


Christopher Beattie · Peter Benner ·
Mark Embree · Serkan Gugercin ·
Sanda Lefteriu *Editors*

Realization and Model Reduction of Dynamical Systems

A Festschrift in Honor of the
70th Birthday of Thanos Antoulas

 Springer

Realization and Model Reduction of Dynamical Systems

Christopher Beattie · Peter Benner · Mark Embree ·
Serkan Gugercin · Sanda Lefteriu
Editors

Realization and Model Reduction of Dynamical Systems

A Festschrift in Honor of the 70th Birthday
of Thanos Antoulas

 Springer

Editors

Christopher Beattie
Department of Mathematics
Virginia Tech
Blacksburg, VA, USA

Peter Benner
Max Planck Institute for Dynamics
of Complex Technical Systems
Magdeburg, Germany

Mark Embree
Department of Mathematics
Virginia Tech
Blacksburg, VA, USA

Serkan Gugercin
Department of Mathematics
Virginia Tech
Blacksburg, VA, USA

Sanda Lefteriu
IMT Lille Douai
Douai Cedex, France

Continental Automotive Romania SRL
Timișoara, Romania

ISBN 978-3-030-95156-6 ISBN 978-3-030-95157-3 (eBook)
<https://doi.org/10.1007/978-3-030-95157-3>

Mathematics Subject Classification: 41A05, 37M99, 93A15, 93A30, 93B15, 93B40

© Springer Nature Switzerland AG 2022

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

*Dedicated to Thanos Antoulas on the
occasion of his 70th birthday*

Rational and Systematic—A Mathematical Biography of Thanos Antoulas

Biographical Sketch

ATHANASIOS (THANOS) ANTOULAS was born in Athens, Greece. Having excelled at physics in high school, he considered moving to England for his undergraduate studies. Counseled by his countryman John Argyris, he instead chose to study in Switzerland. After a year at EPFL, he moved to ETH Zürich, where he completed dual degrees: Diplomas in Electrical Engineering (May 1975) and Mathematics (November 1975). After finishing this program of study, he began as a doctoral student in solid-state physics at ETH, but by 1976 changed plans to work instead with Rudolf Kalman, who had joined ETH as chair of Mathematical Systems Theory in 1973. (Thanos offers several reminiscences from his time as a graduate student in [18].) Thanos received his Ph.D. in Mathematics in 1980, with his dissertation [2] focused on translating state-space-based concepts, such as $F \bmod G$ -invariant subspaces and reachability subspaces, into a transfer function-based input/output formulation. He further pursued and extended these developments in his Habilitation, which he completed at ETH just a few years later, in 1982 [3].

Shortly thereafter, Thanos moved to the Department of Electrical and Computer Engineering at Rice University in Houston, Texas. There he joined a robust group of faculty in control theory and digital signal processing, including Sidney Burrus, Tom Parks, and Boyd Pearson, and quickly rose through the faculty ranks to full professor. (Figure 1 shows Thanos in his early years at Rice.) He was elected an IEEE Fellow in 1991, and 4 years later became a Fellow of the Japan Society for the Promotion of Science. From 2002 to 2015, he also maintained an affiliation with Jacobs University Bremen, where he met several students who would later complete Ph.D.'s with him at Rice. In 2016, he was appointed a Fellow of the Max Planck Society in Magdeburg, Germany and began serving as Adjunct Professor in the Department of Molecular and Cellular Biology at Baylor College of Medicine in Houston. For more than 20 years, he was Editor-in-Chief of *Systems & Control Letters*, serving from October 1994 to December 2015.

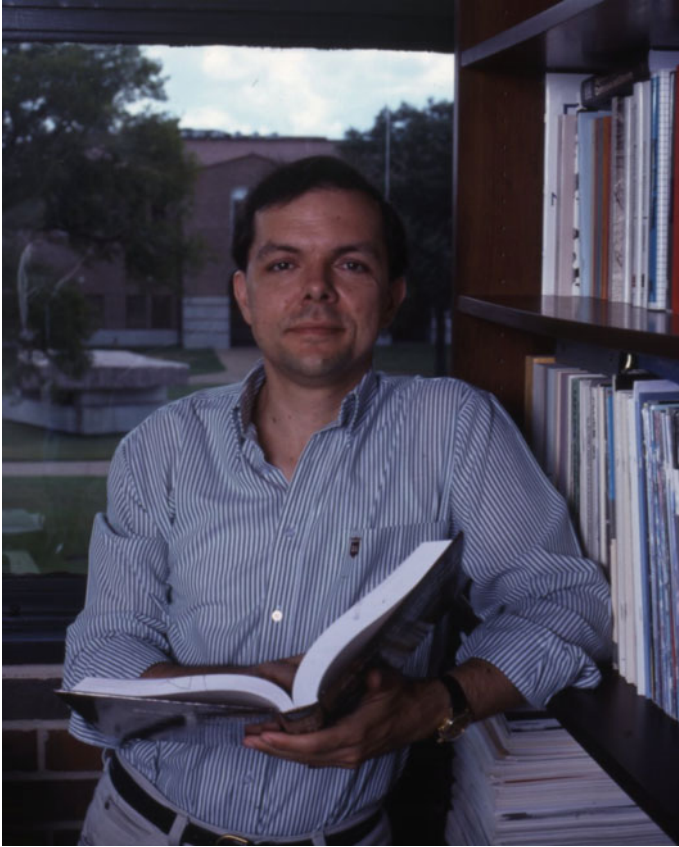


Fig. 1 Thanos in his office in Abercrombie Lab at Rice University, 1980s. (Photograph courtesy of the Woodson Research Center, Fondren Library, Rice University.)

Thanos's research falls under the main umbrellas of approximation theory and systems theory, areas in which he has been an international leader for the past 30 years. Notably, his work includes breakthrough contributions and the development of fundamental methods focused on the rational approximation of large-scale dynamical systems and data-driven modeling. His 2005 SIAM book on *Approximation of Large-Scale Dynamical Systems* is a basic resource for all who seek to learn about and implement model reduction methodology.

It is challenging to give a deserving summary of an academic career as substantial and influential as Thanos's within a brief preface. We will try our best to do so, highlighting only a subset of his fundamental contributions. Articles in this *Festschrift* illustrate clearly how Thanos has influenced many research areas and, to our delight, how he has influenced many of us as well. We hope that he continues to do both long into the future.



Fig. 2 Thanos during a visit to Brian Anderson in Canberra, 1985. Also in the picture are (counter clockwise, starting with Thanos in the red sweater): Brian D.O Anderson, William Andrew Coppel, Ian Petersen, Michael Green, Iven Mareels, and Michel Gevers

The Antoulas–Anderson Collaboration

In 1985, Thanos visited Brian Anderson at the Australian National University (ANU) in Canberra. This research visit had a lasting impact, leading to a breakthrough result that laid the basis for what is now called the Antoulas–Anderson method for rational interpolation of scalar functions [6]. Figure 2 shows young Thanos during this visit, in a photograph taken on his birthday. (Indeed, one can see his cake on the table behind him.)

The paper [6] contains several important results that brought the Loewner matrix into central focus as the right tool to generalize realization theory to rational interpolation. Loewner introduced his eponymous matrix in [46] as an adjunct to the study of monotone matrix functions but also understood the connection this matrix had with rational interpolation and other topics as well. Indeed, the Loewner matrix has been rediscovered over the years; in other contexts is known as the *divided-difference matrix* and the *null-pole coupling matrix*. The article [6] extended and deepened the relationship with rational interpolation by noting a subtle dichotomy in the interpretation of the *rank* of the Loewner matrix: namely, that the minimal degree of a rational interpolant will either be equal to the rank of the Loewner matrix or to the difference between the number of interpolating points and this rank, depending on whether certain ancillary conditions are satisfied. The former case is generic and the minimal interpolating function will be unique, while in the latter case, the minimal

interpolating function will not be unique. This article also developed parameterizations of both minimal and non-minimal rational interpolants starting from the Loewner matrix, and moreover, provided a parametrization of all minimal updates to any given (minimal) rational interpolant.

The sequel paper [1] introduced several other important results: Just as is true for the Hankel matrix, the Loewner matrix can be expressed as the product of a generalized observability matrix and a generalized reachability matrix. For minimal systems, each of these generalized observability and reachability matrices has rank equal to the McMillan degree. A systematic approach for producing state-space realizations of rational functions based on an almost square Loewner matrix was proposed there as well, with related development of a method leading to the construction of rational interpolants from knowledge of the nullspace of a (tall skinny) Loewner matrix. This last point became the basis of what has now become known as the *Adaptive Antoulas–Anderson (AAA) algorithm* [49].¹

Projection-Based Model Reduction

In the mid-1990s, Paul Van Dooren, who has long served as a bridge between the systems theory and numerical linear algebra communities, introduced Thanos to his Rice University colleague Danny Sorensen. An expert in Krylov subspace algorithms, Sorensen had recently developed the Implicitly Restarted Arnoldi Method [54] for large-scale eigenvalue computations. Soon Thanos and Danny were collaborating on projection-based algorithms for model reduction, which have at their core the same Krylov subspace technology used for solving linear systems and eigenvalue problems. This fruitful collaboration yielded numerous important results, including one of the first survey papers on model reduction [15] that covers both systems’ theoretic approaches and Krylov subspace methods. Thanos, Danny, and several of their graduate students also investigated theoretical and algorithmic properties of Lyapunov equations, which play a critical role in control theory and model reduction [14, 16, 37, 56, 57]. Key results from this work include bounds on the decay of Hankel singular values [16] and the approximate structured balanced truncation algorithm [56].

Beyond their joint publications, their symbiotic research influenced other faculty and many graduate students in the electrical engineering and applied math departments. During this period, Thanos was writing his landmark textbook, *Approximation of Large-Scale Dynamical Systems* [4], and test-driving the contents in his ELEC 501

¹ This algorithm had a brief 6-month run as the *Aggressive Antoulas–Anderson algorithm* (see e.g., <https://arxiv.org/abs/1612.00337v1>) but possibly due to the ambiguity of where one may separate compound modifiers in English, it was soon changed to the milder *Adaptive Antoulas–Anderson algorithm*. Of course, to those of us who know Thanos well, “Aggressive Antoulas...” is unseemly and implausible; “(Politely) Assertive Antoulas...” is fitting to the man but perhaps not to the algorithm; and there seems to be general agreement that “Adaptive Antoulas...” is a comfortable compromise.

“Approximation of Systems” graduate course at Rice. Several of us authors fondly remember these lively and dynamic days.

A longstanding and fundamental question in *interpolatory* model reduction (as is the case also for any interpolation-based method) lies in how best to choose the location of interpolation points. Methods for creating interpolatory reduced models were formerly called moment-matching methods or rational Krylov methods; they generally utilized ad hoc choices of interpolation points that were chosen to reduce heuristic error estimates. Systematic approaches for selecting interpolation points were later motivated by expressions for the \mathcal{H}_2 (-error) norm proved in [4, 32, 33]. These expressions decomposed the \mathcal{H}_2 model reduction error into two components: one due to a mismatch of the full and reduced transfer functions at the mirror images of the full-order system poles; the other due to the mismatch at the mirror images of the reduced-order system poles. The first attempts to make use of these expressions in [32, 33] forced interpolation at the mirror images of dominant full-order poles, thus eliminating some terms contributed by the first error component. But just a few years later, it was discovered in [34, 35] that *optimality* with respect to the \mathcal{H}_2 error could be achieved via bitangential Hermite interpolation at the mirror images of the reduced-order poles along tangent directions determined by the reduced-system residues; these are also known as the Meier–Luenberger conditions [48] in the single-input/single-output case. This optimality condition led to the Iterative Rational Krylov Algorithm (IRKA) [34, 35], which iteratively corrects the interpolation locations until the optimality conditions are satisfied. The 2008 paper [34] also showed the equivalence of the Sylvester-equation framework [40] to the \mathcal{H}_2 approximation problem with the interpolatory framework [34, 48]. The IRKA approach, via its connection to numerically efficient rational Krylov methods, has made it possible to compute optimal \mathcal{H}_2 approximations to large-scale dynamical systems and has emerged as a major milestone in model reduction. The success of this approach has motivated the extension of the optimal interpolatory model reduction framework into diverse settings including, e.g., bilinear [21, 27] and quadratic [22] nonlinear dynamical systems; parameterized dynamical systems [19, 39]; structured dynamical systems [25, 36, 58]; data-driven optimal \mathcal{H}_2 approximation [20, 38]; and optimal approximation with respect to related error measures [23, 26]. We have highlighted here Thanos’s contributions to interpolatory model reduction—indeed, his contributions in this area have made the term nearly synonymous with optimal \mathcal{H}_2 approximation (as described above) and the data-driven Loewner framework (as described below). Many of these ideas have formed the foundation of his most recent book, *Interpolatory Methods for Model Reduction*, with the co-authors Beattie and Gugercin [9].

In the early 2000s, Thanos was also developing ideas that brought fundamental advances into the treatment of “passive” systems. When the modeling context that produced the original model contains features that reflect how the system handles energy, it is reasonable to expect that surrogate (reduced) models should retain analogous structural features, in addition to producing a small system error. Indeed, preservation of structural features like this may be necessary for the surrogate model

to behave “physically”—ideally, like a miniature version of the original system. “Passive” systems cannot produce a net gain in energy within the larger systems to which they are coupled, so certainly any prospective reduced-order surrogate system should share that feature and hence also not produce a net gain in energy. For example, when VLSI circuit designs are modeled as RLC circuits, passive reduced models can be synthesized as equivalent RLC circuits, and so can be manifested quite literally as a miniature version of the original RLC model.

Thanos’ first major contribution in this area was an interpolatory reduction method that preserved passivity of the reduced-order model [5]. He showed that if interpolation points were drawn from the set of so-called spectral zeros of the original system, passivity (and hence stability) of the reduced system would be guaranteed. It is interesting to note that this observation converts the model reduction problem to a generalized eigenvalue problem involving the Hamiltonian of the associated full-order system. The key results of [5] were followed by a flurry of subsequent contributions by collaborators and students including Danny Sorensen [55], Roxana Ionutiu, and Joost Rommes [43].

The Loewner Framework for Data-Driven Modeling

Loewner matrices played a fundamental role in Thanos’s earlier work on rational interpolation and realization theory [1, 6–8]. However, it was the breakthrough paper [47], written in collaboration with his Ph.D. student Andrew Mayo, that led to the epoch-making Loewner framework for data-driven modeling. This paper introduced the *shifted Loewner matrix*, enabling the construction of data-driven rational interpolants immediately in descriptor form, thus avoiding the numerically ill-conditioned “inversion” of the \mathbf{E} -matrix during construction (which in earlier works would implicitly enforce a regular state-space realization upon the interpolant). This paper also showed how one could construct *tangential* interpolants directly from data, allowing one to have significantly more flexibility in the construction of rational interpolants. As a final touch and one that is becoming ever more relevant in this age of “big data”, Thanos also demonstrated in [47] how one may remove data redundancy using the *singular value decomposition* (SVD) and thus how *minimal* rational tangential interpolants may be formed when abundant data is available.

While [47] laid out the theoretical foundations of a data-driven approach for frequency-domain modeling of systems with many inputs and outputs given a large amount of data, the principal computational cost that was involved could become formidable, since it required the SVD of a linear combination of Loewner and shifted Loewner matrices of order equal to the number of measurements. To reduce this cost for data sets comprising a large number of samples, Thanos and his Ph.D. student Sanda Lefteriu proposed several approaches for adding measurements in batches, while simultaneously monitoring the error [45]. Notably, one of these approaches is based on an extension of the parameterization of rational interpolants proposed in [6], as described in the previous section.

The 2007 paper [47] has become the foundation for numerous advances in data-driven modeling. For example, [20] employed the Loewner framework in the \mathcal{H}_2 -optimal approximation setting, replacing the projection-based construction of \mathcal{H}_2 -optimal reduced models, such as those in [34], with a data-driven formulation that does not need access to the underlying realization. In [12, 42], Thanos and his Ph.D. students Lefteriu and Ionita developed the Loewner framework for parametric dynamical systems by extending the single-variable Loewner matrices (and single-variable barycentric forms) to multivariable ones. Thus, parametric interpolatory reduced models were directly constructed from parametric transfer function samples, another major milestone in data-driven modeling of dynamical systems.

Another substantial development was the extension of the Loewner framework to special classes of nonlinear dynamical systems by Thanos and his students and collaborators: to bilinear dynamical systems [11] with his Ph.D. students Cosmin Ionita and Ion-Victor Gosea; to quadratic-bilinear dynamical systems [10, 28, 29] with his Ph.D. students Ion-Victor Gosea and Dimitrios S. Karachalios, and his Rice colleague Matthias Heinkenschloss; and to switched dynamical systems [30] with his Ph.D. student Ion-Victor Gosea and his colleague Mihaly Petreczky. The theory from the linear case was elegantly extended to these important classes of nonlinear dynamics by forming divided difference matrices from samples of the multivariate subsystem transfer functions resulting from the Volterra series expansion of nonlinear dynamical systems.

Numerous other works have been motivated by the Loewner framework or further extended it to different settings. For example, [31, 51–53] developed a Loewner-like approach for constructing data-driven structured realizations such as for delay systems or second-order systems. For the case of time-domain data (as opposed to frequency), [41, 44, 50] investigated the Loewner-based data-driven modeling using the time-domain input/output data. In the recent work [24], the Loewner framework was used in contour integral methods for solving nonlinear eigenvalue problems. These are only a small subset of works that are inspired by the influential paper [47].

While the Loewner framework has found applications in diverse settings (see [9, 13] for a variety of examples), we highlight Thanos's collaboration with colleagues at Baylor College of Medicine [17, 59], in which the Loewner-based modeling was utilized to reveal 12-hour cycles appearing in living organisms that are quite distinct from the ubiquitous 24-hour (circadian) cycles. The results of [59] were highlighted in the "Editors' Choice" column of the 30 June 2017 issue of *Science* and constitute one of the most significant scientific applications of the Loewner framework to date. We find that the evolution of this work typifies Thanos's refined taste in interesting

problems and demonstrates his strength in developing wide collaborations engaging diverse areas of application.

Blacksburg, USA
 Magdeburg, Germany
 Blacksburg, USA
 Blacksburg, USA
 Douai, France

Christopher Beattie
 Peter Benner
 Mark Embree
 Serkan Gugercin
 Sanda Lefteriu

Acknowledgements On behalf of all the authors included in this *Festschrift*, we thank Thanos Antoulas for many years of friendship, mentorship, inspiration, and camaraderie. We are also grateful to Thanos for sharing details of his early years in the field, which have informed our biography.

We thank Melissa Kean, University Historian of Rice University, for locating and sharing the photograph in Fig. 1.

References

1. Anderson, B. D. O., Antoulas, A. C.: Rational interpolation and state variable realizations. *Linear Algebra Appl.* **137/138**, 479–509 (1990)
2. Antoulas, A. C.: *A Polynomial Matrix Approach to $F \bmod G$ -Invariant Subspaces*. PhD thesis, Department of Mathematics, ETH Zürich, (1979)
3. Antoulas, A. C.: *New Results on the Algebraic Theory of Linear Systems*. Habilitation thesis, Department of Mathematics, ETH Zürich, (1982)
4. Antoulas, A. C.: *Approximation of Large-Scale Dynamical Systems*, volume 6 of *Advances in Design and Control*. SIAM Philadelphia, (2005)
5. Antoulas, A. C.: A new result on passivity preserving model reduction. *Syst. Control Lett.* **54**, 361–374 (2005)
6. Antoulas, A. C., Anderson, B. D. O.: On the scalar rational interpolation problem. *IMA J. Math. Control Inf.* **3**, 61–88 (1986)
7. Antoulas, A. C., Anderson, B. D. O.: On the problem of stable rational interpolation. *Linear Algebra Appl.* **122–124**, 301–329 (1989)
8. Antoulas, A. C., Ball, J. A., Kang, J., Willems, J. C.: On the solution of the minimal rational interpolation problem. *Linear Algebra Appl.* **137**, 511–573 (1990)
9. Antoulas, A. C., Beattie, C. A., Güğercin, S.: *Interpolatory Methods for Model Reduction*. Computational Science and Engineering 21. SIAM, Philadelphia, (2020)
10. Antoulas, A. C., Gosea, I. V., Heinkenschloss, M.: On the Loewner framework for model reduction of Burgers equation. In *Active Flow and Combustion Control 2018*, pages 255–270. Springer, (2019)
11. Antoulas, A. C., Gosea, I. V., Ionita, A. C.: Model reduction of bilinear systems in the Loewner framework. *SIAM J. Sci. Comput.* **38**(5), B889–B916 (2016)
12. Antoulas, A. C., Ionita, A. C., Lefteriu, S.: On two-variable rational interpolation. *Linear Algebra Appl.* **436**, 2889–2915 (2012)
13. Antoulas, A. C., Lefteriu, S., Ionita, A. C.: A tutorial introduction to the Loewner framework for model reduction. In *Model Reduction and Approximation*, pp. 335–376. SIAM, Philadelphia, (2017)
14. Antoulas, A. C., Sorensen, D. C.: Lyapunov, Lanczos, and inertia. *Linear Algebra Appl.* **326**:137–150 (2001)
15. Antoulas, A. C., Sorensen, D. C., Gugercin, S.: A survey of model reduction methods for large-scale systems. *Contemp. Math.* **280**, 193–219 (2001)

16. Antoulas, A. C., Sorensen, D. C., Zhou, Y.: On the decay rate of Hankel singular values and related issues. *Syst. Control Lett.* **46**(5), 323–342 (2002)
17. Antoulas, A. C., Zhu, B., Zhang, Q., York, B., O'Malley, B. W., Dacso, C. C.: A novel mathematical method for disclosing oscillations in gene transcription: a comparative study. *PLoS One* **13**(9), e0198503 (2018)
18. Antoulas, T.: Recollections of my time as a doctoral student of R. E. Kalman. *IEEE Control Syst. Magazine* **April**, 96–97 (2010)
19. Baur, U., Benner, P., Beattie, C. A., Gugercin, S.: Interpolatory projection methods for parameterized model reduction. *SIAM J. Sci. Comput.* **33**, 2489–2518 (2011)
20. Beattie, C. A., Gugercin, S.: Realization-independent \mathcal{H}_2 -approximation. In *Proceedings of 51st IEEE Conference on Decision and Control*, pp. 4953–4958 (2012)
21. Benner, P., Breiten, T.: Interpolation-based \mathcal{H}_2 -model reduction of bilinear control systems. *SIAM J. Matrix Anal. Appl.* **33**(3), 859–885 (2012)
22. Benner, P., Goyal, P., Gugercin, S.: \mathcal{H}_2 -quasi-optimal model order reduction for quadratic-bilinear control systems. *SIAM J. Matrix Anal. Appl.* **39**(2), 983–1032 (2018)
23. Breiten, T., Beattie, C. A., Gugercin, S.: Near-optimal frequency-weighted interpolatory model reduction. *Syst. Control Lett.* **78**, 8–18 (2015)
24. Brennan, M. C., Embree, M., Gugercin, S.: Contour integral methods for nonlinear eigenvalue problems: a systems theoretic approach. *SIAM Review*, to appear, available as *arXiv preprint arXiv:2012.14979*, (2020)
25. Duff, I. P., Poussot-Vassal, C., Seren, C.: \mathcal{H}_2 -optimal model approximation by input/output-delay structured reduced order models. *Syst. Control Lett.* **117**, 60–67 (2018)
26. Flagg, G., Beattie, C., Gugercin, S.: Interpolatory \mathcal{H}_∞ model reduction. *Syst. Control Lett.* **62**(7), 567–574 (2013)
27. Flagg, G., Gugercin, S.: Multipoint Volterra series interpolation and \mathcal{H}_2 optimal model reduction of bilinear systems. *SIAM J. Matrix Anal. Appl.* **36**, 549–579 (2015)
28. Gosea, I. V., Antoulas, A. C.: Data-driven model order reduction of quadratic-bilinear systems. *Numer. Linear Algebra Appl.* **25**(6), e2200 (2018)
29. Gosea, I. V., Karachalios, D. S., Antoulas, A. C.: Learning reduced-order models of quadratic control systems from input-output data. In *2021 European Control Conference (ECC)*, 1426–1431 (2021)
30. Gosea, I. V., Petreczky, M., Antoulas, A. C.: Data-driven model order reduction of linear switched systems in the Loewner framework. *SIAM J. Sci. Comput.* **40**(2), B572–B610 (2018)
31. Gosea, I. V., Pontes Duff, I., Benner, P., Antoulas, A. C.: Model order reduction of bilinear time-delay systems. In *2019 18th European Control Conference (ECC)*, pp. 2289–2294. IEEE (2019)
32. Gugercin, S.: *Projection Methods for Model Reduction of Large-scale Dynamical Systems*. PhD thesis, Rice University (2003)
33. Gugercin, S., Antoulas, A. C.: An \mathcal{H}_2 error expression for the Lanczos procedure. In *42nd IEEE International Conference on Decision and Control*, volume 2, pp. 1869–1872. IEEE (2003)
34. Gugercin, S., Antoulas, A. C., Beattie, C.: H_2 model reduction for large-scale linear dynamical systems. *SIAM J. Matrix Anal. Appl.* **30**, 609–638 (2008)
35. Gugercin, S., Antoulas, A. C., Beattie, C. A.: An iterative rational Krylov algorithm for optimal \mathcal{H}_2 model reduction. *The 17th International Symposium on Mathematical Theory of Networks and Systems* (July 2006)
36. Gugercin, S., Polyuga, R. V., Beattie, C., van der Schaft, A.: Structure-preserving tangential interpolation for model reduction of port-Hamiltonian systems. *Automatica* **48**(9), 1963–1974 (2012)
37. Gugercin, S., Sorensen, D. C., Antoulas, A. C.: A modified low-rank Smith method for large-scale Lyapunov equations. *Numerical Algorithms* **32**(1), 27–55 (2003)
38. Hokanson, J. M., Magruder, C. C.: \mathcal{H}_2 -optimal model reduction using projected nonlinear least squares. *SIAM J. Sci. Comput.* **42**(6), A4017–A4045 (2020)

39. Hund, M., Mitchell, T., Mlinaric, P., Saak, J.: Optimization-based parametric model order reduction via $\mathcal{H}_2 \otimes \mathcal{L}_2$ first-order necessary conditions. *SIAM Journal on Scientific Computing*, to appear, available as *arXiv preprint* [arXiv:2103.03136](https://arxiv.org/abs/2103.03136) (2021)
40. Hyland, D., Bernstein, D.: The optimal projection equations for model reduction and the relationships among the methods of Wilson, Skelton, and Moore. *IEEE Trans. Auto. Control* **30**(12), 1201–1211 (1985)
41. Ionita, A. C., Antoulas, A. C.: Matrix pencils in time and frequency domain system identification. *Control, Robotics and Sensors. Institution of Engineering and Technology*, pp. 79–88 (2012)
42. Ionita, A. C., Antoulas, A. C.: Data-driven parametrized model reduction in the Loewner framework. *SIAM J. Sci. Comput.* **36**, A984–A1007 (2014)
43. Ionutiu, R., Rommes, J., Antoulas, A. C.: Passivity preserving model reduction using dominant spectral zero interpolation. *IEEE Trans. CAD (Computer-Aided Design of Integrated Circuits and Systems)* **27**, 2250–2263 (2008)
44. Karachalios, D. S., Gosea, I. V., Antoulas, A. C.: On bilinear time-domain identification and reduction in the Loewner framework. In *Model Reduction of Complex Dynamical Systems*, pp. 3–30. Springer (2021)
45. Lefteriu, S., Antoulas, A. C.: A new approach to modeling multiport systems from frequency-domain data. *IEEE Trans. Computer-Aided Design* **29**, 14–27 (2010)
46. Karl L.: Über monotone Matrixfunktionen. *Mathematische Zeitschrift*, **38**(1), 177–216 (1934)
47. Mayo, A. J., Antoulas, A. C.: A framework for the solution of the generalized realization problem. *Linear Algebra Appl.* **425**, 634–662 (2007)
48. Meier III, L., Luenberger, D.: Approximation of linear constant systems. *IEEE Trans. Auto. Control* **12**(5), 585–588 (1967)
49. Nakatsukasa, Y., Sète, O., Trefethen, L. N.: The AAA algorithm for rational approximation. *SIAM J. Sci. Comput.* **40**(3), A1494–A1522 (2018)
50. Peherstorfer, B., Gugercin, S., Willcox, K.: Data-driven reduced model construction with time-domain Loewner models. *SIAM J. Sci. Comput.* **39**(5), A2152–A2178 (2017)
51. Pontes Duff, I., Poussot-Vassal, C., Seren, C.: Realization independent single time-delay dynamical model interpolation and \mathcal{H}_2 -optimal approximation. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pp. 4662–4667. IEEE (2015)
52. Schulze, P., Unger, B.: Data-driven interpolation of dynamical systems with delay. *Syst. Control Lett.* **97**, 125–131 (2016)
53. Schulze, P., Unger, B., Beattie, C., Gugercin, S.: Data-driven structured realization. *Linear Algebra Appl.* **537**, 250–286 (2018)
54. Sorensen, D. C.: Implicit application of polynomial filters in a k -step Arnoldi method. *SIAM J. Matrix Anal. Appl.* **13**, 357–385 (1992)
55. Sorensen, D. C.: Passivity preserving model reduction via interpolation of spectral zeros. *Syst. Control Lett.* **54**(4), 347–360 (2005)
56. Sorensen, D. C., Antoulas, A. C.: On model reduction of structured systems. In P. Benner, V. L. Mehrmann, and Sorensen, D. C., editors, *Dimension Reduction of Large-scale Systems*, pp. 117–130. Springer (2005)
57. Sorensen, D. C., Antoulas, A. C.: The Sylvester equation and approximate balanced reduction. *Linear Algebra Appl.* **351**, 671–700 (2002)
58. Wyatt, S.: *Issues in Interpolatory Model Reduction: Inexact Solves, Second-order Systems and DAEs*. PhD thesis, Virginia Polytechnic Institute and State University (2012)
59. Zhu, B., Zhang, Q., Pan, Y., Mace, E. M., York, B., Antoulas, A. C., Dacso, C. C., Malley, B. W. O.: A cell-autonomous mammalian 12 hr clock coordinates metabolic and stress rhythms. *Cell Metabolism* **25**(6), 1305–1319 (2017)

Contents

Linear Dynamical Systems

| | |
|--|----|
| The Rational Interpolation Problem: Grassmannian and Loewner-Matrix Approaches | 3 |
| Joseph A. Ball and Vladimir Bolotnikov | |
| The Conditioning of a Linear Barycentric Rational Interpolant | 23 |
| Jean-Paul Berrut | |
| Learning Low-Dimensional Dynamical-System Models from Noisy Frequency-Response Data with Loewner Rational Interpolation | 39 |
| Zlatko Drmač and Benjamin Peherstorfer | |
| Pseudospectra of Loewner Matrix Pencils | 59 |
| Mark Embree and A. Cosmin Ioniță | |
| A Loewner Matrix Approach to the Identification of Linear Time-Varying Systems | 79 |
| Paolo Rapisarda | |
| Linear System Matrices of Rational Transfer Functions | 95 |
| Froilán Dopico, María del Carmen Quintana, and Paul Van Dooren | |

Nonlinear Dynamical Systems

| | |
|--|-----|
| Interpolation-Based Model Order Reduction for Quadratic-Bilinear Systems and \mathcal{H}_2 Optimal Approximation | 117 |
| Xingang Cao, Joseph Maubach, Wil Schilders, and Siep Weiland | |
| An Adaptive Sampling Approach for the Reduced Basis Method | 137 |
| Sridhar Chellappa, Lihong Feng, and Peter Benner | |
| Balanced Truncation Model Reduction for Lifted Nonlinear Systems | 157 |
| Boris Kramer and Karen Willcox | |

Modeling the Buck Converter from Measurements of Its Harmonic Transfer Function 175
Sanda Lefteriu

Model Reduction and Realization Theory of Linear Switched Systems 197
Mihály Petreczky and Ion Victor Gosea

Structured Dynamical Systems

Developments in the Computation of Reduced Order Models with the Use of Dominant Spectral Zeros 215
Francisco Damasceno Freitas, Joost Rommes, and Nelson Martins

Structure-Preserving Interpolatory Model Reduction for Port-Hamiltonian Differential-Algebraic Systems 235
Christopher Beattie, Serkan Gugercin, and Volker Mehrmann

Data-Driven Identification of Rayleigh-Damped Second-Order Systems 255
Igor Pontes Duff, Pawan Goyal, and Peter Benner

Balanced Truncation Model Reduction for 3D Linear Magneto-Quasistatic Field Problems 273
Johanna Kerler-Back and Tatjana Stykel

Structure-Preserving Model Reduction of Physical Network Systems 299
Arjan van der Schaft

Model Reduction for Control

\mathcal{H}_2 -gap Model Reduction for Stabilizable and Detectable Systems 317
Tobias Breiten, Christopher Beattie, and Serkan Gugercin

Reduced Order Model Hessian Approximations in Newton Methods for Optimal Control 335
Matthias Heinkenschloss and Caleb Magruder

Interpolation-Based Irrational Model Control Design and Stability Analysis 353
Charles Poussot-Vassal, Pauline Kergus, and Pierre Vuillemin

Applications

Oscillations in Biology 375
Jitendra K. Meena and Clifford C. Dacso

Model-Order Reduction for Coupled Flow and Linear Thermal-Poroplasticity with Applications to Unconventional Reservoirs 387
Horacio Florez, Eduardo Gildin, and Patrick Morkos

Challenges in Model Reduction for Real-Time Simulation of Traction Chain Systems 409
Roxana Ionuțiu

Sparse Representation for Sampled-Data H^∞ Filters 427
Masaaki Nagahara and Yutaka Yamamoto

Analysis of a Reduced Model of Epithelial–Mesenchymal Fate Determination in Cancer Metastasis as a Singularly-Perturbed Monotone System 445
M. Ali Al-Radhawi and Eduardo D. Sontag

Linear Dynamical Systems

The Rational Interpolation Problem: Grassmannian and Loewner-Matrix Approaches



Joseph A. Ball and Vladimir Bolotnikov

*Dedicated to Thanos Antoulas on the occasion of his 70th
birthday with respect and admiration*

Abstract We review the Grassmannian approach to rational interpolation theory and indicate points of contact and divergence with the Loewner-matrix approach. We also provide an introduction to some recent work on free noncommutative function theory, in particular noncommutative rational functions, where some results paralleling the classical single-variable case have been obtained but where many other basic issues remain to be resolved.

Keywords Rational interpolation · Loewner framework · Grassmannian · Noncommutative rational functions

1 Introduction

Starting in the early 1980s there was a lot of interest in the engineering/systems theory community in Nevanlinna-Pick interpolation due to its connection with the Model Matching problem and with the evolving theory of H^∞ -control. Together with Bill Helton the first author was part of the development of the Grassmannian approach to Nevanlinna-Pick interpolation, the natural setting for which was Kreĭn-space operator theory. Later, work of the first author with Israel Gohberg and Leiba

J. A. Ball
Department of Mathematics, Virginia Tech, Blacksburg, VA 24061-0123, USA
e-mail: joball@math.vt.edu

V. Bolotnikov (✉)
Department of Mathematics, William & Mary, Williamsburg, VA 23187-8795, USA
e-mail: vladi@math.wm.edu

Rodman focused on the study of rational solutions and transfer-function realizations for the solutions, with abstract Kreĭn-space arguments replaced by more concrete winding-number arguments. It was much later that we attempted to apply the same ideas to interpolation problems without metric constraints. It was during this period that the first author realized that he and his student (Jeongook Kang) were trying to solve the same problem that Thanos Antoulas was working on, resulting in the joint paper [6]. Thanos's approach has come to be called the *Loewner matrix approach* on which there are now several good overview expositions (see [5, Chap. 4.5] and [8]). It is interesting that the Loewner-matrix approach was applied first to problems without metric constraints and then practitioners saw how to adapt the approach to Nevanlinna-Pick-type interpolation problems (with metric constraints), whereas for the Grassmannian approach the sequence was in reverse. The second author has complementary expertise in yet another school of interpolation theory, namely that of Potapov using reproducing kernel Hilbert space methods and operator equations (see the two articles [42, 43] in a 1997 Potapov-dedicatory volume). Here we review, compare, and contrast these various approaches, and also provide an introduction to the whole new world of freely noncommutative rational matrix functions, a topic of current robust activity.

2 Review of the Grassmannian Approach to Rational Interpolation

Let us consider the following simplest case of the bitangential rational matrix interpolation problem. We are given a data set

$$\mathfrak{D} = \{\boldsymbol{\mu}, \boldsymbol{\ell}, \mathbf{v}; \boldsymbol{\lambda}, \mathbf{r}, \mathbf{w}\} \quad (1)$$

consisting of

- left interpolation nodes $\boldsymbol{\mu} := \{\mu_i\}_{i=1}^q \subset \mathbb{C}$
- left tangential directions $\boldsymbol{\ell} := \{\ell_i\}_{i=1}^q \subset \mathbb{C}^p$
- left tangential values $\mathbf{v} := \{\mathbf{v}_i\}_{i=1}^q \subset \mathbb{C}^m$
- right interpolation nodes $\boldsymbol{\lambda} := \{\lambda_j\}_{j=1}^k \subset \mathbb{C}$
- right tangential directions $\mathbf{r} := \{\mathbf{r}_j\}_{j=1}^k \subset \mathbb{C}^m$
- right tangential values $\mathbf{w} := \{\mathbf{w}_j\}_{j=1}^k \subset \mathbb{C}^p$.

For simplicity and to avoid degeneracies, we assume that the combined set of interpolation nodes $\mu_1, \dots, \mu_q, \lambda_1, \dots, \lambda_k$ consists of distinct points and that the left and right tangential directions $\ell_1, \dots, \ell_q, \mathbf{r}_1, \dots, \mathbf{r}_k$ are all nonzero vectors. Given such a data set, we seek to understand the set of $p \times m$ rational matrix functions \mathbf{H} which are pole-free on the set of interpolation nodes $\boldsymbol{\mu} \cup \boldsymbol{\lambda}$ and satisfy the set of tangential interpolation conditions:

$$\begin{aligned}\ell_i^\top \mathbf{H}(\mu_i) &= \mathbf{v}_i^\top \text{ for } i = 1, \dots, q, \\ \mathbf{H}(\lambda_j) \mathbf{r}_j &= \mathbf{w}_j \text{ for } j = 1, \dots, k.\end{aligned}\quad (2)$$

The idea of the Grassmannian approach to problems of this sort (referred as the *generating system approach* in [5, 8]), as developed starting in the mid-1980s by Ball-Helton [20] and then worked out in state-space coordinates for rational solutions in [14], is to reformulate the problem in terms of the graph space of a solution (viewed as a multiplication operator on an appropriate subspace of rational vector functions) rather than in terms of a solution itself. To get into details, a little more notation will be helpful. For ω any subset of \mathbb{C} and n a positive integer, we use the notation $\mathcal{R}^n(\omega)$ for rational \mathbb{C}^n -valued functions having no poles in ω . We simplify the notation for two special cases: if $\omega = \emptyset$, we write simply \mathcal{R}^n rather than $\mathcal{R}^n(\emptyset)$ for the space of all \mathbb{C}^n -valued rational functions, and if $n = 1$, we write simply $\mathcal{R}(\omega)$ rather than $\mathcal{R}^1(\omega)$ for the space of scalar-valued rational functions with no poles in ω .

Given the data set \mathcal{D} as in (1), let us introduce a quintuple of matrices

$$\begin{aligned}\mathbb{E}_{\mathcal{D}} &= (\mathbf{C}, \mathbf{A}, \mathbf{Z}, \mathbf{B}, \mathbb{L}) \text{ where} \\ \mathbf{C} &= \begin{bmatrix} \mathbf{w}_1 & \cdots & \mathbf{w}_k \\ \mathbf{r}_1 & \cdots & \mathbf{r}_k \end{bmatrix} \in \mathbb{C}^{(p+m) \times k}, \quad \mathbf{A} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_k \end{bmatrix} \in \mathbb{C}^{k \times k}, \\ \mathbf{Z} &= \begin{bmatrix} \mu_1 & & \\ & \ddots & \\ & & \mu_q \end{bmatrix} \in \mathbb{C}^{q \times q}, \quad \mathbf{B} = \begin{bmatrix} \ell_1^\top & -\mathbf{v}_1^\top \\ \vdots & \vdots \\ \ell_q^\top & -\mathbf{v}_q^\top \end{bmatrix} \in \mathbb{C}^{q \times (p+m)}, \\ \mathbb{L} &= \begin{bmatrix} \mathbf{v}_i^\top \mathbf{r}_j - \ell_i^\top \mathbf{w}_j \\ \mu_i - \lambda_j \end{bmatrix}_{1 \leq i \leq q, 1 \leq j \leq k} \in \mathbb{C}^{q \times k}.\end{aligned}\quad (3)$$

The assumptions that each \mathbf{r}_j is nonzero and that the λ_j 's are distinct imply that (\mathbf{C}, \mathbf{A}) is an observable pair, i.e., $\bigcap_{n=0}^{k-1} \text{Ker } \mathbf{C} \mathbf{A}^n = \{0\}$. Similarly the assumptions that each ℓ_i is nonzero and that the μ_i 's are distinct imply that (\mathbf{Z}, \mathbf{B}) is a controllable pair, i.e., $\text{span}_{j=0, \dots, k-1} \text{Ran } \mathbf{Z}^j \mathbf{B} = \mathbb{C}^k$. Furthermore it is easily checked that \mathbb{L} solves the Sylvester equation

$$\mathbb{L} \mathbf{A} - \mathbf{Z} \mathbb{L} = \mathbf{B} \mathbf{C}, \quad (4)$$

and in fact, by our assumption that the eigenvalues of \mathbf{A} ($\lambda_1, \dots, \lambda_k$) are disjoint from the eigenvalues of \mathbf{Z} (μ_1, \dots, μ_q), is uniquely determined from $\mathbf{A}, \mathbf{Z}, \mathbf{B}, \mathbf{C}$ as the unique solution of (4). The reader familiar with the work of Antoulas and collaborators will recognize the matrix \mathbb{L} appearing in (3) as exactly the *Loewner matrix* associated with the bitangential interpolation problem in the *Loewner-matrix approach* to interpolation developed in the many papers of Antoulas and collaborators (see in particular [5, Chap. 4] and [8] for good overview treatments).

In general, given any quintuple of matrices

$$\Xi = (\mathbf{C}, \mathbf{A}, \mathbf{Z}, \mathbf{B}, \mathbb{L})$$

of respective sizes $N \times k, k \times k, q \times q, q \times N, q \times k$, we say that Ξ is an *admissible Sylvester data set* over ω if the following conditions hold:

- (i) (\mathbf{C}, \mathbf{A}) is an observable pair with $\sigma(\mathbf{A}) \subset \omega$,
- (ii) (\mathbf{Z}, \mathbf{B}) is a controllable pair with $\sigma(\mathbf{Z}) \subset \omega$, and
- (iii) \mathbb{L} is a solution of the Sylvester equation

$$\mathbb{L}\mathbf{A} - \mathbf{Z}\mathbb{L} = \mathbf{B}\mathbf{C}. \quad (5)$$

Thus we see in particular that the collection $\Xi_{\mathfrak{D}}$ formed from the bitangential interpolation data set \mathfrak{D} as in (3) is an admissible Sylvester data set over ω (with the input-output space dimension N set equal to $N = p + m$). Given any admissible Sylvester data set Ξ over the subset $\omega \subset \mathbb{C}$, we may define a subspace $\mathcal{S}_{\Xi} \subset \mathcal{R}^N$ by

$$\mathcal{S}_{\Xi} = \left\{ \mathbf{C}(sI - \mathbf{A})^{-1}\mathbf{x} + \mathbf{h}(s) : \mathbf{x} \in \mathbb{C}^k, \mathbf{h} \in \mathcal{R}^N(\omega) \text{ such that} \right. \\ \left. \sum_{z \in \omega} \text{Res}_{s=z}(sI - \mathbf{Z})^{-1}\mathbf{B}\mathbf{h}(s) = \mathbb{L}\mathbf{x} \right\}. \quad (6)$$

Note that the observable pair (\mathbf{C}, \mathbf{A}) determines the pole structure in ω of $\mathbf{f} \in \mathcal{R}^N$ which are in \mathcal{S}_{Ξ} while the controllable pair (\mathbf{Z}, \mathbf{B}) determines the null structure in ω of such functions, and the matrix \mathbb{L} , called the *null-pole coupling matrix* in the work of [14], quantifies how the pole structure and null structure of functions \mathbf{f} must be coupled for admission to the subspace \mathcal{S}_{Ξ} . The significance of the validity of the Sylvester equation (5) is that this is exactly the condition on the data set Ξ required to guarantee that the associated subspace \mathcal{S}_{Ξ} is a module over the ring $\mathcal{R}(\omega)$ of scalar rational functions analytic on ω , i.e., so that

$$r \in \mathcal{R}(\omega), \mathbf{f} \in \mathcal{S}_{\Xi} \Rightarrow r\mathbf{f} \in \mathcal{S}_{\Xi}.$$

$\mathcal{R}(\omega)$ -submodules \mathcal{S} of \mathcal{R}^N of the form $\mathcal{S} = \mathcal{S}_{\Xi}$ have two additional special properties:

- (i) *full-range*: $\dim \mathcal{R}^N(\omega) / [\mathcal{S} \cap \mathcal{R}^N(\omega)] < \infty$, and
- (ii) *simple-invariance*: $\dim[\mathcal{S} + \mathcal{R}^N(\omega)] / \mathcal{R}^N(\omega) < \infty$.

The content of the next result is that these are exactly the free $\mathcal{R}(\omega)$ -submodules of \mathcal{R}^N of rank N (see [14, Theorem 14.1.3]).

Theorem 1 *Let \mathcal{S} be a $\mathcal{R}(\omega)$ submodule of \mathcal{R}^N . Then the following statements are equivalent.*

1. \mathcal{S} is full range and simply invariant.
2. \mathcal{S} has the form $\mathcal{S} = \mathcal{S}_{\Xi}$ for an admissible Sylvester data set Ξ over ω .

3. *There exists a rational matrix function $\Theta \in \mathcal{R}^{N \times N}$ with $\det \Theta$ not identically equal to zero so that $\mathcal{S} = \Theta \cdot \mathcal{R}^N(\omega)$.*

We note that the book [14] uses the terminology *null-pole subspace* for a subspace of the form $\Theta \cdot \mathcal{R}^N(\omega)$ where Θ is some $N \times N$ rational matrix function with determinant not vanishing identically; in module-theoretic language, such a subspace amounts to the $\mathcal{R}(\omega)$ -submodule of \mathcal{R}^N generated by the columns of Θ .

Remark 1 We note that in principle a (possibly descriptor) realization formula for the matrix function Θ in part (3) of Theorem 1 can be computed directly from the admissible null-pole triple Ξ ; in case \mathbb{L} is square and invertible, there is a simple formula for such a Θ , namely:

$$\Theta(s) = I + \mathbf{C}(sI - \mathbf{A})^{-1}\mathbb{L}^{-1}\mathbf{B}. \quad (7)$$

In case \mathbb{L} is not invertible (e.g., if \mathbb{L} is not even square), the idea is to add null-pole data over a subset ω' disjoint from ω (e.g. take ω' to be the point at ∞) so that the combined admissible Sylvester data set does have coupling operator $\mathbb{L}_e = \begin{bmatrix} \mathbb{L} & \mathbb{L}_{12} \\ \mathbb{L}_{21} & \mathbb{L}_0 \end{bmatrix}$ which is invertible. For details on this, we refer to Sects. 4.5 and 4.6 of [14] and the references there.

A finer result is to add pole/null data only at the point at infinity so that the resulting global null/pole triple has invertible coupling matrix and hence there exists a rational matrix function corresponding with global null/pole structure given by this expanded data set (see [34]). As an additional refinement one can demand that the solution be column-reduced at infinity (see [19]).

Let us now return to the more specific case where the admissible Sylvester data set Ξ has the form $\Xi = \Xi_{\mathcal{D}}$ coming from the data set \mathcal{D} (1) for a bitangential interpolation problem (2). Then the formula (6) for the null-pole subspace applied to the case where $\Xi = \Xi_{\mathcal{D}}$ assumes the more detailed form

$$\begin{aligned} \mathcal{S}_{\Xi_{\mathcal{D}}} &= \left\{ \sum_{j=1}^k \begin{bmatrix} \mathbf{w}_j \\ \mathbf{r}_j \end{bmatrix} (s - \lambda_j)^{-1} c_j + \begin{bmatrix} \mathbf{h}_+(s) \\ \mathbf{h}_-(s) \end{bmatrix} : c_j \in \mathbb{C} \text{ for } 1 \leq j \leq k, \right. \\ &\quad \mathbf{h}_+ \in \mathcal{R}^p(\omega), \mathbf{h}_- \in \mathcal{R}^m(\omega) \text{ such that, for each } 1 \leq i \leq q, \\ &\quad \left. \boldsymbol{\ell}_i^\top \mathbf{h}_+(\mu_i) - \mathbf{v}_i^\top \mathbf{h}_-(\mu_i) = \sum_{j=1}^k \frac{\mathbf{v}_i^\top \mathbf{r}_j - \boldsymbol{\ell}_i^\top \mathbf{w}_j}{\mu_i - \lambda_j} c_j \right\}. \quad (8) \end{aligned}$$

It is natural to define a second admissible Sylvester data set over ω , namely,

$$\mathcal{D}_- = (C_-, A, 0, 0, 0) \text{ where } C_- = [r_1 \cdots r_k], \quad A = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_k \end{bmatrix}$$

with associated null-pole subspace

$$\mathcal{S}_{\Xi_{\mathfrak{D}_-}} = \{C_-(sI - A)^{-1}x : x \in \mathbb{C}^k\} + \mathcal{R}^m(\omega).$$

Given a rational matrix function $\mathbf{H} \in \mathcal{R}^{p \times m}$, we let $M_{\mathbf{H}}$ denote the multiplication operator $M_{\mathbf{H}} : \mathbf{h}(s) \mapsto \mathbf{H}(s) \cdot \mathbf{h}(s)$ acting from \mathcal{R}^m into \mathcal{R}^p . We shall be interested in the graph of the operator $M_{\mathbf{H}}|_{\mathcal{S}_{\Xi_{\mathfrak{D}_-}} : \mathcal{S}_{\Xi_{\mathfrak{D}_-}} \rightarrow \mathcal{R}^p$, i.e.,

$$\mathcal{G}_{\mathbf{H}} = \begin{bmatrix} \mathbf{H} \\ I \end{bmatrix} \cdot \mathcal{S}_{\Xi_{\mathfrak{D}_-}} \subset \mathcal{R}^{p+m}. \quad (9)$$

Then the following result gives a reformulation of the interpolation problem (2) as a geometric condition on $\mathcal{G}_{\mathbf{H}}$.

Theorem 2 *A rational matrix function $\mathbf{H} \in \mathcal{R}^{p \times m}$ is analytic on ω and satisfies the interpolation conditions (2) if and only if its graph subspace (9) is contained in the null-pole subspace $\mathcal{S}_{\Xi_{\mathfrak{D}}}$ (6) associated with the interpolation data set \mathfrak{D} (1):*

$$\mathcal{G}_{\mathbf{H}} \subset \mathcal{S}_{\Xi_{\mathfrak{D}}}.$$

Proof (Sketch of proof:) We sketch only one direction: we suppose that $\mathbf{H} \in \mathcal{R}^{p \times m}(\omega)$ and that \mathbf{H} satisfies the interpolation conditions (2). We want to check that then $\mathcal{G}_{\mathbf{H}} \subset \mathcal{S}_{\Xi_{\mathfrak{D}}}$. First note that a generic element of $\mathcal{G}_{\mathbf{H}}$ has the explicit form

$$\begin{bmatrix} \mathbf{H}(s) \\ I \end{bmatrix} \left(\sum_{j=1}^k \mathbf{r}_j (s - \lambda_j)^{-1} c_j + \mathbf{h}(s) \right)$$

for some scalars $c_1, \dots, c_k \in \mathbb{C}$ and some $h \in \mathcal{R}^m(\omega)$. Making use of the left tangential interpolation conditions in (2) we rewrite this as

$$\sum_{j=1}^k \begin{bmatrix} \mathbf{w}_j \\ \mathbf{r}_j \end{bmatrix} (s - \lambda_j)^{-1} c_j + \begin{bmatrix} \sum_{j=1}^k \frac{\mathbf{H}(s) - \mathbf{H}(\lambda_j)}{s - \lambda_j} \mathbf{r}_j c_j + \mathbf{H}(s) \mathbf{h}(s) \\ \mathbf{h}(s) \end{bmatrix}.$$

To check that this quantity is in $\mathcal{S}_{\Xi_{\mathfrak{D}}}$, from the criterion given in (8) we see that we must check that, for each $1 \leq i \leq q$,

$$\sum_{j=1}^k \ell_i^\top \frac{\mathbf{H}(\mu_i) - \mathbf{H}(\lambda_j)}{\mu_i - \lambda_j} \mathbf{r}_j c_j + \ell_i^\top \mathbf{H}(\mu_i) \mathbf{h}(\mu_i) - \mathbf{v}_i^\top \mathbf{h}(\mu_i) = \sum_{j=1}^k \frac{\mathbf{v}_i^\top \mathbf{r}_j - \ell_i^\top \mathbf{w}_j}{\mu_i - \lambda_j} c_j.$$

If we now make use of the interpolation conditions (2), we see that this identity falls out. This computation illustrates the role of the Loewner matrix \mathbb{L} in the Grassmannian approach to the bitangential rational interpolation problem.

The converse direction is not much more difficult but we omit the details here. \square

The next step is to apply Theorem 1: there is a nondegenerate (i.e., with determinant not vanishing identically) $(p + m) \times (p + m)$ rational matrix function Θ so that $\mathcal{S}_{\Xi_{\mathfrak{D}}} = \Theta \cdot \mathcal{R}^{p \times m}(\omega)$. This enables us to arrive at the next result.

Theorem 3 *Let Θ be a nondegenerate $(p \times m) \times (p \times m)$ rational matrix function such that $\mathcal{S}_{\Xi_{\mathfrak{D}}} = \Theta \cdot \mathcal{R}^{p \times m}(\omega)$ as guaranteed by Theorem 1. Then \mathbf{H} is analytic on ω and satisfies the interpolation conditions (2) if and only if there exists a parameter pair of rational functions $\mathbf{P} \in \mathcal{R}^{p \times m}(\omega)$ and $\mathbf{Q} \in \mathcal{R}^{m \times m}(\omega)$ such that the scalar rational function $\det(\Theta_{21}\mathbf{P} + \Theta_{22}\mathbf{Q})$ is analytic with nonzero values on $\omega \setminus \{\lambda_1, \dots, \lambda_d\}$ while having simple poles at the points $\{\lambda_1, \dots, \lambda_d\}$. We note that it can be shown that this latter condition holds for a generic choice of the parameter-pair (\mathbf{P}, \mathbf{Q}) .*

Let us note that higher order bitangential interpolation conditions (where one specifies not only values of $\ell_i(s)^\top \mathbf{H}(s)$ and of $\mathbf{H}(s)\mathbf{r}_j(s)$ at specified points μ_1, \dots, μ_q and $\lambda_1, \dots, \lambda_k$ but also an initial segment of the Taylor-series expansion at these points (where $\ell_i(z)$ and $\mathbf{r}_j(z)$ are specified vector-valued polynomials) can also be handled by this analysis by allowing the matrices \mathbf{A} and \mathbf{Z} to have more general Jordan structure; details can be found in [14] as well as in [13].

In practice the bitangential interpolation problem (2) is considered with additional constraints, as we next discuss.

2.1 Parametrization of Interpolants with Prescribed Macmillan Degree

The additional demand is to characterize the minimal possible Macmillan degree which an interpolant can have and more generally parametrize the interpolants giving any of the possible Macmillan degrees. To solve this problem via the generating system approach, it is important to construct the generating system matrix Θ to be column- (or row,- depending on one's conventions) reduced at infinity. This gives an application of the very problem solved in great generality in [19]. All this is discussed in [6] as well as in the survey [5, Chap. 4] with original results for the simplest case going back to [3, 9]. In particular, one can identify when the minimal Macmillan-degree solution is unique, and, when not, parametrize the set of minimal Macmillan-degree solutions (see Theorem 2.1 in [6] for a concise statement for the scalar case).

There has been much progress in the last couple of decades in the theory of *descriptor realizations*, i.e., in representations of a rational matrix function \mathbf{H} in the form

$$\mathbf{H}(s) = \mathbf{D} + \mathbf{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}$$

with \mathbf{E} not necessarily invertible (see [44] and the references there). In particular, by making use not only of the Loewner matrix \mathbb{L} (as in (3)) but also of the *shifted Loewner matrix* \mathbb{L}_s

$$\mathbb{L}_{\text{sh}} = \left[\frac{\mu_i \mathbf{v}_i^\top \mathbf{r}_j - \boldsymbol{\ell}_i^\top \mathbf{w}_j \lambda_j}{\mu_i - \lambda_j} \right]_{i=1, \dots, q; j=1, \dots, k} \in \mathbb{C}^{q \times k}$$

the result of [44] is that there is a compact descriptor realization formula involving the pencil $\mathbb{L}_{\text{sh}} - s\mathbb{L}$ for the generating system matrix Θ giving the representation $\mathcal{S}_{\Xi_{\mathcal{D}}} = \Theta \cdot \mathcal{R}^{p+k}(\omega)$ generalizing the formula (7) to the case where the unshifted Loewner matrix \mathbb{L} is not invertible. It would be of interest to understand this formula better for the case of a general admissible Sylvester data set Ξ (not necessarily of the form $\Xi_{\mathcal{D}}$ coming from the interpolation data set for the simplest bitangential interpolation problem (2)). Two-variable descriptor realizations for general bivariate rational matrix functions have been obtained in [7] based on the two-variable analogues of the Loewner and shifted Loewner matrices which again leads to more compact realization formulas than those obtained for proper realizations in the sense of Givone-Roesser or Fornasini-Marchesini for multidimensional systems (multidimensional-system analogues of proper (nondesoriptor) realizations); by way of comparison we mention the paper of Galkowski [33].

2.2 Nevanlinna-Pick Interpolation

The bitangential Nevanlinna-Pick interpolation is to assume that all the interpolation nodes $\lambda_1, \dots, \lambda_k, \mu_1, \dots, \mu_q$ are in a specified region Ω (let's say either the unit disk \mathbb{D} or the right half-plane \mathbb{C}_+) and then to demand that a solution \mathbf{H} , in addition to satisfying the interpolation conditions (2), also satisfies the norm constraint $\|\mathbf{H}(s)\| \leq 1$ for all $s \in \Omega$. This fits well with the Grassmannian approach in that the norm constraint on the solution \mathbf{H} translates to the graph $\mathcal{G}_{\mathbf{H}}$ being a maximal negative subspace in an appropriately specified ambient Kreĭn space. Then one imposes the constraint on the generating matrix Θ that it be Ω - J -inner, i.e., J -unitary on the boundary $\partial\Omega$ of Ω and J -contractive inside Ω , where $J = \begin{bmatrix} I_p & 0 \\ 0 & -I_m \end{bmatrix}$. Then it turns out that all parameter pairs of the form $(\mathbf{P}, \mathbf{Q}) = (\mathbf{S}, I_m)$ with \mathbf{S} having contractive values in Ω are all “good” parameter pairs in the sense of Theorem 3, and that restricting to good parameter pairs of this special form parametrizes the set of all solutions of the Nevanlinna-Pick interpolation problem. Another variation is to take $p = m$ and to ask that the solution \mathbf{H} have positive real part on Ω ($\mathbf{H}(s) + \mathbf{H}(s)^* > 0$ for $s \in \Omega$). The solution via the Grassmannian approach proceeds as above, but with $J = \begin{bmatrix} 0 & -I_m \\ -I_m & 0 \end{bmatrix}$; the condition that $\mathbf{H}(s)$ have strictly positive-definite real part for $s \in \Omega$ guarantees that parameter pairs of the form $(\mathbf{P}, \mathbf{Q}) = (\mathbf{G}, I_m)$ with \mathbf{G} a free-parameter with strictly positive-definite real part on Ω are all good, and that restriction to this class of good parameter pairs leads to a parametrization of the set of all solutions of the problem. For details see the book [14] and the references there as well as the expository account of the Kreĭn-space Grassmannian approach [11]. For a behavioral interpretation of the Grassmannian approach to Nevanlinna-Pick interpolation, see [10, 47]. Let us also note that there is also physical motivation

for considering the Nevanlinna-Pick interpolation problem with the addition of a constraint on the Macmillan degree of a solution; a small sample of work on this problem is [30, 45].

There is a connection between the rational interpolation problem without norm constraints and the Nevanlinna-Pick interpolation (with a norm constraint on the values of the interpolant) which can be explained as follows. For specificity we now take $\Omega = \mathbb{C}_+$. The fact that the generating system matrix Θ is J -unitary on $\partial\Omega = i\mathbb{R}$ tells us that $\Theta(-\bar{s})^{*-1} = J\Theta(s)J$ for s on the imaginary line, and then, by meromorphic continuation, on all of \mathbb{C} . Thus the fact that Θ has simple poles at the right interpolation nodes $\lambda_1, \dots, \lambda_k$ and simple zeros at the left interpolation nodes μ_1, \dots, μ_k in \mathbb{C}_+ implies that the Θ has simple zeros at the reflections of the right interpolation nodes $-\bar{\lambda}_1, \dots, -\bar{\lambda}_k$ and has simple poles at the reflections of the left interpolation nodes $-\bar{\mu}_1, \dots, -\bar{\mu}_q$. In fact this is the idea of how to construct a J -inner Θ with prescribed zero/pole structure in \mathbb{C}_+ : imposing the additional reflected zero/pole structure in \mathbb{C}_- determines Θ up to a right constant factor; one can then use the flexibility of this right constant factor to guarantee that Θ has J -unitary values on the imaginary line. Furthermore there is a natural operation of reflection of left tangential interpolation conditions to right tangential interpolation conditions and of right tangential interpolation conditions to left tangential interpolation conditions. Adding these reflected interpolation conditions to the original gives a new enlarged interpolation problem which has its own generating matrix in the sense of rational interpolation without norm constraint. Constructing the generating system matrix for this enlarged problem without the norm constraint leads to a construction of the J -unitary generating system for the original problem with the norm constraint imposed (one can use the invertible constant of freedom to guarantee J -unitary values on the imaginary line). In this way we arrive at the main thesis of Anderson-Antoulas in [4]: *one can view the Nevanlinna-Pick interpolation problem as equivalent to an interpolation problem without norm constraint.*

3 Other Several-Variable Generalizations: Free Noncommutative Rational Functions

We have already mentioned the extension of the Loewner-matrix approach to minimal realization and interpolation problems for bivariate rational functions (at least for the scalar-valued case) in [7] as well as some extensions of Nevanlinna-Pick interpolation to multivariable situations in [1, 12, 24, 25]. Let us mention also the work of [22, 23, 32] on multivariable Nevanlinna-Pick interpolation and connections with H^∞ -control for multidimensional linear systems. What we would like to discuss in the space remaining here, however, is in a somewhat more exotic direction, namely the recently emerging theory of free noncommutative rational matrix functions. This is a particular topic in the broader area of *free noncommutative function theory*, a type of *quantized function theory* whereby one studies functions of

several matrix variables of consistent but arbitrary square sizes satisfying a natural collection of compatibility conditions as one varies the matrix-tuple argument (see conditions (11), (12), (13) below). A first systematization of such a theory appears in the recent book of Kaliuzhnyi-Verbovetskyi and Vinnikov [41]. The theory has origins in the much earlier work of Taylor on noncommutative spectral theory [48, 49] as well as in work of Voiculescu [50–52] done with an eye toward applications to free probability (see e.g. [27, 28, 46] for applications of recent advances in the free noncommutative function theory to free probability). In systems and control theory, noncommutative rational functions and more generally formal power series come up in the theory of automata and formal languages going back to work of Fliess, Schützenberger, Kleene and others from the 1960s (see [29] for a comprehensive treatment), and also as transfer functions of multidimensional linear systems with evolution along a free monoid (see [15, 16, 26]) with applications to robust control of linear systems with Linear-Fractional-Transformation-modeled structured uncertainty (see [17, 18]). Noncommutative functions also come up in the program of Helton and collaborators to formulate optimization problems appearing in systems and control as dimension-independent, i.e., the natural variables are matrices rather than the scalar entries in the matrices, and engineering problems have to do with rational expressions in these matrix variables which have therefore the same form independent of matrix sizes. This has led to the search for understanding which optimization problems can be reduced to LMIs (Linear Matrix Inequalities) with attractive tractable solution algorithms, especially in the context of some convex structure in the statement of the problem (see e.g. [31, 35, 36] for surveys on various aspects of this topic). In addition there are instances where the noncommutative theory sheds light on the more standard function theory of several complex variables: examples are the realization theory (see [15]) and the theory of Nevanlinna-Pick interpolation (see [2, 21]).

3.1 Noncommutative Polynomials

It will be convenient to introduce some notation. We let \mathbb{F}_+^d denote the free monoid (or unital free semigroup) with d generators consisting of the first d integers $1, 2, \dots, d$. Thus elements of \mathbb{F}_+^d are words in the generators

$$\alpha = i_N \cdots i_1 \text{ where } i_j \in \{1, \dots, d\} \text{ for } 1 \leq j \leq N.$$

We let $|\alpha|$ denote the *length* of the word α ; thus $|\alpha| = N$ if $\alpha = i_N \cdots i_1$. Multiplication in \mathbb{F}_+^d is via concatenation

$$\alpha \cdot \beta = i_N \cdots i_1 j_M \cdots j_1 \text{ if } \alpha = i_N \cdots i_1 \text{ and } \beta = j_M \cdots j_1.$$

We let \emptyset denote the word with length equal to 0, so \emptyset is the vacuous word consisting of no letters; the importance of \emptyset is that it serves as the unit element in \mathbb{F}_+^d :

$$\alpha \cdot \emptyset = \emptyset \cdot \alpha = \alpha.$$

For z_1, \dots, z_d be a collection of freely noncommutative indeterminates and $\alpha \in \mathbb{F}_+^d$, we let z^α denote the free noncommutative monomial

$$z^\alpha = z_{i_N} \cdots z_{i_1} \text{ if } \alpha = i_N \cdots i_1. \quad (10)$$

A $p \times q$ -matrix noncommutative polynomial is a formal expression of the form

$$\mathfrak{p}(z) = \sum_{\alpha \in \mathbb{F}_+^d} \mathfrak{p}_\alpha z^\alpha$$

where $\mathfrak{p}_\alpha \in \mathbb{C}^{p \times q}$ and $\mathfrak{p}_\alpha = 0$ for all but finitely many $\alpha \in \mathbb{F}_+^d$. Any $p \times q$ -matrix noncommutative polynomial can be evaluated at a $Z = (Z_1, \dots, Z_d) \in (\mathbb{C}^{n \times n})^d$ (where $n \in \mathbb{N}$ is arbitrary) via a functional calculus based on tensor products:

$$\mathfrak{p}(Z) = \sum_{\alpha \in \mathbb{F}_+^d} \mathfrak{p}_\alpha \otimes Z^\alpha \in \mathbb{C}^{p \times q} \otimes \mathbb{C}^{n \times n} \cong \mathbb{C}^{pn \times qn},$$

where

$$Z^\alpha = Z_{i_N} \cdots Z_{i_1} \in \mathbb{C}^{n \times n} \text{ if } Z = (Z_1, \dots, Z_d) \text{ and } \alpha = i_N \cdots i_1 \in \mathbb{F}_+^d$$

is the extension of the monomial functional calculus (10) to d -tuples of $n \times n$ matrices, where the multiplication now is matrix multiplication in $\mathbb{C}^{n \times n}$ rather than concatenation of freely noncommutative indeterminates. The result is a functional calculus $Z \mapsto \mathfrak{p}(Z)$ for $Z = (Z_1, \dots, Z_d) \in \prod_{n=1}^\infty (\mathbb{C}^{n \times n})^d$ defined on the domain $\text{dom } \mathfrak{p} = \prod_{n=1}^\infty (\mathbb{C}^{n \times n})^d$ which satisfies the following properties:

1. *Gradedness:* For $\mathfrak{p} \in \mathbb{C}^{p \times q} \langle z \rangle$ and $Z \in (\mathbb{C}^{n \times n})^d$ we have

$$\mathfrak{p}(Z) \in \mathbb{C}^{pn \times qn}. \quad (11)$$

2. *Respect for direct sums:* For $Z^{(1)} = (Z_1^{(1)}, \dots, Z_d^{(1)}) \in (\mathbb{C}^{n \times n})^d$ and $Z^{(2)} = (Z_1^{(2)}, \dots, Z_d^{(2)}) \in (\mathbb{C}^{m \times m})^d$, set

$$\begin{bmatrix} Z^{(1)} & 0 \\ 0 & Z^{(2)} \end{bmatrix} = \left(\begin{bmatrix} Z_1^{(1)} & 0 \\ 0 & Z_1^{(2)} \end{bmatrix}, \dots, \begin{bmatrix} Z_d^{(1)} & 0 \\ 0 & Z_d^{(2)} \end{bmatrix} \right) \in (\mathbb{C}^{(n+m) \times (n+m)})^d.$$

Then

$$\mathfrak{p} \left(\begin{bmatrix} Z^{(1)} & 0 \\ 0 & Z^{(2)} \end{bmatrix} \right) = \begin{bmatrix} \mathfrak{p}(Z^{(1)}) & 0 \\ 0 & \mathfrak{p}(Z^{(2)}) \end{bmatrix}. \quad (12)$$

3. *Respect for similarities:* If $Z \in (\mathbb{C}^{n \times n})^d$ and $S \in \mathbb{C}^{n \times n}$ is invertible, set $SZS^{-1} = (SZ_1S^{-1}, \dots, SZ_dS^{-1})$. Then

$$\mathfrak{p}(SZS^{-1}) = (S \otimes I_p)\mathfrak{p}(Z)(S^{-1} \otimes I_q). \quad (13)$$

It is these properties which are taken as the axioms for a free noncommutative function in the comprehensive treatment of a general free noncommutative function theory in [41].

3.2 Noncommutative Rational Matrix Functions

There is a notion of noncommutative rational function which is obtained by completing the ring of freely noncommutative polynomials to the associated universal skew-field of fractions (see [38, 39, 53] and the references there for details and background). A concrete way of constructing such a skew-field developed only recently in [38, 39] is through rational expressions. By a *rational expression* we mean a syntactically valid combination of noncommutative polynomials, arithmetic operators $+$, \cdot , $^{-1}$ and parentheses. For \mathfrak{r} a rational expression define its domain $\text{dom } \mathfrak{r}$ to consist of all $Z = (Z_1, \dots, Z_d) \in \sqcup_{n=1}^{\infty} (\mathbb{C}^{n \times n})^d$ such that substitution of $Z = (Z_1, \dots, Z_d)$ for the formal indeterminates $z = (z_1, \dots, z_d)$ in the formal expression $\mathfrak{r}(z)$ is well-defined (i.e., one is not asked to invert an $n \times n$ matrix which is not invertible). Restrict to rational expressions with nonempty domains. Two rational expressions are equivalent if they agree on the intersection of their domains. The equivalence class of a rational expression \mathfrak{r} defines a rational function r with domain equal to the union of the domains of all the rational expressions in its equivalence class. With an additional adjustment of the domain introduced in [53] (called the *stable extended domain* of \mathfrak{r} , denoted as $\text{edom}^{\text{stab}} \mathfrak{r}$ or $\text{edom}^{\text{stab}} r$ for any $r \in \mathfrak{r}$) there is a well-defined matrix-evaluation $r \mapsto r(Z)$ for $Z \in \text{edom}^{\text{stab}}(r)$ so that the associated functional calculus $Z \mapsto r(Z)$ satisfies (11), (12), (13). Furthermore, all these ideas extend to the case where one starts with noncommutative polynomials with matrix coefficients and considers syntactically valid combinations of matrix noncommutative polynomials and their inverses, sums, products, and parentheses. In this way we arrive at a well-defined notion of *matrix noncommutative rational function* which is also a noncommutative function (i.e., the stable extended domain of r is invariant under direct sums and similarity transformations and r satisfies (11), (12), (13) when considered as a function on its stable extended domain).

Just as in the classical case, an important tool for the study of noncommutative rational functions is *realization theory*. In the multivariable context, it is more convenient to work with proper realizations centered at 0 rather than at an appropriate multivariable version of ∞ . In particular, we shall work with noncommutative Fornasini-Marchesini realizations determined by a system matrix of the form

$$\mathbf{U} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{B}_1 \\ \vdots & \vdots \\ \mathbf{A}_d & \mathbf{B}_d \\ \mathbf{C} & \mathbf{D} \end{bmatrix} : \begin{bmatrix} \mathcal{X} \\ \mathcal{U} \end{bmatrix} \rightarrow \begin{bmatrix} \mathcal{X}^d \\ \mathcal{Y} \end{bmatrix}$$

where the input space \mathcal{U} , the state space \mathcal{X} and the output space \mathcal{Y} are all finite-dimensional. Given such a system matrix we may form the noncommutative rational matrix expression of size $p \times q$ where $p = \dim \mathcal{Y}$ and $q = \dim \mathcal{U}$

$$\mathfrak{r}(z) = \mathbf{D} + \mathbf{C}(I - L_{FM}(z)\mathbf{A})^{-1}L_{FM}(z)\mathbf{B} \quad (14)$$

where we set $L_{FM}(z)$ equal to the linear row expression

$$L_{FM}(z) = [z_1 \cdots z_d] \otimes I_{\mathcal{X}}.$$

When r has a rational matrix function such that some rational matrix expression \mathfrak{r} in its equivalence class has a realization as in (14), we say that r has a noncommutative Fornasini-Marchesini realization. The formal power series associated with such a realization can be understood as a noncommutative Z -transform of the input-output map of an input/state/output linear system with evolution along a rooted tree (with each node having d branches emanating out from it), thereby giving parallel time-domain and frequency-domain universes in which one can work (see [15]), but we shall not have any explicit need for this observation here. We say that the realization (14) is *observable* if it is the case that

$$\bigcap_{\alpha \in \mathbb{F}_+^d} \text{Ker } \mathbf{C}\mathbf{A}^\alpha = \{0\}. \quad (15)$$

We say that the realization (14) is *controllable* if

$$\text{span}_{j: 1 \leq j \leq d, \alpha \in \mathbb{F}_+^d} \text{Ran } \mathbf{A}^\alpha \mathbf{B}_j = \mathcal{X}. \quad (16)$$

We say that the realization (14) is *minimal* if it is both controllable and observable. Then the following properties are known:

Properties of noncommutative rational matrix functions and their realizations:
(P1) (See [15, 29].) *Any rational matrix function r with 0 in $\mathcal{D}(\mathfrak{r})$ has a noncommutative Fornasini-Marchesini realization (14); more precisely, r is given by the rational expression \mathfrak{r} coming from a noncommutative Fornasini-Marchesini realization (14):*

$$\mathfrak{r}(z) = \mathbf{D} + \mathbf{C}(I - L_{FM}(z)\mathbf{A})^{-1}L_{FM}(z)\mathbf{B}.$$

Furthermore one can arrange that the realization is minimal.

(P2) (See [15, 29].) *The minimal possible state-space dimension in a noncommutative Fornasini-Marchesini realization of a given noncommutative rational matrix function*

r is equal to the state-space dimension $\dim \mathcal{X}$ in any controllable and observable (i.e., minimal) realization of r . Furthermore, any two minimal realizations

$$\mathbf{U} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{B}_1 \\ \vdots & \vdots \\ \mathbf{A}_d & \mathbf{B}_d \\ \mathbf{C} & \mathbf{D} \end{bmatrix}, \quad \mathbf{U}' = \begin{bmatrix} \mathbf{A}' & \mathbf{B}' \\ \mathbf{C}' & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{A}'_1 & \mathbf{B}'_1 \\ \vdots & \vdots \\ \mathbf{A}'_d & \mathbf{B}'_d \\ \mathbf{C}' & \mathbf{D} \end{bmatrix}$$

are similar in the following sense: there is an invertible linear transformation $\mathbf{S}: \mathcal{X} \rightarrow \mathcal{X}'$ so that

$$\begin{bmatrix} \mathbf{S} & & & & \\ & \ddots & & & \\ & & \mathbf{S} & & \\ & & & I_y & \\ & & & & \end{bmatrix} \begin{bmatrix} \mathbf{A}_1 & \mathbf{B}_1 \\ \vdots & \vdots \\ \mathbf{A}_d & \mathbf{B}_d \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{A}'_1 & \mathbf{B}'_1 \\ \vdots & \vdots \\ \mathbf{A}'_d & \mathbf{B}'_d \\ \mathbf{C}' & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & I_u \end{bmatrix}.$$

(P3) (See [39, 53].) Suppose that (14) is a minimal Fornasini-Marchesini realization for a noncommutative rational matrix function r . Then the stable extended domain of r is given by

$$\text{edom}^{\text{stab}} r = \coprod_{n=0}^{\infty} \{Z \in (\mathbb{C}^{n \times n})^d : \det L_{\mathbf{A}}(Z) := \det(I_{\mathcal{X}^n} - Z_1 \otimes \mathbf{A}_1 - \cdots - Z_d \otimes \mathbf{A}_d) \neq 0\}.$$

(P4) (See [37].) If $r(z)$ is a noncommutative rational function without any singularities (so $\text{dom } r = \coprod_{n=1}^{\infty} (\mathbb{C}^{n \times n})^d$), then r is a polynomial.

To illustrate the ideas and techniques of this emerging field of study, we offer a modest first step toward the development of a free noncommutative version of the module theory for single-variable rational matrix functions in the spirit of Theorem 1 above. In the discussion to follow we identify a noncommutative rational function with a convenient choice of rational expression in its equivalence class.

Theorem 4 Suppose that $\Theta \in \mathbb{C}^{N \times N} \langle z \rangle$ is a noncommutative square-matrix polynomial such that $\Theta(0) = \Theta_{\emptyset}$ is invertible. Then there is a controllable input pair (\mathbf{Z}, \mathbf{B}) in the sense of (16) (where say \mathbf{Z} has size $dN' \times N'$ and \mathbf{B} has size $dN' \times N$) so that the right noncommutative-polynomial module

$$\mathcal{S}_{\Theta} = \Theta \cdot \mathbb{C}^N \langle z \rangle \tag{17}$$

has the alternative characterization as

$$\mathcal{S}_{\Theta} = \left\{ \mathbf{f} \in \mathbb{C}^N \langle z \rangle : (I - L_{FM}(z)\mathbf{Z})^{-1} L_{FM}(z)\mathbf{B}\mathbf{f}(z) \in \mathbb{C}^{N'} \langle z \rangle \right\}. \tag{18}$$

Proof Let

$$\Theta(z) = \mathbf{D}_0 + \mathbf{C}_0(I - L_{FM}(z)\mathbf{A}_0)^{-1} L_{FM}(z)\mathbf{B}_0$$

be a minimal noncommutative FM-realization for $\Theta(z)$ where by assumption \mathbf{D}_0 is invertible. Then $\Theta(z)^{-1}$ is a noncommutative rational matrix function with minimal FM-realization given by

$$\Theta(z)^{-1} = \mathbf{D}_0^{-1} - \mathbf{D}_0^{-1}\mathbf{C}_0(I - L_{FM}(z)\mathbf{A}_0^\times)^{-1}L_{FM}(z)\mathbf{B}_0\mathbf{D}_0^{-1} \quad (19)$$

where $\mathbf{A}_0^\times := \mathbf{A}_0 - \mathbf{B}_0\mathbf{D}_0^{-1}\mathbf{C}_0$ (see [15]). We shall show that the characterization (18) holds with

$$(\mathbf{Z}, \mathbf{B}) = (\mathbf{A}_0^\times, \mathbf{B}_0\mathbf{D}_0^{-1}). \quad (20)$$

Suppose first that $\mathbf{f} \in \mathbb{C}^N\langle z \rangle$ is a noncommutative polynomial such that $(I - L_{FM}(z)\mathbf{A}_0^\times)^{-1}L_{FM}(z)\mathbf{B}_0\mathbf{D}_0^{-1}\mathbf{f}(z)$ is a noncommutative polynomial (in $\mathbb{C}^{N \times N}\langle z \rangle$). As multiplication on the left by a constant matrix preserves polynomials, we conclude that $-\mathbf{D}_0^{-1}\mathbf{C}_0(I - L_{FM}(z)\mathbf{A}_0^\times)^{-1}L_{FM}(z)\mathbf{B}_0\mathbf{D}_0^{-1}\mathbf{f}(z)$ is also a polynomial. As $\mathbf{f}(z)$ is a polynomial, it follows that $\mathbf{D}_0^{-1}\mathbf{f}(z)$ is a polynomial. Adding these last two expressions together and making use of (19) gives us that

$$\mathbf{g}(z) = \Theta(z)^{-1}\mathbf{f}(z) = \left(\mathbf{D}_0^{-1} - \mathbf{D}_0^{-1}\mathbf{C}_0(I - L_{FM}(z)\mathbf{A}_0^\times)^{-1}L_{FM}(z)\mathbf{B}_0\mathbf{D}_0^{-1} \right) \mathbf{f}(z)$$

is a polynomial. However we can recover \mathbf{f} from \mathbf{g} via $\mathbf{f} = \Theta \cdot \mathbf{g}$. As \mathbf{g} is a polynomial, we conclude that \mathbf{f} has the requisite form for membership in \mathcal{S}_Θ (17).

