used to select the functions L_k (which also applies to (4.8)). Considering the case of a scalar parameter, $d = 1$, one choice is the Lagrange polynomials, which were used in [28]. A computationally more efficient and reliable form is obtained if one uses the barycentric form of Lagrange interpolation [52]. Other choices are rational interpolation (which is used in [31] in barycentric form based on [51]), variants of Hermite interpolation, sinc or spline interpolation, etc.; see [31] for experiments with some of these variants. These interpolation techniques can all be generalized to $d > 1$ using tensorization, but with limited success for growing dimension d of the parameter space. We discuss this issue in some more detail in section 4.2.4.

We will consider one particular case to highlight the advantages and disadvantages of interpolating the reduced-order transfer functions. Consider the case of a scalar parameter, where balanced truncation with a specified error tolerance τ_{BT} is used for generating all the local reduced models, as in [28]. Noting that the representation (4.20) separates the variables s and $\mathbf{p}$, we are now free to put the parametric dependence in any one of the matrices realizing the reduced model. A convenient choice results from the representation (4.20), where a particular realization of $\widehat{\mathbf{H}}_r(s, \mathbf{p})$ is obtained by

$$\begin{aligned}
\widehat{\mathbf{E}}_r &= \operatorname{diag}\big(\mathbf{E}_r(\mathbf{p}_1), \ldots, \mathbf{E}_r(\mathbf{p}_K)\big) \in \mathbb{R}^{r \times r}, \\
\widehat{\mathbf{A}}_r &= \operatorname{diag}\big(\mathbf{A}_r(\mathbf{p}_1), \ldots, \mathbf{A}_r(\mathbf{p}_K)\big) \in \mathbb{R}^{r \times r}, \\
\widehat{\mathbf{B}}_r &= \big[\, \mathbf{B}_r(\mathbf{p}_1)^T, \ldots, \mathbf{B}_r(\mathbf{p}_K)^T \big]^T \in \mathbb{R}^{r \times m}, \\
\widehat{\mathbf{C}}_r(\mathbf{p}) &= [L_1(\mathbf{p})\mathbf{C}_r(\mathbf{p}_1), \ldots, L_K(\mathbf{p})\mathbf{C}_r(\mathbf{p}_K)] \in \mathbb{R}^{q \times r},
\end{aligned} \tag{4.22}$$

with $r = r_1 + r_2 + \cdots + r_K$.

Similar to the approaches of section 4.2.2, the method is flexible in the sense that it does not require affine parameter dependence as in (2.7) for efficient online reduced model generation. In addition, the structure of the approximate transfer function permits using local reduced models of different orders. This is important since balanced truncation with adaptive order will in general yield models with varying r_k. Another advantage is that the main features of balanced truncation are inherited. In particular, if the full model is uniformly stable in the parameter domain Ω, then this property is preserved in the reduced model: for a uniformly stable system, $(\mathbf{E}(\mathbf{p}), \mathbf{A}(\mathbf{p}))$ is an asymptotically stable matrix pencil for all $\mathbf{p} \in \Omega$, and hence $(\mathbf{E}(\mathbf{p}_k), \mathbf{A}(\mathbf{p}_k))$ is asymptotically stable for all k if the parameter samples are taken from Ω. By stability preservation of balanced truncation, then also $(\mathbf{E}_r(\mathbf{p}_k), \mathbf{A}_r(\mathbf{p}_k))$ is asymptotically

stable for all k and, due to the block-diagonal structure, $(\widehat{\mathbf{E}}_r, \widehat{\mathbf{A}}_r)$ inherits this property. Also, instead of balanced truncation, other balancing-based strategies can be employed (e.g., LQG balancing, which is also applicable in the unstable case). See [34, 115, 183] for overviews of possible choices and corresponding properties preserved in the reduced models.

Using balanced truncation for the local models induces an error bound which is obtained by splitting the interpolation error in parameter space from the balanced truncation error. Since the expressions for the error bound depend strongly on the chosen interpolation method, we only provide a version for (polynomial) Lagrange interpolation, but writing the bounds down for other interpolation techniques is straightforward. For details, we refer the reader to the original sources [28, 31]. Assume $\mathbf{H}$ is uniformly stable in Ω, at least K times differentiable with respect to $\mathbf{p}$, and let Λ_{K-1} denote the Lebesgue constant of our node set $\{\mathbf{p}_1, \ldots, \mathbf{p}_K\}$. Then one obtains

$$\left\|\mathbf{H}(s,\hat{\mathbf{p}}) - \widehat{\mathbf{H}}_r(s,\hat{\mathbf{p}})\right\|_2 \leq \rho(\mathbf{H},\mathbf{p},s) + \tau_{\mathrm{BT}}\Lambda_{K-1}, \tag{4.23}$$

where $\rho(\mathbf{H},\hat{\mathbf{p}},s)$ is the interpolation error, which for polynomial interpolation is given by

$$\rho(\mathbf{H},\hat{\mathbf{p}},s) = \frac{1}{K!}\left\|\frac{\partial^K}{\partial \mathbf{p}^K}\mathbf{H}(s,\xi(\mathbf{p}))\right\|_2 \cdot \prod_{k=1}^{K} |\hat{\mathbf{p}} - \mathbf{p}_k|, \tag{4.24}$$

with $\xi(\mathbf{p}) \in (\min_k \mathbf{p}_k, \max_k \mathbf{p}_k)$ (to be understood componentwise). The presence of the Lebesgue constant in the error bound suggests using a node set that produces a small Λ_{K-1}; hence, a uniform distribution of the $\mathbf{p}_k$ should be avoided. A reasonable choice is the Chebyshev nodes, particularly in their second form. As usual, the interpolation error will in general not be determined explicitly, but can be estimated using, e.g., the reduced model instead of $\mathbf{H}$; see, e.g., [28, 31].

A clear disadvantage of the transfer function interpolation approach is that the state-space obtained from the realization $\widehat{\mathbf{A}}_r, \widehat{\mathbf{B}}_r, \widehat{\mathbf{E}}_r, \widehat{\mathbf{C}}_r(\mathbf{p})$ may grow quickly with the number of parameter samples. However, note that for *evaluating* the reduced model online, it is not necessary to explicitly form this realization since the only computation required is the evaluation of the reduced-order transfer functions at the given value s and the evaluation of the (scalar) parametric functions L_k. This amounts to interpolating the outputs computed by the reduced-order models using the given L_k's. This results in an input/output representation of the system and not a state-space realization per se. A similar idea is used in [163], where interpolation is used to create a parametric POD model for a convection problem where the mathematical model is unknown.

Another disadvantage is that by interpolating the reduced-order transfer functions as in (4.20), the poles of $\widehat{\mathbf{H}}_r$ are fixed and do not vary with $\mathbf{p}$ as the poles of $\mathbf{H}$ most likely do. In addition, some poles of the original model may be duplicated in several of the local reduced models, leading to spurious poles of the reduced-order transfer function. These observations are reported in [188], where it is suggested to interpolate the realizations of the local reduced models instead, as already discussed in the previous subsection.

4.2.4. Interpolating Local Information for Systems with Multiple Parameters. For systems with multiple parameters (i.e., $d > 1$) interpolation of local information (as in, e.g., (4.8) and (4.20)) can become more challenging. For regular

grids, tensorization of scalar interpolation methods is fairly easy, resulting, e.g., in bi-/trilinear/-cubic (Hermite, spline) interpolation. Due to the curse of dimensionality, the resulting reduced models quickly become complex even for small d. In [28], the use of sparse grid interpolation [63, 113, 228] is suggested for the interpolating-the-transfer-function approach. The coupling of balanced truncation with piecewise polynomial interpolation using sparse grid points is reported in [28]. Promising results are obtained for this approach for examples with $d < 10$; see [31]. Another possibility to extend the idea of interpolating local objects would be to use irregular grid interpolation approaches such as radial basis functions. Results reported in [213] are also promising and certainly deserve further investigation. It should be noted, though, that dealing with cases $d \gg 10$, e.g., in the case of discretized distributed parameters, with multivariate interpolation parametric model reduction approaches has not been investigated so far. Future research is needed to fuse ideas from high-dimensional approximation theory with parametric model reduction techniques. We hope this review can trigger some of this necessary research. Another important aspect not yet treated in much depth in the existing literature is the regularity of the solution manifold or the transfer function with respect to the parameters. So far, we have assumed a smooth dependency, which, generally speaking, allows the application of all the discussed (multivariate) interpolation approaches, where weaker assumptions would certainly suffice in most cases. But it is not clear yet what type of regularity assumption (weaker than analyticity) with respect to $\mathbf{p}$ is necessary for good approximability in the sense described here, i.e., in the sense that a reduced-order, dynamical, parametric system can be derived yielding good output reconstruction with respect to variations in both $\mathbf{p}$ *and* $\mathbf{u}$.

4.3. Comparing Different Reduced Model Generation Approaches. This section has discussed several approaches for parametric reduced model generation, either using a global basis as in section 4.1 or interpolating local information as in section 4.2. The specific approaches presented are

- using a single global basis;
- interpolating among local bases;
- interpolating among local reduced-order state-space matrices; and
- interpolating among local reduced-order transfer functions.

The following discussion provides some insight into the class of problems for which the different approaches may be most appropriate.

4.3.1. Parameter Sampling. All the presented reduced model generation approaches require an effective sampling of the parameter space, whether samples are used to construct global or local bases. For high-dimensional parameter spaces, the global basis approach is particularly amenable to the greedy adaptive parameter sampling approach discussed in section 3.4, as shown in [61] for building a global POD basis. Similarly, a greedy sampling approach can be easily combined with balanced truncation or rational interpolation methods to construct a global basis. In theory, it would be possible to wrap a greedy sampling approach around the local interpolation methods, although in practice formulation of the problem would be more complicated (and perhaps expensive to solve). Adaptive selection of parameter sample points in conjunction with these local interpolation methods has not been explored in the literature. The parameter selection strategies that are optimal in the composite system-theoretic error measures defined in (2.25) and (2.26) remain an open challenge except for the very special parameter dependence as in [29].

