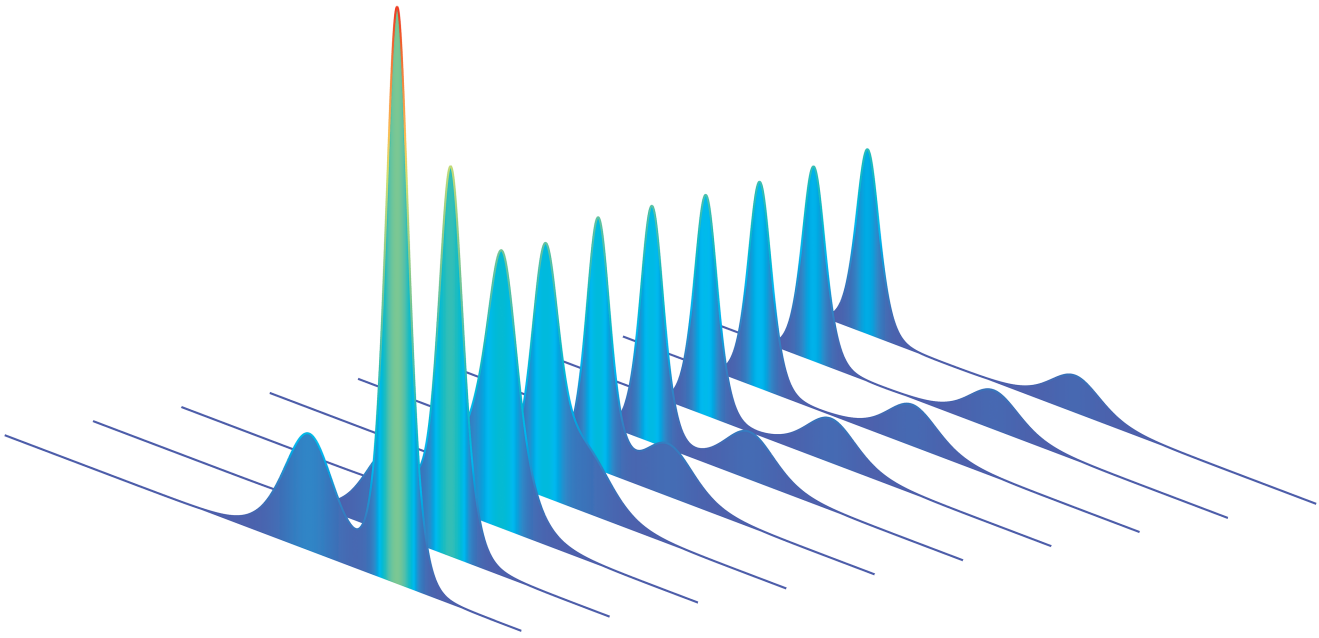
Breaking the Kolmogorov Barrier with Nonlinear Model Reduction



Benjamin Peherstorfer

Introduction. Model reduction is ubiquitous in computational science and engineering. It plays a key role in making computationally tractable outer-loop applications that require simulating systems for many scenarios with different parameters and inputs. Typical outer-loop applications are control, uncertainty quantification, inverse problems, and optimal design [RHP08, BGW15]. With reduced models, one numerically solves the differential equations, which describe the physical system of interest, in problem-dependent, low-dimensional reduced spaces, in contrast to traditional, full models that are formulated in generic, high-dimensional full spaces with, e.g., finite-element/-volume methods. Reduced spaces are constructed in a one-time high-cost training (offline) phase from data and then are leveraged in an online phase to provide approximate solutions often in a fraction of the computation time required for full models, which can greatly speed up the repeated simulations of systems at different scenarios in outer-loop applications.

Model reduction has many attributes of what today is referred to as physics-informed machine learning and scientific machine learning because model reduction simulates physical systems by combining learning from data to construct reduced spaces with traditional numerical methods to solve equations from first principles and physical laws in the learned reduced spaces.

Much progress has been made on deriving reduced models for diffusion-dominated problems governed by specific elliptic/parabolic equations that induce smooth solution manifolds [RHP08, BGW15]. Smooth means that the Kolmogorov n-width decays rapidly so that solutions can be approximated well in low-dimensional spaces [CD16]. However, the important class of problems given by hyperbolic equations, conservation laws,

Benjamin Peherstorfer is an assistant professor at Courant Institute of Mathematical Sciences, New York University. His email address is pehersto@cims.nyu.edu.

Communicated by Notices *Associate Editor Reza Malek-Madani.*

DOI: https://doi.org/10.1090/noti2475

and transport-dominated phenomena—where a coherent structure such as a wave or a phase transition travels through the domain—typically induces rough solution manifolds with slowly decaying Kolmogorov n-widths [OR16, GU19]. Even though lower bounds of the decay of the Kolmogorov n-width are available only for solution manifolds of a limited number of equations, empirical evidence of slowly decaying Kolmogorov n-widths is observed in many applications in science and engineering from pattern formation in biology to storm-surge forecasting in weather modeling to solidification in additive manufacturing to combustion processes in fluid mechanics.

Over the last several years, nonlinear model reduction has started to emerge that seeks nonlinear reduced approximations on manifolds rather than linear approximations in reduced spaces as in classical linear model reduction. The goal of nonlinear model reduction is breaking the Kolmogorov barrier which means achieving a fast error decay even if solution manifolds are not smooth and the Kolmogorov n-width decays slowly as in transport-dominated problems [OR13, PW15, TPQ15, GU19, Peh20, ELMV20, LC20]. There are nonlinear methods that adapt the reduced space explicitly, such as dynamic decompositions [SL09] and the adaptive empirical interpolation method [PW15, Peh20] that will be described in more detail below. Other nonlinear model reduction methods apply transformations to recover (linear) low-rank structures. Examples of such approaches are the method of freezing [OR13] and shifted POD [RSSM18] as well as methods motivated by machine learning such as the approach introduced in [ELMV20] based on Wasserstein spaces and the method proposed in [LC20] that builds on deep autoencoders.