Conversely, suppose that $\mathbf{f} = \Theta \cdot \mathbf{g}$ with \mathbf{g} a polynomial in $\mathbb{C}^N\langle z \rangle$. Then from (19) we see that

$$\mathbf{g} = \Theta^{-1}(z)\mathbf{f}(z) = \left(\mathbf{D}_0^{-1} - \mathbf{D}_0^{-1}\mathbf{C}_0(I - L_{FM}(z)\mathbf{A}_0^\times)^{-1}L_{FM}(z)\mathbf{B}_0\mathbf{D}_0^{-1} \right) \mathbf{f}(z) \quad (21)$$

is a polynomial. As Θ and \mathbf{g} are polynomials, so also is $\mathbf{f} = \Theta\mathbf{g}$ and hence also $\mathbf{D}_0^{-1}\mathbf{f}(z)$. Thus the second term on the right-hand side of (21) is a polynomial:

$$\mathbf{D}_0^{-1}\mathbf{C}_0(I - L_{FM}(z)\mathbf{A}_0^\times)^{-1}L_{FM}(z)\mathbf{B}_0\mathbf{D}_0^{-1}\mathbf{f}(z) \in \mathbb{C}^N\langle z \rangle. \quad (22)$$

Making use of the identity

$$(I - L_{FM}(z)\mathbf{A}_0^\times)^{-1} = I + L_{FM}(z)\mathbf{A}_0^\times(I - L_{FM}(z)\mathbf{A}_0^\times)^{-1}$$

and the fact that $\mathbf{D}_0^{-1}\mathbf{C}_0L_{FM}(z)\mathbf{B}_0\mathbf{D}_0^{-1}\mathbf{f}(z)$ is a polynomial (since $\mathbf{f}(z)$ is a polynomial), we see that then

$$\mathbf{D}_0^{-1}\mathbf{C}_0L_{FM}(z)\mathbf{A}_0^\times(I - L_{FM}(z)\mathbf{A}_0^\times)^{-1}L_{FM}(z)\mathbf{B}_0\mathbf{D}_0^{-1}\mathbf{f}(z)$$

is a polynomial. Rewrite this last expression as

$$\sum_{j=1}^d z_j \mathbf{D}_0^{-1} \mathbf{C}_0 \mathbf{A}_{0,j}^{\times} (I - L_{FM}(z) \mathbf{A}_0^{\times})^{-1} L_{FM}(z) \mathbf{B}_0 \mathbf{D}_0^{-1} \mathbf{f}(z) \in \mathbb{C}^N \langle z \rangle. \quad (23)$$

It is well known that the algebra of matrix-valued noncommutative rational expressions regular at zero can be embedded into the algebra of noncommutative formal power series with matrix coefficients (see e.g. [40, Sect. 4.5]). But in the algebra of formal power series in freely noncommutative indeterminates, monomials of the form $z_j z^{\beta}$ are linearly independent of monomials of the form $z_k z^{\beta'}$ whenever $j \neq k$. (Note that this step fails miserably if the variables (z_1, \dots, z_d) are commutative and $d > 1$!). Hence the fact that the expression (23) is a polynomial implies that each term in the summation is itself a polynomial:

$$z_j \mathbf{D}_0^{-1} \mathbf{C}_0 \mathbf{A}_{0,j}^{\times} (I - L_{FM}(z) \mathbf{A}_0^{\times})^{-1} L_{FM}(z) \mathbf{B}_0 \mathbf{D}_0^{-1} \mathbf{f}(z) \in \mathbb{C}^N \langle z \rangle \text{ for } j = 1, \dots, d.$$

We view this last expression as a formal power series. As it is also a polynomial, when we cancel off the left factor z_j the result is still a polynomial. In other words, we have shown that

$$\mathbf{D}_0^{-1} \mathbf{C}_0 \mathbf{A}_0^{\times \alpha} (I - L_{FM}(z) \mathbf{A}_0^{\times})^{-1} L_{FM}(z) \mathbf{B}_0 \mathbf{D}_0^{-1} \mathbf{f}(z) \in \mathbb{C}^N \langle z \rangle \quad (24)$$

for $\alpha \in \mathbb{F}_+^d$ with $|\alpha| = 1$. Note that the statement (22) is the same statement as (24) but with $\alpha \in \mathbb{F}_+^d$ constrained to satisfy $|\alpha| = 0$.

Inductively assume that (24) holds for all $\alpha \in \mathbb{F}_+^d$ with $|\alpha| = M$ for some $M \in \mathbb{N}$. An argument exactly paralleling the one just completed then gives us that (24) also holds for all $\alpha \in \mathbb{F}_+^d$ with $|\alpha| = M + 1$. We conclude that: for all $M = 0, 1, 2, \dots$,

$$\text{col}_{\alpha \in \mathbb{F}_+^d: |\alpha| \leq M} \left[\mathbf{D}_0^{-1} \mathbf{C}_0 \mathbf{A}_0^{\times \alpha} (I - L_{FM}(z) \mathbf{A}_0^{\times})^{-1} L_{FM}(z) \mathbf{B}_0 \mathbf{D}_0^{-1} \mathbf{f}(z) \right] \in \mathbb{C}^N \langle \sum_{j=0}^M d^j \rangle.$$

But $(-\mathbf{D}_0^{-1} \mathbf{C}_0, \mathbf{A}_0^{\times})$, as the output pair in the minimal realization (19) for Θ^{-1} , is observable in the sense of (15); thus once M is large enough, the block column matrix $\text{col}_{\alpha \in \mathbb{F}_+^d: |\alpha| \leq M} \left[\mathbf{D}_0^{-1} \mathbf{C}_0 \mathbf{A}_0^{\times \alpha} \right]$ is left invertible. Multiplying on the left by this left inverse then gives us that $(I - L_{FM}(z) \mathbf{A}_0^{\times})^{-1} L_{FM}(z) \mathbf{B}_0 \mathbf{D}_0^{-1} \mathbf{f}(z)$ is a polynomial, i.e., that \mathbf{f} satisfies the criterion (18) with (\mathbf{Z}, \mathbf{B}) given by (20). \square

4 Conclusion

We have reviewed the geometric Grassmannian approach to rational matrix-valued interpolation and its connections with the more algebraic approach based on Loewner matrices. We have also indicated how some of these ideas extend to the relatively new and still evolving area of free noncommutative function theory.

References

1. Agler, J.: On the representation of certain holomorphic functions defined on a polydisk. *Topics in Operator Theory: Ernst D. Hellinger Memorial. Operator Theory: Advances and Applications*, vol. 48, pp. 47–66. Birkhäuser, Basel (1990)
2. Agler, J., McCarthy, J.E.: Pick interpolation for free holomorphic functions. *Amer. J. Math.* **137**(6), 241–285 (2015)
3. Anderson, B.D.O., Antoulas, A.C.: On the scalar rational interpolation problem. *IMA J. Math. Control Inf.* **3**, 61–88 (1986)
4. Anderson, B.D.O., Antoulas, A.C.: On the problem of stable rational interpolation. *Linear Algebra Appl.* **122–124**, 301–329 (1989)
5. Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems. *Design and Control 6*. SIAM Philadelphia (2005)
6. Antoulas, A.C., Ball, J.A., Kang, J., Willems, J.C.: On the solution of the minimal rational interpolation problem. *Linear Algebra Appl.* **137**(138), 511–573 (1990)
7. Antoulas, A.C., Ionita, A.C., Lefteriu, S.: On two-variable rational interpolation. *Linear Algebra Appl.* **436**, 2889–2915 (2012)
8. Antoulas, A.C., Lefteriu, S., Ionita, A.C.: A Tutorial Introduction to the Loewner Framework for Model Reduction. *Model Reduction & Approximation. Computer Science Engineering*, vol. 15, pp. 335–376. SIAM, Philadelphia (2017)
9. Antoulas, A.C., Willems, J.C.: Minimal rational interpolation and Prony’s method. *Analysis and Optimization of Systems (Antibes, 1990). Lecture Notes in Control and Information Sciences*, vol. 144, pp. 297–306. Springer, Berlin (1990)
10. Antoulas, A.C., Willems, J.C.: A behavioral approach to linear exact modelling. *IEEE Trans. Automat. Control* **38**, 1776–1802 (1993)
11. Ball, J.A.: Nevanlinna-Pick interpolation: generalizations and applications. *Surveys of Some Recent Results in Operator Theory, Vol. I. Pitman Research Notes in Mathematics Series*, vol. 171, pp. 51–94. Longman Sci. Tech., Harlow (1988)
12. Ball, J.A., Fang, Q.: Nevanlinna-Pick interpolation via graph spaces and Kreĭn-space geometry: a survey. *Mathematical Methods in Systems, Optimization, and Control. Operator Theory: Advances and Applications*, vol. 222, pp. 43–71. Birkhäuser/Springer, Basel (2012)
13. Ball, J.A., Gohberg, I., Rodman, L.: Two-sided Lagrange-Sylvester interpolation problems for rational matrix functions. *Operator Theory and Applications, Part 1 (Durham, NH, 1988). Proceedings of Symposia in Pure Mathematics*, vol. 51, pp. 17–83. Part 1, Amer. Math. Soc., Providence, RI (1990)
14. Ball, J.A., Gohberg, I., Rodman, L.: Interpolation of Rational Matrix Functions. *Operator Theory: Advances and Applications*, vol. 45. Birkhäuser, Basel (1990)
15. Ball, J.A., Groenewald, G., Malakorn, T.: Structured noncommutative multidimensional linear systems. *SIAM J. Control Optim.* **44**, 1474–1528 (2005)
16. Ball, J.A., Groenewald, G., Malakorn, T.: Conservative structured noncommutative multidimensional linear systems. In: *The State Space Method: Generalizations and Applications. Operator Theory: Advances and Applications*, vol. 161, pp. 179–223. Birkhäuser, Basel (2006)
17. Ball, J.A., Groenewald, G., Malakorn, T.: Bounded real lemmas for structured noncommutative multidimensional linear systems and robust control. *Multidimens. Syst. Signal Process.* **17**(2–3), 110–150 (2006)
18. Ball, J.A., Groenewald, G., ter Horst, S.: Bounded real lemma and structured singular value versus diagonal scaling: the free noncommutative setting. *Multidimens. Syst. Signal Process.* **27**(1), 217–254 (2016)
19. Ball, J.A., Kaashoek, M.A., Groenewald, G.J., Kim, J.: Column reduced rational matrix functions with given null-pole data in the complex plane. *Linear Algebra Appl.* **203–204**, 67–110 (1994)
20. Ball, J.A., Helton, J.W.: A Beurling-Lax theorem for the Lie group $U(m, n)$ which contains most classical interpolation theorem. *J. Operator Theory* **9**, 107–142 (1983)

21. Ball, J.A., Marx, G., Vinnikov, V.: Interpolation and transfer-function realization for the non-commutative Schur-Agler class. *Operator Theory in Different Settings and Related Applications. Operator Theory: Advances and Applications*, vol. 115, pp. 23–116. Birkhäuser, Cham (2018)
22. Ball, J.A., Malakorn, T.: Multidimensional linear feedback control systems and interpolation problems for multivariable holomorphic functions. *Multidimens. Systems Signal Process.* **15**(1), 7–36 (2004)
23. Ball, J.A., ter Horst, S.: Robust control, multidimensional systems and multivariable Nevanlinna-Pick interpolation. *Topics in Operator Theory Vol. 2, Systems and Mathematical Physics. Operator Theory: Advances and Applications*, vol. 203, pp. , 13–88. Birkhäuser-Verlag, Basel (2010)
24. Ball, J.A., Trent, T.T.: Unitary colligations, reproducing kernel Hilbert spaces, and Nevanlinna-Pick interpolation in several variables. *J. Funct. Anal.* **157**(1), 1–61 (1998)
25. Ball, J.A., Trent, T.T., Vinnikov, V.: Interpolation and commutant lifting for multipliers on reproducing kernel Hilbert spaces. *Operator Theory and Analysis (Amsterdam, 1997). Operator Theory: Advances and Applications*, vol. 122, pp. 89–138. Birkhäuser, Basel (2001)
26. Ball, J.A., Vinnikov, V.: Lax-Phillips scattering and conservative linear systems: a Cuntz-algebra multidimensional setting. *Memoirs Amer. Math. Soc.* **178**(837) (2005)
27. Belinschi, S.T., Popa, M., Vinnikov, V.: Infinite divisibility and a non-commutative Boolean-to-free Bercovici-Pata bijection. *J. Funct. Anal.* **262**(1), 94–134 (2012)
28. Belinschi, S.T., Popa, M., Vinnikov, V.: On the operator-valued analogues of the semicircle, arcsine and Bernoulli laws. *J. Operator Theory* **70**, 239–258 (2013)
29. Berstel, J., Reutenauer, C.: Noncommutative rational series with applications. *Encyclopedia of Mathematics and its Applications*, vol. 137. Cambridge University Press, Cambridge (2011)
30. Blomqvist, A., Lindquist, A., Nagamune, R.: Matrix-valued Nevanlinna-Pick interpolation with complexity constraint: an optimization approach. *IEEE Trans. Autom. Control* **48**(12), 2172–2190 (2003)
31. de Oliveira, M.C., Helton, J.W., McCullough, S.A., Putinar, M.: Engineering systems and free semi-algebraic geometry. In: Putinar, M., Sullivant, S. (eds.) *Emerging Applications of Algebraic Geometry. The IMA Volumes in Mathematics and its Applications*, pp. 17–61. Springer (2009)
32. Feng, Z.-Y., Wu, Q., Xu, L.: \mathcal{H}_∞ control of linear multidimensional discrete systems. *Multidimens. Syst. Sign. Process.* **23**, 381–411 (2012)
33. Galkowski, K.: Minimal state-space realization for a class of nD systems. *Recent advances in operator theory and its applications. Operator Theory: Advances and Applications*, vol. 160, pp. 179–194. Birkhäuser, Basel (2005)
34. Gohberg, I., Kaashoek, M.A., Ran, A.C.M.: Regular rational matrix functions with prescribed null and pole data except at infinity. *Linear Algebra Appl.* **137**(138), 387–412 (1990)
35. Helton, J.W.: Manipulating matrix inequalities automatically. In: *Mathematical Systems Theory in Biology, Communications, Computation, and Finance (Notre Dame, IN, 2002). IMA Vol. Math. Appl.*, pp. 237–256. Springer, New York (2003)
36. Helton, J.W., Klep, I., McCullough, S.: Free convex algebraic geometry. In: Blekherman, G., Parrilo, P.A., Thomas, R.R. (eds.) *Semidefinite Optimization and Convex Algebraic Geometry*, pp. 341–405. SIAM & MOS, Philadelphia (2013)
37. Helton, J.W., Klep, I., Volčič, J.: Geometry of free loci and factorization of noncommutative polynomials. *Adv. Math.* **331**, 589–626 (2018)
38. Helton, J.W., McCullough, S.A., Vinnikov, V.: Noncommutative convexity arises from linear matrix inequalities. *J. Funct. Anal.* **240**(1), 105–191 (2006)
39. Kaliuzhnyi-Verbovetskyi, D.S., Vinnikov, V.: Singularities of rational functions and minimal factorizations: the noncommutative and the commutative setting. *Linear Algebra Appl.* **430**, 869–889 (2009)
40. Kaliuzhnyi-Verbovetskyi, D.S., Vinnikov, V.: Noncommutative rational functions, their difference-differential calculus and realizations. *Multidimens. Syst. Sign. Process.* **23**, 49–77 (2012)

41. Kaliuzhnyi-Verbovetskyi, D.S., Vinnikov, V.: Foundations of Free Noncommutative Function Theory, Mathematical Surveys and Monographs, vol. 199. American Mathematical Society, Providence, RI (2014)
42. Katsnelson, V.E.: On transformations of Potapov's fundamental matrix inequality. Topics in Interpolation Theory. Operator Theory: Advances and Applications, vol. 95, pp. 253–281 (1997)
43. Katsnelson, V.E., Kheifets, A.Ya., Yuditskii, P.: An abstract interpolation problem and the extension theory of isometric operators, Topics in Interpolation Theory. Operator Theory: Advances and Applications, vol. 95, pp. 283–298 (1997)
44. Mayo, A.J., Antoulas, A.C.: A framework for the solution of the generalized realization problem. Linear Algebra Appl. **425**, 634–662 (2007)
45. Michaletzky, G., Gombani, A.: On the “redundant” null-pairs of functions connected by a general Linear Fractional Transformation. Math. Control Signals Syst. **24**, 443–475 (2012)
46. Popa, M., Vinnikov, V.: Non-commutative functions and non-commutative free Levy-Hincin formula. Adv. Math. **236**, 131–157 (2013)
47. Rapisarda, P., Willems, J.C.: The subspace Nevanlinna interpolation problem and the most powerful unfalsified model. Syst. Control Lett. **32**, 291–300 (1997)
48. Taylor, J.L.: A general framework for a multi-operator functional calculus. Adv. Math. **9**, 183252 (1972)
49. Taylor, J.L.: Functions of several noncommuting variables. Bull. Amer. Mat. Soc. **79**, 1–34 (1973)
50. Voiculescu, D.-V.: Free analysis questions I: duality transform for the Coalgebra of $\partial_{X:B}$. Int. Math. Res. Not. **16**, 793–822 (2004)
51. Voiculescu, D.-V.: Free analysis questions II: the Grassmannian completion and the series expansion at the origin. J. Reine Angew. Math. **645**, 155–236 (2010)
52. Voiculescu, D.-V., Dykema, K.J., Nica, A.: Free random variables. A noncommutative probability approach to free products with applications to random matrices, operator algebras and harmonic analysis on free groups. CRM Monograph Series I. American Mathematical Society, Providence, RI (1992)
53. Volčič, J.: On domains of noncommutative rational functions. Linear Algebra Appl. **516**, 69–81 (2017)

The Conditioning of a Linear Barycentric Rational Interpolant



Jean-Paul Berrut

Abstract We study what a referee has asked us to call the “second Berrut rational interpolant”, introduced thirty years ago, which is linear in the interpolated function values and converges exponentially when the nodes are conformal images of Chebyshev points. We show that the Lebesgue constant for this interpolant grows just logarithmically with the number of nodes under reasonable assumptions, extending results of Bos et al. in 2013 for the corresponding first interpolant.

Keywords Linear rational interpolation · Conditioning · Logarithmic growth · Lebesgue constant · Conformally shifted nodes

1 Introduction

Polynomial interpolation has been a standard computational tool in applied mathematics at least since the time of Newton, Waring and Lagrange [12]. In its simplest version, it consists in associating to a vector $\mathbf{x} = [x_0, \dots, x_n]^T$ of, say, $n + 1$ distinct abscissae, and a vector $\mathbf{f} = [f_0, \dots, f_n]^T$ of corresponding ordinates, which may be values of a function f or not, the unique polynomial $p_n[\mathbf{f}]$ of degree at most n that takes on the value f_j at x_j , $j = 0, \dots, n$. The corresponding mapping has several nice properties. It is linear: the polynomial interpolating a linear combination of \mathbf{f} -vectors is the linear combination of the corresponding polynomials, i.e., $p_n[\alpha\mathbf{f} + \beta\mathbf{g}] = \alpha p_n[\mathbf{f}] + \beta p_n[\mathbf{g}]$, $\forall \alpha, \beta \in \mathbf{R}$; moreover, the uniqueness makes it a projection, a property that permits the study of the conditioning by means of the Lebesgue constant.

But it has its drawbacks, the main one being the fact that it is useful for large n only if the nodes are located in particular ways in the interpolation interval $[a, b] \equiv$

For Thanos Antoulas, on the occasion of his 70th birthday and in remembrance of our undergraduate years together at the ETH in Zurich.

J.-P. Berrut (✉)

Department of Mathematics, University of Fribourg, Pérolles 1700 Fribourg, Switzerland
e-mail: jean-paul.berrut@unifr.ch

© Springer Nature Switzerland AG 2022

C. Beattie et al. (eds.), *Realization and Model Reduction of Dynamical Systems*,
https://doi.org/10.1007/978-3-030-95157-3_2

$[\min x_j, \max x_j]$ [19, Chap. 5]. Long before the advent of electronic computing, for instance, Runge showed that, for the most important case of equidistant nodes, the interpolant of a meromorphic function converges only if the poles of the latter lie outside a certain region containing $[a, b]$ [12]. And even for functions, such as e^x , for which the interpolant converges for equidistant points, it is extremely ill-conditioned for n large [13, Sect. 5.5].

It is therefore appropriate to look for alternative infinitely smooth interpolants which work well for general sets of nodes. In [3], the author took advantage of the observation by Werner [22] that, by modifying the so-called *weights* λ_j in the barycentric formula

$$p_n[\mathbf{f}](x) = \sum_{j=0}^n \frac{\lambda_j}{x - x_j} f_j \bigg/ \sum_{j=0}^n \frac{\lambda_j}{x - x_j}, \quad \lambda_j := 1 \bigg/ \prod_{k \neq j} (x_j - x_k), \quad (1)$$

for p_n , one obtains a rational function, say $r_n[f]$, of degree $\leq n$, i.e., with both numerator and denominator degrees at most n ; the new suggestion was to choose weights β_j different from the λ_j to obtain a good denominator for the given set of nodes, instead of the unique denominator 1 for all sets of nodes in polynomial interpolation. That way the two properties of linearity and projection are preserved [4].

A good choice of the β_j requires the ordering of the nodes according to

$$a = x_0 < x_1 < \dots < x_n = b; \quad (2)$$

indeed, the weights should alternate in sign, for otherwise the interpolant will not reproduce all linear functions [14]. In view of this, the simplest possible choice of the weights, as proposed in [3], is

$$\beta_j = (-1)^j, \quad (3)$$

independently of the values of the nodes. We denote the corresponding interpolant by R_0 ; the referee has suggested that this should be called the “first Berrut rational interpolant” (see [14]). It was shown in [3] that R_0 has no real poles; moreover, its $O(h)$ -convergence (where $h := \max_{i=0, \dots, n-1} (x_{i+1} - x_i)$) and well-conditioning were conjectured. The $O(h)$ -convergence was proved in [11] under a local mesh ratio condition.

For an approximation operator which is a projection, the conditioning is best addressed through its Lebesgue constant Λ_n (see Sect. 2). For R_0 , the latter has been shown to increase only logarithmically for the most frequently used sets of nodes [9], a behaviour as nice as that of polynomial interpolation with the most favourable nodes (see also [21]).

However, by considering the even rational trigonometric interpolant obtained by transferring the polynomial onto the circle with the transform $x = \cos \phi$ [3], one comes to the conclusion that the first and last terms of the sums in both the numerator

and the denominator of the interpolant should appear only once and the other terms twice. After simplifying by a factor of two, one sees that the weights [3]

$$\beta_j = (-1)^j \delta_j, \quad \delta_j := \begin{cases} 1/2, & j = 0 \text{ or } j = n, \\ 1, & \text{otherwise,} \end{cases} \quad (4)$$

should be better than (3). (To be precise, this formula for the weights holds in the present case, where $x_0 = a$ and $x_n = b$, i.e., where the interval $[x_0, x_n]$ covers the whole interval of interpolation: when the interpolant is to be used beyond one or both extremities, slightly more complicated formulae involving trigonometric functions should be used, see [3].) We denote the corresponding interpolant by R_1 , the “second Berrut rational interpolant”. When the nodes are the Chebyshev points of the second kind, the weights (4) coincide with the λ_j [13, p. 252], $R_1 \equiv p_n[\mathbf{f}]$, and exponential convergence is achieved if f is analytic in an ellipse containing $[a, b]$ [20, p. 57]. It was shown in [3] that R_1 does not have any poles in $[a, b]$ for any set of nodes, and conjectured in the same work that R_1 is well-conditioned. Additionally, it was proved in [2] that R_1 retains the exponential convergence when the x_j are conformally shifted Chebyshev nodes, and conjectured that the convergence is $O(h^2)$ in general; conditions guaranteeing the latter are given in [5].

In 2007, Floater and Hormann [11] presented a family of interpolants with a higher order of convergence, which they denote by $d + 1$, $d \in \mathbf{N}^0$. For every d and every set of nodes \mathbf{x} , these authors give a linear rational interpolant $r_n^{(d)}[\mathbf{f}]$, depending on d , such that $\|r_n^{(d)}[\mathbf{f}] - f\| = O(h^{d+1})$. After constructing the interpolant as a blend of the $n - d$ interpolating polynomials corresponding to the subvectors of $d + 1$ consecutive nodes, they provide the weights of its barycentric representation (which always exists, as shown in [7]). For equidistant nodes, $r_n^{(0)}[\mathbf{f}]$ and $r_n^{(1)}[\mathbf{f}]$ correspond to R_0 and R_1 , respectively; however, this is not the case with other sets of nodes.

In the present paper, we study the Lebesgue constant of R_1 for equidistant and well-spaced nodes such as conformally shifted Chebyshev points of the second kind, which are the most important in practice. Our results are extensions of results of Bos et al. [9] for R_0 .

2 The Conditioning of the Interpolant R_1

The conditioning of a problem is the sensitivity of its solution to perturbations of its data. (In many contexts, this is called stability, but in numerical analysis the latter term should be reserved for the stability of an algorithm [13, p. 33].) Here the data are the function values \mathbf{f} and the solution is the interpolant

$$R_1(x) = \sum_{j=0}^n \frac{(-1)^j f_j}{x - x_j} \bigg/ \sum_{i=0}^n \frac{(-1)^i}{x - x_i}, \quad (5)$$

where the double prime attached to a sum indicates that its first and last terms are halved.

The fact that R_1 is a linear projection operator into a linear space of functions with a Lagrange basis (i.e., a basis $\{b_j(x)\}$ with the Lagrange property $b_j(x_i) = \delta_{ij}$) leads to the fact that the norm of that operator, called its Lebesgue constant, provides a good characterization of the conditioning. Indeed, if one perturbs each f_j by a number e_j , i.e., perturbing the vector \mathbf{f} with the vector $\mathbf{e} := (e_j)$, then for a linear interpolant $r_n[\mathbf{f}]$

$$r_n[\mathbf{f} + \mathbf{e}](x) = \sum_{j=0}^n (f_j + e_j)b_j(x).$$

The corresponding perturbation of the interpolant amounts to

$$r_n[\mathbf{f} + \mathbf{e}](x) - r_n[\mathbf{f}](x) = \sum_{j=0}^n e_j b_j(x)$$

so that

$$\begin{aligned} |r_n[\mathbf{f} + \mathbf{e}](x) - r_n[\mathbf{f}](x)| &\leq \sum_{j=0}^n |e_j| \cdot |b_j(x)| \\ &\leq \|\mathbf{e}\| \sum_{j=0}^n |b_j(x)| \end{aligned} \quad (6)$$

where $\|\cdot\|$ denotes the vector infinity norm.

Equation (6) gives a local bound for the perturbation induced by \mathbf{e} . The function

$$\lambda_n(x) := \sum_{j=0}^n |b_j(x)|$$

is called the *Lebesgue function* of the interpolation operator, its infinity norm

$$\Lambda_n := \|\lambda_n\|$$

the *Lebesgue constant*. (It turns out that Λ_n is the norm of the interpolation operator on the space of continuous functions.) Equation (6) then yields

$$\|r_n[\mathbf{f} + \mathbf{e}] - r_n[\mathbf{f}]\| \leq \Lambda_n \cdot \|\mathbf{e}\|;$$

r_n being linear, this means $\|r_n[\mathbf{e}]\|/\|\mathbf{e}\| \leq \Lambda_n$; the global relative perturbation of the interpolant is bounded by Λ_n .

Definition 1 We say that Λ_n grows at most logarithmically with n , if

$$\Lambda_n \leq a + b \ln n \tag{7}$$

for some positive constants a and b and every $n > n_0, n_0 \in \mathbf{N}$.

The rational interpolant $r_n[\mathbf{f}]$ obtained by replacing the weights λ_j in (1) by some β_j independent of \mathbf{f} can be written as $\sum_{j=0}^n f_j b_j(x)$ with the Lagrange basis functions

$$b_j(x) := \frac{\beta_j}{x - x_j} \bigg/ \sum_{i=0}^n \frac{\beta_i}{x - x_i},$$

so that [7]

$$\lambda_n(x) = \sum_{j=0}^n \left| \frac{\beta_j}{x - x_j} \right| \bigg/ \left| \sum_{i=0}^n \frac{\beta_i}{x - x_i} \right|. \tag{8}$$

For R_1 this yields

$$\lambda_n(x) = \sum_{i=0}^n \frac{1}{|x - x_i|} \bigg/ D(x),$$

where

$$D(x) := \left| \sum_{i=0}^n \frac{(-1)^i}{x - x_i} \right|.$$

If x is one of the x_j 's, a limit calculation in (8) shows that $\lambda_n(x) = 1$. Otherwise, let x_k be the node at the immediate left of x , i.e., $x \in (x_k, x_{k+1})$. If k is odd and $\leq n - 5$, then

$$D(x) = \dots - \frac{\delta_{k-1}}{x - x_{k-1}} + \frac{1}{x - x_k} + \frac{1}{x_{k+1} - x} - \frac{1}{x_{k+2} - x} + \dots,$$

where δ_j is defined in (4). Therefore we may write the numerator of $\lambda_n(x)$ as $D(x) + 2N(x)$, where

$$N(x) = \dots + \frac{\delta_{k-3}}{x - x_{k-3}} + \frac{\delta_{k-1}}{x - x_{k-1}} + \frac{1}{x_{k+2} - x} + \frac{1}{x_{k+4} - x} + \dots$$

(setting $\delta_j = 0$ for $j < 0$) to obtain

$$\lambda_n(x) = 1 + 2N(x)/D(x).$$

Note that all terms of N are positive. If k is even, one may multiply all terms in D and N by -1 to get the same expression for λ_n .

At first sight, it seems that, in view of the factors $1/2$ arising in the first and last terms, one must study the two extremal intervals $[x_0, x_1]$ and $[x_{n-1}, x_n]$ separately from the others. However, a shift by $\frac{x_0+x_n}{2} =: m$ of the interval $[x_0, x_n]$ together with the abscissae x_j , i.e.,

$$\tilde{x}_j := x_j - m, \quad \tilde{x} := x - m$$

(so that m becomes 0) without change of the weights just shifts $\lambda_n(x)$. Moreover, a reflection $\hat{x}_j := -x_j$ and $\hat{x} := -x$ reflects $\lambda_n(x)$ as well. Also, $\lambda_n(x)$ is unchanged by a multiplication of all weights by -1 . Thus, if there are at least two points on each side of m , we may bound $\lambda_n(x)$ only on $[x_0, m]$: bounds for $x \in (m, x_n]$ are obtained after a reflection with respect to m and a multiplication of the weights by -1 , if n is odd. This shows that it suffices to treat just the case $x \in [x_0, x_1]$ separately, and explains why the δ_j 's appear to the left only in D and N above.

Lemma 1 For $x \in (x_k, x_{k+1})$, $1 \leq k \leq \lfloor \frac{n}{2} \rfloor$,

$$D(x) \geq A_1 + A_2, \quad (9)$$

where

$$A_1 = \frac{x_k - x_{k-1}}{(x_{k+1} - x_{k-1})(x_{k+1} - x_k)}, \quad A_2 = \frac{x_{k+2} - x_{k+1}}{(x_{k+1} - x_k)(x_{k+2} - x_k)},$$

and

$$N(x) \leq B_1 + B_2 \quad (10)$$

where

$$B_1 = \sum'_{i \geq 0} \frac{1}{x_k - x_{k-1-2i}}, \quad B_2 = \sum'_{i \geq 0} \frac{1}{x_{k+2+2i} - x_{k+1}}$$

and the prime means that, if present, the terms $1/(x_k - x_0)$, resp. $1/(x_n - x_{k+1})$, are to be multiplied by $1/2$.

Proof For the denominator,

$$D(x) \geq D_1(x) + D_2(x),$$

where

$$D_1(x) := -\frac{\delta_{k-1}}{x - x_{k-1}} + \frac{1}{x - x_k} \geq \frac{x_k - x_{k-1}}{(x - x_{k-1})(x - x_k)},$$

and

$$D_2(x) := \frac{1}{x_{k+1} - x} - \frac{1}{x_{k+2} - x} = \frac{x_{k+2} - x_{k+1}}{(x_{k+1} - x)(x_{k+2} - x)}.$$

(In D , all pairs of terms corresponding to nodes to the left of x_{k-1} are positive, and so is the term corresponding to x_0 if k is even; the same holds accordingly to the

right of x_{k+1} . When $k = 1$, the factor δ_0 in front of $1/(x - x_0)$ makes D even larger.) Since D_1 and D_2 are decreasing, respectively increasing, for $x \in (x_k, x_{k+1})$,

$$D_1(x) \geq D_1(x_{k+1}) = A_1, \quad D_2(x) \geq D_2(x_k) = A_2.$$

For the numerator,

$$N(x) = N_1(x) + N_2(x),$$

where

$$N_1(x) = \cdots + \frac{\delta_{k-3}}{x - x_{k-3}} + \frac{\delta_{k-1}}{x - x_{k-1}},$$

and

$$N_2(x) = \frac{1}{x_{k+2} - x} + \frac{1}{x_{k+4} - x} + \cdots.$$

Since N_1 and N_2 are decreasing and increasing, respectively, for $x \in (x_k, x_{k+1})$,

$$N_1(x) \leq N_1(x_k) = B_1, \quad N_2(x) \leq N_2(x_{k+1}) = B_2.$$

■

One proves in a similar way the following corresponding results for $k = 0$.

Lemma 2 For $x \in (x_0, x_1)$,

$$D(x) \geq A_1 + A_2, \tag{11}$$

where

$$A_1 = \frac{1/2}{x_1 - x_0}, \quad A_2 = \frac{x_2 - x_1}{(x_1 - x_0)(x_2 - x_0)},$$

and

$$N(x) \leq B_2 \tag{12}$$

where

$$B_2 = \sum'_{i \geq 0} \frac{1}{x_{2i+2} - x_1}.$$

2.1 Equidistant Nodes

For uniformly spaced points $x_k = a + kh$, $h = (b - a)/n$, the lemmas imply that $D(x) \geq 1/h$ for every $x \in [a, b]$. Moreover,

$$B_1 \leq \frac{1}{h} \left(1 + \frac{1}{3} + \frac{1}{5} + \cdots + \frac{1}{\ell} \right), \quad \ell = \begin{cases} k - 1, & k \text{ even,} \\ k, & k \text{ odd,} \end{cases}$$

and

$$B_2 \leq \frac{1}{h} \left(1 + \frac{1}{3} + \frac{1}{5} + \cdots + \frac{1}{\ell} \right), \quad \ell = \begin{cases} n - k - 1, & n - k \text{ even,} \\ n - k - 2, & n - k \text{ odd,} \end{cases}$$

so that $N(x) < \frac{1}{h}L$ with

$$L := \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{k-1} + \frac{1}{k} \right) + \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n-k-1} \right). \quad (13)$$

As in similar proofs, we now use the fact that $\ln(2j+1)$ is an upper bound of the j th sum of the harmonic series, so that

$$L \leq \ln(2k+1) + \ln(2(n-k-1)+1),$$

which is maximized over real k in $[0, n-1]$ by $k = (n-1)/2$, implying that

$$L \leq 2 \ln(n), \quad (14)$$

and therefore,

$$\lambda_n(x) \leq 1 + 4 \ln(n).$$

Theorem 1 *For uniformly spaced nodes, the Lebesgue constant of the interpolant R_1 grows at most logarithmically with n .*

This result is already known as the case $d = 1$ of [8]. However, the above proof of the bound for D does not make use of the original formula for the Floater–Hormann interpolant, merely of the barycentric formula.

We have confirmed this logarithmic growth numerically. The second column of Table 1 displays Λ_n for doubling numbers of interpolation nodes. To check the constant b in the logarithmic growth, we have also computed after every pair $(n, 2n)$ an approximation \widehat{a} and \widehat{b} in the logarithmic growth formula by solving the system of equations

$$\begin{aligned} \widehat{a} + \widehat{b} \ln(2n+2) &= \Lambda_{2n+2} \\ \widehat{a} + \widehat{b} \ln(n+1) &= \Lambda_{n+1}. \end{aligned}$$

The corresponding values of \widehat{b} are given in the third column of the table. Compared with the number $2/\pi = 0.6366197723675814$, they are in line with the main term of the asymptotic formula for the Lebesgue constant of R_0 given in [15, 23]. (All the computations described in this article were performed in MATLAB on MacBook Pro computers.)

Table 1 The Lebesgue constant Λ_n for equidistant and shifted equidistant nodes. For each of the two we give the constant itself as well as the empirical value \widehat{b} of the factor of $\ln(n)$

| $n + 1$ | Equidistant | | Shifted equidistant | |
|---------|---------------|-----------------|---------------------|--------------------|
| | Λ_n^E | \widehat{b}^E | Λ_n^{SE} | \widehat{b}^{SE} |
| 10 | 2.2220 | | 2.2989 | |
| 20 | 2.6839 | 0.66644 | 2.7574 | 0.66136 |
| 40 | 3.1414 | 0.65992 | 3.2185 | 0.66531 |
| 80 | 3.5904 | 0.64781 | 3.6687 | 0.64944 |
| 160 | 4.0357 | 0.64244 | 4.1142 | 0.64272 |
| 320 | 4.4789 | 0.63947 | 4.5575 | 0.63960 |
| 640 | 4.9212 | 0.63809 | 4.9998 | 0.63808 |
| 1280 | 5.3630 | 0.63735 | 5.4416 | 0.63735 |
| 2560 | 5.8045 | 0.63698 | 5.8831 | 0.63698 |
| 5120 | 6.2459 | 0.63680 | 6.3245 | 0.63680 |
| 10240 | 6.6873 | 0.63671 | 6.7658 | 0.63671 |

2.2 Well-Spaced Nodes

The article [9] introduces the following property of interpolation points.

Definition 2 For each $n \in \mathbf{N}$, let $X_n := \{x_k\}_{k=0}^n$ be a set of interpolation nodes satisfying property (2) and let

$$h_k := x_{k+1} - x_k, \quad k = 0, \dots, n - 1.$$

An array $X = (X_n)_{n \in \mathbf{N}}$ of such sets is called *well-spaced* if there exist two constants K and R , $R \geq 1$, both independent of n , such that the three conditions

$$\frac{x_k - x_{k-1}}{x_k - x_j} \leq \frac{K}{k - j}, \quad k = 1, \dots, n, \quad j = 0, \dots, k - 1, \quad (15)$$

$$\frac{x_{k+1} - x_k}{x_j - x_k} \leq \frac{K}{j - k}, \quad k = 0, \dots, n - 1, \quad j = k + 1, \dots, n, \quad (16)$$

$$\frac{h_k}{h_{k-1}} \leq R, \quad \frac{h_{k-1}}{h_k} \leq R, \quad k = 1, \dots, n - 1, \quad (17)$$

hold for every set of nodes X_n .

We shall indifferently call the array and the nodes well-spaced. Note that the property of well-spacing is invariant with respect to translation or spacing of the nodes.

The authors of [9] established that every such array leads to a logarithmic growth of the Lebesgue constant of R_0 :

Theorem 2 *If $X = (X_n)_{n \in \mathbf{N}}$ is an array of well-spaced nodes, then the Lebesgue constant of the interpolant R_0 grows at most logarithmically with n .*

We shall now use the above lemmas to show the same for R_1 . Let first $k \geq 1$. Then, according to (17),

$$A_1 = \frac{h_{k-1}}{(x_{k+1} - x_{k-1})h_k} \geq \frac{1}{x_{k+1} - x_{k-1}} \cdot \frac{1}{R}$$

and

$$A_2 = \frac{h_{k+1}}{h_k(x_{k+2} - x_k)} \geq \frac{1}{R} \cdot \frac{1}{x_{k+2} - x_k}$$

so that

$$A_1 + A_2 \geq \frac{1}{R} \left(\frac{1}{x_{k+2} - x_k} + \frac{1}{x_{k+1} - x_{k-1}} \right). \quad (18)$$

Moreover, $x_{k+2} - x_k = h_k + h_{k+1} \leq R(h_{k-1} + h_k) = R(x_{k+1} - x_{k-1})$ and $x_{k+1} - x_{k-1} = h_{k-1} + h_k \leq (R+1)h_k$, so that

$$A_1 + A_2 \geq \frac{1}{R} \left(\frac{1}{R} + 1 \right) \frac{1}{x_{k+1} - x_{k-1}} \geq \frac{R+1}{R^2} \frac{1}{(R+1)h_k} = \frac{1}{R^2 h_k}. \quad (19)$$

If $k = 0$, we have in Lemma 2

$$A_1 + A_2 \geq A_1 = \frac{1}{2h_0}.$$

On the whole, we therefore have for $x \in (x_k, x_{k+1})$

$$D(x) \geq \frac{1}{Sh_k}$$

with

$$S := \begin{cases} 1, & k = 0, \\ R^2, & k > 0. \end{cases}$$

In the upper bound for N , possible factors $1/2$ may be ignored. We compute for $k \neq 0$

$$\begin{aligned} B_1 &= \frac{1}{x_k - x_{k-1}} + \frac{1}{x_k - x_{k-3}} + \frac{1}{x_k - x_{k-5}} + \frac{1}{x_k - x_{k-7}} + \dots \\ &= \frac{1}{x_k - x_{k-1}} \left(1 + \frac{x_k - x_{k-1}}{x_k - x_{k-3}} + \frac{x_k - x_{k-1}}{x_k - x_{k-5}} + \frac{x_k - x_{k-1}}{x_k - x_{k-7}} + \dots \right). \end{aligned}$$

By (15), this may be bounded as

$$B_1 \leq \frac{1}{h_{k-1}} \left(1 + \frac{K}{3} + \frac{K}{5} + \frac{K}{7} + \cdots \right).$$

Similarly for B_2 :

$$\begin{aligned} B_2 &= \frac{1}{2} \frac{1}{x_{k+2} - x_{k+1}} + \frac{1}{x_{k+4} - x_{k+1}} + \frac{1}{x_{k+6} - x_{k+1}} + \frac{1}{x_{k+8} - x_{k+1}} + \cdots \\ &= \frac{1}{x_{k+2} - x_{k+1}} \left(\frac{1}{2} + \frac{x_{k+2} - x_{k+1}}{x_{k+4} - x_{k+1}} + \frac{x_{k+2} - x_{k+1}}{x_{k+6} - x_{k+1}} + \frac{x_{k+2} - x_{k+1}}{x_{k+8} - x_{k+1}} + \cdots \right), \end{aligned}$$

which by (16) may be majorized as

$$B_2 \leq \frac{1}{h_{k+1}} \left(1 + \frac{K}{3} + \frac{K}{5} + \frac{K}{7} + \cdots \right)$$

with the same number of terms as in Sect. 2.1, so that with (13)

$$N(x) \leq \left(\frac{1}{h_{k-1}} + \frac{1}{h_{k+1}} \right) KL$$

and thus with (19)

$$\begin{aligned} \lambda_n(x) &\leq 1 + 2R^2 \left(\frac{h_k}{h_{k-1}} + \frac{h_k}{h_{k+1}} \right) 2K \ln(n) \\ &\leq 1 + 8KR^3 \ln(n). \end{aligned}$$

If $k = 0$, $B_1 = 0$ and $B_2 \leq \frac{1}{h_1} \left(1 + \frac{K}{3} + \frac{K}{5} + \cdots \right)$ so that

$$N(x) \leq \frac{2}{h_1} K \ln(n)$$

and

$$\lambda_n(x) \leq 1 + \frac{h_0}{h_1} 4K \ln(n) \leq 1 + 4KR \ln(n).$$

This establishes our first main theorem.

Theorem 3 *If $X = (X_n)_{n \in \mathbf{N}}$ is an array of well-spaced nodes, then the Lebesgue constant of the interpolant R_1 grows at most logarithmically with n .*

2.3 Conformally Shifted Nodes

The complexity of the pseudospectral solution of a boundary value problem can often be drastically reduced by a move of the points, for instance to improve the conditioning of the system of equations to be solved, or to have more points close to a steep gradient (front).

In [16], Kosloff and Tal-Ezer suggested applying such point shifts to loosen the step restriction in the solution of the systems of ordinary differential equations arising from the pseudospectral method of lines applied to time evolution problems. The idea is to introduce, besides the interval I on which the function is to be interpolated, another interval J containing the standard points (such as equidistant or Chebyshev). Then, to obtain a particular set of points in I , a (one-to-one) conformal map of a domain containing J , say g , is used, which applies J onto I , thereby determining the desired shifted points. Since g is infinitely differentiable, the composite function $f \circ g$ inherits the differentiability properties of f .

Recall that R_1 coincides with the interpolating polynomial, when the nodes are Chebyshev points of the second kind, so that it converges exponentially to interpolated functions which are analytic. Recall also that R_1 retains the exponential convergence for the same class of functions, when the Chebyshev points of the second kind are shifted conformally [2]. This example of use of R_1 is likely the most important, as it greatly improves the accuracy of pseudospectral solutions of boundary value ordinary and partial differential equations whose solution displays large gradients: see the applications in explosion modeling [18], fluid infiltration [10] and finance [17]. We shall now see how the results of Bos et al. [9] and the above theorems imply the logarithmic growth of the Lebesgue constant for such nodes.

Bos et al. work on the interval $[0, 1]$. This is indeed general, as one immediately sees that a shift and a contraction/dilation of the interval of interpolation do not modify the Lebesgue function (8) and the Lebesgue constant is thus unaffected by such operations. These authors give the following definition of a (regular) distribution function of nodes.

Definition 3 A function $F \in C[0, 1]$ is a *distribution function*, if it is a strictly increasing bijection of the interval $[0, 1]$. F is *regular*, if $F \in C^1[0, 1]$ and F' has a finite number of zeros in $[0, 1]$ with finite multiplicities.

Given a distribution function F and some $n \in \mathbf{N}$, the authors define the associated interpolation nodes $X_n = X_n(F)$ as the set of x_k with

$$x_k := F(k/n), \quad k = 0, 1, \dots, n. \quad (20)$$

They then prove the following result.

Theorem 4 *If F is a regular distribution function and X_n are the associated interpolation nodes from (20) for every $n \in \mathbf{N}$, then the array of nodes $X = (X_n)_{n \in \mathbf{N}}$ is well-spaced.*

Since a conformal map g cannot have any zero derivative in its domain [1, p. 133], Theorem 2 automatically yields the following results.

Corollary 1 *The Lebesgue constants of R_0 as well as of R_1 with conformally shifted equidistant points grow at most logarithmically with n .*

Proof For R_0 , the result follows immediately from Theorem 3.5 in [9] and Theorem 2 above, for R_1 from Theorem 3. ■

Among their examples of regular distributions, Bos and al. give the extrema of the Chebyshev polynomials, which are the Chebyshev points of the second kind.

Lemma 3 *A composition of regular distribution functions is itself a regular distribution function.*

Proof This follows directly from the chain rule. ■

Corollary 2 *The Lebesgue constant of R_1 with conformally shifted Chebyshev points of the second kind grows at most logarithmically with n .*

Proof Let $g(y)$ be the conformal map of the interval containing the Chebyshev points to the interpolation interval. g being conformal, its derivative does not have any zeros, so that g is a regular distribution function on J . Then $g \circ F$ is a composition of regular distribution functions, thus a regular distribution function as well, according to Lemma 3. The corresponding array of nodes is well-spaced according to Theorem 4 and the Lebesgue constant of R_1 with a well-spaced array grows at most logarithmically (Theorem 3). ■

We have numerically verified this logarithmic behavior of the Lebesgue constant with conformally shifted points. To increase the efficiency of methods for solving differential problems, it is natural to move more points toward the location of a front. A simple map for the case where the front lies in the center of the interval $[-1, 1]$ is that suggested by Kosloff and Tal-Ezer for another purpose and mentioned above,

$$x = g_\alpha(y) = \frac{\arcsin(\alpha y)}{\arcsin(\alpha)}, \quad 0 < \alpha < 1. \quad (21)$$

See the improvement in accuracy in [6]. In the limit $\alpha \rightarrow 0$, the nodes remain the equidistant ones. In the limit $\alpha \rightarrow 1$, the singularities at $\pm 1/\alpha$ of g_α tend toward the extremities of the interval.

First we have repeated the experiment documented in columns 2 and 3 of Table 1 for *equidistant points* shifted with the map (21) and $\alpha = 0.9$. The results are in columns 4 and 5: they demonstrate that the constant b in (7) seems quite independent of the shift.

Then we have repeated the experiment of Table 1, but now with Chebyshev points of the second kind and shifted Chebyshev points of the second kind. The results, given in Table 2, seem to show that the main term of the Lebesgue constant does not change much with a conformal shift of Chebyshev nodes, and is very close to the one corresponding to equidistant points.

Table 2 The Lebesgue constant Λ_n for Chebyshev points of the second kind, and the same shifted. The values of \widehat{b} have the same meaning as in Table 1

| $n + 1$ | Chebyshev 2nd kind | | Shifted Chebyshev | |
|---------|--------------------|-----------------|-------------------|--------------------|
| | Λ_n^C | \widehat{b}^C | Λ_n^{SC} | \widehat{b}^{SC} |
| 10 | 2.3619 | | 2.5474 | |
| 20 | 2.8371 | 0.68568 | 3.0219 | 0.68452 |
| 40 | 3.2948 | 0.66034 | 3.4794 | 0.66008 |
| 80 | 3.7442 | 0.64829 | 3.9287 | 0.64823 |
| 160 | 4.1895 | 0.64241 | 4.3740 | 0.64239 |
| 320 | 4.6328 | 0.63950 | 4.8173 | 0.63950 |
| 640 | 5.0750 | 0.63806 | 5.2595 | 0.63806 |
| 1280 | 5.5168 | 0.63734 | 5.7013 | 0.63735 |
| 2560 | 5.9583 | 0.63698 | 6.1428 | 0.63698 |
| 5120 | 6.3997 | 0.63680 | 6.5842 | 0.63680 |
| 10240 | 6.8410 | 0.63671 | 7.0256 | 0.63671 |

3 Conclusion

This article has been devoted to the study of the behaviour, as the number of nodes increases, of the second ‘‘Berrut’’ interpolant introduced some decades ago. We have shown that, at least in the instances of practical interest, the interpolant enjoys the same perfect condition of its special case of the usual Chebyshev interpolating polynomial, and likewise of its first-order variant R_0 , with a Lebesgue constant growing at most logarithmically. This is true in particular of the important case of the shifted Chebyshev nodes used in practical applications, and justifies their use with arbitrary large numbers of nodes.

Acknowledgements The author wishes to thank Michael Floater for giving him his proof of the logarithmic growth of the Lebesgue constant of R_0 , on which Theorem 1 above is patterned. He is also grateful to Nick Trefethen for his suggestions to the text.

References

- Ahlfors, L.: Complex Analysis, 2nd edn. McGraw-Hill, New York (1966)
- Baltensperger, R., Berrut, J.-P., Noël, B.: Exponential convergence of a linear rational interpolant between transformed Chebyshev points. *Math. Comp.* **68**, 1109–1120 (1999)
- Berrut, J.-P.: Rational functions for guaranteed and experimentally well-conditioned global interpolation. *Comput. Math. Appl.* **15**, 1–16 (1988)
- Berrut, J.-P.: Linear rational interpolation of continuous functions over an interval. In: Gautschi, W. (ed.) *Mathematics of Computation 1943–1993: A Half-Century of Computational Mathematics*. Proceedings of Symposia in Applied Mathematics, pp. 261–264. American Mathematical Society, Providence (1994)

5. Berrut, J.-P.: Properties of a barycentric rational interpolant. *Rend. Sem. Mat. Univ. Politec. Torino* **76**, 39–45 (2018)
6. Berrut, J.-P., Baltensperger, R.: The linear rational pseudospectral method for boundary value problems. *BIT* **41**, 868–879 (2001)
7. Berrut, J.-P., Mittelmann, H.D.: Lebesgue constant minimizing linear rational interpolation of continuous functions over the interval. *Comput. Math. Appl.* **33**, 77–86 (1997)
8. Bos, L., De Marchi, S., Hormann, K., Klein, G.: On the Lebesgue constant of barycentric rational interpolation at equidistant nodes. *Numer. Math.* **121**, 461–471 (2012)
9. Bos, L., De Marchi, S., Hormann, K., Sidon, J.: Bounding the Lebesgue constant for Berrut’s rational interpolant at general nodes. *J. Approx. Theory* **169**, 7–22 (2013)
10. Cueto-Felgueroso, L., Juanes, R.: Adaptive rational spectral methods for the linear stability analysis of nonlinear fourth-order problems. *J. Comput. Phys.* **228**, 6536–6552 (2009)
11. Floater, M.S., Hormann, K.: Barycentric rational interpolation with no poles and high rates of approximation. *Numer. Math.* **107**, 315–331 (2007)
12. Gautschi, W.: Interpolation before and after Lagrange. *Rend. Semin. Mat. Univ. Politec. Torino* **70**, 347–368 (2012)
13. Henrici, P.: *Essentials of Numerical Analysis with Pocket Calculator Demonstrations*. Wiley, New York (1982)
14. Hormann, K.: Barycentric interpolation. In: Fasshauer, G.E., Schumaker, L.L. (eds.) *Approximation Theory XIV: San Antonio 2013*. Springer Proceedings in Mathematics & Statistics, vol. 83, pp. 197–218. Springer, New York (2014)
15. Ibrahimoglu, B.A., Cuyt, A.: Sharp bounds for Lebesgue constants of barycentric rational interpolation at equidistant points. *Exp. Math.* **25**, 347–354 (2016)
16. Kosloff, D., Tal-Ezer, H.: A modified Chebyshev pseudospectral method with an $\mathcal{O}(N^{-1})$ time step restriction. *J. Comput. Phys.* **104**, 457–469 (1993)
17. Pindza, E., Patidar, K., Ngounda, E.: Rational spectral collocation method for pricing American vanilla and butterfly spread options. In: Dimov, I., Faragó, I., Vulkov, L. (eds.) *Finite Difference Methods, Theory and Applications*. Lecture Notes in Computer Science, vol. 9045, pp. 323–331. Springer, Cham (2015)
18. Tee, T.W., Trefethen, L.N.: A rational spectral collocation method with adaptively transformed Chebyshev grid points. *SIAM J. Sci. Comput.* **28**, 1798–1811 (2006)
19. Trefethen, L.N.: *Spectral Methods in MATLAB*. SIAM, Philadelphia (2000)
20. Trefethen, L.N.: *Approximation Theory and Approximation Practice*. SIAM, Philadelphia (2013)
21. Vértesi, P.: On barycentric interpolation. I. (On the T-Lebesgue function and T-Lebesgue constant). *Acta Math. Hungar.* **147**, 396–407 (2015)
22. Werner, W.: Polynomial interpolation: Lagrange versus Newton. *Math. Comp.* **43**, 205–217 (1984)
23. Zhang, R.-J.: Optimal asymptotic Lebesgue constant of Berrut’s rational interpolation operator for equidistant nodes. *Appl. Math. Comput.* **294**, 139–145 (2017)

Learning Low-Dimensional Dynamical-System Models from Noisy Frequency-Response Data with Loewner Rational Interpolation



Zlatko Drmač and Benjamin Peherstorfer

Dedicated to Athanasios C Antoulas on the occasion of his 70th birthday.

Abstract Loewner rational interpolation provides a versatile tool to learn low-dimensional dynamical-system models from frequency-response measurements. This work investigates the robustness of the Loewner approach to noise. The key finding is that if the measurements are polluted with Gaussian noise, then the error due to noise grows at most linearly with the standard deviation with high probability under certain conditions. The analysis gives insights into making the Loewner approach robust against noise via linear transformations and judicious selections of measurements. Numerical results demonstrate the linear growth of the error on benchmark examples.

Keywords Model reduction · Dynamical systems · Concentration inequalities · System identification

1 Introduction

Learning dynamical-system models from measurements is a widely studied task in science, engineering, and machine learning; see, e.g., system identification originating from the systems & control community [1, 10, 11, 15, 22, 27, 29, 38–40, 51],

Z. Drmač
Faculty of Science, Department of Mathematics, University of Zagreb,
Bijenička 30, 10000 Zagreb, Croatia

B. Peherstorfer (✉)
Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, USA
e-mail: pehersto@cims.nyu.edu

sparsity-promoting methods [8, 42–44], dynamic mode decomposition [9, 24, 41, 45, 46, 50], and operator inference [23, 30, 32–34, 37, 49]. Antoulas and collaborators introduced the Loewner approach [2, 5, 25, 28] that constructs models directly from frequency-response measurements, without requiring computationally expensive training phases and without solving potentially non-convex optimization problems. This work investigates the robustness of the Loewner approach to noise in the frequency-response measurements. The main finding of this work is that under certain conditions and with high probability the error introduced by noise into Loewner models grows at most linearly with the standard deviation of the noise.

The Loewner approach [2, 5, 25, 28] derives dynamical-system models from frequency-response data, i.e., from values of the transfer function of the high-dimensional dynamical system of interest. A series of works have extended the Loewner approach from linear time-invariant systems to bilinear systems [4], quadratic-bilinear systems [13], parametrized systems [18], time-delay systems [47], and structured systems [48]. Learning Loewner models from time-domain data, instead of frequency-response measurements, is discussed in [20, 31]. Learning Loewner models from noisy data has received relatively little attention. The work [26] provides a numerical investigation of the effect of noise on the accuracy of Loewner models. The work [17] discusses the rank of Hankel matrices if measurements are polluted by additive Gaussian noise. Numerical experiments in the thesis [19, Sect. 2.1] demonstrate that the selection of frequencies at which to obtain measurements and how to partition the measurements has a significant influence on the robustness against perturbations in the data such as noise. The work [21] applies the Loewner approach to control tasks where only noisy measurements are available. The authors of [6] discuss the robustness of interpolatory model reduction against perturbations in evaluations of the transfer function of the high-dimensional system; however, the perturbations are considered deterministic and stem from, e.g., numerical approximations via iterative methods. During the writing of this manuscript, the authors became aware of the work [12] that studies the sensitivity of Loewner interpolation to perturbations in a deterministic fashion via pseudospectra of the Loewner matrix pencils.

In this work, the robustness of the Loewner approach to Gaussian noise is considered. Rather than a deterministic analysis, the contribution of this work is an analysis that bounds the error introduced by noise in a probabilistic sense. In particular, the analysis shows that the error grows at most linearly in the standard deviation of the noise under certain conditions and with high probability. A relative noise model is considered, which seems realistic because measurement errors typically are relative to the value of the measured quantities. The conditions under which the proposed error bounds hold give insights into selecting the frequencies at which to measure and partitioning the data to reduce the effect of noise. The linear growth of the error with respect to the standard deviation of the noise is observed in numerical examples.

2 Learning Low-Dimensional Dynamical-System Models with Loewner Rational Interpolation

This section recapitulates model reduction with the Loewner approach; see, e.g., [3, 7, 35] for general introductions to model reduction and related concepts.

2.1 Linear Time-Invariant Dynamical Systems

Consider the linear time-invariant system

$$E\dot{x}(t) = Ax(t) + Bu(t), \quad y(t) = Cx(t) \quad (1)$$

of order $N \in \mathbb{N}$ with system matrices $E, A \in \mathbb{R}^{N \times N}$, $B \in \mathbb{R}^{N \times 1}$, and $C \in \mathbb{R}^{1 \times N}$. The state at time t is $x(t) \in \mathbb{R}^N$ and the output at time t is $y(t) \in \mathbb{R}$. The transfer function is

$$H(s) = C(sE - A)^{-1}B, \quad s \in \mathbb{C}.$$

In the following, we only consider systems with full-rank matrix E .

2.2 Loewner Rational Interpolation

To derive a reduced model of dimension $n \in \mathbb{N}$ of system (1), consider $2n$ interpolation points $s_1, \dots, s_{2n} \in \mathbb{C}$. The interpolation points are partitioned into two sets $\{\mu_1, \dots, \mu_n\}$ and $\{\gamma_1, \dots, \gamma_n\}$ of equal size. Define the $n \times n$ Loewner matrix L and the $n \times n$ shifted Loewner matrix $L^{(s)}$ as

$$L_{ij} = \frac{H(\mu_i) - H(\gamma_j)}{\mu_i - \gamma_j}, \quad L_{ij}^{(s)} = \frac{\mu_i H(\mu_i) - \gamma_j H(\gamma_j)}{\mu_i - \gamma_j}, \quad i, j = 1, \dots, n,$$

together with the $n \times 1$ input matrix \widehat{B} and the $1 \times n$ output matrix \widehat{C} with components

$$\widehat{B}_i = H(\mu_i), \quad \widehat{C}_i = H(\gamma_i), \quad i = 1, \dots, n. \quad (2)$$

The Loewner model is

$$\widehat{E}\dot{\hat{x}}(t) = \widehat{A}\hat{x}(t) + \widehat{B}u(t), \quad \hat{y}(t) = \widehat{C}\hat{x}(t)$$

with $\widehat{E} = -L$ and $\widehat{A} = -L^{(s)}$. The n -dimensional state at time t is $\hat{x}(t)$ and the output at time t is $\hat{y}(t)$. The transfer function of the Loewner model is

$$\widehat{H}(s) = \widehat{C}(s\widehat{E} - \widehat{A})^{-1}\widehat{B}, \quad s \in \mathbb{C}.$$

The Loewner approach guarantees that the transfer function \widehat{H} of the Loewner model interpolates the transfer function H of the system (1) at the interpolation points s_1, \dots, s_{2n} , which means

$$H(s_i) = \widehat{H}(s_i), \quad i = 1, \dots, 2n.$$

3 Learning Loewner Models from Noisy Frequency-Response Measurements

We now study the robustness of the Loewner approach to noise in the transfer-function values of the system (1). The key contribution is Theorem 1 that bounds the error that is introduced by noise under certain conditions.

3.1 Noisy Transfer-Function Values

Let $\mu \in \mathbb{C}$ with the real part $\Re(\mu)$ and the imaginary part $\Im(\mu)$ and let $0 < \sigma \in \mathbb{R}$. We denote with $\epsilon \sim \mathcal{CN}(\mu, \sigma)$ a complex random variable, where the real $\Re(\epsilon)$ and the imaginary part $\Im(\epsilon)$ are independently normally distributed. The real part $\Re(\epsilon)$ has mean $\Re(\mu)$, the imaginary part $\Im(\epsilon)$ has mean $\Im(\mu)$. The real and the imaginary part of ϵ have standard deviation σ .

Let $\epsilon_1, \dots, \epsilon_n \sim \mathcal{CN}(0, 1)$ and $\eta_1, \dots, \eta_n \sim \mathcal{CN}(0, 1)$ be independent random variables. Define the noisy transfer-function values as

$$H_\sigma(\mu_i) = H(\mu_i)(1 + \sigma\epsilon_i), \quad H_\sigma(\gamma_i) = H(\gamma_i)(1 + \sigma\eta_i), \quad i = 1, \dots, n,$$

so that

$$H_\sigma(\mu_i) \sim \mathcal{CN}(H(\mu_i), H(\mu_i)\sigma), \quad H_\sigma(\gamma_i) \sim \mathcal{CN}(H(\gamma_i), H(\gamma_i)\sigma), \quad i = 1, \dots, n.$$

The noise pollutes the transfer-function values in a relative sense, i.e., the standard deviation of $H_\sigma(\mu_i)$ is scaled by $H(\mu_i)$. We consider such a relative noise model to be realistic in our situation because measurement errors typically are relative to the value of the quantity that is measured. Note, however, that the techniques used in the following analysis can be extended to an absolute noise model, where the standard deviation of the noise is independent of the transfer-function value; see Sect. 3.5.

3.2 Loewner and Noisy Transfer-Function Values

From the noisy transfer-function values, we derive the noisy Loewner matrices

$$\tilde{L}_{ij} = \frac{H_\sigma(\mu_i) - H_\sigma(\gamma_j)}{\mu_i - \gamma_j}, \quad \tilde{L}_{ij}^{(s)} = \frac{\mu_i H_\sigma(\mu_i) - \gamma_j H_\sigma(\gamma_j)}{\mu_i - \gamma_j}, \quad i, j = 1, \dots, n,$$

which have the same structure as in Sect. 2.2, except that the noisy transfer-function values $H_\sigma(\mu_1), \dots, H_\sigma(\mu_n), H_\sigma(\gamma_1), \dots, H_\sigma(\gamma_n)$ are used rather than the noiseless values $H(\mu_1), \dots, H(\mu_n), H(\gamma_1), \dots, H(\gamma_n)$. We decompose the noisy Loewner and the noisy shifted Loewner matrix into deterministic and random parts as

$$\tilde{L} = L + \sigma \delta L, \quad \tilde{L}^{(s)} = L^{(s)} + \sigma \delta L^{(s)},$$

with

$$\delta L_{ij} = \frac{H(\mu_i)\epsilon_i - H(\gamma_j)\eta_j}{\mu_i - \gamma_j}, \quad \delta L_{ij}^{(s)} = \frac{\mu_i H(\mu_i)\epsilon_i - \gamma_j H(\gamma_j)\eta_j}{\mu_i - \gamma_j}, \quad i, j = 1, \dots, n.$$

We obtain the system matrices

$$\tilde{E} = \hat{E} + \sigma \delta E, \quad \tilde{A} = \hat{A} + \sigma \delta A,$$

where $\delta E = -\delta L_{ij}$ and $\delta A = -\delta L_{ij}^{(s)}$. Similarly, we define $\tilde{B} = \hat{B} + \sigma \delta B$ and $\tilde{C} = \hat{C} + \sigma \delta C$ with

$$\delta B_i = H(\mu_i)\epsilon_i, \quad \delta C_i = H(\gamma_i)\eta_i, \quad i = 1, \dots, n. \quad (3)$$

The Loewner model learned from the noisy transfer-function values is then given by

$$\tilde{E}\dot{\tilde{x}}(t) = \tilde{A}\tilde{x}(t) + \tilde{B}u(t), \quad \tilde{y}(t) = \tilde{C}\tilde{x}(t), \quad (4)$$

with the n -dimensional state $\tilde{x}(t)$ at time t and the output $\tilde{y}(t)$ at time t . The transfer function of the model (4) is

$$\tilde{H}(s) = \tilde{C}(s\tilde{E} - \tilde{A})^{-1}\tilde{B}, \quad s \in \mathbb{C}.$$

We call (4) the noisy Loewner model.

3.3 Noise Structure

The Loewner and the shifted Loewner matrices introduce a structure in the noise that is added to the system matrices of the Loewner model learned from the noisy transfer-function values. Consider the matrix $s\delta E - \delta A$ and obtain

$$\begin{aligned}
s\delta E_{ij} - \delta A_{ij} &= -\frac{1}{\mu_i - \gamma_j} \left(sH(\mu_i)\epsilon_i - sH(\gamma_j)\eta_j - \mu_i H(\mu_i)\epsilon_i + \gamma_j H(\gamma_j)\eta_j \right) \\
&= \frac{1}{\gamma_j - \mu_i} \left((s - \mu_i)H(\mu_i)\epsilon_i + (-s + \gamma_j)H(\gamma_j)\eta_j \right)
\end{aligned}$$

to write in matrix form as

$$\begin{aligned}
s\delta E - \delta A &= \underbrace{\begin{bmatrix} \epsilon_1 & & & \\ & \epsilon_2 & & \\ & & \ddots & \\ & & & \epsilon_n \end{bmatrix}}_{\epsilon} + \underbrace{\begin{bmatrix} H(\mu_1) \frac{s-\mu_1}{\gamma_1-\mu_1} & \dots & H(\mu_1) \frac{s-\mu_1}{\gamma_1-\mu_n} \\ \vdots & \ddots & \vdots \\ H(\mu_n) \frac{s-\mu_n}{\gamma_1-\mu_1} & \dots & H(\mu_n) \frac{s-\mu_n}{\gamma_1-\mu_n} \end{bmatrix}}_{F_E} + \\
&\quad \underbrace{\begin{bmatrix} H(\gamma_1) \frac{\gamma_1-s}{\gamma_1-\mu_1} & \dots & H(\gamma_n) \frac{\gamma_1-s}{\gamma_1-\mu_n} \\ \vdots & \ddots & \vdots \\ H(\gamma_1) \frac{\gamma_1-s}{\gamma_1-\mu_1} & \dots & H(\gamma_n) \frac{\gamma_1-s}{\gamma_1-\mu_n} \end{bmatrix}}_{F_A} \underbrace{\begin{bmatrix} \eta_1 & & & \\ & \eta_2 & & \\ & & \ddots & \\ & & & \eta_n \end{bmatrix}}_{\eta}. \tag{5}
\end{aligned}$$

Equation (5) reveals that the random parts of $s\delta E - \delta A$ can be singled out into the two diagonal random matrices ϵ and η of dimension $n \times n$. The diagonal entries of ϵ and η are independent and have distribution $\mathcal{CN}(0, 1)$.

3.4 Bounding the Error Due to Noisy Transfer-Function Values

The task is to bound

$$\widehat{H}(s) - \widetilde{H}(s) = \widehat{C}(s\widehat{E} - \widehat{A})^{-1}\widehat{B} - \widetilde{C}(s\widetilde{E} - \widetilde{A})^{-1}\widetilde{B}, \tag{6}$$

where s typically takes the values in some specified $\Omega \subset \mathbb{C}$; for instance on the imaginary axis, possibly within a specified frequency range. In any case, we assume in the following that Ω is free of the poles of \widehat{H} . The following main results hold pointwise and therefore are agnostic to additional structure imposed by Ω besides that it excludes poles of \widehat{H} .

If $\widehat{\lambda}_i, \widehat{\phi}_i$ ($\widetilde{\lambda}_i, \widetilde{\phi}_i$) are the poles with the corresponding residues of \widehat{H} (\widetilde{H}), then [14, Proposition 3.3] gives

$$\|\widehat{H} - \widetilde{H}\|_{\mathcal{H}_2}^2 = \sum_{i=1}^n \widehat{\phi}_i (\widehat{H}(-\widehat{\lambda}_i) - \widetilde{H}(-\widehat{\lambda}_i)) + \sum_{j=1}^n \widetilde{\phi}_j (\widetilde{H}(-\widetilde{\lambda}_j) - \widehat{H}(-\widetilde{\lambda}_j)), \tag{7}$$

which shows that the error (6) is of particular interest at the reflected poles of both systems (for (7) to hold, both \widehat{H} and \widetilde{H} are assumed stable, of the same order n , and all poles of both systems are assumed simple). We do not tackle the issue of a probabilistic error bound for the \mathcal{H}_2 norm; this topic is left for our future work.

For an estimate of the error (6), we need to understand the effect of the random noise in the matrices $\widetilde{E} = \widehat{E} + \sigma \delta E$, $\widetilde{A} = \widehat{A} + \sigma \delta A$, $\widetilde{B} = \widehat{B} + \sigma \delta B$ to the solution of the linear system $(s\widetilde{E} - \widetilde{A})^{-1}\widetilde{B}$. To that end, we first briefly review the (deterministic) error bound for a solution of the perturbed system. The key is the condition number $\kappa_2(s\widehat{E} - \widehat{A}) = \|(s\widehat{E} - \widehat{A})^{-1}\|_2 \|s\widehat{E} - \widehat{A}\|_2$, see, e.g., [16, Theorem 7.2].

Proposition 1 *Let s be different from the poles of \widehat{H} , and consider the perturbations (5) and (3) deterministic and bounded as $\|\sigma(s\delta E - \delta A)\|_2 \leq \zeta \|s\widehat{E} - \widehat{A}\|_2$, $\|\sigma\delta B\|_2 \leq \zeta \|\widehat{B}\|_2$, where $\zeta > 0$ is such that $\zeta\kappa_2(s\widehat{E} - \widehat{A}) < 1$. Then $s\widetilde{E} - \widetilde{A}$ is nonsingular and*

$$\frac{\|(s\widetilde{E} - \widetilde{A})^{-1}\widetilde{B} - (s\widehat{E} - \widehat{A})^{-1}\widehat{B}\|_2}{\|(s\widehat{E} - \widehat{A})^{-1}\widehat{B}\|_2} \leq \frac{2\zeta}{1 - \zeta\kappa_2(s\widehat{E} - \widehat{A})} \kappa_2(s\widehat{E} - \widehat{A}). \quad (8)$$

This is the standard perturbation bound for the linear system $\widehat{G}\widehat{x} = \widehat{B}$, $\widehat{G} = s\widehat{E} - \widehat{A}$, under the deterministic perturbations $\Delta\widehat{G} = \sigma(s\delta E - \delta A)$ and $\Delta\widehat{B} = \sigma\delta B$.

Recall that by (5), $s\delta E - \delta A = \epsilon F_E + F_A\eta$, where ϵ and η are diagonal matrices whose diagonals are random vectors. These can be bounded in a probabilistic sense, using concentration inequalities which we briefly review next.

Proposition 2 *Let $Z = [z_1, \dots, z_n]^T$ be a random vector with independent standard normal real components $z_i \sim \mathcal{N}(0, 1)$, $i = 1, \dots, n$. Then with probability at least $1 - \exp(-n/2)$*

$$\|Z\|_2 \leq 2\sqrt{n}. \quad (9)$$

If $Z = [z_1, \dots, z_n]^T$ is a vector of independent complex random variables $z_i \sim \mathcal{CN}(0, 1)$, with \mathcal{CN} defined in Sect. 3.1, then

$$\|Z\|_2 \leq 4\sqrt{n} \quad (10)$$

holds with probability at least $1 - 2\exp(-n/2)$.

The estimate (9) follows from Gaussian concentration and because the χ^2 distribution is sub-Gaussian; see, e.g., [52, Example 2.28]. The statement (10) follows from (9), because $\|Z\|_2 \leq \|\Re(Z)\|_2 + \|\Im(Z)\|_2$, where $\Re(Z)$ is the vector that has as components the real parts of the components of Z and $\Im(Z)$ is the vector that has as components the imaginary parts of Z .

Lemma 1 *Let $s \in \Omega$, i.e., s is different from the poles of \widehat{H} , be such that*

$$0 < \sigma < \frac{1}{\kappa_2(s\widehat{E} - \widehat{A})} \min \left\{ \frac{\|s\widehat{E} - \widehat{A}\|_2}{4\sqrt{n} (\|F_E\|_2 + \|F_A\|_2)}, \frac{\|\widehat{B}\|_2}{4\sqrt{n}\|\widehat{B}\|_\infty} \right\}. \quad (11)$$

Then, with probability at least $1 - 4 \exp(-n/2)$, s is not a pole of \tilde{H} and the error bound (8) in Proposition 1 holds.

Proof Since, by (5), $s\delta E - \delta A = \epsilon F_E + F_A \eta$, we have

$$\|s\delta E - \delta A\|_2 \leq \|\epsilon\|_2 \|F_E\|_2 + \|F_A\|_2 \|\eta\|_2. \quad (12)$$

Because ϵ is diagonal¹ with the elements $\epsilon_1, \dots, \epsilon_n$ on the diagonal, we obtain, using Proposition 2, that

$$\|\epsilon\|_2 = \max_{i=1, \dots, n} |\epsilon_i| = \|[\epsilon_1, \dots, \epsilon_n]^T\|_\infty \leq \|[\epsilon_1, \dots, \epsilon_n]^T\|_2 \leq 4\sqrt{n}, \quad (13)$$

with probability at least $1 - 2 \exp(-n/2)$. Let δB_i denote the i th component of δB , then $\delta B_i = H(\mu_i)\epsilon_i = \widehat{B}_i \epsilon_i$ holds for $i = 1, \dots, n$ and thus Eq. (13) means that

$$\|\delta B\|_2 \leq \|\widehat{B}\|_\infty \|[\epsilon_1, \dots, \epsilon_n]^T\|_2 \leq 4\sqrt{n} \|\widehat{B}\|_\infty \quad (14)$$

$$\|\widetilde{B}\|_2 \leq (1 + \sigma 4\sqrt{n}) \|\widehat{B}\|_2 \quad (15)$$

holds with probability at least $1 - 2 \exp(-n/2)$. Similar arguments show that

$$\|\eta\|_2 \leq 4\sqrt{n} \quad (16)$$

and

$$\|\delta C\|_2 \leq 4\sqrt{n} \|\widehat{C}\|_\infty \quad (17)$$

hold with probability at least $1 - 2 \exp(-n/2)$, where we used $\delta C_i = H(\gamma_i)\eta_i = \widehat{C}_i \eta_i$ for $i = 1, \dots, n$. Set now

$$\zeta = \sigma \widehat{\zeta}, \quad \widehat{\zeta} = \max \left\{ \frac{4\sqrt{n} (\|F_E\|_2 + \|F_A\|_2)}{\|s\widehat{E} - \widehat{A}\|_2}, \frac{4\sqrt{n} \|\widehat{B}\|_\infty}{\|\widehat{B}\|_2} \right\} \quad (18)$$

and observe that (11) guarantees $\zeta \kappa_2(s\widehat{E} - \widehat{A}) < 1$. Thus, together with (12), it follows that

$$\begin{aligned} \|\sigma(s\delta E - \delta A)\|_2 &\leq \sigma 4\sqrt{n} (\|F_E\|_2 + \|F_A\|_2) \\ &= \sigma \frac{4\sqrt{n} (\|F_E\|_2 + \|F_A\|_2)}{\|s\widehat{E} - \widehat{A}\|_2} \|s\widehat{E} - \widehat{A}\|_2 \leq \zeta \|s\widehat{E} - \widehat{A}\|_2 \end{aligned} \quad (19)$$

with probability at least $(1 - 2 \exp(-n/2))^2 \geq 1 - 4 \exp(-n/2)$, where we used (18).

¹ Note that ϵ is a diagonal matrix defined in (5).

With (14) and the definition of ζ in (18), we also obtain that

$$\|\sigma \delta B\|_2 \leq \sigma 4\sqrt{n} \|\widehat{B}\|_\infty = \sigma \frac{4\sqrt{n} \|\widehat{B}\|_\infty}{\|\widehat{B}\|_2} \|\widehat{B}\|_2 \leq \zeta \|\widehat{B}\|_2 \quad (20)$$

holds with probability at least $1 - 4 \exp(-n/2)$. Thus, with (19), (20), and because (11) implies $\zeta \kappa_2(s\widehat{E} - \widehat{A}) < 1$, the error bound (8) is applicable with probability at least $1 - 4 \exp(-n/2)$, which also means that $s\widetilde{E} - \widetilde{A}$ is nonsingular. \square

The following theorem bounds the error due to noise in the transfer-function values.

