Rigorous Error Bounds on the Solutions of Projection-Based Reduced Models

Geoffrey M. Oxberry*, William H. Green,*, Paul I. Barton*
*Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Increasing demands on the physical fidelity of scientific simulations have placed strenuous demands on available computational resources. In many situations, such as combustion, astrophysics, circuit simulation, electromagnetics, climate modeling, and fluid mechanics, the demand for realistic detail in nonlinear ODE and PDE based-models requires model reduction in order to obtain the quality of numerical results desired, subject to constraints on computing power and time.

One popular category of model reduction techniques is projection-based model reduction [1]. This class of techniques takes the right-hand side of an ODE (1)

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), t) \tag{1}$$

and determines a projection matrix $\mathbf{P}$ from this right-hand side and some additional data (typically reference values for state variables and error tolerances). This right-hand side is called the original model. The reduced model (2) constructed by the technique is the right-hand side of the ODE, premultiplied by the projection matrix determined by the technique:

$$\dot{\hat{\mathbf{x}}}(t) = \mathbf{P}\mathbf{f}(\hat{\mathbf{x}}(t), t). \tag{2}$$

Despite the considerable number of projection-based model reduction techniques in the literature, no bounds exist on the approximation error in the solution of a generic projection-based reduced model ODE relative to the solution of its corresponding original model ODE.

In the case of orthogonal projectors, bounds do exist on this approximation error. A classical result on error bounds for proper orthogonal decomposition exists for linear systems [2], [3]. Rathinam and Petzold [4] proved a result on error bounds for nonlinear systems, provided that the projection-based model reduction technique determines an orthogonal projection matrix. However, most projection-based model reduction techniques, including balanced truncation [3], computational singular perturbation [5], and the linearized quasi-steady state approximation [6], construct skew projectors, to which the Rathinam and Petzold result does not apply.

The contribution of this work will be to extend the result by Rathinam and Petzold [4] to general (skew) projection matrices. This result yields the first error bound on solutions of generic projection-based reduced models relative to their corresponding original model solutions. The result places no restrictions on how the projection matrix is constructed, enabling it to be used on all projection-based model reduction techniques. Consequently, provided that bounds on the logarithmic norm of the Jacobian of the right-hand side of the reduced model ODE and bounds on the Lipschitz constant of the right-hand side of the reduced model ODE can both be calculated, bounds on the approximation error due to model reduction can be obtained for a wide range of model reduction techniques, enabling techniques to be compared on the basis of approximation error.

Extension of the result by Rathinam and Petzold [4] requires the following definitions: Let $\mathbf{P} \in \mathbb{R}^{n\times n}$ be a projection matrix of rank k. Consider solving the initial value problem (1) subject to $\mathbf{x}(0) = \mathbf{x}_0 \in \mathbb{R}^n$ using the reduced order model (2) subject to $\hat{\mathbf{x}}(0) = \mathbf{P}(\mathbf{x}_0 - \mathbf{x}_1) + \mathbf{x}_1 \in \mathbb{R}^n$ in the interval $[0, T]$. Let $\mathbf{D} \in \mathbb{R}^{n\times k}$ and $\mathbf{A} \in \mathbb{R}^{k\times n}$ be such that $\mathbf{P} = \mathbf{D}\mathbf{A}$ is a full rank decomposition of $\mathbf{P}$. Let $\mathbf{B} \in \mathbb{R}^{n\times(n-k)}$ and $\mathbf{C} \in \mathbb{R}^{(n-k)\times n}$ be matrices such that $\mathbf{I} - \mathbf{P} = \mathbf{B}\mathbf{C}$ is a full rank decomposition of $\mathbf{I} - \mathbf{P}$. Write the solution (of the original model) $\mathbf{x}(t)$ and the solution $\hat{\mathbf{x}}(t)$ (of the reduced model) as

$$\mathbf{x}(t) = \mathbf{D}\mathbf{u}(t) + \mathbf{B}\mathbf{v}(t) + \mathbf{x}_1 \tag{3}$$
$$\hat{\mathbf{x}}(t) = \mathbf{D}\mathbf{u}(t) + \mathbf{D}\mathbf{w}(t) + \mathbf{x}_1, \tag{4}$$

where $\mathbf{u}(t) \in \mathbb{R}^k$, $\mathbf{w}(t) \in \mathbb{R}^k$, and $\mathbf{v}(t) \in \mathbb{R}^{(n-k)}$.