4.3.2. Reduced Model Construction. In terms of basis construction, any of the methods for computing the basis matrices can be used with any of the reduced model generation approaches. However, the global basis approach is a more compelling choice to use with model reduction by rational (tangential) interpolation since the final parametric reduced model then retains the interpolation property in both the frequency and parameter domains. The global basis and interpolation among local bases approaches both require construction of a new reduced model (via projection) at every parameter value to be solved; hence it is important that the original system has an affine parametrization or is approximated by an affine decomposition so that the repeated reduced model construction is not too costly. However, for the other two methods of interpolating the local reduced matrices and interpolating local transfer functions, reduced models are only constructed at the parameter sample points and not at other parameter values to be solved. Thus, these latter two methods may be more appropriate in the case of general parameter dependence where a DEIM or other affine approximation is inaccurate, or in the case where a DEIM approximation cannot be efficiently applied (as discussed in section 2.3).

4.3.3. Reduced Model Size. The different methods vary in how the overall size of the reduced model behaves, as well as in their flexibility to adapt order in different regions of the parameter space. Interpolating local transfer functions allows for local models with different orders, whereas interpolating the local bases and local reduced matrices both require the same order of local models. The flexibility of having different local orders is important in problems where the system dynamics are significantly more or less rich (i.e., significantly harder or easier to approximate) at different parameter values. In the global basis approach there is only a single order of the basis, but adaptivity can be achieved by uneven sampling of the parameter space (e.g., using the greedy adaptive sampling approach). An advantage of interpolating the local bases or local reduced matrices is that the order of the reduced model does not grow with the number of parameter samples (it is fixed at the local reduced order). This is in contrast to the transfer function interpolation approach, where the order of the parametric reduced model can grow quickly due to the block-diagonal structure of construction. For the global basis approach, the order will also grow if the global basis is obtained by concatenating the local bases. However, if the concatenation is followed by a rank revealing QR or SVD, growth in the overall reduced model order can be managed appropriately. In the nonparametric case, typically, one expects the reduced order r to be at least equal to the minimum of m (the number of inputs) and q (the number of outputs). However, even in the nonparametric case, the relationship between r and the number of inputs and outputs can vary significantly, depending on the particular system dynamics under consideration. In the parametric case, this relationship is similarly problem dependent and is now also affected by the particular parametric dependency of the system. In most cases, especially for the global basis approach where the local bases are concatenated, the dimension r is expected to be larger than q and m. Moreover, in general, as r increases, the approximation is expected to improve, although this cannot usually be guaranteed. In some cases, accuracy behavior with increasing r is well characterized; for example, it can be seen that the $\mathcal{H}_\infty$ error bound for balanced truncation guarantees improvement in a local reduced model at a fixed parameter point. As expected, a definitive answer to these questions in the general parametric model reduction case, such as how quickly error decays as r increases, depends on the particular system dynamics and the parametric dependency under consideration.

4.3.4. Error Estimates. An advantage of interpolating local reduced-order transfer functions is that it separates the interpolation error in the parameter space from the model reduction error (in the frequency domain) and allows an error bound as shown in (4.23), assuming that local reduced models are obtained with an error tolerance. Thus, balanced truncation, which has an a priori error bound, is appropriate to employ in this approach. For other methods, a posteriori error estimates may be possible for certain classes of problems and could be similarly combined with the interpolation error estimate.

4.3.5. Dynamical System Poles. For the global basis, interpolating local bases, and interpolating local reduced matrices approaches, the poles of the reduced models vary with the parameter $\mathbf{p}$, due to the parametric dependency in the reduced $\mathbf{E}_r(\mathbf{p})$ and $\mathbf{A}_r(\mathbf{p})$ matrices. This is desirable, since the poles of the original model also vary with the parameter $\mathbf{p}$. This is not the case for the reduced models obtained by interpolating the local transfer functions, for which the poles are fixed at the poles of the local models, which can lead to spurious poles in the reduced model.

4.3.6. Reduced Basis Method. In recent years, the *reduced basis method* [189] has been widely used for parametric model reduction. This method provides a strategy for sampling the parameter space, generating the basis, and creating the reduced model. We note that we do not explicitly present the reduced basis method in this paper. One reason is that the reduced basis method is usually formulated in the continuous (PDE) domain rather than in the dynamical state-space setting of this paper. Our starting point, the dynamical system in (2.1), is referred to as the "truth model" in the reduced basis framework. The second and more important reason is that this paper presents a general flexible framework for formulating parametric model reduction. The framework involves three distinct steps: parameter sampling, basis generation, and reduced model generation. For each step we have presented a number of different options, many of which can be combined in flexible ways. The reduced basis method implemented in discretized form corresponds to one specific choice for each step: the greedy search strategy for parameter sampling, the POD snapshot method for basis generation, and a global basis approach for reduced model generation. Nonetheless, developments in the reduced basis method have contributed important general advancements to the field of parametric model reduction, especially in methods to efficiently handle nonaffine parametric dependency and in the use of error estimates. These contributions are reflected in the literature cited throughout this paper.

5. Discussion. The preceding sections have presented the different elements for projection-based model reduction of parameterized systems. The major components can be summarized as follows: (1) Choose the parameter values at which to sample. This may be done at once as a first step or in an iterative fashion in conjunction with other reduced model construction components. (2) Evaluate the full model for each sampled parameter, which might involve evaluating system matrices and solving linear systems, solving Lyapunov equations, or evolving the full model dynamics, depending on the reduction method chosen. (3) Build the reduced basis, using a local or global strategy. (4) Project the governing equations to produce the reduced-order state-space matrices. (5) Use the resulting reduced model for simulation, optimization, control, or uncertainty quantification. Within each element of the model construction we have highlighted a number of different methods that can be employed. While in general there is no definitive "recipe" on how to combine elements to achieve the

most effective reduction strategy for a given problem, we highlight in this section some of the relative advantages and disadvantages of the various methods and how that relates to their effectiveness for various classes of problems. We also cite examples from the literature that demonstrate the effectiveness of different reduction strategies for different applications.

5.1. Commonalities among Model Reduction Methods. We begin with the observation that there are significant commonalities among the different methods for computing the reduction bases $\mathbf{V}$ and $\mathbf{W}$. While the specific strategies used in rational interpolation methods, balanced truncation, POD, and the reduced basis method are at surface level quite different, their commonalities are perhaps stronger than the usual presentation of the methods suggests. Rational interpolation methods, POD, and the reduced basis method are all snapshot-based methods. In classical rational interpolation methods, the snapshots correspond to sampled solutions over the complex frequency domain, with extensions to sampling over the parameter domain for the parametric model reduction variants. For the reduced basis method and POD, snapshots correspond to sampled solutions over the parameter domain in the case of static problems, and to sampled solutions over both time and parameter domains in the case of dynamic problems. POD is more commonly applied to dynamic problems, although examples in the literature also include static problems. Different communities seem to prefer different strategies for sampling the parameter domain, but as discussed in section 3.4, there is a great deal of flexibility in combining sampling strategies with basis computation strategies. For example, the combination of POD with a greedy parameter sampling strategy as in [61, 122] results in an overall reduction approach that is essentially the same as the reduced basis method. The reduced basis community largely focuses on formulation of the model reduction problem in the continuous (PDE) domain; however, the resulting numerical algorithms build a projection basis as the span of a set of discrete snapshots over the parameter domain, just as in the POD. As discussed in section 3.3, duality between time and frequency domain formulations for linear systems also reveals the connections between POD and balanced truncation, and between POD and rational interpolation methods; see [30] for a more detailed discussion of these connections. Nonetheless, some important differences remain among the methods, most notably the error bounds associated with balanced truncation (for a fixed parameter sample) and the applicability of POD to general nonlinear systems.

5.2. Applicability of the Basis Computation Methods. POD is the most generally applicable of the methods for computing the basis, since it relies only on snapshots of the underlying simulation code. As a result, the POD basis can be computed easily, even when the simulation is a black-box code (although note that in the black-box case, the projection step to determine the reduced model remains a challenge). The POD can also be applied to general nonlinear problems, since computation of the POD basis does not rely on a specific problem structure. POD can also be used with any of the sampling strategies discussed. POD has shown to be effective in many application domains, including fluid dynamics, structural dynamics, thermal modeling, atmospheric modeling, and many more. Early applications of POD include unsteady flows and turbulence modeling [73, 105, 205], and unsteady fluid-structure interaction [75]. The study in [73] is perhaps the earliest example of the POD being used for a parametric problem. That work considers rotations of POD basis vectors computed at one Reynolds number to provide a POD basis for a different Reynolds number. An early example of POD snapshots being computed over the parameter domain is [163], which considered Rayleigh–Bénard convection with varying Rayleigh number.

In contrast, balanced truncation and rational interpolation methods are based on system-theoretic quantities such as Gramians and transfer functions, and exploit the specific dynamical system structure. This paper has focused on systems that are linear in state. Even though balanced truncation has been elegantly extended to nonlinear systems in [201], the approach is not yet computationally feasible, even for systems with modest dimension. As already discussed in section 3.3, extending approximate balanced truncation via empirical Gramians to nonlinear systems is a feasible approach [146], though the favorable properties of balanced truncation for linear systems cannot be shown in general in this setting. For the case of bilinear systems, a computationally more feasible approach to balanced truncation has been developed in [41]. Rational interpolation methods have been extended to bilinear [24, 36, 37, 39, 58, 92, 98, 192] and quadratic-in-state systems [37, 38, 97, 114]. Even though these methods have proven effective for a wide range of problem settings, they are most widely used in circuit theory, such as [23, 44, 90, 184, 195], e.g., to analyze and predict signal propagation and interference in electric circuits; in structural mechanics, such as [53, 106, 174, 198, 211], to study, e.g., vibration suppression in large structures or behavior of micro-electromechanical systems; and in (optimal) control and controller reduction, such as [11, 21, 126, 185, 215, 231], e.g., in LQR/LQG control design.

5.3. Capturing Parametric Dependence. Section 4.3 discussed the different attributes of the approaches for generating the parametric reduced model, including both global and local approaches. Clearly there is no one "best" approach; rather, the different approaches will work well for different problems depending on the model's underlying parametric structure. As already discussed, the literature shows several examples of successful applications of these approaches; however, it is lacking a rigorous analysis that connects the performance of the methods with the structure and regularity of the parametric dependence in the problem. Since most of the methods rely on interpolation in one form or another, it might be expected that the large body of mathematical literature on analysis of interpolation methods is a fruitful avenue of investigation. The manifold interpolation methods are particularly novel and interesting; again, more work is needed to understand for what class of problems they are best.