This note describes the Kolmogorov barrier of linear model reduction and then outlines how nonlinear methods can overcome the barrier. A numerical example of simulating combustion instabilities in a single-injector element of a rocket engine demonstrates adaptive empirical interpolation [PW15, Peh20] as one example of a nonlinear model reduction method.

Outer-loop applications. Consider a parametrized partial differential equation (PDE)

$$\partial_t q(x;t,\mu) + \mathcal{N}(q;\mu) = 0 \qquad (1)$$

with operator $\mathcal{N}$ and appropriate initial and boundary conditions. The function $q : \Omega \times \mathcal{T} \times \mathcal{D} \to \mathbb{R}$ depends on the spatial coordinate $x \in \Omega \subset \mathbb{R}^d$, time $t \in \mathcal{T} = [0,T]$ and a parameter $\mu \in \mathcal{D}$. The parameter μ describes properties such as conductivity in heat-transfer problems and viscosity and Reynolds number in fluid-mechanics problems. Traditional numerical methods such as finite-difference, finite-element, and finite-volume methods numerically solve (1) by approximating the solution[1] q of (1) in finite-dimensional vector spaces $\mathcal{U}$. Let the space $\mathcal{U}$ be N-dimensional. Further, let $\varphi_1, \dots, \varphi_N$ be a basis of $\mathcal{U}$, which means that the solution function q is approximated as

$$q_N(x;t,\mu) = \sum_{i=1}^{N} \beta_i(t,\mu)\varphi_i(x),$$

with N coefficients $\beta_1(t,\mu), \dots, \beta_N(t,\mu) \in \mathbb{R}$. Numerically solving the PDE (1) for a given parameter $\mu \in \mathcal{D}$ means solving for the N coefficients $\beta_1(t,\mu), \dots, \beta_N(t,\mu)$ via a system of equations such as

$$r(q_N(\cdot;t_k,\mu), q_N(\cdot;t_{k-1},\mu), \varphi_i) = 0, \qquad (2)$$

for $i = 1, \dots, N$ and $k = 1, \dots, K$, where r is an appropriate residual function that includes the time discretization with K time steps $0 = t_0 < t_1 < \cdots < t_K = T$. Thus, the computational costs of numerically solving the PDE (1), i.e., computing the coefficient vector $\beta(t,\mu) = [\beta_1(t,\mu), \dots, \beta_N(t,\mu)]^T \in \mathbb{R}^N$, scale with the dimension N of the space $\mathcal{U}$ and the number of time steps K. If N and K are large, then computing a solution even for a single parameter $\mu \in \mathcal{D}$ can already be computationally demanding. Thus, it can quickly become infeasible to compute solutions for a large number $M \gg 1$ of parameters $\mu_1, \dots, \mu_M$ as needed in outer-loop applications such as optimization, control, inverse problems, and uncertainty quantification.

Model reduction via projection. Model reduction via projection seeks reduced spaces $\mathcal{U}_n$ of low dimension $n \ll N$ so that reduced solutions in $\mathcal{U}_n$ can be rapidly computed for a larger number of parameters [RHP08, BGW15]. The computational procedures of model reduction are typically split into a training (offline) phase, in which a reduced space $\mathcal{U}_n$ is constructed, and an online phase, in which the PDE is numerically solved in the reduced space for parameters $\mu_1, \dots, \mu_M \in \mathcal{D}$ as part of an outer-loop application. The training phase is a one-time, high-cost preprocessing step that is compensated if reduced PDE solutions are computed for a large number M of parameters in the online phase as, for example, in outer-loop applications.

Reduced spaces are problem dependent in the sense that they are constructed to approximate well the elements of the specific solution manifold

$$\mathcal{M} = \{q(\cdot;t,\mu) \mid t \in \mathcal{T}, \mu \in \mathcal{D}\}$$

corresponding to the PDE of interest. A solution manifold $\mathcal{M}$ is visualized in Figure 1, where the manifold $\mathcal{M}$ is depicted as a spiral.

A classical method to numerically construct a reduced space is based on the principal component analysis: first,

[1] *Typically, one considers, e.g., the weak form of the PDE in specific, appropriate spaces; however, to ease exposition and to avoid heavy notation, we refer to q simply as the solution of the PDE in the following.*

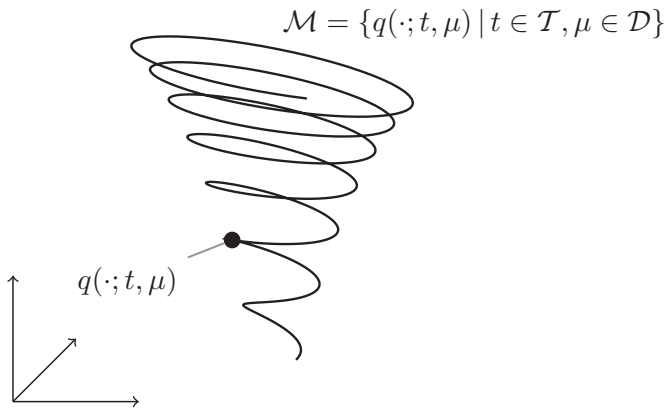


Figure 1. Classical (linear) model reduction seeks to approximate solutions of parametrized PDEs in low-dimensional spaces, which corresponds to a linear approximation of the potentially nonlinear structure of the solution manifold $\mathcal{M}$. In this figure, the spiral depicts a nonlinear solution manifold $\mathcal{M}$, which would be approximated by a straight line with classical model reduction methods.

snapshots are computed, which are numerical PDE solutions at a few training parameters $\mu_1, \dots, \mu_{M_{\text{train}}} \in \mathcal{D}$ obtained with standard numerical methods that solve in $\mathcal{U}$. Then, the first n corresponding principal components of the snapshots are computed to span the reduced space $\mathcal{U}_n$. The principal components depend on a metric that has to be chosen adequately, which leads to weighted principal components; see, e.g., [BGW15]. In model reduction, computing basis vectors via principal component analysis is often referred to as proper orthogonal decomposition (POD) [BGW15].