Let $\mathbf{e}(t) = \hat{\mathbf{x}}(t) - \mathbf{x}(t)$ be the approximation error incurred due to model reduction. Define $\mathbf{e}_i(t)$ and $\mathbf{e}_o(t)$ such that

$$\mathbf{e}_i(t) = \mathbf{P}\mathbf{e}(t) \tag{5}$$
$$\mathbf{e}_o(t) = (\mathbf{I} - \mathbf{P})\mathbf{e}(t), \tag{6}$$

such that $\mathbf{e}(t) = \mathbf{e}_i(t) + \mathbf{e}_o(t)$ and define $\tilde{\mathbf{x}}(t) = \mathbf{P}(\mathbf{x}(t) - \mathbf{x}_1) + \mathbf{x}_1$. Then it follows that

$$\mathbf{e}_o(t) = -\mathbf{B}\mathbf{v}(t), \tag{7}$$
$$\mathbf{e}_i(t) = \mathbf{D}\mathbf{w}(t), \tag{8}$$
$$\tilde{\mathbf{x}}(t) = \mathbf{D}\mathbf{u}(t) + \mathbf{x}_1. \tag{9}$$

Let $\gamma \geq 0$ be the Lipschitz constant of $\mathbf{A}\mathbf{f}(\mathbf{x}, t)$ in the directions corresponding to $\mathcal{N}(\mathbf{A})$ in a region containing $\mathbf{x}(t)$ and $\tilde{\mathbf{x}}(t)$. To be precise, suppose that

$$\|\mathbf{A}\mathbf{f}(\tilde{\mathbf{x}}(t) + \mathbf{B}\mathbf{v}, t) - \mathbf{A}\mathbf{f}(\tilde{\mathbf{x}}(t), t)\| \leq \gamma\|\mathbf{v}\| \tag{10}$$

for all $(\mathbf{v},t) \in D' \subset \mathbb{R}^{(n-k)} \times [0,T]$, where the region D' is such that the associated region $D'' = \{(\tilde{\mathbf{x}}(t) + \mathbf{B}\mathbf{v},t) : (\mathbf{v},t) \in D'\}$ contains $(\tilde{\mathbf{x}}(t),t)$ and $(\mathbf{x}(t),t)$ for all $t \in [0,T]$. Define $\mu(\mathbf{X})$ to be the logarithmic norm (related to the 2-norm) of $\mathbf{X}$:

$$\mu(\mathbf{X}) = \lim_{h\to 0, h>0} \frac{\|\mathbf{I} + h\mathbf{X}\| - 1}{h}. \tag{11}$$

Let $\mu(\mathbf{A}D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_1 + \mathbf{D}\mathbf{z},t)\mathbf{D}) \le \bar{\mu}$ for $(\mathbf{z},t) \in V \subset \mathbb{R}^k \times [0,T]$, where the region V is such that it contains $(\mathbf{u}(t),t)$ and $(\mathbf{u}(t) + \mathbf{w}(t),t)$ for all $t \in [0,T]$. Let ε be defined by

$$\varepsilon = \|\mathbf{e}_o\| = \left(\int_0^T \|\mathbf{e}_o(t)\|^2 \, \mathrm{d}t \right)^{1/2}. \tag{12}$$

Given this background, the main result of this work is as follows:

Theorem 1: The function $\mathbf{e}_i$ satisfies

$$\begin{aligned} &\inf\{C \ge 0 : |e_i(t)| \le C \text{ a.e. on } [0,T] \} = \\ &\|\mathbf{e}_i\|_\infty \le \frac{\varepsilon\gamma}{(2\bar{\mu})^{1/2}} (e^{2\bar{\mu}T} - 1)^{1/2} \|\mathbf{D}\| \|\mathbf{C}\|, \end{aligned} \tag{13}$$

and the 2-norm of the function $\mathbf{e}$ satisfies

$$\begin{aligned} &\left(\int_0^T \|\mathbf{e}(t)\|^2 \, \mathrm{d}t \right)^{1/2} = \\ &\|\mathbf{e}\| \le \varepsilon \left(1 + \frac{\gamma}{2\bar{\mu}} (e^{2\bar{\mu}T} - 1 - 2\bar{\mu}T)^{1/2} \|\mathbf{D}\| \|\mathbf{C}\| \right). \end{aligned} \tag{14}$$