5.4. Preservation of Properties. When choosing a model reduction method, it is also important to consider what mathematical properties are essential to preserve in the reduced model. Rational interpolation methods have the advantage that the transfer function of the parametric reduced model exactly interpolates the transfer function of the full model at sampled frequency and parameter points (as long as the required vectors are included in the global basis). Stability of the parametric reduced model remains an important open question in most cases. Other than for a few exceptions related to specific problem structure, the approach of interpolating the local transfer functions is the only way to guarantee that the reduced model is stable for the full parameter range (provided each local reduced model is stable). Even for linear nonparametric problems, the other approaches do not guarantee in general that an interpolation of stable models yields a globally stable parametric reduced model. Exceptions include the case of negative definite pencils $(\mathbf{E}(\mathbf{p}), \mathbf{A}(\mathbf{p}))$, where one-sided projection, $\mathbf{V} = \mathbf{W}$, combined with either a global basis approach or with positive interpolation weights on local reduced-order matrices will also guarantee stability. However, the general stability preservation property by transfer function interpolation comes at a cost—it results from the fact that the approach does not let

the reduced model poles vary as the parameters change. A number of stabilization methods have been proposed for LTI reduced models in the nonparametric case. For the bilinearization approach discussed in section 4.1.2, some of these stabilization techniques have been adapted in [40]; further extensions of these approaches to the parametric case is an important area of future research. Preservation of other properties such as passivity and contractivity[5] may be important in some applications, but have yet received little attention in parametric model reduction methods.

5.5. Scalability. For problems that involve state dimensions of very large scale and/or very expensive forward model simulations, it is important to consider how amenable the parametric model reduction framework is to high-performance computing. We briefly consider this issue here. The offline stage of parametric model reduction may be data intensive due to extensive large-scale operations. However, if parameter sampling is done using a random or structured grid (e.g., sparse grid) approach, the offline stage lends itself well to parallelism since local model reduction bases corresponding to different parameter samples can be computed independently. For the greedy selection approach, one computes the parameter samples sequentially; however, the basis computation operations for a selected parameter sample could still be done in parallel in some cases. For example, one could parallelize the numerical simulations required to generate POD snapshots or parallelize the linear solves in the rational interpolation case. In addition, the efficiency of the offline phase of rational interpolation methods can be improved by employing subspace recycling, a technique to accelerate the iterative solution of families of linear systems by reusing information obtained in previous solves. See [1, 42, 94] for applications of these techniques to (parametric) model order reduction. For some problems, the state dimension may become sufficiently large-scale that storage becomes an issue. In those cases one would have to use algorithms that trade additional computations for storage (e.g., incremental SVD algorithms that could process data in batches) and that use scalable implementations (e.g., for recent work in this area see [69], which uses MapReduce/Hadoop for scalable parametric model reduction). Given the current research focus on "big data," much attention is being given to such algorithms; parametric model reduction is one area that might benefit from these advances.

5.6. Error Bounds and Error Estimates. Guarantees on the quality of the parametric reduced models remain an important question. The reduced basis community in particular has promoted a strong emphasis on the derivation of error estimates for parametric model reduction [112, 123, 189, 197, 217, 218]. This work has created new methods that certify the reduced model through error estimates that can be computed without recourse to the full model [189]. An important observation is that these error estimates are related to the structure of the system (e.g., properties of the underlying PDEs) being approximated but are not specific to a particular basis computation method. In other words, the basis behind the model reduction step can come from rational interpolation, balanced truncation, or POD. This can be seen by analyzing these error estimates in state-space form as recently presented in [123]. Recall that projection-based model reduction as in (4.1) corresponds to approximating $\mathbf{x}(t;\mathbf{p})$ by $\mathbf{V}\mathbf{x}_r(t;\mathbf{p})$. As in [123], take $\mathbf{E}=\mathbf{I}$ and let $\mathbf{e}(0;\hat{\mathbf{p}})$ denote the error in the state $\mathbf{x}(t;\mathbf{p})$

[5]The concepts of passivity and contractivity appear frequently in circuit and microwave theory. Passive dynamical systems can only absorb power. The transfer function $\mathbf{H}(s)$ of a nonparametric passive system (i) is analytic for $\mathrm{Re}(s)>0$, (ii) satisfies $\overline{\mathbf{H}(s)}=\mathbf{H}(\overline{s})$ for $s\in\mathbb{C}$, and (iii) satisfies $\mathbf{H}(s)+\mathbf{H}^T(\overline{s})$ is positive semidefinite for all $\mathrm{Re}(s)>0$. The transfer function of a nonparametric contractive system, on the other hand, satisfies $\mathbf{I}-\mathbf{H}^T(-\imath\omega)\mathbf{H}(\imath\omega)\geq 0,\ \omega\in\mathbb{R}$.

at time $t = 0$ and at parameter $\mathbf{p} = \hat{\mathbf{p}}$. Then $\mathbf{e}(0;\hat{\mathbf{p}}) = (\mathbf{I} - \mathbf{V}(\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T)\mathbf{x}(0;\hat{\mathbf{p}})$. Similarly define the residual in the state equation

$$\boldsymbol{\rho}(t;\hat{\mathbf{p}}) = \mathbf{A}(\hat{\mathbf{p}})\mathbf{V}\mathbf{x}_r(t;\hat{\mathbf{p}}) + \mathbf{B}(\hat{\mathbf{p}})\mathbf{u}(t) - \mathbf{V}\dot{\mathbf{x}}_r(t;\hat{\mathbf{p}}). \tag{5.1}$$

Let $\xi(\hat{\mathbf{p}}) = \sup_t \left\| e^{\mathbf{A}(\hat{\mathbf{p}})t} \right\|_2$. Then the error in the state variable at time t and at parameter $\hat{\mathbf{p}}$ is bounded by

$$\|\mathbf{x}(t;\hat{\mathbf{p}}) - \mathbf{V}\mathbf{x}_r(t;\hat{\mathbf{p}})\|_2 \leq \xi(\hat{\mathbf{p}}) \left(\|\mathbf{e}(0;\hat{\mathbf{p}})\|_2 + \int_0^t \|\boldsymbol{\rho}(\tau;\hat{\mathbf{p}})\|_2 d\tau \right). \tag{5.2}$$

The output error at time t and the parameter $\hat{\mathbf{p}}$ can be bounded similarly using

$$\|\mathbf{y}(t;\hat{\mathbf{p}}) - \mathbf{y}_r(t;\hat{\mathbf{p}})\|_2 \leq \|\mathbf{C}(\hat{\mathbf{p}})\| \, \|\mathbf{x}(t;\hat{\mathbf{p}}) - \mathbf{V}\mathbf{x}_r(t;\hat{\mathbf{p}})\|_2. \tag{5.3}$$

Although the error estimates apply to a general basis, a key remaining question is the computability of the constant $\xi(\hat{\mathbf{p}})$. This computation must be done in an efficient way, without recourse to the full model, otherwise the error estimate is of limited practical value. Again, the reduced basis community has derived efficient offline/online decomposition approaches to address this issue [189]. These decompositions rely on an affine parametric dependency obtained through an EIM approximation. A second issue is the tightness of the bounds (5.2) and (5.3), which depends on the properties of the underlying full system. In many cases, the effectivity of the error estimate (i.e., the ratio of the estimated error to the true error) can be quite large (see, e.g., [179, 187]).

5.7. Extensions to More General System Structure. The extension of the parametric reduction methods to general nonlinear systems was discussed above. Another assumption on system structure in this paper is that the $\mathbf{E}(\mathbf{p})$ matrix is nonsingular for any parameter selection $\mathbf{p}$. However, in several important applications (e.g., incompressible flows and circuit design) one obtains a system of the form (2.1), where $\mathbf{E}(\mathbf{p})$ could be singular for some or all selections of $\mathbf{p}$. Such systems with a singular $\mathbf{E}(\mathbf{p})$ matrix are called systems of differential-algebraic equations (DAEs). Projection-based model reduction of DAEs has been studied extensively; see, for example, [45, 50, 118, 129, 172, 195, 208, 209]. The theoretical discussions of this paper directly extend to this setting. While in many cases numerical methods can be implemented as effectively as for the ordinary differential equation case, there exist scenarios in which reducing a system of DAEs might prove more costly due to the need for computing projectors on certain deflating subspaces; for details, we refer the reader to the above references.

We also assume that the full-order model $\mathbf{H}(s,\mathbf{p})$ is asymptotically stable for every $\mathbf{p} \in \Omega$; indeed the discussion as presented only requires that $\mathbf{H}(s,\mathbf{p}_i)$ be stable where $\mathbf{p}_i$, for $i = 1,\ldots,K$, are the sampled parameter points. However, even this is unnecessary since most of the model reduction methods presented have already been extended to reducing unstable systems. From a numerical implementation perspective, the rational interpolation methods do not require $\mathbf{H}(s,\mathbf{p}_i)$ to be stable. The only requirement is that the frequency interpolation points $\hat{s}$ should not be a pole of $\mathbf{H}(s,\mathbf{p}_i)$. In the asymptotically stable case, this is easily guaranteed by choosing $\hat{s} \in \mathbb{C}_+$, where $\mathbb{C}_+ = \{z \in \mathbb{C} \mid \mathrm{Re}(z) > 0\}$. In the unstable case, however, one needs to be more careful since some of the poles lie in $\mathbb{C}_+$ and they need to be avoided. Even the optimal rational interpolation-based model reduction methods have been

extended to unstable systems; see, e.g., [143, 165]. Balanced truncation has been also generalized to reducing unstable systems, by either appropriately redefining the system Gramians [25, 232, 233] or by using different balancing techniques, such as LQG balancing [185]. For the POD, the frequency domain formulation will be the appropriate choice for unstable systems since the time-domain snapshots will grow exponentially in this case. This would be a similar approach to the balanced truncation method for unstable systems presented in [232].

We assume zero initial condition for the state-vector, i.e., $\mathbf{x}(0;\mathbf{p}) = \mathbf{0}$. When the initial condition is nonzero but known, this information could be easily included in the time-domain POD method as the initial condition for the numerical simulation. For rational interpolation and balanced truncation approaches, the nonzero initial conditions can be appended to the input-state-matrix $\mathbf{B}(\mathbf{p})$ representing an initial impulse as done in [30, 57, 128]. However, if the initial condition is unknown and/or the reduced model needs to provide a good approximation for a wide range of initial conditions, the corresponding model reduction problem can no longer be easily handled by simple modification of the zero-initial condition case. In these cases, if one treats the initial condition as the parameter set, then the parameter dimension d becomes as large as the system dimension n leading to a high-dimensional parameter space. In some cases this high dimension can be efficiently sampled by exploiting system structure; see, e.g., [27, 154] for recent work that considers initial condition parameters for problems that have linear state dependence.