Theorem 1 *Under the same assumptions as Lemma 1, for each $s \in \Omega$, there exists a constant $C_s > 0$ that may depend on s such that*

$$|\widetilde{H}(s) - \widehat{H}(s)| \leq C_s \sigma \quad (21)$$

holds with probability at least $1 - 4 \exp(-n/2)$.

Proof Consider now

$$\begin{aligned} \widehat{H}(s) - \widetilde{H}(s) &= \widehat{C}(s\widehat{E} - \widehat{A})^{-1} \widehat{B} - \widetilde{C}(s\widetilde{E} - \widetilde{A})^{-1} \widetilde{B} \\ &= \widehat{C} \left((s\widehat{E} - \widehat{A})^{-1} \widehat{B} - (s\widetilde{E} - \widetilde{A})^{-1} \widetilde{B} \right) - \sigma \delta C (s\widetilde{E} - \widetilde{A})^{-1} \widetilde{B} \end{aligned}$$

and take the absolute value to obtain

$$|\widehat{H}(s) - \widetilde{H}(s)| \leq c_1 \frac{\zeta}{1 - \zeta \kappa_2(s\widehat{E} - \widehat{A})} + |\sigma \delta C (s\widetilde{E} - \widetilde{A})^{-1} \widetilde{B}|, \quad (22)$$

where we invoked (8) with

$$c_1 = 2 \|\widehat{C}\|_2 \|(s\widehat{E} - \widehat{A})^{-1} \widehat{B}\|_2 \kappa_2(s\widehat{E} - \widehat{A}) \quad (23)$$

and ζ set as in (18). Note that (8) holds with probability at least $1 - 4 \exp(-n/2)$, and thus (22) holds with the same probability.

We now bound $\|(s\widetilde{E} - \widetilde{A})^{-1}\|_2$ in probability. Consider the Neumann expansion

$$\begin{aligned} (s\widetilde{E} - \widetilde{A})^{-1} &= (s\widehat{E} - \widehat{A} + \sigma (s\delta E - \delta A))^{-1} \\ &= (s\widehat{E} - \widehat{A})^{-1} \sum_{i=0}^{\infty} (-1)^i \sigma^i ((s\delta E - \delta A)(s\widehat{E} - \widehat{A})^{-1})^i, \end{aligned} \quad (24)$$

where the series converges to the inverse of $I + \sigma(s\delta E - \delta A)(s\widehat{E} - \widehat{A})^{-1}$ provided that² $\|\sigma(s\delta E - \delta A)(s\widehat{E} - \widehat{A})^{-1}\|_2 < 1$. Because (19) holds with probability at least $1 - 4\exp(-n/2)$, we obtain that with the same probability of at least $1 - 4\exp(-n/2)$

$$\|\sigma(s\delta E - \delta A)(s\widehat{E} - \widehat{A})^{-1}\|_2 \leq \zeta \|s\widehat{E} - \widehat{A}\|_2 \|(s\widehat{E} - \widehat{A})^{-1}\|_2 < 1,$$

holds, where we used assumption (11) in the second inequality. Note that the second inequality is strict. Set

$$\nu = 4\sqrt{n}(\|F_E\|_2 + \|F_A\|_2) \|(s\widehat{E} - \widehat{A})^{-1}\|_2,$$

and obtain with (19), (24), and $\sigma\nu < 1$ because of (11) that

$$\|(s\widetilde{E} - \widetilde{A})^{-1}\|_2 \leq \|(s\widehat{E} - \widehat{A})^{-1}\|_2 \sum_{i=0}^{\infty} (\nu\sigma)^i = \|(s\widehat{E} - \widehat{A})^{-1}\|_2 \frac{1}{1 - \nu\sigma} \quad (25)$$

holds with probability at least $1 - 4\exp(-n/2)$.

Then, we obtain the bound

$$\begin{aligned} |\sigma\delta C(s\widetilde{E} - \widetilde{A})^{-1}\widetilde{B}| &\leq \sigma\|\delta C\|_2 \|(s\widetilde{E} - \widetilde{A})^{-1}\|_2 \|\widetilde{B}\|_2 \\ &\leq 4\sqrt{n}\|\widehat{C}\|_{\infty}\|\widehat{B}\|_2 \|(s\widehat{E} - \widehat{A})^{-1}\|_2 \frac{\sigma(1 + \sigma 4\sqrt{n})}{1 - \nu\sigma} \\ &\leq c_2 \frac{\sigma + \sigma^2 4\sqrt{n}}{1 - \nu\sigma}, \end{aligned} \quad (26)$$

where we used (15), (17) and (25), which together hold with probability at least $1 - 4\exp(-n/2)$, and we set $c_2 = 4\sqrt{n}\|\widehat{C}\|_{\infty}\|\widehat{B}\|_2 \|(s\widehat{E} - \widehat{A})^{-1}\|_2$.

We obtain with (22), (26) the following bound

$$|\widehat{H}(s) - \widetilde{H}(s)| \leq \sigma \left[\frac{c_1\zeta}{1 - \sigma\zeta\kappa_2(s\widehat{E} - \widehat{A})} + \frac{c_2(1 + \sigma 4\sqrt{n})}{1 - \nu\sigma} \right], \quad (27)$$

which grows at most linearly in σ as long as condition (11) is satisfied, which shows (21). \square

Corollary 1 *Under the same conditions as Theorem 1 and for $s \in \Omega$ that are not zeros of \widehat{H} ,*

$$\frac{|\widetilde{H}(s) - \widehat{H}(s)|}{|\widehat{H}(s)|} \leq C_s \sigma \quad (28)$$

² The necessary and sufficient condition for the convergence is that the spectral radius of $\sigma(s\delta E - \delta A)(s\widehat{E} - \widehat{A})^{-1}$ is strictly less than one.

holds with probability at least $1 - 4 \exp(-n/2)$ and constant $C_s > 0$ that may depend on s .

Proof Note that $\widehat{H}(s)$ is independent of σ and thus dividing (21) by $|\widehat{H}(s)|$ is sufficient to show (28) if s is not a zero of \widehat{H} . To highlight a geometric interpretation of (28), we show (28) via a different approach. Since $|\widehat{H}(s)| = \|\widehat{C}\|_2 \|(s\widehat{E} - \widehat{A})^{-1}\widehat{B}\|_2 |\cos \angle(\widehat{C}^*, (s\widehat{E} - \widehat{A})^{-1}\widehat{B})|$, the factor c_1 defined in (23) can be interpreted as

$$c_1 = |\widehat{H}(s)| \frac{2\kappa_2(s\widehat{E} - \widehat{A})}{|\cos \angle(\widehat{C}^*, (s\widehat{E} - \widehat{A})^{-1}\widehat{B})|}$$

so that the first term on the right-hand side in (27) contains the bound on the relative error $|\widehat{H}(s) - \widetilde{H}(s)|/|\widehat{H}(s)|$ with the two natural condition numbers $\kappa_s(s\widehat{E} - \widehat{A})$ and $|\cos \angle(\widehat{C}^*, (s\widehat{E} - \widehat{A})^{-1}\widehat{B})|$. The interpretation is as follows: Evaluating \widetilde{H} essentially means solving a perturbed linear system $(s\widetilde{E} - \widetilde{A})^{-1}\widetilde{B}$ and then computing an inner product $(\widetilde{C}^*)^*((s\widetilde{E} - \widetilde{A})^{-1}\widetilde{B})$ with a perturbed vector \widetilde{C}^* . The sensitivity of the solution of a system of equations to perturbations is quantified by the condition number $\kappa_2(s\widehat{E} - \widehat{A})$ and the sensitivity of the inner product is quantified by $|\cos \angle(\widehat{C}^*, (s\widehat{E} - \widehat{A})^{-1}\widehat{B})|$. Similarly, the second term on the right-hand side in (27) can be modified as follows: Instead of (26), estimate the second term in (22) as

$$\begin{aligned} |\sigma \delta C(s\widetilde{E} - \widetilde{A})^{-1}\widetilde{B}| &\leq \sigma 4\sqrt{n} \|\widehat{C}\|_\infty \|(s\widehat{E} - \widehat{A})^{-1}\widehat{B}\|_2 \left(1 + \frac{2\zeta\kappa_2(s\widehat{E} - \widehat{A})}{1 - \zeta\kappa_2(s\widehat{E} - \widehat{A})}\right) \\ &= \sigma |\widehat{H}(s)| \frac{4\sqrt{n}}{|\cos \angle(\widehat{C}^*, (s\widehat{E} - \widehat{A})^{-1}\widehat{B})|} \frac{1 + \zeta\kappa_2(s\widehat{E} - \widehat{A})}{1 - \zeta\kappa_2(s\widehat{E} - \widehat{A})} \frac{\|\widehat{C}\|_\infty}{\|\widehat{C}\|_2}. \end{aligned}$$

Hence, because $\widehat{H}(s) \neq 0$ per assumption, we can write

$$\frac{|\widehat{H}(s) - \widetilde{H}(s)|}{|\widehat{H}(s)|} \leq \sigma \left(\frac{1}{|\cos \vartheta_s|} \left(\frac{2\zeta\kappa_2^{(s)}}{1 - \sigma\zeta\kappa_2^{(s)}} + 4\sqrt{n} \frac{1 + \zeta\kappa_2^{(s)}}{1 - \zeta\kappa_2^{(s)}} \frac{\|\widehat{C}\|_\infty}{\|\widehat{C}\|_2} \right) \right) \quad (29)$$

where $\vartheta_s = \angle(\widehat{C}^*, (s\widehat{E} - \widehat{A})^{-1}\widehat{B})$ and $\kappa_2^{(s)} = \kappa_2(s\widehat{E} - \widehat{A})$. \square

3.5 Remarks on Theorem 1

The following remarks are in order. First, note that condition (11) implies that $\widetilde{H}(s)$ exists with probability at least $1 - 4 \exp(-n/2)$. Second, a similar result as in Theorem 1 can be shown for an absolute noise model, i.e., where the noisy transfer-function values $\widetilde{H}(s)$ have a standard deviation that is independent of the transfer-function value $H(s)$. The major change is that matrices F_A and F_E defined in (5) are independent of the transfer-function values when an absolute noise model is used and thus condition (11) holds for a different range of standard deviations σ than for a relative noise model. Third, condition (11) in Theorem 1 depends on the scaling of

the entries of the system matrices \widehat{E} , \widehat{A} , \widehat{B} , and \widehat{C} . Consider two nonsingular matrices D_1 and D_2 of size $n \times n$. Then, the Loewner model given by \widehat{E} , \widehat{A} , \widehat{B} , \widehat{C} , derived from noiseless transfer-function values, can be transformed as

$$\check{E} = D_1 \widehat{E} D_2, \quad \check{A} = D_1 \widehat{A} D_2, \quad \check{B} = D_1 \widehat{B}, \quad \check{C} = \widehat{C} D_2,$$

with transfer function $\check{H}(s) = \check{C}(s\check{E} - \check{A})^{-1}\check{B}$. It holds $\check{H}(s) = \widehat{H}(s)$ for $s \in \mathbb{C}$; however, the condition number $\kappa_2(s\widehat{E} - \widehat{A})$ is *not* invariant under the transformation given by D_1 and D_2 , which means that condition (11) in Theorem 1 is not invariant if the system is transformed. A component-wise analysis [16, Sect. 7.2] could be an option to derive a version of Theorem 1 that is invariant under diagonal linear transformations D_1 and D_2 . Fourth, condition (11) in Theorem 1 depends on the interpolation points s_1, \dots, s_{2n} and on the partition $\{\mu_1, \dots, \mu_n\}, \{\gamma_1, \dots, \gamma_n\}$, which shows that the interpolation points and their partition can influence the robustness of the Loewner approach to noise; cf., e.g., [19, Sect. 2.1]. Our numerical results demonstrate that different choices of interpolation points indeed influence condition (11) and thus when the linear growth of the error (21) with the standard deviation σ of the noise holds. The numerical results below suggest that taking interpolation points that keep the condition number of $s\widehat{E} - \widehat{A}$ low seems reasonable. However, additional analyses and numerical experiments are necessary before one can give a definitive recommendation. The issue of selection and partition of the interpolation points is an important challenging problem that will be addressed in our future work.

4 Numerical Results

In this section, we demonstrate the error bound of Theorem 1 on numerical experiments with some well known benchmark examples.

4.1 CD Player

The system of a CD player is a common benchmark problem for model reduction and can be downloaded from the SLICOT website.³

4.1.1 Problem Setup

We consider single-input-single-output (SISO) systems and therefore we use only the first input and the second output of the CD-player system. The frequency range is $[2\pi, 200\pi]$, which contains some of the major dynamics of the CD-player system.

³ <http://slicot.org/20-site/126-benchmark-examples-for-model-reduction>.

The order of the system is $N = 120$. To derive a Loewner model of order n , we select $2n$ interpolation points as follows. First, n points w_1, \dots, w_n are selected logarithmically equidistant in the range $[2\pi, 200\pi]$ on the imaginary axis. Then, $s_i = (-1)^i w_i$ for $i = 1, \dots, n$ and $s_{n+i} = (-1)^{i+1} w_i$. The first n points s_1, \dots, s_n are put into the set $\{\mu_1, \dots, \mu_n\}$ and the following n points s_{n+1}, \dots, s_{2n} are put into $\{\gamma_1, \dots, \gamma_n\}$. To test the Loewner models, we select 200 test points $s_1^{\text{test}}, \dots, s_{200}^{\text{test}}$ on the imaginary axis in the range $[2\pi, 200\pi]$.

From the noiseless transfer-function values $H(\mu_1), \dots, H(\mu_n)$ and $H(\gamma_1), \dots, H(\gamma_n)$ a Loewner model is derived with transfer function \widehat{H} . Then, the transfer-function values are polluted with noise with standard deviation σ as described in Sect. 3.1 and a noisy Loewner model is derived with transfer function \widetilde{H}_σ . Note that we now explicitly denote standard deviation as subscript in the transfer functions of noisy Loewner models.

4.1.2 Results

We consider the error

$$e(\sigma) = \frac{1}{200} \sum_{i=1}^{200} |\widehat{H}(s_i^{\text{test}}) - \widetilde{H}_\sigma(s_i^{\text{test}})|, \quad (30)$$

which is an average of the error (21) over all 200 test points. Figure 1a shows the mean of $e(\sigma)$ over 10 replicates of independent noise samples. The standard deviation σ is in the range $[10^{-15}, 10^5]$ and the dimension is $n = 20$. The error bars in Fig. 1a show the minimum and maximum of $e(\sigma)$ over the 10 replicates. A linear growth of the mean of error (30) with the standard deviation σ is observed for $\sigma < 10^{-5}$. Figure 1b shows the number of test points that violate condition (11) of Theorem 1. The results indicate that for $\sigma \geq 10^{-7}$ condition (11) is violated for all 200 test points, which seems to align with Fig. 1a that shows a linear growth for $\sigma < 10^{-5}$. Thus, the results in Fig. 1a are in agreement with Theorem 1. Similar observations can be made for $n = 28$ in Fig. 1c, d.

4.2 Penzl

The Penzl system is a benchmark problem that has been introduced in [36, Example 3] and is used in, e.g., [19, 31] to demonstrate the Loewner approach.

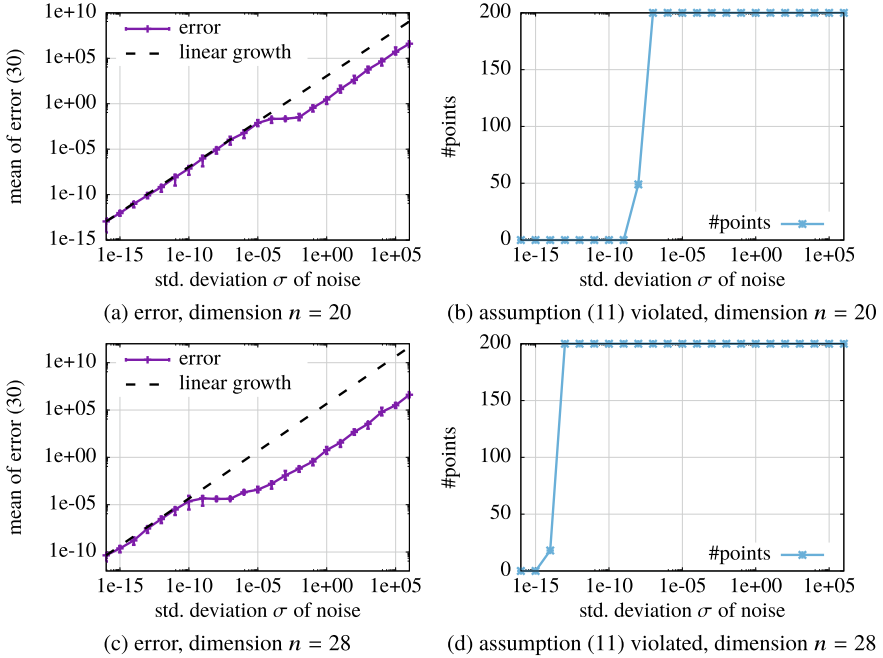


Fig. 1 CD player: Plots **a** and **c** show the growth of the mean of error (30) over 10 replicates of independent noise samples. Plots **b** and **d** show the number of test points for which condition (11) of Theorem 1 is violated. The mean of the error (30) grows linearly with σ as long as condition (11) is satisfied, which is in agreement with Theorem 1. The error bars in **a** and **c** show the minimum and maximum of the error (30) over the 10 replicates of the noise samples

4.2.1 Problem Setup

The Penzl system is of order $N = 1006$ and we consider the frequency range $[10, 1000]$. We consider two different sets of interpolation points in this example. First, we select n logarithmically equidistant points on the imaginary axis in the range $[10, 1000]$ and include their complex conjugates as described in Sect. 4.1.1. We denote the corresponding set of $2n$ interpolation points as \mathcal{E}_n . Second, we select points randomly in the complex plane in the range $[10, 1000] \times \iota[10, 1000]$, where ι is the complex unit $\iota = \sqrt{-1}$. To select the points randomly, we first select 10^6 logarithmically equidistant points w_1, \dots, w_{10^6} in $[10, 1000]$ and then draw uniformly $2n$ points from w_1, \dots, w_{10^6} for the real and imaginary parts of the n interpolation points s_1, \dots, s_n . Then, their complex conjugates are $s_{n+i} = \bar{s}_i$ for $i = 1, \dots, n$. The interpolation points are partitioned into the two sets $\{\mu_1, \dots, \mu_n\}$ and $\{\gamma_1, \dots, \gamma_n\}$ with $\mu_i = s_{2i-1}$ and $\gamma_i = s_{2i}$ for $i = 1, \dots, n$ such that conjugate pairs are in the same set. The set of interpolation points is denoted as \mathcal{R}_n . The test points are 200 points selected on the imaginary axis in the range $[10, 1000]$.

4.2.2 Results

Figure 2a shows the error (30) corresponding to the Loewner model with interpolation points \mathcal{R}_n (random) and of dimension $n = 16$. Figure 2c shows that condition (11) is violated for all test points and all standard deviations σ in the range $[10^{-15}, 10^5]$, which means that Theorem 1 is not applicable. In contrast, Fig. 2b shows a linear growth for the error (30) corresponding to the models learned from logarithmically equidistant points \mathcal{E}_n . Figure 2d indicates that up to $\sigma = 10^{-10}$ the condition (11) for Theorem 1 is satisfied, which explains the linear growth for $\sigma \leq 10^{-10}$. For $\sigma > 10^{-10}$, Theorem 1 is not applicable, even though a linear growth is observed, which demonstrates that Theorem 1 is rather pessimistic in this example. Figure 3 shows the magnitude of the transfer functions for $\sigma = 10^{-6}$ and demonstrates that the logarithmically equidistant points \mathcal{E}_n seem to provide more robustness against noise than the random points \mathcal{R}_n in this example. The numerical observations seem to be in alignment with Theorem 1 because assumption (11) is violated for interpolation points \mathcal{R}_n for all σ considered in Fig. 2c, whereas assumption (11) is satisfied for

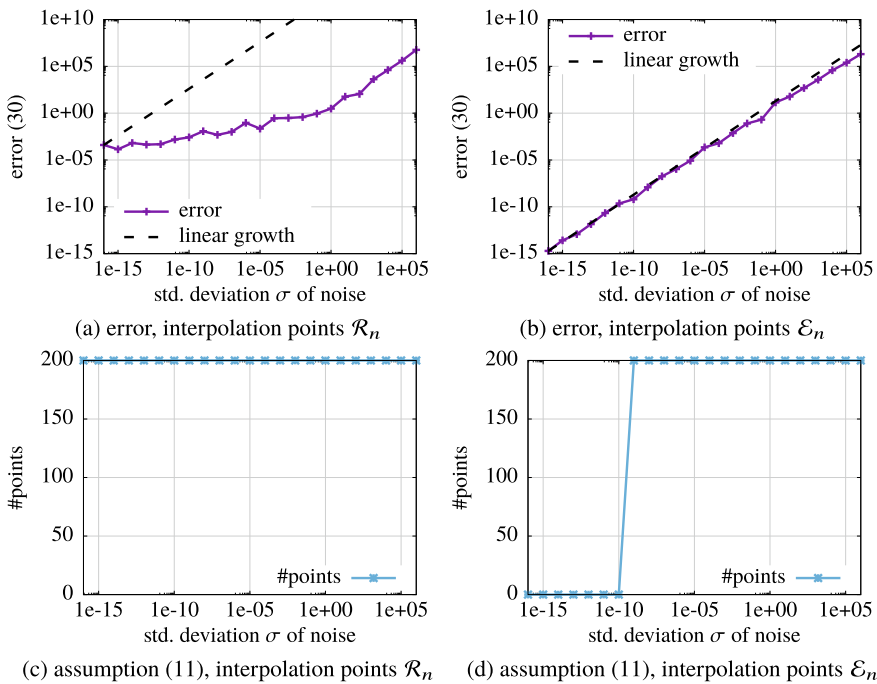


Fig. 2 Penzl: The choice of the interpolation points can have a significant effect on the robustness of the Loewner approach to noise. Plots **a** and **c** show results for points selected randomly and plots **b** and **d** for points selected logarithmically equidistantly. Learning Loewner models from the logarithmically equidistant points seems to be more robust in this example than learning from the randomly selected points

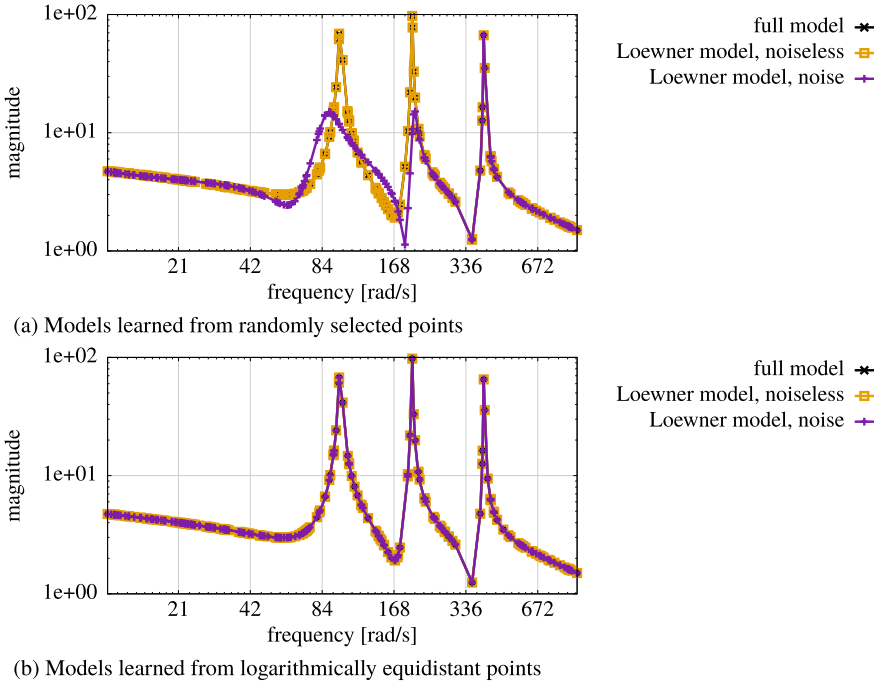


Fig. 3 Penzl: Plots **a** and **b** show the magnitude of the transfer function of Loewner models learned from noiseless transfer-function values and from transfer-function values polluted with noise with standard deviation $\sigma = 10^{-6}$. The interpolation points are randomly selected in **a** and logarithmically equidistant in **b**; compare with Fig. 2

small σ for points \mathcal{E}_n . Note that the condition number of $s\widehat{E} - \widehat{A}$ plays a critical role in the assumption (11). Thus, it seems that taking interpolation points that keep the condition number of $s\widehat{E} - \widehat{A}$ low is reasonable. However, additional analyses and numerical experiments are necessary before one can give a definitive recommendation.

Acknowledgements Drmač was supported in parts by the DARPA Contract HR0011-18-9-0033, the ONR Contract N00014-19-C-1053, and the Croatian Science Foundation through Grant IP-2019-04-6268 “*Randomized low rank algorithms and applications to parameter dependent problems.*” Peherstorfer was partially supported by US Department of Energy, Office of Advanced Scientific Computing Research, Applied Mathematics Program (Program Manager Dr. Steven Lee), DOE Award DESC0019334 and by the National Science Foundation under Grant No. 1901091. The numerical experiments were computed with support through the NYU IT High Performance Computing resources, services, and staff expertise.

References

1. Abderrahim, K., Mathlouthi, H., Msahli, F.: New approaches to finite impulse response systems identification using higher-order statistics. *IET Signal Proc.* **4**(5), 488–501 (2010)
2. Antoulas, A., Anderson, B.D.O.: On the scalar rational interpolation problem. *IMA J. Math. Control Inf.* **3**(2–3), 61–88 (1986)
3. Antoulas, A., Beattie, C., Gugercin, S.: Interpolatory model reduction of large-scale dynamical systems. In: Mohammadpour, J., Grigoriadis, K. (eds.) *Efficient Modeling and Control of Large-Scale Systems*. Springer (2010)
4. Antoulas, A., Gosea, I., Ionita, A.: Model reduction of bilinear systems in the Loewner framework. *SIAM J. Sci. Comput.* **38**(5), B889–B916 (2016)
5. Beattie, C., Gugercin, S.: Realization-independent \mathcal{H}_2 -approximation. In: *Proceedings of IEEE Conference on Decision Control*, pp. 4953–4958, Maui, HI, USA (2012)
6. Beattie, C., Gugercin, S., Wyatt, S.: Inexact solves in interpolatory model reduction. *Linear Algebra Appl.* **436**(8), 2916–2943 (2012). Special Issue dedicated to Danny Sorensen's 65th birthday
7. Benner, P., Gugercin, S., Willcox, K.: A survey of projection-based model reduction methods for parametric dynamical systems. *SIAM Rev.* **57**(4), 483–531 (2015)
8. Brunton, S.L., Proctor, J.L., Kutz, J.N.: Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci.* **113**(15), 3932–3937 (2016)
9. Drmač, Z., Mezić, I., Mohr, R.: Data driven modal decompositions: analysis and enhancements. *SIAM J. Sci. Comput.* **40**(4), A2253–A2285 (2018)
10. Drmač, Z., Gugercin, S., Beattie, C.: Quadrature-based vector fitting for discretized \mathcal{H}_2 approximation. *SIAM J. Sci. Comput.* **37**(2), A625–A652 (2015)
11. Drmač, Z., Gugercin, S., Beattie, C.: Vector fitting for matrix-valued rational approximation. *SIAM J. Sci. Comput.* **37**(5), A2346–A2379 (2015)
12. Embree, M., Ionita, A.C.: Pseudospectra of Loewner matrix pencils, pp. 1–18 (2019). [arXiv:1910.12153](https://arxiv.org/abs/1910.12153)
13. Gosea, I.V., Antoulas, A.C.: Data-driven model order reduction of quadratic-bilinear systems. *Numer. Linear Algebra with Appl.* **25**(6), e2200 (2018)
14. Gugercin, S., Antoulas, A., Beattie, C.: \mathcal{H}_2 model reduction for large-scale linear dynamical systems. *SIAM J. Matrix Anal. Appl.* **30**(2), 609–638 (2008)
15. Gustavsen, B., Semlyen, A.: Rational approximation of frequency domain responses by vector fitting. *IEEE Trans. Power Deliv.* **14**(3), 1052–1061 (1999)
16. Higham, N.J.: *Accuracy and Stability of Numerical Algorithms*. SIAM (2002)
17. Hokanson, J.M.: A data-driven McMillan degree lower bound. *SIAM J. Sci. Comput.* **42**(5), A3447–A3461 (2020)
18. Ionita, A., Antoulas, A.: Data-driven parametrized model reduction in the Loewner framework. *SIAM J. Sci. Comput.* **36**(3), A984–A1007 (2014)
19. Ionita, A.C.: Lagrange rational interpolation and its applications to approximation of large-scale dynamical systems. Ph.D. thesis, Rice University (2013)
20. Ionita, A.C., Antoulas, A.C.: Matrix pencils in time and frequency domain system identification. In: *Control, Robotics and Sensors*, pp. 79–88. Institution of Engineering and Technology (2012)
21. Kergus, P., Formentin, S., Pousset-Vassal, C., Demourant, F.: Data-driven control design in the Loewner framework: Dealing with stability and noise. In: *2018 European Control Conference (ECC)*, pp. 1704–1709 (2018)
22. Kramer, B., Gugercin, S.: Tangential interpolation-based eigensystem realization algorithm for MIMO systems. *Math. Comput. Model. Dyn. Syst.* **22**(4), 282–306 (2016)
23. Kramer, B., Peherstorfer, B., Willcox, K.: Feedback control for systems with uncertain parameters using online-adaptive reduced models. *SIAM J. Appl. Dyn. Syst.* **16**(3), 1563–1586 (2017)
24. Kutz, J.N., Brunton, S.L., Brunton, B.W., Proctor, J.L.: *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems*. SIAM (2016)

25. Lefteriu, S., Antoulas, A.: A new approach to modeling multiport systems from frequency-domain data. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **29**(1), 14–27 (2010)
26. Lefteriu, S., Ionita, A.C., Antoulas, A.C.: Modeling systems based on noisy frequency and time domain measurements. In: Willems, J.C., Hara, S., Ohta, Y., Fujioka, H. (eds.) *Perspectives in Mathematical System Theory, Control, and Signal Processing: A Festschrift in Honor of Yutaka Yamamoto on the Occasion of his 60th Birthday*, pp. 365–378. Springer, Berlin, Heidelberg (2010)
27. Ljung, L.: *System Identification*. Prentice Hall (1987)
28. Mayo, A., Antoulas, A.: A framework for the solution of the generalized realization problem. *Linear Algebra Appl.* **425**(2–3), 634–662 (2007)
29. Mendel, J.: Tutorial on higher-order statistics (spectra) in signal processing and system theory: theoretical results and some applications. *Proc. IEEE* **79**, 278–305 (1991)
30. Peherstorfer, B.: Sampling low-dimensional Markovian dynamics for preasymptotically recovering reduced models from data with operator inference. *SIAM J. Sci. Comput.* **42**(5), A3489–A3515 (2020)
31. Peherstorfer, B., Gugercin, S., Willcox, K.: Data-driven reduced model construction with time-domain Loewner models. *SIAM J. Sci. Comput.* **39**(5), A2152–A2178 (2017)
32. Peherstorfer, B., Willcox, K.: Dynamic data-driven reduced-order models. *Comput. Methods Appl. Mech. Eng.* **291**, 21–41 (2015)
33. Peherstorfer, B., Willcox, K.: Data-driven operator inference for nonintrusive projection-based model reduction. *Comput. Methods Appl. Mech. Eng.* **306**, 196–215 (2016)
34. Peherstorfer, B., Willcox, K.: Dynamic data-driven model reduction: adapting reduced models from incomplete data. *Adv. Model. Simul. Eng. Sci.* **3**(11) (2016)
35. Peherstorfer, B., Willcox, K., Gunzburger, M.: Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Rev.* **60**(3), 550–591 (2018)
36. Penzl, T.: Algorithms for model reduction of large dynamical systems. *Linear Algebra Appl.* **415**(2–3), 322–343 (2006)
37. Qian, E., Kramer, B., Marques, A.N., Willcox, K.E.: Transform & learn: a data-driven approach to nonlinear model reduction. In: *AIAA Aviation 2019 Forum* (2019)
38. Qin, S.J.: An overview of subspace identification. *Comput. Chem. Eng.* **30**(10–12), 1502–1513 (2006)
39. Rabiner, L., Crochiere, R., Allen, J.: FIR system modeling and identification in the presence of noise and with band-limited inputs. *IEEE Trans. Acoust. Speech Signal Process.* **26**(4), 319–333 (1978)
40. Reynders, E.: System identification methods for (operational) modal analysis: review and comparison. *Arch. Comput. Methods Eng.* **19**(1), 51–124 (2012)
41. Rowley, C., Mezić, I., Bagheri, S., Schlatter, P., Henningson, D.: Spectral analysis of nonlinear flows. *J. Fluid Mech.* **641**, 115–127 (2009)
42. Rudy, S.H., Brunton, S.L., Proctor, J.L., Kutz, J.N.: Data-driven discovery of partial differential equations. *Sci. Adv.* **3**(4) (2017)
43. Schaeffer, H., Cafilisch, R., Hauck, C.D., Osher, S.: Sparse dynamics for partial differential equations. *Proc. Natl. Acad. Sci.* **110**(17), 6634–6639 (2013)
44. Schaeffer, H., Tran, G., Ward, R.: Extracting sparse high-dimensional dynamics from limited data. *SIAM J. Appl. Math.* **78**(6), 3279–3295 (2018)
45. Schmid, P.: Dynamic mode decomposition of numerical and experimental data. *J. Fluid Mech.* **656**, 5–28, 8 (2010)
46. Schmid, P., Sesterhenn, J.: Dynamic mode decomposition of numerical and experimental data. In: *61st APS Meeting Bulletin of American Physical Society*, p. 208. American Physical Society (2008)
47. Schulze, P., Unger, B.: Data-driven interpolation of dynamical systems with delay. *Syst. Control Lett.* **97**, 125–131 (2016)
48. Schulze, P., Unger, B., Beattie, C., Gugercin, S.: Data-driven structured realization. *Linear Algebra Appl.* **537**, 250–286 (2018)

49. Swischuk, R., Kramer, B., Huang, C., Willcox, K.: Learning physics-based reduced-order models for a single-injector combustion process. *AIAA J.* **58**(6), 2658–2672 (2020)
50. Tu, J.H., Rowley, C.W., Luchtenburg, D.M., Brunton, S.L., Kutz, J.N.: On dynamic mode decomposition: theory and applications. *J. Comput. Dyn.* **1**(2), 391–421 (2014)
51. Viberg, M.: Subspace-based methods for the identification of linear time-invariant systems. *Automatica* **31**(12), 1835–1851 (1995). *Trends in System Identification*
52. Wainwright, M.J.: *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press (2019)

Pseudospectra of Loewner Matrix Pencils



Mark Embree and A. Cosmin Ioniță

Abstract Loewner matrix pencils play a central role in the system realization theory of Mayo and Antoulas, an important development in data-driven modeling. The eigenvalues of these pencils reveal system poles. How robust are the poles recovered via Loewner realization? With several simple examples, we show how pseudospectra of Loewner pencils can be used to investigate the influence of interpolation point location and partitioning on pole stability, the transient behavior of the realized system, and the effect of noisy measurement data. We include an algorithm to efficiently compute such pseudospectra by exploiting Loewner structure.

Keywords Generalized eigenvalue problem · Pseudospectra · System realization

1 Introduction

The landmark systems realization theory of Mayo and Antoulas [19] shows how to construct a dynamical system that interpolates tangential frequency domain measurements of a multi-input, multi-output system. Central to this development is the *matrix pencil* $z\mathbb{L} - \mathbb{L}_s$, composed of Loewner and shifted Loewner matrices \mathbb{L} and \mathbb{L}_s that encode the interpolation data. When this technique is used for exact system recovery (as opposed to data-driven model reduction), the eigenvalues of this pencil match the poles of the transfer function of the original system. However other spectral properties, including the sensitivity of the eigenvalues to perturbation, can differ greatly, depending on the location of the interpolation points in the complex plane

Dedicated to Thanos Antoulas.

M. Embree (✉)

Department of Mathematics and Computational Modeling and Data Analytics Division,
Academy of Integrated Science, Virginia Tech, Blacksburg, VA 24061, USA
e-mail: embree@vt.edu

A. C. Ioniță

The MathWorks Inc., Natick, MA 01760, USA
e-mail: cionita@mathworks.com

© Springer Nature Switzerland AG 2022

C. Beattie et al. (eds.), *Realization and Model Reduction of Dynamical Systems*,
https://doi.org/10.1007/978-3-030-95157-3_4

relative to the system poles, and how the interpolation points are partitioned. Since one uses $z\mathbb{L} - \mathbb{L}_s$ to learn about the original system, such subtle differences matter.

Pseudospectra are sets in the complex plane that contain the eigenvalues but provide additional insight about the sensitivity of those eigenvalues to perturbation and the transient behavior of the underlying dynamical system. While most often used to analyze single matrices, pseudospectral concepts have been extended to matrix pencils (generalized eigenvalue problems).

This introductory note shows several ways to use pseudospectra to investigate spectral questions involving Loewner pencils derived from system realization problems. Using simple examples, we explore the following questions.

- How do the locations of the interpolation points and their partition into “left” and “right” points affect the sensitivity of the eigenvalues of $z\mathbb{L} - \mathbb{L}_s$?
- Do solutions to the dynamical system $\mathbb{L}\dot{\mathbf{x}}(t) = \mathbb{L}_s\mathbf{x}(t)$ mimic solutions to the original system $\mathbf{E}\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t)$, especially in the transient regime? Does this agreement depend on the interpolation points?
- How do noisy measurements affect the eigenvalues of $z\mathbb{L} - \mathbb{L}_s$?

We include an algorithm for computing pseudospectra of an n -dimensional Loewner pencil in $\mathcal{O}(n^2)$ operations, improving the $\mathcal{O}(n^3)$ cost for generic matrix pencils; the appendix gives a MATLAB implementation.

Throughout this note, we use $\sigma(\cdot)$ to denote the spectrum (eigenvalues) of a matrix or matrix pencil, and $\|\cdot\|$ to denote the vector 2-norm and the matrix norm it induces. (All definitions here can readily be adapted to other norms, as needed. The algorithm, however, is designed for use with the 2-norm.)

2 Loewner Realization Theory in a Nutshell

We briefly summarize Loewner realization theory, as developed by Mayo and Antoulas [19]; see also [1]. Consider the linear, time-invariant dynamical system

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

for $\mathbf{A}, \mathbf{E} \in \mathbb{C}^{n \times n}$, $\mathbf{B} \in \mathbb{C}^{n \times m}$, and $\mathbf{C} \in \mathbb{C}^{p \times n}$, with which we associate, via the Laplace transform, the transfer function $\mathbf{H}(z) = \mathbf{C}(z\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}$.

Given *tangential measurements* of $\mathbf{H}(z)$ we seek a realization of the system that interpolates the given data. More precisely, consider the *right interpolation data*

- distinct interpolation points $\lambda_1, \dots, \lambda_\rho \in \mathbb{C}$;
- interpolation directions $\mathbf{r}_1, \dots, \mathbf{r}_\rho \in \mathbb{C}^m$;
- function values $\mathbf{w}_1, \dots, \mathbf{w}_\rho \in \mathbb{C}^p$;

and *left interpolation data*

- distinct interpolation points $\mu_1, \dots, \mu_\nu \in \mathbb{C}$;

- interpolation directions $\ell_1, \dots, \ell_\nu \in \mathbb{C}^p$;
- function values $\mathbf{v}_1, \dots, \mathbf{v}_\nu \in \mathbb{C}^m$.

Assume the left and right interpolation points are disjoint, $\{\lambda_i\}_{i=1}^\varrho \cap \{\mu_j\}_{j=1}^\nu = \emptyset$; (in our examples, all $\varrho + \nu$ points are distinct).

The interpolation problem seeks matrices $\widehat{\mathbf{A}}, \widehat{\mathbf{E}}, \widehat{\mathbf{B}}, \widehat{\mathbf{C}}$ for which the transfer function $\widehat{\mathbf{H}}(z) = \widehat{\mathbf{C}}(z\widehat{\mathbf{E}} - \widehat{\mathbf{A}})^{-1}\widehat{\mathbf{B}}$ interpolates the data: for $i = 1, \dots, \varrho$ and $j = 1, \dots, \nu$,

$$\widehat{\mathbf{H}}(\lambda_i)\mathbf{r}_i = \mathbf{w}_i, \quad \ell_j^*\widehat{\mathbf{H}}(\mu_j) = \mathbf{v}_j^*.$$

Two structured matrices play a crucial role in the development of Mayo and Antoulas [19]. From the data, construct the *Loewner* and *shifted Loewner* matrices

$$\mathbb{L} = \begin{bmatrix} \frac{\mathbf{v}_1^*\mathbf{r}_1 - \ell_1^*\mathbf{w}_1}{\mu_1 - \lambda_1} & \dots & \frac{\mathbf{v}_1^*\mathbf{r}_\varrho - \ell_1^*\mathbf{w}_\varrho}{\mu_1 - \lambda_\varrho} \\ \vdots & \ddots & \vdots \\ \frac{\mathbf{v}_\nu^*\mathbf{r}_1 - \ell_\nu^*\mathbf{w}_1}{\mu_\nu - \lambda_1} & \dots & \frac{\mathbf{v}_\nu^*\mathbf{r}_\varrho - \ell_\nu^*\mathbf{w}_\varrho}{\mu_\nu - \lambda_\varrho} \end{bmatrix}, \quad \mathbb{L}_s = \begin{bmatrix} \frac{\mu_1\mathbf{v}_1^*\mathbf{r}_1 - \lambda_1\ell_1^*\mathbf{w}_1}{\mu_1 - \lambda_1} & \dots & \frac{\mu_1\mathbf{v}_1^*\mathbf{r}_\varrho - \lambda_\varrho\ell_1^*\mathbf{w}_\varrho}{\mu_1 - \lambda_\varrho} \\ \vdots & \ddots & \vdots \\ \frac{\mu_\nu\mathbf{v}_\nu^*\mathbf{r}_1 - \lambda_1\ell_\nu^*\mathbf{w}_1}{\mu_\nu - \lambda_1} & \dots & \frac{\mu_\nu\mathbf{v}_\nu^*\mathbf{r}_\varrho - \lambda_\varrho\ell_\nu^*\mathbf{w}_\varrho}{\mu_\nu - \lambda_\varrho} \end{bmatrix}, \quad (1)$$

i.e., the (i, j) entries of these $\nu \times \varrho$ matrices have the form

$$(\mathbb{L})_{i,j} = \frac{\mathbf{v}_i^*\mathbf{r}_j - \ell_i^*\mathbf{w}_j}{\mu_i - \lambda_j}, \quad (\mathbb{L}_s)_{i,j} = \frac{\mu_i\mathbf{v}_i^*\mathbf{r}_j - \lambda_j\ell_i^*\mathbf{w}_j}{\mu_i - \lambda_j}.$$

Now collect the data into matrices. The right interpolation points, directions, and data are stored in

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_\varrho \end{bmatrix} \in \mathbb{C}^{\varrho \times \varrho}, \quad \mathbf{R} = \begin{bmatrix} \mathbf{r}_1 & \dots & \mathbf{r}_\varrho \end{bmatrix} \in \mathbb{C}^{m \times \varrho}, \quad \mathbf{W} = \begin{bmatrix} \mathbf{w}_1 & \dots & \mathbf{w}_\varrho \end{bmatrix} \in \mathbb{C}^{p \times \varrho},$$

while the left interpolation points, directions, and data are stored in

$$\mathbf{M} = \begin{bmatrix} \mu_1 & & \\ & \ddots & \\ & & \mu_\nu \end{bmatrix} \in \mathbb{C}^{\nu \times \nu}, \quad \mathbf{L} = \begin{bmatrix} \ell_1 & \dots & \ell_\nu \end{bmatrix} \in \mathbb{C}^{p \times \nu}, \quad \mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \dots & \mathbf{v}_\nu \end{bmatrix} \in \mathbb{C}^{m \times \nu}.$$

2.1 Selecting and Arranging Interpolation Points

As Mayo and Antoulas observe, Sylvester equations connect these matrices:

$$\mathbb{L}\mathbf{A} - \mathbf{M}\mathbb{L} = \mathbf{L}^*\mathbf{W} - \mathbf{V}^*\mathbf{R}, \quad \mathbb{L}_s\mathbf{A} - \mathbf{M}\mathbb{L}_s = \mathbf{L}^*\mathbf{W}\mathbf{A} - \mathbf{M}\mathbf{V}^*\mathbf{R}. \quad (2)$$

Just using the dimensions of the components, note that

$$\begin{aligned} \text{rank}(\mathbf{L}^*\mathbf{W}), \text{rank}(\mathbf{L}^*\mathbf{W}\mathbf{A}), &\leq \min\{v, q, p\} \\ \text{rank}(\mathbf{V}^*\mathbf{R}), \text{rank}(\mathbf{V}^*\mathbf{R}\mathbf{M}) &\leq \min\{v, q, m\}. \end{aligned}$$

Thus for modest m and p , the Sylvester equations (2) must have low-rank right-hand sides.¹ This situation often implies the rapid decay of singular values of solutions to the Sylvester equation [2, 3, 21, 22, 26]. While this phenomenon is convenient for balanced truncation model reduction (enabling low-rank approximations to the controllability and observability Gramians), it is less welcome in the Loewner realization setting, where the rank of \mathbb{L} should reveal the order of the original system: *fast decay of the singular values of \mathbb{L} makes this rank ambiguous*. Since \mathbf{A} and \mathbf{M} are diagonal, they are normal matrices, and hence Theorem 2.1 of [3] gives

$$\frac{s_{qk+1}(\mathbb{L})}{s_1(\mathbb{L})} \leq \inf_{\phi \in \mathcal{R}_{k,k}} \frac{\max\{|\phi(\lambda)| : \lambda \in \{\lambda_1, \dots, \lambda_q\}\}}{\min\{|\phi(\mu)| : \mu \in \{\mu_1, \dots, \mu_v\}\}}, \quad (3)$$

where $s_j(\cdot)$ denotes the j th largest singular value, $q = \text{rank}(\mathbf{L}^*\mathbf{W} - \mathbf{V}^*\mathbf{R})$, and $\mathcal{R}_{k,k}$ denotes the set of irreducible rational functions whose numerators and denominators are polynomials of degree k or less.² The right hand side of (3) will be *small* when there exists some $\phi \in \mathcal{R}_{k,k}$ for which all $|\phi(\lambda_i)|$ are small, while all $|\phi(\mu_j)|$ are large: a good separation of $\{\lambda_i\}$ from $\{\mu_j\}$ is thus *sufficient* to ensure the rapid decay of the singular values of \mathbb{L} and \mathbb{L}_s . (Beckermann and Townsend give an explicit bound for the singular values of Loewner matrices when the interpolation points fall in disjoint real intervals [3, Corollary 4.2].)

Here, if we want the singular values of \mathbb{L} and \mathbb{L}_s to reveal the system's order (without decay of singular values as an accident of the arrangement of interpolation points), it is *necessary* for one $|\phi(\lambda_i)|$ to be about the same size as the smallest value of $|\phi(\mu_j)|$ for all $\phi \in \mathcal{R}_{k,k}$. Roughly speaking, we want the left and right interpolation points to be close together (even interleaved). While this arrangement is *necessary* for slow decay of the singular values, it does not alone prevent such decay, as will be seen in examples in Fig. 4 (since (3) is only an upper bound).

Another heuristic, based on the Cauchy-like structure of \mathbb{L} and \mathbb{L}_s , also suggests the left and right interpolation points should be close together. Namely, \mathbb{L} and \mathbb{L}_s are

¹ For single-input, single-output (SISO) systems, $m = p = 1$, so the rank of the right-hand sides in (2) cannot exceed two. The same will apply for multi-input, multi-output systems with identical left and right interpolation directions: $\ell_i \equiv \ell$ for all $i = 1, \dots, v$ and $\mathbf{r}_j \equiv \mathbf{r}$ for all $j = 1, \dots, q$.

² The same bound holds for $s_{qk+1}(\mathbb{L}_s)/s_1(\mathbb{L}_s)$ with $q = \text{rank}(\mathbf{L}^*\mathbf{W}\mathbf{A} - \mathbf{M}\mathbf{V}^*\mathbf{R})$.

a more general form of the Cauchy matrix $(\mathbf{C})_{i,j} = 1/(\mu_i - \lambda_j)$, whose determinant has the elegant formula (e.g., [16, p. 38])

$$\det(\mathbf{C}) = \frac{\prod_{1 \leq i < j \leq n} (\mu_j - \mu_i)(\lambda_i - \lambda_j)}{\prod_{1 \leq i \leq j \leq n} (\mu_i - \lambda_j)}. \quad (4)$$

It is an open question if $\det(\mathbb{L})$ and $\det(\mathbb{L}_s)$ have similarly elegant formulas. Nevertheless, $\det(\mathbb{L})$ and $\det(\mathbb{L}_s)$ do have the same denominator as $\det(\mathbf{C})$ (which can be checked by recursively subtracting the first row from all other rows when computing the determinant). This observation suggests that to avoid artificially small determinants for \mathbb{L} and \mathbb{L}_s (which, up to sign, are the products of the singular values) it is *necessary* for the denominator of (4) to be small, and, thus, for the left and right interpolation points to be close together.

In practice, we often start with initial interpolation points x_1, \dots, x_{2n} that we want to partition into left and right interpolation points to form \mathbb{L} and \mathbb{L}_s . Our analysis of (4) suggests a simple way to arrange the interpolation points such that the denominator of $\det(\mathbb{L})$ and $\det(\mathbb{L}_s)$ is small: relabel the points to satisfy

$$x_k = \arg \min_{k \leq q \leq 2n} |x_{k-1} - x_q|, \quad \text{for } k = 2, \dots, 2n, \quad (5)$$

$$\{\mu_i\} = \{x_1, x_3, \dots, x_{2n-1}\}, \quad \{\lambda_j\} = \{x_2, x_4, \dots, x_{2n}\}.$$

The greedy reordering in (5) ensures that $|x_k - x_{k-1}|$ is small and allows us to simply interleave the left and right interpolation points. Moreover, when x_1, \dots, x_{2n} are located on a line, the reordering in (5) simplifies to directly interleaving μ_i and λ_j and, thus, it can be skipped. This ordering need not be optimal, as we do not visit all possible combinations of $\mu_i - \lambda_j$; it simply seeks a partition that yields a large determinant (which must also depend on the interpolation *data*). We note its simplicity, effectiveness, and efficiency (requiring only $\mathcal{O}(n^2)$ operations). (For SISO systems, the state-space representation is equivalent to a barycentric form [17, p. 77]. We thank the referee for observing that in barycentric rational Remez approximations [9], a similar interleaving of support points from the set of reference points produces favorable numerical conditioning. Recent work by Gosea and Antoulas gives further numerical evidence for interleaving Loewner points [13].)

2.2 Construction of Interpolants

Throughout we make the fundamental assumptions that for all $\widehat{z} \in \{\lambda_i\}_{i=1}^{\ell} \cup \{\mu_j\}_{j=1}^{\nu}$,

$$\text{rank}(\widehat{z}\mathbb{L} - \mathbb{L}_s) = \text{rank} \left(\begin{bmatrix} \mathbb{L} \\ \mathbb{L}_s \end{bmatrix} \right) = \text{rank}([\mathbb{L} \ \mathbb{L}_s]) =: r,$$

and we presume the underlying dynamical system is controllable and observable.

When $r = \nu = \varrho$ is the order of the system, Mayo and Antoulas [19, Lemma 5.1] show that the transfer function $\widehat{\mathbf{H}}(z) := \widehat{\mathbf{C}}(z\widehat{\mathbf{E}} - \widehat{\mathbf{A}})^{-1}\widehat{\mathbf{B}}$ defined by

$$\widehat{\mathbf{E}} = -\mathbb{L}, \quad \widehat{\mathbf{A}} = -\mathbb{L}_s, \quad \widehat{\mathbf{B}} = \mathbf{V}^*, \quad \widehat{\mathbf{C}} = \mathbf{W} \quad (6)$$

interpolates the $\varrho + \nu$ data values.

When $r < \max(\nu, \varrho)$, fix some $\widehat{z} \in \{\lambda_i\}_{i=1}^{\varrho} \cup \{\mu_j\}_{j=1}^{\nu}$ and compute the (economy-sized) singular value decomposition

$$\widehat{z}\mathbb{L} - \mathbb{L}_s = \mathbf{Y}\mathbf{\Sigma}\mathbf{X}^*,$$

with $\mathbf{Y} \in \mathbb{C}^{\nu \times r}$, $\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$, and $\mathbf{X} \in \mathbb{C}^{\varrho \times r}$. Then with

$$\widehat{\mathbf{E}} = -\mathbf{Y}^*\mathbb{L}\mathbf{X}, \quad \widehat{\mathbf{A}} = -\mathbf{Y}^*\mathbb{L}_s\mathbf{X}, \quad \widehat{\mathbf{B}} = \mathbf{Y}^*\mathbf{V}^*, \quad \widehat{\mathbf{C}} = \mathbf{W}\mathbf{X}, \quad (7)$$

$\widehat{\mathbf{H}}(z) := \widehat{\mathbf{C}}(z\widehat{\mathbf{E}} - \widehat{\mathbf{A}})^{-1}\widehat{\mathbf{B}}$ gives an order- r system interpolating the data [19, Theorem 5.1].

3 Pseudospectra for Matrix Pencils

Though introduced decades earlier, in the 1990s pseudospectra emerged as a popular tool for analyzing the behavior of dynamical systems (see, e.g., [30]), eigenvalue perturbations (see, e.g., [4]), and stability of uncertain linear time-invariant (LTI) systems (see, e.g., [15]).

Definition 1 For a matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$ and $\varepsilon > 0$, the ε -pseudospectrum of \mathbf{A} is

$$\sigma_\varepsilon(\mathbf{A}) = \{z \in \mathbb{C} \text{ is an eigenvalue of } \mathbf{A} + \mathbf{\Gamma} \text{ for some } \mathbf{\Gamma} \in \mathbb{C}^{n \times n} \text{ with } \|\mathbf{\Gamma}\| < \varepsilon\}. \quad (8)$$

For all $\varepsilon > 0$, $\sigma_\varepsilon(\mathbf{A})$ is a bounded, open subset of the complex plane that contains the eigenvalues of \mathbf{A} . (A popular variation uses the weak inequality $\|\mathbf{\Gamma}\| \leq \varepsilon$; the strict inequality has favorable properties for operators on infinite-dimensional spaces [5].) Definition 1 motivates pseudospectra via eigenvalues of perturbed matrices. A numerical analyst studying accuracy of a backward stable eigenvalue algorithm might be concerned with ε on the order of $n\|\mathbf{A}\|_{\varepsilon_{\text{mach}}}$, where $\varepsilon_{\text{mach}}$ denotes the *machine epsilon* for the floating point system [20]. An engineer or scientist might consider $\sigma_\varepsilon(\mathbf{A})$ for much larger ε values, corresponding to uncertainty in parameters or data that contribute to the entries of \mathbf{A} .

Via the singular value decomposition, one can show that (8) is equivalent to

$$\sigma_\varepsilon(\mathbf{A}) = \{z \in \mathbb{C} : \|(z\mathbf{I} - \mathbf{A})^{-1}\| > 1/\varepsilon\}; \quad (9)$$

see, e.g., [29, Chap. 2]. The presence of the resolvent $(z\mathbf{I} - \mathbf{A})^{-1}$ in this definition suggests a connection to the transfer function $\mathbf{H}(z) = \mathbf{C}(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$ for the system

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

Indeed, definition (1) readily leads to bounds on $\|e^{t\mathbf{A}}\|$, and hence transient growth of solutions to $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t)$; see [29, Part IV].

Various extensions of pseudospectra have been proposed to handle more general eigenvalue problems and dynamical systems; see [7] for a concise survey. The first elaborations addressed the generalized eigenvalue problem $\mathbf{A}\mathbf{x} = \lambda\mathbf{E}\mathbf{x}$ (i.e., the matrix pencil $z\mathbf{E} - \mathbf{A}$) [10, 23, 24]. Here we focus on the definition proposed by Frayssé, Gueury, Nicoud, and Toumazou [10], which is ideally suited to analyzing eigenvalues of nearby matrix pencils. To permit the perturbations to \mathbf{A} and \mathbf{E} to be scaled independently, this definition includes two additional parameters, γ and δ .

Definition 2 Let $\gamma, \delta > 0$. For a pair of matrices $\mathbf{A}, \mathbf{E} \in \mathbb{C}^{n \times n}$ and any $\varepsilon > 0$, the ε - (γ, δ) -pseudospectrum $\sigma_\varepsilon^{(\gamma, \delta)}(\mathbf{A}, \mathbf{E})$ of the matrix pencil $z\mathbf{E} - \mathbf{A}$ is the set

$$\sigma_\varepsilon^{(\gamma, \delta)}(\mathbf{A}, \mathbf{E}) = \{z \in \mathbb{C} \text{ is an eigenvalue of the pencil } z(\mathbf{E} + \mathbf{\Delta}) - (\mathbf{A} + \mathbf{\Gamma}) \\ \text{for some } \mathbf{\Gamma}, \mathbf{\Delta} \in \mathbb{C}^{n \times n} \text{ with } \|\mathbf{\Gamma}\| < \varepsilon\gamma, \|\mathbf{\Delta}\| < \varepsilon\delta\}.$$

This definition has been extended to matrix polynomials in [14, 27].

Remark 1 Note that $\sigma_\varepsilon^{(\gamma, \delta)}(\mathbf{A}, \mathbf{E})$ is an open, nonempty subset of the complex plane, but it need not be bounded.

- (a) If $z\mathbf{E} - \mathbf{A}$ is a singular pencil ($\text{rank}(z\mathbf{E} - \mathbf{A}) < n$ for all $z \in \mathbb{C}$), then $\sigma(\mathbf{A}, \mathbf{E}) = \mathbb{C}$.
- (b) If $z\mathbf{E} - \mathbf{A}$ is a regular pencil but \mathbf{E} is not invertible, then $\sigma(\mathbf{A}, \mathbf{E})$ contains an infinite eigenvalue, and hence $\sigma_\varepsilon^{(\gamma, \delta)}(\mathbf{A}, \mathbf{E})$ is unbounded.
- (c) If \mathbf{E} is nonsingular but $\varepsilon\delta$ exceeds the distance of \mathbf{E} to singularity (the smallest singular value of \mathbf{E}), then $\sigma_\varepsilon^{(\gamma, \delta)}(\mathbf{A}, \mathbf{E})$ contains the point at infinity.

Since these pseudospectra can be unbounded, Lavallée [18] and Higham and Tisseur [14] visualize $\sigma_\varepsilon^{(\gamma, \delta)}(\mathbf{A}, \mathbf{E})$ as stereographic projections on the Riemann sphere.

Remark 2 Just as the conventional pseudospectrum $\sigma_\varepsilon(\mathbf{A})$ can be characterized using the resolvent of \mathbf{A} in (9), Frayssé et al. [10] show that Definition 2 is equivalent to

$$\sigma_\varepsilon^{(\gamma, \delta)}(\mathbf{A}, \mathbf{E}) = \left\{ z \in \mathbb{C} : \|(z\mathbf{E} - \mathbf{A})^{-1}\| > \frac{1}{\varepsilon(\gamma + |z|\delta)} \right\}. \quad (10)$$

This formula suggests a way to compute $\sigma_\varepsilon^{(\gamma, \delta)}(\mathbf{A}, \mathbf{E})$: evaluate $\|(z\mathbf{E} - \mathbf{A})^{-1}\|$ on a grid of points covering a relevant region of the complex plane (or the Riemann sphere) and use a contour plotting routine to draw boundaries of $\sigma_\varepsilon^{(\gamma, \delta)}(\mathbf{A}, \mathbf{E})$. The accuracy of the resulting pseudospectra depends on the density of the grid. Expedient algorithms

for computing $\|(z\mathbf{E} - \mathbf{A})^{-1}\|$ can be derived by computing a unitary simultaneous triangularization (generalized Schur form) of \mathbf{A} and \mathbf{E} in $\mathcal{O}(n^3)$ operations, then using inverse iteration or inverse Lanczos, as described by Trefethen [28] and Wright [31], to compute $\|(z\mathbf{E} - \mathbf{A})^{-1}\|$ at each grid point in $\mathcal{O}(n^2)$ operations. For the structured Loewner pencils of interest here, one can compute $\|(z\mathbf{E} - \mathbf{A})^{-1}\|$ in $\mathcal{O}(n^2)$ operations without recourse to the $\mathcal{O}(n^3)$ preprocessing step, as proposed in Sect. 4.

Remark 3 Definition 2 can be extended to $\delta = 0$ by only perturbing \mathbf{A} [23]:

$$\begin{aligned}\sigma_\varepsilon^{(1,0)}(\mathbf{A}, \mathbf{E}) &= \{z \in \mathbb{C} \text{ is an eigenvalue of the pencil } z\mathbf{E} - (\mathbf{A} + \mathbf{\Gamma}) \\ &\quad \text{for some } \mathbf{\Gamma} \in \mathbb{C}^{n \times n} \text{ with } \|\mathbf{\Gamma}\| < \varepsilon\}. \\ &= \{z \in \mathbb{C} : \|(z\mathbf{E} - \mathbf{A})^{-1}\| > 1/\varepsilon\}.\end{aligned}$$

This definition may be more suitable for cases where \mathbf{E} is fixed and uncertainty in the system only emerges, e.g., through physical parameters that appear in \mathbf{A} .

Remark 4 Since we ultimately intend to study the pseudospectra $\sigma_\varepsilon^{(\gamma,\delta)}(\mathbb{L}_s, \mathbb{L})$ of Loewner matrix pencils, one might question the use of generic perturbations $\mathbf{\Gamma}, \mathbf{\Delta} \in \mathbb{C}^{n \times n}$ in Definition 2. Should we restrict $\mathbf{\Gamma}$ and $\mathbf{\Delta}$ to maintain Loewner structure, i.e., so that $\mathbb{L}_s + \mathbf{\Gamma}$ and $\mathbb{L} + \mathbf{\Delta}$ maintain the coupled shifted Loewner–Loewner form in (1)? Such sets are called *structured pseudospectra*.

Three considerations motivate the study of generic perturbations $\mathbf{\Gamma}, \mathbf{\Delta} \in \mathbb{C}^{n \times n}$: one practical, one speculative, and one philosophical. (a) Beyond repeatedly computing the eigenvalues of Loewner pencils with randomly perturbed data, no systematic method is known to compute the coupled Loewner structured pseudospectra, i.e., no analogue of the resolvent-like definition (10) is known. (b) Rump [25] showed that in many cases, preserving structure has little effect on the standard matrix pseudospectra. For example, the structured ε -pseudospectrum of a Hankel matrix \mathbf{H} allowing only complex Hankel perturbations exactly matches the unstructured ε -pseudospectrum $\sigma_\varepsilon(\mathbf{H})$ based on generic complex perturbations [25, Theorem 4.3]. *Whether a similar results holds for Loewner structured pencils is an interesting open question.* (c) If one seeks to analyze the *behavior* of dynamical systems (as opposed to eigenvalues of nearby matrices), then generic perturbations give much greater insight; see [29, p. 456] for an example where *real-valued* perturbations do not move the eigenvalues much toward the imaginary axis (hence the real structured pseudospectra are benign), yet the stable system still exhibits strong transient growth.

As we shall see in Sect. 6, Definition 2 provides a helpful tool for investigating the sensitivity of eigenvalues of matrix pencils. A different generalization of Definition 1 gives insight into the transient behavior of solutions of $\mathbf{E}\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t)$. This approach is discussed in [29, Chap. 45], following [23, 24], and has been extended to handle singular \mathbf{E} in [8] (for differential-algebraic equations and descriptor systems). Restricting our attention here to nonsingular \mathbf{E} , we analyze the conventional (single matrix) pseudospectra $\sigma_\varepsilon(\mathbf{E}^{-1}\mathbf{A})$. From these sets one can develop various upper and lower bounds on $\|e^{t\mathbf{E}^{-1}\mathbf{A}}\|$ and $\|\mathbf{x}(t)\|$ [29, Chap. 15]. Here we shall just

state one basic result. If $\sup\{\operatorname{Re}(z) : z \in \sigma_\varepsilon(\mathbf{E}^{-1}\mathbf{A})\} = K\varepsilon$ for some $K \geq 1$ (where $\operatorname{Re}(\cdot)$ denotes the real part of a complex number), then

$$\sup_{t \geq 0} \|e^{t\mathbf{E}^{-1}\mathbf{A}}\| \geq K. \quad (11)$$

This statement implies that there exists some unit-length initial condition $\mathbf{x}(0)$ such that $\|\mathbf{x}(t)\| \geq K$, even though $\sigma(\mathbf{A}, \mathbf{E})$ may be contained in the left half-plane. (Optimizing this bound over $\varepsilon > 0$ yields the Kreiss Matrix Theorem [29, (15.9)].)

Pseudospectra of matrix pencils provide a natural vehicle to explore that stability of the matrix pencil associated with the Loewner realization in (6). We shall thus investigate eigenvalue perturbations via $\sigma_\varepsilon^{(\gamma, \delta)}(\widehat{\mathbf{A}}, \widehat{\mathbf{E}}) = \sigma_\varepsilon^{(\gamma, \delta)}(\mathbb{L}_s, \mathbb{L})$ and transient behavior via $\sigma_\varepsilon(\mathbb{L}^{-1}\mathbb{L}_s)$.

4 Efficient Computation of Loewner Pseudospectra

We first present a novel technique for efficiently computing pseudospectra of large Loewner matrix pencils, $\sigma_\varepsilon^{(\gamma, \delta)}(\mathbb{L}_s, \mathbb{L})$, using the equivalent definition given in (10). When the Loewner matrix \mathbb{L} is nonsingular, we employ inverse iteration to exploit the structure of the Loewner pencil to compute $\|(z\mathbb{L} - \mathbb{L}_s)^{-1}\|$ (in the two-norm) using only $\mathcal{O}(n^2)$ operations. This avoids the need to compute an initial simultaneous unitary triangularization of \mathbb{L}_s and \mathbb{L} using the QZ algorithm, an $\mathcal{O}(n^3)$ operation.

Inverse iteration (and inverse Lanczos) for $\|(z\mathbb{L} - \mathbb{L}_s)^{-1}\|$ requires computing

$$(z\mathbb{L} - \mathbb{L}_s)^{-*}(z\mathbb{L} - \mathbb{L}_s)^{-1}\mathbf{u} \quad (12)$$

for a series of vectors $\mathbf{u} \in \mathbb{C}^n$ (e.g., see [29, Chap. 39]). We invoke a property observed by Mayo and Antoulas [19], related to (2): by construction, the Loewner and shifted Loewner matrices satisfy $\mathbb{L}_s - \mathbb{L}\mathbf{A} = \mathbf{V}^*\mathbf{R}$. Thus the resolvent can be expressed using only \mathbb{L} and not \mathbb{L}_s :

$$(z\mathbb{L} - \mathbb{L}_s)^{-1} = (\mathbb{L}(z\mathbf{I} - \mathbf{A}) - \mathbf{V}^*\mathbf{R})^{-1}.$$

We now use the Sherman–Morrison–Woodbury formula (see, e.g., [12]) to get

$$(z\mathbb{L} - \mathbb{L}_s)^{-1} = (\mathbf{I} + \Upsilon(z)(\mathbf{I} - \mathbf{R}\Upsilon(z))^{-1}\mathbf{R})(z\mathbf{I} - \mathbf{A})^{-1}\mathbb{L}^{-1} = \Theta(z)\mathbb{L}^{-1},$$

where $\Upsilon(z) := (z\mathbf{I} - \mathbf{A})^{-1}\mathbb{L}^{-1}\mathbf{V}^*$ and $\Theta(z) := (\mathbf{I} + \Upsilon(z)(\mathbf{I} - \mathbf{R}\Upsilon(z))^{-1}\mathbf{R})(z\mathbf{I} - \mathbf{A})^{-1}$. As a result, we can compute the inverse iteration vectors in (12) as

$$(z\mathbb{L} - \mathbb{L}_s)^{-*}(z\mathbb{L} - \mathbb{L}_s)^{-1}\mathbf{u} = \mathbb{L}^{-*}\Theta(z)^*\Theta(z)\mathbb{L}^{-1}\mathbf{u}, \quad (13)$$

which requires solving several linear systems given by the same Loewner matrix \mathbb{L} , e.g., $\mathbb{L}^{-1}\mathbf{u}$, $\mathbb{L}^{-1}\mathbf{V}^*$.

Crucially, solving a linear system involving a Loewner matrix $\mathbb{L} \in \mathbb{C}^{n \times n}$ can be done efficiently in only $2(m+p+1)n^2$ operations, since \mathbb{L} has displacement rank $m+p$. More precisely, \mathbb{L} is a Cauchy-like matrix that satisfies the Sylvester equation (2) given by diagonal generator matrices \mathbf{A} and \mathbf{M} and a right-hand side of rank at most $m+p$, i.e., $\text{rank}(\mathbf{L}^*\mathbf{W} - \mathbf{V}^*\mathbf{R}) \leq m+p$. The displacement rank structure of \mathbb{L} can be exploited to compute its LU factorization in only $2(m+p)n^2$ flops (see [11] and [12, Sect. 12.1]). Given the LU factorization of \mathbb{L} , solving $\mathbb{L}^{-1}\mathbf{u}$ in (13) via standard forward and backward substitution requires another $2n^2$ operations.

Next, multiplying $\Theta(z)$ with the solution of $\mathbb{L}^{-1}\mathbf{u}$ requires a total of $2mn^2 + (4m^2 + 6m + 2)n + \frac{2}{3}m^3 - m^2$ operations, namely (to leading order on the factorizations):

- $2mn^2 + 2mn$ operations to compute $\Upsilon(z) \in \mathbb{C}^{n \times m}$;
- $m^2(2n-1) + m$ operations to compute $\mathbf{I} - \mathbf{R}\Upsilon(z) \in \mathbb{C}^{m \times m}$;
- $\frac{2}{3}m^3 + 2m^2n$ operations to solve $(\mathbf{I} - \mathbf{R}\Upsilon(z))^{-1}\mathbf{R} \in \mathbb{C}^{m \times n}$ via an LU factorization followed by n forward and backward substitutions;
- $n + m(2n-1) + (2m-1)n + 2n$ operations to multiply $\Theta(z)$ with $\mathbb{L}^{-1}\mathbf{u}$.

Finally, multiplying with $\mathbb{L}^{-*}\Theta(z)^*$ in (13) requires an additional $2n^2 + 2mn^2 + (4m^2 + 6m + 2)n + \frac{2}{3}m^3 - m^2$ operations, bringing the total cost of computing (13) to $2(3m+p+1)n^2 + 4(2m^2 + 3m + 1)n + \frac{4}{3}m^3 - 2m^2$ operations.

In practice, the sizes of the right and left tangential directions are much smaller than the size of the Loewner pencil, i.e., $m, p \ll n$. For example, for scalar data (associated with SISO systems), $m = p = 1$. Therefore, in practice, computing (13) can be done in only $\mathcal{O}(n^2)$ operations.

Partial pivoting can be included in the LU factorization of the Loewner matrix \mathbb{L} to overcome numerical difficulties. Adding partial pivoting maintains the $\mathcal{O}(n^2)$ operation count for the LU factorization of \mathbb{L} (see [12, Sect. 12.1]), and hence computing (13) can still be done in $\mathcal{O}(n^2)$ operations. The appendix gives a MATLAB implementation of this efficient inverse iteration.

We measure these performance gains for a Loewner pencil generated by sampling $f(x) = \sum_{k=1}^8 (-1)^{k+1} \left(1 + 100(x-k)^2\right)^{-1/2} + (-1)^{k+1} \left(1 + 100(x-k-1/2)^2\right)^{-1/2}$ at $2n$ points uniformly spaced in the interval $[1, 8]$. We compare our new $\mathcal{O}(n^2)$ Loewner pencil inverse iteration against a standard implementation (see [29, p. 373]) applied to a simultaneous triangularization of \mathbb{L}_s and \mathbb{L} . (The simultaneous triangularization costs $\mathcal{O}(n^3)$ but is fast, as MATLAB's `qz` routine invokes LAPACK code. For a fair comparison, we test against a C++ implementation of the fast Loewner code, compiled into a MATLAB `.mex` file.) Table 1 shows timings for both implementations by computing $\|(z\mathbb{L} - \mathbb{L}_s)^{-1}\|$ on a 200×200 grid of points. Exploiting the Loewner structure gives a significant performance improvement for large n .

We next examine two simple examples involving full-rank realization of SISO systems, to illustrate the kinds of insights one can draw from pseudospectra of Loewner pencils. (For these small examples we use the standard $\mathcal{O}(n^3)$ algorithm.)

Table 1 Comparison of speed of computing $\sigma_\varepsilon^{(1,1)}(\mathbb{L}_s, \mathbb{L})$ using the fast Loewner algorithm versus a generic inverse iteration method applied to simultaneous triangularizations of \mathbb{L}_s and \mathbb{L}

| n | $\mathcal{O}(n^2)$ algorithm (s) | $\mathcal{O}(n^3)$ algorithm (s) |
|-----|----------------------------------|----------------------------------|
| 100 | 1.85 | 1.65 |
| 200 | 6.75 | 6.75 |
| 300 | 17.24 | 33.30 |
| 400 | 49.15 | 65.05 |

5 Example 1: Eigenvalue Sensitivity and Transient Behavior

We first consider a simple controllable and observable SISO system with $n = 2$:

$$\mathbf{E} = \mathbf{I}, \quad \mathbf{A} = \begin{bmatrix} -1.1 & 1 \\ 1 & -1.1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \mathbf{C} = [0 \ 1]. \quad (14)$$

This \mathbf{A} is symmetric negative definite, with eigenvalues $\sigma(\mathbf{A}) = \{-0.1, -2.1\}$. Since the system is SISO, the transfer function $\mathbf{H}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$ maps \mathbb{C} to \mathbb{C} , and hence the choice of “interpolation directions” is trivial (though the division into “left” and “right” points matters). We take $\varrho = \nu = 2$ left and right interpolation points, with $\mathbf{r}_1 = \mathbf{r}_2 = 1$ and $\ell_1 = \ell_2 = 1$. We will study various choices of interpolation points, all of which satisfy, for each $\widehat{z} \in \{\lambda_1, \lambda_2, \mu_1, \mu_2\}$,

$$\text{rank}(\widehat{z} \mathbb{L} - \mathbb{L}_s) = \text{rank}(\mathbb{L}) = \text{rank}(\mathbb{L}_s) = n = 2. \quad (15)$$

This basic set-up makes it easy to focus on the influence of the interpolation points $\lambda_1, \lambda_2, \mu_1, \mu_2$. We will use the pseudospectra $\sigma_\varepsilon^{(1,1)}(\mathbb{L}_s, \mathbb{L})$ to examine how the interpolation points affect the stability of the eigenvalues of the Loewner pencil.

Table 2 records four different choices of $\{\lambda_1, \lambda_2, \mu_1, \mu_2\}$; Fig. 1 shows the corresponding pseudospectra $\sigma_\varepsilon^{(1,1)}(\mathbb{L}_s, \mathbb{L})$. All four Loewner realizations match the eigenvalues of \mathbf{A} and satisfy the interpolation conditions. However, the pseudospectra show how the *stability* of the eigenvalues -0.1 and -2.1 differs across these

Table 2 Right and left interpolation points for the two-dimensional SISO system (14). The two right columns report the singular values of the Loewner matrix \mathbb{L}

| Example | λ_1 | λ_2 | μ_1 | μ_2 | $s_1(\mathbb{L})$ | $s_2(\mathbb{L})$ |
|---------|-------------|-------------|---------|---------|-------------------|-------------------|
| (a) | 0 | 1 | 1i | -1i | 6.9871212 | 0.0731542 |
| (b) | 0.25 | 0.75 | 2i | -2i | 1.0021659 | 0.0296996 |
| (c) | 0.40 | 0.60 | 4i | -4i | 0.3605151 | 0.0057490 |
| (d) | 8 | 9 | 10 | 11 | 0.0035344 | 0.0000019 |

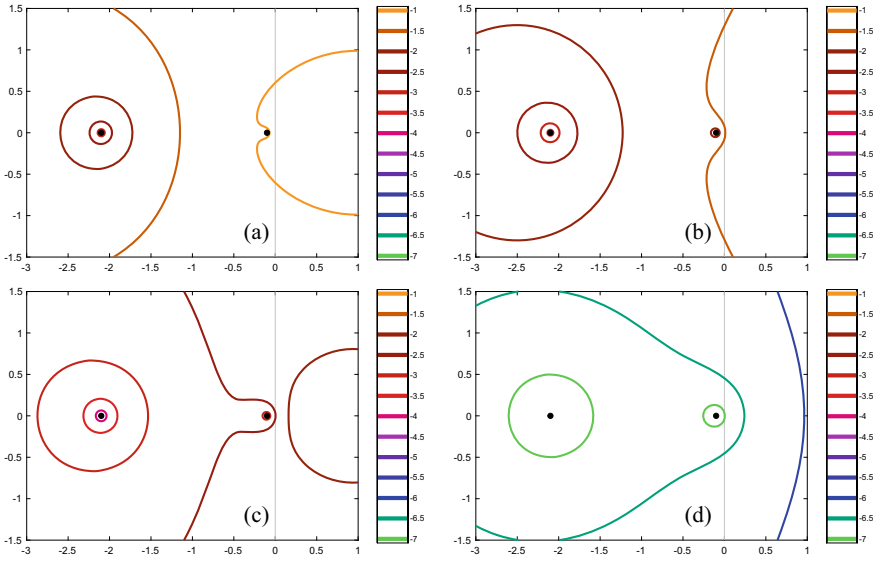


Fig. 1 Boundaries of pseudospectra $\sigma_\epsilon^{(1,1)}(\mathbb{L}_s, \mathbb{L})$ for four Loewner realizations of the system (14) using the interpolation points in Table 2. All four realizations correctly give $\sigma(\mathbb{L}_s, \mathbb{L}) = \sigma(\mathbf{A})$, but $\sigma_\epsilon^{(1,1)}(\mathbb{L}_s, \mathbb{L})$ show how the stability of the realized eigenvalues depends on the choice of interpolation points. *In this and all similar plots, the colors denote $\log_{10}(\epsilon)$.* Thus, in plot (d), there exist perturbations to \mathbb{L}_s and \mathbb{L} of norm $10^{-6.5}$ that move an eigenvalue into the right half-plane

four realizations. From example (a) to (d), these eigenvalues become increasingly sensitive as the interpolation points move farther from $\sigma(\mathbf{A})$. Table 2 also shows the singular values of the Loewner matrix \mathbb{L} , demonstrating how the second singular value $s_2(\mathbb{L})$ decreases as the eigenvalues become increasingly sensitive. (Taken to a greater extreme, it would eventually be difficult to determine if \mathbb{L} truly is rank 2.)

Remark 5 By Remark 1, note that if $\epsilon < s_{\min}(\mathbb{L})$, then $\sigma_\epsilon^{(1,1)}(\mathbb{L}_s, \mathbb{L})$ will be unbounded. Thus the decreasing values of $s_{\min}(\mathbb{L})$ in Table 2 suggest the enlarging pseudospectra seen in Fig. 1. For example, in case (d) the $\epsilon = 10^{-5}$ pseudospectrum $\sigma_\epsilon^{(1,1)}(\mathbb{L}_s, \mathbb{L})$ must contain the point at infinity.