Proof: The proof follows the development of Proposition 4.2 in [4]. Since $\mathbf{e}_i(t) = \mathbf{D}\mathbf{w}(t)$ and $\mathbf{A}\mathbf{D} = \mathbf{I}_k$, it follows that $\mathbf{A}\mathbf{e}_i(t) = \mathbf{w}(t)$, so

$$\dot{\mathbf{w}}(t) = \mathbf{A}\dot{\mathbf{e}}_i(t) = \mathbf{A}\mathbf{f}(\hat{\mathbf{x}}(t),t) - \mathbf{A}\mathbf{f}(\mathbf{x}(t),t) \tag{15}$$

It follows that

$$\begin{aligned} \dot{\mathbf{w}}(t) = {} & \mathbf{A}\mathbf{f}(\mathbf{x}_1 + \mathbf{D}\mathbf{u}(t) + \mathbf{D}\mathbf{w}(t),t) \\ & - \mathbf{A}\mathbf{f}(\mathbf{x}_1 + \mathbf{D}\mathbf{u}(t) + \mathbf{B}\mathbf{v}(t),t). \end{aligned} \tag{16}$$

Applying a Taylor expansion for $h > 0$, $\mathbf{w}(t+h) = \mathbf{w}(t) + h\dot{\mathbf{w}}(t) + O(h^2)$, from which it follows that

$$\begin{aligned} \|\mathbf{w}(t+h)\| &= \|\mathbf{w}(t) + h\dot{\mathbf{w}}(t) + O(h^2)\|, \\ &= \|\mathbf{w}(t) + h\mathbf{A}\mathbf{f}(\mathbf{x}_1 + \mathbf{D}\mathbf{u}(t) + \mathbf{D}\mathbf{w}(t),t) \\ &\quad - h\mathbf{A}\mathbf{f}(\mathbf{x}_1 + \mathbf{D}\mathbf{u}(t) + \mathbf{B}\mathbf{v}(t),t) \\ &\quad + O(h^2)\|. \end{aligned} \tag{17}$$

Applying the triangle inequality to the previous inequality (17) yields

$$\begin{aligned} \|\mathbf{w}(t+h)\| \le {} & \|\mathbf{w}(t) + h\mathbf{A}\mathbf{f}(\mathbf{x}_1 + \mathbf{D}\mathbf{u}(t) + \mathbf{D}\mathbf{w}(t),t) \\ & - h\mathbf{A}\mathbf{f}(\mathbf{x}_1 + \mathbf{D}\mathbf{u}(t),t)\| \\ & + h\|\mathbf{A}\mathbf{f}(\mathbf{x}_1 + \mathbf{D}\mathbf{u}(t) + \mathbf{B}\mathbf{v}(t),t) \\ & - \mathbf{A}\mathbf{f}(\mathbf{x}_1 + \mathbf{D}\mathbf{u}(t),t)\| + O(h^2). \end{aligned} \tag{18}$$

Let $\mathbf{g} : \mathbb{R}^k \to \mathbb{R}^k$ be the function

$$\mathbf{g}(\boldsymbol{\eta}) = \boldsymbol{\eta} + h\mathbf{A}\mathbf{f}(\mathbf{x}_1 + \mathbf{D}\boldsymbol{\eta},t). \tag{19}$$

Then

$$\begin{aligned} &\|\mathbf{w}(t) + h\mathbf{A}\mathbf{f}(\mathbf{x}_1 + \mathbf{D}\mathbf{u}(t) + \mathbf{D}\mathbf{w}(t),t) \\ &\quad - h\mathbf{A}\mathbf{f}(\mathbf{x}_1 + \mathbf{D}\mathbf{u}(t),t)\| = \\ &\|\mathbf{g}(\mathbf{u}(t) + \mathbf{w}(t)) - \mathbf{g}(\mathbf{u}(t))\|. \end{aligned} \tag{20}$$

Applying the multivariate mean value theorem to $\mathbf{g}$ yields

$$\|\mathbf{g}(\mathbf{u}(t) + \mathbf{w}(t)) - \mathbf{g}(\mathbf{u}(t))\| \le \kappa \|\mathbf{w}(t)\|, \tag{21}$$

where

$$\begin{aligned} \kappa &\ge \sup_{\boldsymbol{\eta} \in [\mathbf{u}(t), \mathbf{u}(t) + \mathbf{w}(t)]} \|D\mathbf{g}(\boldsymbol{\eta})\| \\ &\ge \sup_{\boldsymbol{\eta} \in [\mathbf{u}(t), \mathbf{u}(t) + \mathbf{w}(t)]} \|\mathbf{I}_k + h\mathbf{A}D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_1 + \mathbf{D}\boldsymbol{\eta},t)\mathbf{D}\|. \end{aligned} \tag{22}$$