Numerically, a basis of a POD reduced space can be computed, for example, with the singular value decomposition: For $t_1, \dots, t_K$ and $\mu_1, \dots, \mu_{M_{\text{train}}}$, let

$$q_N(t_i, \mu_j) = [\beta_1(t_i, \mu_j), \dots, \beta_N(t_i, \mu_j)]^T \in \mathbb{R}^N, \qquad (3)$$

be a snapshot and let

$$Q = [q_N(t_1, \mu_1), \dots, q_N(t_K, \mu_{M_{\text{train}}})] \in \mathbb{R}^{N \times K M_{\text{train}}} \qquad (4)$$

be the snapshot matrix. Computing the singular value decomposition of Q and taking the n left-singular vectors corresponding to the largest singular values as columns of the basis matrix U_n leads to the reduced space $\mathcal{U}_n$ spanned by the columns of U_n. The error of projecting a snapshot, i.e., a column of Q, onto the space $\mathcal{U}_n$ is bounded by the sum of the squared singular values with index greater than n

$$\sum_{i=1}^{K} \sum_{j=1}^{M_{\text{train}}} \left\| q_N(t_i, \mu_j) - U_n U_n^T q_N(t_i, \mu_j) \right\|_2^2 = \sum_{i=n+1}^{r} \sigma_i^2,$$

where $r > n$ is the rank of the snapshot matrix Q and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ are the singular values. Thus, the decay of the singular values of the snapshot matrix Q indicates how well the snapshots can be approximated in $\mathcal{U}_n$.

There is a large number of other methods for constructing reduced spaces, such as greedy methods and interpolatory methods; we refer to the surveys [RHP08, BGW15] for more details. All these methods have in common that reduced spaces $\mathcal{U}_n$ are constructed in the training phase and the approximate PDE solutions are then sought in the reduced space for different parameters and initial conditions in the online phase.

Two motivating numerical experiments. Let us apply projection-based model reduction, as described above, to a diffusion problem such as the heat equation with a forcing term

$$\partial_t q(x;t,\mu) - \mu \partial_x^2 q(x;t,\mu) = 1, \qquad x \in \Omega, \qquad (5)$$

with spatial domain $\Omega = (0,1) \subset \mathbb{R}$. We impose homogeneous Dirichlet boundary conditions on the left and right boundary. The initial condition is 0. The parameter $\mu \in \mathcal{D} = [0.1, 10] \subset \mathbb{R}$ is the heat conductivity coefficient and we set it to $\mu = 1$ in this experiment. The equation (5) is discretized with $N = 1024$ linear finite elements in space and implicit Euler in time with time-step size 10^{-3}. The numerical solution up to time $T = 0.4$ is show in Figure 2a.

We collect snapshots (3) over time for parameter $\mu = 1$ and assemble the snapshot matrix (4). Recall that the singular values of the snapshot matrix indicate how well the snapshots can be approximated in the reduced space constructed with the POD procedure. Let $\sigma_1 \geq \cdots \geq \sigma_n$ be the first $n = 150$ singular values of the corresponding snapshot matrix. Figure 2b shows the normalized singular values, where normalized means that the first normalized singular value is one. The decay of the singular values shows that the first 15 left-singular vectors span a reduced space in which the snapshots can be approximated up to machine precision, which results in a dimensionality reduction of a factor of almost 70, namely from dimension $N = 1024$ of the finite-element approximation to $n = 15$ dimensions of the reduced model. The decay of the singular values does not tell us anything about the approximation quality of the space $\mathcal{U}_n$ for solutions of the PDE at other parameters than the ones used for creating the snapshots. However, the decay of the singular values often serves as a useful empirical heuristic for how much reduction can be achieved; a more formal description follows in the next section.

Let us now consider a transport-dominated problem given by the linear advection equation

$$\partial_t q(x;t,\mu) + \mu \partial_x q(x;t,\mu) = 0, \qquad x \in \Omega, \qquad (6)$$

with $\Omega = (0,1)$ and periodic boundary conditions. The parameter μ is the transport speed and fixed to $\mu = 0.8$ in this experiment. The initial condition is a Gaussian probability density function with mean 0.1 and standard deviation 1.5×10^{-2}. The linear advection equation propagates the