Contrast these results with the standard pseudospectra of \mathbf{A} itself, $\sigma_\epsilon(\mathbf{A}) = \sigma_\epsilon^{(1,0)}(\mathbf{A}, \mathbf{I})$ shown at the top of Fig. 2. Since \mathbf{A} is real symmetric (hence normal), $\sigma_\epsilon(\mathbf{A})$ is the union of open ϵ -balls surrounding the eigenvalues. Figure 2 compares these pseudospectra to $\sigma_\epsilon(\mathbb{L}^{-1}\mathbb{L}_s)$, which give insight into the transient behavior of solutions to $\mathbb{L}\dot{\mathbf{x}}(t) = \mathbb{L}_s\mathbf{x}(t)$, e.g., via the bound (11). (Since $\mathbb{L}_s - \mathbb{L}\mathbf{A} = \mathbf{V}^*\mathbf{R}$, and $m = 1$, $\mathbb{L}^{-1}\mathbb{L}_s = \mathbf{A} + \mathbb{L}^{-1}\mathbf{V}^*\mathbf{R}$ is a rank-1 update of the matrix \mathbf{A} [19, p. 643].) The top plot shows $\sigma_\epsilon(\mathbf{A})$, whose rightmost extent in the complex plane is always $\epsilon - 0.1$: no transient growth is possible for this system. However, in all four Loewner realizations, $\sigma_\epsilon(\mathbb{L}^{-1}\mathbb{L}_s)$ extends more than ϵ into the right-half plane for $\epsilon = 10^0$

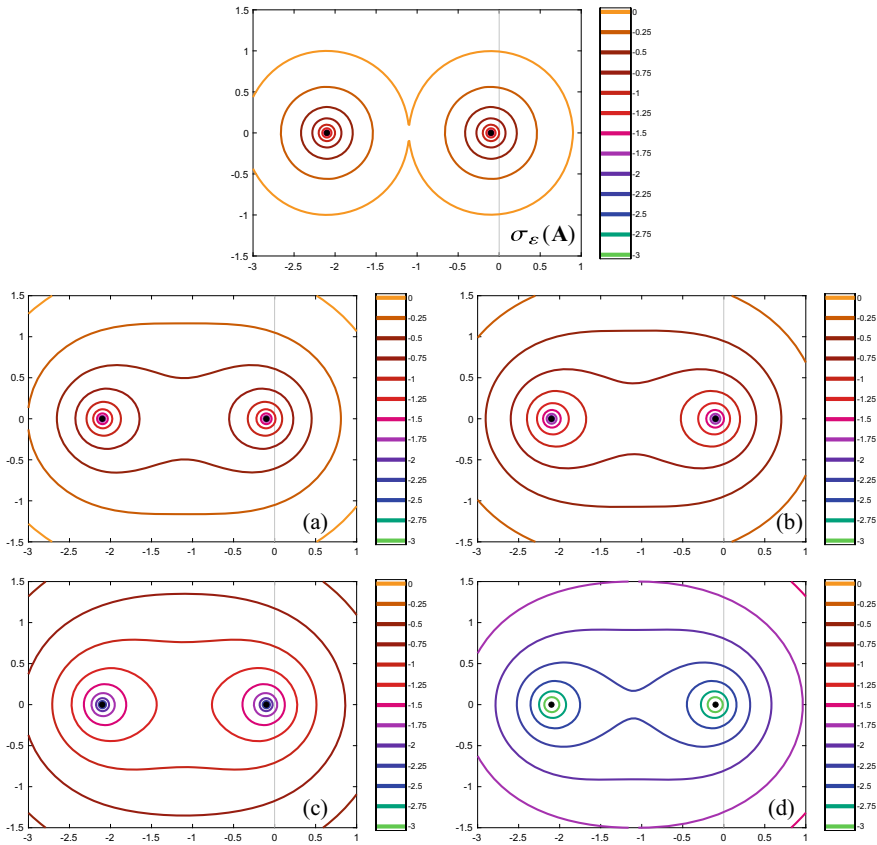


Fig. 2 The pseudospectra $\sigma_\varepsilon(\mathbf{A})$ (top), compared to $\sigma_\varepsilon(\mathbb{L}^{-1}\mathbb{L}_s)$ for four Loewner realizations of the system (14) using the interpolation points in Table 2. In all cases $\sigma(\mathbb{L}^{-1}\mathbb{L}_s) = \sigma(\mathbf{A})$, but the pseudospectra of $\mathbb{L}^{-1}\mathbb{L}_s$ are all quite a bit larger than $\sigma_\varepsilon(\mathbf{A})$

(orange level curve), indicating by (11) that transient growth must occur for some initial condition. Figure 3 shows this growth for all four realizations: *the more remote interpolation points lead to Loewner realizations with greater transient growth.*

6 Example 2: Partitioning Interpolation Points and Noisy Data

To further investigate how the interpolation points influence eigenvalue stability for the Loewner pencil, consider the SISO system of order 10 given by

$$\mathbf{E} = \mathbf{I}, \quad \mathbf{A} = \text{diag}(-1, -2, \dots, -10), \quad \mathbf{B} = [1, 1, \dots, 1]^T, \quad \mathbf{C} = [1, 1, \dots, 1].$$

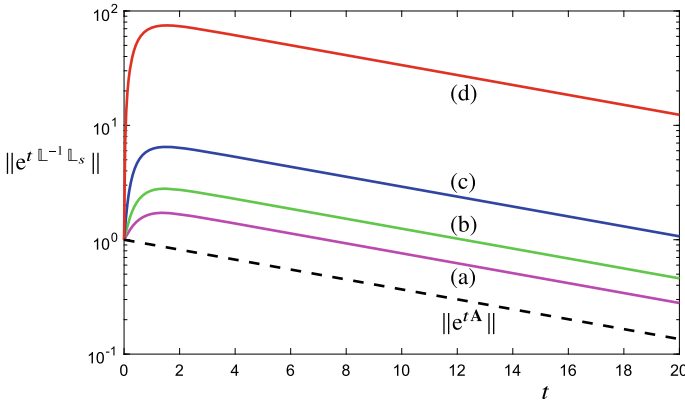


Fig. 3 Evolution of the norm of the solution operator for the original system (black dashed line) and the four interpolating Loewner models. The instability revealed by the pseudospectra in Fig. 2 corresponds to transient growth in the Loewner systems

Figure 4 shows $\sigma_\varepsilon^{(1,1)}(\mathbb{L}_s, \mathbb{L})$ for six configurations of the interpolation points. Plots (a) and (b) use the points $\{-10.25, -9.75, -9.25, \dots, -1.25, -0.75\}$; in plot (a) the left and right points interleave, suggesting slower decay of the singular values of \mathbb{L} , as discussed in Sect. 2.1; in plot (b) the left and right points are separated, leading to faster decay of the singular values of \mathbb{L} and considerably larger pseudospectra. (The Beckermann–Townsend bound [3, Corollary 4.2] applies to this case.) Plot (c) further separates the left and right points, giving even larger pseudospectra. In plots (d) and (f), complex interpolation points $\{-5 \pm 0.5i, -5 \pm 1.0i, \dots, -5 \pm 5i\}$ are mostly interleaved (d) (keeping conjugate pairs together) and separated (f): the latter significantly enlarges the pseudospectra. Plot (f) uses the same relative arrangement that gave such nice results in plot (a) (the singular value bound (3) is the same for (a) and (e)), but their locations relative to the poles of the original system differ. The pseudospectra are now much larger, showing that a large upper bound in (3) is not alone enough to guarantee small pseudospectra. (Indeed, the pseudospectra are so large in (e) and (f) that the plots are dominated by numerical artifacts of computing $\|(z\mathbb{L} - \mathbb{L}_s)^{-1}\|$.) *Pseudospectra reveal the great influence interpolation point location and partition can have on the stability of the realized pencils.*

Pseudospectra also give insight into the consequences of inexact measurement data. (For a recent study of how noisy data affects the recovered transfer function, see [6].) Consider the following experiment. Take the scenario in Fig. 4a, the most robust of these examples. Subject each right and left measurement $\{\mathbf{w}_1, \dots, \mathbf{w}_\ell\} \subset \mathbb{C}$ and $\{\mathbf{v}_1, \dots, \mathbf{v}_\nu\} \subset \mathbb{C}$ to random complex noise of magnitude 10^{-1} , then build the Loewner pencil $z\widehat{\mathbb{L}} - \widehat{\mathbb{L}}_s$ from this noisy data. How do the badly polluted measurements affect the computed eigenvalues? Figure 5 shows the results of 1,000 random trials, which can depart from the true matrices significantly:

$$3.06 \leq \|\widehat{\mathbb{L}}_s - \mathbb{L}_s\| \leq 5.88, \quad 0.49 \leq \|\widehat{\mathbb{L}} - \mathbb{L}\| \leq 0.63.$$

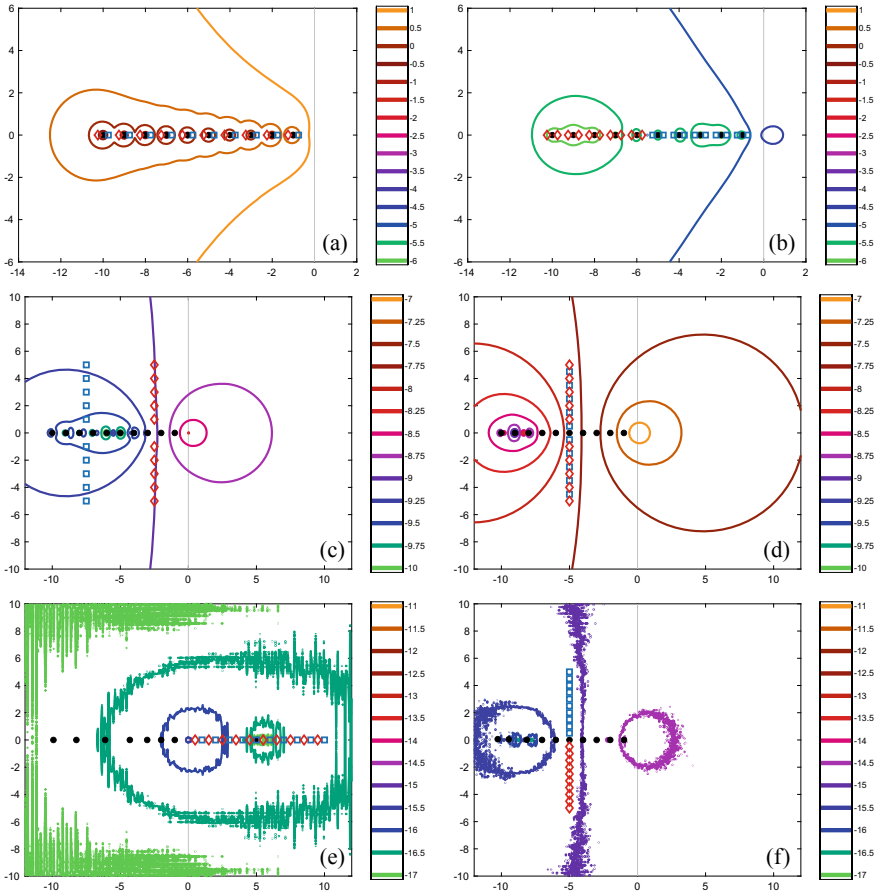


Fig. 4 Pseudospectra $\sigma_\varepsilon^{(1,1)}(\mathbb{L}_s, \mathbb{L})$ for six Loewner realizations of a SISO system of order $n = 10$ with poles $\sigma(\mathbf{A}) = \{-1, -2, \dots, -10\}$. Black dots show computed eigenvalues of the pencil $z\mathbb{L} - \mathbb{L}_s$ (which should agree with $\sigma(\mathbf{A})$, but a few are off axis in plot (e)); blue squares show the right interpolation points $\{\lambda_i\}$; red diamonds show the left points $\{\mu_j\}$

Despite these large perturbations, the recovered eigenvalues are remarkably accurate: in 99.99% of cases, the eigenvalues have absolute accuracy of at least 10^{-2} , indeed more accurate than the measurements themselves. The pseudospectra in Fig. 4a suggest good robustness (though the pseudospectral level curves are pessimistic by one or two orders of magnitude). Contrast this with the complex interleaved interpolation points used in Fig. 4d. Now we only perturb the data by a small amount, $5 \cdot 10^{-9}$, for which the perturbed Loewner matrices (over 1,000 trials) satisfy

$$1.52 \cdot 10^{-7} \leq \|\widehat{\mathbb{L}}_s - \mathbb{L}_s\| \leq 2.03 \cdot 10^{-7}, \quad 2.57 \cdot 10^{-8} \leq \|\widehat{\mathbb{L}} - \mathbb{L}\| \leq 3.14 \cdot 10^{-8}.$$

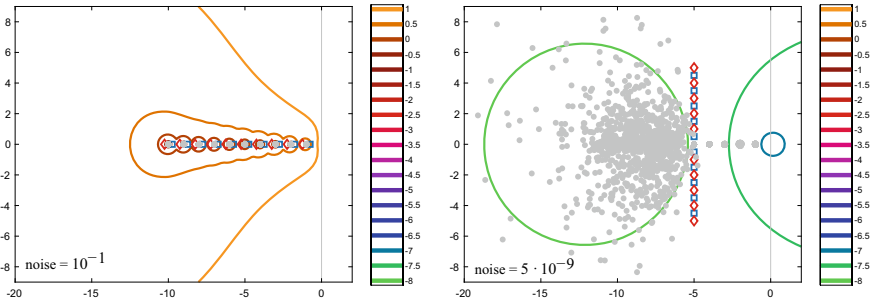


Fig. 5 Eigenvalues of the perturbed Loewner pencil $\widehat{\mathbb{L}}_s - z\widehat{\mathbb{L}}$ (gray dots), constructed from measurements that have been perturbed by random complex noise of magnitude 10^{-1} (left) and $5 \cdot 10^{-9}$ (right) (1,000 trials). As the pseudospectra in Fig. 4a, d indicate, the interleaved interpolation points on the left are remarkably stable, while the similarly interleaved complex interpolation points on the right give a Loewner pencil that is highly sensitive to small changes in the data

With this mere hint of noise, the eigenvalues of the recovered system erupt: only 36.73% of the eigenvalues are correct to two digits. (Curiously, -4 and -5 are always computed correctly, while -8 , -9 , and -10 are never computed correctly.) The pseudospectra indicate that the leftmost eigenvalues are more sensitive, and again hint at the effect of the perturbation (though off by roughly an order of magnitude in ε).³ Measurements of real systems (or even numerical simulations of nontrivial systems) are unlikely to produce such high accuracy; pseudospectra can reveal the virtue or folly of a given interpolation point configuration.

In these simple experiments, pseudospectra have been most helpful for indicating the sensitivity of eigenvalues when the left and right interpolation points are favorably

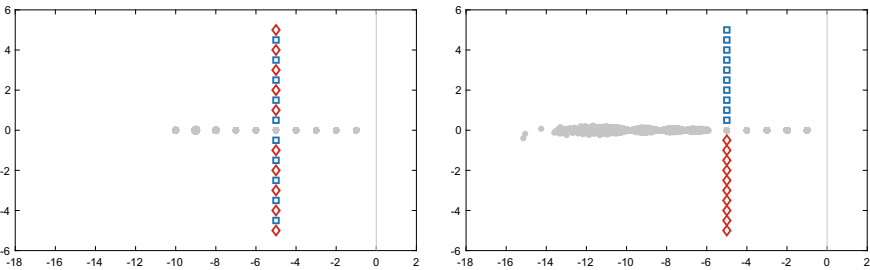


Fig. 6 Eigenvalues of the perturbed Loewner pencil $\widehat{\mathbb{L}}_s - z\widehat{\mathbb{L}}$ (gray dots), constructed from measurements that have been perturbed by random complex noise of magnitude 10^{-10} (10,000 trials). As suggested by the pseudospectra plots, the interleaved interpolation points (left) are more robust to perturbations than the separated points (right), though the difference is not as acute as suggested by Fig. 4d, f. For example, in these 10,000 trials, the least stable pole (-9) is computed accurately (absolute error less than 0.01) in 10.97% of trials on the left, and 0.29% on the right

³ One might also consider the $\varepsilon = 1$ level curve of $\sigma_\varepsilon^{(\gamma_\star, \delta_\star)}(\mathbb{L}_s, \mathbb{L})$ for $\gamma_\star = \max\{\|\mathbb{L}_s - \widehat{\mathbb{L}}_s\|\}$ and $\delta_\star = \max\{\|\mathbb{L} - \widehat{\mathbb{L}}\|\}$.

partitioned (e.g., interleaved). They seem to be less precise at predicting the sensitivity to noise of poor left/right partitions of the interpolation points. Figure 6 gives an example, based on the two partitions of the same interpolation points in Fig. 4d, f. The pseudospectra suggest that the eigenvalues for plot (f) should be much more sensitive to noise than those for the interleaved points in plot (d). In fact, the configuration in plot (f) appears to be only marginally less stable to noise of size 10^{-10} , over 10,000 trials. This is a case where one could potentially glean additional insight from *structured* Loewner pseudospectra.

7 Conclusion

Pseudospectra provide a tool for analyzing the stability of eigenvalues of Loewner matrix pencils. Elementary examples show how pseudospectra can inform the selection and partition of interpolation points, and bound the eigenvalues of Loewner pencils in the presence of noisy data. Using a different approach to pseudospectra, we showed that while the realized Loewner pencil matches the poles of the original system, it need not replicate transient dynamics of $\mathbf{E}\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t)$; pseudospectra can reveal potential transient growth, which varies with the interpolation points.

In this initial study we have intentionally used simple examples involving small, symmetric \mathbf{A} and $\mathbf{E} = \mathbf{I}$. Realistic examples, e.g., with complex poles, nonnormal \mathbf{A} , singular \mathbf{E} , multiple inputs and outputs, and rank-deficient Loewner matrices, will add additional complexity. Moreover, we have only sought to realize a system whose order is known; we have not addressed pseudospectra of the reduced pencils (7) in the context of data-driven model reduction.

Structured Loewner pseudospectra provide another avenue for future study. Structured *matrix pencil* pseudospectra have not been much investigated, especially with Loewner structure. Rump's results for standard pseudospectra [25] suggest the following problem; its positive resolution would imply that the Loewner pseudospectrum $\sigma_\varepsilon^{(\gamma, \delta)}(\mathbb{L}_s, \mathbb{L})$ matches the structured Loewner matrix pseudospectrum.

Given any $\varepsilon, \gamma, \delta > 0$, Loewner matrix \mathbb{L} and associated shifted Loewner matrix \mathbb{L}_s , suppose $z \in \sigma_\varepsilon^{(\gamma, \delta)}(\mathbb{L}_s, \mathbb{L})$. *Does there exist some Loewner matrix $\widehat{\mathbb{L}}$ and associated shifted Loewner matrix $\widehat{\mathbb{L}}_s$ such that $z \in \sigma(\widehat{\mathbb{L}}_s, \widehat{\mathbb{L}})$ and*

$$\|\widehat{\mathbb{L}}_s - \mathbb{L}_s\| < \varepsilon\gamma, \quad \|\widehat{\mathbb{L}} - \mathbb{L}\| < \varepsilon\delta \quad ?$$

Acknowledgements This work was motivated by a question posed by Thanos Antoulas, who we thank not only for suggesting this investigation, but also for his many profound contributions to systems theory and his inspiring teaching and mentorship. We also thank Serkan Gugercin and an anonymous referee for many helpful comments. (Mark Embree was supported by the U.S. National Science Foundation under grant DMS-1720257.)

Appendix

We provide a MATLAB implementation that computes the inverse iteration vectors \mathbf{u} in (12) in only $\mathcal{O}(n^2)$ operations by exploiting the Cauchy-like rank displacement structure of the Loewner pencil, as shown in (13). Namely, we start from the general $\mathcal{O}(n^3)$ MATLAB code from [29, p. 373] and modify it to account for the Loewner structure, and hence achieve $\mathcal{O}(n^2)$ efficiency. This code computes $\|(z\mathbb{L} - \mathbb{L}_s)^{-1}\|$ for a fixed z . To compute pseudospectra, one applies this algorithm on a grid of z values. In that case, the $\mathcal{O}((m+p)n^2)$ structured LU factorization in the first line need only be computed *once* for all z values (just as the standard algorithm computes an $\mathcal{O}(n^3)$ simultaneous triangularization using the QZ algorithm once for all z).

```
[L1,U1,piv1] = LUdispPiv(mu,lambda,[V' L'],[R.' -W.']);
L1t = L1'; U1t = U1';
z = 1./(z-lambda); Upz = z.*(U1\(L1\V(:,piv1)'));
[L2,U2,piv2] = lu(eye(m)-R*Upz,'vector'); R2 = R(piv2,:);
L2t = L2'; U2t = U2'; R2t = R2'; Upzt = Upz';
applyTheta = @(x) x+Upz*(U2\(L2\(R2*x)));
applyThetaTranspose = @(x) x+R2t*(L2t\(U2t\(Upzt*x)));

sigold = 0;
for it = 1:maxit
    u = U1\(L1\u(piv1));
    u = conj(z).*applyThetaTranspose(applyTheta(z.*u));
    u(piv1) = L1t\(U1t\u);
    sig = 1/norm(u);
    if abs(sigold/sig-1) < 1e-2, break, end
    u = sig*u; sigold = sig;
end
sigmin = sqrt(sig);
```

The function `LUdispPiv` computes the LU factorization (with partial pivoting) of the Loewner matrix \mathbb{L} in $\mathcal{O}((m+p)n^2)$ operations. The Loewner matrix is not formed explicitly; instead, the function uses the raw interpolation data $\lambda_i, \mathbf{r}_i, \mathbf{w}_i$ and $\mu_j, \ell_j, \mathbf{v}_j$. The implementation details for `LUdispPiv` can be found in [12, Sect. 12.1]. The LU factorization of $\mathbf{I} - \mathbf{R}\mathbf{V}^*(z) \in \mathbb{C}^{m \times m}$ is given by $L2$ and $U2$, while the first three lines of the loop represent the computation of $\mathbb{L}^{-*}\Theta(z)*\Theta(z)\mathbb{L}^{-1}\mathbf{u}$, as defined in (13). Note the careful grouping of terms and the use of elementwise multiplication `.*` to keep the total operation count at $\mathcal{O}(n^2)$.

References

1. Antoulas, A.C., Lefteriu, S., Ionita, A.C.: A tutorial introduction to the Loewner framework for model reduction. In: Benner, P., Cohen, A., Ohlberger, M., Willcox, K. (eds.) *Model Reduction and Approximation: Theory and Algorithms*, pp. 335–376. SIAM, Philadelphia (2017)
2. Antoulas, A.C., Sorensen, D.C., Zhou, Y.: On the decay rate of Hankel singular values and related issues. *Syst. Control Lett.* **46**, 323–342 (2002)
3. Beckermann, B., Townsend, A.: On the singular values of matrices with displacement structure. *SIAM J. Matrix Anal. Appl.* **38**, 1227–1248 (2017)
4. Chaitin-Chatelin, F., Frayssé, V.: *Lectures on Finite Precision Computations*. SIAM, Philadelphia (1996)
5. Chaitin-Chatelin, F., Harrabi, A.: About definitions of pseudospectra of closed operators in Banach spaces. Technical Report TR/PA/98/08, CERFACS (1998)
6. Drmač, Z., Peherstorfer, B.: Learning low-dimensional dynamical-system models from noisy frequency-response data with Loewner rational interpolation. In: C. Beattie et al. (eds.) *Realization and Model Reduction of Dynamical Systems*, pp. 39–57. Springer, Cham (2022)
7. Embree, M.: Pseudospectra (Chap. 23). In: Hogben, L. (ed.) *Handbook of Linear Algebra*, 2nd edn. CRC/Taylor & Francis, Boca Raton (2014)
8. Embree, M., Keeler, B.: Pseudospectra of matrix pencils for transient analysis of differential-algebraic equations. *SIAM J. Matrix Anal. Appl.* **38**, 1028–1054 (2017)
9. Filip, S.-I., Nakatsukasa, Y., Trefethen, L.N., Beckermann, B.: Rational minimax approximation via adaptive barycentric approximations. *SIAM J. Sci. Comput.* **40**, A2427–A2455 (2018)
10. Frayssé, V., Gueury, M., Nicoud, F., Toumazou, V.: Spectral portraits for matrix pencils. Technical Report TR/PA/96/19, CERFACS (1996)
11. Gohberg, I., Kailath, T., Olshevsky, V.: Fast Gaussian elimination with partial pivoting for matrices with displacement structure. *Math. Comp.* **64**, 1557–1576 (1995)
12. Golub, G.H., Van Loan, C.F.: *Matrix Computations*, 4th edn. Johns Hopkins University Press, Baltimore (2012)
13. Gosea, I.V., Antoulas, A.C.: Rational approximation of the absolute value function from measurements: a numerical study of recent methods (2020). [arXiv:2005.02736](https://arxiv.org/abs/2005.02736)
14. Higham, N.J., Tisseur, F.: More on pseudospectra for polynomial eigenvalue problems and applications in control theory. *Linear Algebra Appl.* **351–352**, 435–453 (2002)
15. Hinrichsen, D., Pritchard, A.J.: Real and complex stability radii: a survey. In: Hinrichsen, D., Mårtensson, B. (eds.) *Control of Uncertain Systems*. Birkhäuser, Boston (1990)
16. Horn, R.A., Johnson, C.R.: *Matrix Analysis*, 2nd edn. Cambridge University Press, Cambridge (2013)
17. Ioniă, A.C.: Lagrange rational interpolation and its applications to approximation of large-scale dynamical systems. Ph.D. Thesis, Rice University (2013)
18. Lavallée, P.-F.: *Nouvelles Approches de Calcul du ε -Spectre de Matrices et de Faisceaux de Matrices*. Ph.D. Thesis, Université de Rennes 1 (1997)
19. Mayo, A.J., Antoulas, A.C.: A framework for the solution of the generalized realization problem. *Linear Algebra Appl.* **425**, 634–662 (2007)
20. Overton, M.L.: *Numerical Computing with IEEE Floating Point Arithmetic*. SIAM, Philadelphia (2001)
21. Penzl, T.: A cyclic low-rank Smith method for large sparse Lyapunov equations. *SIAM J. Sci. Comput.* **21**, 1401–1418 (2000)
22. Penzl, T.: Eigenvalue decay bounds for solutions of Lyapunov equations: the symmetric case. *Syst. Control Lett.* **40**, 139–144 (2000)
23. Riedel, K.S.: Generalized epsilon-pseudospectra. *SIAM J. Numer. Anal.* **31**, 1219–1225 (1994)
24. Ruhe, A.: The rational Krylov algorithm for large nonsymmetric eigenvalues — mapping the resolvent norms (pseudospectrum). Unpublished manuscript, March 1995
25. Rump, S.M.: Eigenvalues, pseudospectrum and structured perturbations. *Linear Algebra Appl.* **413**, 567–593 (2006)

26. Sabino, J.: Solution of large-scale Lyapunov equations via the block modified Smith method. Ph.D. Thesis, Rice University (2006)
27. Tisseur, F., Higham, N.J.: Structured pseudospectra for polynomial eigenvalue problems, with applications. *SIAM J. Matrix Anal. Appl.* **23**, 187–208 (2001)
28. Trefethen, L.N.: Computation of pseudospectra. *Acta Numer.* **8**, 247–295 (1999)
29. Trefethen, L.N., Embree, M.: *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press, Princeton (2005)
30. Trefethen, L.N., Trefethen, A.E., Reddy, S.C., Driscoll, T.A.: Hydrodynamic stability without eigenvalues. *Science* **261**, 578–584 (1993)
31. Wright, T.G.: Algorithms and software for pseudospectra. D.Phil. Thesis, Oxford University (2002)

A Loewner Matrix Approach to the Identification of Linear Time-Varying Systems



Paolo Rapisarda

Abstract We show that if enough “sufficiently informative” input-output trajectories generated by a linear, time-varying, finite-dimensional system and its dual are given, then *state* trajectories corresponding to them can be computed by factorizing a time-varying matrix directly constructed from the data. From such input-state-output trajectories an “unfalsified” linear time-varying model can be obtained solving a system of functional equations. Our approach is particularly relevant when the data-producing system is self-dual (e.g. if it is conservative).

Keywords Interpolation · Loewner matrices · Duality · Time-varying linear systems

1 Introduction

It is a pleasure for me to contribute to Prof. Antoulas’ 70th birthday *Festschrift*. Thanos was a member of the Reading Committee of my Ph.D. dissertation, and I vividly recall meeting him for the first time on the occasion of my Ph.D. defence. His personal warmth and infectious enthusiasm for research have been a pleasant ingredient of all our interactions, and the elegance and powerful simplicity of his work have strongly influenced me in all stages of my career.

The present contribution is an *hommage* to some of Thanos’s own pioneering ideas. In his work on rational interpolation and data modeling (see [1–5]) and part of his *opus* about model-order reduction (see e.g. [6]) he introduced the concept of mirroring of vector-exponential trajectories, and the all-important Loewner matrix. Many years later, Thanos and I showed in [7–9] that such concepts can also be expressed in the language of bilinear- and quadratic differential forms, and pointed out the role played by the concept of *duality* of trajectories and of systems in the Loewner approach. Such intuition makes explicit the connection between the approach to data modeling

P. Rapisarda (✉)

School of Electronics and Computer Science, University of Southampton, Southampton, UK
e-mail: pr3@ecs.soton.ac.uk

© Springer Nature Switzerland AG 2022

C. Beattie et al. (eds.), *Realization and Model Reduction of Dynamical Systems*,
https://doi.org/10.1007/978-3-030-95157-3_5

and model order reduction that I have independently pursued with various co-authors (see [10–15]) and Thanos’s own.

With Thanos I applied Loewner and duality concepts to the modelling of two-dimensional vector-exponential trajectories and of parametric systems (see [16–18]). In this paper I present a further application of ideas inspired by his Loewner approach. Building on previous work by myself and collaborators (see [19, 20]), I consider the modelling of input-output data produced by linear, time-varying systems. In doing this I will emphasise the essential coherence of my recent and ongoing work with Thanos’s original idea, thus paying a modest, but hopefully fitting tribute to his far-sighted intellect and to the strength of his intuition.

2 Problem Statement

We are given N (a fixed, “large” number) of input-output trajectories

$$\begin{bmatrix} u_k(\cdot) \\ y_k(\cdot) \end{bmatrix} : \mathbb{R} \rightarrow \mathbb{R}^{m+p}, \quad k = 1, \dots, N, \quad (1)$$

generated by an unknown linear, time-varying state-space system

$$\begin{aligned} \frac{d}{dt}x(\cdot) &= A(\cdot)x(\cdot) + B(\cdot)u(\cdot) \\ y(\cdot) &= C(\cdot)x(\cdot) + D(\cdot)u(\cdot). \end{aligned} \quad (2)$$

In (2) the matrix functions $A(\cdot)$, $B(\cdot)$, $C(\cdot)$, $D(\cdot)$ are respectively $n \times n$, $n \times m$, $p \times n$ and $p \times m$, with analytic entries. In the following we denote the space of $j \times k$ matrices with real analytic entries by $\mathcal{A}^{j \times k}$. In the rest of this paper we assume that the representation (2) is *controllable* and *observable*, see [25].

We are also given N input-output trajectories

$$\begin{bmatrix} u'_k(\cdot) \\ y'_k(\cdot) \end{bmatrix} : \mathbb{R} \rightarrow \mathbb{R}^{p+m}, \quad k = 1, \dots, N, \quad (3)$$

of the *dual* system of (2), a state-space system with state variable x' associated with analytic matrix functions $(A'(\cdot), B'(\cdot), C'(\cdot), D'(\cdot))$ defining similar equations as (1), such that the following property holds.

Definition 1 (*Duality*) Two linear, time-varying state-space systems are *dual* if there exists a $n \times n$ matrix-function $Q(\cdot)$ with $\det Q(t) \neq 0$ for all $t \in \mathbb{R}$, such that for every pair of trajectories (x, u, y) satisfying (2) and (x', u', y') satisfying the equations defined by $(A'(\cdot), B'(\cdot), C'(\cdot), D'(\cdot))$ it holds that

$$u^\top y' + y^\top u' = \frac{d}{dt} (x^\top Q x'). \quad (4)$$

In the following we call (4) the *duality* relation.

Remark 1 Definition 1 is analogous to the characteristic property of *adjoint* non-linear system stated in Lemma 2.1 of [21]; see also [22] for the definition of this concept in the linear, time-invariant case, and Definition 1 and Proposition 1 p. 605 of [23] for the linear, time-varying case. It is a *power-conservation relation* such as that occurring in lossless systems. Consider for example the relation between the voltage and current pairs at the ports of LC electrical circuits (where the bilinear product on the left-hand side of (4) is electrical power), and the capacitor voltages and inductor currents (where the bilinear product on the right-hand side of (4) is the stored energy). Power-conservation relations are also at the core of the notion of port-Hamiltonian system (see [24]) and the approach presented here can also be applied to linear, time-varying, *dissipative* systems, provided that the bilinear form associated with the dissipation rate is known (see [20]). In Sect. 6 of this paper we illustrate several current and future research directions related to this aspect of our work.

We want to identify from the data (1) and (3) an *unfalsified* i-s-o model for the primal system, i.e. matrices $\widehat{A}(\cdot)$, $\widehat{B}(\cdot)$, $\widehat{C}(\cdot)$, $\widehat{D}(\cdot)$ such that input-state-output Eq. (2) hold for *some* (to be computed) state trajectories $x_k(\cdot)$, $k = 1, \dots, N$ and the given $(u_k(\cdot), y_k(\cdot))$.

In this problem formulation it is assumed that data from both the primal system (2) and from its dual are known. While we recognize that it is unrealistic that measurements from the dual system are available, in order to emphasize the connections with the Loewner framework we shall find it convenient to work on the problem as stated. It is worthwhile to remark that for many classes of systems, *self-duality* is either implied by energy conservation or by other physical properties, for example those implied by port-Hamiltonicity.

3 A Loewner Matrix for Time-Varying Systems

We begin with the following fundamental result.

Proposition 1 *Let $t \in \mathbb{R}$, and let (x, u, y) and (x', u', y') be trajectories of the primal, respectively of the dual system. Assume that $x(-\infty) := \lim_{t \rightarrow -\infty} x(t) = 0$ and $x'(-\infty) := \lim_{t \rightarrow -\infty} x'(t) = 0$; then*

$$\int_{-\infty}^t u(\tau)^\top y'(\tau) + y(\tau)^\top u'(\tau) d\tau = x(t)^\top Q(t)x'(t) = x(t)^\top Q(t)x(t) . \quad (5)$$

Proof Integrating both sides of (4), obtain

$$\int_{-\infty}^t u(\tau)^\top y'(\tau) + y(\tau)^\top u'(\tau) d\tau = x(t)^\top Q(t)x'(t) - x(-\infty)^\top Q(0)x'(-\infty) .$$

Now apply the assumption $x(-\infty) = 0 = x'(-\infty)$. \square

Definition 2 Assume that for every $t \in \mathbb{R}$ all trajectories (x_k, u_k, y_k) and (x'_i, u'_i, y'_i) of the data (1), (3) satisfy $x_k(-\infty) = 0 = x'_i(-\infty)$. The *time-varying Loewner matrix* associated with the data (1), (3) is the $N \times N$ symmetric matrix function $E(\cdot) : (-\infty, t] \rightarrow \mathbb{R}^{N \times N}$ defined by

$$[E(t)]_{i,k=1,\dots,N} := \int_{-\infty}^t u_k(\tau)^\top y'_i(\tau) + y_k(\tau)^\top u'_i(\tau) d\tau . \quad (6)$$

3.1 Relation with the Classical Loewner Matrix

In order to justify the terminology introduced in Definition 2, let us consider the case in which the data-generating system (and its dual) are *time-invariant*, and the primal and dual data consists of *vector-exponential* trajectories:

$$\begin{bmatrix} u'_i(t) \\ y'_i(t) \end{bmatrix} =: \begin{bmatrix} \bar{u}'_i \\ \bar{y}'_i \end{bmatrix} e^{\mu_i t} \quad \text{and} \quad \begin{bmatrix} u_k(\cdot) \\ y_k(\cdot) \end{bmatrix} =: \begin{bmatrix} \bar{u}_k \\ \bar{y}_k \end{bmatrix} e^{\lambda_k t} , \quad (7)$$

with \bar{u}'_i, \bar{y}'_i are p -dimensional constant real vectors, \bar{y}'_i, \bar{u}_k are m -dimensional constant real vectors, and μ_i, λ_k are positive real numbers, $i, k = 1, \dots, N$. The case of complex vectors and exponentials is easily dealt with, but the notation is slightly more involved and we do not consider it here. Note that the input-output trajectories (7) satisfy the condition $x_k(-\infty) = 0 = x'_i(-\infty)$, $i, k = 1, \dots, N$. In the rest of this chapter we assume that

$$\lambda_k + \mu_i \neq 0 , \quad i, k = 1, \dots, N .$$

Proposition 2 Assume that the data-generating system and its dual are time-invariant, and that the data is of the form (7), with $\bar{u}'_i, \bar{y}'_i \in \mathbb{R}^p$, $\bar{y}'_i, \bar{u}_k \in \mathbb{R}^m$, and $\mu_i, \lambda_k \in \mathbb{R}_+$, $i, k = 1, \dots, N$. Then

$$[E(t)]_{i,k} = e^{(\lambda_k + \mu_i)t} \frac{\bar{u}_k^\top \bar{y}'_i + \bar{y}_k^\top \bar{u}'_i}{\lambda_k + \mu_i} \quad (8)$$

Proof From the definition of the data, for every $k, i = 1, \dots, N$ it holds that

$$\begin{aligned} E_{k,i}(t) &= \int_{-\infty}^t e^{\lambda_k \tau} \bar{u}_k^\top \bar{y}'_i e^{\mu_i \tau} + e^{\lambda_k \tau} \bar{y}_k^\top \bar{u}'_i e^{\mu_i \tau} d\tau = \int_{-\infty}^t e^{(\lambda_k + \mu_i)\tau} (\bar{u}_k^\top \bar{y}'_i + \bar{y}_k^\top \bar{u}'_i) d\tau \\ &= \frac{e^{(\lambda_k + \mu_i)t}}{\lambda_k + \mu_i} \Big|_{-\infty}^t (\bar{u}_k^\top \bar{y}'_i + \bar{y}_k^\top \bar{u}'_i) = e^{(\lambda_k + \mu_i)t} \frac{\bar{u}_k^\top \bar{y}'_i + \bar{y}_k^\top \bar{u}'_i}{\lambda_k + \mu_i} . \end{aligned}$$

□

Denote the transfer function of the primal system by $Y(s)$, and let $Y(s) = N(s)D(s)^{-1} = P(s)^{-1}Q(s)$ be respectively a right-coprime and a left-coprime factorization of $Y(s)$. The following result holds.

Proposition 3 *The transfer function of the dual is*

$$Y'(s) := -Y(-s)^\top = -D(-s)^{-\top}N(-s)^\top = -Q(-s)^\top P(-s)^{-\top} .$$

Proof The proof follows from the material in Sect. 10 of [26], in particular from Proposition 10.1 p. 1730. □

Observe that for $t = 0$ the result of Proposition 2 implies that

$$[E(0)]_{i,k} = \frac{\bar{u}_k^\top \bar{y}'_i + \bar{y}_k^\top \bar{u}'_i}{\lambda_k + \mu_i} , i, k = 1, \dots, N . \tag{9}$$

Assume now that $m = p = 1$, i.e. the single-input, single-output case (the general case is considered in Sect. 2.3 of [7]); then one can assume without loss of generality that $\bar{u}_k = 1 = \bar{u}'_i, i, k = 1, \dots, N$. It follows that the matrix $E(0)$ in (9) is the Loewner matrix \mathbb{L} associated with the interpolation data

$$Z = \{-\mu_i\}_{i=1,\dots,N} \cup \{\lambda_k\}_{k=1,\dots,N} \text{ and } Y = \{y(-\mu_i)\}_{i=1,\dots,N} \cup \{y(\lambda_k)\}_{k=1,\dots,N} ,$$

see p. 639 of [4].

Moreover, note that in the time-invariant case, since for each pair $(i, k), i, k = 1, \dots, N$ the associated frequencies $-\mu_i$ and λ_k are known, the Loewner matrix (9) and its time-varying version (8) embody the same information. Observe also that from the result of Proposition 3 it follows that the trajectories of the dual system associated with a frequency μ_i can be computed by *mirroring*: namely, if

$$\begin{bmatrix} u_i \\ y_i \end{bmatrix} \in \mathbb{C}^{m+p}$$

arises from the primal transfer function evaluated at $-\mu_i$, then

$$\begin{bmatrix} u'_i \\ y'_i \end{bmatrix} := \begin{bmatrix} y_i \\ u_i \end{bmatrix} \in \mathbb{C}^{p+m}$$

arises from the dual transfer function evaluated at μ_i . It follows that in the linear, time-invariant case there is no need to assume that the dual trajectories are available from measurements of the dual system, since they can be generated by mirroring.

4 Properties of the Time-Varying Loewner Matrix

In this section we follow the structure of Sect. 3: we first establish some properties of the time-varying Loewner matrix (6), and subsequently we show that they are generalisations of well-known properties of the time-invariant case established by Thanos and his collaborators.

The first result is a straightforward consequence of the definition.

Proposition 4 *Let (x_k, u_k, y_k) and (x'_i, u'_i, y'_i) , $i, k = 1, \dots, N$ be i -s-o trajectories of the primal, respectively dual system. Assume that all trajectories (x_k, u_k, y_k) and (x'_i, u'_i, y'_i) satisfy the condition $x_k(-\infty) = 0 = x'_i(-\infty)$, $i, k = 1, \dots, N$. Then for all $t \in \mathbb{R}$ it holds that*

$$E(t) = \underbrace{\begin{bmatrix} x'_1(t)^\top \\ \vdots \\ x'_N(t)^\top \end{bmatrix}}_{=: X'(t)^\top} Q(t) \underbrace{[x_1(t) \dots x_N(t)]}_{=: X(t)}. \quad (10)$$

Proof The result follows in a straightforward way from Definition 2 and from Proposition 1. \square

An important consequence of Proposition 4 is that for “almost all choices” of the input-output trajectories (u_k, y_k) and (u'_i, y'_i) , $i, k = 1, \dots, N$, and “almost all” $t \in \mathbb{R}$ the time-varying Loewner matrix has rank n , the dimension of the state space of the primal and dual system. Before proving this result, we need to formalize the concept of “almost all choices”; in order to do this, we use the algebraic notion of *algebraic genericity*.

Let \mathcal{L} be some linear space of finite-dimension d ; then given a basis $\{\ell_i\}_{i=1, \dots, d}$ for \mathcal{L} , every $\ell \in \mathcal{L}$ can be written as $\ell = \sum_{i=1}^d x_i \ell_i$ for some coefficients x_i in the field on which \mathcal{L} is defined. A map $p : \mathcal{L} \rightarrow \mathbb{R}$ is a *polynomial* if $p(\ell)$ is a polynomial in the variables x_i , $i = 1, \dots, d$. An *algebraic variety* is a subset \mathcal{V} of \mathcal{L} consisting of all zeroes of some polynomial p . A subset $\mathcal{S} \subset \mathcal{L}$ is called *generic* if there is a proper algebraic variety $\mathcal{V} \subsetneq \mathcal{L}$ such that $\mathcal{S} \supset (\mathcal{L} \setminus \mathcal{V})$.

With these definitions, we can now state the following important result.

Theorem 1 *Assume that $N > n$; then for all $t \in \mathbb{R}$ it holds generically that $\text{rank } E(t) = \text{rank } E_{11}(t) = n$.*

Proof We show that $\text{rank } E(t) = n$; the equality $\text{rank } E(t) = \text{rank } E_{11}(t)$ is a straightforward consequence of the fact that $\det E_{11}(t)$ is one of the $n \times n$ minors of the $N \times N$ rank n matrix $E(t)$. This minor is generically nonzero if $\text{rank } E(t) = n$.

Conclude from Eq. (10) and Definition 1 that

$$\text{rank } E(t) \leq \min \{ \text{rank } [x_1(t) \dots x_N(t)], \text{rank } [x'_1(t) \dots x'_N(t)] \} .$$

We now show that generically $\text{rank} [x_1(t) \dots x_N(t)] = n$; an analogous argument yields that generically also $\text{rank} [x'_1(t) \dots x'_N(t)] = n$. Using the genericity assumption we will then prove the claim.

In order to show that generically $\text{rank} [x_1(t) \dots x_N(t)] = n$, we prove the following, stronger result, which will be useful later. \square

Lemma 1 *Let $\begin{bmatrix} x_i(\cdot) \\ u_i(\cdot) \end{bmatrix}$, $i = 1, \dots, N > n + m$ be trajectories satisfying the state equation in (2). Then generically for all $t \in \mathbb{R}$ it holds that*

$$\text{rank} \begin{bmatrix} x_1(t) \dots x_N(t) \\ u_1(t) \dots u_N(t) \end{bmatrix} = n + m$$

Proof To simplify the argument, we use a nonsingular transformation of the state-space basis as in [25] to transform the Eq. (2) to a more manageable version. Denote by $X(\cdot)$ a fundamental matrix solution of the free response, i.e. a matrix having full rank almost everywhere, such that solves the matrix differential equation $\frac{d}{dt} X(t) = A(t)X(t)$ for every $t \in \mathbb{R}$. Now let τ be a fixed, but otherwise arbitrary real number, and denote by $\phi(t, \tau)$ the transition matrix of (2) in the interval $[\tau, t]$, i.e. $\phi(t, \tau) = X(t)X(\tau)^{-1}$. Apply the transformation

$$x(\cdot) \rightarrow \phi(\cdot, \tau)^{-1}x(\cdot) = \phi(\tau, \cdot)x(\cdot) =: z(\cdot),$$

to the state variable x , and verify with straightforward manipulations that in this new basis of the state space, the state equation in (2) becomes

$$\frac{d}{dt}z(\cdot) = \phi(\tau, \cdot)B(\cdot)u(\cdot). \quad (11)$$

We now prove that generically

$$\text{rank} \begin{bmatrix} z_1(t) \dots z_N(t) \\ u_1(t) \dots u_N(t) \end{bmatrix} = \begin{bmatrix} \phi(\tau, t) & 0_{n \times m} \\ 0_{m \times n} & I_m \end{bmatrix} \begin{bmatrix} x_1(t) \dots x_N(t) \\ u_1(t) \dots u_N(t) \end{bmatrix} = n + m; \quad (12)$$

note that such equality implies the claim of the Lemma, since $\det \begin{bmatrix} \phi(\tau, t) & 0_{n \times m} \\ 0_{m \times n} & I_m \end{bmatrix} \neq 0$.

To prove (12), we use the concept of *controllability* matrix of the matrix pair $(A(\cdot), B(\cdot))$ (see p. 66 of [25]). Define the sequence of $n \times m$ matrices

$$P_0(\cdot) := B(\cdot) \\ P_{k+1}(\cdot) := -A(\cdot)P_k(\cdot) + \frac{d}{dt}(P_k(\cdot));$$

and from such sequence define the $n \times n \cdot m$ *controllability* matrix of the matrix pair $(A(\cdot), B(\cdot))$ by

$$C(A(\cdot), B(\cdot)) := [P_0(\cdot) \ P_1(\cdot) \ \dots \ P_{n-1}(\cdot)] . \quad (13)$$

It is straightforward to verify that from (11) it follows that

$$\frac{d^k z}{dt^k}(\cdot) = \phi(\tau, \cdot) [P_0(\cdot) \ \dots \ P_{k-1}(\cdot)] \begin{bmatrix} \frac{d^{k-1}u}{dt^{k-1}}(\cdot) \\ \vdots \\ u(\cdot) \end{bmatrix} ,$$

$k = 1, \dots$ The Wronskian matrix of $\text{col}(z(\cdot), u(\cdot))$ (see p. 67 of [25]) is defined by

$$\mathcal{W}(z(\cdot), u(\cdot)) := \begin{bmatrix} z(\cdot) & \frac{dz}{dt}(\cdot) & \dots & \frac{d^{n+m-1}z}{dt^{n+m-1}}(\cdot) \\ u(\cdot) & \frac{d}{dt}u(\cdot) & \dots & \frac{d^{n+m-1}u}{dt^{n+m-1}}(\cdot) \end{bmatrix} .$$

Note that

$$\begin{aligned} & \mathcal{W}(z(\cdot), u(\cdot)) & (14) \\ & = \begin{bmatrix} z(\cdot) & \phi(\tau, \cdot)P_0(\cdot) & \phi(\tau, \cdot)P_1(\cdot) & \dots & \phi(\tau, \cdot)P_{n+m-1}(\cdot) \\ 0 & I_n & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} I_n & 0 & 0 & \dots & 0 \\ 0 & u(\cdot) & \frac{du}{dt}(\cdot) & \dots & \frac{d^{n+m-1}u}{dt^{n+m-1}}(\cdot) \\ 0 & 0 & u(\cdot) & \dots & \frac{d^{n+m-2}u}{dt^{n+m-2}}(\cdot) \\ 0 & 0 & 0 & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & u \end{bmatrix} . \end{aligned}$$

Since $u(\cdot)$ is arbitrarily chosen in the space of m -dimensional time functions, generically the second matrix on the right-hand side of (14) has full column rank at t . The first n rows of the first matrix on the right-hand side of (14) contain as submatrix a nonsingular transformation (via $\phi(\tau, \cdot)$) of the controllability matrix (13), which has full row rank n on a set of points everywhere dense on any finite interval of \mathbb{R} (see Theorem 4 p. 69 of [25]). Consequently, it has generically full row rank.

From this argument it follows that for every $i = 1, \dots, n$, at every $t \in \mathbb{R}$ generically the Wronskian $\mathcal{W}(\text{col}(z_i(t), u_i(t)))$ has rank $n + m$. It follows that generically the Wronskian of $\begin{bmatrix} z_1(t) & \dots & z_N(t) \\ u_1(t) & \dots & u_N(t) \end{bmatrix}$ also has rank $n + m$. \square

We resume the proof of Theorem 1. Lemma 1 implies that the trajectories $z_i(\cdot)$ are linearly independent (see Lemma 3 p. 68 of [25]). Since they are obtained from a nonsingular transformation of the state trajectories $x_i(\cdot)$ of the primal system, it follows that generically

$$\text{rank} [x_1(t) \ \dots \ x_N(t)] = n .$$

A result analogous to that of Lemma 1 can be established for the matrix constructed from the state and input trajectories of the dual system, namely that generically

$$\text{rank} \begin{bmatrix} x'_1(t) & \dots & x'_N(t) \\ u'_1(t) & \dots & u'_N(t) \end{bmatrix} = n + m ,$$

from which it follows that generically

$$\text{rank} [x'_1(t) \dots x'_N(t)] = n .$$

Now observe that $\text{rank } E(t) < n$ if and only if

$$\text{im} ([x'_1(t) \dots x'_N(t)])^\perp \cap \text{im} ([x_1(t) \dots x_N(t)]) \neq \{0\} ,$$

which is generically satisfied. Consequently, generically also $\text{rank } E(t) = n$, as was to be proved. \square

The following result is a straightforward consequence of Propositions 4 and 1.

Corollary 1 *Define*

$$\begin{aligned} U(t) &:= [u_1(t) \dots u_N(t)] , & Y(t) &:= [y_1(t) \dots y_N(t)] \\ U'(t) &:= [u'_1(t) \dots u'_N(t)] , & Y'(t) &:= [y'_1(t) \dots y'_N(t)] . \end{aligned}$$

Then $E(\cdot)$ satisfies the matrix differential equation

$$\frac{d}{dt} E(\cdot) = [U'(\cdot)^\top \ Y'(\cdot)^\top] Q(\cdot) \begin{bmatrix} U(\cdot)^\top \\ Y(\cdot)^\top \end{bmatrix} \quad (15)$$

Proof The claim is a straightforward consequence of Propositions 4 and 1. \square

The result of Proposition 4 and Theorem 1 can be used to compute an *unfalsified* first-order linear, time-varying model for the primal and dual data.

Proposition 5 *The trajectories* $\begin{bmatrix} x_k \\ u_k \\ y_k \end{bmatrix}$, $k = 1, \dots, N$ *satisfy the first-order equations*

$$K(t) \frac{d}{dt} X(t) + F(t) X(t) + G(t) \begin{bmatrix} Y(t) \\ U(t) \end{bmatrix} = 0 , \quad (16)$$

where

$$\begin{aligned} K(\cdot) &:= X'(\cdot)^\top Q(\cdot) \\ F(\cdot) &:= \left(\frac{d}{dt} X'(\cdot) \right)^\top Q(\cdot) + X'(\cdot)^\top \left(\frac{d}{dt} Q(\cdot) \right) \\ G(\cdot) &:= [-U'(\cdot)^\top \ -Y'(\cdot)^\top] . \end{aligned}$$

Moreover, generically for all $t \in \mathbb{R}$ the matrix $K(t) = X'(t)^\top Q(t)$ has a left inverse $K(t)^\dagger$, and the trajectories $\begin{bmatrix} x_k \\ u_k \\ y_k \end{bmatrix}$, $k = 1, \dots, N$ also satisfy the equations

$$\frac{d}{dt} X(t) = [-K(t)^\dagger F(t)] X(t) - K(t)^\dagger G(t) \begin{bmatrix} Y(t) \\ U(t) \end{bmatrix}.$$

Proof Equation (16) follows by differentiating the left-hand side of the Eq. (15), using the equality (10).

The claim on the generic existence of a left inverse for $K(t)$ follows from the nonsingularity of $Q(t)$, and the fact that $X'(t)$ has generically full column rank because of the dual version of Lemma 1. The claim follows pre-multiplying both sides of (16) with such left-inverse matrix. \square

4.1 Relation with the Time-Invariant Case

We consider the linear, time-invariant case, with the same notation and under the same assumptions as in Sect. 3.1, and we review the results illustrated in the linear, time-varying case.

The result of Proposition 4 follows in a straightforward way from an argument similar to that used in Proposition 10.1 p. 1730; note that the assumption therein is that the bases of the state-space of the primal and of the dual system are *matched*, i.e. that $Q(t) = I_n$. In the time-invariant case for every $k, i = 1, \dots, N$ there exist constant vectors \bar{x}_k, \bar{x}'_i such that

$$x_k(t) = \bar{x}_k e^{\lambda_k t} \text{ and } x'_i(t) = \bar{x}'_i e^{\mu_i t},$$

and consequently (10) can be written as

$$E(t) = \underbrace{\text{diag} (e^{\mu_i t})_{i=1, \dots, N}}_{=: X'(t)^\top} \begin{bmatrix} \bar{x}'_1 \\ \vdots \\ \bar{x}'_N \end{bmatrix} \underbrace{Q[\bar{x}_1 \dots \bar{x}_N]}_{=: X(t)} \text{diag} (e^{\lambda_k t})_{k=1, \dots, N}. \quad (17)$$

The result of Theorem 1 implies, for $t = 0$, that generically $E(0) = \mathbb{L}$ has rank n ; in the classical Loewner framework this conclusion was proved in Lemma 2.1 p. 639 of [4].

The result of Proposition 5 is equivalent to that of Proposition 3.1 p. 641 of [4]. Indeed, for the time-invariant case it holds that

$$\begin{aligned} \frac{d}{dt} \left(X'(t)^\top \right) &= \frac{d}{dt} \left(\text{diag} \left(e^{\mu_i t} \right)_{i=1, \dots, N} \begin{bmatrix} \bar{x}'_1{}^\top \\ \vdots \\ \bar{x}'_N{}^\top \end{bmatrix} \right) = \underbrace{\text{diag} (\mu_i)_{i=1, \dots, N}}_{=:M} X'(t)^\top \\ \frac{d}{dt} (X(t)) &= \frac{d}{dt} \left([\bar{x}_1 \dots \bar{x}_N] \text{diag} \left(e^{\lambda_k t} \right)_{k=1, \dots, N} \right) = X(t) \underbrace{\text{diag} (\lambda_k)_{k=1, \dots, N}}_{=:A}, \end{aligned}$$

and the “shifted Loewner matrix” equals $\frac{d}{dt} E(t) |_{t=0}$.

5 From Loewner Matrix to State Equations

In the classical, time-invariant approach, the Loewner matrix is used to compute a transfer-function model for the data (i.e. a rational interpolant). If \mathbb{L} is of full rank this can be performed directly (see e.g. Theorem 4.1 p. 642 of [4]); in the general case, the factorization of a matrix computed from the Loewner one is necessary (see assumption (20) p. 645 and Theorem 5.1 p. 646 of [4]). We illustrated analogous procedures in [13–16, 18]. We now show that also in the time-varying case similar results hold.

First, some terminology. Given $E(\cdot) : \mathbb{R} \rightarrow \mathbb{R}^{N \times N}$, we call $E(\cdot) = Z'(\cdot)^\top R(\cdot) Z(\cdot)$ a *rank-revealing* factorisation if $Z' : \mathbb{R} \rightarrow \mathbb{R}^{r \times N}$, $Z : \mathbb{R} \rightarrow \mathbb{R}^{r \times N}$, $R : \mathbb{R} \rightarrow \mathbb{R}^{r \times r}$, with $r := \text{rank } E(t) = \text{rank } Z'(t) = \text{rank } Z(t) = \text{rank } R(t)$ on a set of points everywhere dense on any finite interval of \mathbb{R} .

We now show that minimal state trajectories corresponding to given input-output data can be computed from a rank-revealing factorisation of the Loewner matrix.

Theorem 2 *Let (2) be a minimal representation of a LTV system and let (u_k, y_k, x_k) satisfy (2), $k = 1, \dots, N$. Assume that all trajectories (x_k, u_k, y_k) and (x'_i, u'_i, y'_i) are such that $x_k(-\infty) = 0 = x'_i(-\infty)$, and define $E(\cdot)$ by (6). Assume that $\text{rank } E(t) = n$ for all $t \in \mathbb{R}$. Define*

$$\mathcal{B} := \left\{ \begin{bmatrix} u \\ y \end{bmatrix} \mid \exists x \text{ such that (2) holds} \right\}.$$

Let $E(\cdot) = Z'(\cdot)^\top R(\cdot) Z(\cdot)$ be a rank-revealing factorisation of $E(\cdot)$. Denote the k th column of Z by z_k ; then there exists an i - s - o representation with state variable z such that (u_k, y_k, z_k) satisfies the equations, $k = 1, \dots, N$.

Proof The claim is evidently correct for the factorisation (10) of $E(t)$ in Proposition 4. Use a standard linear-algebraic argument to conclude that for every $t \in \mathbb{R}$, any rank-revealing factorisation $E(t) = Z'(t)^\top R(t) Z(t)$ is related to (10) by

$$Z(t) = T(t) X(t) \text{ and } Z'(t) = T'(t) Z(t),$$

where $T, T'(\cdot) : \mathbb{R} \rightarrow \mathbb{R}^{n \times n}$ are nonsingular almost everywhere, and $T'(t)^\top R(t)T(t) = R(t)$. This proves that the k th column z_k of $Z(\cdot)$ is a state trajectory corresponding to (u_k, y_k) , $k = 1, \dots, N$. It can be verified that the i-s-o representation (2) corresponding to such state variable is induced by the matrices

$$\left(T(\cdot)(A(\cdot)T(\cdot)^{-1} - \frac{d}{dt}(T(\cdot)^{-1}), T(\cdot)B(\cdot), C(\cdot)T(\cdot)^{-1}, D(\cdot) \right).$$

□

In the time-invariant case, matrix factorization can be performed with standard numerical algorithms (e.g. the singular value decomposition); however, in the time-varying case, the factorization of a *matrix function* of dimension N is necessary. This is a non-trivial problem with considerable implications of a theoretical and computational nature. We now show that in the generic case, the result of Theorem 1 can be used to achieve a factorization of $E(\cdot)$ by inverting its $(1, 1) n \times n$ submatrix. As in Theorem 1, we partition the Loewner matrix as

$$E(t) = \begin{bmatrix} E_{11}(t) & E_{12}(t) \\ E_{21}(t) & E_{22}(t) \end{bmatrix}, \quad (18)$$

where $E_{11}(t)$ is $n \times n$, with $n = \text{rank}(E(t))$, $E_{12}(t)$ is $n \times (N - n)$, $E_{21}(t)$ is $(N - n) \times n$, and $E_{22}(t)$ is $(N - n) \times (N - n)$. The following result holds.

Theorem 3 *Let $E(\cdot)$ be partitioned as in (18); generically at t the following equation holds:*

$$E(t) = \begin{bmatrix} I_n & \\ E_{21}(t)E_{11}(t)^{-1} & \end{bmatrix} E_{11}(t) \begin{bmatrix} I_n & E_{11}(t)^{-1}E_{12}(t) \end{bmatrix}. \quad (19)$$

Proof The fact that $E_{11}(t)$ is generically nonsingular is stated in Theorem 1; consequently, the claim follows if we prove that $E_{22}(t) = E_{21}(t)E_{11}(t)^{-1}E_{12}(t)$. Consider the equality

$$\begin{aligned} & \begin{bmatrix} I_n & 0_{n \times N} \\ -E_{21}(t)^{-1}E_{11}(t) & I_{N-n} \end{bmatrix} \begin{bmatrix} E_{11}(t) & E_{12}(t) \\ E_{21}(t) & E_{22}(t) \end{bmatrix} \begin{bmatrix} I_n & -E_{11}(t)^{-1}E_{12}(t) \\ 0_{(N-n) \times n} & I_{N-n} \end{bmatrix} \\ &= \begin{bmatrix} E_{11}(t) & 0_{n \times n} \\ 0_{(N-n) \times n} & E_{22}(t) - E_{21}(t)E_{11}(t)^{-1}E_{12}(t) \end{bmatrix}. \end{aligned}$$

Now apply Theorem 1 to conclude that since the matrix on the right-hand side of the equation has rank $n = \text{rank}(E_{11}(t))$, the $(2, 2)$ -block must be zero. □

From Theorems 2 and 3 it follows that the k th column of any of the two right-factors

$$\begin{bmatrix} E_{11}(\cdot) & E_{12}(\cdot) \end{bmatrix} \text{ or } \begin{bmatrix} I_n & E_{11}(\cdot)^{-1}E_{12}(\cdot) \end{bmatrix}$$

of $E(\cdot)$ can be used to associate a state trajectory to the input-output data $\begin{bmatrix} u_k(\cdot) \\ y_k(\cdot) \end{bmatrix}$; without loss of generality in the following we discuss on the basis of the first choice. The derivatives of these state trajectories are the columns of the matrix

$$\left[\frac{d}{dt} E_{11}(\cdot) \quad \frac{d}{dt} E_{12}(\cdot) \right];$$

define also the matrices

$$\begin{aligned} U(\cdot) &:= [u_1(\cdot) \dots u_N(\cdot)] =: [U_1(\cdot) \ U_2(\cdot)] \\ Y(\cdot) &:= [y_1(\cdot) \dots y_N(\cdot)] =: [Y_1(\cdot) \ Y_2(\cdot)], \end{aligned}$$

with U_1 a $m \times n$ matrix function, U_2 $m \times (N - n)$, Y_1 $p \times n$, and Y_2 $p \times (N - n)$. Theorem 2 implies that matrix functions $\widehat{A}(\cdot)$, $\widehat{B}(\cdot)$, $\widehat{C}(\cdot)$, $\widehat{D}(\cdot)$ exist such that the following equation holds true:

$$\begin{bmatrix} \frac{d}{dt} E_{11}(\cdot) & \frac{d}{dt} E_{12}(\cdot) \\ Y_1(\cdot) & Y_2(\cdot) \end{bmatrix} = \begin{bmatrix} \widehat{A}(\cdot) & \widehat{B}(\cdot) \\ \widehat{C}(\cdot) & \widehat{D}(\cdot) \end{bmatrix} \begin{bmatrix} E_{11}(\cdot) & E_{12}(\cdot) \\ U_1(\cdot) & U_2(\cdot) \end{bmatrix}. \quad (20)$$

The following result states how such an unfalsified model can be computed.

Theorem 4 *Generically for every $t \in \mathbb{R}$ there exists a right-inverse of*

$$\begin{bmatrix} E_{11}(t) & E_{12}(t) \\ U_1(t) & U_2(t) \end{bmatrix}, \quad (21)$$

denoted in the following by

$$\begin{bmatrix} E_{11}(t) & E_{12}(t) \\ U_1(t) & U_2(t) \end{bmatrix}^\dagger.$$

Consequently, for every $t \in \mathbb{R}$ the values $\widehat{A}(t)$, $\widehat{B}(t)$, $\widehat{C}(t)$, $\widehat{D}(t)$ of an unfalsified state-space model (20) for the data (1) can be computed via the equation

$$\begin{bmatrix} \widehat{A}(t) & \widehat{B}(t) \\ \widehat{C}(t) & \widehat{D}(t) \end{bmatrix} = \begin{bmatrix} \frac{d}{dt} E_{11}(t) & \frac{d}{dt} E_{12}(t) \\ Y_1(t) & Y_2(t) \end{bmatrix} \begin{bmatrix} E_{11}(t) & E_{12}(t) \\ U_1(t) & U_2(t) \end{bmatrix}^\dagger.$$

Proof The first part of the claim follows from Theorem 2 and Lemma 1. The second part follows multiplying both sides of Eq. (20) on the right by the right-inverse of (21). \square

6 Conclusions and Ongoing Research

We have generalised the rational interpolation approach based on the classical (constant) Loewner matrix, originally developed by Thanos and his collaborators, to the case of data generated by a time-varying system, by introducing a time-varying Loewner matrix. Time and again (see Sects. 3.1 and 4.1) we have shown that the results we obtain in our framework are *mutatis mutandis* generalisations of those already known in the classical one. The festive occasion dictated that we concentrate on illustrating such common features, but a couple of remarks are in order, if only to emphasise that Thanos’s intuition is more far-reaching than perhaps even he realises.

Firstly, we note that our identification approach is conceptually analogous to that of *subspace identification* (see [27, 28]); from data we first compute state trajectories, and on the basis of those and the measurements we compute state equations. In the rational interpolation setting the intermediate step is neither conceptually nor computationally necessary, since the problem formulation is different. However, an explicit computation of state trajectories through the factorisation of Loewner matrices opens up several interesting opportunities. Since different factorizations generate different state trajectories and consequently different bases of the state space, “special” factorizations can be performed to achieve special (e.g. balanced) state representations. Moreover, besides an “exact” factorization, also an approximate one (obtained for example selecting only a lower rank approximation of a matrix, as happens in the case of singular value decompositions) can be performed. This opens up the possibility of performing model-order reduction of time-varying systems from data (see [13] for the linear, time-invariant case).

Secondly, we emphasize that *duality* plays an essential role in our approach, whether explicitly (as in the present contribution and in [7, 16, 19]) or implicitly, where we have investigated the identification of conservative (see [13–15]) or port-Hamiltonian systems (see [20]). In the latter cases duality arises as a consequence of either *energy conservation* or *power* relations and their effect on the system dynamics. Such properties arise in a variety of systems, for example nonlinear ones, on which our current research is focused. Many of the current techniques to compute nonlinear models from data seem to fall in the class of *parametric*, rather than *system* identification methods: the structure of the model is known, and system identification is reduced to determining numerical values for the undetermined parameters, so that the resulting model explains the given measurements. The potential of a duality (equivalently: energy-, or power relation-based) point of view lies in providing procedures that can identify the *structure* of a system.

Finally, we point out the interesting problems arising from the application of the abstract mathematical approach outlined here to the real world. Computing continuous-time models from sampled data while maintaining fundamental system properties and structures is an important current area of research in system identification (see [29]). In the framework presented here, analogous issues arise in the use of suitable numerical procedures to compute integrals, and in the factorization of matrices of functions, that are guaranteed to produce e.g. self-dual systems.

References

1. Anderson, B.D.O., Antoulas, A.C.: Rational interpolation and state variable realizations. *Linear Algebra Appl.* **137**(138), 479–509 (1990)
2. Antoulas, A.C., Ball, J.A., Kang, J., Willems, J.C.: On the solution of the minimal rational interpolation problem. *Linear Algebra Appl.* **137**(138), 511–573 (1990)
3. Mayo, A.J., Antoulas, A.C.: A behavioural approach to positive real interpolation. *J. Math. Comp. Mod. Dyn. Syst.* **8**, 445–455 (2002)
4. Mayo, A.J., Antoulas, A.C.: A framework for the solution of the generalized realization problem. *Linear Algebra Appl.* **425**, 634–662 (2007)
5. Antoulas, A.C., Willems, J.C.: A behavioral approach to linear exact modeling. *IEEE Trans. Autom. Control* **AC-38**, 1776–1802 (1993)
6. Antoulas, A.C.: A new result on passivity-preserving model reduction. *Syst. Control Lett.* **54**, 361–374 (2005)
7. Rapisarda, P., Antoulas, A.C.: Bilinear differential forms and the Loewner framework for rational interpolation. In: Belur, M.N., Çamlıbel, M.K., Rapisarda, P., Scherpen, J.M.A. (eds.) *Mathematical Control Theory II: Behavioral Systems and Robust Control. Lecture Notes in Control and Information Sciences*, vol. 462, pp. 23–43. Springer, Berlin (2015)
8. Rapisarda, P., Antoulas, A.C.: A duality perspective on Loewner rational interpolation and state-space modelling of vector-exponential trajectories. In: *Proceedings of the 54th IEEE CDC, Osaka, Japan*, pp. 2096–2100 (2015)
9. Rapisarda, P., Antoulas, A.C.: On bilinear differential forms and Loewner rational interpolation. In: *Proceedings of the 54th IEEE CDC, Osaka, Japan* (2015)
10. Rapisarda, P., Willems, J.C.: The subspace Nevanlinna interpolation problem and the most powerful unfalsified model. *Syst. Control Lett.* **32**, 291–300 (1997)
11. Kaneko, O., Rapisarda, P.: Recursive exact H_∞ -identification from impulse response measurements. *IFAC Proc. Vol.* **36**(16), 1855–1860 (2003)
12. Ha, M.B., Trentelman, H.L., Rapisarda, P.: Dissipativity preserving model reduction by retention of trajectories of minimal dissipation. *Math. Syst. Signals Control* **21**, 171–201 (2009)
13. Rapisarda, P., Trentelman, H.L.: Identification and data-driven model reduction of state-space representations of lossless and dissipative systems from noise-free data. *Automatica* **47**, 1721–1728 (2010)
14. Rao, S., Rapisarda, P.: Realization of lossless systems via constant matrix factorizations. *IEEE Trans. Autom. Control* **58**(10), 2632–2636 (2013)
15. Rapisarda, P., van der Schaft, A.J.: Identification and data-driven model order reduction for linear conservative port- and self-adjoint Hamiltonian systems. In: *Proceedings of the 52nd IEEE CDC, Firenze, Italy* (2013)
16. Rapisarda, P., Antoulas, A.C.: State-space modeling of two-dimensional vector-exponential trajectories. *SIAM J. Control Opt.* **54**(5), 2734–2753 (2016)
17. Rapisarda, P., Antoulas, A.C.: A bilinear differential forms approach to parametric structured state-space modelling. *Syst. Control Lett.* **92**, 14–19 (2016)
18. Rapisarda, P.: Discrete Roesser state models from 2D frequency data. *Mult. Syst. Sign. Proc.* **30**(2), 591–610 (2019)
19. Rapisarda, P.: On the identification of self-adjoint linear time-varying state models. *IFAC-PapersOnLine* **51**(15), 251–256 (2018)
20. Branford, E., Rapisarda, P.: From Dirac structure to state model: identification of linear time-varying port-Hamiltonian systems. Submitted to *Proceedings of the 58th IEEE CDC, Nice, France* (2019)
21. Crouch, P.E., Van der Schaft, A.J.: Variational and Hamiltonian control systems. *Lecture Notes in Control and Information Sciences*. Springer, Berlin (1987)
22. van der Schaft, A.J.: Duality for linear systems: external and state space characterization of the adjoint system. In: Bonnard B., Bride B., Gauthier JP., Kupka I. (eds.) *Analysis of Controlled Dynamical Systems. Progress in Systems and Control Theory*, vol. 8. Birkhäuser, Boston (1991)

23. Crouch, P.E., Lamnabhi-Lagarrigue, F., van der Schaft, A.J.: Adjoint and Hamiltonian input-output differential equations. *IEEE Trans. Autom. Control* **40**(4), 603–615 (1995)
24. van der Schaft, A.J., Jeltsema, D.: Port-Hamiltonian systems theory: an introductory overview. *Found. Trends Syst. Control* **1**(2–3), 173–378 (2014)
25. Silverman, L.M., Meadows, H.E.: Controllability and observability in time-variable linear systems. *SIAM J. Control* **5**(1), 64–73 (1967)
26. Willems, J.C., Trentelman, H.L.: On quadratic differential forms. *SIAM J. Control Opt.* **36**(5), 1707–1749 (1998)
27. van Overschee, P., de Moor, B.: *Subspace Identification for Linear Systems*. Springer, Berlin (1996)
28. Verhaegen, M., Yu, X.: A class of subspace model identification algorithms to identify periodically and arbitrarily time-varying systems. *Automatica* **31**(2), 201–216 (1995)
29. Bruschetta, M., Picci, G., Saccon, A.: A variational integrators approach to second order modeling and identification of linear mechanical systems. *Automatica* **50**(3), 727–736 (2014)

Linear System Matrices of Rational Transfer Functions



Froilán Dopico, María del Carmen Quintana, and Paul Van Dooren

Abstract In this paper we derive new sufficient conditions for a linear system matrix $S(\lambda) := \begin{bmatrix} T(\lambda) & -U(\lambda) \\ V(\lambda) & W(\lambda) \end{bmatrix}$, where $T(\lambda)$ is assumed regular, to be strongly irreducible. In particular, we introduce the notion of strong minimality, and the corresponding conditions are shown to be sufficient for a polynomial system matrix to be strongly minimal. A strongly irreducible or minimal system matrix has the same structural elements as the rational matrix $R(\lambda) := W(\lambda) + V(\lambda)T(\lambda)^{-1}U(\lambda)$, which is also known as the transfer function connected to the system matrix $S(\lambda)$. The pole structure, zero structure and null space structure of $R(\lambda)$ can be then computed with the staircase algorithm and the QZ algorithm applied to pencils derived from $S(\lambda)$. We also show how to derive a strongly minimal system matrix from an arbitrary linear system matrix by applying to it a reduction procedure, that only uses unitary equivalence transformations. This implies that numerical errors performed during the reduction procedure remain bounded. Since we use unitary transformations in both the reduction procedure and the computation of the eigenstructure, this guarantees that we computed the exact eigenstructure of a perturbed linear system matrix, but where the perturbation is of the order of the machine precision.

Keywords System matrix · Strong minimality · Strong irreducibility

F. Dopico · M. del Carmen Quintana
Universidad Carlos III de Madrid, Getafe, Spain
e-mail: dopico@math.uc3m.es

M. del Carmen Quintana
e-mail: maquinta@math.uc3m.es

P. Van Dooren (✉)
Université catholique de Louvain, Ottignies-Louvain-la-Neuve, Belgium
e-mail: paul.vandooren@uclouvain.be

1 Introduction

Already in the seventies, Rosenbrock [6] introduced the concept of a polynomial system matrix

$$S(\lambda) := \begin{bmatrix} T(\lambda) & -U(\lambda) \\ V(\lambda) & W(\lambda) \end{bmatrix}, \quad (1)$$

where $T(\lambda)$ is assumed to be regular. He showed that the finite pole and zero structure of its transfer function matrix $R(\lambda) = W(\lambda) + V(\lambda)T(\lambda)^{-1}U(\lambda)$ can be retrieved from the polynomial matrices $T(\lambda)$ and $S(\lambda)$, respectively, provided it is *irreducible* or *minimal*, meaning that the matrices

$$\begin{bmatrix} T(\lambda) & -U(\lambda) \end{bmatrix}, \quad \begin{bmatrix} T(\lambda) \\ V(\lambda) \end{bmatrix}, \quad (2)$$

have, respectively, full row and column rank for all finite λ . This was already well known for state-space models of a proper transfer function $R_p(\lambda)$, where the system matrix takes the special form

$$S_p(\lambda) := \begin{bmatrix} \lambda I - A & -B \\ C & D \end{bmatrix}$$

where (A, B) is controllable and (A, C) is observable, meaning that $S_p(\lambda)$ is minimal. That is, $\begin{bmatrix} \lambda I - A & -B \end{bmatrix}$ and $\begin{bmatrix} \lambda I - A \\ C \end{bmatrix}$ both satisfy the conditions in (2), respectively. The poles of such a proper transfer function are all finite and are the eigenvalues of A , while the finite zeros are the finite generalized eigenvalues of the pencil $S_p(\lambda)$. The main advantage of using state-space models is that there are algorithms to compute the eigenstructure using unitary transformations only. There are also algorithms available to derive a minimal state-space model from a non-minimal one, and these algorithms are also based on unitary transformations only [8].

When allowing generalized state space models, then all transfer functions can be realized by a system matrix of the type

$$S_g(\lambda) := \begin{bmatrix} \lambda E - A & -B \\ C & D \end{bmatrix}, \quad (3)$$

since the matrix E is allowed to be singular. Moreover, when the pencils

$$\begin{bmatrix} \lambda E - A & -B \end{bmatrix}, \quad \begin{bmatrix} \lambda E - A \\ C \end{bmatrix}, \quad (4)$$

have, respectively, full row rank and column rank for all finite λ , then we retrieve the irreducibility or minimality conditions of Rosenbrock in (2), which imply that the finite poles of $R(\lambda) := D + C(\lambda E - A)^{-1}B$ are the finite eigenvalues of $\lambda E - A$

and the finite zeros of $R(\lambda)$ are the finite zeros of $S_g(\lambda)$. It was shown in [10] that when imposing also the conditions that the pencil in (3) is *strongly* irreducible, meaning that the matrices in (4) have full row rank for all finite and infinite λ , then also the infinite pole and zero structure of $R(\lambda)$ can be retrieved from the infinite structure of $\lambda E - A$ and $S_g(\lambda)$, respectively, and that the left and right minimal indices of $R(\lambda)$ and $S_g(\lambda)$ are also the same. Moreover, a reduction procedure to derive a strongly irreducible generalized state-space model from a reducible one was also given in [8], and it is also based on unitary transformations only.

In [11] these results were then extended to arbitrary polynomial models, but the procedure required irreducibility tests that were more involved. In this paper we will show that these conditions can again be simplified (and also made more uniform) when the system matrix is linear, i.e.,

$$S(\lambda) := \begin{bmatrix} A(\lambda) & -B(\lambda) \\ C(\lambda) & D(\lambda) \end{bmatrix} := \begin{bmatrix} \lambda A_1 - A_0 & B_0 - \lambda B_1 \\ \lambda C_1 - C_0 & \lambda D_1 - D_0 \end{bmatrix}. \quad (5)$$

We will define the notion of strongly minimal polynomial system matrix, and we will prove that the strong minimality conditions imply the strong irreducibility conditions in [11]. We remark that, although the notions of irreducible or minimal polynomial system matrix refer to the same conditions in (2), the conditions for a polynomial system matrix to be strongly irreducible or strongly minimal are different in general. We will also show that when the strong minimality conditions are not satisfied, we can reduce the system matrix to one where they are satisfied, and this without modifying the transfer function. Such a procedure was already derived in [9], but only for linear system matrices that were already minimal at finite points. In this paper we thus extend this to arbitrary linear system matrices.

In the next Section we briefly recall the background material for this paper and introduce the basic notation. In Sect. 3 we also recall the definition of strongly irreducible polynomial system matrix in [11], and we introduce the notion of strong minimality. In addition, we establish the relation between them. We then give, in Sect. 4, an algorithm to construct a strongly minimal linear system matrix from an arbitrary one, and we discuss the computational aspects in Sect. 5. Finally, we end with some numerical experiments in Sect. 6 and some concluding remarks in Sect. 7.