Here, for any two vectors $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \mathbb{R}^k$, $[\boldsymbol{\eta}_1, \boldsymbol{\eta}_2]$ denotes the line segment joining the two. (Traditionally, this bracket notation refers to intervals; however, the convention used by Petzold and Rathinam [4] is followed here.) Since the line $[\mathbf{u}(t), \mathbf{u}(t) + \mathbf{w}(t)]$ is a compact subset of $\mathbb{R}^k$,

$$\begin{aligned} &\sup_{\boldsymbol{\eta} \in [\mathbf{u}(t), \mathbf{u}(t) + \mathbf{w}(t)]} \|\mathbf{I}_k + h\mathbf{A}D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_1 + \mathbf{D}\boldsymbol{\eta},t)\mathbf{D}\| = \\ &\max_{\boldsymbol{\eta} \in [\mathbf{u}(t), \mathbf{u}(t) + \mathbf{w}(t)]} \|\mathbf{I}_k + h\mathbf{A}D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_1 + \mathbf{D}\boldsymbol{\eta},t)\mathbf{D}\|. \end{aligned}$$

It follows from (20), (21), (22), and (18) that

$$\begin{aligned} &\|\mathbf{w}(t+h)\| - \|\mathbf{w}(t)\| \le \\ &\left(\max_{\boldsymbol{\eta} \in [\mathbf{u}(t), \mathbf{u}(t) + \mathbf{w}(t)]} \|\mathbf{I}_k + h\mathbf{A}D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_1 + \mathbf{D}\boldsymbol{\eta},t)\mathbf{D}\| - 1 \right) \cdot \\ &\|\mathbf{w}(t)\| + h\|\mathbf{A}\mathbf{f}(\mathbf{x}_1 + \mathbf{D}\mathbf{u}(t) + \mathbf{B}\mathbf{v}(t),t) \\ &- \mathbf{A}\mathbf{f}(\mathbf{x}_1 + \mathbf{D}\mathbf{u}(t),t)\|, \\ &\le \left(\max_{\boldsymbol{\eta} \in [\mathbf{u}(t), \mathbf{u}(t) + \mathbf{w}(t)]} \|\mathbf{I}_k + h\mathbf{A}D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_1 + \mathbf{D}\boldsymbol{\eta},t)\mathbf{D}\| - 1 \right) \cdot \\ &\|\mathbf{w}(t)\| + h\gamma\|\mathbf{v}(t)\| + O(h^2), \end{aligned} \tag{23}$$

which implies that

$$\frac{\|\mathbf{w}(t+h)\| - \|\mathbf{w}(t)\|}{h} \leq \bar{\mu}\|\mathbf{w}(t)\| + \gamma\|\mathbf{v}(t)\| + O(h), \tag{24}$$

where the $O(h)$ term may be uniformly bounded independent of $\mathbf{w}(t)$ (see [7], Equations 10.17 and 10.18). Then it follows from Theorem 10.6 of [7] that

$$\|\mathbf{w}(t)\| \leq \gamma \int_0^t e^{\bar{\mu}(t-\tau)}\|\mathbf{v}(t)\|\,\mathrm{d}\tau. \tag{25}$$

Since $\mathbf{e}_i(t) = \mathbf{D}\mathbf{w}(t)$, it follows that

$$\|\mathbf{e}_i(t)\| \leq \|\mathbf{D}\|\|\mathbf{w}(t)\| \leq \|\mathbf{D}\|\gamma \int_0^t e^{\bar{\mu}(t-\tau)}\|\mathbf{v}(t)\|\,\mathrm{d}\tau. \tag{26}$$

After applying the Cauchy-Schwarz inequality on the right-hand side, it follows that

$$\|\mathbf{e}_i(t)\| \leq \|\mathbf{D}\|\frac{\gamma}{(2\bar{\mu})^{1/2}}(e^{2\bar{\mu}t} - 1)^{1/2}\left(\int_0^t \|\mathbf{v}(t)\|^2\,\mathrm{d}\tau\right)^{1/2}. \tag{27}$$