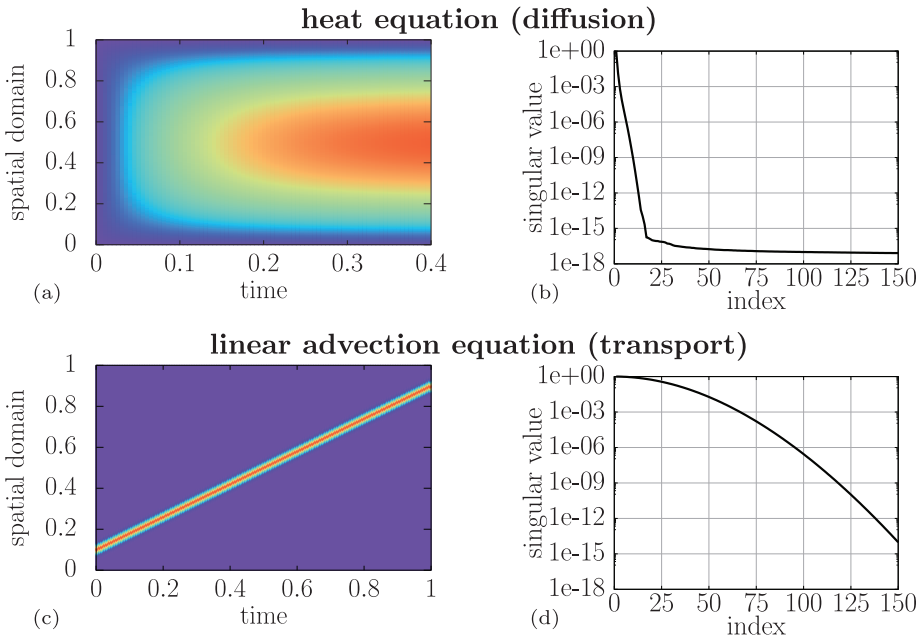


Figure 2. For the heat equation, which describes diffusion-dominated problems, the decay of the singular values indicates that a reduced space of dimension $n = 15$ is sufficient to approximate the snapshots up to machine precision in this example. In contrast, the singular values decay orders of magnitude slower in the case of the linear advection equation, which indicates that classical model reduction that derives linear approximations in spaces can be inefficient for such transport-dominated problems.

initial condition to the right as shown in Figure 2c. We collect snapshots for this problem and compute the normalized singular values, which are shown in Figure 2d. The decay of the singular values is orders of magnitude slower than for the diffusion problem.

In this numerical experiment, the transport-dominated problem requires a higher-dimensional reduced space than the diffusion-dominated problem to achieve a comparable error of approximating the snapshots in the reduced space. Thus, the experiment indicates that projection-based model reduction that computes linear approximations of solutions in a reduced space $\mathcal{U}_n$ obtained with POD, as described above, is less efficient for transport- than for diffusion-dominated problems. In fact, as we will see in the next section, this observation holds more generally and shows that different model reduction methods are needed for diffusion-dominated than for transport-dominated problems.

Limitations of model reduction based on linear approximations. Let us now formalize the numerical observation of the previous section. The key is to understand the lowest error that can be achieved when approximating elements of a PDE-solution manifold $\mathcal{M}$ in vector spaces of dimension n; independent of how the reduced space is constructed. The best-approximation error is given by the Kolmogorov n-width, see, e.g., [MPT02, CD16],

$$d_n(\mathcal{M}) = \inf_{\substack{\mathcal{U}_n \\ \dim(\mathcal{U}_n)=n}} \sup_{q(\cdot;t,\mu)\in\mathcal{M}} \inf_{\tilde{q}\in\mathcal{U}_n} \|q(\cdot;t,\mu) - \tilde{q}\|, \qquad (7)$$

which is the lowest error that any n-dimensional space $\mathcal{U}_n$ can achieve over all elements in $\mathcal{M}$ with respect to the norm $\|\cdot\|$. Other types of Kolmogorov n-widths have been proposed that look at the average error over all elements in $\mathcal{M}$ and that are formulated directly via a metric; see, e.g., [MPT02, CD16, ELMV20]. If $d_n(\mathcal{M})$ decays quickly with n, then there exist low-dimensional spaces $\mathcal{U}_n$ that approximate well the elements of $\mathcal{M}$. For example, the authors of [MPT02] have shown that the Kolmogorov n-width of the solution manifold of a specific elliptic PDE in an appropriate norm decays exponentially fast in the dimension n; more general results for elliptic problems have been derived in [CD16]. In Figure 2a-b, we also observe *numerically* an exponential decay of the projection error of the snapshots. However, it is important to note that the singular values do *not*, in general, correspond to the Kolmogorov n-width, because, e.g., the singular values depend on the snapshots and only lead to a bound on the projection error corresponding to the POD space. In contrast, the Kolmogorov n-width gives the best-approximation error over all possible spaces and is not tied to a specific way of constructing reduced spaces. It can be exceedingly difficult to construct spaces that achieve the best-approximation error given by the Kolmogorov n-width; however, sequences of spaces that obtain the same error rate can be constructed with greedy methods in certain situations [RHP08].

Let us now consider the linear advection equation (6) with a Heaviside step initial condition

$$q_0(x) = \begin{cases} 1, & x \leq 0, \\ 0, & \text{otherwise}. \end{cases}$$

It has been shown in [OR16] that the corresponding solution manifold has a Kolmogorov n-width that cannot decay faster than $1/\sqrt{n}$,

$$d_n(\mathcal{M}) \geq c\frac{1}{\sqrt{n}},$$

where $c > 0$ is a constant independent of n. Even though the decay of the singular values is insufficient to draw conclusions about lower bounds on the Kolmogorov n-width, see comment above, in our numerical experiments, the projection error of the snapshots also decays slower for the transport-dominated problem than for the diffusion-dominated problem. In general, lower bounds on the Kolmogorov n-width of solution manifolds of transport-dominated problems suggest a slow decay. For example, the authors of [GU19] show similarly slow decays for problems governed by the wave equation.

Kolmogorov barrier. A slow decay of the Kolmogorov n-width is sometimes referred to as the *Kolmogorov barrier* because it limits the decay of the error that can be achieved with projection-based model reduction methods that seek linear approximations in spaces.

Nonlinear approximations and model reduction. Nonlinear model reduction methods seek to overcome the Kolmogorov barrier via nonlinear approximations. Let us first consider a *linear* reduced approximation $\tilde{q} \in \mathcal{U}_n$, which we can write as a linear combination

$$\tilde{q}(x; \tilde{\beta}(t,\mu)) = \sum_{i=1}^{n} \tilde{\beta}_i(t,\mu)\phi_i(x) \tag{8}$$

that makes the dependence on the coefficients $\tilde{\beta}(t,\mu) = [\tilde{\beta}_1(t,\mu), \dots, \tilde{\beta}_n(t,\mu)]^T$ explicit. The coefficients $\tilde{\beta}_1(t,\mu), \dots, \tilde{\beta}_n(t,\mu)$ enter linearly in the approximation $\tilde{q}$. Stated differently, the space $\mathcal{U}_n$ spanned by the set of basis functions $\{\phi_i\}_{i=1}^n$ is fixed independent of which element of $q(\cdot;t,\mu) \in \mathcal{M}$ is to be approximated—changing the coefficients $\tilde{\beta}(t,\mu)$ based on the to-be-approximated element $q(\cdot;t,\mu)$ does not change the basis functions. This means that the Kolmogorov n-width applies and it lower bounds the best-approximation error that can be achieved with any reduced space of dimension n.