5.8. Equation-Free Model Reduction. In this paper, we have focused on projection-based model reduction techniques that assume the availability of a state-space description of the original model (2.1) and that apply an explicit projection to state-space dimension. However, there are settings where equations describing the evolution of a dynamical system are not explicitly specified (i.e., $\mathbf{E}(\mathbf{p}), \mathbf{A}(\mathbf{p})$, etc., are unavailable) and the only access to dynamics is via input/output measurements. This might be in the form of system responses such as measurements of $\mathbf{H}(s,\mathbf{p})$, or simulation outputs of a black-box code. In these cases, a reduced parametric model as in (2.2) is obtained directly from measurements or from simulation outputs without access to internal dynamics. For the case of dynamical systems with no parametric dependence, the Loewner-based data-driven framework [148, 149, 171] has been applied with great success to construct reduced models using only transfer function samples. Recently, this approach has been used to produce even locally optimal reduced models; see [33]. The data-driven Loewner framework has been extended to parameter-dependent systems as well [16, 134], where the reduced parametric model is a rational function not only in the frequency s but also in every parameter p_k, $k = 1, \ldots, d$. This allows choosing the order of approximation in s and $\mathbf{p}$ independently. Even though the theoretical discussion extends to the multiple parameter case directly, the numerical computation and construction of the reduced model might present a challenge, and the order of the reduced model might grow undesirably as d, the number of parameters, grows. Another set of nonintrusive approaches represents the parametric solution in a reduced subspace (usually using POD) and then interpolates those solutions without recourse to the underlying full system. Interpolation can be achieved using polynomial or spline interpolation [60, 163, 69], least squares fitting [59, 178], or radial basis function models [20].

6. Outlook. We close this review paper with a brief discussion of promising current research directions and some open challenges. As already mentioned, recent development of the EIM and DEIM methods has led to significant progress in model

reduction for nonlinear systems using POD and the reduced basis method. It remains an open question whether the rational interpolation approaches and balanced truncation can be extended beyond bilinear- and quadratic-in-state systems to handle general nonlinear dependence in the state variable $\mathbf{x}(t)$. A balancing method for nonlinear systems was proposed in [201], but a scalable algorithmic implementation remains an open challenge. Even for the linear-in-state systems we consider in this paper, despite the existence of effective sampling strategies discussed above, the optimal parameter selection strategies for the composite system-theoretic error measures defined in (2.25) and (2.26) remain an open challenge for the case of general parametric dependence. Further open challenges include parametric model reduction for systems with time-dependent and/or stochastic parameters. The bilinearization approach discussed in section 4.1.2 permits time dependency of the parameters since it treats the parameters as auxiliary inputs; see [35]. The POD approach, in principal, will also allow time-dependent parameters; the snapshot generation will then require simulating a dynamical system with time-dependent coefficients. In general, though, extending parametric model reduction methods to these broader classes of systems will require new approaches for defining the reduced subspaces, as well as new methods for constructing and solving the reduced model. Handling high-dimensional parameter spaces remains another challenging problem. While some progress has been made in this area, in particular with greedy sampling approaches, further work is needed to develop methods that exploit system structure to avoid the curse of dimensionality. Promising recent efforts towards this goal use tensor techniques [182, 186]. The combination of tensor calculus [125] and parametric model reduction techniques for time-dependent problems is still in its infancy, but offers a promising research direction.

REFERENCES

[1] K. Ahuja, E. de Sturler, S. Gugercin, and E. R. Chang, *Recycling BiCG with an application to model reduction*, SIAM J. Sci. Comput., 34 (2012), pp. A1925–A1949.

[2] N. Alexandrov, J. Dennis, Jr., R. Lewis, and V. Torczon, *A trust region framework for managing the use of approximation models in optimization*, Struct. Optim., 15 (1998), pp. 16–23.

[3] N. Alexandrov, R. Lewis, C. Gumbert, L. Green, and P. Newman, *Approximation and model management in aerodynamic optimization with variable-fidelity models*, J. Aircraft, 38 (2001), pp. 1093–1101.

[4] D. Allaire, D. Kordonowy, M. Lecerf, L. Mainini, and K. Willcox, *Multifidelity DDDAS methods with application to a self-aware aerospace vehicle*, Procedia Comput. Sci., 29 (2014), pp. 1182–1192.

[5] D. Amsallem, J. Cortial, K. Carlberg, and C. Farhat, *A method for interpolating on manifolds structural dynamics reduced-order models*, Internat. J. Numer. Methods Engrg., 80 (2009), pp. 1241–1258.

[6] D. Amsallem, J. Cortial, and C. Farhat, *Towards real-time computational-fluid-dynamics-based aeroelastic computations using a database of reduced-order information*, AIAA J., 48 (2010), pp. 2029–2037.

[7] D. Amsallem, S. Deolalikar, F. Gurrola, and C. Farhat, *Model predictive control under coupled fluid-structure constraints using a database of reduced-order models on a tablet*, in Proceedings of the 43rd AIAA Fluid Dynamics and Colocated Conferences and Exhibit, San Diego, CA, 2013, paper 2013-2588.

[8] D. Amsallem and C. Farhat, *Interpolation method for the adaptation of reduced-order models to parameter changes and its application to aeroelasticity*, AIAA J., 46 (2008), pp. 1803–1813.

[9] D. Amsallem and C. Farhat, *An online method for interpolating linear parametric reduced-order models*, SIAM J. Sci. Comput., 33 (2011), pp. 2169–2198.

[10] D. Amsallem, M. J. Zahr, and C. Farhat, *Nonlinear model order reduction based on local reduced-order bases*, Internat. J. Numer. Methods Engrg., 92 (2012), pp. 891–916.

[11] B. Anderson and Y. Liu, *Controller reduction: Concepts and approaches*, IEEE Trans. Automat. Control, 34 (2002), pp. 802–812.
[12] H. Antil, M. Heinkenschloss, and R. H. W. Hoppe, *Domain decomposition and balanced truncation model reduction for shape optimization of the Stokes system*, Optim. Methods Softw., 26 (2011), pp. 643–669.
[13] H. Antil, M. Heinkenschloss, R. H. W. Hoppe, C. Linsenmann, and A. Wixforth, *Reduced order modeling based shape optimization of surface acoustic wave driven microfluidic biochips*, Math. Comput. Simulation, 82 (2012), pp. 1986–2003.
[14] A. C. Antoulas, *Approximation of Large-Scale Dynamical Systems*, SIAM, Philadelphia, PA, 2005.
[15] A. Antoulas, C. Beattie, and S. Gugercin, *Interpolatory model reduction of large-scale dynamical systems*, in Efficient Modeling and Control of Large-Scale Systems, J. Mohammadpour and K. Grigoriadis, eds., Springer-Verlag, New York, 2010, pp. 2–58.
[16] A. Antoulas, A. Ionita, and S. Lefteriu, *On two-variable rational interpolation*, Linear Algebra Appl., 436 (2012), pp. 2889–2915.
[17] E. Arian, M. Fahl, and E. Sachs, *Trust-region proper orthogonal decomposition models by optimization methods*, in Proceedings of the 41st IEEE Conference on Decision and Control, Las Vegas, NV, 2002, pp. 3300–3305.
[18] S. Arridge, J. Kaipio, V. Kolehmainen, M. Schweiger, E. Somersalo, T. Tarvainen, and M. Vauhkonen, *Approximation errors and model reduction with an application in optical diffusion tomography*, Inverse Problems, 22 (2006), pp. 175–195.
[19] P. Astrid, S. Weiland, K. Willcox, and T. Backx, *Missing point estimation in models described by proper orthogonal decomposition*, IEEE Trans. Automat. Control, 53 (2008), pp. 2237–2251.
[20] C. Audouze, F. De Vuyst, and P. Nair, *Reduced-order modeling of parameterized PDEs using time–space-parameter principal component analysis*, Internat. J. Numer. Methods Engrg., 80 (2009), pp. 1025–1057.
[21] J. Badía, P. Benner, R. Mayo, E. Quintana-Ortí, G. Quintana-Ortí, and J. Saak, *Parallel order reduction via balanced truncation for optimal cooling of steel profiles*, in Euro-Par 2005 Parallel Processing, Lecture Notes in Comput. Sci. 3648, Springer-Verlag, Berlin, Heidelberg, 2005, pp. 857–866.
[22] Z. Bai, *Krylov subspace techniques for reduced-order modeling of large-scale dynamical systems*, Appl. Numer. Math., 43 (2002), pp. 9–44.
[23] Z. Bai, P. Feldmann, and R. Freund, *How to make theoretically passive reduced-order models passive in practice*, in Custom Integrated Circuits Conference, Proc. IEEE, 1998, pp. 207–210.
[24] Z. Bai and D. Skoogh, *A projection method for model reduction of bilinear dynamical systems*, Linear Algebra Appl., 415 (2006), pp. 406–425.
[25] M. Barahona, A. Doherty, M. Sznaier, H. Mabuchi, and J. Doyle, *Finite horizon model reduction and the appearance of dissipation in Hamiltonian systems*, in Proceedings of the 41st IEEE Conference on Decision and Control, Vol. 4, 2002, pp. 4563–4568.
[26] M. Barrault, Y. Maday, N. Nguyen, and A. Patera, *An "empirical interpolation" method: Application to efficient reduced-basis discretization of partial differential equations*, C. R. Math. Acad. Sci. Paris, 339 (2004), pp. 667–672.
[27] O. Bashir, K. Willcox, O. Ghattas, B. van Bloemen Waanders, and J. Hill, *Hessian-based model reduction for large-scale systems with initial condition inputs*, Internat. J. Numer. Methods Engrg., 73 (2008), pp. 844–868.
[28] U. Baur and P. Benner, *Model reduction for parametric systems using balanced truncation and interpolation*, at–Automatisierungstechnik, 57 (2009), pp. 411–420.
[29] U. Baur, C. Beattie, P. Benner, and S. Gugercin, *Interpolatory projection methods for parameterized model reduction*, SIAM J. Sci. Comput., 33 (2011), pp. 2489–2518.
[30] U. Baur, P. Benner, and L. Feng, *Model order reduction for linear and nonlinear systems: A system-theoretic perspective*, Arch. Comput. Methods Eng., 21 (2014), pp. 331–358.
[31] U. Baur, P. Benner, A. Greiner, J. Korvink, J. Lienemann, and C. Moosmann, *Parameter preserving model order reduction for MEMS applications*, Math. Comput. Model. Dyn. Syst., 17 (2011), pp. 297–317.
[32] C. Beattie and S. Gugercin, *Interpolatory projection methods for structure-preserving model reduction*, Systems Control Lett., 58 (2009), pp. 225–232.
[33] C. Beattie and S. Gugercin, *Realization-independent $\mathcal{H}_2$-approximation*, in Proceedings of the 51st IEEE Conference on Decision and Control, 2012, pp. 4953–4958.
[34] P. Benner, *System-theoretic methods for model reduction of large-scale systems: Simulation, control, and inverse problems*, in Proceedings of MathMod, Vienna, I. Troch and F. Breitenecker, eds., of ARGESIM Report 35, 2009, pp. 126–145.