2 Background

We will restrict ourselves here to polynomial and rational matrices with coefficients in the field of complex numbers \mathbb{C} . The set of $m \times n$ polynomial matrices, denoted by $\mathbb{C}[\lambda]^{m \times n}$ and the set of $m \times n$ rational matrices, denoted by $\mathbb{C}(\lambda)^{m \times n}$, can both be viewed as matrices over the field of rational functions with complex coefficients, denoted by $\mathbb{C}(\lambda)$.

Every rational matrix can have poles and zeros and has a right and a left null space (these can be trivial, i.e., equal to $\{0\}$). Via the local Smith-McMillan form, one can

associate structural indices to the poles and zeros, and via the notion of minimal polynomial bases for rational vector spaces, one can associate so called right and left minimal indices to the right and left null spaces. We briefly recall here these different types of indices. Since we assumed (for simplicity) that the coefficients of the rational matrix are in \mathbb{C} , the poles and zeros are in the same set.

Definition 1 A square rational matrix $M(\lambda) \in \mathbb{C}(\lambda)^{m \times m}$ is said to be regular at a point $\lambda_0 \in \mathbb{C}$ if the matrix $M(\lambda_0)$ is bounded (i.e., $M(\lambda_0) \in \mathbb{C}^{m \times m}$) and is invertible. This is equivalent to both rational matrices $M(\lambda)$ and $M(\lambda)^{-1}$ having a convergent Taylor expansion around the point $\lambda = \lambda_0$. Namely,

$$\begin{aligned} M(\lambda) &:= M_0 + (\lambda - \lambda_0)M_1 + (\lambda - \lambda_0)^2M_2 + (\lambda - \lambda_0)^3M_3 + \dots, \\ M(\lambda)^{-1} &:= M_0^{-1} + (\lambda - \lambda_0)H_1 + (\lambda - \lambda_0)^2H_2 + (\lambda - \lambda_0)^3H_3 + \dots. \end{aligned}$$

If $\lambda = \infty$, $M(\lambda)$ is said to be biproper or regular at infinity if the Taylor expansions above are in terms of $1/\lambda$ instead of the factor $(\lambda - \lambda_0)$.

Definition 2 Let $R(\lambda)$ be an arbitrary $m \times n$ rational matrix of normal rank r . Then its *local Smith-McMillan form* at a point $\lambda_0 \in \mathbb{C}$ is the diagonal matrix obtained under rational left and right transformations $M_\ell(\lambda)$ and $M_r(\lambda)$, that are *regular* at λ_0 :

$$M_\ell(\lambda)R(\lambda)M_r(\lambda) = \begin{bmatrix} \text{diag}((\lambda - \lambda_0)^{d_1}, \dots, (\lambda - \lambda_0)^{d_r}) & 0 \\ 0 & 0_{(m-r) \times (n-r)} \end{bmatrix}, \quad (6)$$

where $d_1 \leq d_2 \leq \dots \leq d_r$. If $\lambda_0 = \infty$, the basic factor $(\lambda - \lambda_0)$ is replaced by $\frac{1}{\lambda}$ and the transformation matrices are then biproper. The latter can be viewed as a change of variable $\mu = \frac{1}{\lambda}$ which transform $\lambda_0 = \infty$ to $\mu_0 = 0$.

Remark 1 The normal rank of a rational matrix is the size of its largest nonidentically zero minor. The indices d_i are unique and are called the *structural indices* of $R(\lambda)$ at λ_0 . In particular, the strictly positive indices correspond to a zero at λ_0 , and the strictly negative indices correspond to a pole at λ_0 . The *zero degree* is defined as the sum of all structural indices of all zeros (infinity included), and the *polar degree* is the sum of all structural indices (in absolute value) of all poles (infinity included).

Example 1 Let us consider the 2×2 rational matrix

$$R(\lambda) = \begin{bmatrix} e_5(\lambda) & 0 \\ c/\lambda & e_1(\lambda) \end{bmatrix} \quad (7)$$

where $e_5(\lambda)$ is a monic polynomial of degree 5 and $e_1(\lambda)$ is a monic polynomial of degree 1, with $e_5(0) \neq 0$ and $e_1(0) \neq 0$. If $c \neq 0$, the only poles are 0 and infinity, and the corresponding local Smith-McMillan forms for these two points are

$$\lambda_0 = 0 : \text{diag}(\lambda^{-1}, \lambda^1), \quad \lambda_0 = \infty (\mu_0 = 0) : \text{diag}(\mu^{-5}, \mu^{-1}),$$

indicating that $\lambda_0 = 0$ is a zero as well as a pole. The other finite zeros are the six finite roots of $e_5(\lambda)$ and $e_1(\lambda)$. The polar degree and the zero degree for this example are thus both equal to 7. When $c = 0$, the pole and zero at $\lambda = 0$ disappear and the matrix is polynomial instead of rational. The polar and zero degree are then both equal to 6.

The above definitions of pole and zero structure of a rational matrix $R(\lambda)$ are those that are commonly used in linear systems theory (see [6]) and are due to McMillan. They describe the spectral properties of a rational matrix. But when applying them to matrix pencils $S(\lambda)$ we may wonder if they coincide with definitions of eigenvalues and generalized eigenvalues and their multiplicities, i.e. the Kronecker structure of $S(\lambda)$ (see [3]).

Definition 3 The Kronecker canonical form of an arbitrary $m \times n$ pencil $\lambda B - A$ of normal rank r is a block diagonal form obtained via invertible transformations S and T :

$$S(\lambda B - A)T = \text{diag}(L_{\eta_1}^T(\lambda), \dots, L_{\eta_{m-r}}^T(\lambda), \lambda I_{r_f} - A_J, \lambda N - I_{r_\infty}, L_{\epsilon_1}(\lambda), \dots, L_{\epsilon_{n-r}}(\lambda))$$

where A_J is in Jordan form, N is nilpotent and in Jordan form, and

$$L_k(\lambda) := \begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & \\ & & & \lambda & 1 \end{bmatrix}$$

is a $k \times (k + 1)$ singular pencil. The finite eigenvalues of $(\lambda B - A)$ are the r_f eigenvalues of A_J and its r_∞ infinite eigenvalues are the generalized eigenvalues of $\lambda N - I_{r_\infty}$.

For this comparison, we only need to look at zeros, since a pencil has only one pole (namely, infinity) and its multiplicity is the rank of the coefficient of λ . In other words, its polar structure is trivial. But what about the correspondence of the *zero structure* of $S(\lambda)$ (in the McMillan sense) and the eigenvalue structure of $S(\lambda)$ (in the sense of Kronecker)? It turns out that for finite eigenvalues of $S(\lambda)$ there is a complete isomorphism with the zero structure of $S(\lambda)$: every Jordan block of size k at an eigenvalue λ_0 in the Kronecker canonical form of $S(\lambda)$ corresponds to an elementary divisor $(\lambda - \lambda_0)^k$ in the Smith-McMillan form of $S(\lambda)$. But for $\lambda = \infty$, there is a difference. It is well known (see [10]) that a Kronecker block of size k at $\lambda = \infty$ corresponds to an elementary divisor $(\frac{1}{\lambda})^{(k-1)}$ in the Smith-McMillan form. For the point at infinity there is thus a shift of 1 in the structural indices. For this reason we want to make a clear distinction between both index sets. Whenever we talk about *zeros*, we refer to the McMillan structure, and whenever we talk about *eigenvalues*, we refer to the Kronecker structure.

It is well known that every rational vector subspace \mathcal{V} , i.e., every subspace $\mathcal{V} \subseteq \mathbb{C}(\lambda)^n$ over the field $\mathbb{C}(\lambda)$, has bases consisting entirely of polynomial vectors. Among

them some are minimal in the following sense introduced by Forney [2]: a *minimal basis* of \mathcal{V} is a basis of \mathcal{V} consisting of polynomial vectors whose sum of degrees is minimal among all bases of \mathcal{V} consisting of polynomial vectors. The fundamental property [2, 5] of such bases is that the ordered list of degrees of the polynomial vectors in any minimal basis of \mathcal{V} is always the same. Therefore, these degrees are an intrinsic property of the subspace \mathcal{V} and are called the *minimal indices* of \mathcal{V} . This leads to the definition of the minimal bases and indices of a rational matrix. An $m \times n$ rational matrix $R(\lambda)$ of normal rank r smaller than m and/or n has non-trivial left and/or right rational null-spaces, respectively, over the field $\mathbb{C}(\lambda)$:

$$\begin{aligned}\mathcal{N}_\ell(R) &:= \{y(\lambda)^T \in \mathbb{C}(\lambda)^{1 \times m} : y(\lambda)^T R(\lambda) \equiv 0^T\}, \\ \mathcal{N}_r(R) &:= \{x(\lambda) \in \mathbb{C}(\lambda)^{n \times 1} : R(\lambda)x(\lambda) \equiv 0\}.\end{aligned}$$

Rational matrices with non-trivial left and/or right null-spaces are said to be *singular*. If the rational subspace $\mathcal{N}_\ell(R)$ is non-trivial, it has minimal bases and minimal indices, which are called the *left minimal bases and indices* of $R(\lambda)$. Analogously, the *right minimal bases and indices* of $R(\lambda)$ are those of $\mathcal{N}_r(R)$, whenever this subspace is non-trivial. Notice that an $m \times n$ rational matrix of normal rank r has $m - r$ left minimal indices $\{\eta_1, \dots, \eta_{m-r}\}$, and $n - r$ right minimal indices $\{\epsilon_1, \dots, \epsilon_{n-r}\}$.

The *McMillan degree* $\delta(R)$ of a rational matrix $R(\lambda)$ is the polar degree introduced in Remark 1. The following degree sum theorem was proven in [10], and relates the McMillan degree to the other structural elements of $R(\lambda)$: to the *zero degree* $\delta_z(R)$, to the *left nullspace degree* $\delta_\ell(R)$, that is the sum of all left minimal indices, and to the *right nullspace degree* $\delta_r(R)$, that is the sum of all right minimal indices.

Theorem 1 *Let $R(\lambda) \in \mathbb{C}(\lambda)^{m \times n}$. Then*

$$\delta(R) := \delta_p(R) = \delta_z(R) + \delta_\ell(R) + \delta_r(R).$$

3 Strong Irreducibility and Minimality

In this section we recall the strong irreducibility conditions in [11] for polynomial system matrices, and we introduce the notion of strong minimality. Then, we study the relation between them for the case of linear system matrices.

Definition 4 A polynomial system matrix $S(\lambda)$ as in (1) is said to be *strongly controllable* and *strongly observable*, respectively, if the polynomial matrices

$$\begin{bmatrix} T(\lambda) & -U(\lambda) & 0 \\ V(\lambda) & W(\lambda) & -I \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} T(\lambda) & -U(\lambda) \\ V(\lambda) & W(\lambda) \\ 0 & I \end{bmatrix}, \quad (8)$$

have no finite or infinite zeros. If both conditions are satisfied, $S(\lambda)$ is said to be *strongly irreducible*.

Let us now consider the transfer function matrix $R(\lambda) = W(\lambda) + V(\lambda)T(\lambda)^{-1}U(\lambda)$ of the polynomial system matrix in (1). In such a case, we also say that the system quadruple $\{T(\lambda), U(\lambda), V(\lambda), W(\lambda)\}$ realizes $R(\lambda)$. Moreover, we say that the system quadruple is *strongly irreducible* if the polynomial system matrix is *strongly irreducible*. It was shown in [11] that the pole/zero and null space structure of $R(\lambda)$ can be retrieved from a strongly irreducible system quadruple $\{T(\lambda), U(\lambda), V(\lambda), W(\lambda)\}$ as follows.

Theorem 2 *If the polynomial system matrix $S(\lambda)$ in (1) is strongly irreducible, then*

1. *the zero structure of $R(\lambda)$ at finite and infinite λ is the same as the zero structure of $S(\lambda)$ at finite and infinite λ ,*
2. *the pole structure of $R(\lambda)$ at finite λ is the same as the zero structure at λ of $T(\lambda)$,*
3. *the pole structure of $R(\lambda)$ at infinity is the same as the zero structure at infinity of*

$$\begin{bmatrix} T(\lambda) & -U(\lambda) & 0 \\ V(\lambda) & W(\lambda) & -I \\ 0 & I & 0 \end{bmatrix},$$

4. *the left and right minimal indices of $R(\lambda)$ and $S(\lambda)$ are the same.*

If one specializes this to the generalized state space model (3) one retrieves the results of [10], which are simpler and only involve the pencils $(\lambda E - A)$, (3) and (4). We now show that the above conditions can be simplified when the system matrices are linear as in (5). First, we present the definition of strongly minimal polynomial system matrix.

Definition 5 Let d be the degree of the polynomial system matrix $S(\lambda)$ in (1). $S(\lambda)$ is said to be *strongly E-controllable* and *strongly E-observable*, respectively, if the polynomial matrices

$$[T(\lambda) \quad -U(\lambda)], \quad \text{and} \quad \begin{bmatrix} T(\lambda) \\ V(\lambda) \end{bmatrix}, \tag{9}$$

have no finite or infinite¹ eigenvalues, considered as polynomial matrices of grade d . If both conditions are satisfied, $S(\lambda)$ is said to be *strongly minimal*.

The letter E in the definition of strong E-controllability and E-observability refers to the condition of the matrices in (9) not having eigenvalues, finite or infinite. We prove in Proposition 1 that the strong irreducibility conditions hold if the strong minimality conditions are satisfied. For this, we need to recall Lemma 1 of [10], which we give here in its transposed form. Then, we prove Theorems 3 and 4, and Proposition 1 as a corollary of them.

¹ The eigenvalues at infinity of a polynomial matrix $P(\lambda)$ considered as a polynomial matrix of grade g , with $g \geq \text{degree } P(\lambda)$, are the eigenvalues at zero of $\text{rev}_g P(\lambda) := \lambda^g P(1/\lambda)$ (see [4]).

Lemma 1 *The zero structure at infinity of the pencil $[\lambda K_1 - K_0 \mid -L_0]$ where K_1 has full column rank, is isomorphic to the zero structure at zero of the pencil $[K_1 - \mu K_0 \mid -L_0]$. Moreover, if the pencil has full row normal rank, then it has no zeros at infinity, provided the constant matrix $[K_1 \mid -L_0]$ has full row rank.*

Proof The first part is proven in [10]. The second part is a direct consequence of the first part, when filling in $\mu = 0$. \square

Theorem 3 *The pencil*

$$\left[\begin{array}{ccc|c} \lambda A_1 - A_0 & B_0 - \lambda B_1 & 0 & \\ \lambda C_1 - C_0 & \lambda D_1 - D_0 & -I & \end{array} \right], \quad (10)$$

where $\lambda A_1 - A_0$ is regular, has no zeros at infinity if the pencil

$$[\lambda A_1 - A_0 \quad B_0 - \lambda B_1] \quad (11)$$

has no eigenvalues at infinity.

Proof Clearly the pencils in (10) and (11) have full row normal rank since $\lambda A_1 - A_0$ is regular. We can thus apply the result of Lemma 1 as follows. If we use an invertible matrix V to “compress” the columns of the coefficient of λ in the following pencil

$$\left[\begin{array}{ccc|c} \lambda A_1 - A_0 & B_0 - \lambda B_1 & 0 & \\ \lambda C_1 - C_0 & \lambda D_1 - D_0 & -I & \end{array} \right] \left[\begin{array}{c|c} V & 0 \\ \hline 0 & I \end{array} \right] = \left[\begin{array}{ccc|c} \lambda K_1 - K_0 & -L_0 & 0 & \\ \lambda \widehat{K}_1 - \widehat{K}_0 & -\widehat{L}_0 & -I & \end{array} \right],$$

such that the matrix $\begin{bmatrix} \widehat{K}_1 \\ \widehat{K}_1 \end{bmatrix}$ has full column rank, then this pencil has no zeros at infinity provided the constant matrix $\begin{bmatrix} \widehat{K}_1 & -\widehat{L}_0 & 0 \\ \widehat{K}_1 & -\widehat{L}_0 & -I \end{bmatrix}$ has full row rank. But if $[\lambda A_1 - A_0 \quad B_0 - \lambda B_1]$ has no infinite eigenvalues, it follows that $[A_1 \quad -B_1]$ has full row rank. And since $[A_1 \quad -B_1] V = [K_1 \quad 0]$, K_1 must have full row rank as well (in fact, it is invertible). It then follows from Lemma 1 that the pencil in (10) has no zeros at infinity. \square

In the next theorem, we state without proof the transposed version of Theorem 3.

Theorem 4 *The pencil*

$$\left[\begin{array}{ccc|c} \lambda A_1 - A_0 & B_0 - \lambda B_1 & & \\ \lambda C_1 - C_0 & \lambda D_1 - D_0 & & \\ 0 & & & I \end{array} \right],$$

where $\lambda A_1 - A_0$ is regular, has no zeros at infinity if the pencil

$$\begin{bmatrix} \lambda A_1 - A_0 \\ \lambda C_1 - C_0 \end{bmatrix} \quad (12)$$

has no eigenvalues at infinity.

Let us now consider a linear system matrix

$$L(\lambda) := \lambda L_1 - L_0 := \begin{bmatrix} \lambda A_1 - A_0 & B_0 - \lambda B_1 \\ \lambda C_1 - C_0 & \lambda D_1 - D_0 \end{bmatrix}, \quad (13)$$

with $\lambda A_1 - A_0$ regular. Notice that if $L(\lambda)$ is minimal (i.e., satisfies (2)) and, in addition, satisfies the conditions in (11) and (12), then it is strongly minimal. By Theorems 3 and 4, we have that these conditions imply strong irreducibility on linear system matrices. We state such result in Proposition 1.

Proposition 1 *A linear system matrix as in (13) is strongly irreducible if it is strongly minimal.*

Remark 2 Notice that conditions (11) and (12) are only sufficient, not necessary. But they are easy to test, and also to obtain after a reduction procedure, as we show in Sect. 4.

Theorems 3 and 4 and Proposition 1 can be extended to polynomial system matrices. However, we do not state these results here since, in this paper, we are focusing on linear system matrices. If we recapitulate the results of this section, we obtain the following theorem.

Theorem 5 *A linear system pencil $L(\lambda)$ as in (13), realizing the transfer function $R(\lambda) := (\lambda D_1 - D_0) + (\lambda C_1 - C_0)(\lambda A_1 - A_0)^{-1}(\lambda B_1 - B_0)$, is strongly irreducible if it is strongly minimal. Moreover, if $L(\lambda)$ is strongly irreducible then*

1. *the zero structure of $R(\lambda)$ at finite and infinite λ is the same as the zero structure of $L(\lambda)$ at finite and infinite λ ,*
2. *the left and right minimal indices of $R(\lambda)$ and $L(\lambda)$ are the same,*
3. *the finite polar structure of $R(\lambda)$ is the same as the finite zero structure of $\lambda A_1 - A_0$, and*
4. *the infinite polar structure of $R(\lambda)$ is the same as the infinite zero structure of the pencil*

$$\begin{bmatrix} \lambda A_1 - A_0 & -\lambda B_1 & 0 \\ \lambda C_1 & \lambda D_1 & -I \\ 0 & I & 0 \end{bmatrix}. \quad (14)$$

Remark 3 It follows from this theorem and the degree sum theorem in Theorem 1 that the rank of L_1 equals the McMillan degree of $R(\lambda)$, and that there can be no linear system matrix for $R(\lambda)$ with a smaller rank of L_1 that satisfies Theorem 5.

It may look strange that there is such a difference in the treatment of finite and infinite poles of $R(\lambda)$ in Theorem 5, but it should be pointed out that the matrices (B_1, C_1, D_1) contribute to the infinite polar structure of $R(\lambda)$, and not to the finite polar structure. Notice that in (14) we have eliminated the matrices B_0, C_0 and D_0 with strict equivalence transformations using the identity matrices as pivots.

4 Reducing to a Strongly Minimal Linear System Matrix

In this section we give an algorithm to reduce an arbitrary linear system matrix to a strongly minimal one. Given a linear system quadruple $\{A(\lambda), B(\lambda), C(\lambda), D(\lambda)\}$, where $A(\lambda) \in \mathbb{C}(\lambda)^{d \times d}$, $B(\lambda) \in \mathbb{C}(\lambda)^{d \times n}$, $C(\lambda) \in \mathbb{C}(\lambda)^{m \times d}$, $D(\lambda) \in \mathbb{C}(\lambda)^{m \times n}$ and $A(\lambda)$ is assumed to be regular, we describe first how to obtain a strongly E-controllable quadruple $\{A_c(\lambda), B_c(\lambda), C_c(\lambda), D_c(\lambda)\}$ of smaller state dimension $(d - r)$. For that, our reduction procedure deflates finite and infinite “uncontrollable eigenvalues” by proceeding in three different steps. Then the reduction to a strongly E-observable one is dual and can be obtained by mere transposition of the system matrix and application of the first method for obtaining a strongly E-controllable system.

Step 1: We first show that there exist unitary transformations U and V that yield a decomposition of the type

$$\begin{bmatrix} U & 0 \\ 0 & I_m \end{bmatrix} \begin{bmatrix} A(\lambda) & -B(\lambda) \\ C(\lambda) & D(\lambda) \end{bmatrix} \begin{bmatrix} V & 0 \\ 0 & I_n \end{bmatrix} = \begin{bmatrix} X(\lambda) \widehat{W}_{11} & 0 & X(\lambda) W_{13} \\ \widetilde{Y}(\lambda) & \widetilde{A}(\lambda) & -\widetilde{B}(\lambda) \\ \widetilde{Z}(\lambda) & \widetilde{C}(\lambda) & D(\lambda) \end{bmatrix}, \quad (15)$$

where $\widehat{W}_{11} \in \mathbb{C}^{r \times r}$ and $W_{13} \in \mathbb{C}^{r \times n}$ are constant, and \widehat{W}_{11} is invertible. This will allow us in step 2 to deflate the block $X(\lambda)$ and construct a lower order model that is strongly E-controllable. In order to prove this, we start from the generalized Schur decomposition for singular pencils (see [7])

$$U \begin{bmatrix} A(\lambda) & -B(\lambda) \end{bmatrix} W^* = \begin{bmatrix} X(\lambda) & 0 & 0 \\ Y(\lambda) & \widehat{A}(\lambda) & -\widehat{B}(\lambda) \end{bmatrix}, \quad (16)$$

where $X(\lambda) \in \mathbb{C}[\lambda]^{r \times r}$ is the regular part of $\begin{bmatrix} A(\lambda) & -B(\lambda) \end{bmatrix}$, $\widehat{A}(\lambda) \in \mathbb{C}[\lambda]^{(d-r) \times (d-r)}$, and $\begin{bmatrix} \widehat{A}(\lambda) & -\widehat{B}(\lambda) \end{bmatrix}$ has no finite or infinite eigenvalues anymore. The decomposition in (16) can be obtained by using unitary transformations U and W . If we partition U as $\begin{bmatrix} U_1 \\ U_2 \end{bmatrix}$, with $U_1 \in \mathbb{C}^{r \times d}$, then

$$U_1 \begin{bmatrix} A(\lambda) & -B(\lambda) \end{bmatrix} = \begin{bmatrix} X(\lambda) W_{11} & X(\lambda) W_{12} & X(\lambda) W_{13} \end{bmatrix},$$

where $W_{11} \in \mathbb{C}^{r \times r}$, $W_{12} \in \mathbb{C}^{r \times (d-r)}$ and $W_{13} \in \mathbb{C}^{r \times n}$ are the corresponding submatrices of W . Since $A(\lambda)$ is regular, $X(\lambda) \begin{bmatrix} W_{11} & W_{12} \end{bmatrix}$ must be full normal rank, and hence $\begin{bmatrix} W_{11} & W_{12} \end{bmatrix}$ must be full row rank as well. Therefore, there must exist a unitary matrix V such that $\begin{bmatrix} W_{11} & W_{12} \end{bmatrix} V = \begin{bmatrix} \widehat{W}_{11} & 0 \end{bmatrix}$, where \widehat{W}_{11} is invertible. Hence, we have

$$\begin{bmatrix} U & 0 \\ 0 & I_m \end{bmatrix} \begin{bmatrix} A(\lambda) & -B(\lambda) \\ C(\lambda) & D(\lambda) \end{bmatrix} \begin{bmatrix} V & 0 \\ 0 & I_n \end{bmatrix} = \begin{bmatrix} X(\lambda)\widehat{W}_{11} & 0 & X(\lambda)W_{13} \\ \widetilde{Y}(\lambda) & \widetilde{A}(\lambda) & -\widetilde{B}(\lambda) \\ \widetilde{Z}(\lambda) & \widetilde{C}(\lambda) & D(\lambda) \end{bmatrix},$$

where

$$W \begin{bmatrix} V & 0 \\ 0 & I_n \end{bmatrix} = \begin{bmatrix} \widehat{W}_{11} & 0 & W_{13} \\ \widehat{W}_{21} & \widehat{W}_{22} & W_{23} \\ \widehat{W}_{31} & \widehat{W}_{32} & W_{33} \end{bmatrix}.$$

Step 2: We now define $E := -\widehat{W}_{11}^{-1}W_{13}$ and perform the following non-unitary transformation on the pencil:

$$\begin{bmatrix} X(\lambda)\widehat{W}_{11} & 0 & X(\lambda)W_{13} \\ \widetilde{Y}(\lambda) & \widetilde{A}(\lambda) & -\widetilde{B}(\lambda) \\ \widetilde{Z}(\lambda) & \widetilde{C}(\lambda) & D(\lambda) \end{bmatrix} \begin{bmatrix} I_r & 0 & E \\ 0 & I_{d-r} & 0 \\ 0 & 0 & I_n \end{bmatrix} = \begin{bmatrix} X(\lambda)\widehat{W}_{11} & 0 & 0 \\ \widetilde{Y}(\lambda) & \widetilde{A}(\lambda) & \widetilde{Y}(\lambda)E - \widetilde{B}(\lambda) \\ \widetilde{Z}(\lambda) & \widetilde{C}(\lambda) & \widetilde{Z}(\lambda)E + D(\lambda) \end{bmatrix}.$$

We have obtained an equivalent system representation in which the $(1, 1)$ -block, $X(\lambda)\widehat{W}_{11}$, can be deflated since it does not contribute to the transfer function. We then obtain a smaller linear system pencil:

$$\begin{bmatrix} \widetilde{A}(\lambda) & \widetilde{Y}(\lambda)E - \widetilde{B}(\lambda) \\ \widetilde{C}(\lambda) & \widetilde{Z}(\lambda)E + D(\lambda) \end{bmatrix},$$

that has the same transfer function. One can also perform this elimination by another unitary transformation \widetilde{W} constructed to eliminate W_{13} :

$$\begin{bmatrix} \widehat{W}_{11} & 0 & W_{13} \end{bmatrix} \begin{bmatrix} \widetilde{W}_{11} & 0 & \widetilde{W}_{13} \\ 0 & I_{d-r} & 0 \\ \widetilde{W}_{31} & 0 & \widetilde{W}_{33} \end{bmatrix} = \begin{bmatrix} I_r & 0 & 0 \end{bmatrix}, \quad (17)$$

implying $\widetilde{W}_{11} = \widehat{W}_{11}^*$, $\widetilde{W}_{31} = W_{13}^*$, and $\widetilde{W}_{13} = -\widehat{W}_{11}^{-1}W_{13}\widetilde{W}_{33}$. This then yields

$$\begin{aligned} & \begin{bmatrix} X(\lambda)\widehat{W}_{11} & 0 & X(\lambda)W_{13} \\ \widetilde{Y}(\lambda) & \widetilde{A}(\lambda) & -\widetilde{B}(\lambda) \\ \widetilde{Z}(\lambda) & \widetilde{C}(\lambda) & D(\lambda) \end{bmatrix} \begin{bmatrix} \widetilde{W}_{11} & 0 & \widetilde{W}_{13} \\ 0 & I_{d-r} & 0 \\ \widetilde{W}_{31} & 0 & \widetilde{W}_{33} \end{bmatrix} \\ &= \begin{bmatrix} X(\lambda) & 0 & 0 \\ \widetilde{Y}(\lambda)\widetilde{W}_{11} - \widetilde{B}(\lambda)\widetilde{W}_{31} & \widetilde{A}(\lambda) & \widetilde{Y}(\lambda)\widetilde{W}_{13} - \widetilde{B}(\lambda)\widetilde{W}_{33} \\ \widetilde{Z}(\lambda)\widetilde{W}_{11} + D(\lambda)\widetilde{W}_{31} & \widetilde{C}(\lambda) & \widetilde{Z}(\lambda)\widetilde{W}_{13} + D(\lambda)\widetilde{W}_{33} \end{bmatrix}. \end{aligned}$$

Notice that the new transfer function has now changed, but only by postmultiplication by the constant matrix \tilde{W}_{33} , which moreover is invertible. This follows from

$$\begin{bmatrix} E \\ I_n \end{bmatrix} \tilde{W}_{33} = \begin{bmatrix} \tilde{W}_{13} \\ \tilde{W}_{33} \end{bmatrix},$$

expressing that both matrices span the null-space of the same matrix $\begin{bmatrix} \hat{W}_{11} & W_{13} \end{bmatrix}$ and where the right hand side matrix has full rank since it has orthonormal columns. This also implies that

$$\begin{bmatrix} \tilde{A}(\lambda) & \tilde{Y}(\lambda)E - \tilde{B}(\lambda) \\ \tilde{C}(\lambda) & \tilde{Z}(\lambda)E + D(\lambda) \end{bmatrix} \begin{bmatrix} I_{d-r} & 0 \\ 0 & \tilde{W}_{33} \end{bmatrix} = \begin{bmatrix} \tilde{A}(\lambda) & \tilde{Y}(\lambda)\tilde{W}_{13} - \tilde{B}(\lambda)\tilde{W}_{33} \\ \tilde{C}(\lambda) & \tilde{Z}(\lambda)\tilde{W}_{13} + D(\lambda)\tilde{W}_{33} \end{bmatrix},$$

which shows that their Schur complements are related by the constant matrix \tilde{W}_{33} .

Step 3: Finally, we show that the submatrix

$$\begin{bmatrix} \tilde{A}(\lambda) & \tilde{Y}(\lambda)E - \tilde{B}(\lambda) \\ 0 & \tilde{W}_{33} \end{bmatrix} \begin{bmatrix} I_{d-r} & 0 \\ 0 & \tilde{W}_{33} \end{bmatrix} = \begin{bmatrix} \tilde{A}(\lambda) & \tilde{Y}(\lambda)\tilde{W}_{13} - \tilde{B}(\lambda)\tilde{W}_{33} \\ 0 & \tilde{W}_{33} \end{bmatrix},$$

has no finite or infinite eigenvalues anymore. For this, we first point out that the following product of unitary matrices has the form given below

$$W \begin{bmatrix} V & 0 \\ 0 & I_n \end{bmatrix} \begin{bmatrix} \tilde{W}_{11} & 0 & \tilde{W}_{13} \\ 0 & I_{d-r} & 0 \\ \tilde{W}_{31} & 0 & \tilde{W}_{33} \end{bmatrix} =: \begin{bmatrix} I_r & 0 & 0 \\ 0 & \tilde{V}_{22} & \tilde{V}_{23} \\ 0 & \tilde{V}_{32} & \tilde{V}_{33} \end{bmatrix} =: \begin{bmatrix} I_r & 0 \\ 0 & \tilde{V} \end{bmatrix}$$

because the identity (17) implies that the first block column equals $\begin{bmatrix} I_r & 0 & 0 \end{bmatrix}$. This then implies the equality

$$\begin{aligned} & \begin{bmatrix} X(\lambda) & 0 & 0 \\ \tilde{Y}(\lambda)\tilde{W}_{11} - \tilde{B}(\lambda)\tilde{W}_{31} & \tilde{A}(\lambda) & \tilde{Y}(\lambda)\tilde{W}_{13} - \tilde{B}(\lambda)\tilde{W}_{33} \end{bmatrix} \\ & = \begin{bmatrix} X(\lambda) & 0 & 0 \\ Y(\lambda) & \hat{A}(\lambda) & -\hat{B}(\lambda) \end{bmatrix} \begin{bmatrix} I_r & 0 \\ 0 & \tilde{V} \end{bmatrix}, \end{aligned}$$

which in turn implies that $\begin{bmatrix} \tilde{A}(\lambda) & \tilde{Y}(\lambda)\tilde{W}_{13} - \tilde{B}(\lambda)\tilde{W}_{33} \\ \tilde{C}(\lambda) & \tilde{Z}(\lambda)\tilde{W}_{13} + D(\lambda)\tilde{W}_{33} \end{bmatrix}$ has no finite or infinite eigenvalues. We thus have shown that the system matrix

$$S_c(\lambda) := \begin{bmatrix} A_c(\lambda) & -B_c(\lambda) \\ C_c(\lambda) & D_c(\lambda) \end{bmatrix} := \begin{bmatrix} \tilde{A}(\lambda) & \tilde{Y}(\lambda)\tilde{W}_{13} - \tilde{B}(\lambda)\tilde{W}_{33} \\ \tilde{C}(\lambda) & \tilde{Z}(\lambda)\tilde{W}_{13} + D(\lambda)\tilde{W}_{33} \end{bmatrix}$$

is now strongly E-controllable and that its transfer function $R_c(\lambda)$ equals $R(\lambda)\tilde{W}_{33}$, where $R(\lambda)$ is the transfer function of the original quadruple and \tilde{W}_{33} is invertible.

We summarize the result obtained by the three-step procedure above in Theorem 6, where we denote $d - r$ by d_c , to indicate that it is the size of $A_c(\lambda)$ in the new strongly E-controllable system, and r is replaced by $d_{\bar{c}}$, so that $d = d_{\bar{c}} + d_c$.

Theorem 6 *Let $\{A(\lambda), B(\lambda), C(\lambda), D(\lambda)\}$ be a linear system quadruple, with $A(\lambda) \in \mathbb{C}[\lambda]^{d \times d}$ regular, realizing the rational matrix $R(\lambda) := C(\lambda)A(\lambda)^{-1}B(\lambda) + D(\lambda) \in \mathbb{C}(\lambda)^{m \times n}$. Then there exist unitary transformations $U, V \in \mathbb{C}^{d \times d}$ and $\tilde{W} \in \mathbb{C}^{(d+n) \times (d+n)}$ such that the following identity holds*

$$\begin{bmatrix} U & 0 \\ 0 & I_m \end{bmatrix} \begin{bmatrix} A(\lambda) & -B(\lambda) \\ C(\lambda) & D(\lambda) \end{bmatrix} \begin{bmatrix} V & 0 \\ 0 & I_n \end{bmatrix} \tilde{W} = \begin{bmatrix} X_{\bar{c}}(\lambda) & 0 & 0 \\ Y_{\bar{c}}(\lambda) & A_c(\lambda) & -B_c(\lambda) \\ Z_{\bar{c}}(\lambda) & C_c(\lambda) & D_c(\lambda) \end{bmatrix},$$

where \tilde{W} is of the form $\tilde{W} := \begin{bmatrix} \tilde{W}_{11} & 0 & \tilde{W}_{13} \\ 0 & I_{d_c} & 0 \\ \tilde{W}_{31} & 0 & \tilde{W}_{33} \end{bmatrix} \in \mathbb{C}^{(d_{\bar{c}}+d_c+n) \times (d_{\bar{c}}+d_c+n)}$, $d_{\bar{c}}$ is the number of (finite and infinite) eigenvalues of $\begin{bmatrix} A(\lambda) & -B(\lambda) \end{bmatrix}$, and $X_{\bar{c}}(\lambda) \in \mathbb{C}[\lambda]^{d_{\bar{c}} \times d_{\bar{c}}}$ is a regular pencil. Moreover,

- (a) the eigenvalues of $\begin{bmatrix} A(\lambda) & -B(\lambda) \end{bmatrix}$ are the eigenvalues of $X_{\bar{c}}(\lambda)$,
- (b) $\begin{bmatrix} A_c(\lambda) & -B_c(\lambda) \end{bmatrix} \in \mathbb{C}[\lambda]^{d_c \times (d_c+n)}$ has no (finite or infinite) eigenvalues,
- (c) the quadruple $\{A_c(\lambda), B_c(\lambda), C_c(\lambda), D_c(\lambda)\}$ is a realization of the transfer function $R_c(\lambda) := R(\lambda)\tilde{W}_{33}$, with $\tilde{W}_{33} \in \mathbb{C}^{n \times n}$ invertible, and
- (d) if $\begin{bmatrix} A(\lambda) \\ C(\lambda) \end{bmatrix}$ has no finite or infinite eigenvalues, then $\begin{bmatrix} A_c(\lambda) \\ C_c(\lambda) \end{bmatrix}$ also has no finite or infinite eigenvalues.

Remark 4 Notice that conditions (b) and (d) in Theorem 6 imply that the system quadruple $\{A_c(\lambda), B_c(\lambda), C_c(\lambda), D_c(\lambda)\}$ is strongly minimal.

Proof The decomposition and the three properties (a), (b) and (c) were shown in the discussion above. The only part that remains to be proven is property (d). This follows from the identity (15), which yields

$$\begin{bmatrix} U & 0 \\ 0 & I_m \end{bmatrix} \begin{bmatrix} A(\lambda) \\ C(\lambda) \end{bmatrix} V = \begin{bmatrix} X(\lambda)\tilde{W}_{11} & 0 \\ \tilde{Y}(\lambda) & A_c(\lambda) \\ \tilde{Z}(\lambda) & C_c(\lambda) \end{bmatrix}.$$

This clearly implies that if $\begin{bmatrix} A(\lambda) \\ C(\lambda) \end{bmatrix}$ has full rank for all λ (including infinity), then so does $\begin{bmatrix} A_c(\lambda) \\ C_c(\lambda) \end{bmatrix}$. □

We state below a dual theorem that constructs, from an arbitrary linear system quadruple $\{A(\lambda), B(\lambda), C(\lambda), D(\lambda)\}$, a subsystem $\{A_o(\lambda), B_o(\lambda), C_o(\lambda), D_o(\lambda)\}$ where $\begin{bmatrix} A_o(\lambda) \\ C_o(\lambda) \end{bmatrix}$ has no finite or infinite eigenvalues. Its proof is obtained by applying

the previous theorem on the transposed system $\{A^T(\lambda), C^T(\lambda), B^T(\lambda), D^T(\lambda)\}$ and then transposing back the result.

Theorem 7 *Let $\{A(\lambda), B(\lambda), C(\lambda), D(\lambda)\}$ be a linear system quadruple, with $A(\lambda) \in \mathbb{C}[\lambda]^{d \times d}$ regular, realizing the rational matrix $R(\lambda) := C(\lambda)A(\lambda)^{-1}B(\lambda) + D(\lambda) \in \mathbb{C}(\lambda)^{m \times n}$. Then there exist unitary transformations $U, V \in \mathbb{C}^{d \times d}$ and $\tilde{W} \in \mathbb{C}^{(d+m) \times (d+m)}$ such that the following identity holds*

$$\tilde{W} \begin{bmatrix} U & 0 \\ 0 & I_m \end{bmatrix} \begin{bmatrix} A(\lambda) & -B(\lambda) \\ C(\lambda) & D(\lambda) \end{bmatrix} \begin{bmatrix} V & 0 \\ 0 & I_n \end{bmatrix} = \begin{bmatrix} X_{\bar{\sigma}}(\lambda) & Y_{\bar{\sigma}}(\lambda) & Z_{\bar{\sigma}}(\lambda) \\ 0 & A_o(\lambda) & -B_o(\lambda) \\ 0 & C_o(\lambda) & D_o(\lambda) \end{bmatrix},$$

where \tilde{W} is of the form $\tilde{W} := \begin{bmatrix} \tilde{W}_{11} & 0 & \tilde{W}_{13} \\ 0 & I_{d_o} & 0 \\ \tilde{W}_{31} & 0 & \tilde{W}_{33} \end{bmatrix} \in \mathbb{C}^{(d_{\bar{\sigma}}+d_o+m) \times (d_{\bar{\sigma}}+d_o+m)}$, $d_{\bar{\sigma}}$ is the

number of (finite and infinite) eigenvalues of $\begin{bmatrix} A(\lambda) \\ C(\lambda) \end{bmatrix}$, and $X_{\bar{\sigma}}(\lambda) \in \mathbb{C}[\lambda]^{d_{\bar{\sigma}} \times d_{\bar{\sigma}}}$ is a regular pencil. Moreover,

- the eigenvalues of $\begin{bmatrix} A(\lambda) \\ C(\lambda) \end{bmatrix}$ are the eigenvalues of $X_{\bar{\sigma}}(\lambda)$,
- $\begin{bmatrix} A_o(\lambda) \\ C_o(\lambda) \end{bmatrix} \in \mathbb{C}[\lambda]^{(d_o+m) \times d_o}$ has no (finite or infinite) eigenvalues,
- the quadruple $\{\tilde{A}_o(\lambda), B_o(\lambda), \tilde{C}_o(\lambda), D_o(\lambda)\}$ is a realization of the transfer function $R_o(\lambda) := \tilde{W}_{33}R(\lambda)$, with $\tilde{W}_{33} \in \mathbb{C}^{m \times m}$ invertible, and
- if $\begin{bmatrix} A(\lambda) & -B(\lambda) \end{bmatrix}$ has no finite or infinite eigenvalues then $\begin{bmatrix} A_o(\lambda) & -B_o(\lambda) \end{bmatrix}$ also has no finite or infinite eigenvalues.

In order to extract from the system quadruple $\{A(\lambda), B(\lambda), C(\lambda), D(\lambda)\}$ a subsystem $\{A_{co}(\lambda), B_{co}(\lambda), C_{co}(\lambda), D_{co}(\lambda)\}$ that is both strongly E-controllable and E-observable (and hence also strongly minimal), we only need to apply the above two theorems one after the other. The resulting subsystem would then be a realization of the transfer function $R_{co} = C_{co}(\lambda)A_{co}(\lambda)^{-1}B_{co}(\lambda) + D_{co}(\lambda) = W_\ell R(\lambda)W_r \in \mathbb{C}(\lambda)^{m \times n}$. Since the transfer function was changed only by left and right transformations that are constant and invertible, the left and right nullspace will be transformed by these invertible transformations, but their minimal indices will be unchanged.

5 Computational Aspects

In this section we give a more “algorithmic” description of the procedure described in Sect. 4 to reduce a given system quadruple $\{A(\lambda), B(\lambda), C(\lambda), D(\lambda)\}$ to a strongly E-controllable quadruple $\{A_c(\lambda), B_c(\lambda), C_c(\lambda), D_c(\lambda)\}$ of smaller size. We describe the essence of the three steps that were discussed in that section.

Step 1 : Compute the staircase reduction of the submatrix $[A(\lambda) \ -B(\lambda)]$

$$U [A(\lambda) \ -B(\lambda)] W^* = \left[\begin{array}{c|cc} X(\lambda) & 0 & 0 \\ \hline Y(\lambda) & \hat{A}(\lambda) & -\hat{B}(\lambda) \end{array} \right].$$

Step 2 : Compute the unitary matrices V and \tilde{W} to compress the first block row of W

$$\begin{bmatrix} W_{11} & W_{12} & W_{13} \end{bmatrix} \begin{bmatrix} V & 0 \\ 0 & I_n \end{bmatrix} \begin{bmatrix} \tilde{W}_{11} & 0 & \tilde{W}_{13} \\ 0 & I_{d-r} & 0 \\ \tilde{W}_{31} & 0 & \tilde{W}_{33} \end{bmatrix} = \begin{bmatrix} I_r & 0 & 0 \end{bmatrix},$$

where V does the compression $[W_{11} \ W_{12}] V = [\tilde{W}_{11}^* \ 0]$ of the first two blocks and \tilde{W} does the further reduction of the first block row to $[I_r \ 0 \ 0]$.

Step 3 : Display the uncontrollable part $X(\lambda)$ using the transformations U , V and \tilde{W}

$$\begin{bmatrix} U & 0 \\ 0 & I_m \end{bmatrix} \begin{bmatrix} A(\lambda) & -B(\lambda) \\ C(\lambda) & D(\lambda) \end{bmatrix} \begin{bmatrix} V & 0 \\ 0 & I_n \end{bmatrix} \tilde{W} = \begin{bmatrix} X_{\bar{c}}(\lambda) & 0 & 0 \\ \times & A_c(\lambda) & -B_c(\lambda) \\ \times & C_c(\lambda) & D_c(\lambda) \end{bmatrix},$$

where we have used the notations introduced in Sect. 4, and the resulting \times entries are of no interest because they do not contribute to the transfer function $R_c(\lambda) := C_c(\lambda)A_c(\lambda)^{-1}B_c(\lambda) + D_c(\lambda)$.

The computational complexity of these three steps is cubic in the dimensions of the matrices that are involved, provided that the staircase algorithm is implemented in an efficient manner [1]. But it is also important to point out that the reduction procedure to extract a strongly minimal linear system matrix from an arbitrary one, can be done with unitary transformations only, and that only one staircase reduction is needed when one knows that the pencil $[A(\lambda) \ -B(\lambda)]$ has normal rank equal to its number of rows. Indeed, this pencil then does not have any left null space or left minimal indices and only the regular part has to be separated from the right null space structure. This can be obtained by performing one staircase reduction on the *rotated* pencil $[\tilde{A}(\mu) \ -\tilde{B}(\mu)]$, where the coefficient matrices

$$\begin{bmatrix} \tilde{A}_0 \\ \tilde{A}_1 \end{bmatrix} = \begin{bmatrix} cI & sI \\ -sI & cI \end{bmatrix} \begin{bmatrix} A_0 \\ A_1 \end{bmatrix}, \quad \begin{bmatrix} \tilde{B}_0 \\ \tilde{B}_1 \end{bmatrix} = \begin{bmatrix} cI & sI \\ -sI & cI \end{bmatrix} \begin{bmatrix} B_0 \\ B_1 \end{bmatrix}, \quad c^2 + s^2 = 1$$

correspond to a change of variable $\lambda = (c\mu - s)/(s\mu + c)$. If one now chooses the rotation such that the rotated pencil has no eigenvalues at $\mu = \infty$, then only the finite spectrum has to be separated from the right minimal indices, which can be done with one staircase reduction [7].

6 Numerical Results

We illustrate the results of this paper with a polynomial example and a rational one.

Example 2 We consider the 2×2 polynomial matrix $P(\lambda) = \text{diag}(e_1(\lambda), e_5(\lambda))$, where $e_5(\lambda)$ is a polynomial of degree 5 with coefficients $[9.6367e - 01, -5.4026e - 07, 2.6333e - 01, -1.1101e - 04, -2.9955e - 04, 4.4650e - 02]$, ordered by descending powers of λ , and $e_1(\lambda)$ is a polynomial of degree 1 with coefficients $[-2.1886e - 03, -1.0000e + 00]$, that were randomly chosen. Expanding this fifth order polynomial matrix as

$$P(\lambda) = P_0 + P_1\lambda + \dots + P_5\lambda^5,$$

a linear system matrix $S_P(\lambda)$ of $P(\lambda)$ is given by the following 10×10 pencil:

$$S_P(\lambda) = \left[\begin{array}{ccc|c} I_2 & -\lambda I_2 & & P_1 \\ & I_2 & -\lambda I_2 & P_2 \\ & & I_2 & -\lambda I_2 \\ & & & I_2 \\ \hline & & & P_4 + \lambda P_5 \\ -\lambda I_2 & & & P_0 \end{array} \right].$$

The six finite Smith zeros of $P(\lambda)$ are clearly those of the scalar polynomials $e_1(\lambda)$ and $e_5(\lambda)$. These are also the finite zeros of $S_P(\lambda)$, since $S_P(\lambda)$ is minimal. However, $S_P(\lambda)$ is not strongly minimal if P_5 is singular and, in fact, it has 4 eigenvalues at infinity (in the sense of [4]). But in the McMillan sense, $P(\lambda)$ has *no* infinite zeros. The deflation procedure that we derived in this paper precisely gets rid of the *extraneous* infinite eigenvalues of $S_P(\lambda)$. The numerical tests show that the sensitivity of the true McMillan zeros also can benefit from this.

In this example we compare the roots computed by four different methods:

1. computing the roots of the scalar polynomials and appending four ∞ roots,
2. computing the generalized eigenvalues of $S_P(\lambda)$,
3. computing the roots of $QS_P(\lambda)Z$ for random orthogonal matrices Q and Z ,
4. computing the roots of the *minimal* pencil obtained by our method.

The first column are the so-called ‘‘correct’’ eigenvalues λ_i , corresponding to the first method, the next three columns are the corresponding errors $\delta_i^{(k)} := |\lambda_i - \hat{\lambda}_i^{(k)}|$, $k = 2, 3, 4$, of the above three methods.² The *extraneous* eigenvalues that are deflated in our approach are put between brackets (Table 1).

We notice that for the largest finite eigenvalue of the order of 10^2 the QZ algorithm applied to $S_P(\lambda)$ gets 14 digits of relative accuracy but, when deflating the four uncontrollable eigenvalues at ∞ , our method recovers a relative accuracy of 16 digits.

² An error $\delta_i^{(k)}$ is NaN when it is the indeterminate form $\text{Inf} - \text{Inf}$. However, some of the eigenvalues at ∞ are computed as a large but finite number and, then, the corresponding error is Inf.

Table 1 The correct generalized λ_i and the corresponding accuracies δ_i^k for the three different calculations

| λ_i | $\delta_i^{(2)}$ | $\delta_i^{(3)}$ | $\delta_i^{(4)}$ |
|-----------------------------|------------------|------------------|------------------|
| -4.5811e-01 | 2.7756e-16 | 4.4409e-16 | 1.1102e-16 |
| 3.5076e-01 + 3.5785e-01i | 9.5020e-16 | 1.1102e-16 | 4.0030e-16 |
| 3.5076e-01 - 3.5785e-01i | 9.5020e-16 | 1.1102e-16 | 4.0030e-16 |
| -1.2170e-01 + 6.2287e - 01i | 6.7589e-16 | 7.8945e-16 | 2.2248e-16 |
| -1.2170e-01 - 6.2287e-01i | 6.7589e-16 | 7.8945e-16 | 2.2248e-16 |
| -4.5691e+02 | 2.9559e-12 | 2.7285e-12 | 5.6843e-14 |
| Inf | NaN | NaN | (Inf) |
| Inf | NaN | NaN | (Inf) |
| Inf | NaN | NaN | (Inf) |
| Inf | NaN | NaN | (Inf) |

Example 3 The second example is the rational matrix $R(\lambda)$ in (7) with $c = 1$.

$$R(\lambda) = \begin{bmatrix} e_5(\lambda) & 0 \\ 1/\lambda & e_1(\lambda) \end{bmatrix} = P_0 + P_1\lambda + \dots + P_5\lambda^5 + \begin{bmatrix} 0 & 0 \\ 1/\lambda & 0 \end{bmatrix},$$

by using the notation of the example above. In this case, $e_5(\lambda)$ has the row vector $[4.7865e - 02, 1.4279e - 04, 2.4361e - 03, -1.5336e - 02, -9.9155e - 01, 1.1948e - 01]$ as coefficients, and $e_1(\lambda)$ has the row vector $[6.5250e - 03, 9.9997e - 01]$. We consider the 12×12 linear system matrix

$$S_R(\lambda) = \left[\begin{array}{cccc|c} \lambda I_2 - A & & & & -B \\ & I_2 & -\lambda I_2 & & P_1 \\ & & I_2 & -\lambda I_2 & P_2 \\ & & & I_2 & -\lambda I_2 \\ & & & & I_2 \\ \hline C & -\lambda I_2 & & & P_0 \end{array} \right],$$

where

$$A = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \quad C = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

is a non-minimal realization of the strictly proper rational function $1/\lambda$. In fact, the matrix A in the realization triple (A, B, C) has two eigenvalues at $\lambda = 0$, of which one is uncontrollable since $1/\lambda$ only has a pole at 0 of order 1. This is an artificial example since we could have realized the strictly proper part by using a minimal triple (A, B, C) by removing the uncontrollable eigenvalue, but this is precisely what our reduction procedure does simultaneously for finite and infinite uncontrollable eigenvalues. The quantities given in the following table are defined

Table 2 The correct generalized λ_i and the corresponding accuracies δ_i^k for the three different calculations

| λ_i | $\delta_i^{(2)}$ | $\delta_i^{(3)}$ | $\delta_i^{(4)}$ |
|---------------------------|------------------|------------------|------------------|
| 0 | 0 | 8.1752e-09 | (4.5874e-16) |
| 0 | 3.6752e-18 | 8.1752e-09 | 5.3729e-16 |
| 1.2028e-01 | 1.8041e-16 | 9.7145e-17 | 9.7145e-17 |
| 2.1135e+00 | 1.7764e-15 | 2.6645e-15 | 1.3323e-15 |
| -2.1404e+00 | 1.7764e-15 | 2.2204e-15 | 8.8818e-16 |
| -4.8180e-02 + 2.1412e+00i | 2.3216e-15 | 1.7990e-15 | 4.0614e-15 |
| -4.8180e-02 - 2.1412e+00i | 2.3216e-15 | 1.7990e-15 | 4.0614e-15 |
| -1.5325e+02 | 2.5580e-13 | 1.5321e-07 | 5.6843e-14 |
| Inf | NaN | Inf | (Inf) |
| Inf | NaN | Inf | (Inf) |
| Inf | NaN | NaN | (NaN) |
| Inf | NaN | NaN | (NaN) |

as in the previous example, except that we added two roots at 0 corresponding to the “exact” eigenvalues (Table 2).

In this example the QZ algorithm applied to $S_R(\lambda)$ recovers well all generalized eigenvalues. When applying the QZ algorithm to an orthogonally equivalent pencil $QS_R(\lambda)Z$, the Jordan block at 0 gets perturbed to two roots of the order of the square root of the machine precision, which can be expected. But when deflating the uncontrollable eigenvalue at 0, this Jordan block is reduced to a single eigenvalue and part of the accuracy gets restored.

These two examples show that deflating uncontrollable eigenvalues may improve the sensitivity of the remaining eigenvalues which may improve the accuracy of their computation.

7 Conclusion

In this paper we looked at quadruple realizations $\{A(\lambda), B(\lambda), C(\lambda), D(\lambda)\}$ for a given rational transfer function $R(\lambda) = C(\lambda)A(\lambda)^{-1}B(\lambda) + D(\lambda)$, where the matrices $A(\lambda)$, $B(\lambda)$, $C(\lambda)$ and $D(\lambda)$ are pencils, and where $A(\lambda)$ is assumed to be regular. We showed that under certain minimality assumptions on this quadruple, the poles, zeros and left and right null space structure of the rational matrix $R(\lambda)$ can be recovered from the generalized eigenstructure of two block pencils constructed from the quadruple. We also showed how to obtain such a minimal quadruple from a non-minimal one, by applying a reduction procedure that is based on the staircase algorithm. These results extend those previously obtained for generalized state space systems and polynomial matrices.

Acknowledgements We would like to thank the anonymous reviewer whose helpful comments and suggestions have greatly improved this manuscript. The first author was supported by “Ministerio de Economía, Industria y Competitividad (MINECO)” of Spain and “Fondo Europeo de Desarrollo Regional (FEDER)” of EU through grants MTM2015-65798-P and MTM2017-90682-REDT. The second author was funded by the “contrato predoctoral” BES-2016-076744 of MINECO. This work was developed while the third author held a “Chair of Excellence UC3M - Banco de Santander” at Universidad Carlos III de Madrid in the academic year 2017–2018.

References

1. Beelen, T., Van Dooren, P.: An improved algorithm for the computation of Kronecker’s canonical form of a singular pencil. *Linear Algebra Appl.* **105**, 9–65 (1988)
2. Forney, G.D.: Minimal bases of rational vector spaces, with applications to multivariable linear systems. *SIAM J. Control* **13**, 493–520 (1975)
3. Gantmacher, F.R.: *The Theory of Matrices*, Vol. I and II (transl.). Chelsea, New York (1959)
4. Gohberg, I., Lancaster, P., Rodman, L.: *Matrix Polynomials*. SIAM Publications (2009). Originally published: Academic Press, New York (1982)
5. Kailath, T.: *Linear Systems*. Prentice Hall, Englewood Cliffs, NJ (1980)
6. Rosenbrock, H.: *State-Space and Multivariable Theory*. Wiley, New York (1970)
7. Van Dooren, P.: The computation of Kronecker’s canonical form of a singular pencil. *Linear Algebra Appl.* **27**, 103–141 (1979)
8. Van Dooren, P.: The generalized eigenstructure problem in linear system theory. *IEEE Trans. Aut. Control* **26**(1), 111–129 (1981)
9. Van Dooren, P., Dewilde, P.: The eigenstructure of an arbitrary polynomial matrix: computational aspects. *Linear Algebra Appl.* **50**, 545–579 (1983)
10. Verghese, G., Van Dooren, P., Kailath, T.: Properties of the system matrix of a generalized state-space system. *Int. J. Control* **30**(2), 235–243 (1979)
11. Verghese, G.: Comments on ‘Properties of the system matrix of a generalized state-space system’. *Int. J. Control* **31**(5), 1007–1009 (1980)

Nonlinear Dynamical Systems

Interpolation-Based Model Order Reduction for Quadratic-Bilinear Systems and \mathcal{H}_2 Optimal Approximation



Xingang Cao, Joseph Maubach, Wil Schilders, and Siep Weiland

Abstract The work of this paper focuses on model order reduction for a special class of nonlinear dynamical systems, that is, the class of quadratic-bilinear dynamical systems. This kind of systems can be used to represent other nonlinear dynamical systems with strong nonlinearities such as exponent and high-order polynomials. This paper addresses the \mathcal{H}_2 optimal model approximation problem for this class of systems. To solve the model order reduction problem, a notion of generalized transfer functions and the \mathcal{H}_2 norm are first discussed. A Volterra series interpolation scheme is proposed to interpolate the system from both the input-to-output and the output-to-input directions. In contrast to existing methods, we propose to interpolate all Volterra kernels, which can be achieved by solving Sylvester equations. The necessary \mathcal{H}_2 optimality conditions are fulfilled by the proposed interpolation scheme. A fixed point method is applied to solve the nonlinear Sylvester equations. A numerical example demonstrates the effectiveness of the proposed methods.

Keywords Model order reduction · Quadratic bilinear dynamical systems · \mathcal{H}_2 Optimal approximation · Volterra series interpolation · Generalized Sylvester equations

X. Cao (✉) · J. Maubach · W. Schilders
Department of Mathematics and Computer Science, Eindhoven University of Technology,
Eindhoven, The Netherlands
e-mail: j.m.l.maubach@tue.nl

W. Schilders
e-mail: w.h.a.schilders@tue.nl

S. Weiland
Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The
Netherlands
e-mail: s.weiland@tue.nl

1 Introduction

In many applications of systems and control including twin rotor system [27], continuous stirred tank reactors [13], nonlinear dynamical systems are often approximated by quasi-linear parameter varying (LPV) systems for controller synthesis. However, the nature of these systems is nonlinear. From the modeling point of view, directly considering the nonlinear dynamics is more reasonable and also enables us to reduce the complexity of these systems. By making polynomial reformulations of the nonlinearities [23] and augmenting the state space, many nonlinear dynamical systems such as the nonlinear transmission line circuit [23], the FitzHugh-Nagumo model [11] and the spring-mass-damper system [25] with a duffing spring, can be brought into the quadratic-bilinear (QB) dynamical system form

$$\Sigma: \begin{cases} \dot{x}(t) = Ax(t) + H(x(t) \otimes x(t)) + \sum_{q=1}^{n_i} N_q u_q(t)x(t) + Bu(t), \\ y(t) = Cx(t), \quad x(t_0) = x_0. \end{cases} \quad (1)$$

where $x(t) \in \mathbb{R}^n$ is the state, $u(t) = (u_1(t), \dots, u_{n_i}(t))^\top \in \mathbb{R}^{n_i}$ is the input, $y(t) \in \mathbb{R}^{n_o}$ is the output and \otimes denotes Kronecker product. The real-valued matrices A, B, H, N_q and C are of compatible dimensions. The sparse matrix $H \in \mathbb{R}^{n \times n^2}$ is called the Hessian matrix and usually, by construction, is symmetric, which, in this context, means that for all vectors $v, u \in \mathbb{R}^n$, we have

$$H(v \otimes u) = H(u \otimes v). \quad (2)$$

If H is not symmetric, it can be symmetrized by the method proposed in [4]. The model order reduction problem is to find another QB system

$$\widehat{\Sigma}: \begin{cases} \hat{x}(t) = \widehat{A}\hat{x}(t) + \widehat{H}(\hat{x}(t) \otimes \hat{x}(t)) + \sum_{q=1}^{n_i} \widehat{N}_q u_q(t)\hat{x}(t) + \widehat{B}u(t), \\ \hat{y}(t) = \widehat{C}\hat{x}(t), \quad \hat{x}(t_0) = \hat{x}_0, \end{cases} \quad (3)$$

with state space dimension $r \ll n$ meanwhile the system response $\hat{y}(t)$ is close enough to $y(t)$ in (1).

In recent years, model order reduction attracts lots of attention [2], especially for nonlinear systems, see, for example [3, 8, 21]. In [23], the author proposed a method to reformulate nonlinear vector fields by quadratic nonlinearities. Then by applying the *variational equation approach* in [28], the generalized transfer functions based on Laplace transform of Volterra kernels can be obtained. A nonlinear Krylov subspace method is then used to construct Galerkin projections to reduce the system. The problem was again addressed in [1, 4]. Nonlinear Krylov subspace methods are considered to construct Petrov-Galerkin projections for model order reduction. The Petrov-Galerkin projection method is more accurate, but typically does not preserve the stability properties of the system. See, for example, [17, 24]. In [10], it is shown that if the first two transfer functions and the quadratic part of the third transfer function are of interest, rational Krylov method, such as the rational interpolation

[22] for single-input single-output (SISO) systems and the tangential interpolation [20] for multiple-input multiple-output (MIMO) systems, are adequate to reduce the system. This offers an interesting computationally attractive alternative to reduce QB systems. From controllability and observability energy point of view [5, 18, 29], balanced truncation and its truncated version were proposed in [6]. In [8], again the first two transfer functions and the quadratic part of the third generalized transfer function are considered and a quasi-optimal \mathcal{H}_2 norm approximation method was proposed.

In this paper, we assume zero initial condition, i.e., $x_0 = 0$ in (1) and then generalize the work of [15] for bilinear dynamical systems to reduce QB systems. An interpolation scheme is proposed to interpolate the zeroth and first order moments. With the help of the \mathcal{H}_2 norm defined for QB systems, the proposed interpolation scheme can be modified to approximate the system so as to satisfy the necessary conditions for \mathcal{H}_2 -optimal approximants. Both the interpolation and the \mathcal{H}_2 optimal approximation are achieved by solving the generalized QB Sylvester equations.

We concentrate the discussion with SISO QB systems. One can see that the proposed method can be extended to MIMO systems by considering the tangential interpolation approach [20]. For details and numerical examples, we refer to [9]. Organization of this paper is as follows. In Sect. 2, the generalized transfer functions of the original QB system and its adjoint are presented. So is the definition of the \mathcal{H}_2 norm of a QB system. This section is closed by discussing the pole-residue representation of the generalized transfer functions and an alternative to express the \mathcal{H}_2 norm. Section 3 begins with the discussion on the moments definition and the Volterra series interpolation of QB systems. Then it is shown that \mathcal{H}_2 optimal approximation of a QB system can be achieved by the proposed interpolation scheme. In Sect. 4, a numerical case study illustrates the performance of the proposed methods. Section 5 concludes the paper.

2 Generalized Transfer Functions and the \mathcal{H}_2 Norm

Consider a SISO QB system Σ given in (1), i.e., with $n_i = n_o = 1$. By rewriting the quadratic term $H(x \otimes x)$ as $H(x \otimes I)x$, the state equation can be expressed as

$$\dot{x}(t) = (A + H(x(t) \otimes I))x(t) + (Nx(t) + B)u(t),$$

which is in the form of a linear analytic system as advocated in [28]. Therefore, by applying the *variational equation approach* [28], the generalized transfer functions can be derived. The input-to-state generalized transfer matrices are [8]

$$\begin{aligned}
F_1(s_1) &= (s_1 I - A)^{-1} B, \\
F_2(s_1, s_2) &= (s_2 I - A)^{-1} N (s_1 I - A)^{-1} B, \\
F_k(s_1, \dots, s_k) &= \left((s_k I - A)^{-1} H \left(\sum_{\substack{i,j \geq 1 \\ i+j=k-1}} F_i(s_{k-i-1}, \dots, s_{k-1}) \otimes F_j(s_1, \dots, s_j) \right), \right. \\
&\quad \left. (s_k I - A)^{-1} N F_{k-1}(s_1, \dots, s_{k-1}) \right), \quad k \geq 3.
\end{aligned} \tag{4}$$

Correspondingly, the input-to-output generalized transfer functions are

$$G_k(s_1, \dots, s_k) = C F_k(s_1, \dots, s_k), \quad k \in \mathbb{Z}_+. \tag{5}$$

To construct a Petrov-Galerkin projection scheme for model order reduction, we also consider the adjoint system. Based on the work of [19], the authors of [6] showed that the input-to-state generalized transfer functions of the adjoint system are

$$\begin{aligned}
F_{a,1}(s_1) &= (s_1 I - A^\top)^{-1} C^\top, \\
F_{a,2}(s_1, s_2) &= (s_2 I - A^\top)^{-1} N^\top (s_1 I - A^\top)^{-1} C^\top, \\
F_{a,k}(s_1, \dots, s_k) &= \left((s_k I - A^\top)^{-1} H^{(2)} \left(\sum_{\substack{i,j \geq 1 \\ i+j=k-1}} F_i(s_{k-i-1}, \dots, s_{k-1}) \otimes F_{a,j}(s_1, \dots, s_j) \right), \right. \\
&\quad \left. (s_k I - A^\top)^{-1} N^\top F_{a,k-1}(s_1, \dots, s_{k-1}) \right), \quad k \geq 3,
\end{aligned} \tag{6}$$

where $H^{(2)}$ is the mode-2 matricization of an order-3 tensor whose mode-1 matricization is given by the Hessian matrix H . For detailed explanations, we refer to [4, 6]. The input-to-output generalized transfer functions of the adjoint system are simply

$$G_{a,k}(s_1, \dots, s_k) = B^\top F_{a,k}(s_1, \dots, s_k), \quad k \in \mathbb{Z}_+.$$

To quantify the approximation accuracy of the reduced-order model, an appropriate measure of the approximation error is required. In this paper, the focus is on the \mathcal{H}_2 norm.

Definition 1 Consider the SISO QB system Σ given in (1), the \mathcal{H}_2 norm of Σ is defined as

$$\|\Sigma\|_{\mathcal{H}_2} := \left(\sum_{k=1}^{\infty} \frac{1}{(2\pi)^k} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \langle G_k(i\omega_1, \dots, i\omega_k), G_k(-i\omega_1, \dots, -i\omega_k) \rangle d\omega_1 \cdots d\omega_k \right)^{\frac{1}{2}}, \tag{7}$$

where $G_k(s_1, \dots, s_k)$ is the k th generalized transfer function defined in (4) and (5) and $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product of two complex-valued vectors.

From Parseval's Identity of several variables, the \mathcal{H}_2 norm of Σ is equal to the \mathcal{L}_2 norm of the Volterra kernels associated in (4) and (5) in the time domain. In [8], it is shown that the \mathcal{H}_2 norm of QB systems defined above can be computed in terms of the Gramians of the system as

$$\|\Sigma\|_{\mathcal{H}_2}^2 = \|\Sigma_a\|_{\mathcal{H}_2}^2 = \|\Sigma\|_{\mathcal{L}_2}^2 = CRC^\top = B^\top QB, \quad (8)$$

where Σ_a is the adjoint of Σ , R and Q solve the generalized Lyapunov equations

$$AR + RA^\top + H(R \otimes R)H^\top + NRN^\top + BB^\top = 0, \quad (9)$$

$$A^\top Q + QA + H^{(2)}(R \otimes Q)(H^{(2)})^\top + N^\top QN + C^\top C = 0. \quad (10)$$

Note that the \mathcal{H}_2 norm is finite if and only if the Gramians R and Q exist and are positive definite. Sufficient conditions on the existence of the Gramians are discussed in [9], which are equivalent to the conditions presented in [6]. This analysis is out of scope for this paper. The interested reader is referred to the aforementioned references.

2.1 Pole-Residue Representation of the Generalized Transfer Functions

In the transfer functions of both the original system and its adjoint, the frequency variables $s_k \in \mathbb{C}$, $k \in \mathbb{Z}_+$ are independent from each other. Hence, the generalized transfer functions in (4) and (6) admit representations in the pole-residue form.

Definition 2 The residues of the k th-order generalized transfer function $G_k(s_1, \dots, s_k)$ which only has simple poles are defined as

$$\phi_{l_k \dots l_1} := \lim_{s_k \rightarrow \lambda_{l_k}} (s_k - \lambda_{l_k}) \cdots \lim_{s_1 \rightarrow \lambda_{l_1}} (s_1 - \lambda_{l_1}) G_k(s_1, s_2, \dots, s_k).$$

Theorem 1 (pole-residue representation) *Re-express the generalized transfer functions $G_k(s_1, \dots, s_k)$ as adjugate over determinant. Then*

$$G_k(s_1, \dots, s_k) = \sum_{l_k=1}^n \cdots \sum_{l_1=1}^n \frac{\phi_{l_k \dots l_1}}{\prod_{i=1}^k (s_i - \lambda_{l_i})}. \quad (11)$$

Proof The proof is given in a constructive way with computation of residues. Assume that the matrix $A = \text{diag}(\lambda_1, \dots, \lambda_n)$ is in diagonal form. For $k = 1, 2$, the computations are quite trivial. For $k = 3$, $G_3 = (G_3^{QD}, G_3^{BL})$. The quadratic part has the transfer function

$$G_3^{QD}(s_1, s_2, s_3) = \sum_{l_3=1}^n \sum_{l_2=1}^n \sum_{l_1=1}^n \frac{c_{l_3} H_{\mathcal{T}l_3l_2l_1} b_{l_2} b_{l_1}}{(s_3 - \lambda_{l_3})(s_2 - \lambda_{l_2})(s_1 - \lambda_{l_1})},$$

where $H_{\mathcal{T}}$ is the order-3 tensor which defines the quadratic coefficient matrix H in system Σ and $H_{\mathcal{T}ijk}$ is the i th row j th column k th layer of the tensor $H_{\mathcal{T}}$, b_{l_i} and c_{l_i} , $i = 1, 2, 3$ are the l_i -th entry of B and C . The bilinear recursive part can be represented as

$$G_3^{BL}(s_1, s_2, s_3) = \sum_{l_3=1}^n \sum_{l_2=1}^n \sum_{l_1=1}^n \frac{c_{l_3} N_{l_3l_2} N_{l_2l_1} b_{l_1}}{(s_3 - \lambda_{l_3})(s_2 - \lambda_{l_2})(s_1 - \lambda_{l_1})}.$$

Correspondingly, the 3rd residues are

$$\phi_{l_3l_2l_1} = (c_{l_3} H_{\mathcal{T}l_3l_2l_1} b_{l_2} b_{l_1} \quad c_{l_3} N_{l_3l_2} N_{l_2l_1} b_{l_1}). \quad (12)$$

Let ψ^k denote the k th residue of the input-to-state transfer functions $F_k(s_1, s_2, \dots, s_k)$. Then for $k \geq 3$, the sum of the residues $\phi_{l_k l_{k-1} \dots l_1}$ satisfies

$$\sum_{l_k=1}^n \dots \sum_{l_1=1}^n \phi_{l_k \dots l_1} = \left(C H_{\mathcal{T}} \left(\sum_{\substack{i,j \geq 1 \\ i+j=k-1}} \psi^i \otimes \psi^j \right), C N \psi^{k-1} \right). \quad (13)$$

Since $G_k(s_1, s_2, \dots, s_k)$ only has simple roots, we can define

$$\check{C}_{l_k} = \left(\frac{c_{l_k}}{\prod_{i=1}^n (s_{l_k} - \lambda_{l_i})} \dots \frac{c_n}{\prod_{k=1}^n (s_k - \lambda_{l_k})} \right).$$

Then from (13) one can show that

$$G_k(s_1, s_2, \dots, s_k) = \left(\check{C}_{l_k} H_{\mathcal{T}} \left(\sum_{\substack{i,j \geq 1 \\ i+j=k-1}} \psi^i \otimes \psi^j \right), \check{C}_{l_k} N \psi^{k-1} \right).$$

Hence, the pole-residue representations in (11) hold for $k \in \mathbb{Z}_+$. \square

The pole-residue representation of the generalized transfer functions gives another possibility to express the \mathcal{H}_2 norm.

Theorem 2 *Let Σ be a SISO QB system given by (1) with a finite \mathcal{H}_2 norm defined by (7). The \mathcal{H}_2 norm satisfies*

$$\|\Sigma\|_{\mathcal{H}_2}^2 = \sum_{k=1}^{\infty} \sum_{l_k=1}^n \dots \sum_{l_1=1}^n \langle \phi_{l_k l_{k-1} \dots l_1}, G_k(-\lambda_{l_k}, -\lambda_{l_{k-1}}, \dots, -\lambda_{l_1}) \rangle. \quad (14)$$

Proof The proof is similar to the proof of Theorem 2.10 in [16]. It is a generalization of the linear system case by applying the multivariate Cauchy's Theorem. \square

Based on the \mathcal{H}_2 norm expression given by (14), the \mathcal{H}_2 approximation error between two QB systems can be expressed in a novel manner. Suppose that the reduced-order approximation of Σ defined in (1) is given by $\widehat{\Sigma}$ defined in (3) with $r \ll n$. Also, assume that the reduced-order system only has simple poles at $\{\tilde{\lambda}_i\}_{i=1}^r$. Then the following holds

$$\begin{aligned} \|\Sigma - \widehat{\Sigma}\|_{\mathcal{H}_2}^2 &= \sum_{k=1}^{\infty} \sum_{l_k=1}^n \cdots \sum_{l_1=1}^n \langle \phi_{l_k \dots l_1}, G_k(-\lambda_{l_k}, \dots, -\lambda_{l_1}) - \widehat{G}_k(-\lambda_{l_k}, \dots, -\lambda_{l_1}) \rangle - \\ &\quad \sum_{k=1}^{\infty} \sum_{l_k=1}^r \cdots \sum_{l_1=1}^r \langle \tilde{\phi}_{l_k \dots l_1}, G_k(-\tilde{\lambda}_{l_k}, \dots, -\tilde{\lambda}_{l_1}) - \widehat{G}_k(-\tilde{\lambda}_{l_k}, \dots, -\tilde{\lambda}_{l_1}) \rangle, \end{aligned}$$

where $\tilde{\phi}_{l_k \dots l_1}$ is the k th residue of the reduced-order system $\widehat{\Sigma}$ and $\widehat{G}_k(s_1, \dots, s_k)$ is the k th reduced-order generalized transfer function. Hence, the approximation error is fully characterized by the mismatches between the original and the reduced-order generalized transfer functions at the mirror images of the poles of the original system and the poles of the reduced-order system. In the \mathcal{H}_2 optimal approximation framework, minimizing the mismatches at the poles of the reduced-order system is of more interest according to the 1st-order optimality conditions [9]. The 1st-order necessary optimality conditions of the \mathcal{H}_2 approximation require to interpolate the zeroth and first-order derivative of the generalized transfer functions at the mirror images of the poles of the reduced-order system. In the following, we propose to solve such an interpolation problem by interpolating the Volterra series.

3 Volterra Series Interpolation and \mathcal{H}_2 Optimal Approximation

The state-of-the-art interpolation methods in [4, 10] only consider to interpolate the first a few transfer functions [1, 8]. In this work, we propose a method to interpolate the weighted sum of all the generalized transfer functions, which is generalized from the bilinear dynamical systems case developed in [15]. In the interpolation scheme proposed in this section, we assume that the interpolation points are neither in the spectrum of A nor in the spectrum of A_r .

3.1 Volterra Series Interpolation

The \mathcal{H}_2 error expression given in Sect. 2 shows that the optimal approximation requires to interpolate the weighted sum of the generalized transfer functions and their first-order derivatives, i.e., the zeroth and first-order moments [9]. Let us start the discussion with the definition of the weighted zeroth-order moments of the

generalized input-to-state transfer matrices (4). Given a set of interpolation points $\{\sigma_i\}_{i=1}^r$ together with a matrix $U \in \mathbb{R}^{r \times r}$ and a symmetric 3-tensor $\Xi_{\mathcal{T}} \in \mathbb{R}^{r \times r \times r}$, we define

$$v_l^1 = (\sigma_l I - A)^{-1} B, \quad (15a)$$

$$v_l^2 = (\sigma_l I - A)^{-1} N \sum_{l_1=1}^r u_{ll_1} (\sigma_{l_1} I - A)^{-1} B, \quad (15b)$$

where u_{ll_1} is the (l, l_1) -entry of U and

$$v_l^k = (\sigma_l I - A)^{-1} \left(N \sum_{l_{k-1}=1}^r u_{ll_{k-1}} v_{l_{k-1}}^{k-1} + H \sum_{\substack{i,j \geq 1 \\ i+j=k-1}} \sum_{l_{k-1}=1}^r \sum_{l_{k-i-1}=1}^r \xi_{ll_{k-1}l_{k-i-1}} \left(v_{l_{k-1}}^i \otimes v_{l_{k-i-1}}^j \right) \right), \quad (15c)$$

with $k \geq 3$, $l = 1, 2, \dots, r$ and $\xi_{ll_{k-1}l_{k-i-1}}$ is the (l, l_{k-1}, l_{k-i-1}) -entry of $\Xi_{\mathcal{T}}$. Assuming convergence, a vector v_l is then defined as $v_l := \sum_{k=1}^{\infty} v_l^k$. Then from input to output, the zeroth-order moments are defined as

$$\gamma_l = \sum_{k=1}^{\infty} \gamma_l^k, \quad \gamma_l^k = C v_l^k, \quad k \in \mathbb{Z}_+, \quad l = 1, 2, \dots, r. \quad (16)$$

We also consider the generalized transfer functions of the adjoint system in (6). Suppose that the interpolation points are $\{\hat{\sigma}_i\}_{i=1}^r$ and the weighting matrix and tensor are given as \hat{U} and $\hat{\Xi}_{\mathcal{T}}$, respectively. Similar to v_l^k , another set of vectors is defined as

$$w_l^1 = (\hat{\sigma}_l I - A)^{-\top} C^{\top}, \quad (17a)$$

$$w_l^2 = (\hat{\sigma}_l I - A)^{-\top} N^{\top} \sum_{l_1=1}^r \hat{u}_{ll_1} (\hat{\sigma}_{l_1} I - A)^{-\top} C^{\top}, \quad (17b)$$

$$w_l^k = (\hat{\sigma}_l I - A)^{-\top} \left(N^{\top} \sum_{l_{k-1}=1}^r \hat{u}_{ll_{k-1}} w_{l_{k-1}}^{k-1} + H^{(2)} \sum_{\substack{i,j \geq 1 \\ i+j=k-1}} \sum_{l_{k-1}=1}^r \sum_{l_{k-i-1}=1}^r \hat{\xi}_{ll_{k-1}l_{k-i-1}} \left(w_{l_{k-1}}^i \otimes w_{l_{k-i-1}}^j \right) \right), \quad (17c)$$

with $k \geq 3$ and $l = 1, 2, \dots, r$. Assuming convergence, a vector w_l is defined as $w_l := \sum_{k=1}^{\infty} w_l^k$. Hence, from the output-to-input direction, the zeroth-order moments are defined as

$$\zeta_l = \sum_{k=1}^{\infty} \zeta_l^k, \quad \zeta_l^k = B^{\top} w_l^k, \quad k \in \mathbb{Z}_+, \quad l = 1, 2, \dots, r. \quad (18)$$

Note that here, ζ_l can be considered as the frequency response of the *adjoint* system at frequency $\hat{\sigma}_l$. Later on, one can see that interpolating ζ_l means the generalized transfer functions of the *adjoint* system are interpolated. We use the vectors v_l and w_l to construct the Krylov subspaces that define the Petrov-Galerkin spaces for the model order reduction. In addition, we will show that the recursive formula gives a possibility to construct the vectors v_l and w_l by solving generalized Sylvester equations.