In contrast, consider now a nonlinear approximation of the form

$$\tilde{q}(x; \tilde{\alpha}(t,\mu), \tilde{\beta}(t,\mu)) = \sum_{i=1}^{n} \tilde{\beta}_i(t,\mu)\phi_i(x; \tilde{\alpha}(t,\mu)), \tag{9}$$

where $\tilde{\alpha}(t,\mu)$ enters nonlinearly in the basis functions $\phi_1, \dots, \phi_n$. Thus, there is a nonlinear dependence of $\tilde{q}$ on $\tilde{\alpha}(t,\mu)$, which is in stark contrast to the linear approximation (8) that depends on $\tilde{\beta}(t,\mu)$ alone and where the coefficients $\tilde{\beta}(t,\mu)$ enter linearly. Stated differently, the nonlinear approximation (9) is a linear combination with functions $\{\phi_i(\cdot; \tilde{\alpha}(t,\mu))\}_{i=1}^n$ that depend through $\tilde{\alpha}(t,\mu)$ on the element $q(\cdot;t,\mu) \in \mathcal{M}$ that is to be approximated, which is different from the linear approximation (8) where the basis functions are fixed independent of which element of $\mathcal{M}$ is approximated.

Even though nonlinear approximations of the form (9) have been studied from a theoretical perspective for a long time, we want to note that they are closely related to deep neural networks, where $\tilde{\alpha}(t,\mu)$ are typically referred to as features, which are learned together with the coefficients $\tilde{\beta}(t,\mu)$. Another class of nonlinear approximation methods selects basis functions from a large dictionary based on the to-be-approximated element. These dictionary-based methods are typically formulated via sparse regression and compressed sensing. In the context of model reduction, dictionary-based methods are sometimes referred to as localized model reduction because reduced spaces are locally varied depending on time, parameters, and/or spatial coordinates [BGW15].

Nonlinear approximations (9) are more expressive than linear approximations (8) in the sense that nonlinear approximations can lead to lower errors than the Kolmogorov n-width for the same number of degrees of freedom; thus, nonlinear approximations can break the Kolmogorov barrier. To see this, consider the linear advection problem (6). In the case of this simple example, the analytic solution can be obtained with the method of characteristics $q(x;t,\mu) = q_0(x - t\mu)$, where q_0 is the initial condition. Building on the nonlinear approximation (9), set $n = 1$ and the function ϕ_1 to

$$\phi_1(x;\theta) = q_0(x - \theta).$$

Then, the nonlinear reduced model

$$\tilde{q}(x; \tilde{\alpha}(t,\mu), \tilde{\beta}(t,\mu)) = \tilde{\beta}_1(t,\mu)\phi_1(x; \tilde{\alpha}_1(t,\mu)) \tag{10}$$

with fixed coefficient $\tilde{\beta}(t,\mu) = [\tilde{\beta}_1(t,\mu)] = [1]$ and feature $\tilde{\alpha}(t,\mu) = [\tilde{\alpha}_1(t,\mu)] = [t\mu]$ exactly represents the solution. Thus, for this example, the nonlinear reduced model (10) breaks the Kolmogorov barrier of linear approximations (8), for which the error cannot decay faster than $1/\sqrt{n}$, where n is the number of degrees of freedom.

Stability and online efficiency. Increasing the expressiveness by breaking the Kolmogorov barrier with nonlinear approximations is only a first step towards nonlinear model reduction of transport-dominated problems. Just as in numerical analysis in general, increasing expressiveness alone is insufficient. Rather, nonlinear reduced models and their underlying nonlinear approximations have to be stable to be useful for numerical computations. Additionally, the goal of model reduction is achieving speedups compared to solving the original, full model, which means that the computational complexity of solving the reduced model online has to scale independently of the dimension N of the full-model approximation space. Thus, a truly practical nonlinear model reduction approach breaks the Kolmogorov barrier in a numerically stable and online efficient way.

Adaptive empirical interpolation: Nonlinear approximations via adaptive spaces. Formulation (9) of nonlinear approximations is typically too general to work with numerically; see also the previous remark on stability and online efficiency. For example, there are no restrictions on the basis functions and their dependence on the features $\tilde{\alpha}(t,\mu)$. We now describe a concrete numerical method for nonlinear model reduction: the adaptive empirical interpolation method (ADEIM) introduced in [PW15, Peh20], where the basis of the reduced space is adapted with low-rank updates.