[35] P. Benner and T. Breiten, *On $\mathcal{H}_2$-model reduction of linear parameter-varying systems*, Proc. Appl. Math. Mech., 11 (2011), pp. 805–806.
[36] P. Benner and T. Breiten, *Interpolation-based $\mathcal{H}_2$-model reduction of bilinear control systems*, SIAM J. Matrix Anal. Appl., 33 (2012), pp. 859–885.
[37] P. Benner and T. Breiten, *Krylov-subspace based model reduction of nonlinear circuit models using bilinear and quadratic-linear approximations*, in Progress in Industrial Mathematics at ECMI 2010, Springer, Berlin, Heidelberg, 2012, pp. 153–159.
[38] P. Benner and T. Breiten, *Two-sided projection methods for nonlinear model order reduction*, SIAM J. Sci. Comput., 37 (2015), pp. B239–B260.
[39] P. Benner, T. Breiten, and T. Damm, *Generalised tangential interpolation for model reduction of discrete-time MIMO bilinear systems*, Internat. J. Control, 84 (2011), pp. 1398–1407.
[40] P. Benner and A. Bruns, *Parametric model order reduction of thermal models using the bilinear interpolatory rational Krylov algorithm*, Math. Comput. Model. Dyn. Syst., 21 (2014), pp. 103–129.
[41] P. Benner and T. Damm, *Lyapunov equations, energy functionals, and model order reduction of bilinear and stochastic systems*, SIAM J. Control Optim., 49 (2011), pp. 686–711.
[42] P. Benner and L. Feng, *On recycling Krylov subspaces for solving linear systems with successive right-hand sides with applications in model reduction*, in Model Reduction for Circuit Simulation, P. Benner, M. Hinze, and E. ter Maten, eds., Lecture Notes in Electrical Engrg. 74, Springer-Verlag, Dordrecht, The Netherlands, 2011, pp. 125–140.
[43] P. Benner and L. Feng, *A robust algorithm for parametric model order reduction based on implicit moment matching*, in Reduced Order Methods for Modeling and Computational Reduction, A. Quarteroni and G. Rozza, eds., MS&A. Model. Simul. Appl. 9, Springer-Verlag, Cham, Switzerland, 2014, pp. 159–185.
[44] P. Benner, M. Hinze, and E. ter Maten, eds., *Model Reduction for Circuit Simulation*, Lecture Notes in Electrical Engrg. 74, Springer-Verlag, Dordrecht, The Netherlands, 2011.
[45] P. Benner, M.-S. Hossain, and T. Stykel, *Model reduction of periodic descriptor systems using balanced truncation*, in Model Reduction for Circuit Simulation, P. Benner, M. Hinze, and E. ter Maten, eds., Lecture Notes in Electrical Engrg. 74, Springer-Verlag, Dordrecht, The Netherlands, 2011, pp. 193–206.
[46] P. Benner, P. Kürschner, and J. Saak, *Efficient handling of complex shift parameters in the low-rank Cholesky factor ADI method*, Numer. Algorithms, 62 (2013), pp. 225–251.
[47] P. Benner, J. Li, and T. Penzl, *Numerical solution of large-scale Lyapunov equations, Riccati equations, and linear-quadratic optimal control problems*, Numer. Linear Algebra Appl., 15 (2008), pp. 755–777.
[48] P. Benner and J. Saak, *Numerical solution of large and sparse continuous time algebraic matrix Riccati and Lyapunov equations: A state of the art survey*, GAMM-Mitt., 36 (2013), pp. 32–52.
[49] P. Benner, E. Sachs, and S. Volkwein, *Model order reduction for PDE constrained optimization*, in Trends in PDE Constrained Optimization, G. Leugering, P. Benner, S. Engell, A. Griewank, H. Harbrecht, M. Hinze, R. Rannacher, and S. Ulbrich, eds., Internat. Ser. Numer. Math. 165, Springer International, 2014, pp. 303–326.
[50] P. Benner and V. Sokolov, *Partial realization of descriptor systems*, Systems Control Lett., 55 (2006), pp. 929–938.
[51] J.-P. Berrut, R. Baltensperger, and H. Mittelmann, *Recent developments in barycentric rational interpolation*, in Trends and Applications in Constructive Approximation, Internat. Ser. Numer. Math. 151, Birkhäuser, Basel, 2005, pp. 27–51.
[52] J.-P. Berrut and L. N. Trefethen, *Barycentric Lagrange interpolation*, SIAM Rev., 46 (2004), pp. 501–517.
[53] D. Bindel, Z. Bai, and J. Demmel, *Model reduction for RF MEMS simulation*, in Applied Parallel Computing. State of the Art in Scientific Computing, Springer, 2006, pp. 286–295.
[54] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk, *Convergence rates for greedy algorithms in the reduced basis methods*, SIAM J. Math. Anal., 43 (2011), pp. 1457–1472.
[55] B. Bond and L. Daniel, *A piecewise-linear moment-matching approach to parameterized model-order reduction for highly nonlinear systems*, IEEE Trans. Computer-Aided Design Integrated Circuits Syst., 26 (2007), pp. 2116–2129.
[56] B. N. Bond and L. Daniel, *Parameterized model order reduction of nonlinear dynamical systems*, in IEEE/ACM International Conference on Computer-Aided Design, 2005, pp. 487–494.
[57] T. Breiten, *Krylov Subspace Methods for Model Order Reduction of Bilinear Control Systems*, Master's thesis, TU Kaiserslautern, Germany, 2009.

[58] T. Breiten and T. Damm, *Krylov subspace methods for model order reduction of bilinear control systems*, Systems Control Lett., 59 (2010), pp. 443–450.
[59] T. Bui-Thanh, M. Damodaran, and K. Willcox, *Aerodynamic data reconstruction and inverse design using proper orthogonal decomposition*, AIAA J., 42 (2004), pp. 1505–1516.
[60] T. Bui-Thanh, D. Murali, and K. Willcox, *Proper orthogonal decomposition extensions for parametric applications in compressible aerodynamics*, in Proceedings of the 21st Applied Aerodynamics AIAA Conference, Orlando, FL, 2003, paper 2003-4213.
[61] T. Bui-Thanh, K. Willcox, and O. Ghattas, *Model reduction for large-scale systems with high-dimensional parametric input space*, SIAM J. Sci. Comput., 30 (2008), pp. 3270–3288.
[62] T. Bui-Thanh, K. Willcox, and O. Ghattas, *Parametric reduced-order models for probabilistic analysis of unsteady aerodynamic applications*, AIAA J., 46 (2008), pp. 2520–2529.
[63] H.-J. Bungartz and M. Griebel, *Sparse grids*, Acta Numer., 13 (2004), pp. 147–269.
[64] A. Bunse-Gerstner, D. Kubalinska, G. Vossen, and D. Wilczek, *H_2-norm optimal model reduction for large scale discrete dynamical MIMO systems*, J. Comput. Appl. Math., 233 (2010), pp. 1202–1216.
[65] R. Cameron and W. Martin, *The orthogonal development of nonlinear functionals in series of Fourier-Hermite functionals*, Ann. of Math., 48 (1947), pp. 385–392.
[66] K. Carlberg, C. Farhat, J. Cortial, and D. Amsallem, *The GNAT method for nonlinear model reduction: Effective implementation and application to computational fluid dynamics and turbulent flows*, J. Comput. Phys., 242 (2013), pp. 623–647.
[67] S. Chaturantabut and D. C. Sorensen, *Nonlinear model reduction via discrete empirical interpolation*, SIAM J. Sci. Comput., 32 (2010), pp. 2737–2764.
[68] C. Chia-Chi, L. Ming-Hong, and F. Wu-Shiung, *MIMO interconnects order reductions by using the multiple point adaptive-order rational global Arnoldi algorithm*, IEICE Trans. Electronics, 89 (2006), pp. 792–802.
[69] P. G. Constantine, D. F. Gleich, Y. Hou, and J. Templeton, *Model reduction with MapReduce-enabled tall and skinny singular value decomposition*, SIAM J. Sci. Comput., 36 (2014), pp. S166–S191.
[70] P. G. Constantine and Q. Wang, *Residual minimizing model interpolation for parameterized nonlinear dynamical systems*, SIAM J. Sci. Comput., 34 (2012), pp. A2118–A2144.
[71] L. Daniel, O. Siong, S. Low, K. Lee, and J. White, *A multiparameter moment matching model reduction approach for generating geometrically parameterized interconnect performance models*, IEEE Trans. Computer-Aided Design Integrated Circuits Syst., 23 (2004), pp. 678–693.
[72] E. de Sturler, S. Gugercin, M. E. Kilmer, S. Chaturantabut, C. Beattie, and M. O'Connell, *Nonlinear parametric inversion using interpolatory model reduction*, SIAM J. Sci. Comput., 37 (2015), pp. B495–B517.
[73] A. Deane, I. Kevrekidis, G. E. Karniadakis, and S. Orszag, *Low-dimensional models for complex geometry flows: Application to grooved channels and circular cylinders*, Phys. Fluids A, 3 (1991), pp. 2337–2354.
[74] J. Degroote, J. Vierendeels, and K. Willcox, *Interpolation among reduced-order matrices to obtain parameterized models for design, optimization and probabilistic analysis*, Internat. J. Numer. Methods Fluids, 63 (2010), pp. 207–230.
[75] E. Dowell and K. Hall, *Modeling of fluid-structure interaction*, Ann. Rev. Fluid Mech., 33 (2001), pp. 445–490.
[76] M. Drohmann, B. Haasdonk, and M. Ohlberger, *Adaptive reduced basis methods for nonlinear convection-diffusion equations*, in Finite Volumes for Complex Applications. VI. Problems & Perspectives, J. Fořt, J. Fürst, J. Halama, R. Herbin, and F. Hubert, eds., Springer Proc. Math. 4, Springer, Heidelberg, 2011, pp. 369–377.
[77] M. Drohmann, B. Haasdonk, and M. Ohlberger, *Reduced basis approximation for nonlinear parametrized evolution equations based on empirical operator interpolation*, SIAM J. Sci. Comput., 34 (2012), pp. A937–A969.
[78] V. Druskin, L. Knizhnerman, and V. Simoncini, *Analysis of the rational Krylov subspace and ADI methods for solving the Lyapunov equation*, SIAM J. Numer. Anal., 49 (2011), pp. 1875–1898.
[79] V. Druskin, C. Lieberman, and M. Zaslavsky, *On adaptive choice of shifts in rational Krylov subspace reduction of evolutionary problems*, SIAM J. Sci. Comput., 32 (2010), pp. 2485–2496.
[80] V. Druskin, V. Simoncini, and M. Zaslavsky, *Solution of the time-domain inverse resistivity problem in the model reduction framework, part* I. *One-dimensional problem with SISO data*, SIAM J. Sci. Comput., 35 (2013), pp. A1621–A1640.