Lemma 1 *Let Σ be a stable SISO QB system given by (1) with dimension n . Given the interpolation points $\{\sigma_i\}_{i=1}^r \in \mathbb{C}$ and $\{\hat{\sigma}_i\}_{i=1}^r \in \mathbb{C}$ with $r \ll n$ together with two matrices $U, \hat{U} \in \mathbb{R}^{r \times r}$ and two symmetric tensors $\Xi_{\mathcal{T}}, \hat{\Xi}_{\mathcal{T}} \in \mathbb{R}^{r \times r \times r}$ such that the vectors v_l^k are defined according to (15a)–(15c) and the vectors w_l^k are defined as in (17a)–(17c). For each σ_l and $\hat{\sigma}_l$, the expressions $v_l = \sum_{k=1}^{\infty} v_l^k$ and $w_l = \sum_{k=1}^{\infty} w_l^k$ converge. Let the matrices $\text{diag}(\sigma_1, \dots, \sigma_r)$ and $\text{diag}(\hat{\sigma}_1, \dots, \hat{\sigma}_r)$ be denoted as Λ and $\hat{\Lambda}$, respectively. Then the matrices $V = (v_1, \dots, v_r) \in \mathbb{R}^{n \times r}$ and $W = (w_1, \dots, w_r) \in \mathbb{R}^{n \times r}$ solve the generalized Sylvester equations*

$$V\Lambda - AV - NVU^{\top} - H(V \otimes V)(\Xi_{\mathcal{T}}^{(1)})^{\top} = B\mathbb{1}_r^{\top}, \quad (19)$$

$$W\hat{\Lambda} - A^{\top}W - N^{\top}W\hat{U}^{\top} - H^{(2)}(V \otimes W)(\hat{\Xi}_{\mathcal{T}}^{(1)})^{\top} = C^{\top}\mathbb{1}_r^{\top}, \quad (20)$$

where $\mathbb{1}_r$ is a column vector of r ones.

Proof The proof follows the construction of $v_l = \sum_{k=1}^{\infty} v_l^k$ and $w_l = \sum_{k=1}^{\infty} w_l^k$ directly. \square

Theorem 3 *Let Σ be a stable SISO QB system given by (1) with dimension n . Given the interpolation conditions such that V, W solve the generalized Sylvester equations (19) and (20), respectively. If $V^{\top}W$ is invertible, the system matrices of the reduced-order model $\hat{\Sigma}$ in (3) are computed by*

$$\begin{aligned} \hat{A} &= (W^{\top}V)^{-1}W^{\top}AV, \quad \hat{N} = (W^{\top}V)^{-1}W^{\top}NV, \quad \hat{H} = (W^{\top}V)^{-1}W^{\top}H(V \otimes V), \\ \hat{B} &= (W^{\top}V)^{-1}W^{\top}B, \quad \hat{C} = CV. \end{aligned} \quad (21)$$

Then the resulting reduced-order system (3) interpolates the zeroth-order moments of both the original system and its adjoint, i.e., $\hat{\gamma}_l = \gamma_l$ and $\hat{\zeta}_l = \zeta_l$ for $l = 1, 2, \dots, r$.

Proof We follow the method in the proof of Theorem 3.1 in [15]. Define the projector $\Pi = V(W^{\top}V)^{-1}W^{\top}$. Then left-multiplying (19) by Π we have

$$\begin{aligned} &\Pi \left(V\Lambda - AV - NVU^{\top} - H(V \otimes V)(\Xi_{\mathcal{T}}^{(1)})^{\top} - B\mathbb{1}_r^{\top} \right) \\ &= V \left(\Lambda - \hat{A} - \hat{N}U^{\top} - \hat{H}(\Xi_{\mathcal{T}}^{(1)})^{\top} - \hat{B}\mathbb{1}_r^{\top} \right) = 0. \end{aligned}$$

Since V has full column rank, it is immediate that $\widehat{V} = I_r$ solves

$$\widehat{V}\Lambda - \widehat{A}\widehat{V} - \widehat{N}\widehat{V}U^\top - \widehat{H}(\widehat{V} \otimes \widehat{V})(\widehat{\Xi}_{\mathcal{T}}^{(1)})^\top - \widehat{B}\mathbb{1}_r^\top = 0.$$

Hence, the construction of \widehat{v}_l directly follows the construction method of v_l . Then

$$\widehat{\gamma}_l := \widehat{C}\widehat{v}_l = CV\widehat{v}_l = Cv_l := \gamma_l.$$

For the matrix W , left-multiplying (20) by Π^\top , it can be obtained that

$$\begin{aligned} & \Pi^\top \left(W\widehat{\Lambda} - A^\top W - N^\top W\widehat{U}^\top - H^{(2)}(V \otimes W)(\widehat{\Xi}_{\mathcal{T}}^{(1)})^\top - C^\top \mathbb{1}_r^\top \right) \\ &= W(V^\top W)^{-1}. \\ & \left((V^\top W)\widehat{\Lambda} - \widehat{A}^\top (V^\top W) - \widehat{N}^\top (V^\top W)\widehat{U}^\top - \widehat{H}^{(2)}(I_r \otimes (V^\top W))(\widehat{\Xi}_{\mathcal{T}}^{(1)})^\top - \widehat{C}^\top \mathbb{1}_r^\top \right) \\ &= 0. \end{aligned}$$

Note that for the tensor product term we used the fact that

$$\begin{aligned} V^\top H^{(2)}(V \otimes W) &= V^\top H^{(2)}(V \otimes W(V^\top W)^{-1}(V^\top W)) \\ &= V^\top H^{(2)}(V \otimes W(V^\top W)^{-1})(I_r \otimes (V^\top W)) = \widehat{H}^{(2)}(I_r \otimes (V^\top W)). \end{aligned}$$

Denote $V^\top W$ as Θ . Since $W(V^\top W)^{-1}$ has full column rank, Θ solves

$$\Theta\widehat{\Lambda} - \widehat{A}^\top \Theta - \widehat{N}\Theta\widehat{U}^\top - \widehat{H}^{(2)}(I_r \otimes \Theta)(\widehat{\Xi}_{\mathcal{T}}^{(1)})^\top - \widehat{C}^\top \mathbb{1}_r^\top = 0.$$

Hence, the construction of θ_l directly follows the construction of w_l . Then we have

$$\widehat{\zeta}_l := \widehat{B}^\top \theta_l = B^\top W(V^\top W)^{-1} \theta_l = B^\top w_l := \zeta_l,$$

which completes the proof. \square

As a result, by specifying suitable interpolation conditions, the interpolation-based model reduction problem of QB systems can be solved by solving the generalized Sylvester equations (19) and (20). Furthermore, by applying the Petrov-Galerkin projection scheme, generalized transfer functions of both the original and the *adjoint* systems are interpolated.

The following shows that the zeroth-order moments defined for the original system and its adjoint coincide with each other for QB systems if they are defined in the aforementioned way.

Corollary 1 *For the interpolation conditions proposed in Lemma 1, if $\widehat{\sigma}_l = \sigma_l$, $l = 1, 2, \dots, r$, $\widehat{U} = U^\top$ and $\widehat{\Xi}_{\mathcal{T}}^{(1)} = \Xi_{\mathcal{T}}^{(2)}$, then $\gamma_l = \zeta_l$, $l = 1, 2, \dots, r$.*

Proof The detailed proof can be found in [9, Chap. 5]. \square

3.2 \mathcal{H}_2 Optimal Approximation

Assume that the reduced-order system given by (3) only has simple poles $\{\tilde{\lambda}_l\}_{l=1}^r$ and the matrix \widehat{A} is diagonalizable. Then we can assume that the reduced-order system is already in the form with a diagonal matrix $\widehat{A} = \Lambda = \text{diag}(\tilde{\lambda}_1, \dots, \tilde{\lambda}_r)$ as desired. Correspondingly, the other system matrices are all transformed by the eigenvectors of \widehat{A} . The following analysis shows that for the proposed interpolation scheme in Sect. 3.1, the first-order necessary optimality conditions for \mathcal{H}_2 approximation are generalizations of those proposed for LTI systems by Meier and Luenberger in [26].

Lemma 2 *Let Σ and $\widehat{\Sigma}$ be stable SISO QB systems with dimensions n and r with $r \ll n$. And $T^{-1}\widehat{A}T = \Lambda$ is a diagonal matrix. Correspondingly, $\tilde{N} = T^{-1}\widehat{N}T$, $\tilde{H} = T^{-1}\widehat{H}(T \otimes T)$, $\tilde{B} = T^{-1}\widehat{B}$ and $\tilde{C} = \widehat{C}T$. If V solves the generalized Sylvester equation*

$$V(-\Lambda) - AV - NV\tilde{N}^\top - H(V \otimes V)\tilde{H}^\top = B\tilde{B}^\top, \quad (22)$$

then

$$\tilde{C}(CV)^\top = \sum_{k=1}^{\infty} \sum_{l_1=1}^r \cdots \sum_{l_1=1}^r \langle \tilde{\phi}_{ll_{k-1}\dots l_1}^k, G_k(-\tilde{\lambda}_l, -\tilde{\lambda}_{l_{k-1}}, \dots, -\tilde{\lambda}_{l_1}) \rangle, \quad (23)$$

where $\tilde{\phi}_{ll_{k-1}\dots l_1}^k$ are the residues of $\widehat{\Sigma}$ and G_k are the generalized transfer functions of Σ .

Proof Let v_l denote the l th column of V , since $v_l = \sum_{k=1}^{\infty} v_l^k$ (we only show it for $k = 1, 2, 3$ for demonstration purposes), we have

$$Cv_l^1 = \tilde{b}_l C(-\tilde{\lambda}_l I - A)^{-1} B,$$

$$Cv_l^2 = \sum_{l_1=1}^r \tilde{N}_{ll_1} \tilde{b}_{l_1} C(-\tilde{\lambda}_l I - A)^{-1} N(-\tilde{\lambda}_{l_1} I - A)^{-1} B,$$

$$Cv_l^3 = \sum_{l_2=1}^r \sum_{l_1=1}^r \tilde{N}_{ll_2} \tilde{N}_{l_2 l_1} \tilde{b}_{l_1} C(-\tilde{\lambda}_l I - A)^{-1} N(-\tilde{\lambda}_{l_2} I - A)^{-1} N(-\tilde{\lambda}_{l_1} I - A)^{-1} B +$$

$$\sum_{l_2=1}^r \sum_{l_1=1}^r \tilde{H}_{ll_2 l_1} \tilde{b}_{l_2} \tilde{b}_{l_1} C(-\tilde{\lambda}_l I - A)^{-1} H \left((-\tilde{\lambda}_{l_2} I - A)^{-1} B \otimes (-\tilde{\lambda}_{l_1} I - A)^{-1} B \right),$$

Since the system under consideration is SISO, we have

$$\tilde{C}(CV)^\top = \sum_{k=1}^{\infty} \sum_{l=1}^r \tilde{c}_l Cv_l^k = \sum_{k=1}^{\infty} \sum_{l=1}^r \cdots \sum_{l_1=1}^r \langle \tilde{\phi}_{ll_{k-1}\dots l_1}^k, G_k(-\tilde{\lambda}_l, -\tilde{\lambda}_{l_{k-1}}, \dots, -\tilde{\lambda}_{l_1}) \rangle,$$

where $\tilde{C} = (\tilde{c}_1 \dots \tilde{c}_r)$ is the reduced-order C matrix. Then, clearly (23) holds. \square

Notice that comparing to the original interpolation scheme, where the interpolation is in the average direction $\mathbb{1}_r$, the above lemma requires the interpolation direction to become the reduced-order system matrices \tilde{B} and \tilde{C} . Another lemma is also essential to prove the necessary \mathcal{H}_2 optimality conditions.

Lemma 3 For any $1 \leq k_1, k_2 \leq k \in \mathbb{Z}_+$ satisfying $k_1 + k_2 = k + 1$ and the interpolation conditions satisfying those in Lemma 2, if W solves

$$W\Lambda + A^\top W + N^\top W\tilde{N} + 2H^{(2)}(V \otimes W) \left(\tilde{H}^{(2)} \right)^\top + C^\top \tilde{C} = 0, \quad (24)$$

then

$$\sum_{k_1+k_2=k+1} (w_l^{k_1})^\top v_l^{k_2} = \sum_{l_{k-1}=1}^r \cdots \sum_{l_1=1}^r \left\langle \tilde{\phi}_{ll_{k-1}\dots l_1}^k, \frac{\partial}{\partial s_{k_2}} G_k(-\tilde{\lambda}_{l_k}, \dots, -\tilde{\lambda}_{l_1}) \right\rangle. \quad (25)$$

Proof We first consider a simple case where $\tilde{C}^\top = \tilde{B} = \mathbb{1}_r$ in (22) and (24), respectively. From Corollary 1 we know that if V and W solves (19) and (20) respectively, there holds $B^\top w_l = C v_l$. Since the system is SISO, there also holds $(w_l)^\top B = C v_l$. Let $C v_l = \sum_{k=1}^\infty C v_l^k = (w_l)^\top B = \sum_{k=1}^\infty \tilde{v}_l^k B$. The term \tilde{v}_l^k is defined later in the proof. Introducing this new term helps to simplify the proof and this new term can be considered as a replacement of w_l^k which has been used to construct W earlier in the paper. Notice that \tilde{v}_l^k is based on the generalized transfer function of the original system rather than the adjoint system, so it does not satisfy the same Sylvester equations as w_l^k . However, by definition, $\sum_{k=1}^\infty \tilde{v}_l^k B$ and $\sum_{k=1}^\infty (w_l^k)^\top B$ are equivalent. Now reconsider the case where the quadratic term in (20) is scaled by a factor of 2 (this scaling factor is related to the derivative of the quadratic term since $d(f_1(x) \otimes f_2(x))/dx = df_1(x)/dx \otimes f_2(x) + f_1(x) \otimes df_2(x)/dx$), we can define \tilde{v}_l^k as

$$\begin{aligned} \tilde{v}_l^1 &= C(\sigma_l I - A)^{-1}, \\ \tilde{v}_l^2 &= C(\sigma_l I - A)^{-1} N \sum_{l_1=1}^r u_{ll_1} (\sigma_{l_1} I - A)^{-1}, \\ \tilde{v}_l^k &= C(\sigma_l I - A)^{-1} \left(N \sum_{l_{k-1}=1}^r u_{ll_{k-1}} v_{l_{k-1}}^{k-1} B^\dagger + \right. \\ &\quad \left. H \sum_{\substack{i,j \geq 1 \\ i+j=k-1}}^r \sum_{l_{k-1}=1}^r \sum_{l_{k-i-1}=1}^r \xi_{ll_{k-1}l_{k-i-1}} (v_{l_{k-1}}^i \otimes v_{l_{k-i-1}}^j B^\dagger + v_{l_{k-1}}^i B^\dagger \otimes v_{l_{k-i-1}}^j) \right), \quad k \geq 3, \end{aligned}$$

where B^\dagger stands for the right pseudo inverse of B , which may not be unique, but any B^\dagger satisfying $B^\dagger B = I_n$ can be used here. For $k = 1$, i.e., $k_1 + k_2 = 2$, it's not difficult to calculate that

$$\tilde{v}_l^1 v_l^1 = C(\sigma_l I - A)^{-2} B = \left. \frac{\partial}{\partial s_1} \right|_{s_1=\sigma_l} G_1(s_1).$$

Similarly, for $k = 2$, i.e., $k_1 + k_2 = 3$, one can obtain that

$$\tilde{v}_l^1 v_l^2 = \left. \frac{\partial}{\partial s_1} \right|_{\substack{s_1=\sigma_l, \\ s_2=\sigma_{l_1}}} G_2(s_1, s_2), \quad \tilde{v}_l^2 v_l^1 = \left. \frac{\partial}{\partial s_2} \right|_{\substack{s_1=\sigma_l, \\ s_2=\sigma_{l_1}}} G_2(s_1, s_2).$$

Notice that in the second equation of above calculation one has to re-arrange the indices of σ in v_l^2 . For $k_1 + k_2 = 4$, i.e., $k = 3$, the pairs (k_1, k_2) can be taken as $(1, 3)$ or $(2, 2)$ or $(3, 1)$. By re-arranging the indices of σ in both $\tilde{v}_l^{k_1}$ and $v_l^{k_2}$, it can be calculated that

$$\begin{aligned} \tilde{v}_l^1 v_l^3 &= C(\sigma_l I - A)^{-2} N \sum_{l_2=1}^r u_{ll_2} (\sigma_{l_2} I - A)^{-1} N \sum_{l_1=1}^r u_{l_2 l_1} (\sigma_{l_1} I - A)^{-1} B + \\ &C(\sigma_l I - A)^{-2} H \sum_{l_2=1}^r \sum_{l_1=1}^r \xi_{ll_2 l_1} ((\sigma_{l_2} I - A)^{-1} B \otimes (\sigma_{l_1} I - A)^{-1} B), \\ \tilde{v}_l^2 v_l^2 &= C(\sigma_l I - A)^{-1} N \sum_{l_2=1}^r u_{ll_2} (\sigma_{l_2} I - A)^{-2} N \sum_{l_1=1}^r u_{l_2 l_1} (\sigma_{l_1} I - A)^{-1} B, \\ \tilde{v}_l^3 v_l^1 &= C(\sigma_l I - A)^{-1} N \sum_{l_2=1}^r u_{ll_2} (\sigma_{l_2} I - A)^{-1} N \sum_{l_1=1}^r u_{l_2 l_1} (\sigma_{l_1} I - A)^{-2} B + \\ &C(\sigma_l I - A)^{-1} H \sum_{l_2=1}^r \sum_{l_1=1}^r \xi_{ll_2 l_1} ((\sigma_{l_2} I - A)^{-2} B \otimes (\sigma_{l_1} I - A)^{-1} B) + \\ &C(\sigma_l I - A)^{-1} H \sum_{l_2=1}^r \sum_{l_1=1}^r \xi_{ll_2 l_1} ((\sigma_{l_2} I - A)^{-1} B \otimes (\sigma_{l_1} I - A)^{-2} B). \end{aligned}$$

Sum these three equations together, it can be calculated that

$$\sum_{\substack{1 \leq k_1, k_2 \leq 3, \\ k_1 + k_2 = 4}} \tilde{v}_l^{k_1} v_l^{k_2} = \sum_{k_2=1}^3 \left. \frac{\partial}{\partial s_{k_2}} \right|_{\substack{s_i = \sigma_{l_i}, i=1,2, \\ s_3 = \sigma_l}} G_3(s_1, s_2, s_3).$$

From $\tilde{v}_l^3 v_l^1$ one can see there are two terms with Kronecker product, which requires the scaling factor 2 to be present in (24). Now let's choose two arbitrary integers $1 \leq k_1, k_2 \leq K$ satisfying $k_1 + k_2 = k + 1$ for $k \leq K$ and $K \in \mathbb{N}_{\geq 3}$, then we have

$$\begin{aligned}
\sum_{\substack{1 \leq k_1, k_2 \leq k, \\ k_1 + k_2 = k+1}} \tilde{v}_l^{k_1} v_l^{k_2} &= C(\sigma_l I - A)^{-1} v_l^{k_2} + C(\sigma_l I - A)^{-1} N \sum_{l_1=1}^r u_{ll_1} (\sigma_l I - A)^{-1} v_l^{k_2} + \dots + \\
C(\sigma_l I - A)^{-1} N \sum_{l_{k_1-1}=1}^r u_{ll_{k_1-1}} v_{l_{k_1-1}}^{k_1-1} B^\dagger v_l^{k_2} &+ C(\sigma_l I - A)^{-1} H \sum_{\substack{i, j \geq 1 \\ i+j=k_1-1}} \\
\sum_{l_{k_1-1}=1}^r \sum_{l_{k_1-i-1}=1}^r \xi_{ll_{k_1-1}l_{k_1-i-1}} (v_{l_{k_1-1}}^i \otimes v_{l_{k_1-i-1}}^j B^\dagger v_l^{k_2} &+ v_{l_{k_1-1}}^i B^\dagger v_l^{k_2} \otimes v_{l_{k_1-i-1}}^j) + \dots
\end{aligned}$$

Re-arranging the indices of σ in $\tilde{v}_l^{k_1}$ and $v_l^{k_2}$, the first two terms of the above equation are equivalent to taking derivatives of the first two interpolation points in $G^k(s_1, s_2, \dots, s_k)$. Then by using the recursive expression of v_l^k in (15c), derivatives of all interpolation points are taken in the other terms. Hence, there holds

$$\sum_{\substack{1 \leq k_1, k_2 \leq k, \\ k_1 + k_2 = k+1}} \tilde{v}_l^{k_1} v_l^{k_2} = \sum_{k_2=1}^k \frac{\partial}{\partial s_{k_2}} \Bigg|_{\substack{s_i = \sigma_i, i=1, 2, \dots, k-1, \\ s_i = \sigma_i}} G^k(s_1, s_2, \dots, s_k).$$

For simplicity, the above proof assumed that the interpolation directions are $\mathbb{1}_r$ from both the input and the output side. Instead, if the interpolation directions are \tilde{C} and \tilde{B} , which are defined by the reduced-order system matrices, by following the same method as in Lemma 2, (25) can be proved. \square

The reason why the factor 2 appears in (24) but not in (20) is when matching the first moments is required, derivatives of $G_k(s_1, \dots, s_k)$ requires to take derivatives of Kronecker product terms, which simply doubles the number of Kronecker product terms in the equation. A typical example is $\tilde{v}_l^3 v_l^1$. That's why the scaling factor 2 in (24) is required.

Applying Lemma 2 together with Lemma 3, the \mathcal{H}_2 necessary optimality conditions are achieved by interpolation.

Theorem 4 *Let Σ be a stable SISO QB system with dimension n and Σ_r be its \mathcal{H}_2 optimal approximation of dimension $r \ll n$. The projection matrices V and W solve (22) and (24), respectively. Then Σ_r interpolates the zeroth and first-order moments of Σ , i.e.,*

$$\begin{aligned}
&\sum_{k=1}^{\infty} \sum_{l_k=1}^r \dots \sum_{l_1=1}^r \langle \tilde{\phi}_{l_k \dots l_1}^k, G_k(-\tilde{\lambda}_{l_k}, \dots, -\tilde{\lambda}_{l_1}) \rangle \\
&= \sum_{k=1}^{\infty} \sum_{l_k=1}^r \dots \sum_{l_1=1}^r \langle \tilde{\phi}_{l_k \dots l_1}^k, G_{rk}(-\tilde{\lambda}_{l_k}, \dots, -\tilde{\lambda}_{l_1}) \rangle,
\end{aligned} \tag{26a}$$

$$\begin{aligned}
& \sum_{k=1}^{\infty} \sum_{l_k=1}^r \cdots \sum_{l_1=1}^r \langle \tilde{\phi}_{l_k \dots l_1}^k, \sum_{j=1}^k \frac{\partial}{\partial s_j} G_k(-\tilde{\lambda}_{l_k}, \dots, -\tilde{\lambda}_{l_1}) \rangle \\
& = \sum_{k=1}^{\infty} \sum_{l_k=1}^r \cdots \sum_{l_1=1}^r \langle \tilde{\phi}_{l_k \dots l_1}^k, \sum_{j=1}^k \frac{\partial}{\partial s_j} G_{rk}(-\tilde{\lambda}_{l_k}, \dots, -\tilde{\lambda}_{l_1}) \rangle,
\end{aligned} \tag{26b}$$

where G_{rk} is the k th generalized transfer function of Σ_r which has a diagonalizable A_r matrix.

Proof The above theorem is proved by using Lemmas 2 and 3. For details, we refer to [9]. \square

3.3 Computational Aspects

As one has noticed, the interpolation scheme requires to solve QB Sylvester equations, which is not a computationally trivial task. In many cases, the existence of the solutions is even unclear. In this work, we propose to solve the scaled version of these nonlinear matrix equations [8, 12]:

$$V\Lambda + AV + \alpha^2 NV\tilde{N}^\top + \alpha^2 H(V \otimes V)\tilde{H}^\top + B\tilde{B}^\top = 0, \tag{27}$$

$$W\Lambda + A^\top W + \alpha^2 N^\top W\tilde{N} + 2\alpha^2 H^{(2)}(V \otimes W) \left(\tilde{H}^{(2)} \right)^\top + C^\top \tilde{C} = 0, \tag{28}$$

by a fixed point method. The scaling factor α^2 shrinks the nonlinear effect of the original system such that the fixed point method converges. Define two operators as:

$$\mathcal{L}(X) = AX + X\Lambda, \quad \mathcal{N}(X) = -NX\tilde{N}^\top - H(X \otimes X)\tilde{H}^\top.$$

As long as the spectral radius $\rho(\mathcal{L}^{-1}\mathcal{N}) < 1/\alpha^2$ with α^2 the scaling factor, the fixed point method is guaranteed to converge. However, in real cases, trials and errors may be needed for selecting α . If α is too close to 1, the fixed point method does not converge; if it is too close to 0, although the fixed point method converges rapidly (1 or 2 iterations), the nonlinear part of the system may not be considered enough. As a consequence, the resulted reduced-order model can not capture the nonlinear behavior of the original system well.

Certainly, solving QB Sylvester equations by the fixed point method can be more expensive than only solving the Sylvester equations corresponding to the first two and the quadratic part of the generalized transfer functions. However, as long as the fixed point method converges in a few iterations, the computational cost is still acceptable.

4 Numerical Examples

The proposed methods are demonstrated on the 1D Fisher-KPP equation, which has an exact quadratic nonlinearity. For more case studies, such as the FitzHugh-Nagumo system, we refer to [9]. The system under consideration is described by partial differential equations, and it is semi-discretized by using the central difference method.

In the tests, we compare the rational Krylov method in [10], the \mathcal{H}_2 optimal approximation method proposed in this paper, which is referred to as α QB-IRKA and the truncated-QB-IRKA algorithms in [8]. The truncated-QB-IRKA method approximates the system by interpolating the first three generalized transfer functions G_1 , G_2 and G_3 , so it is slightly different from the original method in [8]. And no scaling factor is required for this method. To make the legend compact in figures, we use the following abbreviations: Full-order model simulation is denoted as QB FOM; Reduced-order model simulations are denoted as

- α QBIRKA if it is obtained by the method proposed in this work;
- truncQBIRKA if it is obtained by the method proposed in [8];
- LK if it is obtained by the rational Krylov method proposed in [10].

4.1 Fisher-KPP Equation

The Fisher-KPP equation is a reaction-diffusion system widely used in population dynamics [14], plasma physics and combustion models. It is a spatially 1D equation with its boundary conditions and initial condition described by

$$\begin{aligned} w_t(s, t) &= \alpha w_{ss}(s, t) + \beta w(s, t)(1 - w(s, t)) + u(s, t), \quad w(s, 0) = 0, \\ w(0, t) &= 0 = w(L, t). \end{aligned}$$

In the equation, the parameter values are taken as $L = 1$, $\alpha = 1$ and $\beta = 3$. The input is assumed to be put on the first node and the state of the last node scaled by 10^6 is the quantity of interest. By applying the central difference method on the spatial domain, the system can be represented by an ODE model with $n = 100$ states as

$$\dot{x}(t) = Ax(t) + H(x(t) \otimes x(t)) + Bu(t), \quad y(t) = Cx(t).$$

The detailed system matrices can be found in [9]. Here we choose the norm of C to be 10^6 to scale the output such that it is comparable with the input. In such a setting, the largest real part of the eigenvalues of A is $-6.8688 \in \mathbb{C}^-$. Thus the system is locally stable near the origin. It can be shown that the fixed point method of solving Sylvester equations is guaranteed to converge [9]. The \mathcal{H}_2 norm of the original system is then $\|\Sigma\|_{\mathcal{H}_2} = 2.8462$. To test the performance of the model order reduction, we reduced

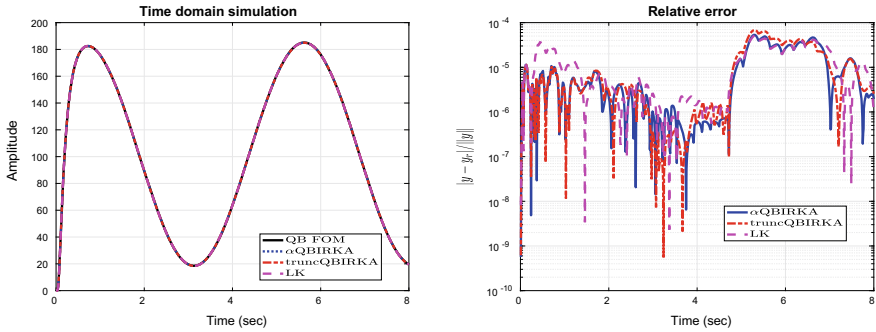


Fig. 1 Time domain simulations and the relative errors of the proposed model order reduction methods for the Fisher-KPP equation

Table 1 Relative \mathcal{H}_2 errors of three different model order reduction methods for the Fisher-KPP equation

| Methods | α QBIRKA | truncQBIRKA | LK |
|--------------------------------|-------------------------|-------------------------|-------------------------|
| Relative \mathcal{H}_2 error | 8.2436×10^{-7} | 1.5830×10^{-6} | 5.4389×10^{-4} |

the system from order $n = 100$ to order $r = 12$. The full-order and reduced-order systems are simulated with a sinusoidal input

$$u(t) = 50 \sin(0.4\pi t + \pi/3) + 60,$$

over a given time interval $t \in [0, 8]$. The time domain responses and the relative approximation errors are depicted in Fig. 1. To avoid division by zero, the relative error (or the absolute error normalized by the \mathcal{L}_2 norm of the original output) is defined as

$$e(t) = |y(t) - y_r(t)| / \|y\|, \quad \|y\| := \|y\|_{\mathcal{L}_2}.$$

When applying the rational Krylov method, we generate $r = 12$ interpolation points $\lambda_1, \dots, \lambda_{12}$ from -10^3 to -1 in the logarithmic scale. Hence, we only interpolate the values of the first two generalized transfer functions in [10] rather than their derivatives. For the α QB-IRKA and the truncated-QB-IRKA methods, they both converge in 50 iterations. It can be seen from Fig. 1 that the performance of the three methods are quite close to each other. In the \mathcal{H}_2 norm sense, the relative \mathcal{H}_2 error defined as

$$\text{error} = \|\Sigma - \Sigma_r\|_{\mathcal{H}_2} / \|\Sigma\|_{\mathcal{H}_2},$$

is calculated in Table 1. Clearly, our method gives the smallest relative \mathcal{H}_2 error, which is expected, since it is the only one among the three demonstrated methods which satisfies the necessary \mathcal{H}_2 optimality condition.

5 Conclusions

In this paper we first discussed the class of quadratic bilinear dynamical systems together with a definition of their \mathcal{H}_2 norm. We characterized the \mathcal{H}_2 norm by means of a pole-residual representation of the generalized QB transfer functions and characterized the \mathcal{H}_2 norm of the approximation error between two QB systems in terms of pole-residual functions. On the basis of this characterization, we proposed an \mathcal{H}_2 optimal model order reduction scheme that employs an interpolation framework on complex-valued functions. The interpolation problem can be tackled by solving two generalized Sylvester equations. In such an interpolation framework, we showed that by selecting the interpolation variables carefully, an \mathcal{H}_2 optimal approximation can be achieved using a classical Petrov-Galerkin projection on the state space matrices of the QB model. We pointed out that the computation of the Sylvester equations involved in the reduction process may not be feasible. A numerical example is given to demonstrate the quality of the proposed method. Results of the numerical examples show accurate approximations. For higher-order polynomial systems, [7] provides a general framework for interpolation-based model order reduction.

References

1. Ahmad, M.I., Benner, P., Jaimoukha, I.M.: Krylov subspace methods for model reduction of quadratic-bilinear systems. *IET Control Theory Appl.* **10**(16), 2010–2018 (2016)
2. Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems, vol. 6. SIAM (2005)
3. Antoulas, A.C., Gosea, I.V., Ionita, A.C.: Model reduction of bilinear systems in the Loewner framework. *SIAM J. Sci. Comput.* **38**(5), B889–B916 (2016)
4. Benner, P., Breiten, T.: Two-sided projection methods for nonlinear model order reduction. *SIAM J. Sci. Comput.* **37**(2), B239–B260 (2015)
5. Benner, P., Damm, T.: Lyapunov equations, energy functionals, and model order reduction of bilinear and stochastic systems. *SIAM J. Control Optim.* **49**(2), 686–711 (2011)
6. Benner, P., Goyal, P.: Balanced truncation model order reduction for quadratic-bilinear control systems (2017). [arXiv:1705.00160](https://arxiv.org/abs/1705.00160)
7. Benner, P., Goyal, P.: Interpolation-based model order reduction for polynomial parametric systems (2019). [arXiv:1904.11891](https://arxiv.org/abs/1904.11891)
8. Benner, P., Goyal, P., Gugercin, S.: \mathcal{H}_2 -quasi-optimal model order reduction for quadratic-bilinear control systems. *SIAM J. Matrix Anal. Appl.* **39**(2), 983–1032 (2018)
9. Cao, X.: Optimal model order reduction for parametric nonlinear systems. Ph.D. thesis, Eindhoven University of Technology (2019)
10. Cao, X., Maubach, J., Weiland, S., Schilders, W.H.A.: A novel Krylov method for model order reduction of quadratic bilinear systems. In: Proceedings of the 2018 57th IEEE Conference on Decision and Control (CDC), pp. 3217–3222. IEEE (2018)
11. Chaturantabut, S., Sorensen, D.C.: Nonlinear model reduction via discrete empirical interpolation. *SIAM J. Sci. Comput.* **32**(5), 2737–2764 (2010)
12. Condon, M., Ivanov, R.: Nonlinear systems-algebraic Gramians and model reduction. *COMPEL Int. J. Comput. Math. Electr. Electron. Eng.* **24**(1), 202–219 (2005)
13. Ding, B., Ping, X., Pan, H.: On dynamic output feedback robust MPC for constrained quasi-LPV systems. *Int. J. Control* **86**(12), 2215–2227 (2013)
14. Fisher, R.A.: The wave of advance of advantageous genes. *Ann. Eugen.* **7**(4), 355–369 (1937)

15. Flagg, G., Gugercin, S.: Multipoint Volterra series interpolation and \mathcal{H}_2 optimal model reduction of bilinear systems. *SIAM J. Matrix Anal. Appl.* **36**(2), 549–579 (2015)
16. Flagg, G.M.: Interpolation methods for the model reduction of bilinear systems. Ph.D. thesis, Virginia Tech (2012)
17. Freund, R.W.: Model reduction methods based on Krylov subspaces. *Acta Numer.* **12**, 267–319 (2003)
18. Fujimoto, K., Scherpen, J.M.A.: Balanced realization and model order reduction for nonlinear systems based on singular value analysis. *SIAM J. Control Optim.* **48**(7), 4591–4623 (2010)
19. Fujimoto, K., Scherpen, J.M.A., Gray, W.S.: Hamiltonian realizations of nonlinear adjoint operators. *Automatica* **38**(10), 1769–1775 (2002)
20. Gallivan, K., Vandendorpe, A., van Dooren, P.: Model reduction of MIMO systems via tangential interpolation. *SIAM J. Matrix Anal. Appl.* **26**(2), 328–349 (2004)
21. Gosea, I.V., Antoulas, A.C.: Data-driven model order reduction of quadratic-bilinear systems. *Numer. Linear Algebra Appl.* **25**(6), e2200 (2018)
22. Grimme, E.J.: Krylov projection methods for model reduction. Ph.D. thesis, University of Illinois at Urbana-Champaign (1997)
23. Gu, C.: QLMOR: a projection-based nonlinear model order reduction approach using quadratic-linear representation of nonlinear systems. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **30**(9), 1307–1320 (2011)
24. Gugercin, S., Antoulas, A.C., Beattie, C.: \mathcal{H}_2 model reduction for large-scale linear dynamical systems. *SIAM J. Matrix Anal. Appl.* **30**(2), 609–638 (2008)
25. Gugercin, S., Polyuga, R.V., Beattie, C., van der Schaft, A.: Structure-preserving tangential interpolation for model reduction of port-Hamiltonian systems. *Automatica* **48**(9), 1963–1974 (2012)
26. Meier, L., Luenberger, D.: Approximation of linear constant systems. *IEEE Trans. Autom. Control* **12**(5), 585–588 (1967)
27. Rotondo, D., Nejjari, F., Puig, V.: Quasi-LPV modeling, identification and control of a twin rotor MIMO system. *Control Eng. Pract.* **21**(6), 829–846 (2013)
28. Rugh, W.J.: *Nonlinear System Theory*. Johns Hopkins University Press Baltimore (1981)
29. Scherpen, J.M.A.: Balancing for nonlinear systems. *Syst. Control Lett.* **21**(2), 143–153 (1993)

An Adaptive Sampling Approach for the Reduced Basis Method



Sridhar Chellappa, Lihong Feng, and Peter Benner

Abstract The offline time of the reduced basis method can be very long given a large training set of parameter samples. This usually happens when the system has more than two independent parameters. On the other hand, if the training set includes fewer parameter samples, the greedy algorithm might produce a reduced-order model with large errors at the samples outside of the training set. We introduce a method based on a surrogate error model to efficiently sample the parameter domain such that the training set is adaptively updated starting from a coarse set with a small number of parameter samples. A sharp a posteriori error estimator is evaluated on a coarse training set. Radial Basis Functions are used to interpolate the error estimator over a separate fine training set. Points from the fine training set are added into the coarse training set at every iteration based on a user defined criterion. In parallel, parameter samples satisfying a defined tolerance are adaptively removed from the coarse training set. The approach is shown to avoid high computational costs by using a small training set and to provide a reduced-order model with guaranteed accuracy over a fine training set. Further, we show numerical evidence that the reduced-order model meets the defined tolerance over an independently sampled test set from the parameter domain.

Keywords Reduced basis method · Training set sampling · Adaptivity · Error estimation · Radial basis interpolation

S. Chellappa—Supported by the International Max Planck Research School for Advanced Methods in Process and Systems Engineering (IMPRS-ProEng)

S. Chellappa (✉) · L. Feng · P. Benner
Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany
e-mail: chellappa@mpi-magdeburg.mpg.de

L. Feng
e-mail: feng@mpi-magdeburg.mpg.de

P. Benner
e-mail: benner@mpi-magdeburg.mpg.de

P. Benner
Fakultät für Mathematik, Otto-von-Guericke Universität Magdeburg, Magdeburg, Germany

© Springer Nature Switzerland AG 2022

C. Beattie et al. (eds.), *Realization and Model Reduction of Dynamical Systems*,
https://doi.org/10.1007/978-3-030-95157-3_8

1 Introduction

The need of modeling physical processes accurately very often leads to mathematical models of large-scale. Simulating such large-scale systems poses challenges on computer memory and computational power. Model Order Reduction (MOR) aims to speedup simulation of the large-scale full order model (FOM) by constructing a small-scale system, called the reduced-order model (ROM), which serves as a good approximation of the FOM.

For FOMs arising from discretization of parametric partial differential equations (PDEs), the Reduced Basis (RB) method is a popular choice [17]. It is suitable especially when the FOM needs to be solved repeatedly for a range of parameters. The RB method involves two stages. The first is an expensive *offline stage* in which a subspace of the solution manifold is constructed by simulations of the FOM at certain chosen parameters. Then, using Galerkin projection, the solution to the PDE is projected onto this subspace to generate the ROM. In the second *online stage*, instead of solving the FOM, one solves the ROM for any parameter of interest. The RB method works well when the solution manifold is supposed to have a small Kolmogorov n -width. In this work, we focus on the offline stage of the RB method.

The offline stage in the RB method requires the construction of a representative set of parameters from the domain of interest. This is called the *training set*. The subspace approximating the solution manifold is constructed by solving the FOM at a set of selected points from this training set, picked using a greedy algorithm [11]. The choice of a training set is nontrivial. On the one hand, if it includes too few parameters, the original solution manifold may not be adequately represented, leading to a poor ROM with large error. On the other hand, if it is too fine, the offline time can be prohibitively long. When the PDE involves several parameters, properly defining the training set can be a severe computational issue.

Several authors have attempted to address this issue. The earliest work to consider an adaptive sampling of the training set was [19], where the author proposes a multi-stage method. The greedy algorithm is run several times over randomly sampled small training sets to generate the RB basis. Then, the ROM is tested over a much larger training set and the greedy algorithm is re-run only on those points failing the tolerance criterion. The authors of [10] address the issue of large training sets by two approaches. The first one is a procedure to monitor the error over an additional validation parameter set. If a large error is detected, then the training set is further refined, either uniformly or locally. The second approach, similar to the one presented in [6], is based on partitioning the parameter domain adaptively, and generating a local basis for each partition. Other approaches for adaptive sampling are proposed in [12, 14]. More recently, an approach based on Kriging interpolation and clustering is proposed in [16], to tackle the problem of high-dimensional parameter spaces. An interpolant of the residual norm is calculated over a fine grid of parameters. Then, k-means clustering is used to identify parameters that have high probability of presenting larger errors.

The approach we propose here is based on interpolating a sharp a posteriori output error estimator originally proposed in [5]. The RB method is initialized with a coarse training set. We update this set progressively by adding or removing points from it. At each iteration of the greedy algorithm, we estimate the error of the ROM at every parameter in the coarse training set. Note that the evaluation of the estimator does not need to evaluate the FOM. In order to further reduce the computational costs, we then interpolate the estimated error over a fine training set and use the interpolant as the error surrogate, which will replace the error estimator for estimating the ROM error over the fine training set. To achieve this, a surrogate model based on Radial Basis Functions (RBFs) is used.

At each iteration of the greedy algorithm, the ROM errors at the parameters in the fine training set are checked by the error surrogate. Those parameters corresponding to large values of the error surrogate are selected and added into the coarse training set. The error surrogate is much cheaper to compute than the error estimator. Therefore, using the former instead of the latter to check the ROM error over the fine training set, will reduce the computational cost. Additionally, if any parameter in the coarse training set achieves the required ROM accuracy, we remove it from the coarse training set. Such an approach is able to construct a small, representative training set by fully exploring the parameter domain with reduced computational cost. Forming the RBF interpolant is a relatively cheap operation with the most expensive contribution coming from the need to solve an $\ell \times \ell$ linear system of equations at each parameter sample, where ℓ is the cardinality of the coarse training set, which is expected to be small. The algorithm proposed in our work is based on the use of the RBF interpolant and offers certain advantages over the most relevant methods in [12, 16]. Our proposed approach differs in the following aspects, when compared to [12]:

- We use a primal-dual a posteriori error estimator on a coarse training set and a cheaply computable error surrogate on the fine training set. No cheap error estimator is used on the fine training set in [12], potentially leading to a larger computational effort.
- At each iteration of the greedy algorithm, a saturation assumption is introduced in [12] in order to avoid calculation of the ROM at certain parameters. However, this requires estimation of a saturation constant which needs to be defined a priori by the user. In our method, no such constant needs to be estimated.

Furthermore, when compared to [16],

- We consider both addition and removal of parameter samples from the coarse training set. The method in [16] is restricted to adding new samples only.
- The use of Kriging interpolation involves the estimation of several hyper-parameters. This can be a computational bottleneck [20]. In the case of RBF, there is only one free parameter at most. If using some special kernel functions, e.g. polyharmonic spline kernels, there are no free parameters to tune.
- Finally, a nonlinear model was considered in [16] in order to demonstrate the adaptive sampling approach. An offline hyper-reduction step was used to reduce

the complexity of the online nonlinearity evaluations, where a second training set for the nonlinear function is used in the hyper-reduction step, and is likely to be separately given and fixed. In our approach, we employ the Adaptive POD-Greedy-(D)EIM algorithm proposed in [5]. This avoids the need for a potentially separate training set for the hyper-reduction phase. Instead, the (D)EIM basis generation is carried out as a part of the greedy loop. In this way, we propose a fully adaptive approach in Algorithm 4.

The rest of the paper is organized as follows. In Sect. 2, the basic idea behind the RB method is reviewed. An adaptive parameter sampling approach using the RBF interpolation is proposed and elaborated in Sect. 3. There, we begin with a brief introduction to the theory of RBFs. Then an adaptive sampling approach for the standard POD-Greedy algorithm is proposed to show a general framework for the proposed technique. In Sect. 3.3, a fully adaptive algorithm is proposed, which integrates the adaptive sampling technique with an adaptive basis construction method: the adaptive POD-Greedy-(D)EIM algorithm for general parametric nonlinear/nonaffine systems. A strategy of computing the shape parameter for RBF interpolation is detailed in Sect. 3.4. In Sect. 4, we show numerical results for a series of benchmark examples, validating the proposed algorithms. We conclude in Sect. 5 by summarizing our results and suggesting future research directions.

2 Reduced Basis Method

For the sake of completeness, we briefly highlight the key idea of the RB method. Consider a nonlinear parametric dynamical system arising from space-time discretization of a model of PDEs,

$$\begin{aligned} \mathbf{E}(\mu)x^{k+1}(\mu) &= \mathbf{A}(\mu)x^k(\mu) + f(x^k(\mu); \mu) + \mathbf{B}(\mu)u^k, \\ y^{k+1}(\mu) &= \mathbf{C}(\mu)x^{k+1}(\mu) \end{aligned} \quad (1)$$

where $x^k(\mu) \in \mathbb{R}^n$ is the state vector at the k -th time step t^k in the time interval $[0, T]$ divided as $t^0 < t^1 < \dots < t^k < \dots < t^K = T$. $\mathbf{E}(\mu)$, $\mathbf{A}(\mu) \in \mathbb{R}^{n \times n}$ are the parameter dependent system matrices. $f(x^k(\mu); \mu) \in \mathbb{R}^n$ depicts the state, parameter dependent system nonlinearity. $\mathbf{B}(\mu) \in \mathbb{R}^{n \times q}$, $\mathbf{C}(\mu) \in \mathbb{R}^{m \times n}$ are the input and output matrices, respectively, while u^k denotes the input at the k -th time step and $y^{k+1}(\mu)$ denotes the quantities-of-interest. The vector of parameters $\mu \in \mathbb{R}^d$ belongs to a parameter domain $\mathcal{P} \subset \mathbb{R}^d$. The space discretization can be done through any method of choice such as Finite Element Method (FEM), Finite Volume Method (FVM) or the Finite Difference Method (FDM). For the time discretization, we adopt a semi-implicit scheme for ease of using the a posteriori error estimator in [5]. The RB method attempts to generate a subspace \mathcal{X} with dimension $r \ll n$, that best approximates the solution manifold. Let $\mathbf{V} \subset \mathbb{R}^{n \times r}$ be an orthogonal basis of \mathcal{X} . The ROM is then given via Galerkin projection,

$$\begin{aligned}\mathbf{E}_r(\mu)x_r^{k+1}(\mu) &= \mathbf{A}_r(\mu)x_r^k(\mu) + f_r(x_r^k(\mu); \mu) + \mathbf{B}_r(\mu)u^k, \\ y_r^{k+1}(\mu) &= \mathbf{C}_r(\mu)x_r^{k+1}(\mu),\end{aligned}\quad (2)$$

where $x_r^k \in \mathbb{R}^r$ is the reduced state vector at the k -th time step. $\mathbf{E}_r(\mu) := \mathbf{V}^T \mathbf{E}(\mu) \mathbf{V} \in \mathbb{R}^{r \times r}$, $\mathbf{A}_r(\mu) := \mathbf{V}^T \mathbf{A}(\mu) \mathbf{V} \in \mathbb{R}^{r \times r}$ are the reduced system matrices. $\mathbf{B}_r(\mu) := \mathbf{V}^T \mathbf{B}(\mu) \in \mathbb{R}^{r \times q}$, $\mathbf{C}_r(\mu) := \mathbf{C}(\mu) \mathbf{V} \in \mathbb{R}^{m \times r}$ are the reduced input and output matrices, respectively. The reduced nonlinear term is given by $f_r(\mathbf{V}x_r^k(\mu); \mu) := \mathbf{V}^T f(\mathbf{V}x_r^k(\mu); \mu) \in \mathbb{R}^r$. An important assumption is that the system matrices and the input, output matrices have affine decomposition, i.e., any of them can be formulated as a sum of products of parameter dependent and parameter independent components:

$$\begin{aligned}\mathbf{E}(\mu) &= \sum_{i=1}^{Q_E} \theta_i^E(\mu) \mathbf{E}_i, \\ \mathbf{B}(\mu) &= \sum_{i=1}^{Q_B} \theta_i^B(\mu) \mathbf{B}_i, \quad \mathbf{A}(\mu) = \sum_{i=1}^{Q_A} \theta_i^A(\mu) \mathbf{A}_i, \\ \mathbf{C}(\mu) &= \sum_{i=1}^{Q_C} \theta_i^C(\mu) \mathbf{C}_i.\end{aligned}$$

In the offline stage, the expensive parameter independent terms are evaluated once and stored for repeated use in the online stage.

The standard process of the RB method for time-dependent parametric systems is presented in Algorithm 1 which is known as the POD-Greedy algorithm. It constructs the projection matrix \mathbf{V} by selecting a set of samples $\{\mu_i\}_{i=1}^r$ from a given parameter training set, through a greedy algorithm. The FOM solutions $x^k(\mu)$, $k = 1, \dots, K$, at a newly selected sample μ_i are assembled into a matrix $\mathbf{X} := [x^1(\mu_i), \dots, x^K(\mu_i)]$. The matrix \mathbf{V} is iteratively enriched by projecting the solution vectors in \mathbf{X} onto the current subspace $\text{range}(\mathbf{V})$, see Steps 3–5 in Algorithm 1. Crucial to the success of the RB method is the a posteriori error estimator,

$$\Delta^{k+1}(\mu) \geq \|y^{k+1}(\mu) - y_r^{k+1}(\mu)\|.$$

It quantifies the error between the FOM and ROM for each parameter, without having to solve the FOM. The parameters at which the FOM is to be solved to enrich \mathbf{V} , is picked iteratively by maximizing the error estimator over the training set. Although Eq. (2) is of reduced order, the complexity of evaluating the nonlinear term $f_r(\mathbf{V}x_r^k(\mu); \mu)$ remains of order $\mathcal{O}(n)$. This is the so-called *lifting bottleneck*. Techniques like Empirical Interpolation Method (EIM) or its discrete variation, the Discrete Empirical Interpolation Method (DEIM), can be used to obtain a ROM whose computational complexity is independent of the full order n . For more details on this approach, we refer the reader to [2, 4]. In the following section we discuss the idea of adaptive parameter sampling.

3 Adaptive Parameter Sampling

The quality of the ROM depends on how well the training set represents the parameter domain. In the standard approach, if a finely sampled training set is used for a parameter domain with high-dimension, it can be computationally expensive. If otherwise, a coarse training set is used, it will not guarantee that a reliable ROM can be derived. Consequently, one faces the situation of trial and error. The approach we propose here is based on a surrogate error model generated through RBF interpolation. It aims to overcome the drawbacks of the standard approach using a fixed training set.

3.1 Radial Basis Functions

RBF interpolation is an efficient way to approximate scattered data in high-dimensional space. Consider the domain $\Omega \subseteq \mathbb{R}^d$ and the function $h := \mathbb{R}^d \rightarrow \mathbb{R}$. Assume that h is difficult to evaluate, or known only at a few points in Ω . Let Φ be the *kernel function* having the property that $\Phi(\mu_1, \mu_2) = \Phi(\|\mu_1 - \mu_2\|)$, $\forall \mu_1, \mu_2 \in \Omega$. Typically, the Euclidean norm is used. Such kernels have the name of *radial basis functions* since they are radially symmetric. We seek an approximation $s : \mathbb{R}^d \rightarrow \mathbb{R}$ to the function h given as

$$s(\mu) := \sum_{i=1}^{\ell} c_i \Phi(\|\mu - \mu_i\|), \quad \forall \mu \in \Omega. \quad (3)$$

The coefficients $\{c_i\}_{i=1}^{\ell}$ are determined by enforcing the interpolation condition $h(\mu_i) = s(\mu_i)$, $i = 1, 2, \dots, \ell$. This amounts to solving the linear system of equations

$$\underbrace{\begin{bmatrix} \Phi(\mu_1, \mu_1) & \Phi(\mu_1, \mu_2) & \cdots & \Phi(\mu_1, \mu_\ell) \\ \Phi(\mu_2, \mu_1) & \Phi(\mu_2, \mu_2) & \cdots & \Phi(\mu_2, \mu_\ell) \\ \vdots & \vdots & \ddots & \vdots \\ \Phi(\mu_\ell, \mu_1) & \Phi(\mu_\ell, \mu_2) & \cdots & \Phi(\mu_\ell, \mu_\ell) \end{bmatrix}}_{\mathbf{R}} \underbrace{\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_\ell \end{bmatrix}}_c = \underbrace{\begin{bmatrix} h(\mu_1) \\ h(\mu_2) \\ \vdots \\ h(\mu_\ell) \end{bmatrix}}_d \quad (4)$$

The points $\{\mu_i\}_{i=1}^{\ell} \in \Omega$ are called the *centers* of the radial basis functions. Under the assumption that the centers are pairwise distinct, it can be proven that the RBF kernel matrix \mathbf{R} is positive definite for some suitable choice of radial basis functions Φ and thus Eq. (4) has a unique solution. We refer to [21, Chap. 6] for the proof and detailed treatment. Table 1 provides a list of commonly used radial basis functions. All of the radial basis functions in Table 1 have global support. There exist also radial basis functions with local support. For more details we refer the reader to [3].

In practice, the class of radial basis functions that have the positive definite property (\mathbf{R} being positive definite) is limited to a few, such as Gaussian and Inverse

Table 1 Common radial basis kernels

| Kernel | $\Phi(r) := \Phi(\ \mu - \mu_i\)$ |
|----------------------|------------------------------------|
| Gaussian | $e^{-\sigma^2 r^2}$ |
| Thin-plate splines | $r^2 \ln(r)$ |
| Multiquadric | $\sqrt{r^2 + \sigma^2}$ |
| Inverse multiquadric | $\frac{1}{\sqrt{r^2 + \sigma^2}}$ |

multiquadric. The class of admissible basis functions ($\Phi(\cdot)$) can be expanded by defining the so called *conditionally positive definite* functions, by which the positive definiteness can be satisfied by imposing some additional constraints given as,

$$\sum_{j=1}^M c_j p_j(\mu) = 0, \quad i = 1, 2, \dots, \ell.$$

The functions p_1, p_2, \dots, p_M are a basis of the polynomial space with suitable degree. In practice, we choose M to be equal to the number of scalar parameters d . With the new conditions imposed, the radial basis interpolant now becomes,

$$s(\mu) := \sum_{i=1}^{\ell} c_i \Phi(\|\mu - \mu_i\|) + \sum_{j=1}^M \lambda_j p_j(\mu). \quad (5)$$

We then obtain a saddle-point system of dimension $N_{\text{RBF}} := (M + \ell) \times (M + \ell)$:

$$\begin{bmatrix} \mathbf{R} & \mathbf{P} \\ \mathbf{P}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} c \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathfrak{d} \\ \mathbf{0} \end{bmatrix} \quad (6)$$

With a proper choice of p_1, p_2, \dots, p_M , the augmented coefficient matrix is positive definite for all choices of radial basis functions in Table 1 and this ensures the uniqueness of the interpolant. For a detailed discussion of the rationale behind the idea of conditionally positive definite functions, we refer to [21, Chap. 6].

3.2 *POD-Greedy Algorithm with Adaptive Sampling*

Now we detail the proposed approach of adaptively constructing the training set for the RB method. The proposed adaptive parameter sampling scheme is general and can be combined with the standard greedy algorithm for steady parametric systems, or with the POD-greedy algorithm (Algorithm 1) for time-dependent systems, to adaptively update the training set and reduce the offline costs. Since the frame-

work is similar, we only present the adaptive parameter sampling approach with POD-Greedy for time-dependent problems. The standard POD-Greedy algorithm is presented in Algorithm 1. POD-greedy with adaptive sampling is presented as Algorithm 2, where it can be seen that the training set (Ξ_c) is updated at each iteration.

Algorithm 2 is initialized using a coarse training set Ξ_c , with cardinality N_c . At the end of each iteration, we compute the error estimator *only* over the *coarse* training set Ξ_c , i.e., $\Delta(\mu)$, $\forall \mu \in \Xi_c$ (compare Step 6 in Algorithm 1 with Step 4 in Algorithm 2). We use this as the *input data* to form the radial basis interpolant over the fine training set, i.e., $s(\mu)$, $\forall \mu \in \Xi_f$, where Ξ_f is a finely sampled training set of cardinality $N_f \gg N_c$. The evaluation costs are $\mathcal{O}((N_c + M)^3)$ for identifying the interpolant coefficients, and $\mathcal{O}(N_f N_c)$ for interpolating over the fine training set. Here M is the number of polynomial functions added to make the augmented coefficient matrix in Eq. (6) conditionally positive definite. Since N_c, M are both small numbers, the cost of evaluating $s(\mu)$ is much smaller than the cost of computing $\Delta(\mu)$ which includes solving the ROM for each μ . We enrich the coarse training set by adding new parameters from Ξ_f to Ξ_c . We add those n_{add} new parameters that have the largest magnitude of $s(\mu)$, where n_{add} can be fixed and user-defined.

One heuristic way of varying n_{add} adaptively is $n_{\text{add}} = \log_{10} \left(\left\lfloor \frac{\max_{\mu \in \Xi_c} \Delta(\mu)}{\text{tolerance}} \right\rfloor \right)$.

Additionally, we also monitor the coarse training set to identify those points $\check{\mu} \in \Xi_c$ with $\Delta(\check{\mu}) < \text{tol}$. Those points are then removed from Ξ_c , meaning that their corresponding full solutions need not be computed to enrich the RB basis. In this way, the coarse training set gets updated iteratively and its size remains as small as possible, avoiding many unnecessary full simulations of the FOM.

In this work, we make use of the estimator proposed in [5]. It is given by,

$$\Delta^{k+1}(\mu) := \bar{\Psi} \|r_{\text{pr}}^{k+1}\|,$$

where r_{pr}^{k+1} is the residual obtained by substituting the approximate solution $\mathbf{V}\mathbf{x}_r^{k+1}$ in Eq. (1), at each time instance t^{k+1} . The term $\bar{\Psi}$ involves the norm of the approximate solution to the dual system and its residual. Additionally, it involves two terms that need to be estimated. The first is the *inf-sup* constant, given by the smallest singular value of the matrix $\mathbf{E}(\mu)$ (when considering the Euclidean norm). The second term to be estimated involves the residual of an auxiliary system. For a more detailed discussion, we refer the reader to [5]. In the implementation of Algorithms 3 and 4, we have made use of the method proposed in [15] to estimate the inf-sup constant efficiently.

Remark 1 (*Interpolating small values*) The magnitudes of the errors we interpolate are often very small, especially at the latter iterations of the greedy algorithm. To ensure good, stable interpolation, we consider the input data to the RBF as the base 10 logarithm of the estimated errors. After interpolation, we perform the anti-logarithm and project the logarithm value back to the actual error value.

Algorithm 1 Standard POD-Greedy**Input:** Training set Ξ , Tolerance (tol).**Output:** \mathbf{V}

- 1: Initialize. $\mathbf{V} = []$, $\mu^* \in \Xi$.
- 2: **while** $\Delta(\mu^*) > \text{tol}$ **do**
- 3: Compute full order solution at μ^* . $\mathbf{X} = [x(t^1, \mu^*), x(t^2, \mu^*), \dots, x(t^K, \mu^*)]$.
- 4: Form $\tilde{\mathbf{X}} := \mathbf{X} - \mathbf{V}\mathbf{V}^T\mathbf{X}$; Obtain $\tilde{\mathbf{X}} \xrightarrow{\text{svd}} \mathbf{U}\Sigma\mathbf{W}^T$.
- 5: Update $\mathbf{V} := \text{orth}[\mathbf{V}, \mathbf{U}(:, 1)]$.
- 6: Solve the ROM and compute the error estimator $\Delta(\mu)$, $\forall \mu \in \Xi$.
- 7: Next μ^* is chosen as, $\mu^* := \arg \max_{\mu \in \Xi} \Delta(\mu)$.
- 8: **end while**

Algorithm 2 POD-Greedy with adaptive parameter sampling**Input:** Coarse training set Ξ_c , Fine training set Ξ_f , Tolerance (tol).**Output:** \mathbf{V} .

- 1: Initialize. $\mathbf{V} = []$, $\mu^* \in \Xi_c$.
- 2: **while** $\Delta(\mu^*) > \text{tol}$ **do**
- 3: Perform Steps 3-5 in Algorithm 1.
- 4: Solve the ROM and compute the error estimator $\Delta(\mu)$, $\forall \mu \in \Xi_c$.
- 5: Form RBF interpolant $s(\mu)$ of $\Delta(\mu)$ over Ξ_f .
- 6: Calculate n_{add} , pick $\{\mu_1, \dots, \mu_{\text{add}}\}$ from Ξ_f with largest errors measured by $s(\mu)$.
- 7: Update the training set. $\Xi_c := [\Xi_c \cup \{\mu_1, \dots, \mu_{\text{add}}\}]$ (Add samples to the current training set Ξ_c). Find all $\hat{\mu}_i$, $\forall i = 1, \dots, n_{\text{del}}$ with $\Delta(\hat{\mu}_i) < \text{tol}$. Set $\Xi_c := \Xi_c \setminus \{\hat{\mu}_i, \forall i = 1, \dots, n_{\text{del}}\}$ (Remove unnecessary samples from the current training set Ξ_c).
- 8: Next μ^* is chosen as, $\mu^* := \arg \max_{\mu \in \Xi_c} \Delta(\mu)$.
- 9: **end while**

3.3 A Fully Adaptive POD-Greedy-(D)EIM Algorithm

An adaptive version of Algorithm 1, called the Adaptive POD-Greedy-(D)EIM algorithm, is proposed in [5] for nonlinear systems, which aims to update the RB, (D)EIM basis with an adaptively chosen number of basis vectors at each iteration of the greedy algorithm. In case of a nonlinear or nonaffine system, the standard POD-Greedy algorithm usually precomputes the (D)EIM basis and the interpolation index matrix, \mathbf{U}_f and \mathbf{P}_f . Unlike the standard POD-Greedy algorithm (Algorithm 1), the Adaptive POD-Greedy algorithm (Algorithm 3) updates the (D)EIM matrix inside the greedy algorithm. This avoids many additional FOM simulations caused by the separate (D)EIM computation outside the POD-Greedy loop. To ensure a reliable ROM, an efficient a posteriori error estimator ($\Delta(\mu)$) is used.

The algorithm updates the RB vectors and (D)EIM basis vectors at every iteration (Steps 5–8 in Algorithm 3). The number of RB basis vectors, δ_{RB} , and that of (D)EIM basis vectors, δ_{DEIM} , to be added to/removed from the current basis are computed adaptively, see Step 8 in Algorithm 3. For details of computing δ_{RB} and δ_{DEIM} , we refer the reader to [5]. We note that for this algorithm, the training set is fixed.

Algorithm 3 Adaptive POD-Greedy-(D)EIM algorithm

Input: Training set Ξ , Tolerance (tol).

Output: \mathbf{V} , DEIM matrices $\mathbf{U}_f, \mathbf{P}_f$.

- 1: Initialize. $\mathbf{V} = [], \mu^* \in \Xi, \delta_{\text{RB}} = 1, \ell_{\text{DEIM}} = 1, \delta_{\text{DEIM}} = 0, \mathbf{U}_f = [], \mathbf{P}_f = [], \mathbf{F} = []$.
 - 2: **while** $\Delta(\mu^*) > \text{tol}$ **do**
 - 3: Compute full order solution at μ^* . Form the snapshot matrices:
 $\mathbf{X} = [x(t^1, \mu^*), x(t^2, \mu^*), \dots, x(t^K, \mu^*)]$ and
 $\mathbf{F}^{\text{new}} = [f(x(t^1, \mu^*)), f(x(t^2, \mu^*)), \dots, f(x(t^K, \mu^*))]$.
 - 4: Form $\bar{\mathbf{X}} := \mathbf{X} - \mathbf{V}\mathbf{V}^T\mathbf{X}$; obtain $\bar{\mathbf{X}} \xrightarrow{\text{svd}} \mathbf{U}\Sigma\mathbf{W}^T$.
 - 5: Update $\mathbf{V} := \text{orth}([\mathbf{V}, \mathbf{U}(:, 1 : \delta_{\text{RB}})])$ if $\delta_{\text{RB}} \geq 0$; remove δ_{RB} vectors from \mathbf{V} if $\delta_{\text{RB}} < 0$.
 - 6: Compute ℓ_{DEIM} (D)EIM interpolation basis vectors from snapshots of the nonlinear function,
 $\mathbf{F} := [\mathbf{F}, \mathbf{F}^{\text{new}}]$.
 - 7: Solve ROM and compute the error estimator $\Delta(\mu), \forall \mu \in \Xi$.
 - 8: Based on $\Delta(\mu)$, decide the new number δ_{RB} and δ_{DEIM} using the adaptive scheme in [5].
 $\ell_{\text{DEIM}} = \ell_{\text{DEIM}} + \delta_{\text{DEIM}}$.
 - 9: Next μ^* is chosen as, $\mu^* := \arg \max_{\mu \in \Xi} \Delta(\mu)$.
 - 10: **end while**
-

Algorithm 4 Adaptive POD-Greedy-(D)EIM with adaptive parameter sampling

Input: Coarse training set Ξ_c , Fine training set Ξ_f , Tolerance (tol).

Output: \mathbf{V} , DEIM basis $\mathbf{U}_f, \mathbf{P}_f$.

- 1: Initialize. $\mathbf{V} = [], \mu^* \in \Xi_c, \ell_{\text{RB}} = 1, \ell_{\text{DEIM}} = 1, \delta_{\text{DEIM}} = 0, \mathbf{U}_f = [], \mathbf{P}_f = []$.
 - 2: **while** $\Delta(\mu^*) > \text{tol}$ **do**
 - 3: Perform Steps 3-6 in Algorithm 3.
 - 4: Solve ROM and compute the error estimator $\Delta(\mu), \forall \mu \in \Xi_c$.
 - 5: Based on $\Delta(\mu)$, decide on the new number δ_{RB} and δ_{DEIM} using the adaptive scheme in [5].
 $\ell_{\text{DEIM}} = \ell_{\text{DEIM}} + \delta_{\text{DEIM}}$.
 - 6: Perform Steps 5-8 in Algorithm 2
 - 7: **end while**
-

With the proposed adaptive sampling technique, we present a fully adaptive POD-Greedy-(D)EIM algorithm: Algorithm 4. It is a combination of Algorithm 3 with the adaptive sampling approach as presented in Algorithm 2. Since the adaptive sampling scheme is the same as in Algorithm 2, we omit the detailed explanation for Algorithm 4. One only needs to keep in mind that the error estimator is computed *only* over the *coarse* training set, see Step 4. To estimate the errors at samples in the fine training set, the error surrogate $s(\mu)$ is computed and used, see Step 6.

3.4 A Strategy for Shape Parameter Selection

In [16], the authors propose a technique to adaptively sample parameters based on Kriging interpolation, where several hyper-parameters need to be estimated. As an alternative, we use RBF to interpolate an efficient output error estimator. Based on the choice of the kernel used, RBF involves at most one free parameter the user has to specify. This is the case for Gaussian or Multiquadric kernels, where the shape

parameter σ needs to be determined. For the case of Thin-plate spline kernels, there are no free parameters to choose. A good choice of the shape parameter is essential to ensure that the RBF kernel matrix (\mathbf{R}) remains well-conditioned [7]. In [18], a heuristic approach is introduced to determine the optimal shape parameter based on the idea of the Leave One Out Cross-Validation (LOOCV) strategy commonly used in the field of statistics [13]. The main idea of LOOCV is to utilize the available data and find a σ that best fits the data. Below, we briefly describe how we use it to estimate the shape parameter. For more details on the method and its generalization - the k-fold cross validation, we refer to [13].

- (i) We start with the available ℓ centers and data points $(\mu_i, \Delta(\mu_i))$, $\forall i = 1, \dots, \ell$ for the RBF interpolation.
- (ii) For every μ_i , $\forall i = 1, \dots, \ell$, compute a ‘less accurate’ radial basis interpolant $s^{\mu_i}(\mu)$ by removing the i th row and i th column of the RBF kernel matrix \mathbf{R} and the i th row of the input data vector \mathfrak{d} . Note that $s^{\mu_i}(\mu_i)$ is actually an approximation of $h(\mu_i)$, whereas, $s(\mu_i) = h(\mu_i)$ (recall Eq. (4)).
- (iii) Following this, we have the error between $h(\mu_i)$ and $s^{\mu_i}(\mu_i)$: $e^{\mu_i} := h(\mu_i) - s^{\mu_i}(\mu_i)$. Since $h(\mu_i) = s(\mu_i)$, we have

$$e^{\mu_i} = s(\mu_i) - s^{\mu_i}(\mu_i), \quad i = 1, \dots, \ell.$$

- (iv) Form the error vector $\mathbf{e} = [e^{\mu_1}, e^{\mu_2}, \dots, e^{\mu_\ell}]$.
- (v) Choose the optimal σ as $\sigma^* = \arg \min_{\sigma \in [\sigma_{\min}, \sigma_{\max}]} \|\mathbf{e}(\sigma)\|_2$.

We solve the minimization problem using the MATLAB[®] function `fminbnd`, as suggested in [7]. For the success of the adaptive sampling approach, the following criteria are crucial:

1. *Good data*: The input data to the RBF should be ‘good’. This means that we are interpolating something meaningful to the problem we are trying to approximate.
2. *Good shape parameter*: The RBF interpolation should be robust, ensuring that the interpolated values can act as ‘good’ surrogates of the actual values.

We ensure the first criterion by using an efficient output error estimator for the RB method from [5]. This is a residual based estimator suitable for nonlinear dynamical systems and tailored for output error estimation. Meanwhile, the second criterion is ensured by adopting the LOOCV strategy described above.

4 Numerical Results

In this section, we validate the proposed adaptive parameter sampling method using three examples. The first is a nonlinear Burgers’ equation model with one parameter. The second is a two-parameter linear convection-diffusion model. The last one is a three-parameter model of a microthruster unit. All numerical tests were performed

in MATLAB[®] 2015a, on a laptop with Intel[®] Core[™] i5-7200U @ 2.5 GHZ, with 8 GB of RAM. For all the examples,

- The Adaptive POD-Greedy-(D)EIM algorithm proposed in [5] is used for basis construction. Therefore, we compare the results of Algorithm 3 with fixed training set with those of Algorithm 4 with adaptive training set.
- The maximal output error over the test set Ξ_{test} is defined as

$$\varepsilon_{\max} := \max_{\mu \in \Xi_{\text{test}}} \left(\frac{1}{K} \sum_{k=1}^K \|y^k(\mu) - \bar{y}^k(\mu)\| \right). \quad (7)$$

- The mean (over time steps) output error over Ξ_{test} is defined as

$$\varepsilon(\mu) := \left(\frac{1}{K} \sum_{k=1}^K \|y^k(\mu_i) - \bar{y}^k(\mu_i)\| \right), \quad \forall \mu \in \Xi_{\text{test}}. \quad (8)$$

- The semi-implicit Euler scheme is used for time discretization.
- If not particularly pointed out, the *initial* coarse training set Ξ_c for Algorithm 4 is the same as the fixed training set Ξ for Algorithm 3.

4.1 Burgers' Equation

We consider the 1-D model of the viscous Burgers' equation. The PDE is defined in the spatial domain $w \in \Omega := [0, 1]$ and for time $T := [0, 2]$ as

$$\frac{dv}{dt} + v \frac{\partial v}{\partial w} = q \frac{\partial^2 v}{\partial w^2} + s(w, t), \quad v(0, t) = 0, \quad \frac{\partial v(1, t)}{\partial w} = 0.$$

where $v(w, t)$ is the unknown variable and $q \in \mathcal{P} := [0.001, 1]$ is the viscosity parameter. The initial condition is set as zero and a constant input to the system is $s(w, t) \equiv 1$. The output is monitored on the last spatial point in the domain $y = v(1, t)$. The model has $N = 500$ equations after discretization in space.

We show the results of Algorithms 3 and 4, respectively. The tolerance is set as $\text{tol} = 10^{-5}$. To implement the RB method without adaptive parameter sampling, i.e. Algorithm 3, we choose an initial training set consisting of 10 random parameter samples, using the `rng` command in MATLAB[®], with the seed value 112 for the random number generator `twister`. The fine training set Ξ_f for Algorithm 4 consists of 300 random samples from the same parameter domain, sampled using the seed value at 114 for the random number generator `twister`. Finally, 100 different random samples are considered for the test parameter set Ξ_{test} , sampled using `simdTwister` with the seed value 200. To construct the error surrogate model $s(\mu)$ in Algorithm 4, the Inverse Multiquadric (IMQ) kernel function is used for the RBF interpolation. Cross validation LOOCV has been applied to specify the shape

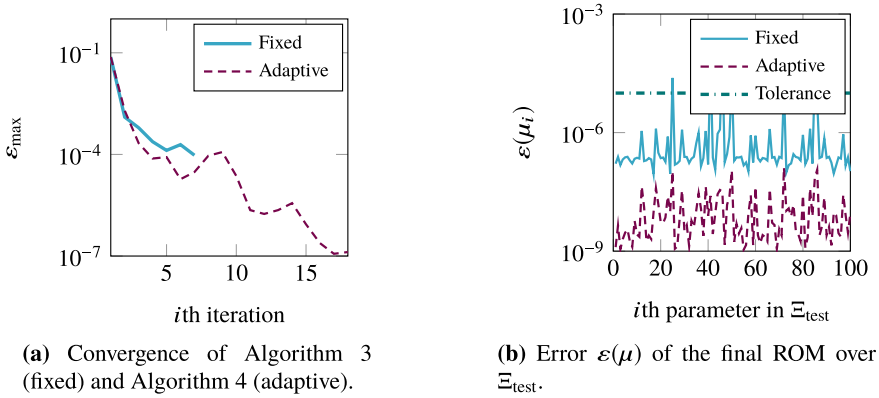


Fig. 1 Results for the Burgers' equation

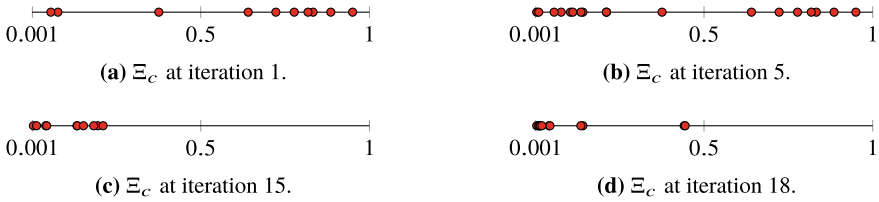


Fig. 2 Burgers' equation: training set evolution

parameter. Since the model is nonlinear, DEIM [4] is used in order to efficiently compute the nonlinear term.

In Fig. 1a shows the decay of the maximal output error, i.e. ε_{\max} , over the test set Ξ_{test} . It is calculated at every iteration of either Algorithm 3 or 4. The former results in a ROM of dimension 11, while the latter results in a ROM of dimension 28. Evidently, using an adaptive training set leads to a better convergence. Algorithm 3 converges in 7 iterations, while Algorithm 4 requires 18 iterations to converge. However, as seen from the output error $\varepsilon(\mu)$ of the final ROM over the test set Ξ_{test} in Fig. 1b, Algorithm 4 produces a ROM with errors being uniformly below the tolerance, whereas, the ROM computed using Algorithm 3 has large errors above the tolerance. In Fig. 2, we show the evolution of the coarse training set Ξ_c for different iterations of Algorithm 4. It is seen that the algorithm tends to pick parameters in the low-viscosity regions (close to 0.001), as expected, since the solution of the PDE tends to be 'less smooth', requiring more basis functions to approximate.

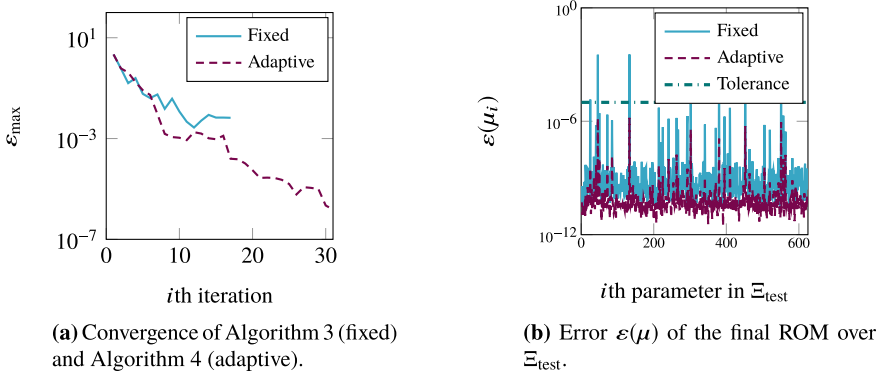


Fig. 3 Results for the Convection-diffusion equation

4.2 Convection-Diffusion Equation

Next, we consider a 1-D model of brain transport, originally discussed in [1] and also considered in [9, 22]. The transport is modelled as a linear convection-diffusion PDE defined in the spatial domain $w \in \Omega := [0, 1]$ and for time $T := [0, 1]$,

$$\frac{dv}{dt} = q_1 \frac{\partial^2 v}{\partial w^2} + q_2 \frac{\partial v}{\partial w} - q_2, \quad (9)$$

where $v(w, t)$ is the state vector and the two parameters $(q_1, q_2) \in \mathcal{P} := [0.001, 1] \times [0.5, 5]$ are the diffusion and convection constants, respectively. The boundary conditions are given by,

$$v(w, 0) = \begin{cases} 1, & w \leq 0.5 \\ 0, & \text{otherwise} \end{cases}, \quad v(0, t) = v(1, t) = 0.$$

We discretize the equation using the FDM on a grid yielding $n = 800$. The output is calculated as the average value of the state in a small interval Ω_o centered around the midpoint of the domain at $w = 0.5$.

$$y(t) := \frac{1}{|\Omega_o|} \int_{\Omega_o} v(w, t) dw, \quad \Omega_o := [0.495, 0.505].$$

The tolerance is set as $\text{tol} = 10^{-5}$. For Algorithm 3, we set Ξ as a set including 25 random samples from the parameter domain \mathcal{P} , picked using the `rng` command and making use of `twister`, with a seed of 112. The fine training set Ξ_f for Algorithm 4 includes 1600 equidistant samples in \mathcal{P} . We choose a test set Ξ_{test} consisting of 625 random samples. The test parameters are generated using the same random number generator as for the Burgers' equation example, viz., `simdTwister` with a seed

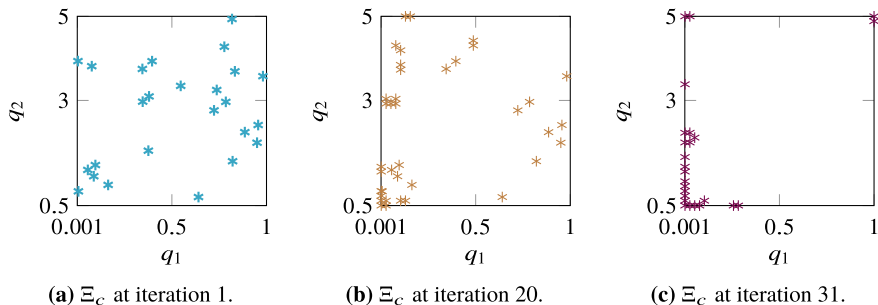


Fig. 4 Convection-diffusion equation: training set evolution

of 200. The error surrogate model $s(\mu)$ in Algorithm 4 is constructed using IMQ kernel function, where cross validation LOOCV has been applied to specify the shape parameter.