Consider a time-discrete spatially discretized systems of nonlinear equations

$$q_N^{(k)} = f(q_N^{(k+1)};\mu)\,, \qquad k = 0,\dots,K-1\,,$$

that arises from (2) via, e.g., an implicit Euler discretization. The state vector $q_N^{(k)}$ at time step k is of dimension N and the dynamics are given by the function $f : \mathbb{R}^N \times \mathcal{D} \to \mathbb{R}^N$. Recall that $U_n \in \mathbb{R}^{N\times n}$ is a basis matrix with columns that span the reduced space $\mathcal{U}_n$. Consider first linear model reduction with empirical interpolation [BMNP04, CS10] to obtain the approximation

$$\tilde{f}(\tilde{q};\mu) = (P^T U_n)^{-1} P^T f(U_n\tilde{q};\mu)$$

where $P = [e_{i_1},\dots,e_{i_n}] \in \{0,1\}^{N\times n}$ is a selection matrix with N-dimensional canonical unit vectors $e_{i_1},\dots,e_{i_n}$ that have 1 at components $i_1,\dots,i_n \in \{1,\dots,N\}$, respectively. This means $P^T f(U_n\tilde{q};\mu)$ requires evaluating only the n component functions of f corresponding to the components $i_1,\dots,i_n$ selected by P. The selection matrix is obtained from U_n typically with greedy approaches [BMNP04, CS10]. The corresponding linear, static reduced model is

$$\tilde{q}^{(k)} = \tilde{f}(\tilde{q}^{(k+1)};\mu)\,, \qquad k = 0,\dots,K-1\,. \tag{11}$$

In adaptive empirical interpolation [PW15, Peh20], the basis matrix depends on the time step k and is adapted via low-rank updates as

$$U_n^{(k+1)} = U_n^{(k)} + \alpha_k \beta_k^T\,,$$

where $\alpha_k \in \mathbb{R}^{N\times z}$ and $\beta_k \in \mathbb{R}^{n\times z}$ and z is the rank of the update, which is in contrast to static (linear) empirical interpolation (11) where the space is independent of the time step. The update $\alpha_k\beta_k^T$ is obtained via an optimization problem

$$\min_{\alpha\in\mathbb{R}^{N\times z},\beta\in\mathbb{R}^{n\times z}} \left\| S_k^T\left((U_n^{(k)} + \alpha_k\beta_k^T)C_k - F_k\right)\right\|_F^2$$

where $S_k \in \{0,1\}^{N\times m}$ is a sampling matrix that selects m components, similarly to the selection matrix in static empirical interpolation. The coefficient matrix is $C_k = (P_k^T U_n^{(k)})^{-1} P_k^T F_k$ and $F_k \in \mathbb{R}^{N\times w}$ is the right-hand side matrix of window size w. The update $\alpha_k\beta_k^T$ can be obtained via a singular value decomposition of an $n\times w$ matrix. The selection matrix P_k also depends on the time step k and is adapted by either re-running the greedy selection procedures [BMNP04, CS10] or via low-rank updates [PW15]. The right-hand side matrix $F_k = [\hat{q}^{(k-w-1)},\dots,\hat{q}^{(k)}]$ is assembled by evaluating the full-model right-hand side function f at the m components selected by S_k and approximating all other components as

$$\begin{aligned} S_k\hat{q}^{(k)} &= S_k f(U_n^{(k)}\tilde{q}^{(k)};\mu),\\ \check{S}_k\hat{q}^{(k)} &= \check{S}_k U_n^{(k)}(P_k^T U_n^{(k)})^{-1} P_k^T f(U_n^{(k)}\tilde{q}^{(k)};\mu)\,, \end{aligned}$$

where $\check{S}_k$ is the complementary sampling points matrix that selects the components not selected by S_k. The sampling points S_k are adapted via greedy strategies, for which several strategies have been proposed, including a computationally efficient strategy in [Peh20].

In summary, the process of the adaptive empirical interpolation method is to adapt the space $\mathcal{U}_n^{(k)}$ at time step k to the space $\mathcal{U}_n^{(k+1)}$ at time step $k+1$. The adaptation is achieved by applying a low-rank update to the basis matrix $U_n^{(k)}$ to obtain the basis matrix $U_n^{(k+1)}$ of the adapted space $\mathcal{U}_n^{(k+1)}$. The update is computed from sparse evaluations of the full-model right-hand side function f at a few selected components; we refer to [Peh20] for technical details.

Adaptive empirical interpolation: Nonlinear model reduction for predicting limit cycle oscillations in a combustor. We apply the adaptive empirical interpolation as a nonlinear model reduction approach to a quasi-1D model of a single-element rocket combustor, which is described

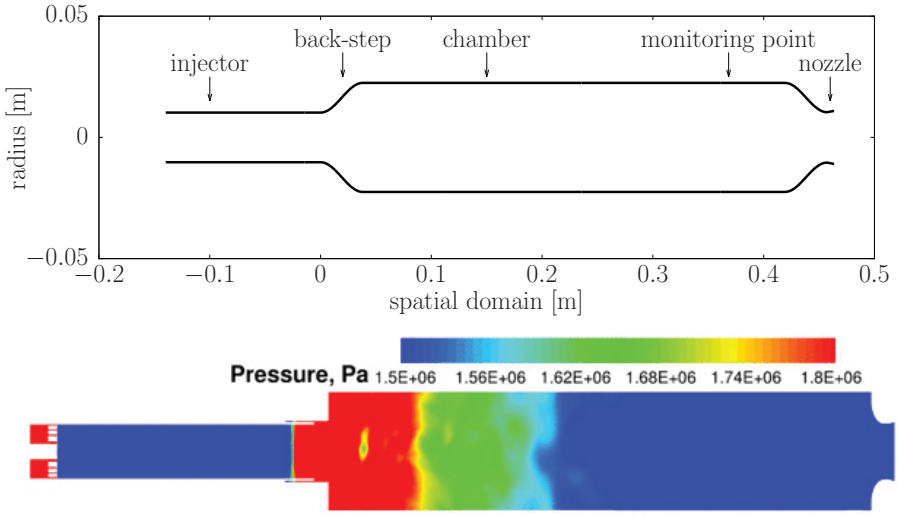


Figure 3. Top: Numerically predicting the growth of the amplitude of pressure oscillations at the monitoring point in the combustor chamber helps to derive designs that prevent combustion instabilities. Bottom: Pressure field of a 2D version of the quasi-1D model combustor considered in this experiment. Pressure waves traveling through the combustion chamber make this problem transport dominated, which motivates the reduction with nonlinear methods such as adaptive empirical interpolation [PW15, Peh20].

in [XD17]. The goal is to predict the growth of the amplitude of pressure oscillations at a monitoring point, which provides critical insights for designing engines that avoid combustion instabilities that are caused by unbounded growth of the amplitude of the pressure oscillations. The pressure oscillations lead to waves traveling through the combustion chamber that make this problem transport dominated and thus linear model reduction methods fail for this problem; cf. [XD17, Peh20].