[81] A. Edelman, T. A. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 303–353.
[82] J. L. Eftang and B. Stamm, *Parameter multi-domain "hp" empirical interpolation*, Internat. J. Numer. Methods Engrg., 90 (2012), pp. 412–428.
[83] R. Eid, B. Salimbahrami, B. Lohmann, E. Rudnyi, and J. Korvink, *Parametric order reduction of proportionally damped second-order systems*, in Sensors and Materials, Tokyo, Vol. 19, 2007, pp. 149–164.
[84] M. Eldred and D. Dunlavy, *Formulations for surrogate-based optimization with data-fit, multifidelity, and reduced-order models*, in Proceedings of the 11th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Portsmouth, VA, 2006, paper 2006-7117.
[85] M. Eldred, A. Giunta, and S. Collis, *Second-order corrections for surrogate-based optimization with model hierarchies*. in Proceedings of the 10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, 2004, paper 2004-4457.
[86] H. C. Elman and Q. Liao, *Reduced basis collocation methods for partial differential equations with random coefficients*, SIAM/ASA J. Uncertain. Quantif., 1 (2013), pp. 192–217.
[87] R. Everson and L. Sirovich, *The Karhunen-Loeve procedure for gappy data*, J. Optical Soc. Amer., 12 (1995), pp. 1657–1664.
[88] M. Fahl and E. Sachs, *Reduced order modelling approaches to PDE-constrained optimization based on proper orthogonal decomposition*, in Large-Scale PDE-Constrained Optimization, L. T. Biegler, O. Ghattas, M. Heinkenschloss, and B. van Bloemen Waanders, eds., Springer, 2003, pp. 268–280.
[89] O. Farle, V. Hill, P. Ingelström, and R. Dyczij-Edlinger, *Multi-parameter polynomial order reduction of linear finite element models*, Math. Comput. Model. Dyn. Syst., 14 (2008), pp. 421–434.
[90] P. Feldmann and R. Freund, *Efficient linear circuit analysis by Padé approximation via the Lanczos process*, IEEE Trans. Computer-Aided Design Integrated Circuits Syst., 14 (1995), pp. 639–649.
[91] L. Feng, *Parameter independent model order reduction*, Math. Comput. Simulation, 68 (2005), pp. 221–234.
[92] L. Feng and P. Benner, *A note on projection techniques for model order reduction of bilinear systems*, in International Conference of Numerical Analysis and Applied Mathematics, AIP Conf. Proc. 936, Springer-Verlag, 2007, pp. 208–211.
[93] L. Feng and P. Benner, *A robust algorithm for parametric model order reduction based on implicit moment matching*, Proc. Appl. Math. Mech., 7 (2008), pp. 10215.01–10215.02.
[94] L. Feng, P. Benner, and J. Korvink, *Subspace recycling accelerates the parametric macromodeling of MEMS*, Internat. J. Numer. Methods Engrg., 94 (2013), pp. 84–110.
[95] L. Feng, D. Koziol, E. Rudnyi, and J. Korvink, *Parametric model reduction for fast simulation of cyclic voltammograms*, Sensor Lett., 4 (2006), pp. 165–173.
[96] L. Feng, E. Rudnyi, and J. Korvink, *Preserving the film coefficient as a parameter in the compact thermal model for fast electrothermal simulation*, IEEE Trans. Computer-Aided Design Integrated Circuits Syst., 24 (2005), pp. 1838–1847.
[97] L. Feng, X. Zeng, C. Chiang, D. Zhou, and Q. Fang, *Direct nonlinear order reduction with variational analysis*, in Design, Automation and Test in Europe Conference and Exhibition, Vol. 2, 2004, pp. 1316–1321.
[98] G. Flagg, *Interpolation Methods for Model Reduction of Bilinear Systems*, Ph.D. thesis, Department of Mathematics, Virginia Tech, 2012.
[99] G. Flagg and S. Gugercin, *Multipoint Volterra series interpolation and $\mathcal{H}_2$ optimal model reduction of bilinear systems*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 549–579.
[100] M. Frangos, Y. Marzouk, K. Willcox, and B. van Bloemen Waanders, *Surrogate and reduced-order modeling: A comparison of approaches for large-scale statistical inverse problems*, in Computational Methods for Large-Scale Inverse Problems and Quantification of Uncertainty, L. Biegler, G. Biros, O. Ghattas, M. Heinkenschloss, D. Keyes, B. Mallick, Y. Marzouk, L. Tenorio, B. van Bloemen Waanders, and K. Willcox, eds., Wiley, 2011.
[101] R. Freund, *Model reduction methods based on Krylov subspaces*, Acta Numer., 12 (2003), pp. 267–319.
[102] D. Galbally, K. Fidkowski, K. Willcox, and O. Ghattas, *Nonlinear model reduction for uncertainty quantification in large-scale inverse problems*, Internat. J. Numer. Methods Engrg., 81 (2010), pp. 1581–1608.
[103] K. Gallivan, E. Grimme, and P. Van Dooren, *Asymptotic waveform evaluation via a Lanczos method*, Appl. Math. Lett., 7 (1994), pp. 75–80.
[104] K. Gallivan, A. Vandendorpe, and P. Van Dooren, *Model reduction of MIMO systems via tangential interpolation*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 328–349.

[105] T. Gatski and M. Glauser, *Proper orthogonal decomposition based turbulence modeling*, in Instability, Transition, and Turbulence, Springer, 1992, pp. 498–510.
[106] W. Gawronski and J. Juang, *Model reduction for flexible structures*, Control Dynam. Syst., 36 (1990), pp. 143–222.
[107] R. Ghanem and P. Spanos, *Stochastic Finite Elements: A Spectral Approach*, Springer-Verlag, New York, 1991.
[108] A. Giunta, V. Balabanov, M. Kaufman, S. Burgee, B. Grossman, R. Haftka, W. Mason, and L. Watson, *Variable-complexity response surface design of an HSCT configuration*, in Multidisciplinary Design Optimization: State of the Art, N. M. Alexandrov and M. Y. Hussaini, eds., SIAM, Philadelphia, PA, 1997, pp. 348–367.
[109] G. Golub and C. van Loan, *Matrix Computations*, 4th ed., Johns Hopkins University Press, Baltimore, MD, 2013.
[110] W. S. Gray and J. Mesko, *Energy functions and algebraic Gramians for bilinear systems*, in Preprints of the 4th IFAC Nonlinear Control Systems Design Symposium, Enschede, The Netherlands, 1998, pp. 103–108.
[111] M. Grepl, *Reduced-Basis Approximation and A Posteriori Error Estimation for Parabolic Partial Differential Equations*, Ph.D. thesis, MIT, Cambridge, MA, 2005.
[112] M. Grepl and A. Patera, A posteriori *error bounds for reduced-basis approximations of parametrized parabolic partial differential equations*, M2AN Math. Model. Numer. Anal., 39 (2005), pp. 157–181.
[113] M. Griebel, *Sparse grids and related approximation schemes for higher dimensional problems*, in Foundations of Computational Mathematics (FoCM05), Santander, L. Pardo, A. Pinkus, E. Suli, and M. Todd, eds., Cambridge University Press, Cambridge, UK, 2006, pp. 106–161.
[114] C. Gu, *QLMOR: A projection-based nonlinear model order reduction approach using quadratic-linear representation of nonlinear systems*, IEEE Trans. Computer-Aided Design Integrated Circuits Syst., 30 (2011), pp. 1307–1320.
[115] S. Gugercin and A. Antoulas, *A survey of model reduction by balanced truncation and some new results*, Internat. J. Control, 77 (2004), pp. 748–766.
[116] S. Gugercin, A. C. Antoulas, and C. Beattie, $\mathcal{H}_2$ *model reduction for large-scale linear dynamical systems*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 609–638.
[117] S. Gugercin, D. Sorensen, and A. Antoulas, *A modified low-rank Smith method for large-scale Lyapunov equations*, Numer. Algorithms, 32 (2003), pp. 27–55.
[118] S. Gugercin, T. Stykel, and S. Wyatt, *Model reduction of descriptor systems by interpolatory projection methods*, SIAM J. Sci. Comput., 35 (2013), pp. B1010–B1033.
[119] P. Gunupudi, R. Khazaka, and M. Nakhla, *Analysis of transmission line circuits using multidimensional model reduction techniques*, IEEE Trans. Adv. Packaging, 25 (2002), pp. 174–180.
[120] P. Gunupudi, R. Khazaka, M. Nakhla, T. Smy, and D. Celo, *Passive parameterized time-domain macromodels for high-speed transmission-line networks*, IEEE Trans. Microwave Theory Tech., 51 (2003), pp. 2347–2354.
[121] B. Haasdonk, M. Dihlmann, and M. Ohlberger, *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 (2011), pp. 423–442.
[122] B. Haasdonk and M. Ohlberger, *Reduced basis method for finite volume approximations of parametrized linear evolution equations*, M2AN Math. Model. Numer. Anal., 42 (2008), pp. 277–302.
[123] B. Haasdonk and M. Ohlberger, *Efficient reduced models and a posteriori error estimation for parametrized dynamical systems by offline/online decomposition*, Math. Comput. Model. Dyn. Syst., 17 (2011), pp. 145–161.
[124] W. Hackbusch, *On the fast solving of parabolic boundary control problems*, SIAM J. Control Optim., 17 (1979), pp. 231–244.
[125] W. Hackbusch, *Tensor Spaces and Numerical Tensor Calculus*, Springer, Berlin, 2012.
[126] Y. Halevi, D. Bernstein, and W. Haddad, *On stable full-order and reduced-order LQG controllers*, Optim. Control Appl. Methods, 12 (1991), pp. 163–172.
[127] A. Hay, J. Borggaard, and D. Pelletier, *Local improvements to reduced-order models using sensitivity analysis of the proper orthogonal decomposition*, J. Fluid Mech., 629 (2009), pp. 41–72.
[128] M. Heinkenschloss, T. Reis, and A. Antoulas, *Balanced truncation model reduction for systems with inhomogeneous initial conditions*, Automatica, 47 (2011), pp. 559–564.
[129] M. Heinkenschloss, D. C. Sorensen, and K. Sun, *Balanced truncation model reduction for a class of descriptor systems with application to the Oseen equations*, SIAM J. Sci. Comput., 30 (2008), pp. 1038–1063.