Figure 3a shows the maximal error decay (ε_{\max}) over Ξ_{test} , computed at each iteration of Algorithms 3 and 4, respectively. It is clear that adaptively enriching the training set leads to orders of magnitude faster convergence to the required accuracy. For Algorithm 3, the ROM size is 38 and for Algorithm 4, it is 63. Figure 3b plots the error $\varepsilon(\mu)$ (Eq. (8)) of the final ROM at every parameter in the test set Ξ_{test} , for the case of an adaptive training set and a fixed training set. We see that for some test parameters, the required tolerance is not met by the ROM computed using the fixed training set. In Fig. 4, the evolution of the training set is shown at different stages of Algorithm 4. It can be seen that samples from the left boundary of the parameter domain are added or kept. This has physical sense as this corresponds to the lower viscosity regions where the convective part of the solution dominates and the solution is ‘less smooth’. Thus, more basis functions are required for a good approximation.

In order to achieve a better ROM by using the fixed training set, we tried a Ξ with 225 random samples for Algorithm 3, leading to a ROM with error below the tolerance. In Table 2a, we provide the runtime results of both algorithms, where Algorithm 3 uses the refined training set with 225 samples. On the other hand, the initial coarse training set and the fine training set for Algorithm 4 remain the same. Algorithm 4 with adaptive sampling outperforms Algorithm 3, though both approaches produce accurate ROMs.

4.3 A Thermal Model

The final example is the Thermal model taken from the MORwiki benchmark collection.¹ It models the heat transfer in a microthruster unit. The system is governed by the following semi-discretized ODE,

¹https://morwiki.mpi-magdeburg.mpg.de/morwiki/index.php/Thermal_Model.

$$\mathbf{E}\dot{x} = \left(\mathbf{A} - \sum_{i=1}^3 h_i \mathbf{A}_i \right) x + \mathbf{B}u.$$

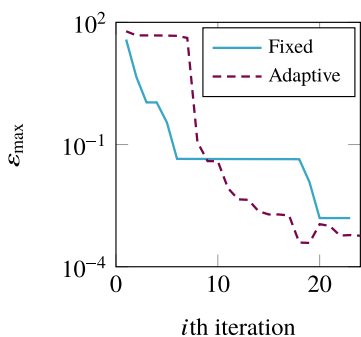
The model has been discretized in space using the FEM and the matrices are available from the Oberwolfach Benchmark Collection hosted at MORWiki. The dimension of the system is $n = 4257$. The parameters $\{h_i\}_{i=1}^3$ are the film coefficients at the top, bottom and side of the microthruster unit, respectively. They have a range between $[1, 10^8]$ in our simulations. The output is the temperature at the center of the polysilicon heater in the unit, which corresponds to the first row of the output matrix C .

Since this is a three-parameter problem with very large parameter domain, we use a large fixed training set (Ξ) with 6^3 logarithmically equidistant parameter samples for Algorithm 3 for a fair comparison. For Algorithm 4, the initial coarse training set includes only $N_c = 10$ randomly chosen samples. To generate these random samples, we first generate a uniform sampling of each parameter h_j , $j = 1, 2, 3$, given as, $h_{ij} = 10^{\frac{i}{N_c/8}}$, $i = 1, 2, \dots, N_c$, which leads to a matrix $(h_{ij}) \in \mathbb{R}^{N_c \times 3}$. Then, the rows of the matrix are permuted using the `rng` and `randperm` commands. The seeds and random number generators used are 100 (`twister`), 120 (`combRecursive`) and 600 (`combRecursive`), respectively. In this way, we get random samples which spread through the 3-D parameter domain. The fine training set Ξ_f for Algorithm 4 is made up of 16^3 equidistant samples, while the test set Ξ_{test} consists of 8^3 logarithmically equidistant samples. For the RBF interpolation, Thin-plate spline kernel is used. No shape parameter needs to be determined. The tolerance for this model is set as $\text{tol} = 10^{-3}$. In Fig. 5a, we show error decay ε_{max} over Ξ_{test} at each iteration of Algorithms 3 and Algorithm 4, respectively. Algorithm 3 takes 23 iterations to converge with the fixed training set Ξ . The resulting order of the ROM is 44. However, the maximum error ε_{max} (Eq. (7)) over the test set Ξ_{test} is still above the tolerance. Algorithm 4 converges in 24 iterations, to a ROM of order 74. The maximum error ε_{max} over Ξ_{test} is below the tolerance upon convergence. We note that, in the first few iterations, Algorithm 3 has a faster convergence in comparison to Algorithm 4. However, since the training set is fixed, the convergence eventually saturates. In Fig. 5b, we plot the error $\varepsilon(\mu)$ of the final ROM over Ξ_{test} , computed by the two algorithms respectively. Algorithm 4 once again outperforms Algorithm 3.

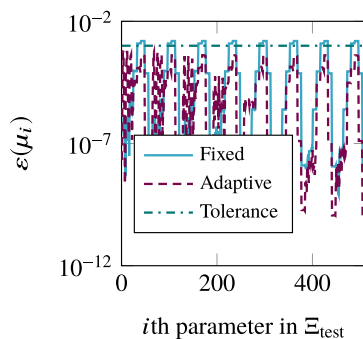
In Table 2b, we provide the runtime comparison between Algorithms 3 and 4 for the thermal model, where an obvious speed-up by Algorithm 4 is observed. More importantly, with the reduced runtime, Algorithm 4 produces a ROM with sufficient accuracy, whereas, the ROM computed by Algorithm 3 still does not meet the accuracy requirement. This further justifies the motivation of using adaptive sampling for models with two or more parameters, especially when the parameter domain is very large.

Table 2 Runtime comparison between Algorithms 3 and 4

| (a) Convection-diffusion example | |
|-------------------------------------|-------------------|
| Algorithm | Runtime (seconds) |
| Fixed training set (Algorithm 3) | 74.09 |
| Adaptive training set (Algorithm 4) | 41.52 |
| (b) Thermal example | |
| Algorithm | Runtime (seconds) |
| Fixed training set (Algorithm 3) | 151.16 |
| Adaptive training set (Algorithm 4) | 38.76 |



(a) Convergence of Algorithm 3 (fixed) and Algorithm 4 (adaptive).



(b) Error $\varepsilon(\mu_i)$ of the final ROM over Ξ_{test} .

Fig. 5 Results for the thermal model

5 Conclusion

Using a fixed training set in the POD-Greedy algorithm may either entail high computational costs or lead to large errors if the parameter domain is not properly sampled. In this work, we introduce a method to construct the training set for the RB method in an adaptive manner. To achieve this, we make use of an error surrogate model based on RBF interpolation.

The use of an error surrogate model enables sufficient exploration of the parameter space, since it avoids the need to solve the ROM at every parameter in the training set, as required by the error estimator. Furthermore, we implement a cross validation strategy in order to choose good shape parameters for the RBF kernels. The proposed algorithm is tested on several examples and it is shown to be effective in constructing a ROM with required accuracy. Furthermore, the adaptive parameter sampling method is integrated with the adaptive POD-Greedy-(D)EIM algorithm in [5] to achieve a fully adaptive scheme for the RB method.

As future work, we propose to extend the adaptive sampling approach to the frequency domain model reduction methods, such as multi-moment matching [8], in order to adaptively sample the interpolation points.

References

1. Banks, H.T., Kunisch, K.: Estimation techniques for distributed parameter systems. In: *Systems and Control: Foundations and Applications*, vol. 6. Birkhäuser, Boston (1989). <https://doi.org/10.1007/978-1-4612-3700-6>
2. Barrault, M., Maday, Y., Nguyen, N.C., Patera, A.T.: An ‘empirical interpolation’ method: application to efficient reduced-basis discretization of partial differential equations. *C.R. Acad. Sci. Paris* **339**(9), 667–672 (2004)
3. Buhmann, M.D.: *Radial Basis Functions: Theory and Implementations*. Cambridge Monographs on Applied and Computational Mathematics, vol. 12. Cambridge University Press, Cambridge (2003)
4. Chaturantabut, S., Sorensen, D.C.: Nonlinear model reduction via discrete empirical interpolation. *SIAM J. Sci. Comput.* **32**(5), 2737–2764 (2010). <https://doi.org/10.1137/090766498>
5. Chellappa, S., Feng, L., Benner, P.: Adaptive basis construction and improved error estimation for parametric nonlinear dynamical systems. *Int. J. Numer. Methods Eng.* **121**(23), 5320–5349 (2020). <https://doi.org/10.1002/nme.6462>
6. Eftang, J.L., Knezevic, D.J., Patera, A.T.: An hp certified reduced basis method for parametrized parabolic partial differential equations. *Math. Comput. Model. Dyn. Syst.* **17**(4), 395–422 (2011)
7. Fasshauer, G.E., Zhang, J.G.: On choosing “optimal” shape parameters for RBF approximation. *Numer. Algorithms* **45**(1), 345–368 (2007)
8. Feng, L., Benner, P.: A robust algorithm for parametric model order reduction based on implicit moment matching. In: Quarteroni, A., Rozza, G. (eds.) *Reduced Order Methods for Modeling and Computational Reduction*, MS & A, vol. 9, pp. 159–186. Springer, Berlin, Heidelberg, New York (2014). https://doi.org/10.1007/978-3-319-02090-7_6
9. Grepl, M.: Reduced-basis approximation a posteriori error estimation for parabolic partial differential equations. Ph.D. thesis, Massachusetts Institute of Technology (MIT), Cambridge, USA (2005)
10. Haasdonk, B., Dihlmann, M., Ohlberger, M.: 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**(4), 423–442 (2011)
11. Haasdonk, B., Ohlberger, M.: Reduced basis method for finite volume approximations of parametrized linear evolution equations. *ESAIM: Math. Model. Numer. Anal.* **42**(2), 277–302 (2008)
12. Hesthaven, J.S., Stamm, B., Zhang, S.: Efficient greedy algorithms for high-dimensional parameter spaces with applications to empirical interpolation and reduced basis methods. *ESAIM: Math. Model. Numer. Anal.* **48**(1), 259–283 (2014)
13. James, G., Witten, D., Hastie, T., Tibshirani, R.: *An Introduction to Statistical Learning?: With Applications in R*. Springer Texts in Statistics, vol. 103. Springer, New York (2013)
14. Maday, Y., Stamm, B.: Locally adaptive greedy approximations for anisotropic parameter reduced basis spaces. *SIAM J. Sci. Comput.* **35**(6), A2417–A2441 (2013)
15. Manzoni, A., Negri, F.: Heuristic strategies for the approximation of stability factors in quadratically nonlinear parametrized PDEs. *Adv. Comput. Math.* **41**(5), 1255–1288 (2015). <https://doi.org/10.1007/s10444-015-9413-4>
16. Paul-Dubois-Taine, A., Amsellem, D.: An adaptive and efficient greedy procedure for the optimal training of parametric reduced-order models. *Int. J. Numer. Methods Eng.* **102**(5), 1262–1292 (2015)

17. Quarteroni, A., Manzoni, A., Negri, F.: *Reduced Basis Methods for Partial Differential Equations*, *La Matematica per il 3+2*, vol. 92. Springer International Publishing (2016)
18. Rippa, S.: An algorithm for selecting a good value for the parameter c in radial basis function interpolation. *Adv. Comput. Math.* **11**(2), 193–210 (1999)
19. Sen, S.: Reduced-basis approximation and a posteriori error estimation for many-parameter heat conduction problems. *Numer. Heat Transf. Part B: Fundam.* **54**(5), 369–389 (2008)
20. van Stein, B., Wang, H., Kowalczyk, W., Emmerich, M.T.M., Bäck, T.: Cluster-based Kriging approximation algorithms for complexity reduction. *Appl. Intell.* **50**, 778–791 (2020)
21. Wendland, H.: *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics, vol. 17. Cambridge University Press, Cambridge (2005)
22. Zhang, Y., Feng, L., Li, S., Benner, P.: An efficient output error estimation for model order reduction of parametrized evolution equations. *SIAM J. Sci. Comput.* **37**(6), B910–B936 (2015)

Balanced Truncation Model Reduction for Lifted Nonlinear Systems



Boris Kramer and Karen Willcox

Abstract We present a balanced truncation model reduction approach for a class of nonlinear systems with time-varying and uncertain inputs. First, our approach brings the nonlinear system into quadratic-bilinear (QB) form via a process called lifting, which introduces transformations via auxiliary variables to achieve the specified model form. Second, we extend a recently developed QB balanced truncation method to be applicable to such lifted QB systems that share the common feature of having a system matrix with zero eigenvalues. We illustrate this framework and the multi-stage lifting transformation on a tubular reactor model. In the numerical results we show that our proposed approach can obtain reduced-order models that are more accurate than proper orthogonal decomposition reduced-order models in situations where the latter are sensitive to the choice of training data.

Keywords Balanced truncation · Nonlinear model reduction · Lifting transformation · Quadratic-bilinear models

1 Introduction

We consider reduced-order modeling for large-scale nonlinear systems with time-dependent and uncertain inputs, as such models appear in many practical engineering applications. A reduced-order model (ROM) has to capture the rich dynamics resulting from time-dependent inputs [1]. Proper orthogonal decomposition (POD) [18] is the most common model reduction framework for nonlinear systems. As a trajectory-

B. Kramer (✉)
Department of Mechanical and Aerospace Engineering, University of California San Diego,
San Diego, CA, USA
e-mail: bmramer@ucsd.edu

K. Willcox
Oden Institute for Computational Engineering Science, University of Texas at Austin,
Austin, TX, USA
e-mail: kwillcox@oden.utexas.edu

based method, POD relies on user-specified training data, leaving the method vulnerable to poor choices in training inputs.

Interpolatory model reduction approaches that focus on the transfer function mapping from inputs to outputs—and do not require training data—are mature for linear systems [1, 2]. Moreover, significant progress has been made for input-output model reduction for nonlinear systems with known governing equations [4, 6, 8]. In the situation where the governing equations are unknown and only frequency-domain input-output data is available, Antoulas and Gosea [13] proposed a data-driven reduced-order modeling framework for quadratic-bilinear (QB) systems.

Balanced truncation methods offer another promising avenue for model reduction. Initially proposed by Moore [24] for linear systems, many extensions are available, such as snapshot-based approximations (in the stable [29] and unstable case [11]), as well as weighted and time-limited balanced truncation methods, see e.g., the survey [16]. Extending balancing transformations to large-scale nonlinear systems is an open problem, since the balancing transformations become state-dependent (see [12, 27]) and are hence impractical when the model dimension is large. To overcome this computational bottleneck, approximate Gramians that are state-independent (as they are in the linear case) can be used. Balanced truncation for nonlinear systems based on algebraic Gramians [14] requires the solution of Lyapunov-type equations, and is hence appealing from a computational perspective. A computationally efficient framework via truncated Gramians for QB systems has been proposed in [5]. Empirical Gramians for nonlinear systems [21] can also be used, yet their computation requires as many simulations of the full systems as there are inputs and outputs: each simulation uses an impulse disturbance per input channel while setting the other inputs to zero. A method for approximating the nonlinear balanced truncation reduction map via reproducing kernel Hilbert spaces has been proposed in [7].

In this work, we propose a balanced truncation model reduction method for a class of nonlinear systems with time-varying inputs. As a first step, we use variable transformations and *lifting*, which allows us to bring the nonlinear system in QB form via the introduction of auxiliary variables. Lifting transformations that transform general nonlinear systems into polynomial models have been explored over several decades in different communities [4, 15, 19, 20, 22, 23, 26, 28]. We show that the lifted QB models often have a special structure, namely that the linear system matrix becomes has zero eigenvalues, making the balancing algorithm for QB systems [5] not directly applicable due to the typical requirement that the system matrix be stable. We thus propose a modified approach for balanced truncation of lifted QB systems, which uses the definition of the lifted variable to artificially introduce a stabilization to the system. As an example, we consider a nonlinear tubular reactor model for which we present a lifting transformation that brings the model into QB form, and subsequently perform balancing.

This paper is organized as follows. Section 2 reviews truncated Gramians for QB systems. Section 3 illustrates that lifted QB models often have a common structure and also introduces the tubular reactor model and its lifted QB model. Section 4 proposes our balancing approach for lifted QB systems with special structure. Section 5 presents numerical results for the tubular reactor model.

2 Quadratic-Bilinear Systems and Balancing

A QB system of ordinary differential equations can be written as

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{H}(\mathbf{x} \otimes \mathbf{x}) + \sum_{k=1}^m (\mathbf{N}_k \mathbf{x}) \mathbf{u}_k + \mathbf{B}\mathbf{u}, \quad (1)$$

$$\mathbf{y} = \mathbf{C}\mathbf{x}, \quad (2)$$

where $\mathbf{x}(t) \in \mathbb{R}^n$ is the state, $t \geq 0$ denotes time, the initial condition is $\mathbf{x}(0) = \mathbf{x}_0$, $\mathbf{u}(t) \in \mathbb{R}^m$ is a time-dependent input and \mathbf{u}_k its k th component, $\mathbf{B} \in \mathbb{R}^{n \times m}$ is the input matrix, $\mathbf{A} \in \mathbb{R}^{n \times n}$ is the system matrix, $\mathbf{C} \in \mathbb{R}^{p \times n}$ is the output matrix and the $\mathbf{N}_k \in \mathbb{R}^{n \times n}$, $k = 1, \dots, m$ represent the bilinear coupling between input and state. The matrix \mathbf{H} is viewed as a mode-1 matricization of a tensor $\mathcal{H} \in \mathbb{R}^{n \times n \times n}$, so $\mathcal{H}^{(1)} = \mathbf{H}$. Moreover, $\mathcal{H}^{(2)}$ denotes the mode-2 matricization of the tensor \mathcal{H} . We assume without loss of generality (see [4, Sect. 3.1]) that the matrix \mathbf{H} is symmetric in that $\mathbf{H}(\mathbf{x}_1 \otimes \mathbf{x}_2) = \mathbf{H}(\mathbf{x}_2 \otimes \mathbf{x}_1)$.

For system (1)–(2) the controllability and observability energy functions are defined as

$$\mathcal{E}_c(\mathbf{x}_0) := \min_{\mathbf{u} \in L_2(-\infty, 0), \mathbf{x}(-\infty) = \mathbf{0}, \mathbf{x}(0) = \mathbf{x}_0} \frac{1}{2} \int_{-\infty}^0 \|\mathbf{u}(t)\|^2 dt,$$

$$\mathcal{E}_o(\mathbf{x}_0) := \frac{1}{2} \int_0^{\infty} \|\mathbf{y}(t)\|^2 dt,$$

where \mathcal{E}_c quantifies the minimum amount of energy required to steer the system from $\mathbf{x}(-\infty) = \mathbf{0}$ to $\mathbf{x}(0) = \mathbf{x}_0$, and \mathcal{E}_o quantifies the output energy generated by the nonzero initial condition \mathbf{x}_0 and $\mathbf{u}(t) \equiv 0$.

Balanced truncation achieves reduction in the dimension of the state space by eliminating those states that are hard to control (large \mathcal{E}_c) and hard to observe (small \mathcal{E}_o). Thus, an important component of balanced truncation model reduction is to find the coordinate transformation that ranks the controllability and observability of the states. The *Gramian* matrices (formally introduced below) are central to finding this balanced coordinate transformation for system (1)–(2). In particular, if symmetric positive definite Gramians $\mathbf{P} = \mathbf{L}_P \mathbf{L}_P^\top$ and $\mathbf{Q} = \mathbf{L}_Q \mathbf{L}_Q^\top$ for the QB system (1)–(2) are available, we can obtain transformation matrices \mathbf{V} and \mathbf{W} , with $\mathbf{V} = \mathbf{W}^{-1}$ that diagonalize the Gramians, $\mathbf{V}^\top \mathbf{P} \mathbf{V} = \mathbf{W}^\top \mathbf{Q} \mathbf{W} = \mathcal{S} = \text{diag}(\sigma_1, \dots, \sigma_n)$, where σ_i are the singular values¹ of $\mathbf{L}_Q^\top \mathbf{L}_P$. The energy functions in the balanced state $\tilde{\mathbf{x}} = \mathbf{W}\mathbf{x}$ can then be bounded as $\mathcal{E}_c(\tilde{\mathbf{x}}) \geq \frac{1}{2} \tilde{\mathbf{x}}^\top \mathcal{S}^{-1} \tilde{\mathbf{x}}$, and $\mathcal{E}_o(\tilde{\mathbf{x}}) \leq \frac{1}{2} \tilde{\mathbf{x}}^\top \mathcal{S} \tilde{\mathbf{x}}$ in a neighborhood of the origin [5, Sect. 4]. From these inequalities, we see that states corresponding to small singular values of $\mathbf{L}_Q^\top \mathbf{L}_P$ are *hard to control* and *hard to observe*. The subspaces spanned by the singular vectors corresponding to small singular values can thus be discarded in a ROM.

¹ In the linear case those would be the Hankel singular values.

We now turn to the computation of the Gramian matrices \mathbf{P} and \mathbf{Q} . For the special case of QB systems, the next proposition states conditions under which approximate algebraic Gramians exist, and suggests a computational framework to find those.

Proposition 1 ([5], Cor. 3.4) *Consider the QB system (1)–(2). If \mathbf{A} is a stable matrix, then the truncated Gramians \mathbf{P}_T , \mathbf{Q}_T are defined as solutions to*

$$\mathbf{A}\mathbf{P}_T + \mathbf{P}_T\mathbf{A}^\top + \mathbf{H}(\mathbf{P}_1 \otimes \mathbf{P}_1)\mathbf{H}^\top + \sum_{k=1}^m \mathbf{N}_k \mathbf{P}_1 \mathbf{N}_k^\top + \mathbf{B}\mathbf{B}^\top = \mathbf{0}, \quad (3)$$

$$\mathbf{A}^\top \mathbf{Q}_T + \mathbf{Q}_T \mathbf{A} + \mathcal{H}^{(2)}(\mathbf{P}_1 \otimes \mathbf{Q}_1)(\mathcal{H}^{(2)})^\top + \sum_{k=1}^m \mathbf{N}_k^\top \mathbf{Q}_1 \mathbf{N}_k + \mathbf{C}^\top \mathbf{C} = \mathbf{0}, \quad (4)$$

where \mathbf{P}_1 and \mathbf{Q}_1 are solutions to the standard linear Lyapunov equations

$$\mathbf{A}\mathbf{P}_1 + \mathbf{P}_1\mathbf{A}^\top + \mathbf{B}\mathbf{B}^\top = \mathbf{0}, \quad \mathbf{A}^\top \mathbf{Q}_1 + \mathbf{Q}_1\mathbf{A} + \mathbf{C}^\top \mathbf{C} = \mathbf{0}. \quad (5)$$

The Gramians \mathbf{P}_T , \mathbf{Q}_T are used to obtain a QB reduced-order model (QB-ROM) as follows: Compute the singular value decomposition $\mathbf{L}_{\mathbf{Q}_T}^\top \mathbf{L}_{\mathbf{P}_T} = \mathcal{U}\mathcal{S}\mathcal{V}^\top$ and let its rank r approximation be denoted as $\mathcal{U}_r \mathcal{S}_r \mathcal{V}_r^\top$. The projection matrices are $\mathbf{W} = \mathbf{L}_{\mathbf{Q}_T} \mathcal{U}_r \mathcal{S}_r^{-1/2}$ and $\mathbf{V} = \mathbf{L}_{\mathbf{P}_T} \mathcal{V}_r \mathcal{S}_r^{-1/2}$. The balanced QB-ROM has the form

$$\dot{\hat{\mathbf{x}}} = \hat{\mathbf{A}}\hat{\mathbf{x}} + \hat{\mathbf{H}}(\hat{\mathbf{x}} \otimes \hat{\mathbf{x}}) + \sum_{k=1}^m \hat{\mathbf{N}}_k \hat{\mathbf{x}} \mathbf{u}_k + \hat{\mathbf{B}}\mathbf{u}, \quad (6)$$

$$\hat{\mathbf{y}} = \hat{\mathbf{C}}\hat{\mathbf{x}}. \quad (7)$$

with the matrices $\hat{\mathbf{A}} = \mathbf{W}^\top \mathbf{A}\mathbf{V}$, $\hat{\mathbf{B}} = \mathbf{W}^\top \mathbf{B}$, $\hat{\mathbf{C}} = \mathbf{C}\mathbf{V}$, $\hat{\mathbf{N}}_k = \mathbf{W}^\top \mathbf{N}_k \mathbf{V}$, $\hat{\mathbf{H}} = \mathbf{W}^\top \mathbf{H}(\mathbf{V} \otimes \mathbf{V})$. The Gramians of the QB-ROM are balanced, i.e., the matrices $\hat{\mathbf{P}}_T = \hat{\mathbf{Q}}_T$ are equal and diagonal.

3 Lifting Systems to QB Form: Introducing Structure

In Sect. 3.1 we illustrate that lifting transformations for nonlinear systems often lead to a specific structure of the lifted system. Section 3.2 introduces a nonlinear tubular reactor model, and its subsequent lifted QB model with this special structure.

3.1 Structure in Lifted QB Systems

As an example of a higher-order dynamical system that can be lifted to QB form, consider the nonlinear ordinary differential equation

$$\dot{x} = \sum_{k=1}^d a_k x^k + bu, \quad (8)$$

where $x(t)$ is the one-dimensional state variable, a_1, \dots, a_d are known coefficients, x^k is the k th power of $x(t)$, $u(t)$ is an input function, and $b \in \mathbb{R}$ is a parameter. We consider here a one-dimensional ordinary differential equation (ODE) for illustration and note that the lifting framework directly applies to n -dimensional nonlinear systems. Our goal is to rewrite this system in QB form by introducing auxiliary variables. For this example, we define auxiliary variables $w_\ell := x^{\ell+1}$, so that the governing equation can be written as a linear system

$$\dot{x} = a_1 x + \sum_{k=1}^{d-1} a_{k-1} w_k + bu. \quad (9)$$

The auxiliary state dynamics are computed via the chain rule (or Lie derivative) as

$$\dot{w}_\ell = \frac{d}{dt} x^{\ell+1} = (\ell+1)x^\ell \dot{x} = (\ell+1)w_{\ell-1} \left(a_1 x + \sum_{k=1}^{d-1} a_{k-1} w_k + bu \right),$$

and here $w_0 := x$. Note that the right-hand-side contains quadratic products $w_{\ell-1}(t)w_k(t)$ as well as bilinear products $b(\ell+1)u(t)w_{\ell-1}(t)$. Another choice to write the auxiliary dynamics would be to replace $w_{\ell-1}x$ with w_ℓ . Next, we show two examples of polynomial and non-polynomial models where lifting leads to a similar structure.

Example 1 Consider the ODE

$$\dot{x} = ax^3 + bu. \quad (10)$$

Our goal is to lift this system to QB form by introducing auxiliary variables. Let

$$w_1 = x^2, \quad w_2 = x^3,$$

be the auxiliary variables. Note that the original dynamics then become linear, i.e., $\dot{x} = aw_2 + bu$. We compute the auxiliary state dynamics $\dot{w}_1 = 2x\dot{x} = 2axw_2 + 2xbu$ and $\dot{w}_2 = 3x^2\dot{x} = 3w_1(aw_2 + bu)$. Taken together, the nonlinear equation (10) with one state variable is equivalent to the QB-ODE with three state variables

$$\dot{x} = aw_2 + bu, \quad \dot{w}_1 = 2axw_2 + 2bxu, \quad \dot{w}_2 = 3aw_1w_2 + 3bw_1u. \quad (11)$$

The system (11) is equivalent to the original nonlinear equation (10), in the sense that both systems yield the same solution $x(t)$. We can write (11) in the QB form of equation (1) with the lifted state $\mathbf{x} = [x, w_1, w_2]^\top$ and matrices

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & a \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2a & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3a & 0 & 0 \end{bmatrix}, \quad \mathbf{N}_1 = \begin{bmatrix} 0 & 0 & 0 \\ 2b & 0 & 0 \\ 0 & 3b & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}. \quad (12)$$

Example 2 Consider an ODE with non-polynomial term and linear term:

$$\dot{x} = ax + e^{-x} + bu,$$

which we lift to QB form. We introduce $w_1 = e^{-x}$ as the auxiliary variable, so that $\dot{w}_1 = -w_1(ax + w_1 + bu) = -aw_1x - w_1^2 - w_1bu$, which can be written in QB form as

$$\begin{bmatrix} \dot{x} \\ \dot{w}_1 \end{bmatrix} = \begin{bmatrix} a & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ w_1 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix} \left(\begin{bmatrix} x \\ w_1 \end{bmatrix} \otimes \begin{bmatrix} x \\ w_1 \end{bmatrix} \right) + \begin{bmatrix} 0 & 0 \\ 0 & -b \end{bmatrix} \begin{bmatrix} x \\ w_1 \end{bmatrix} u + \begin{bmatrix} b \\ 0 \end{bmatrix} u. \quad (13)$$

We observe that the lifted matrices in the previous two examples (Eqs. (12) and (13)) have the following block-structure:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} \mathbf{0} \\ \mathbf{H}_2 \end{bmatrix}, \quad \mathbf{N} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{N}_{21} & \mathbf{N}_{22} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{0} \end{bmatrix}, \quad (14)$$

where $\mathbf{0}$ is a matrix of zeros of appropriate dimensions. Note, that the system matrix of the lifted system (14) has zero eigenvalues. Moreover, the original input $\mathbf{B}u(t)$ only affects the original state, but appears as a bilinear term in the auxiliary states after application of the chain rule. Thus, the \mathbf{N} matrix is nonzero only in the rows corresponding to the auxiliary states. The matrix structure in Eq. (14) commonly occurs in QB systems that were obtained from lifting transformations, for instance the systems given in Example 1 and Example 2. Nevertheless, due to the non-uniqueness of lifting transformations, the block structure in (14) is both a result of the original dynamical system form and the chosen lifting transformation.

3.2 Tubular Reactor Model

We consider a non-adiabatic tubular reactor model with single reaction as in [17] and follow the discretization in [30]. The model describes the evolution of the species concentration $\psi(s, t)$ and temperature $\theta(s, t)$ with spatial variable $s \in (0, 1)$ and time $t > 0$. The PDE model is

$$\frac{\partial \psi}{\partial t} = \frac{1}{Pe} \frac{\partial^2 \psi}{\partial s^2} - \frac{\partial \psi}{\partial s} - \mathcal{D}f(\psi, \theta; \gamma), \quad (15)$$

$$\frac{\partial \theta}{\partial t} = \frac{1}{Pe} \frac{\partial^2 \theta}{\partial s^2} - \frac{\partial \theta}{\partial s} - \beta(\theta - \theta_{\text{ref}}) + \mathcal{B}\mathcal{D}f(\psi, \theta; \gamma) + bu, \quad (16)$$

with input function $u = u(t)$ and a function $b = b(s)$ encoding the influence of the input to the computational domain. The parameters are the Damköhler number \mathcal{D} , Péclet number Pe as well as known constants \mathcal{B} , β , θ_{ref} , γ . The polynomial nonlinear term that drives the reaction is

$$f(\psi, \theta; \gamma) = \psi(c_0 + c_1\theta + c_2\theta^2 + c_3\theta^3)$$

with given constants c_0, \dots, c_3 . Robin boundary conditions are imposed on the left boundary of the domain and Neumann boundary conditions on the right

$$\begin{aligned} \frac{\partial \psi}{\partial s}(0, t) &= Pe(\psi(0, t) - 1), & \frac{\partial \theta}{\partial s}(0, t) &= Pe(\theta(0, t) - 1), \\ \frac{\partial \psi}{\partial s}(1, t) &= 0, & \frac{\partial \theta}{\partial s}(1, t) &= 0. \end{aligned}$$

The initial conditions are prescribed as $\psi(s, 0) = \psi_0(s)$, and $\theta(s, 0) = \theta_0(s)$. The output of interest is the temperature oscillation at the reactor exit, i.e., the quantity $y(t) = \theta(s = 1, t)$. The diffusive and convective terms are approximated with second-order centered differences in the interior of the domain. Second-order forward and backward difference schemes are used for the inflow and outflow boundary conditions, respectively. The finite difference approximation of dimension $2n$ of the tubular reactor PDE is then

$$\dot{\psi} = \mathbf{A}_\psi \psi + \mathbf{b}_\psi - \mathcal{D} \psi \odot (c_0 + c_1 \odot \theta + c_2 \odot \theta^2 + c_3 \odot \theta^3), \quad (17)$$

$$\dot{\theta} = \mathbf{A}_\theta \theta + \mathbf{b}_\theta + \mathbf{b}u + \mathcal{B}\mathcal{D} \psi \odot (c_0 + c_1 \odot \theta + c_2 \odot \theta^2 + c_3 \odot \theta^3). \quad (18)$$

Here, the (Hadamard) componentwise product of two vectors is denoted as $[\psi \odot \theta]_i = \psi_i \theta_i$, and the powers of vectors are also taken componentwise. The constant vector \mathbf{b}_ψ encodes the boundary condition, and \mathbf{b}_θ is the sum of contributions from the boundary conditions and the $\beta \cdot \theta_{\text{ref}}$ term in Eq. (16). The term $\mathbf{A}_\psi \psi$ is a discretized version of the advection-diffusion terms $\frac{1}{Pe} \psi_{,ss} - \psi_s$ and $\mathbf{A}_\theta \theta$ is a discretized representation of the $\frac{1}{Pe} \theta_{,ss} - \theta_s - \beta\theta$ terms.

We quadratic-bilinearize the finite dimensional system via a lifting transformation.²³ To lift the system, we introduce the auxiliary variables

$$\mathbf{w}_1 = \psi \odot \theta, \quad \mathbf{w}_2 = \psi \odot \theta^2, \quad \mathbf{w}_3 = \psi \odot \theta^3, \quad \mathbf{w}_4 = \theta^2, \quad \mathbf{w}_5 = \theta^3,$$

so that we have the equivalent QB system of dimension $7n$:

² Lifting is not unique, and there might be lower-order quadratic systems than the one suggested here. However, how to obtain them is an open problem.

³ Lifting for a tubular reactor model with Arrhenius reaction term has been considered by the authors in [20], however that lifting transformation resulted in QB-DAEs, for which balancing model reduction is an open problem.

$$\dot{\boldsymbol{\psi}} = [\mathbf{A}_\psi - \mathcal{D}\text{diag}(c_0)]\boldsymbol{\psi} + \mathbf{b}_\psi - \mathcal{D}(c_1 \odot \mathbf{w}_1 + c_2 \odot \mathbf{w}_2 + c_3 \odot \mathbf{w}_3), \quad (19)$$

$$\dot{\boldsymbol{\theta}} = \mathbf{A}_\theta \boldsymbol{\theta} + \mathcal{B}\mathcal{D}c_0 \odot \boldsymbol{\psi} + \mathbf{b}_\theta + \mathbf{b}u + \mathcal{B}\mathcal{D}(c_1 \odot \mathbf{w}_1 + c_2 \odot \mathbf{w}_2 + c_3 \odot \mathbf{w}_3), \quad (20)$$

$$\dot{\mathbf{w}}_1 = \dot{\boldsymbol{\psi}} \odot \boldsymbol{\theta} + \boldsymbol{\psi} \odot \dot{\boldsymbol{\theta}}, \quad (21)$$

$$\dot{\mathbf{w}}_2 = \dot{\boldsymbol{\psi}} \odot \mathbf{w}_4 + 2\mathbf{w}_1 \odot \dot{\boldsymbol{\theta}}, \quad (22)$$

$$\dot{\mathbf{w}}_3 = \dot{\boldsymbol{\psi}} \odot \mathbf{w}_5 + 3\mathbf{w}_2 \odot \dot{\boldsymbol{\theta}}, \quad (23)$$

$$\dot{\mathbf{w}}_4 = 2\boldsymbol{\theta} \odot \dot{\boldsymbol{\theta}}, \quad (24)$$

$$\dot{\mathbf{w}}_5 = 3\mathbf{w}_4 \odot \dot{\boldsymbol{\theta}}. \quad (25)$$

Note, that the original state equations for species and temperature, (19)–(20), are linear in the new state variables, whereas the differential equations for the added state variables (21)–(25) are quadratic in the state variables (after inserting $\boldsymbol{\psi}$, $\boldsymbol{\theta}$, which are linear). Hence, the system is of QB form (1)–(2), where $\mathbf{x}(t) = [\boldsymbol{\psi}^\top \boldsymbol{\theta}^\top \mathbf{w}_1^\top \mathbf{w}_2^\top \mathbf{w}_3^\top \mathbf{w}_4^\top \mathbf{w}_5^\top]^\top \in \mathbb{R}^{7n}$ is the new lifted state. The lifted state Eqs. (19)–(25) again induce a specific structure of the system matrices:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0}_{5n \times 2n} & \mathbf{0}_{5n} \end{bmatrix} \in \mathbb{R}^{7n \times 7n}, \quad \mathbf{A}_{11} = \begin{bmatrix} \mathbf{A}_\psi - \mathcal{D}\text{diag}(c_0) & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_\theta \end{bmatrix}, \quad (26a)$$

$$\mathbf{H} = \begin{bmatrix} \mathbf{0}_{2n \times 4n^2} & \mathbf{0}_{2n \times 45n^2} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{bmatrix} \in \mathbb{R}^{7n \times (7n)^2}, \quad \mathbf{N} = \begin{bmatrix} \mathbf{0}_{2n \times 2n} & \mathbf{0}_{2n \times 5n} \\ \mathbf{N}_{21} & \mathbf{N}_{22} \end{bmatrix} \in \mathbb{R}^{7n \times 7n}, \quad (26b)$$

$$\mathbf{B} = \begin{bmatrix} \mathbf{b}_\psi & \mathbf{0} \\ \mathbf{b}_\theta & \mathbf{b} \\ \mathbf{0}_{5n} & \mathbf{0}_{5n} \end{bmatrix} \in \mathbb{R}^{7n \times 2}, \quad \mathbf{C} = [\mathbf{C}_1 \ \mathbf{0}_{5n}] \in \mathbb{R}^{1 \times 7n}, \quad \mathbf{C}_1 = [\mathbf{0}_{2n-1} \ 1]. \quad (26c)$$

Note that the original dynamics are linear after the lifting is applied, and the auxiliary states have no linear parts, as the time derivative of the auxiliary variables consistently results in purely quadratic dynamics. Thus, the matrix \mathbf{A} has zero eigenvalues, and \mathbf{H} is such that there is no contribution in the rows corresponding to the original state. Moreover, the bilinear state-input interactions only occur in the auxiliary states. This structure is exactly the one presented in Eq. (14).

4 Balancing for Lifted Systems with Special Structure

Section 4.1 introduces an artificial numerical parameter that guarantees existence of truncated Gramians. Section 4.2 derives the Gramians for the linear part of the system, while Sect. 4.3 presents the truncated Gramians for the QB system.

4.1 Artificial Stabilization

The special structure of the system matrix \mathbf{A} that arises from lifting transformations leads to zero eigenvalues. Hence, there are no unique solutions to the Lyapunov equations (3)–(5) and the standard QB balanced truncation method from [5] cannot be applied. To circumvent this problem, we observe that if the original nonlinear system is polynomial (cf. (8) and its lifted version (9)) then $-\alpha w_\ell + \alpha x^{\ell+1} = 0$ for any $\alpha \in \mathbb{R}$. We can use this equality to artificially stabilize the linear term. For example, in Eq. (21) of the tubular reactor ODE, we can introduce a term $-\alpha \mathbf{w}_1 + \alpha(\boldsymbol{\psi} \odot \boldsymbol{\theta})$. Applying this to every auxiliary variable, the linear system matrix becomes

$$\mathbf{A}(\alpha) = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & -\alpha \mathbf{I} \end{bmatrix}, \quad (27)$$

and the quadratic tensor $\mathbf{H}(\alpha) = \mathbf{H} + \tilde{\mathbf{H}}(\alpha)$, where $\tilde{\mathbf{H}}(\alpha)$ contains the artificially added terms such that $-\alpha[\mathbf{0}^\top, \mathbf{w}^\top]^\top + \tilde{\mathbf{H}}(\alpha)(\mathbf{x} \otimes \mathbf{x}) = \mathbf{0}$. With this system matrix $\mathbf{A}(\alpha)$, the Lyapunov equations (3)–(5) have unique positive semi-definite solutions if \mathbf{A}_{11} is stable. Similar concepts have been used recently in [25]. For all results in this section to hold, we require that \mathbf{A}_{11} is stable.

4.2 Balancing with Gramians of Linearized System

The Lyapunov equations for the linear system part, Eq. (5) are needed to compute the truncated Gramians for the QB system. The next two propositions detail the computation of those linear Lyapunov equations.

Proposition 2 *The solutions to the standard linear Lyapunov equations (5) with $\mathbf{A}(\alpha)$ from equation (27) are given by*

$$\mathbf{P}_1 = \begin{bmatrix} \mathbf{P}_{11} & \boldsymbol{\theta} \\ \mathbf{0} & \boldsymbol{\theta} \end{bmatrix}, \quad \mathbf{Q}_1 = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12}(\alpha) \\ \mathbf{Q}_{12}^\top(\alpha) & \frac{1}{2\alpha} \tilde{\mathbf{Q}}_{22}(\alpha) \end{bmatrix}, \quad (28)$$

where

$$\begin{aligned} \boldsymbol{\theta} &= \mathbf{A}_{11} \mathbf{P}_{11} + \mathbf{P}_{11} \mathbf{A}_{11}^\top + \mathbf{B}_1 \mathbf{B}_1^\top, \\ \boldsymbol{\theta} &= \mathbf{A}_{11}^\top \mathbf{Q}_{11} + \mathbf{Q}_{11} \mathbf{A}_{11} + \mathbf{C}_1^\top \mathbf{C}_1, \\ \mathbf{Q}_{12}(\alpha) &= -(\mathbf{A}_{11}^\top - \alpha \mathbf{I})^{-1} \mathbf{Q}_{11} \mathbf{A}_{12}, \\ \tilde{\mathbf{Q}}_{22}(\alpha) &= \mathbf{A}_{12}^\top \mathbf{Q}_{12}(\alpha) + \mathbf{Q}_{12}(\alpha)^\top \mathbf{A}_{12}. \end{aligned}$$

Proof For the controllability Gramian computation we have that

$$\mathbf{0} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & -\alpha\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{12}^\top & \mathbf{P}_{22} \end{bmatrix} + \begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{12}^\top & \mathbf{P}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11}^\top & \mathbf{0} \\ \mathbf{A}_{12}^\top & -\alpha\mathbf{I} \end{bmatrix} + \begin{bmatrix} \mathbf{B}_1\mathbf{B}_1^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

which results in the four equations

$$\begin{aligned} \mathbf{A}_{11}\mathbf{P}_{11} + \mathbf{A}_{12}\mathbf{P}_{12}^\top + \mathbf{P}_{11}\mathbf{A}_{11}^\top + \mathbf{P}_{12}\mathbf{A}_{12}^\top + \mathbf{B}_1\mathbf{B}_1^\top &= \mathbf{0}, \\ (\mathbf{A}_{11} - \alpha\mathbf{I})\mathbf{P}_{12} + \mathbf{A}_{12}\mathbf{P}_{22} &= \mathbf{0}, \\ \mathbf{P}_{12}^\top(\mathbf{A}_{11}^\top - \alpha\mathbf{I}) + \mathbf{P}_{22}\mathbf{A}_{12}^\top &= \mathbf{0}, \\ -2\alpha\mathbf{P}_{22} &= \mathbf{0}. \end{aligned}$$

Since $(\mathbf{A}_{11}^\top - \alpha\mathbf{I})$ is invertible, the last three equations yield that $\mathbf{P}_{22} = \mathbf{0}$ and $\mathbf{P}_{12} = \mathbf{0}$. Next, we consider the observability Lyapunov equation

$$\mathbf{0} = \begin{bmatrix} \mathbf{A}_{11}^\top & \mathbf{0} \\ \mathbf{A}_{12}^\top & -\alpha\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{12}^\top & \mathbf{Q}_{22} \end{bmatrix} + \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{12}^\top & \mathbf{Q}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & -\alpha\mathbf{I} \end{bmatrix} + \begin{bmatrix} \mathbf{C}_1^\top\mathbf{C}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Blockwise multiplication results in the three distinct equations

$$\begin{aligned} \mathbf{A}_{11}^\top\mathbf{Q}_{11} + \mathbf{Q}_{11}\mathbf{A}_{11} + \mathbf{C}_1^\top\mathbf{C}_1 &= \mathbf{0}, \\ (\mathbf{A}_{11}^\top - \alpha\mathbf{I})\mathbf{Q}_{12} + \mathbf{Q}_{11}\mathbf{A}_{12} &= \mathbf{0}, \\ \mathbf{A}_{12}^\top\mathbf{Q}_{12} + \mathbf{Q}_{12}^\top\mathbf{A}_{12} - 2\alpha\mathbf{Q}_{22} &= \mathbf{0}, \end{aligned}$$

which we can solve successively to get \mathbf{Q}_{11} , \mathbf{Q}_{12} , \mathbf{Q}_{22} . We define $\tilde{\mathbf{Q}}_{22}(\alpha) = \mathbf{A}_{12}^\top\mathbf{Q}_{12}(\alpha) + \mathbf{Q}_{12}^\top(\alpha)\mathbf{A}_{12}$ so that the explicit dependence on $1/\alpha$ appears as $\mathbf{Q}_{22}(\alpha) = \frac{1}{2\alpha}\tilde{\mathbf{Q}}_{22}(\alpha)$. \square

Proposition 3 Consider the Gramians from Proposition 2, and let $\mathbf{P}_{11} = \mathbf{L}_{\mathbf{P}_{11}}\mathbf{L}_{\mathbf{P}_{11}}^\top$ and $\mathbf{Q}_{11} = \mathbf{L}_{\mathbf{Q}_{11}}\mathbf{L}_{\mathbf{Q}_{11}}^\top$ be Cholesky factorizations. Let the singular value decomposition approximate the product $\mathbf{L}_{\mathbf{Q}_{11}}^\top\mathbf{L}_{\mathbf{P}_{11}} \approx \mathcal{U}_r\mathcal{S}_r\mathcal{V}_r^\top$. Then the projection matrices for the balancing transformation are given as

$$\mathbf{V} = \begin{bmatrix} \mathbf{L}_{\mathbf{P}_{11}}\mathcal{V}_r\mathcal{S}_r^{-1/2} \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} \mathbf{L}_{\mathbf{Q}_{11}}\mathcal{U}_r\mathcal{S}_r^{-1/2} \\ \mathbf{Q}_{12}(\alpha)^\top\mathbf{L}_{\mathbf{Q}_{11}}^{-\top}\mathcal{U}_r\mathcal{S}_r^{-1/2} \end{bmatrix}. \quad (29)$$

Proof To obtain the balancing transformation, we need to compute the product $\mathbf{L}_{\mathbf{Q}_{11}}^\top\mathbf{L}_{\mathbf{P}_{11}}$. We have that

$$\mathbf{P}_1 = \mathbf{L}_{\mathbf{P}_1}\mathbf{L}_{\mathbf{P}_1}^\top = \begin{bmatrix} \mathbf{L}_{\mathbf{P}_{11}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{L}_{\mathbf{P}_{11}}^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

For the Cholesky factorization of \mathbf{Q}_1 we introduce the Schur complement $\mathbf{S}(\alpha) = \frac{1}{2\alpha} \tilde{\mathbf{Q}}_{22}(\alpha) - \mathbf{Q}_{12}(\alpha)^\top \mathbf{Q}_{11}^{-1} \mathbf{Q}_{12}(\alpha) = \mathbf{L}_{\mathbf{S}(\alpha)} \mathbf{L}_{\mathbf{S}(\alpha)}^\top$, so that we have

$$\mathbf{Q}_1 = \mathbf{L}_{\mathbf{Q}_1} \mathbf{L}_{\mathbf{Q}_1}^\top = \begin{bmatrix} \mathbf{L}_{\mathbf{Q}_{11}} & \mathbf{0} \\ \mathbf{Q}_{12}(\alpha)^\top \mathbf{L}_{\mathbf{Q}_{11}}^{-\top} & \mathbf{L}_{\mathbf{S}(\alpha)} \end{bmatrix} \begin{bmatrix} \mathbf{L}_{\mathbf{Q}_{11}}^\top & \mathbf{L}_{\mathbf{Q}_{11}}^{-1} \mathbf{Q}_{12}(\alpha) \\ \mathbf{0} & \mathbf{L}_{\mathbf{S}(\alpha)}^\top \end{bmatrix},$$

as follows from the results for the block-diagonal form of the Cholesky factorization (which can be verified by direct computation). To obtain the balancing transformation, we consider the approximation via singular value decomposition

$$\mathbf{L}_{\mathbf{Q}_1}^\top \mathbf{L}_{\mathbf{P}_1} = \begin{bmatrix} \mathbf{L}_{\mathbf{Q}_{11}}^\top & \mathbf{L}_{\mathbf{Q}_{11}}^{-1} \mathbf{Q}_{12}(\alpha) \\ \mathbf{0} & \mathbf{L}_{\mathbf{S}(\alpha)}^\top \end{bmatrix} \begin{bmatrix} \mathbf{L}_{\mathbf{P}_{11}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{L}_{\mathbf{Q}_{11}}^\top \mathbf{L}_{\mathbf{P}_{11}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \approx \begin{bmatrix} \mathcal{U}_r \\ \mathbf{0} \end{bmatrix} \mathcal{S}_r [\mathcal{V}_r^\top \ \mathbf{0}].$$

We then compute

$$\mathbf{V} = \begin{bmatrix} \mathbf{L}_{\mathbf{P}_{11}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathcal{V}_r \\ \mathbf{0} \end{bmatrix} \mathcal{S}_r^{-1/2} = \begin{bmatrix} \mathbf{L}_{\mathbf{P}_{11}} \mathcal{V}_r \mathcal{S}_r^{-1/2} \\ \mathbf{0} \end{bmatrix},$$

$$\mathbf{W} = \begin{bmatrix} \mathbf{L}_{\mathbf{Q}_{11}} & \mathbf{0} \\ \mathbf{Q}_{12}(\alpha)^\top \mathbf{L}_{\mathbf{Q}_{11}}^{-\top} & \mathbf{L}_{\mathbf{S}(\alpha)} \end{bmatrix} \begin{bmatrix} \mathcal{U}_r \\ \mathbf{0} \end{bmatrix} \mathcal{S}_r^{-1/2} = \begin{bmatrix} \mathbf{L}_{\mathbf{Q}_{11}} \mathcal{U}_r \mathcal{S}_r^{-1/2} \\ \mathbf{Q}_{12}(\alpha)^\top \mathbf{L}_{\mathbf{Q}_{11}}^{-\top} \mathcal{U}_r \mathcal{S}_r^{-1/2} \end{bmatrix},$$

which completes the proof. \square

The two propositions show that only Lyapunov equations in the original model dimensions are necessary (and not in the lifted dimensions), which is computationally appealing. For example, for the tubular reactor model in Sect. 3.2, the original model dimensions are $2n$ and the lifted dimensions are $7n$ and we only solve $2n$ -dimensional Lyapunov equations.

4.3 Balancing with Truncated (Quadratic) Gramians

The next two propositions consider the solution of the Lyapunov equations (3)–(4) for the truncated Gramians, which take into consideration the QB nature of the problem by incorporating \mathbf{H} and \mathbf{N} .

Proposition 4 *Let \mathbf{P}_1 be the solution from Proposition 2. The truncated Gramian from equation (3) for the QB system is given as*

$$\mathbf{P}_{\mathcal{T}} = \mathbf{P}_1 + \frac{1}{2\alpha} \begin{bmatrix} \tilde{\mathbf{P}}_{11}(\alpha) & \tilde{\mathbf{P}}_{12}(\alpha) \\ \tilde{\mathbf{P}}_{12}(\alpha)^\top & \tilde{\mathbf{P}}_{22}(\alpha) \end{bmatrix} \quad (30)$$

via the following equations:

$$\begin{aligned}\tilde{\mathbf{P}}_{22}(\alpha) &= \mathbf{H}_{21}(\alpha)(\mathbf{P}_{11} \otimes \mathbf{P}_{11})\mathbf{H}_{21}^\top(\alpha) + \mathbf{N}_{21}\mathbf{P}_{11}\mathbf{N}_{21}^\top, \\ \tilde{\mathbf{P}}_{12}(\alpha) &= -(\mathbf{A}_{11} - \alpha\mathbf{I})^{-1}\mathbf{A}_{12}\tilde{\mathbf{P}}_{22}(\alpha), \\ \mathbf{0} &= \mathbf{A}_{11}\tilde{\mathbf{P}}_{11}(\alpha) + \tilde{\mathbf{P}}_{11}(\alpha)\mathbf{A}_{11}^\top + [\mathbf{A}_{12}\tilde{\mathbf{P}}_{12}(\alpha)^\top + \tilde{\mathbf{P}}_{12}(\alpha)\mathbf{A}_{12}^\top].\end{aligned}$$

Proof The generalized controllability Lyapunov equation (3) is a linear equation, therefore we can decompose the truncated Gramian $\mathbf{P}_T = \mathbf{P}_1 + \mathbf{P}_2(\alpha)$, where \mathbf{P}_1 is the solution from Proposition 2. Therefore, $\mathbf{P}_2(\alpha)$ is the solution to

$$\mathbf{A}(\alpha)\mathbf{P}_2 + \mathbf{P}_2\mathbf{A}(\alpha)^\top + \mathbf{H}(\alpha)(\mathbf{P}_1 \otimes \mathbf{P}_1)\mathbf{H}(\alpha)^\top + \mathbf{N}\mathbf{P}_1\mathbf{N}^\top = \mathbf{0}.$$

With the matrices from above, we compute

$$\begin{aligned}\mathbf{N}\mathbf{P}_1\mathbf{N}^\top &= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{N}_{21} & \mathbf{N}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{P}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{0} & \mathbf{N}_{21}^\top \\ \mathbf{0} & \mathbf{N}_{22}^\top \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{N}_{21}\mathbf{P}_{11}\mathbf{N}_{21}^\top \end{bmatrix}, \\ \mathbf{H}(\alpha)(\mathbf{P}_1 \otimes \mathbf{P}_1)\mathbf{H}(\alpha)^\top &= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_{21}(\alpha)(\mathbf{P}_{11} \otimes \mathbf{P}_{11})\mathbf{H}_{21}(\alpha)^\top \end{bmatrix}.\end{aligned}$$

The symmetric positive semi-definite solution $\mathbf{P}_2(\alpha)$ needs to satisfy

$$\begin{aligned}\mathbf{0} &= \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & -\alpha\mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{P}}_{11} & \hat{\mathbf{P}}_{12} \\ \hat{\mathbf{P}}_{12}^\top & \hat{\mathbf{P}}_{22} \end{bmatrix} + \begin{bmatrix} \hat{\mathbf{P}}_{11} & \hat{\mathbf{P}}_{12} \\ \hat{\mathbf{P}}_{12}^\top & \hat{\mathbf{P}}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11}^\top & \mathbf{0} \\ \mathbf{A}_{12}^\top & -\alpha\mathbf{I} \end{bmatrix} \\ &+ \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_{21}(\alpha)(\mathbf{P}_{11} \otimes \mathbf{P}_{11})\mathbf{H}_{21}(\alpha)^\top \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{N}_{21}\mathbf{P}_{11}\mathbf{N}_{21}^\top \end{bmatrix},\end{aligned}$$

which yields the equations

$$\mathbf{A}_{11}\hat{\mathbf{P}}_{11} + \hat{\mathbf{P}}_{11}\mathbf{A}_{11}^\top + \mathbf{A}_{12}\hat{\mathbf{P}}_{12}^\top + \hat{\mathbf{P}}_{12}\mathbf{A}_{12}^\top = \mathbf{0}, \quad (31)$$

$$(\mathbf{A}_{11} - \alpha\mathbf{I})\hat{\mathbf{P}}_{12} + \mathbf{A}_{12}\hat{\mathbf{P}}_{22} = \mathbf{0}, \quad (32)$$

$$\hat{\mathbf{P}}_{12}^\top(\mathbf{A}_{11}^\top - \alpha\mathbf{I}) + \hat{\mathbf{P}}_{22}\mathbf{A}_{12}^\top = \mathbf{0}, \quad (33)$$

$$-2\alpha\hat{\mathbf{P}}_{22} + \mathbf{H}_{21}(\alpha)(\mathbf{P}_{11} \otimes \mathbf{P}_{11})\mathbf{H}_{21}(\alpha)^\top + \mathbf{N}_{21}\mathbf{P}_{11}\mathbf{N}_{21}^\top = \mathbf{0}. \quad (34)$$

We define $\tilde{\mathbf{P}}_{22}(\alpha) = [\mathbf{H}_{21}(\alpha)(\mathbf{P}_{11} \otimes \mathbf{P}_{11})\mathbf{H}_{21}(\alpha)^\top + \mathbf{N}_{21}\mathbf{P}_{11}\mathbf{N}_{21}^\top]$, so that $\hat{\mathbf{P}}_{22}(\alpha) = \frac{1}{2\alpha}\tilde{\mathbf{P}}_{22}(\alpha)$. Due to the symmetry of $\hat{\mathbf{P}}_{22}(\alpha)$, equations (32) and (33) are the same, and after substitution, we obtain $\hat{\mathbf{P}}_{12}(\alpha) = -\frac{1}{2\alpha}(\mathbf{A}_{11} - \alpha\mathbf{I})^{-1}\mathbf{A}_{12}\tilde{\mathbf{P}}_{22}(\alpha) := \frac{1}{2\alpha}\tilde{\mathbf{P}}_{12}(\alpha)$, with $\tilde{\mathbf{P}}_{12}(\alpha) := -(\mathbf{A}_{11} - \alpha\mathbf{I})^{-1}\mathbf{A}_{12}\tilde{\mathbf{P}}_{22}(\alpha)$. Substitution of $\hat{\mathbf{P}}_{12}(\alpha)$ into equation (31) yields the result. \square

Proposition 5 *Let the matrix \mathbf{Q}_1 be as in Proposition 2. The truncated Gramian \mathbf{Q}_T that solves equation (4) is given by*

$$\mathbf{Q}_T = \mathbf{Q}_1 + \begin{bmatrix} \hat{\mathbf{Q}}_{11}(\alpha) & \hat{\mathbf{Q}}_{12}(\alpha) \\ \hat{\mathbf{Q}}_{12}(\alpha)^\top & \hat{\mathbf{Q}}_{22}(\alpha) \end{bmatrix}, \quad (35)$$

where

$$\begin{aligned}\mathbf{0} &= \mathbf{A}_{11}^\top \widehat{\mathbf{Q}}_{11}(\alpha) + \widehat{\mathbf{Q}}_{11}(\alpha) \mathbf{A}_{11} + \widetilde{\mathbf{H}}_{11}(\alpha) + \frac{1}{2\alpha} \mathbf{N}_{21}^\top \widetilde{\mathbf{Q}}_{22}(\alpha) \mathbf{N}_{21}, \\ \widehat{\mathbf{Q}}_{12}(\alpha) &= -(\mathbf{A}_{11}^\top - \alpha \mathbf{I})^{-1} \left(\widetilde{\mathbf{H}}_{12}(\alpha) + \frac{1}{2\alpha} \mathbf{N}_{21}^\top \widetilde{\mathbf{Q}}_{22}(\alpha) \mathbf{N}_{22} + \widehat{\mathbf{Q}}_{11}(\alpha) \mathbf{A}_{12} \right), \\ \widehat{\mathbf{Q}}_{22}(\alpha) &= \frac{1}{2\alpha} \left[\mathbf{A}_{12}^\top \widehat{\mathbf{Q}}_{12}(\alpha) + \widehat{\mathbf{Q}}_{12}^\top(\alpha) \mathbf{A}_{12} + \widetilde{\mathbf{H}}_{22}(\alpha) + \frac{1}{2\alpha} \mathbf{N}_{22}^\top \widetilde{\mathbf{Q}}_{22}(\alpha) \mathbf{N}_{22} \right].\end{aligned}$$

Proof We begin by computing

$$\mathbf{N}^\top \mathbf{Q}_1 \mathbf{N} = \begin{bmatrix} \mathbf{0} & \mathbf{N}_{21}^\top \\ \mathbf{0} & \mathbf{N}_{22}^\top \end{bmatrix} \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12}(\alpha) \\ \mathbf{Q}_{12}^\top(\alpha) & \frac{1}{2\alpha} \widetilde{\mathbf{Q}}_{22}(\alpha) \end{bmatrix} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{N}_{21} & \mathbf{N}_{22} \end{bmatrix} = \frac{1}{2\alpha} \begin{bmatrix} \mathbf{N}_{21}^\top \widetilde{\mathbf{Q}}_{22} \mathbf{N}_{21} & \mathbf{N}_{21}^\top \widetilde{\mathbf{Q}}_{22} \mathbf{N}_{22} \\ \mathbf{N}_{22}^\top \widetilde{\mathbf{Q}}_{22} \mathbf{N}_{21} & \mathbf{N}_{22}^\top \widetilde{\mathbf{Q}}_{22} \mathbf{N}_{22} \end{bmatrix},$$

and for notational convenience partition the full matrix

$$\mathcal{H}^{(2)}(\alpha)(\mathbf{P}_1 \otimes \mathbf{Q}_1)(\mathcal{H}^{(2)}(\alpha))^\top =: \begin{bmatrix} \widetilde{\mathbf{H}}_{11}(\alpha) & \widetilde{\mathbf{H}}_{12}(\alpha) \\ \widetilde{\mathbf{H}}_{12}(\alpha)^\top & \widetilde{\mathbf{H}}_{22}(\alpha) \end{bmatrix},$$

which is a symmetric matrix since \mathbf{P}_1 and \mathbf{Q}_1 are symmetric Gramians. We then solve for the truncated Gramian via

$$\begin{aligned}\mathbf{0} &= \begin{bmatrix} \mathbf{A}_{11}^\top & \mathbf{0} \\ \mathbf{A}_{12}^\top & -\alpha \mathbf{I} \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{Q}}_{11} & \widehat{\mathbf{Q}}_{12} \\ \widehat{\mathbf{Q}}_{12}^\top & \widehat{\mathbf{Q}}_{22} \end{bmatrix} + \begin{bmatrix} \widehat{\mathbf{Q}}_{11} & \widehat{\mathbf{Q}}_{12} \\ \widehat{\mathbf{Q}}_{12}^\top & \widehat{\mathbf{Q}}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & -\alpha \mathbf{I} \end{bmatrix} \\ &+ \begin{bmatrix} \widetilde{\mathbf{H}}_{11}(\alpha) & \widetilde{\mathbf{H}}_{12}(\alpha) \\ \widetilde{\mathbf{H}}_{12}(\alpha)^\top & \widetilde{\mathbf{H}}_{22}(\alpha) \end{bmatrix} + \frac{1}{2\alpha} \begin{bmatrix} \mathbf{N}_{21}^\top \widetilde{\mathbf{Q}}_{22} \mathbf{N}_{21} & \mathbf{N}_{21}^\top \widetilde{\mathbf{Q}}_{22} \mathbf{N}_{22} \\ \mathbf{N}_{22}^\top \widetilde{\mathbf{Q}}_{22} \mathbf{N}_{21} & \mathbf{N}_{22}^\top \widetilde{\mathbf{Q}}_{22} \mathbf{N}_{22} \end{bmatrix}.\end{aligned}$$

Thus, the Gramian solution has to satisfy the equations

$$\begin{aligned}\mathbf{0} &= \mathbf{A}_{11}^\top \widehat{\mathbf{Q}}_{11} + \widehat{\mathbf{Q}}_{11} \mathbf{A}_{11} + \widetilde{\mathbf{H}}_{11}(\alpha) + \frac{1}{2\alpha} \mathbf{N}_{21}^\top \widetilde{\mathbf{Q}}_{22} \mathbf{N}_{21}, \\ \mathbf{0} &= (\mathbf{A}_{11}^\top - \alpha \mathbf{I}) \widehat{\mathbf{Q}}_{12} + \widehat{\mathbf{Q}}_{11} \mathbf{A}_{12} + \widetilde{\mathbf{H}}_{12}(\alpha) + \frac{1}{2\alpha} \mathbf{N}_{21}^\top \widetilde{\mathbf{Q}}_{22} \mathbf{N}_{22}, \\ \mathbf{0} &= \mathbf{A}_{12}^\top \widehat{\mathbf{Q}}_{12} - 2\alpha \widehat{\mathbf{Q}}_{22} + \widehat{\mathbf{Q}}_{12}^\top \mathbf{A}_{12} + \widetilde{\mathbf{H}}_{22}(\alpha) + \frac{1}{2\alpha} \mathbf{N}_{22}^\top \widetilde{\mathbf{Q}}_{22} \mathbf{N}_{22},\end{aligned}$$

from which we get the stated result. \square

By decomposing the Lyapunov solution into block form, we can again compute the truncated Gramians by only working in the original dimensions (not the lifted dimensions). Due to the dependence on $1/\alpha$ the Gramians \mathbf{Q}_T , \mathbf{P}_T are not defined for $\alpha \rightarrow 0$ and taking the limit $\alpha \rightarrow 0$ can therefore result in numerical difficulties. We summarize our proposed approach below.

Proposition 6 *An approximate balancing transformation for the QB structure arising in lifting of nonlinear systems can be computed as follows:*

1. Choose $\alpha \in (0, \infty)$ and compute $\mathbf{P}_T, \mathbf{Q}_T$ from Propositions 4 and 5 with $\mathbf{A}(\alpha)$.
2. Compute Cholesky factors $\mathbf{P}_T = \mathbf{L}_{\mathbf{P}_T} \mathbf{L}_{\mathbf{P}_T}^\top$ and $\mathbf{Q}_T = \mathbf{L}_{\mathbf{Q}_T} \mathbf{L}_{\mathbf{Q}_T}^\top$.
3. Compute SVD $\mathbf{L}_{\mathbf{Q}_T}^\top \mathbf{L}_{\mathbf{P}_T} = \mathcal{U} \mathcal{S} \mathcal{V}^\top$; denote rank r approximation as $\mathcal{U}_r \mathcal{S}_r \mathcal{V}_r^\top$.
4. Projection matrices $\mathbf{W} = \mathbf{L}_{\mathbf{Q}_T} \mathcal{U}_r \mathcal{S}_r^{-1/2}$ and $\mathbf{V} = \mathbf{L}_{\mathbf{P}_T} \mathcal{V}_r \mathcal{S}_r^{-1/2}$.
5. Matrices for QB-ROM (6)–(7) are

$$\widehat{\mathbf{A}} = \mathbf{W}^\top \mathbf{A} \mathbf{V}, \quad \widehat{\mathbf{B}} = \mathbf{W}^\top \mathbf{B}, \quad \widehat{\mathbf{C}} = \mathbf{C} \mathbf{V}, \quad \widehat{\mathbf{N}}_k = \mathbf{W}^\top \mathbf{N}_k \mathbf{V}, \quad \widehat{\mathbf{H}} = \mathbf{W}^\top \mathbf{H} (\mathbf{V} \otimes \mathbf{V}).$$

5 Numerical Results

For the tubular reactor model in Sect. 3.2, the parameters are $\mathcal{D} = 0.17$, $Pe = 25$, $\mathcal{B} = 0.5$, $\beta = 2.5$, $\theta_{\text{ref}} \equiv 1$, $\gamma = 5$, and $n = 199$. We compare the approximate balanced truncation ROM, denoted QB-BT, from Proposition 6 with a ROM computed from POD with discrete empirical interpolation [9], denoted as POD-DEIM. For the predictive comparison, the full-order model and ROMs are simulated until $t_f = 30s$ with different inputs $u(t)$ and initial conditions $\mathbf{x}(0) = \mathbf{x}_0$ as given below.

The POD-DEIM models are obtained as follows: We simulate the full-order model until $t_{\text{train}} = 15s$ with training input $u_{\text{train}}(t)$ and training initial condition $\mathbf{x}_{0,\text{train}}$. We record a snapshot every 0.01s and store the snapshots in a matrix \mathbf{X} , compute the POD basis $\mathbf{V} \in \mathbb{R}^{n \times r}$ of dimension r , and project the system matrices in (17)–(18) onto \mathbf{V} . For every r -dimensional POD-DEIM ROM we use r DEIM interpolation points using the QDEIM algorithm from [10].

For the QB-BT models, we discussed in Sect. 4.1 that the parameter $\alpha \in (0, \infty)$. Here, we found that $\alpha = 20$ resulted in the most accurate balanced ROMs and that for $\alpha \ll 1$ the Gramians' subspaces would lead to unstable ROMs. This could be related to the influence of the field of values of $\mathbf{A}(\alpha)$ on the decay of the singular values of the solution to the Lyapunov equation [3]. Here, for $\alpha > \alpha_c = 2.6$, the field of values of $\mathbf{A}(\alpha)$ is in the open left-half plane, which we can choose as a selection criterion for α .

Reduced-order models constructed via POD-DEIM depend on the snapshot data that is obtained from training simulations. In contrast, balancing transformations only depend on the matrices defining the full-order QB system $\mathbf{A}, \mathbf{H}, \mathbf{N}, \mathbf{B}, \mathbf{C}$; the basis generation does not require a choice of training data. To illustrate this point, we consider four test cases with different testing conditions and for which the POD-DEIM training choices vary. For each test case we compare the output error $\sum_{i=1}^s |y(t_i) - y_r(t_i)|$ with $s = 3,000$ for the POD-DEIM and QB-BT ROMs. We detail each case below.

Case 1: $u(t) = \cos(t)$, $u_{\text{train}}(t) = u(t)$, $\mathbf{x}_{0,\text{train}} = \mathbf{x}_0$. For this test case, training and testing conditions are the same, so we expect POD-DEIM ROMs to accurately reproduce the full-order model (FOM) simulations. The output quantity of interest of the FOM compared to two ROMs with $r = 20$ basis functions for this test case is

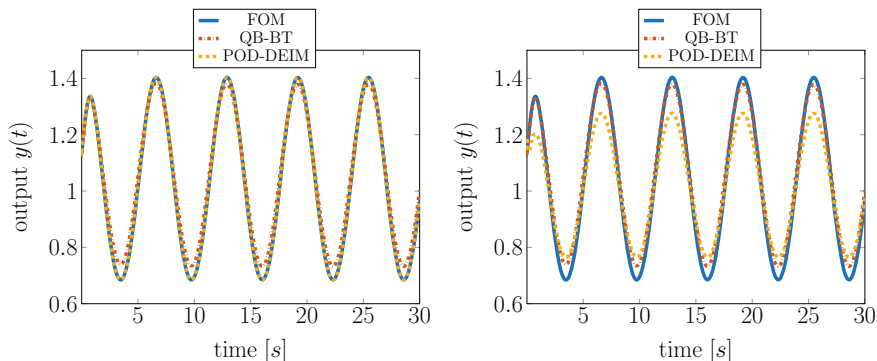


Fig. 1 Model outputs for test Case 1 (left) and Case 2 (right): Comparison of FOM with two ROMs for $r = 20$ basis functions, respectively

Table 1 Case 1: Output error $\sum_{i=1}^s |y(t_i) - y_r(t_i)|$ in ROMs

| ROM | $r = 4$ | $r = 6$ | $r = 8$ | $r = 10$ | $r = 12$ | $r = 14$ | $r = 16$ | $r = 18$ | $r = 20$ |
|----------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| QB-BT | $6.51 \times$ E-04 | $7.00 \times$ E-04 | $6.98 \times$ E-04 | $7.09 \times$ E-04 | $7.06 \times$ E-04 | $6.84 \times$ E-04 | $6.66 \times$ E-04 | $7.05 \times$ E-04 | $6.95 \times$ E-04 |
| POD-DEIM | $2.53 \times$ E-05 | $1.17 \times$ E-05 | $7.39 \times$ E-06 | $6.93 \times$ E-06 | $5.61 \times$ E-06 | $5.62 \times$ E-06 | $5.62 \times$ E-06 | $5.63 \times$ E-06 | $5.62 \times$ E-06 |

Table 2 Case 2: Output error $\sum_{i=1}^s |y(t_i) - y_r(t_i)|$ in ROMs

| ROM | $r = 4$ | $r = 6$ | $r = 8$ | $r = 10$ | $r = 12$ | $r = 14$ | $r = 16$ | $r = 18$ | $r = 20$ |
|----------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| QB-BT | $6.51 \times$ E-04 | $7.00 \times$ E-04 | $6.98 \times$ E-04 | $7.09 \times$ E-04 | $7.06 \times$ E-04 | $6.84 \times$ E-04 | $6.66 \times$ E-04 | $7.05 \times$ E-04 | $6.95 \times$ E-04 |
| POD-DEIM | $1.47 \times$ E-03 | $1.41 \times$ E-03 | $1.40 \times$ E-03 | $1.41 \times$ E-03 | $1.39 \times$ E-03 | $1.40 \times$ E-03 | $1.40 \times$ E-03 | $1.40 \times$ E-03 | $1.40 \times$ E-03 |

shown in Fig. 1, left. Table 1 shows the output error for various ROM model sizes. We see that the POD-DEIM ROM outperforms the QB-BT ROM in this case.

Case 2: $u(t) = \cos(t)$, $u_{\text{train}}(t) = u(t)$, $\mathbf{x}_{0,\text{train}} = \mathbf{x}_0$; noise = 10% Here we use the same training input and training initial conditions as in Case 1. However, we add noise to the training data, reflecting a situation that practitioners often face where measurements are noise-corrupted. In particular, we use $\tilde{X} = X + 0.1X(-1 + 2\Xi)$ where $\Xi = \mathcal{N}(0, 1)$ is a normally distributed random variable. Figure 1, right, shows the model output for the FOM and both ROMs for $r = 20$ where we see that the POD-DEIM model does not capture the output amplitude well. Table 2 shows the output errors for increasing ROM sizes. We observe that the QB-BT ROM outperforms the POD-DEIM model in this case, as the POD-DEIM model suffers from a two orders of magnitude loss in accuracy (compared to Table 1) due to noisy snapshots.

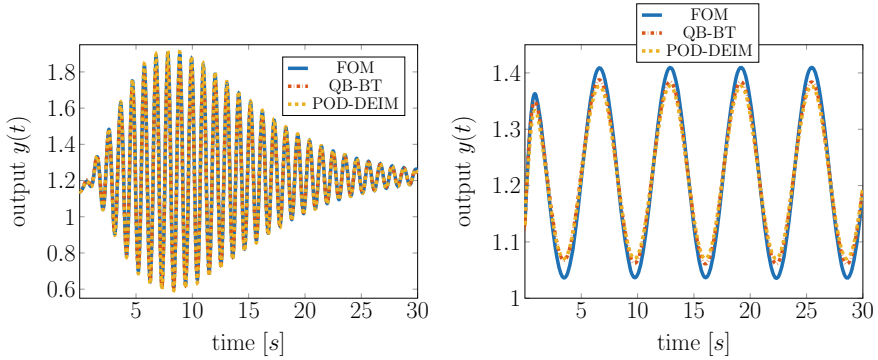


Fig. 2 Model outputs for Case 3 (left) and Case 4 (right): Comparison of FOM with two ROMs for $r = 20$ basis functions, respectively

Table 3 Case 3: Output error $\sum_{i=1}^s |y(t_i) - y_r(t_i)|$ in ROMs

| ROM | $r = 4$ | $r = 6$ | $r = 8$ | $r = 10$ | $r = 12$ | $r = 14$ | $r = 16$ | $r = 18$ | $r = 20$ |
|----------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| QB-BT | $5.30 \times$ E-04 | $6.02 \times$ E-04 | $5.41 \times$ E-04 | $4.96 \times$ E-04 | $5.03 \times$ E-04 | $7.18 \times$ E-04 | $6.48 \times$ E-04 | $5.63 \times$ E-04 | $5.37 \times$ E-04 |
| POD-DEIM | $2.42 \times$ E-03 | $2.22 \times$ E-03 | $2.15 \times$ E-03 | $2.15 \times$ E-03 | $1.96 \times$ E-03 | $9.11 \times$ E-04 | $6.55 \times$ E-04 | $6.16 \times$ E-04 | $2.37 \times$ E-04 |

Case 3: $u(t) = 0.5(1 + t^2 \exp(-t/4) \sin(6t))$, $u_{\text{train}}(t) = 0.5$, $\mathbf{x}_{0,\text{train}} = \mathbf{x}_0$ This case illustrates a scenario where a practitioner trained a model assuming that input conditions are operating at a constant equilibrium. However during operation of the plant the inputs indeed vary but the oscillations are damped and revert to the equilibrium position of 0.5. Figure 2, left, shows the model output for the FOM and both ROMs for $r = 20$; both ROMs are reproducing the output well. Table 3 shows the output errors for increasing ROM sizes. The QB-BT ROM is again more accurate, yet at one model order, $r = 20$, the POD-DEIM model is more accurate.

Case 4: $u(t) = \cos(t)$, $u_{\text{train}}(t) = 0.5$, $\mathbf{x}_{0,\text{train}}(1:n) = 0$, $\mathbf{x}_{0,\text{train}}(n+1:2n) = 1$. Here, the training initial condition is different from the initial condition used for prediction, i.e., $\mathbf{x}_{0,\text{train}} \neq \mathbf{x}_0$. Figure 2, right, shows the model output for the FOM and both ROMs for $r = 20$; the QB-BT model slightly outperforms the POD-DEIM model, but both fall short of predicting the full amplitude. Table 4 shows the output errors for increasing ROM sizes, where the QB-BT model again seems to outperform the POD-DEIM model for all basis sizes.

Table 4 Case 4: Output error $\sum_{i=1}^s |y(t_i) - y_r(t_i)|$ in ROMs

| ROM | $r = 4$ | $r = 6$ | $r = 8$ | $r = 10$ | $r = 12$ | $r = 14$ | $r = 16$ | $r = 18$ | $r = 20$ |
|----------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| QB-BT | $3.46 \times$ E-04 | $3.47 \times$ E-04 | $3.77 \times$ E-04 | $3.76 \times$ E-04 | $3.75 \times$ E-04 | $3.71 \times$ E-04 | $3.41 \times$ E-04 | $3.76 \times$ E-04 | $3.66 \times$ E-04 |
| POD-DEIM | $8.86 \times$ E-04 | $9.23 \times$ E-04 | $9.26 \times$ E-04 | $9.25 \times$ E-04 | $9.29 \times$ E-04 | $7.05 \times$ E-04 | $4.51 \times$ E-04 | $4.51 \times$ E-04 | $4.38 \times$ E-04 |

6 Conclusions

We presented a balanced truncation model reduction approach that is applicable to a large class of nonlinear systems with time-varying and uncertain inputs. As a first step, our approach lifts the nonlinear system to quadratic-bilinear form via the introduction of auxiliary variables. As lifted systems often have system matrices with zero eigenvalues, we first introduced an artificial stabilization parameter and then derived a balancing algorithm for those lifted quadratic-bilinear systems that only requires expensive matrix computations in the original—and not in the lifted—dimension. The method was illustrated by the model reduction problem for a tubular reactor model, for which we derived the multi-stage lifting transformations and performed balanced model reduction. The numerical results showed that through our proposed method, we can obtain ROMs for nonlinear systems that are superior to POD-DEIM models in situations where a good choice of training data is not feasible.

Acknowledgements This work has been supported in part by the Air Force Center of Excellence on Multi-Fidelity Modeling of Rocket Combustor Dynamics under award FA9550-17-1-0195, by the Air Force Office of Scientific Research (AFOSR) MURI on managing multiple information sources of multi-physics systems, Award Numbers FA9550-15-1-0038 and FA9550-18-1-0023, and the AEOLUS MMICC center under the Department of Energy grant DE-SC0019303. The authors thank the reviewer for the very valuable comments which helped us improve our previous stabilization approach in Sect. 4.1. The authors thank Dr. Pawan Goyal for valuable discussions about the truncated Gramian framework. The authors thank Prof. Mark Embree for insightful feedback on the manuscript, and pointing out the connection to the field of values and the choice of α in our algorithm.