Figure 3 shows the setup of the problem. The oxidizer is induced and meets the fuel at the back-step, where it reacts instantaneously. The combustion products exit the chamber through the nozzle. The combustion follows a one-step reaction model,

$$CH_4 + 2O_2 \to CO_2 + 2H_2O\,,$$

where the fuel is gaseous methane and the oxidizer is a mixture of oxygen and water. The parameter μ of the problem controls the heat release. The governing equations of the model combustor are described in detail in [XD17]. The following numerical results summarize the experiments conducted in [Peh20]. Figure 4(top) shows the pressure at a monitoring point for heat release $\mu = 3.0$, where the combustor enters a steady state. In contrast, for heat release parameter $\mu = 3.8$, the system enters a limit cycle oscillation as shown in Figure 4(bottom). In both cases, the adaptive reduced model faithfully approximates the full model while achieving a speedup of a factor 6–8 over various heat-release parameters. Thus, the nonlinear reduced model enables quickly sweeping over a large range of parameters for informing early design decisions to prevent an unbounded growth of the pressure amplitude.

Conclusions and open questions. There is a clear need for nonlinear model reduction methods to derive efficient reduced models of transport-dominated problems in science and engineering. This note focused on increasing expressiveness compared to linear model reduction to break the Kolmogorov barrier. However, increasing expressiveness alone is insufficient for truly practical nonlinear model reduction methods. Rather, nonlinear model reduction methods also have to be numerically stable, just as traditional methods in scientific computing, which has received little attention in nonlinear model reduction. Additionally, the purpose of model reduction is to obtain speedups: First, constructing nonlinear reduced models in the training phase has to be cheaper in terms of, e.g., data volume and training time than solving the outer-loop task with the original, full model in the first place. This can be challenging to achieve with data-hungry machine learning methods. Second, it is paramount that solving nonlinear reduced models at new parameters in the online phase is computationally cheaper than solving the full model. The ultimate goal is to achieve online efficiency in nonlinear model reduction in the sense that the cost complexity of solving the nonlinear reduced model at new parameters scales independently of the dimension of the full approximation space. An often overlooked aspect is that nonlinear model reduction methods have to be easy to use for achieving wide acceptance in the domain sciences and engineering communities, which is getting increasingly more attention via nonintrusive methods that learn reduced models from data [IA14, PW16, HU18].

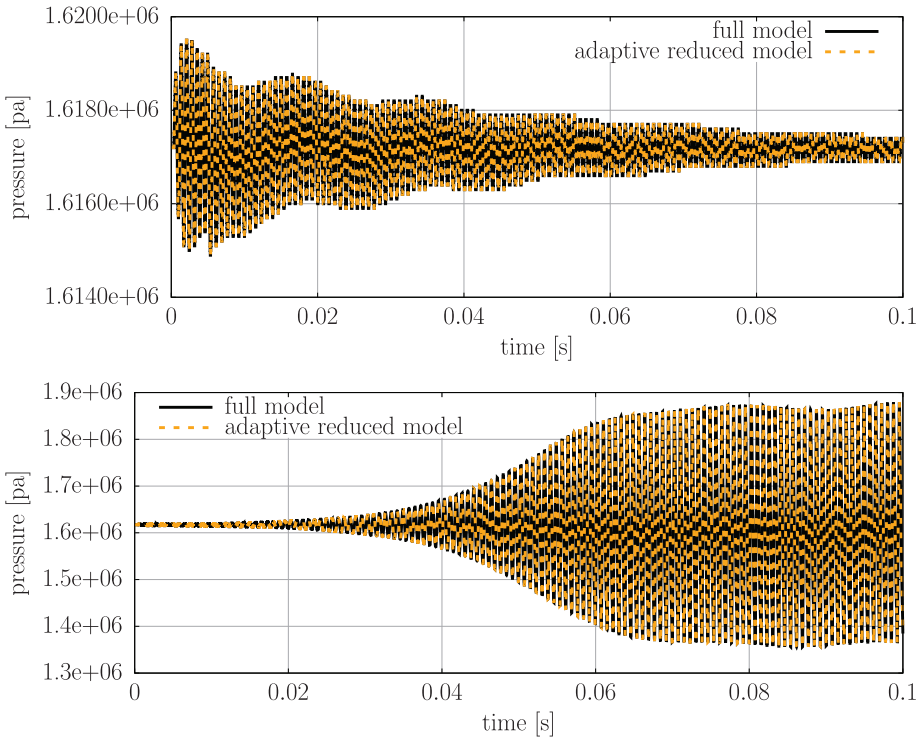


Figure 4. The nonlinear reduced model based on adaptive empirical interpolation faithfully predicts the pressure oscillation in this model combustor for low (top) and high (bottom) heat release, which enables quickly sweeping over parameters to support decision-making in early design stages.

Nonlinear model reduction is at its early stages. It will require considerable progress of mathematical theory and computational methods—bringing together machine learning and scientific computing—to advance nonlinear model reduction into a similarly rigorous, reliable, flexible, and ubiquitous tool of science and engineering as linear model reduction is today.

This manuscript contains only a limited number of references because journal rules restrict the maximum number of references to 20; additional references to other nonlinear model reduction methods are cited in the manuscript [Peh20].

ACKNOWLEDGMENTS. The author was partially supported by NSF under award DMS-2046521 and the Air Force Center of Excellence on Multi-Fidelity Modeling of Rocket Combustor Dynamics under Award Number FA9550-17-1-0195.