[130] M. Hess and P. Benner, *A reduced basis method for microwave semiconductor devices with geometric variations*, COMPEL, 33 (2014), pp. 1071–1081.
[131] M. Hinze and S. Volkwein, *Error estimates for abstract linear-quadratic optimal control problems using proper orthogonal decomposition*, Comput. Optim. Appl., 39 (2008), pp. 319–345.
[132] H. Hotelling, *Analysis of a complex of statistical variables with principal components*, J. Educ. Psychology, 24 (1933), pp. 417–441, 498–520.
[133] S. Hovland, J. Gravdahl, and K. Willcox, *Explicit model predictive control for large-scale systems via model reduction*, J. Guidance Control Dynam., 31 (2008), pp. 918–926.
[134] A. C. Ionita and A. C. Antoulas, *Data-driven parametrized model reduction in the Loewner framework*, SIAM J. Sci. Comput., 36 (2014), pp. A984–A1007.
[135] I. M. Jaimoukha and E. M. Kasenally, *Krylov subspace methods for solving large Lyapunov equations*, SIAM J. Numer. Anal., 31 (1994), pp. 227–251.
[136] I. Jolliffe, *Principal component analysis*, in Encyclopedia of Statistics in Behavioral Science, John Wiley & Sons, 2005.
[137] M. Kaufman, V. Balabanov, A. Giunta, B. Grossman, W. Mason, S. Burgee, R. Haftka, and L. Watson, *Variable-complexity response surface approximations for wing structural weight in HSCT design*, Comput. Mech., 18 (1996), pp. 112–126.
[138] M. Kavousanakis, R. Erban, A. Boudouvis, C. Gear, and I. Kevrekidis, *Projective and coarse projective integration for problems with continuous symmetries*, J. Comput. Phys., 225 (2007), pp. 382–407.
[139] A. Keane and P. Nair, *Computational Approaches for Aerospace Design: The Pursuit of Excellence*, John Wiley & Sons, 2005.
[140] M. Kennedy and A. O'Hagan, *Bayesian calibration of computer models*, J. R. Stat. Soc. Ser. B Stat. Methodol., 63 (2001), pp. 425–464.
[141] T. Kim, *Frequency-domain Karhunen-Loeve method and its application to linear dynamic systems*, AIAA J., 36 (1998), pp. 2117–2123.
[142] D. Kosambi, *Statistics in function space*, J. Indian Math. Soc., 7 (1943), pp. 76–88.
[143] D. Kubalińska, *Optimal Interpolation-Based Model Reduction*, Ph.D. thesis, Universität Bremen, 2008.
[144] K. Kunisch and S. Volkwein, *Proper orthogonal decomposition for optimality systems*, M2AN Math. Model. Numer. Anal., 42 (2008), pp. 1–23.
[145] K. Kunisch and S. Volkwein, *Optimal snapshot location for computing POD basis functions*, M2AN Math. Model. Numer. Anal., 44 (2010), pp. 509–529.
[146] S. Lall, J. Marsden, and S. Glavaski, *A subspace approach to balanced truncation for model reduction of nonlinear control systems*, Internat. J. Robust Nonlinear Control, 12 (2002), pp. 519–535.
[147] T. Lassila, A. Manzoni, A. Quarteroni, and G. Rozza, *A reduced computational and geometrical framework for inverse problems in hemodynamics*, Internat. J. Numer. Methods Biomed. Engrg., 29 (2013), pp. 741–776.
[148] S. Lefteriu and A. Antoulas, *A new approach to modeling multiport systems from frequency-domain data*, IEEE Trans. Computer-Aided Design Integrated Circuits Syst., 29 (2010), pp. 14–27.
[149] S. Lefteriu and A. C. Antoulas, *A new adaptive approach to modeling measured multiport scattering parameters*, in Scientific Computing in Electrical Engineering, SCEE 2008, J. Roos and L. Costa, eds., Math. Indust. 14, Springer-Verlag, 2010, pp. 21–28.
[150] A.-M. Leung and R. Khazaka, *Parametric model order reduction technique for design optimization*, in IEEE International Symposium on Circuits and Systems (ISCAS), 2005, pp. 1290–1293.
[151] R. M. Lewis and S. G. Nash, *Model problems for the multigrid optimization of systems governed by differential equations*, SIAM J. Sci. Comput., 26 (2005), pp. 1811–1837.
[152] J.-R. Li and J. White, *Low-rank solution of Lyapunov equations*, SIAM Rev., 46 (2004), pp. 693–713.
[153] Y. Li, Z. Bai, and Y. Su, *A two-directional Arnoldi process and its application to parametric model order reduction*, J. Comput. Appl. Math., 226 (2009), pp. 10–21.
[154] C. Lieberman, K. Fidkowski, K. Willcox, and B. Bloemen Waanders, *Hessian-based model reduction: Large-scale inversion and prediction*, Internat. J. Numer. Methods Fluids, 71 (2013), pp. 135–150.
[155] C. Lieberman and K. Willcox, *Goal-oriented inference: Approach, linear theory, and application to advection diffusion*, SIAM Rev., 55 (2013), pp. 493–519.
[156] C. Lieberman, K. Willcox, and O. Ghattas, *Parameter and state model reduction for large-scale statistical inverse problems*, SIAM J. Sci. Comput., 32 (2010), pp. 2523–2542.

[157] J. Liesen and Z. Strakos, *Krylov Subspace Methods: Principles and Analysis*, Oxford University Press, 2012.
[158] T. Lieu and C. Farhat, *Adaptation of aeroelastic reduced-order models and application to an F-16 configuration*, AIAA J., 45 (2007), pp. 1244–1257.
[159] T. Lieu, C. Farhat, and M. Lesoinne, *Reduced-order fluid/structure modeling of a complete aircraft configuration*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 5730–5742.
[160] M. Loéve, *Probability Theory*, D. Van Nostrand, New York, 1955.
[161] B. Lohmann and R. Eid, *Efficient order reduction of parametric and nonlinear models by superposition of locally reduced models*, in Methoden und Anwendungen der Regelungstechnik. Erlangen-Münchener Workshops 2007 und 2008, G. Roppencker and B. Lohmann, eds., Aachen, 2009, Shaker Verlag, pp. 27–36.
[162] J. Lumley, *The Structures of Inhomogeneous Turbulent Flow*, Atmospheric Turbulence and Radio Wave Propagation, (1967), pp. 166–178.
[163] H. Ly and H. Tran, *Modeling and control of physical processes using proper orthogonal decomposition*, J. Math. Comput. Model., 33 (2001), pp. 223–236.
[164] M. Ma, A.-M. Leung, and R. Khazaka, *Sparse macromodels for parametric networks*, in IEEE International Symposium on Circuits and Systems (ISCAS), 2006, pp. 2545–2548.
[165] C. Magruder, C. Beattie, and S. Gugercin, *Rational Krylov methods for optimal $\mathcal{L}_2$ model reduction*, in Proceedings of the 49th IEEE Conference on Decision and Control, 2010, pp. 6797–6802.
[166] L. Mainini and K. Willcox, *A surrogate modeling approach to support real-time structural assessment and decision making*, AIAA J., 53 (2015), pp. 1612–1626.
[167] A. March and K. Willcox, *Provably convergent multifidelity optimization algorithm not requiring high-fidelity derivatives*, AIAA J., 50 (2012), pp. 1079–1089.
[168] Y. M. Marzouk and H. N. Najm, *Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems*, J. Comput. Phys., 228 (2009), pp. 1862–1902.
[169] Y. M. Marzouk, H. N. Najm, and L. A. Rahn, *Stochastic spectral methods for efficient Bayesian solution of inverse problems*, J. Comput. Phys., 224 (2007), pp. 560–586.
[170] L. Mathelin and O. L. Maitre, *Robust control of uncertain cylinder wake flows based on robust reduced order models*, Comput. Fluids, 38 (2009), pp. 1168–1182.
[171] A. Mayo and A. Antoulas, *A framework for the solution of the generalized realization problem*, Linear Algebra Appl., 425 (2007), pp. 634–662.
[172] V. Mehrmann and T. Stykel, *Balanced truncation model reduction for large-scale systems in descriptor form*, in Dimension Reduction of Large-Scale Systems, P. Benner, V. Mehrmann, and D. Sorensen, eds., Lect. Notes Comput. Sci. Eng. 45, Springer, Berlin, 2005, pp. 83–115.
[173] L. Meier III and D. Luenberger, *Approximation of linear constant systems*, IEEE Trans. Automat. Control, 12 (1967), pp. 585–588.
[174] D. Meyer and S. Srinivasan, *Balancing and model reduction for second-order form linear systems*, IEEE Trans. Automat. Control, 41 (1996), pp. 1632–1644.
[175] B. Moore, *Principal component analysis in linear systems: Controllability, observability, and model reduction*, IEEE Trans. Automat. Control, 26 (1981), pp. 17–32.
[176] K. Moosmann and J. Korvink, *Automatic parametric MOR for MEMS design*, in Tagungsband GMA-FA 1.30, Modellbildung, Identifikation und Simulation in der Automatisierungstechnik, Workshop am Bostalsee, 27.–29.9.2006, B. Lohmann and A. Kugi, eds., 2006, pp. 89–99.
[177] C. Mullis and R. Roberts, *Synthesis of minimum roundoff noise fixed point digital filters*, IEEE Trans. Circuits Syst., 23 (1976), pp. 551–562.
[178] P. Nair, *Physics based surrogate modeling of parameterized PDEs for optimization and uncertainty analysis*, in Proceedings of the 16th AIAA Computational Fluid Dynamics Conference, Orlando, FL, 2002, paper 2002-1586.
[179] F. Negri, G. Rozza, A. Manzoni, and A. Quarteroni, *Reduced basis method for parametrized elliptic optimal control problems*, SIAM J. Sci. Comput., 35 (2013), pp. A2316–A2340.
[180] N. Nguyen, A. Patera, and J. Peraire, *A "best points" interpolation method for efficient approximation of parametrized functions*, Internat. J. Numer. Methods Engrg., 73 (2008), pp. 521–543.
[181] G. North, T. Bell, R. Cahalan, and F. Moeng, *Sampling errors in the estimation of empirical orthogonal functions*, Monthly Weather Rev., 110 (1982), pp. 699–706.
[182] A. Nouy, *A priori model reduction through proper generalized decomposition for solving time-dependent partial differential equations*, Comput. Methods Appl. Mech. Engrg., 199 (2010), pp. 1603–1626.