Since $\mathbf{v}(t) = -\mathbf{C}\mathbf{e}_o(t)$, it follows that

$$\|\mathbf{e}_i(t)\| \leq \|\mathbf{D}\|\|\mathbf{C}\|\frac{\gamma}{(2\bar{\mu})^{1/2}}(e^{2\bar{\mu}t} - 1)^{1/2}\left(\int_0^t \|\mathbf{e}_o(t)\|^2\,\mathrm{d}\tau\right)^{1/2}, \tag{28}$$

from which it follows that

$$\|\mathbf{e}_i\|_\infty \leq \frac{\varepsilon\gamma}{(2\bar{\mu})^{1/2}}(e^{2\bar{\mu}T} - 1)^{1/2}\|\mathbf{D}\|\|\mathbf{C}\|. \tag{29}$$

Substituting 12 and then squaring (28) and integrating yields the bound

$$\|\mathbf{e}_i\| \leq \frac{\varepsilon\gamma}{2\bar{\mu}}(e^{2\bar{\mu}T} - 1 - 2\bar{\mu}T)^{1/2}\|\mathbf{D}\|\|\mathbf{C}\|. \tag{30}$$

Applying the triangle inequality yields

$$\begin{aligned}\|\mathbf{e}\| &\leq \|\mathbf{e}_i\| + \|\mathbf{e}_o\| \\ &\leq \varepsilon\left(1 + \frac{\gamma}{2\bar{\mu}}(e^{2\bar{\mu}T} - 1 - 2\bar{\mu}T)^{1/2}\|\mathbf{D}\|\|\mathbf{C}\|\right).\end{aligned} \tag{31}$$

■

It is worth noting that the bounds obtained for the skew projector case are weaker than the case for the orthogonal projector case because the matrices $\mathbf{C}$ and $\mathbf{D}$ no longer have norm 1, and the Pythagorean theorem cannot be used to obtain a bound on $\|\mathbf{e}\|$, because $\mathbf{e}_i(t)$ and $\mathbf{e}_o(t)$ are not necessarily orthogonal for any given t.

Using this new result, error bounds are illustrated for a linear case study and a simple nonlinear case study, demonstrating its validity.

In the future, this result will be augmented by extending from orthogonal projectors to general (skew) projection matrices a computational technique used by Homescu, *et al.* [8] to estimate the approximation error incurred by projection-based model reduction techniques. In the event that it is not possible to calculate bounds on the logarithmic norm of the Jacobian of the reduced model or the Lipschitz constant of the right-hand side of the reduced model ODE, this new technique can be used to give modelers a better sense of the physical and numerical fidelity of their reduced models. Both the rigorous error bounds presented here and estimates to be developed in future work will enable the construction of reduced models that reduce the computational effort needed to carry out their simulations while providing the error information needed for reduced models to be useful in applications.

REFERENCES

[1] G. M. Oxberry, W. H. Green, and P. I. Barton, "Affine Lumping Formalism for Comparison of Projection-Based Model Reduction Techniques," in *2009 AIChE Annual Meeting*, (Nashville, Tennessee), 2009.

[2] M. T. Chu, R. E. Funderlic, and G. H. Golub, "A Rank-One Reduction Formula and its Applications to Matrix Factorizations," *SIAM Review*, vol. 37, pp. 512–530, December 1995.

[3] A. Antoulas and D. Sorensen, "Approximation of large-scale dynamical systems: An overview," in *Large Scale Systems 2004: Theory and Applications (LSS'04): a Proceedings Volume from the 10th IFAC/IFORS/IMACS/IFIP Symposium, Osaka, Japan, 26-28 July 2004*, vol. 11, p. 19, Elsevier for the International Federation of Automatic Control, 2005.

[4] M. Rathinam and L. Petzold, "A new look at proper orthogonal decomposition," *SIAM Journal on Numerical Analysis*, vol. 41, no. 5, pp. 1893–1925, 2004.

[5] S. Lam and D. Goussis, "The CSP method for simplifying kinetics," *International Journal of Chemical Kinetics*, vol. 26, no. 4, pp. 461–486, 1994.

[6] T. Lu and C. Law, "Systematic approach to obtain analytic solutions of quasi steady state species in reduced mechanisms," *Journal of Physical Chemistry A*, vol. 110, pp. 13202–13208, December 2006.

[7] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I: Nonstiff Problems*. Springer, 2008.

[8] C. Homescu, L. Petzold, and R. Serban, "Error estimation for reduced-order models of dynamical systems," *SIAM Journal on Numerical Analysis*, vol. 43, no. 4, pp. 1693–1714, 2006.