References

[BMNP04] Maxime Barrault, Yvon Maday, Ngoc Cuong Nguyen, and Anthony T. Patera, *An 'empirical interpolation' method: application to efficient reduced-basis discretization of partial differential equations* (English, with English and French summaries), C. R. Math. Acad. Sci. Paris **339** (2004), no. 9, 667–672, DOI 10.1016/j.crma.2004.08.006. MR2103208

[BGW15] Peter Benner, Serkan Gugercin, and Karen Willcox, *A survey of projection-based model reduction methods for parametric dynamical systems*, SIAM Rev. **57** (2015), no. 4, 483–531, DOI 10.1137/130932715. MR3419868

[CS10] Saifon Chaturantabut and Danny C. Sorensen, *Nonlinear model reduction via discrete empirical interpolation*, SIAM J. Sci. Comput. **32** (2010), no. 5, 2737–2764, DOI 10.1137/090766498. MR2684735

[CD16] Albert Cohen and Ronald DeVore, *Kolmogorov widths under holomorphic mappings*, IMA J. Numer. Anal. **36** (2016), no. 1, 1–12. MR3463432

[ELMV20] Virginie Ehrlacher, Damiano Lombardi, Olga Mula, and François-Xavier Vialard, *Nonlinear model reduction on metric spaces. Application to one-dimensional conservative PDEs in Wasserstein spaces*, ESAIM Math. Model. Numer. Anal. **54** (2020), no. 6, 2159–2197, DOI 10.1051/m2an/2020013. MR4169690

[GU19] Constantin Greif and Karsten Urban, *Decay of the Kolmogorov N-width for wave problems*, Appl. Math. Lett. **96** (2019), 216–222, DOI 10.1016/j.aml.2019.05.013. MR3953419

[HU18] J. S. Hesthaven and S. Ubbiali, *Non-intrusive reduced order modeling of nonlinear problems using neural networks*, J. Comput. Phys. **363** (2018), 55–78, DOI 10.1016/j.jcp.2018.02.037. MR3784416

[IA14] A. C. Ionita and A. C. Antoulas, *Data-driven parametrized model reduction in the Loewner framework*, SIAM J. Sci. Comput. **36** (2014), no. 3, A984–A1007. MR3209730

[LC20] Kookjin Lee and Kevin T. Carlberg, *Model reduction of dynamical systems on nonlinear manifolds using deep convolutional autoencoders*, J. Comput. Phys. **404** (2020), 108973, 32, DOI 10.1016/j.jcp.2019.108973. MR4043884

[MPT02] Yvon Maday, Anthony T. Patera, and G. Turinici, *Global a priori convergence theory for reduced-basis approximations of single-parameter symmetric coercive elliptic partial differential equations* (English, with English and French summaries), C. R. Math. Acad. Sci. Paris **335** (2002), no. 3, 289–294, DOI 10.1016/S1631-073X(02)02466-4. MR1933676

[OR13] Mario Ohlberger and Stephan Rave, *Nonlinear reduced basis approximation of parameterized evolution equations via the method of freezing* (English, with English and French summaries), C. R. Math. Acad. Sci. Paris **351** (2013), no. 23-24, 901–906, DOI 10.1016/j.crma.2013.10.028. MR3133601

[OR16] Mario Ohlberger and Stephan Rave, *Reduced basis methods: Success, limitations and future challenges*, Proceedings of the Conference Algoritmy (2016), 1–12.

[Peh20] Benjamin Peherstorfer, *Model reduction for transport-dominated problems via online adaptive bases and adaptive sampling*, SIAM J. Sci. Comput. **42** (2020), no. 5, A2803–A2836, DOI 10.1137/19M1257275. MR4151479

[PW15] Benjamin Peherstorfer and Karen Willcox, *Online adaptive model reduction for nonlinear systems via low-rank updates*, SIAM J. Sci. Comput. **37** (2015), no. 4, A2123–A2150, DOI 10.1137/140989169. MR3384839

[PW16] Benjamin Peherstorfer and Karen Willcox, *Data-driven operator inference for nonintrusive projection-based model reduction*, Comput. Methods Appl. Mech. Engrg. **306** (2016), 196–215, DOI 10.1016/j.cma.2016.03.025. MR3502565

[RSSM18] J. Reiss, P. Schulze, J. Sesterhenn, and V. Mehrmann, *The shifted proper orthogonal decomposition: a mode decomposition for multiple transport phenomena*, SIAM J. Sci. Comput. **40** (2018), no. 3, A1322–A1344. MR3799062

[RHP08] G. Rozza, D. B. P. Huynh, and A. T. 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), no. 3, 229–275, DOI 10.1007/s11831-008-9019-9. MR2430350

[SL09] Themistoklis P. Sapsis and Pierre F. J. Lermusiaux, *Dynamically orthogonal field equations for continuous stochastic dynamical systems*, Phys. D **238** (2009), no. 23-24, 2347–2360, DOI 10.1016/j.physd.2009.09.017. MR2576078

[TPQ15] T. Taddei, S. Perotto, and A. Quarteroni, *Reduced basis techniques for nonlinear conservation laws*, ESAIM Math. Model. Numer. Anal. **49** (2015), no. 3, 787–814, DOI 10.1051/m2an/2014054. MR3342228

[XD17] Jiayang Xu and Karthik Duraisamy, *Reduced-Order Modeling of Model Rocket Combustors*, 53rd AIAA/SAE/ASEE Joint Propulsion Conference, July 2017.

Benjamin Peherstorfer

Credits

Opening image and Figures 1, 2, and 4 are courtesy of Benjamin Peherstorfer.

Figure 3 is courtesy of Benjamin Peherstorfer and Cheng Huang.

Photo of Benjamin Peherstorfer is courtesy of NYU/DanCreighton.com.