[183] G. Obinata and B. Anderson, *Model Reduction for Control System Design*, Communications and Control Engineering Series, Springer-Verlag, London, UK, 2001.
[184] A. Odabasioglu, M. Celik, and L. Pileggi, *PRIMA: Passive reduced-order interconnect macromodeling algorithm*, in Proceedings of the 1997 IEEE/ACM International Conference on Computer-Aided Design, 1997, pp. 58–65.
[185] P. Opdenacker and E. Jonckheere, *LQG balancing and reduced LQG compensation of symmetric passive systems*, Internat. J. Control, 41 (1985), pp. 73–109.
[186] I. V. Oseledets, *Tensor-train decomposition*, SIAM J. Sci. Comput., 33 (2011), pp. 2295–2317.
[187] P. Pacciarini and G. Rozza, *Stabilized reduced basis method for parametrized advection–diffusion PDEs*, Comput. Methods Appl. Mech. Engrg., 274 (2014), pp. 1–18.
[188] H. Panzer, J. Mohring, R. Eid, and B. Lohmann, *Parametric model order reduction by matrix interpolation*, at–Automatisierungstechnik, 58 (2010), pp. 475–484.
[189] A. Patera and G. Rozza, *Reduced basis approximation and* a posteriori *error estimation for parametrized partial differential equations*, Version 1.0, MIT, Cambridge, MA, 2006.
[190] B. Peherstorfer, D. Butnaru, K. Willcox, and H.-J. Bungartz, *Localized discrete empirical interpolation method*, SIAM J. Sci. Comput., 36 (2014), pp. A168–A192.
[191] T. Penzl, *A cyclic low-rank Smith method for large sparse Lyapunov equations*, SIAM J. Sci. Comput., 21 (2000), pp. 1401–1418.
[192] J. Phillips, *Projection-based approaches for model reduction of weakly nonlinear, time-varying systems*, IEEE Trans. Computer-Aided Design Integrated Circuits Syst., 22 (2003), pp. 171–187.
[193] C. Prud'homme, D. Rovas, K. Veroy, Y. Maday, A. Patera, and G. Turinici, *Reliable real-time solution of parameterized partial differential equations: Reduced-basis output bound methods*, J. Fluids Engrg., 124 (2002), pp. 70–80.
[194] M. Rathinam and L. R. Petzold, *A new look at proper orthogonal decomposition*, SIAM J. Numer. Anal., 41 (2003), pp. 1893–1925.
[195] T. Reis and T. Stykel, *Positive real and bounded real balancing for model reduction of descriptor systems*, Internat. J. Control, 83 (2010), pp. 74–88.
[196] C. Rowley, *Model reduction for fluids, using balanced proper orthogonal decomposition*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 15 (2005), pp. 997–1013.
[197] G. Rozza, D. Huynh, and A. Patera, *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 Eng., 15 (2008), pp. 229–275.
[198] E. Rudnyi and J. Korvink, *Review: Automatic model reduction for transient simulation of MEMS-based devices*, Sensors Update, 11 (2002), pp. 3–33.
[199] J. Sabino, *Solution of Large-Scale Lyapunov Equations via the Block Modified Smith Method*, Ph.D. thesis, Rice University, 2007.
[200] A. Sartori, D. Baroli, A. Cammi, L. Luzzi, and G. Rozza, *A reduced order model for multi-group time-dependent parametrized reactor spatial kinetics*, in Proceedings of the 2014 22nd International Conference on Nuclear Engineering (ICONE22), Prague, Czech Republic, ASME, 2014, paper 30707.
[201] J. Scherpen, *Balancing for nonlinear systems*, Systems Control Lett., 21 (1993), pp. 143–153.
[202] W. Schilders, H. van der Vorst, and J. Rommes, eds., *Model Order Reduction: Theory, Research Aspects and Applications*, Math. Indust. 13, Springer-Verlag, Berlin, 2008.
[203] V. Simoncini, *A new iterative method for solving large-scale Lyapunov matrix equations*, SIAM J. Sci. Comput., 29 (2007), pp. 1268–1288.
[204] T. W. Simpson, T. Mauery, J. Korte, and F. Mistree, *Kriging models for global approximation in simulation-based multidisciplinary design optimization*, AIAA J., 39 (2001), pp. 2233–2241.
[205] L. Sirovich, *Turbulence and the dynamics of coherent structures. Part* 1: *Coherent structures*, Quart. Appl. Math., 45 (1987), pp. 561–571.
[206] R. A. Smith, *Matrix equation* $XA + BX = C$, SIAM J. Appl. Math., 16 (1968), pp. 198–201.
[207] N. Son, *A real time procedure for affinely dependent parametric model order reduction using interpolation on Grassmann manifolds*, Internat. J. Numer. Methods Engrg., 93 (2013), pp. 818–833.
[208] T. Stykel, *Gramian-based model reduction for descriptor systems*, Math. Control Signals Systems, 16 (2004), pp. 297–319.
[209] T. Stykel, *Low-rank iterative methods for projected generalized Lyapunov equations*, Electron. Trans. Numer. Anal., 30 (2008), pp. 187–202.
[210] T. Stykel and V. Simoncini, *Krylov subspace methods for projected Lyapunov equations*, Appl. Numer. Math., 62 (2012), pp. 35–50.

[211] T.-J. Su and R. Craig, Jr., *Model reduction and control of flexible structures using Krylov vectors*, J. Guidance Control Dynam., 14 (1991), pp. 260–267.
[212] L. N. Trefethen and D. Bau III, *Numerical Linear Algebra*, SIAM, Philadelphia, PA, 1997.
[213] G. van Bühren, N. Hornung, T. Clees, and L. Nikitinaon, *Aspects of adaptive hierarchical RBF metamodels for optimization*, J. Comput. Methods Sci. Engrg., 12 (2012), pp. 5–23.
[214] P. Van Dooren, K. Gallivan, and P.-A. Absil, *$\mathcal{H}_2$-optimal model reduction of MIMO systems*, Appl. Math. Lett., 21 (2008), pp. 1267–1273.
[215] A. Varga and B. Anderson, *Accuracy-enhancing methods for balancing-related frequency-weighted model and controller reduction*, Automatica, 39 (2003), pp. 919–927.
[216] G. Venter, R. Haftka, and J. Starnes, *Construction of response surface approximations for design optimization*, AIAA J., 36 (1998), pp. 2242–2249.
[217] K. Veroy and A. Patera, *Certified real-time solution of the parametrized steady incompressible Navier-Stokes equations: Rigorous reduced-basis* a posteriori *error bounds*, Internat. J. Numer. Methods Fluids, 47 (2005), pp. 773–788.
[218] K. Veroy, C. Prud'homme, D. Rovas, and A. Patera, *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, Orlando, FL, 2003, paper 2003-3847.
[219] E. Wachspress, *The ADI minimax problem for complex spectra*, Appl. Math. Lett., 1 (1988), pp. 311–314.
[220] J. Wang and N. Zabaras, *Using Bayesian statistics in the estimation of heat source in radiation*, Internat. J. Heat Mass Transfer, 48 (2005), pp. 15–29.
[221] K. Washabaugh, D. Amsallem, M. Zahr, and C. Farhat, *Nonlinear model reduction for CFD problems using local reduced-order bases*, in Proceedings of the 42nd AIAA Fluid Dynamics Conference and Exhibit, Fluid Dynamics and Co-located Conferences, AIAA, 2012, paper 2012-2686.
[222] D. Weile, E. Michielssen, E. Grimme, and K. Gallivan, *A method for generating rational interpolant reduced order models of two-parameter linear systems*, Appl. Math. Lett., 12 (1999), pp. 93–102.
[223] N. Wiener, *The homogeneous chaos*, Amer. J. Math., 60 (1938), pp. 897–936.
[224] S. M. Wild, R. G. Regis, and C. A. Shoemaker, *ORBIT: Optimization by radial basis function interpolation in trust-regions*, SIAM J. Sci. Comput., 30 (2008), pp. 3197–3219.
[225] K. Willcox and J. Peraire, *Balanced model reduction via the proper orthogonal decomposition*, AIAA J., 40 (2002), pp. 2323–2330.
[226] D. Xiu and G. E. Karniadakis, *The Wiener–Askey polynomial chaos for stochastic differential equations*, SIAM J. Sci. Comput., 24 (2002), pp. 619–644.
[227] Y. Yue and K. Meerbergen, *Accelerating optimization of parametric linear systems by model order reduction*, SIAM J. Optim., 23 (2013), pp. 1344–1370.
[228] C. Zenger, *Sparse grids*, in Parallel Algorithms for Partial Differential Equations (Kiel, 1990), Notes Numer. Fluid Mech. 31, Vieweg, Braunschweig, 1991, pp. 241–251.
[229] L. Zhang and J. Lam, *On H_2 model reduction of bilinear systems*, Automatica, 38 (2002), pp. 205–216.
[230] Y. Zhang, L. Feng, S. Li, and P. Benner, *Reduced-order modeling and ROM-based optimization of batch chromatography*, in Numerical Mathematics and Advanced Applications - ENUMATH 2013, A. Abdulle, S. Deparis, D. Kressner, F. Nobile, and M. Picasso, eds., Lect. Notes Comput. Sci. Eng. 103, Springer International Publishing, 2014, pp. 427–435.
[231] K. Zhou, J. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice-Hall, Upper Saddle River, NJ, 1996.
[232] K. Zhou, G. Salomon, and E. Wu, *Balanced realization and model reduction for unstable systems*, Internat. J. Robust Nonlinear Control, 9 (1999), pp. 183–198.
[233] A. Zilouchian, *Balanced structures and model reduction of unstable systems*, in IEEE Proceedings of Southeastcon '91, 1991, pp. 1198–1201.