References

1. Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems. Advances in Design and Control. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (2005)
2. Antoulas, A.C., Beattie, C.A., Gugercin, S.: Interpolatory model reduction of large-scale dynamical systems. In: Efficient modeling and control of large-scale systems, pp. 3–58. Springer (2010)
3. Baker, J., Embree, M., Sabino, J.: Fast singular value decay for Lyapunov solutions with nonnormal coefficients. SIAM J. Matrix Anal. Appl. **36**(2), 656–668 (2015)
4. Benner, P., Breiten, T.: Two-sided projection methods for nonlinear model order reduction. SIAM J. Sci. Comput. **37**(2), B239–B260 (2015)

5. Benner, P., Goyal, P.: Balanced truncation model order reduction for quadratic-bilinear control systems (2017). [arXiv:1705.00160](https://arxiv.org/abs/1705.00160)
6. Benner, P., Goyal, P., Gugercin, S.: H2-quasi-optimal model order reduction for quadratic-bilinear control systems. *SIAM J. Matrix Anal. Appl.* **39**(2), 983–1032 (2018)
7. Bouvrie, J., Hamzi, B.: Kernel methods for the approximation of nonlinear systems. *SIAM J. Control Optim.* **55**(4), 2460–2492 (2017)
8. Cao, X., Maubach, J., Weiland, S., Schilders, W.: A novel Krylov method for model order reduction of quadratic bilinear systems. In: 2018 IEEE Conference on Decision and Control (CDC), pp. 3217–3222. IEEE (2018)
9. Chaturantabut, S., Sorensen, D.: Nonlinear model reduction via discrete empirical interpolation. *SIAM J. Sci. Comput.* **32**(5), 2737–2764 (2010)
10. Drmac, Z., Gugercin, S.: A new selection operator for the discrete empirical interpolation method—Improved a priori error bound and extensions. *SIAM J. Sci. Comput.* **38**(2), A631–A648 (2016)
11. Flinois, T.L., Morgans, A.S., Schmid, P.J.: Projection-free approximate balanced truncation of large unstable systems. *Phys. Rev. E* **92**(2), 023012 (2015)
12. Fujimoto, K., Scherpen, J.M.: Balanced realization and model order reduction for nonlinear systems based on singular value analysis. *SIAM J. Control Optim.* **48**(7), 4591–4623 (2010)
13. Gosea, I.V., Antoulas, A.C.: Data-driven model order reduction of quadratic-bilinear systems. *Numer. Linear Algebra Appl.* **25**(6), e2200 (2018)
14. Gray, W.S., Verriest, E.L.: Algebraically defined Gramians for nonlinear systems. In: 2006 45th IEEE Conference on Decision and Control, pp. 3730–3735. IEEE (2006)
15. Gu, C.: QLMOR: a projection-based nonlinear model order reduction approach using quadratic-linear representation of nonlinear systems. *IEEE Trans. Comput-Aided Des. Integr. Circuits Syst.* **30**(9), 1307–1320 (2011)
16. Gugercin, S., Antoulas, A.C.: A survey of model reduction by balanced truncation and some new results. *Int. J. Control* **77**(8), 748–766 (2004)
17. Heinemann, R.F., Poore, A.B.: Multiplicity, stability, and oscillatory dynamics of the tubular reactor. *Chem. Eng. Sci.* **36**(8), 1411–1419 (1981)
18. Holmes, P., Lumley, J.L., Berkooz, G.: *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*. Cambridge Monographs on Mechanics. Cambridge University Press (1996)
19. Kerner, E.H.: Universal formats for nonlinear ordinary differential systems. *J. Math. Phys.* **22**(7), 1366–1371 (1981)
20. Kramer, B., Willcox, K.: Nonlinear model order reduction via lifting transformations and proper orthogonal decomposition. *AIAA J.* **57**(6), 2297–2307 (2019)
21. Lall, S., Marsden, J.E., Glavaški, S.: A subspace approach to balanced truncation for model reduction of nonlinear control systems. *Int. J. Robust Nonlinear Control: IFAC-Affil. J.* **12**(6), 519–535 (2002)
22. Liu, J., Zhan, N., Zhao, H., Zou, L.: Abstraction of elementary hybrid systems by variable transformation. In: *International Symposium on Formal Methods*, pp. 360–377. Springer (2015)
23. McCormick, G.P.: Computability of global solutions to factorable nonconvex programs: part I—convex underestimating problems. *Math. Program.* **10**(1), 147–175 (1976)
24. Moore, B.: Principal component analysis in linear systems: controllability, observability, and model reduction. *IEEE Trans. Autom. Control* **26**(1), 17–32 (1981)
25. Pulch, R., Narayan, A.: Balanced truncation for model order reduction of linear dynamical systems with quadratic outputs. *SIAM J. Sci. Comput.* **41**(4), A2270–A2295 (2019)
26. Savageau, M.A., Voit, E.O.: Recasting nonlinear differential equations as S-systems: a canonical nonlinear form. *Math. Biosci.* **87**(1), 83–115 (1987)
27. Scherpen, J.M.: Balancing for nonlinear systems. *Syst. Control Lett.* **21**(2), 143–153 (1993)
28. Swischuk, R., Kramer, B., Huang, C., Willcox, K.: Learning physics-based reduced-order models for a single-injector combustion process. *AIAA J.* **58**(6), 2658–2672 (2020)
29. Willcox, K., Peraire, J.: Balanced model reduction via the proper orthogonal decomposition. *AIAA J.* **40**(11), 2323–2330 (2002)
30. Zhou, Y.B.: *Model Reduction for Nonlinear Dynamical Systems with Parametric Uncertainties*. Ph.D. thesis, Massachusetts Institute of Technology (2012)

Modeling the Buck Converter from Measurements of Its Harmonic Transfer Function



Sanda Lefteriu

It is an honor to serve as one of the Editors of this Festschrift celebrating the 70th anniversary of Thanos Antoulas. His guidance during my time as his PhD student was invaluable and helped shape my academic career.

Abstract Buck converters are DC-DC converters which provide a step-down of the input voltage produced by the supply to the output voltage delivered to the load. They contain two semiconductor devices (typically a transistor and a diode) acting as switches which open and close periodically, together with an output filter in the form of an RLC circuit. These considerations render the buck converter as a periodically switched linear system. Thanks to the periodic nature, the Harmonic Transfer Function (HTF) is a suitable tool for modeling the buck converter. This chapter shows how to identify a frequency domain model for the buck converter from measurements of the HTF through the Loewner framework. The measurements are obtained by computing the Fast Fourier Transform (FFT) of the outputs in a small signal analysis. The model corresponds to a single-input multiple-output system with the the number of outputs given by the desired number of harmonics.

Keywords Periodically switched linear system · Harmonic transfer function · Loewner framework

1 Introduction

Power converters are electronic circuits that allow the electric power source (a battery, the electrical network, a solar panel, etc.) to be adapted to the needs of the receiver (an electric motor, an asynchronous machine, etc.). One way of classifying converters is

S. Lefteriu (✉)

Center for Digital Systems, IMT Lille Douai, Inisut Mines-Télécom, 59000 Lille, France
e-mail: s.lefteriu@gmail.com

© Springer Nature Switzerland AG 2022

C. Beattie et al. (eds.), *Realization and Model Reduction of Dynamical Systems*,
https://doi.org/10.1007/978-3-030-95157-3_10

175

Fig. 1 Circuit diagram for the Buck converter

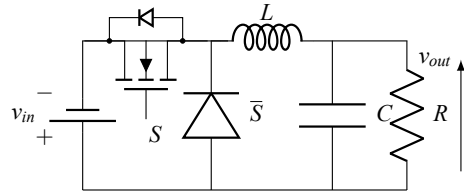
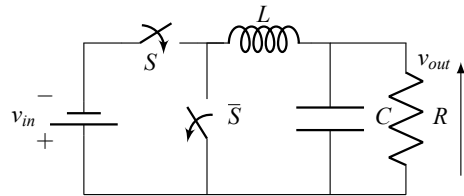


Fig. 2 The Buck converter as a switched system

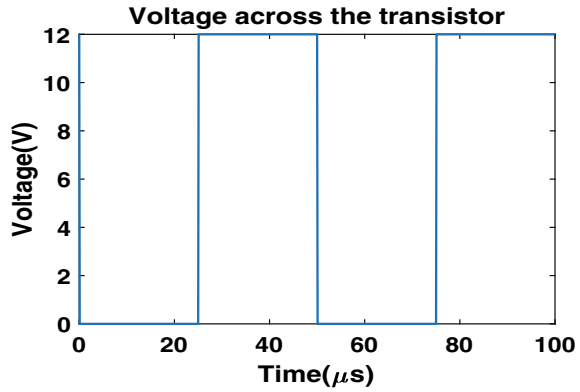


with respect to the type of current accepted by the supply and the load: alternating current (AC) or direct current (DC). Hence, there are four possible combinations: DC to DC, DC to AC, AC to DC and AC to AC converters. DC-DC converters can be found in mobile phones, laptops, wind turbines, photovoltaic systems, electric cars, etc.

This chapter focuses on a type of DC-DC converters that are the easiest to model, namely buck converters. The circuit diagram of an open-loop buck converter is shown in Fig. 1. In continuous current mode (CCM), the transistor S is open and closed with a certain switching frequency, while the diode \bar{S} follows the opposite behavior (Fig. 2). Figure 3 shows the voltage drop over the transistor when assuming ideal switching (in reality, switching is not an instantaneous process). The switching frequency is much larger than the resonant frequency of the RLC filter to ensure correct functioning of the converter. The presence of the two semiconductor devices render the converter as a switched system. On the other hand, the periodic behavior of the two switches dictates the periodicity of the overall system. The output filter is a linear RLC circuit, hence, the buck converter can be regarded as a periodic switched linear (PSL) system.

The goal of this chapter is to model the buck converter in the frequency domain from physically realizable measurements. Using the small signal paradigm, an AC voltage of small amplitude and varying frequency is added to the DC input. An FFT analysis of the output reveals that, besides the perturbation frequency, other frequencies are also present (typically referred to as harmonic distortions [1]). These frequencies are predicted by the Harmonic Transfer Function (HTF), which was developed as a tool for analyzing periodic systems in the frequency domain. The HTF is expressed in terms of doubly-infinite state-space matrices containing the Fourier coefficients of the periodic state-space matrices. Based on the application, one decides on the number of relevant harmonics and can, thus, truncate the infinite matrices to only a few terms. From the spectral analysis of the response to a range of sinusoidal perturbations, measurements at the expected peaks are gathered, which represent, after post-processing, measurements of the HTFs. These can be employed,

Fig. 3 Ideal switching on two switching periods



together with the Loewner framework [2], to identify the continuous-form of these few HTFs.

Typically, the buck converter is characterized by the average model [3]. This is a linear time-invariant model describing the dynamics of the system in terms of the average values for the current and voltage variables, neglecting effects due to switching. Hence, a plethora of works focus on the identification of these low order LTI average models, an example being [4]. Periodic switching causes the appearance of higher order harmonics and aliasing effects appear for perturbation frequencies larger than half of the switching frequency. Hence, the average model is valid only until one half of the switching frequency, before aliasing becomes an issue. High frequency harmonics are, therefore, not captured by the average model. Figure 4 shows the response to a DC input of the ideal buck converter detailed in Sect. 4.3. This response resembles that of a second-order system and is precisely the response predicted by the average model. However, when considering a time scale in the same order of magnitude as the switching frequency (Figs. 5 and 6), oscillations with the

Fig. 4 Response of the Buck converter

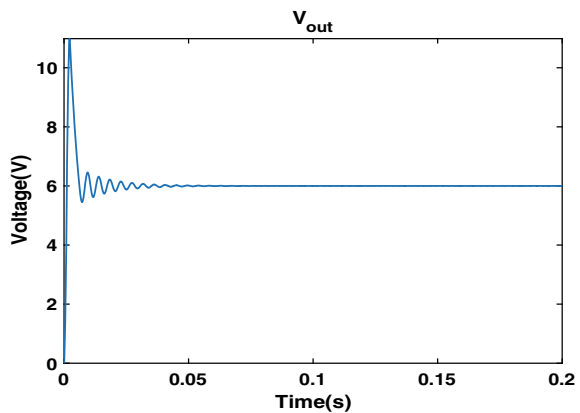


Fig. 5 Response between 0.0298 s and 0.03 s

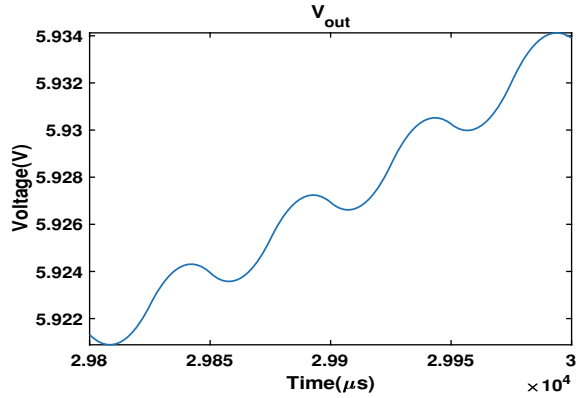
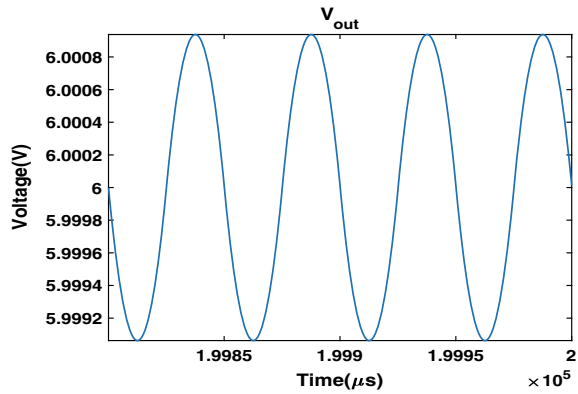


Fig. 6 Response between 0.1998 s and 0.2 s



same frequency as the switching frequency can be detected. These oscillations cannot be captured by the average model.

To account for harmonics induced by the periodic behavior, the concept of Harmonic Transfer Functions (HTFs) was developed in [5], with one transfer function modeling each harmonic's response. For an application consisting of a helicopter rotor, [6] showed that the theoretical HTFs match the quantities obtained from processing the Fourier transform of outputs and inputs. The work of [7] was the first to apply the harmonic state-space (HSS) framework from [5] to model a number of electrical systems from power electronics. In this framework, state-space matrices are doubly-infinite with entries given by the Fourier coefficients of the periodic state-space matrices. A common technique is to lift the system through the application of a Floquet transformation. This is a periodic similarity transformation which yields a constant system matrix \mathbf{A} , thus facilitating the stability analysis of such systems. An identification algorithm for linear time-periodic (LTP) systems using Floquet transformations applied to the discretized LTP system is presented in [8]. The paper

[9] provides a review of frequency-domain modeling of converters, including Floquet theory, Harmonic Transfer Functions and Equivalent Signal theory. Last, but not least, [10] assumes full state measurements and identifies the HSS matrices from input-output data.

The Loewner framework is typically employed for identifying a descriptor-form representation of linear time invariant (LTI) systems [2, 11] from frequency-domain measurements of their transfer function. Authors in [12] extended it to the class of linear switched systems. In particular, the data used in [12] consist of frequency domain samples of input-output mappings in the form of a series of multivariate rational functions obtained by taking the multivariate Laplace transform. However, it is not clear how to measure these quantities in practice.

This paper is structured as follows. Section 2 first recalls the problem of modeling LTI systems in the frequency domain in Sect. 2.1 and, afterwards, reviews the concept of harmonic transfer function for modeling periodic systems in Sect. 2.2. Section 3 provides a short review of the classical Loewner framework. Section 4 shows the application of the proposed methodology to an ideal buck converter, as well as when integrating the ON resistance of the switches into the model, together with the numerical experiments. Finally, Sect. 5 concludes the paper and presents directions for future research.

2 Modeling Periodic Systems in the Frequency Domain

This section discusses modeling dynamical systems in the frequency domain. First, the case of linear time-invariant systems is quickly reviewed. Afterwards, the theory of harmonic transfer functions is presented for periodic switched systems.

2.1 Linear Time-Invariant Systems in the Frequency Domain

For simplicity, we review the case of single-input single-output (SISO) systems. We seek to recover a descriptor realization consisting of matrices $\mathbf{E} \in \mathbb{R}^{n \times n}$, $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{b} \in \mathbb{R}^{n \times 1}$, $\mathbf{c} \in \mathbb{R}^{1 \times n}$ for an underlying linear time invariant (LTI) system of the form

$$\mathbf{E}\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}u(t) \quad (1)$$

$$y(t) = \mathbf{c}\mathbf{x}(t). \quad (2)$$

The excitation is $u(t)$, while the response is denoted by $y(t)$. The column vector $\mathbf{x}(t) \in \mathbb{R}^n [t]$ is an internal variable referred to as the state if the matrix \mathbf{E} is invertible. Equations (1)–(2) provide the time-domain representation, while the frequency-domain representation is given in terms of the transfer function obtained by taking the Laplace transform of (1)–(2) and computing the ratio between the Laplace transform

of the output to that of the input. The transfer function is

$$\mathbf{H}(s) = \frac{\mathbf{Y}(s)}{\mathbf{U}(s)} = \mathbf{c}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{b}, \quad (3)$$

where s is the Laplace variable, a complex number.

Remark 1 The descriptor-form representation (1)–(2) of an LTI system is more general than the state-space form

$$\dot{\mathbf{x}}(t) = \mathbf{A}_{ss}\mathbf{x}(t) + \mathbf{b}_{ss}u(t) \quad (4)$$

$$y(t) = \mathbf{c}_{ss}\mathbf{x}(t) + d_{ss}u(t), \quad (5)$$

as it can describe systems with a constant d_{ss} , as well as higher-order polynomial terms, by incorporating these into \mathbf{b} or \mathbf{c} and making \mathbf{E} singular. If the matrix \mathbf{E} in (1) is invertible, then $\mathbf{A}_{ss} = \mathbf{E}^{-1}\mathbf{A}$ and $\mathbf{b}_{ss} = \mathbf{E}^{-1}\mathbf{b}$.

2.1.1 Review of Modeling LTI Systems from Frequency-Domain Measurements

Frequency-domain measurements are obtained by considering the input $u(t)$ as a sinusoidal excitation $u(t) = \cos(2\pi f_j t)$ for a range of perturbation frequencies f_j , $j = 1, \dots, P$. At steady state, the output is also a sinusoidal signal with the same frequency, but a different magnitude and phase: $y_{ss}(t) = a \cos(2\pi f_j t + \phi)$. The magnitude and phase-shift encode a complex number $H_j = a \exp^{i\phi}$, where i is the unit imaginary number. The quantity H_j represents the measurement at frequency f_j , so, by collecting these pairs of points (f_j, H_j) , $j = 1, \dots, P$, the goal is to recover the unknown transfer function such that, when evaluating $H(s)$ for $s = i2\pi f_j$, the result is close to the corresponding frequency response measurement H_j . In other words, given the discrete points (f_j, H_j) , $j = 1, \dots, P$, we wish to find the continuous-time rational transfer function $H(s)$ such that the points are interpolated: $H_j = H(i2\pi f_j)$, or well approximated: $H_j \approx H(i2\pi f_j)$ (depending on the application).

2.2 Periodic Systems in the Frequency Domain

Due to the periodic opening and closing of the switches, the converter's dynamics can be expressed as a periodic system in state-space form:

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{b}(t)v_{in}(t) \quad (6)$$

$$v_{out}(t) = \mathbf{c}(t)\mathbf{x}(t) + d(t)v_{in}(t) \quad (7)$$

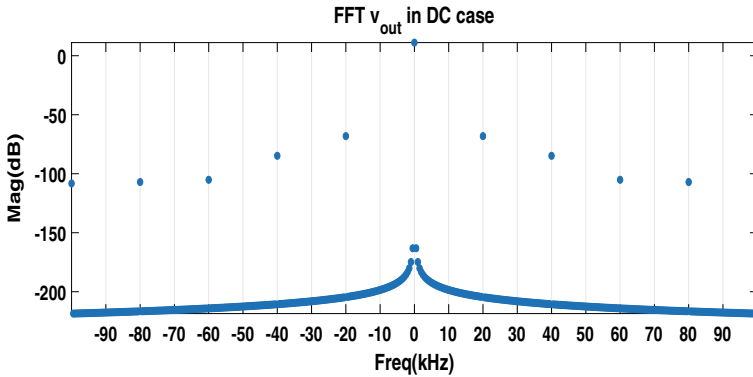


Fig. 7 Fourier analysis of the buck converter’s response to a DC input

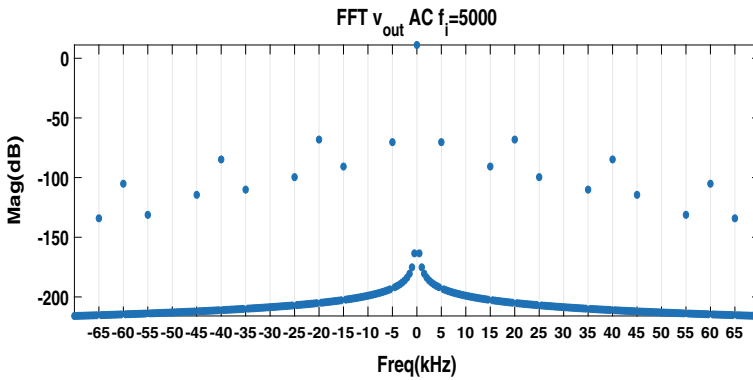


Fig. 8 Fourier analysis of the buck converter’s response to a sum of DC and 5 kHz AC inputs (26)

where the state-space quantities $\mathbf{A}(t)$, $\mathbf{b}(t)$, $\mathbf{c}(t)$, $d(t)$ are periodic, with a fixed and known period $T_s > 0$, called the switching period, such that $\mathbf{A}(t) = \mathbf{A}(t + mT_s)$, $\forall m \in \mathbb{Z}$, and the same for $\mathbf{b}(t)$, $\mathbf{c}(t)$, $d(t)$.

For a linear periodically time-varying (LPTV) system, a sinusoidal input with excitation frequency f_j produces an infinite sum of sinusoidal signals with different magnitudes and phase-shifts, at frequencies that are the sum of f_j and multiples of the switching frequency f_s : $f_i + mf_s, \forall m \in \mathbb{Z}$. This can be visualized in Fig. 7 from the Fourier analysis of the converter’s response to a DC input and in Fig. 8 for an AC input with perturbation frequency $f_i = 5 \text{ kHz}$ superimposed to a DC signal. The converter is operating with switching frequency $f_s = 20 \text{ kHz}$.

The periodic switched system changes between different discrete modes with period $T_s = \frac{1}{f_s}$, the switching period. For simplicity, and because it fits the case of the buck converter operating in continuous conduction mode (as will be explained in Sect. 4), we will discuss the simple case of two modes, but results are equally applicable to systems with more than two discrete modes.

1. When the switch is ON, the first mode is activated. This lasts for DT_s , where $D < 1$ is the duty cycle (the percentage of time that the switch is ON), in the time interval $t \in [mT_s, (m + D)T_s)$, $\forall m \in \mathbb{Z}$. The system dynamics can be expressed in terms of the state-space matrices $\mathbf{A}^{(1)}$, $\mathbf{b}^{(1)}$, $\mathbf{c}^{(1)}$, $d^{(1)}$.
2. The second mode is activated when the switch is OFF, lasting for the remaining $(1 - D)T_s$, namely for $t \in [(m + D)T_s, (m + 1)T_s)$, with state-space matrices $\mathbf{A}^{(2)}$, $\mathbf{b}^{(2)}$, $\mathbf{c}^{(2)}$, $d^{(2)}$.

This yields the following expression for $\mathbf{A}(t)$, and similarly for $\mathbf{b}(t)$, $\mathbf{c}(t)$ and $d(t)$:

$$\mathbf{A}(t) = \begin{cases} \mathbf{A}^{(1)}, & t \in [mT_s, (m + D)T_s) \\ \mathbf{A}^{(2)}, & t \in [(m + D)T_s, (m + 1)T_s) \end{cases}, \forall m \in \mathbb{Z}. \quad (8)$$

As for any periodic signal, we can express $\mathbf{A}(t)$ in terms of its Fourier Series

$$\mathbf{A}(t) = \sum_{j=-\infty}^{\infty} \mathbf{A}_j \exp^{ij\omega_s t}, \quad \omega_s = 2\pi f_s, \quad (9)$$

and similarly for $\mathbf{b}(t)$, $\mathbf{c}(t)$ and $d(t)$. The Fourier coefficients \mathbf{A}_j can be computed with the classical formula

$$\mathbf{A}_j = \frac{1}{T_s} \int_0^{T_s} \mathbf{A}(t) \exp^{-ij\omega_s t} dt \quad (10)$$

which reduces to

$$\mathbf{A}_j = \frac{1}{T_s} \int_0^{DT_s} \mathbf{A}^{(1)} \exp^{-ij\omega_s t} dt + \frac{1}{T_s} \int_{DT_s}^{T_s} \mathbf{A}^{(2)} \exp^{-ij\omega_s t} dt \quad (11)$$

$$= (\mathbf{A}^{(1)} - \mathbf{A}^{(2)}) \frac{1 - \exp^{-i2\pi j D}}{i2\pi j}, \quad \text{for } j \neq 0 \quad (12)$$

and

$$\mathbf{A}_0 = \frac{1}{T_s} \int_0^{T_s} \mathbf{A}(t) dt = D\mathbf{A}^{(1)} + (1 - D)\mathbf{A}^{(2)}, \quad (13)$$

after taking into account the two switching phases in Eq. (8). Similar expressions are available for the Fourier coefficients of the \mathbf{b} , \mathbf{c} and d quantities. Substituting into (6)–(7), we obtain

$$\dot{\mathbf{x}}(t) = \sum_{j=-\infty}^{\infty} \mathbf{A}_j \exp^{ij\omega_s t} \mathbf{x}(t) + \sum_{j=-\infty}^{\infty} \mathbf{b}_j \exp^{ij\omega_s t} v_{in}(t) \quad (14)$$

$$v_{out}(t) = \sum_{j=-\infty}^{\infty} \mathbf{c}_j \exp^{ij\omega_s t} \mathbf{x}(t) + \sum_{j=-\infty}^{\infty} d_j \exp^{ij\omega_s t} v_{in}(t) \quad (15)$$

which, after applying the Fourier Transform, becomes

$$i\omega \mathbf{X}(i\omega) = \sum_{j=-\infty}^{\infty} \mathbf{A}_j \mathbf{X}(i\omega - ij\omega_s) + \sum_{j=-\infty}^{\infty} \mathbf{b}_j V_{in}(i\omega - ij\omega_s) \quad (16)$$

$$V_{out}(i\omega) = \sum_{j=-\infty}^{\infty} \mathbf{c}_j \mathbf{X}(i\omega - ij\omega_s) + \sum_{j=-\infty}^{\infty} d_j V_{in}(i\omega - ij\omega_s). \quad (17)$$

Fourier Series can also be used for the time-domain signals $\mathbf{x}(t)$, $v_{in}(t)$ and $v_{out}(t)$ in (14)–(15), yielding more complicated equations in (16)–(17) which are omitted here. Eventually, they can be re-written compactly in matrix-form:

$$s\mathcal{X}(s) = (\mathcal{A} - \mathcal{N}) \mathcal{X}(s) + \mathcal{B}\mathcal{V}_{in}(s) \quad (18)$$

$$\mathcal{V}_{out}(s) = \mathcal{C}\mathcal{X}(s) + \mathcal{D}\mathcal{V}_{in}(s), \quad (19)$$

by defining infinite vectors \mathcal{X} , \mathcal{V}_{in} and \mathcal{V}_{out} :

$$\mathcal{X}(s) = \begin{bmatrix} \vdots \\ \mathbf{X}_{-1}(s + i\omega_s) \\ \mathbf{X}_0(s) \\ \mathbf{X}_1(s - i\omega_s) \\ \vdots \end{bmatrix}, \quad \mathcal{V}_{in}(s) = \begin{bmatrix} \vdots \\ V_{in,-1}(s + i\omega_s) \\ V_{in,0}(s) \\ V_{in,1}(s - i\omega_s) \\ \vdots \end{bmatrix}, \quad \mathcal{V}_{out}(s) = \begin{bmatrix} \vdots \\ V_{out,-1}(s + i\omega_s) \\ V_{out,0}(s) \\ V_{out,1}(s - i\omega_s) \\ \vdots \end{bmatrix}, \quad (20)$$

where \mathbf{X}_j , $V_{in,j}$, $V_{out,j}$, $j = -\infty, \dots, \infty$ are the Fourier coefficients of the corresponding time-domain signals in the ω_s -basis. Moreover, \mathcal{A} , \mathcal{B} , \mathcal{C} , \mathcal{D} are doubly-infinite matrices:

$$\mathcal{A} = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \ddots \\ \dots & \mathbf{A}_0 & \mathbf{A}_{-1} & \mathbf{A}_{-2} & \dots & \\ \dots & \mathbf{A}_1 & \mathbf{A}_0 & \mathbf{A}_{-1} & \dots & \\ \dots & \mathbf{A}_2 & \mathbf{A}_1 & \mathbf{A}_0 & \dots & \\ \ddots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad \mathcal{B} = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \ddots \\ \dots & \mathbf{b}_0 & \mathbf{b}_{-1} & \mathbf{b}_{-2} & \dots & \\ \dots & \mathbf{b}_1 & \mathbf{b}_0 & \mathbf{b}_{-1} & \dots & \\ \dots & \mathbf{b}_2 & \mathbf{b}_1 & \mathbf{b}_0 & \dots & \\ \ddots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad (21)$$

$$\mathcal{C} = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \ddots \\ \dots & \mathbf{c}_0 & \mathbf{c}_{-1} & \mathbf{c}_{-2} & \dots \\ \dots & \mathbf{c}_1 & \mathbf{c}_0 & \mathbf{c}_{-1} & \dots \\ \dots & \mathbf{c}_2 & \mathbf{c}_1 & \mathbf{c}_0 & \dots \\ \ddots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad \mathcal{D} = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \ddots \\ \dots & d_0 & d_{-1} & d_{-2} & \dots \\ \dots & d_1 & d_0 & d_{-1} & \dots \\ \dots & d_2 & d_1 & d_0 & \dots \\ \ddots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \quad (22)$$

These are block Toeplitz matrices generated by T_s -periodic functions $\mathbf{A}(t)$, $\mathbf{b}(t)$, $\mathbf{c}(t)$ and $d(t)$, respectively, containing their Fourier coefficients. Last, but not least, \mathcal{N} is the block diagonal matrix

$$\mathcal{N} = \begin{bmatrix} \ddots & & & & \\ & -i\omega_s \mathbf{I} & & & \\ & & \mathbf{0} & & \\ & & & i\omega_s \mathbf{I} & \\ & & & & \ddots \end{bmatrix}. \quad (23)$$

Finally, the Harmonic Transfer Function (HTF), relating harmonics of the output to harmonics of the input signal: $\mathcal{V}_{out}(s) = \mathcal{H}(s)\mathcal{V}_{in}(s)$, is defined as

$$\mathcal{H}(s) = \mathcal{C}(\mathbf{sI} - (\mathcal{A} - \mathcal{N}))^{-1} \mathcal{B} + \mathcal{D} \quad (24)$$

$$= \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \ddots \\ \dots & \mathbf{H}_0(s + i\omega_s) & \mathbf{H}_{-1}(s) & \mathbf{H}_{-2}(s - i\omega_s) & \dots \\ \dots & \mathbf{H}_1(s + i\omega_s) & \mathbf{H}_0(s) & \mathbf{H}_{-1}(s - i\omega_s) & \dots \\ \dots & \mathbf{H}_2(s + i\omega_s) & \mathbf{H}_1(s) & \mathbf{H}_0(s - i\omega_s) & \dots \\ \ddots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \quad (25)$$

The HTF involves infinite state-space matrices, which cannot be implemented on a machine without truncation. Unfortunately, it may be possible that low frequency behavior is affected by truncation because the matrix inverse $(\mathbf{sI} - (\mathcal{A} - \mathcal{N}))^{-1}$ may result in wrapping high frequency components back to low frequency. Truncation effects are significant when the dynamics matrix $\mathbf{A}(t)$ contains an impulse. The Fourier components of an impulse do not roll off (they are constant), causing significant truncation errors and [5] noted that the number of harmonic components needed to produce a good match is larger for systems with discontinuous dynamics equations. However, when the converter is operating in continuous current mode, as shown in [13], truncation does not cause large modeling errors. Hence, in our application, only a limited number of harmonics are of interest (say N), so the infinite quantities are truncated to dimension $2N + 1$, and HTFs in Eq. (25) range from $-N$ to N .

2.2.1 Modeling the Buck Converter from Frequency-Domain Measurements

For DC-DC converters, the load requires a constant supply voltage with minimal disturbance, so a small signal analysis should be considered. In this setting, the input is taken as a constant voltage together with an AC component with amplitude one or several orders of magnitude smaller than the DC voltage:

$$v_{in}(t) = v_{DCin} + v_{ACin} \cos(2\pi f_j t) = v_{DCin} + \frac{v_{ACin}}{2} (\exp^{i\omega_j t} + \exp^{-i\omega_j t}), \quad (26)$$

with f_j , the perturbation frequency of choice, and $v_{DCin} \gg v_{ACin}$. In the Laplace domain, this corresponds to the vector

$$\mathcal{V}_{in}(s) = \begin{bmatrix} \vdots \\ 0 \\ V_{in_0}(s) = \frac{v_{DCin}}{s} + v_{ACin} \frac{s}{s^2 + \omega_j^2} \\ 0 \\ \vdots \end{bmatrix}. \quad (27)$$

The response is a superposition of the output due to the DC input and the output due to the AC input: $v_{out}(t) = v_{DCout}(t) + v_{ACout}(t)$. The DC output contains, besides the 0-frequency, harmonics at integer multiples of the switching frequency (as shown in Fig. 7):

$$v_{DCout}(t) = v_{DC0} + \sum_{\substack{m=-\infty \\ m \neq 0}}^{\infty} v_{DCm} \cos(2\pi m f_s t + \phi_{DCm}). \quad (28)$$

Similarly, the AC output, contains, besides $-f_j$ and f_j , harmonics at $-f_j - m f_s$ and $f_j + m f_s$, $m \in \mathbb{Z}$ (recall Fig. 8):

$$v_{ACout}(t) = v_{AC0} \cos(2\pi f_k t + \phi_{AC0}) + \sum_{\substack{m=-\infty \\ m \neq 0}}^{\infty} v_{ACm} \cos(2\pi (f_k + m f_s) t + \phi_{ACm}). \quad (29)$$

In the Laplace domain, this amounts to expressing $\mathcal{V}_{out}(s) = \mathcal{H}(s)\mathcal{V}_{in}(s)$, namely

$$\mathcal{V}_{out}(s) = \begin{bmatrix} \vdots \\ V_{out_{-1}}(s + i\omega_s) \\ V_{out_0}(s) \\ V_{out_1}(s - i\omega_s) \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ H_{-1}(s) \\ H_0(s) \\ H_1(s) \\ \vdots \end{bmatrix} V_{in_0}(s) \quad (30)$$

In practice, only the information contained in the AC output will be used, for different perturbation frequencies f_j . Depending on the number of HTFs to be identified (e.g., N), the perturbation frequencies should contain enough points and span at least N multiples of f_s . However, values that are equal to multiples of the switching frequency should be avoided, as components add up. Through an FFT analysis of the output, the amplitude and phase of the harmonics are determined and the measurement of the HTF established:

$$H_m(i\omega_j) = \frac{v_{ACm}}{v_{ACin}} \exp^{i\phi_{ACm}}, m = -N, \dots, 0, \dots, N. \quad (31)$$

3 Review of the Loewner Framework

This section reviews the Loewner framework as a tool for solving the problem of modeling LTI systems in the frequency domain introduced in Sect. 2.1 for SISO systems. Here, we consider the general case of multiple-input multiple output (MIMO) systems with p inputs and q outputs. Given pairs of the form (f_j, \mathbf{H}_j) , $j = 1, \dots, P$, where $\mathbf{H}_j \in \mathbb{C}^{q \times p}$ is the matrix measured at f_j , we partition the set of points $\{i\omega_1, \dots, i\omega_P\} = \{\lambda_1, \dots, \lambda_{P/2}\} \cup \{\mu_1, \dots, \mu_{P/2}\}$ into right $\lambda_k, k = 1, \dots, \frac{P}{2}$ and left points $\mu_h, h = 1, \dots, \frac{P}{2}$. We select right tangential directions as column vectors \mathbf{r}_k and left directions as row vectors \mathbf{l}_h . They can be chosen, for simplicity, as vectors of the identity matrix [11]. Matrix data \mathbf{H}_j are converted to right vector data $\mathbf{H}_k \mathbf{r}_k = \mathbf{w}_k$ and left vector data $\mathbf{l}_h \mathbf{H}_h = \mathbf{v}_h$. These quantities are collected into matrices $\mathbf{\Lambda} = \text{diag}[\lambda_1 \dots \lambda_{\frac{P}{2}}]$, $\mathbf{M} = \text{diag}[\mu_1 \dots \mu_{\frac{P}{2}}]$, $\mathbf{R} = [\mathbf{r}_1 \dots \mathbf{r}_{\frac{P}{2}}]$, $\mathbf{W} = [\mathbf{w}_1 \dots \mathbf{w}_{\frac{P}{2}}]$, $\mathbf{L} = [\mathbf{l}_1 \dots \mathbf{l}_{\frac{P}{2}}]^T$, $\mathbf{V} = [\mathbf{v}_1 \dots \mathbf{v}_{\frac{P}{2}}]^T$. Next, the Loewner and shifted Loewner matrices are defined entry-wise as

$$\mathbb{L}_{hk} = \frac{\mathbf{v}_h \mathbf{r}_k - \ell_h \mathbf{w}_k}{\mu_h - \lambda_k}, \quad \sigma \mathbb{L}_{hk} = \frac{\mu_h \mathbf{v}_h \mathbf{r}_k - \lambda_k \ell_h \mathbf{w}_k}{\mu_h - \lambda_k}, \quad h, k = 1, \dots, \frac{P}{2}. \quad (32)$$

We can write a (non-minimal) descriptor realization as in (1)–(2) of the transfer function model

$$\mathbf{H}(s) = \mathbf{W}(\sigma \mathbb{L} - s\mathbb{L})^{-1} \mathbf{V} \quad (33)$$

satisfying the right and left interpolation conditions $\mathbf{H}(\lambda_k) \mathbf{r}_k = \mathbf{w}_k$ and $\mathbf{l}_h \mathbf{H}(\mu_h) = \mathbf{v}_h$ [2]. To obtain a minimal realization, we perform a singular value decomposition (SVD) of a linear combination of the Loewner and shifted Loewner matrices:

$$[\mathbf{Y}, \mathbf{\Sigma}, \mathbf{X}] = \text{svd}(\sigma \mathbb{L} - x\mathbb{L}), \quad x \in \{f_i\}. \quad (34)$$

Choosing n as the singular value where the largest drop between two consecutive singular values takes place (n is application-dependent), we truncate the SVD and define (in Matlab notation) $\mathbf{X}_n = \mathbf{X}(:, 1:n)$ and $\mathbf{Y}_n = \mathbf{Y}(:, 1:n)^*$. The model of size

n in descriptor form is

$$\mathbf{E} = -\mathbf{Y}_n \mathbb{L} \mathbf{X}_n, \quad \mathbf{A} = -\mathbf{Y}_n \sigma \mathbb{L} \mathbf{X}_n, \quad \mathbf{B} = \mathbf{Y}_n \mathbf{V}, \quad \mathbf{C} = \mathbf{W} \mathbf{X}_n, \quad \mathbf{D} = \mathbf{0}. \quad (35)$$

4 Analysis of the Buck Converter as a PSL System

The Buck converter is a hybrid dynamical system with the continuous behavior dictated by the linear time-invariant elements (resistor, capacitor, inductor) and the discrete behavior given by the two switches. Figure 1 shows the topology of a Buck converter which supplies a passive load resistor with voltage v_{out} . The switch is controlled by a binary input signal S : when S is ON, the switch is closed (conducting), and for S OFF, the switch is non-conducting (open), as illustrated in Figs. 2 and 3. We consider the converter to be operating in continuous conduction mode (CCM), meaning that the current through the inductor never reaches zero, hence there are two linear working modes: S ON, \bar{S} OFF in the time interval $t \in [mT_s, (m + D)T_s]$ and S OFF, \bar{S} ON for $t \in [(m + D)T_s, (m + 1)T_s]$, $\forall m \in \mathbb{Z}$.

4.1 Ideal System

This section covers the case of ideal switches, which turn ON or OFF instantaneously. The circuit can be modeled by differential equations involving the capacitor's voltage $v_C(t)$ and the inductor current $i_L(t)$. Equation (36) describes the dynamics of the Buck converter in each working mode $i = 1, 2$ in state-space form:

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}^{(i)}v_{in}(t), \quad \mathbf{x}(t) = \begin{bmatrix} V_C(t) \\ I_L(t) \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} -\frac{1}{RC} & \frac{1}{C} \\ -\frac{1}{L} & 0 \end{bmatrix}, \quad \mathbf{b}^{(1)} = \begin{bmatrix} 0 \\ \frac{1}{L} \end{bmatrix}, \quad \mathbf{b}^{(2)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \quad (36)$$

The response is $v_{out}(t) = v_C(t)$, yielding $\mathbf{c} = [1 \ 0]$ and $d = 0$ for the output equation $v_{out}(t) = \mathbf{c}\mathbf{x}(t) + dv_{in}(t)$. In terms of the analysis described in Sect. 2.2 for periodic systems, we notice that \mathbf{A} , \mathbf{c} and d are the same for both modes. This allows to simplify the matrices appearing in the HTF: $\mathcal{A} = \text{blkdiag}[\dots, \mathbf{A}, \mathbf{A}, \dots]$, $\mathcal{C} = \text{blkdiag}[\dots, \mathbf{c}, \mathbf{c}, \dots]$, $\mathcal{D} = \mathbf{0}$, while \mathcal{B} is full with Fourier coefficients given by $\mathbf{b}_0 = D\mathbf{b}^{(1)}$ and $\mathbf{b}_j = \mathbf{b}^{(1)} \frac{1 - \exp^{-i2\pi j D}}{i2\pi j}$, for $j \neq 0$. This allows to decouple the different HTFs we seek to identify: $H_j(s) = \mathbf{c}(s\mathbf{I} - (\mathbf{A} - ij\omega_s\mathbf{I}))^{-1} \mathbf{b}_j$. One could attempt to identify each of the HTFs separately.

Remark 2 The average model is precisely the HTF $H_0(s) = \mathbf{c}(s\mathbf{I} - \mathbf{A})^{-1} D\mathbf{b}^{(1)}$ for this ideal system.

Fig. 9 Updated circuit diagram of the Buck converter

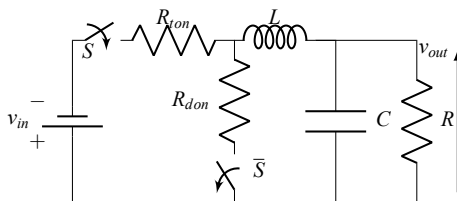
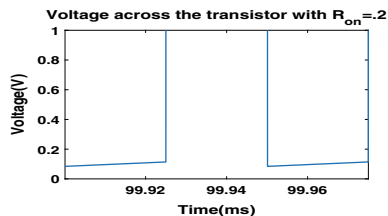


Fig. 10 Influence of R_{ton} on switching



4.2 Integrating the ON Resistance of the Semiconductor Devices

Semiconductor devices are not perfect open circuits when OFF and, similarly, they are not perfect short circuits when ON. In practice, their equivalent circuits would contain resistors, inductors and capacitors. In this section, we account for the small resistance R_{on} when devices are ON. Figure 9 shows the updated circuit diagram when considering R_{ton} , the ON resistance for the transistor, and R_{don} , the ON resistance for the diode. They are each in series with the corresponding ideal switch. The effect of R_{ton} can be seen in Fig. 10, from the voltage across the transistor, with the switching no longer being instantaneous.

In terms of the model in state-space form, accounting for the resistances R_{on} results in different \mathbf{A} matrices for each discrete mode:

$$\mathbf{A}^{(1)} = \begin{bmatrix} -\frac{1}{RC} & \frac{1}{C} \\ -\frac{1}{L} & -\frac{1}{L} \end{bmatrix} \text{ and } \mathbf{A}^{(2)} = \begin{bmatrix} -\frac{1}{RC} & \frac{1}{C} \\ -\frac{1}{L} & -\frac{1}{L} \end{bmatrix}, \tag{37}$$

while the rest of the matrices stay unchanged from Sect. 4.1. In this case, the doubly-infinite matrix \mathcal{A} in (21) appearing in the HTF is no longer block-diagonal, but full, with Fourier coefficients equal to

$$\mathbf{A}_0 = D\mathbf{A}^{(1)} + (1 - D)\mathbf{A}^{(2)} \text{ and } \mathbf{A}_j = (\mathbf{A}^{(1)} - \mathbf{A}^{(2)}) \frac{1 - \exp^{-i2\pi jD}}{i2\pi j} \text{ for } j \neq 0. \tag{38}$$

Hence, the HTFs can no longer be decoupled, as it was the case for ideal switches, and the harmonic transfer functions should be modeled altogether.

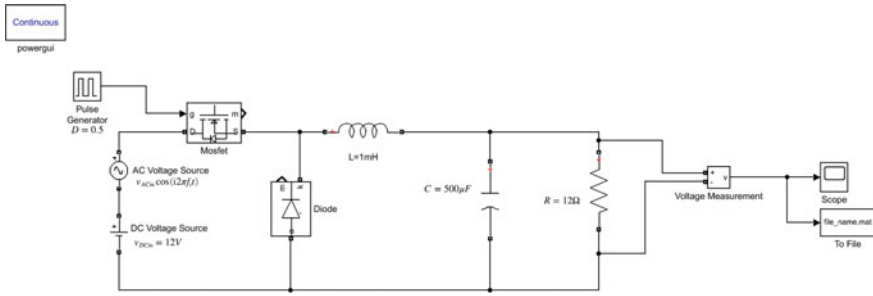


Fig. 11 Matlab simulink/simscape electrical model for the buck converter

4.3 Numerical Experiments

A Matlab 2019a [14] Simulink/Simscape Electrical model [15] (Fig. 11) of a Buck converter with parameters: $L = 1 \text{ mH}$, $C = 500 \mu\text{F}$, $R = 12 \Omega$, $f_s = 20 \text{ kHz}$, $v_{DCin} = 12 \text{ V}$ and duty cycle $D = \frac{1}{2}$ was employed for the numerical experiments. When integrating the R_{on} resistance of the semiconductor devices in the simulation, the chosen values are $R_{ton} = 0.2$ and $R_{don} = 0.01$.

The output for several perturbation frequencies: 10 Hz, 20 Hz, ..., 100 Hz, 200 Hz, ..., 1000 Hz, 2000 Hz, ..., 19000 Hz, 19100 Hz, ..., 19900 Hz, 19910 Hz, 19920 Hz, ..., 19990 Hz, yielding a total of 54 frequencies, was saved to files for subsequent post-processing. A large number of frequencies were considered, from low up to just below the switching frequency, despite the low order model we expect to identify, to ensure that all dynamics are modeled accurately. From Fig. 8, which shows the magnitude of the FFT of v_{out} for $v_{in}(t) = 12 + \cos(2\pi \cdot 5000 \cdot t)$, we notice that components at frequencies 0, 20 kHz, 40 kHz, etc., are present, together with components at frequencies 5 kHz, $20 - 5 = 15 \text{ kHz}$, $20 + 5 = 25 \text{ kHz}$, $40 - 5 = 35 \text{ kHz}$, $40 + 5 = 45 \text{ kHz}$, etc. Using the insight from Fig. 8, the magnitude of AC components superior to 35 kHz are below -100 dB , typically considered as engineering precision. Hence, the infinite sum in the HTF is truncated to $N = 1$, leading to the following HTFs to be identified: $H_{-1}(s)$, $H_0(s)$ and $H_1(s)$ from the FFT analysis of $V_{out_{-1}}(i\omega_j + i\omega_s)$, $V_{out_0}(i\omega_j)$ and $V_{out_1}(i\omega_j - i\omega_s)$ in (30), respectively. Moreover, recall that, for our application, the converter is operating in continuous current mode and truncation of the infinite HSS matrices does not cause large errors. This fact is verified in Tables 1 and 2, where the $N = 100$ truncation offers no improvement or only a slight improvement in the accuracy with respect to the $N = 1$ truncation.

Making use of the complex conjugate information from the negative frequencies

$$\begin{cases} V_{out_{-1}}(-i\omega_j + i\omega_s) = \overline{V_{out_1}(i\omega_j - i\omega_s)} \text{ together with (30)} \Rightarrow H_{-1}(-i\omega_j) = \overline{H_1(i\omega_j)} \\ V_{out_0}(-i\omega_j) = \overline{V_{out_0}(i\omega_j)} \text{ together with (30)} \Rightarrow H_0(-i\omega_j) = \overline{H_0(i\omega_j)} \\ V_{out_1}(-i\omega_j - i\omega_s) = \overline{V_{out_{-1}}(i\omega_j + i\omega_s)} \text{ together with (30)} \Rightarrow H_1(-i\omega_j) = \overline{H_{-1}(i\omega_j)}. \end{cases}$$

and applying the output transformation [10]

Fig. 12 SVD drop of the Loewner matrix pencil for the ideal system

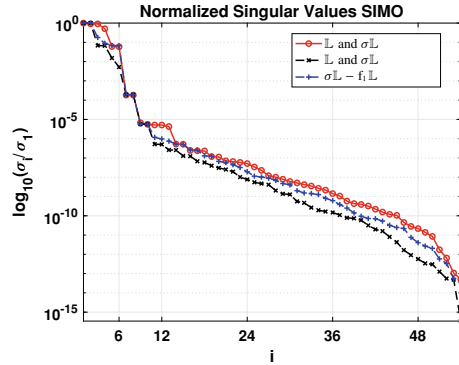
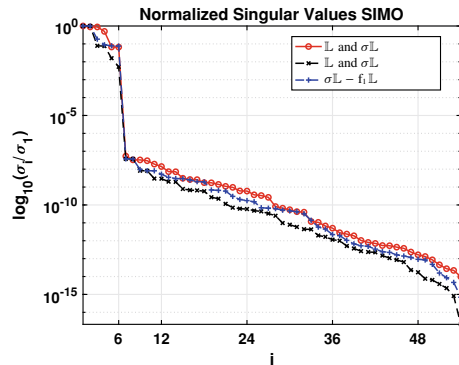


Fig. 13 SVD drop of the Loewner matrix pencil for the system with parasitics



$$\mathbf{T}_y = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ -i & 0 & i \end{bmatrix} \Rightarrow \mathbf{T}_y \begin{bmatrix} H_{-1}(s) \\ H_0(s) \\ H_1(s) \end{bmatrix} = \begin{bmatrix} H_{-1}(s) + H_1(s) \\ H_0(s) \\ -iH_1(s) + iH_1(s) \end{bmatrix} \triangleq \hat{\mathbf{H}}(s) \quad (39)$$

yields the usual complex conjugate relationship for $\hat{\mathbf{H}}$: $\hat{\mathbf{H}}(-i\omega_j) = \overline{\hat{\mathbf{H}}(i\omega_j)}$, allowing one to obtain state-space matrices with real entries from the single input multiple output measurements $\hat{\mathbf{H}}_j$. In the Loewner framework (Sect. 3), frequencies with odd indices together with their negatives are used as right data, while those with even indices and their negatives as left data. We expect a system of order 6, due to the matrix $\mathcal{A} - \mathcal{N}$ having size 6, when truncating (21) such that the HTF contains only $H_{-1}(s)$, $H_0(s)$ and $H_1(s)$. Figures 12 and 13 show the singular value drop of the matrices \mathbb{L} , $\sigma\mathbb{L}$ and a linear combination thereof, exhibiting a drop of several orders of magnitude after the 6th singular value (more clear for the system with parasitics). This allows to identify an order 6 system, as expected. The final realization is obtained by applying the inverse of the output transformation and premultiplying the resulting \mathbf{C} matrix with \mathbf{T}_y^{-1} .

Fig. 14 $H_{-1}(s)$ in the ideal case

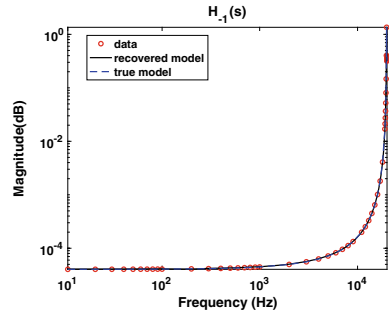


Fig. 15 $H_0(s)$ in the ideal case

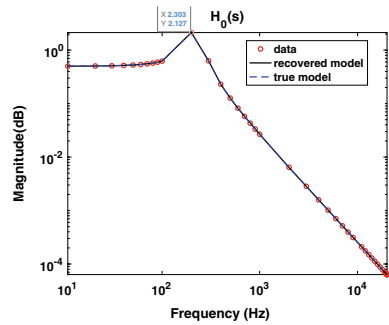
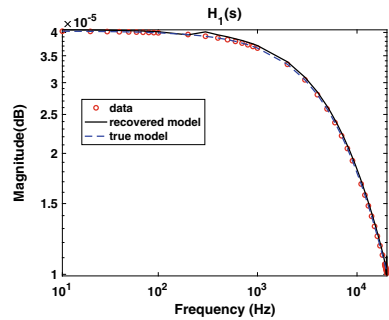


Fig. 16 $H_1(s)$ in the ideal case



Figures 14, 15 and 16 show the magnitude of the three HTFs $H_{-1}(s)$, $H_0(s)$ and $H_1(s)$ in the ideal case, plotted only for positive frequencies. The recovered models match the data, as well as the theoretical HTFs.

The same good agreement is noticed in Figs. 17, 18 and 19 from the magnitude of the three HTFs $H_{-1}(s)$, $H_0(s)$ and $H_1(s)$ when taking into account the ON resistance of the switches. The effect of these resistances can be seen in the damping of the low frequency resonance of $H_0(s)$ and in the appearance of dips at the same low frequency resonance in the response of $H_{-1}(s)$ and $H_1(s)$. The increase in damping can also be explained from the pole plots of the two systems in Figs. 20 and 21. While

Fig. 17 $H_{-1}(s)$ with parasitics

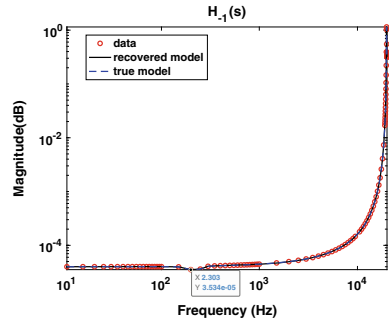


Fig. 18 $H_0(s)$ with parasitics

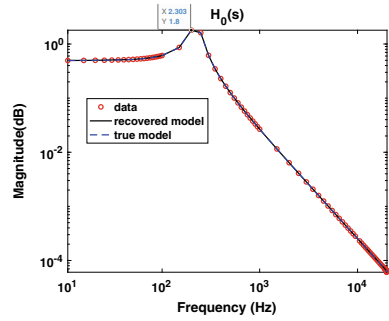
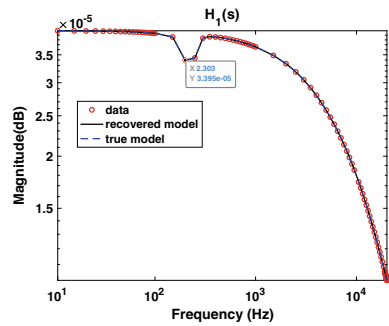


Fig. 19 $H_1(s)$ with parasitics



the imaginary parts of the poles remain largely unaffected (on the default scale), the real part has decreased from taking into account parasitics.

Remark 3 Since the HTFs can no longer be decoupled as in the ideal case, modeling each HTF individually would not allow to recover the system of order 6.

Fig. 20 Pole plot for the ideal system

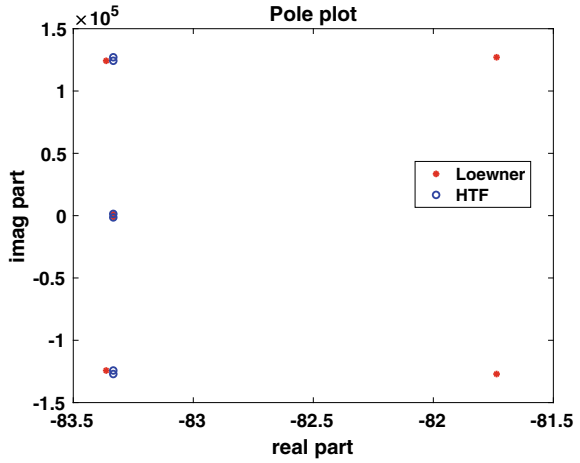
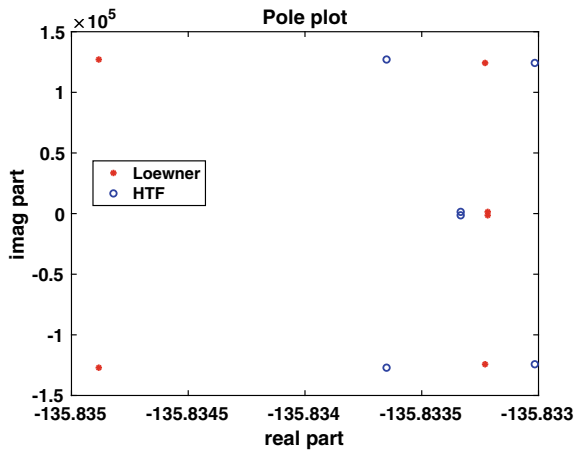


Fig. 21 Pole plot for the system with parasitics



Remark 4 Even though the true model obtained by truncating the infinite harmonic state-space matrices matches the data in Figs. 14, 15, 16, 17, 18 and 19, when computing the errors

$$\mathcal{H}_\infty \text{ error} = \frac{\max_{j=1 \dots P} (\mathbf{H}_j - \mathbf{H}(i\omega_j))}{\max_{j=1 \dots P} (\mathbf{H}_j)} \tag{40}$$

and

Table 1 Errors for the ideal system

| | \mathcal{H}_∞ error | \mathcal{H}_2 error |
|-------------------|----------------------------|------------------------|
| Loewner | $2.5112 \cdot 10^{-4}$ | $1.6839 \cdot 10^{-4}$ |
| HTF ($N = 1$) | $7.9216 \cdot 10^{-3}$ | $6.7037 \cdot 10^{-3}$ |
| HTF ($N = 100$) | $7.9216 \cdot 10^{-3}$ | $6.7037 \cdot 10^{-3}$ |

Table 2 Errors for the system with parasitics

| | \mathcal{H}_∞ error | \mathcal{H}_2 error |
|-------------------|----------------------------|------------------------|
| Loewner | $3.2559 \cdot 10^{-7}$ | $2.8704 \cdot 10^{-7}$ |
| HTF ($N = 1$) | $7.9498 \cdot 10^{-3}$ | $6.7565 \cdot 10^{-3}$ |
| HTF ($N = 100$) | $7.9201 \cdot 10^{-3}$ | $6.7442 \cdot 10^{-3}$ |

$$\mathcal{H}_2 \text{ error} = \sqrt{\frac{\sum_{j=1 \dots P} \|\mathbf{H}_j - \mathbf{H}(i\omega_j)\|_F^2}{\sum_{j=1 \dots P} \|\mathbf{H}_j\|_F^2}}, \text{ where } \|\mathbf{H}(i\omega_j)\|_F^2 = \sum_{m=-N \dots N} |\mathbf{H}_m(i\omega_j)|^2, \quad (41)$$

we obtain the values displayed in the last two rows of Table 1 for the ideal system and Table 2 for the system with parasitics. Surprisingly, these errors are rather large. For the ideal system, increasing the number of terms in the truncation to $N = 100$ yields precisely the same errors as with $N = 1$. For the system with parasitics, increasing the number of terms in the truncation to $N = 100$ improves the errors marginally with respect to the truncation with $N = 1$ discussed in this paper. Last, but not least, the accuracy of the models built with the Loewner framework from data is in the order of the first neglected singular value (the 7th in our case), as seen from Figs. 12 and 13, namely 10^{-4} for the ideal system and 10^{-7} for the system with parasitics.

5 Conclusion and Future Work

This chapter presents a first step towards modeling periodic switched linear systems from frequency domain measurements which can be physically achievable. The theory presented relies on the identification of harmonic transfer functions in the Loewner framework.

While the descriptor-form of the model can be used in simulations, the realization for the vector of HTFs is of the form (35) and can no longer be interpreted physically in terms of Fourier coefficients. Hence, future work involves applying suitable similarity transformations to retrieve matrices from which the Fourier coefficients of the periodic state-space matrices $\mathbf{A}(t)$, $\mathbf{b}(t)$, $\mathbf{c}(t)$ and $d(t)$ can be easily identi-

fied. Unfortunately, as shown by the errors in Tables 1 and 2, data obtained from the Fourier analysis of the output signals from Simulink do not perfectly match the harmonic transfer function model and the need to obtain direct, instead of simulation measurements, is evident.

This chapter focused on the open-loop Buck converter as an application case. In practice, the system is typically operating in closed-loop, with the switching signal obtained through Pulse Width Modulation, by comparing state variables to the clocking carrier waveform. The control signal can be taken into consideration as an extra input [16] and a similar analysis through the HTF can be performed. In future work, we will also account for the control-to-output transfer function in the modeling procedure. Moreover, more realistic models integrating high-frequency parasitic elements of the semiconductor devices and of the output filter will be considered.

Acknowledgements This work has been achieved within the framework of the CE2I project (Convertisseur d'Énergie Intégré Intelligent). CE2I is co-financed by European Union with the financial support of European Regional Development Fund (ERDF), the French State and the French Region of Hauts-de-France.

References

1. Erickson, R.W.: Large signals in switching converters, Ph.D. thesis, California Institute of Technology (1983)
2. Mayo, A.J., Antoulas, A.C.: A framework for the solution of the generalized realization problem. *Linear Algebra Appl.* **425**(2–3), 634–662 (2007)
3. Middlebrook, R.D., Cuk, S.: A general unified approach to modelling switching-converter power stages. In: *Proceedings of the IEEE Power Electronics Specialists Conference*, pp. 18–34 (1976)
4. Amedo, L., Burgos, R., Wang, F., Boroyevich, D.: Black-Box Terminal Characterization Modeling of DC-to-DC Converters. In: *Proceedings of the IEEE Applied Power Electronics Conference and Exposition*, pp. 457–463 (2007)
5. Wereley, N.W.: Analysis and Control of Linear Periodically Time Varying Systems, Ph.D. thesis, Massachusetts Institute of Technology (1991)
6. Siddiqi, A.: Identification of the Harmonic Transfer Functions of a Helicopter Rotor, MSc thesis, Massachusetts Institute of Technology (1999)
7. Mollerstedt, E.: Dynamic Analysis of Harmonics in Electrical Systems, Ph.D. thesis, Department of Automatic Control, Lund Institute of Technology (2000)
8. Uyanık, İ., Saranlı, U., Ankaralı, M.M., Cowan, N.J., Morgül, Ö.: Frequency-domain subspace identification of linear time-periodic (LTP) systems. *IEEE Trans. Autom. Control* **64**(6), 2529–2536 (2019)
9. Ramirez, A., Abdel-Rahaman, M., Noda, T.: Frequency-domain modeling of time-periodic switched electrical networks: a review. *Ain Shams Eng. J.* **9**(4), 2527–2533 (2018)
10. Uyanık, İ., Saranlı, U., Morgül, Ö., Ankaralı, M.M.: Parametric identification of hybrid linear-time-periodic systems *IFAC-PapersOnLine* **49**(9), 7–12 (2016)
11. Lefteriu, S., Antoulas, A.C.: A new approach to modeling multiport systems from frequency-domain data. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **29**(1), 14–27 (2010)
12. Gosea, I.V., Petreczky, M., Antoulas, A.C.: Data-driven model order reduction of linear switched systems in the Loewner framework. *SIAM J. Sci. Comput.* **40**(2), B572–B610 (2018)

13. Love, G.N.: Small signal modelling of power electronic converters, for the study of time-domain waveforms, harmonic domain spectra, and control interactions, Ph.D. thesis, University of Canterbury (2007)
14. MATLAB R2019a, Natick, Massachusetts: The MathWorks Inc. (2019)
15. Srinivasan, K.: Buck-Converter Open loop, MATLAB Central File Exchange (<https://www.mathworks.com/matlabcentral/fileexchange/69285-buck-converter-open-loop>). Retrieved August 27 (2019)
16. Yue, X., Wang, X., Blaabjerg, F.: Review of small-signal modeling methods including frequency-coupling dynamics of power converters. *IEEE Trans. Power Electron.* **34**(4), 3313–3328 (2019)

Model Reduction and Realization Theory of Linear Switched Systems



Mihály Petreczky and Ion Victor Gosea

Abstract The goal of this chapter is to present an overview of some recent results on model reduction of linear switched systems and their interplay with realization theory of these systems. The emphasis will be on those results on model reduction which are directly related to realization theory, we do not aim at being exhaustive. In particular, we will review some recent results on balanced truncation and moment matching, focusing on the theoretical aspects rather than on the computational ones.

Keywords Hybrid systems · Realization theory · Model reduction · Balanced truncation · Moment matching

1 Introduction

In this chapter we will present an overview of some recent results on model reduction of hybrid systems which rely heavily on realization theory. Both model reduction and realization theory have been central to Prof. Antoulas's work, so we feel that showing the interaction between these two topics for hybrid systems is a fitting tribute to his scientific contribution.

Hybrid systems [13] are non-linear systems which combine continuous and discrete behavior. More precisely, a hybrid system is a finite collection of continuous-state dynamical systems, indexed by a set of so called *discrete modes (or states)*. The state of each dynamical system is governed by a set of differential or difference equations. The discrete mode in any time instant can be chosen arbitrarily or

M. Petreczky (✉)

Centre de Recherche en Informatique, Signal et Automatique de Lille (CRISTAL), CNRS, Ecole Centrale, CNRS (Centre national de la recherche scientifique), Avenue Carl Gauss, 59650 Villeneuve-d'Ascq, France
e-mail: mihaly.petreczky@centralelille.fr

I. V. Gosea

Data-Driven System Reduction and Identification (DRI) Group, Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstrasse 1, 39106 Magdeburg, Germany
e-mail: gosea@mpi-magdeburg.mpg.de

© Springer Nature Switzerland AG 2022

C. Beattie et al. (eds.), *Realization and Model Reduction of Dynamical Systems*,
https://doi.org/10.1007/978-3-030-95157-3_11

it may depend on the value of the continuous state and possibly other constraints, which are referred to as *guards*. The transitions between the discrete states may result in a jump in the state of the underlying continuous dynamical system. This jump is defined by the application of the so called *reset maps*. *Linear switched systems (LSSs)* [19, 37] are the simplest and most widely studied subclass of hybrid systems where the continuous subsystems are linear systems, and the change of the discrete state is externally generated.

While there is a large literature on control of hybrid systems in general, and of LSSs in particular, the computational complexity of the existing algorithms for hybrid systems is high, and hence they cannot be applied to large scale systems. In order to address this problem, model reduction methods were proposed for hybrid systems. Model reduction methods for LSSs can be grouped into the following categories.

LMI-based methods. These methods compute the matrices of the reduced order model by solving a set of linear matrix inequalities (LMIs). The disadvantage is that the proposed conditions are only sufficient, and the trade-off between the dimension of the reduced model and the error bound is not clear. Moreover, the computational complexity of solving those LMIs might be too high. Without claiming completeness, we mention [12, 38–40].

Methods based on local Gramians. These algorithms are based on finding observability/controllability Gramians for each linear subsystem. For these methods often there are no error bounds and the reduced order model need not be well-posed. Examples of such papers include [7–9, 14, 16, 20, 21]. Note that to the best of our knowledge, the only algorithm which always yields a well-posed LSS of the same type as the original one and for which there exists an analytic error bound (which holds only for slow switching) is the one of [14]. For the case of jump-linear systems with a stochastic switching a similar approach was taken in [18] and an error bound was derived.

Methods based on common Gramians. These methods rely on finding the same observability/controllability Gramians for each linear subsystem. These Gramians are derived as solutions of a suitable LMI. Such algorithms were described in [34, 35] and an analytic error bound was derived in [29]. These algorithms apply only to LSSs which have a global quadratic Lyapunov function. Moreover, the computational complexity of solving the corresponding LMIs is high. In order to address this problem [31] proposes to replace LMIs by Lyapunov equations. The downside of the latter approach is that the error bounds of [29] do not always apply. Another approach was proposed in [33], where for a specific subclass of LSSs, the original LSS is replaced by a linear time-invariant (LTI) system and classical balanced truncation is applied.

Moment matching. The idea behind these algorithms is to find a reduced order linear switched system such that certain coefficients of the series expansions of the input-output maps of the original and the reduced order system coincide. The series expansion can be the Taylor series with respect to switching times, in which case a number of the so-called Markov parameters coincide. Alternatively, the series expansion can be a Laurent-series expansion of a multivariate Laplace transform of the input-output map around a certain frequency. The former approach was pursued

in [1, 4, 5], the latter in [15]. While those methods do not allow for analytical error bounds, under suitable assumption it can be guaranteed that the reduced model will have the same input-output behavior for certain switching signals [1, 4, 5]. A somewhat different approach is that of [32], which considers LSSs with state-dependent switching and it proposes a model reduction procedure which guarantees that the reduced model has the same steady-state output response to certain inputs as the original model.

In this chapter we discuss some recent results on balanced truncation and moment matching for LSSs. For the sake of simplicity, *we consider LSSs only in continuous time, and we will assume that all the linear subsystems are defined on the same state-space and the reset maps are identity*. Most of the presented results are true for the discrete-time case too, and some can be extended to include LSSs with reset maps which are not identity. We will cite the relevant literature on these extensions.

The model reduction methods to be discussed in this chapter rely heavily on realization theory. The goal of realization theory is to understand the relationship between input-output behaviors and internal (state-space) representations. A fairly complete realization theory was developed for LSSs, see the discussion and references in [22–25, 28]. The results on realization theory of LSSs rely on realization theory of bilinear systems and recognizable formal power series [6, 17, 36].

Realization theory is relevant for model reduction in many ways. First, minimization and realization algorithms can be viewed as simple model reduction algorithms. Moreover, the relationship between span-reachability, observability and minimality is closely related to the existence of Gramians which are used in balanced truncation. Realization theory is even more critical for moment matching, as the latter can be viewed as partial realization algorithm. We will elaborate on the precise relationship later on.

The chapter is structured as follows. In Sect. 2 we present the formal definition of the class of LSSs and the corresponding terminology. In Sect. 3 we present a brief overview of the relevant results on realization theory of switched systems. In Sect. 4 we discuss model reduction: In Sect. 4.1 we present balanced truncation and in Sect. 4.2 we discuss moment matching for LSSs.

2 Linear Switched Systems: Basic Definitions

A *linear switched system* (LSS) is a control system of the form

$$\Sigma \begin{cases} \dot{x}(t) = A_{\sigma(t)}x(t) + B_{\sigma(t)}u(t) \\ y(t) = C_{\sigma(t)}x(t), \end{cases} \quad (1)$$

where $x(t) \in \mathbb{R}^n$ is the continuous-valued state at time t , $\sigma(t) \in \mathcal{Q}$ is the discrete mode at time t , $y(t) \in \mathbb{R}^p$ is the output at time t , and $u(t) \in \mathbb{R}^m$ is the continuous-

valued input at time t . The set Q is a finite one, and it is referred to as the set of discrete modes or states. Moreover, $A_q \in \mathbb{R}^{n \times n}$, $B_q \in \mathbb{R}^{n \times m}$, $C_q \in \mathbb{R}^{p \times n}$ are the matrices of the linear system in the discrete state $q \in Q$. The number n is called the *dimension (order)* of Σ and will be denoted by $\dim(\Sigma)$. The short-hand notation

$$\Sigma = \{A_q, B_q, C_q\}_{q \in Q}$$

is used for LSSs of the form (1).

Let \mathbb{R}_+ be the *real time-axis*, i.e. $\mathbb{R}_+ = [0, +\infty)$. Denote by $AC(\mathbb{R}_+, \mathbb{R}^k)$ respectively $PC(\mathbb{R}_+, \mathbb{R}^k)$ the set of all absolutely continuous respectively piecewise-continuous functions of the form $h : \mathbb{R}_+ \rightarrow \mathbb{R}^k$.¹ Let $\mathcal{X} = AC(\mathbb{R}_+, \mathbb{R}^n)$, $\mathcal{U} = PC(\mathbb{R}_+, \mathbb{R}^m)$, $\mathcal{Y} = PC(\mathbb{R}_+, \mathbb{R}^p)$, and let \mathcal{Q} be the set of all piecewise-constants functions $g : \mathbb{R}_+ \rightarrow Q$. A tuple $(x, u, \sigma, y) \in \mathcal{X} \times \mathcal{U} \times \mathcal{Q} \times \mathcal{Y}$ is called a solution, if (x, u, σ, y) satisfy (1). For any switching signal $\sigma \in \mathcal{Q}$, input $u \in \mathcal{U}$ and initial state $x_0 \in \mathbb{R}^n$, there exists a unique solution (x, u, σ, y) of Σ such that $x(0) = x_0$. This prompts us to define the *input-to-state* map $\mathcal{X}_\Sigma : \mathcal{U} \times \mathcal{Q} \rightarrow \mathcal{X}$ and the *input-output* map $Y_\Sigma : \mathcal{U} \times \mathcal{Q} \rightarrow \mathcal{Y}$ of an LSS Σ as follows: $X_\Sigma(u, \sigma) = x$, and $Y_\Sigma(u, \sigma) = y$ if and only if (x, u, σ, y) is the unique solution of Σ such that $x(0) = 0 \in \mathbb{R}^n$.²

Intuitively, an LSS is just a control system which switches among finitely many linear time-invariant systems. The switching signal is part of the input. Whenever a switch occurs, the continuous state remains the same, only the differential equation governing the state and output evolution changes. That is, whenever we switch to a new linear system, we start the new linear system from the state which is the final state of the previous linear system.

We model potential input-output behaviors of LSSs as functions

$$f : \mathcal{U} \times \mathcal{Q} \rightarrow \mathcal{Y}, \quad (2)$$

and call them *input-output maps*. They capture the behavior of a black-box, which reacts to piecewise-continuous inputs and switching sequences by generating outputs in \mathbb{R}^p . Next, we define what it means that this black-box can be modelled as an LSS, i.e. that an LSS is a realization of f . The LSS Σ is a *realization of an input-output map f of the form (2)*, if $Y_\Sigma = f$, i.e. if the input-output map of Σ coincides with f . If Σ is a realization of f , then Σ is a *minimal realization of f* , if for any LSS realization $\hat{\Sigma}$ of f , $\dim \Sigma \leq \dim \hat{\Sigma}$. Two LSSs Σ_1, Σ_2 are said to be *input-output equivalent*, if their input-output maps are equal, i.e. $Y_{\Sigma_1} = Y_{\Sigma_2}$. An LSS Σ is said to be *minimal*, if it is a minimal realization of its own input-output map $f = Y_\Sigma$.

¹ Piecewise-continuous functions have a finite number of discontinuities on each finite interval and at each point of discontinuity, the left- and right-hand side limits exist and are finite.

² The definition of the input-to-state and input-output map can be extended to include non-zero initial states [23, 24]. We prefer to stick to zero initial state to avoid excessive notation and terminology.

3 Realization Theory of Linear Switched Systems

Below we present the main results on minimality, existence of a realization, and a Ho-Kalman-like realization algorithm for LSSs. These results are identical for discrete-time LSSs [25], and can be extended to LSSs with linear reset maps [24, 27]. In order to present these results, and later throughout the chapter, we will use the following notation from automata theory [11].

Notation (Q^* , Q^+ , ϵ , q^k) Denote by Q^+ the set of all finite sequences of elements of Q , i.e. each element $w \in Q^+$ is of the form $w = a_1 a_2 \cdots a_k$ for some $a_1, a_2, \dots, a_k \in Q$, $k \in \mathbb{N}$, $k > 0$. The integer k is called the *length* of w and it is denoted by $|w|$. Let $\epsilon \notin Q^+$ be a symbol, which we will call the *empty sequence or empty word*. By convention, the length of ϵ is defined to be zero. Denote by Q^* the set $Q^+ \cup \{\epsilon\}$. For any two sequences $w, v \in Q^*$, we denote by wv the concatenation of w and v . If $w, v \in Q^+$ are of the form $v = v_1 v_2 \cdots v_k$, $k > 0$ and $w = w_1 w_2 \cdots w_m$, $m > 0$, $v_1, v_2, \dots, v_k, w_1, w_2, \dots, w_m \in Q$, then define $vw = v_1 v_2 \cdots v_k w_1 w_2 \cdots w_m$. If $v = \epsilon$ and $w \in Q^*$, then define $vw = w$, and if $w = \epsilon$ and $v \in Q^*$, then define $vw = v$. For $q \in Q$ and $k \in \mathbb{N}$, $k > 0$, we denote by q^k the sequence $\overbrace{qq \cdots q}^{k\text{-times}}$; by convention $q^0 = \epsilon$.

3.1 Minimality of Linear Switched Systems

We start by presenting the main results on minimality of LSSs. To this end, we introduce the notions of observability, span-reachability and isomorphism. Let Σ be an LSS of the form (1). Then Σ is said to be *observable*, if for any two distinct states $x_{1,0} \neq x_{2,0} \in \mathbb{R}^n$, there exists an input u and a switching signal σ , such that if (x_i, u, σ, y_i) , $i = 1, 2$ are two solutions of Σ with $x_i(0) = x_{i,0}$, $i = 1, 2$, then $y_1 \neq y_2$, i.e., the outputs induced by the initial states $\{x_{i,0}\}_{i=1,2}$ under the input u and switching signal σ are different. Let $\mathcal{R}_0(\Sigma) \subseteq \mathbb{R}^n$ denote the reachable set of the Σ from the zero initial state, i.e., $\mathcal{R}_0(\Sigma)$ is the set of all vectors x_f such that for some $t \in \mathbb{R}_+$, $x_f = X_\Sigma(u, \sigma)(t)$ for some $u \in \mathcal{U}$, $\sigma \in \mathcal{Q}$. We say that Σ is *span-reachable*, if $\mathbb{R}^n = \text{Span } \mathcal{R}_0(\Sigma)$, i.e., if \mathbb{R}^n is the smallest vector space containing $\mathcal{R}_0(\Sigma)$. Note that span-reachability and reachability are the same in continuous-time [37].

Two LSSs $\Sigma_1 = \{A_q, B_q, C_q\}_{q \in Q}$, and $\Sigma_2 = \{A_q^a, B_q^a, C_q^a\}_{q \in Q}$ are said to be *isomorphic*, if there exists a non-singular square matrix \mathcal{S} such that $A_q^a = \mathcal{S}A_q\mathcal{S}^{-1}$, $B_q^a = \mathcal{S}B_q$, and $C_q^a = C_q\mathcal{S}^{-1}$ for all $q \in Q$.

Theorem (Minimality, [22, 23]) *An LSS is minimal, if and only if it is span-reachable and observable. If Σ_1 and Σ_2 are two minimal LSSs and they are input-output equivalent, then they are isomorphic.* \square

Note that minimality of an LSS does not imply minimality of any of its linear subsystems, see [23] for a counter-example. Hence, realization theory of LSSs cannot be reduced to realization theory of linear subsystems.

We present an algorithm for converting any LSS to a minimal one while preserving its input-output map. To this end, we define the subspaces

$$\mathcal{V}^* = \text{Span}\mathcal{R}_0(\Sigma), \quad (3)$$

$$\mathcal{W}^* = \text{Span}\{x_0 \in \mathbb{R}^n \mid \forall \sigma \in \mathcal{Q} : \exists x \in \mathcal{X} : \quad (4)$$

$$(x, 0, \sigma, 0) \text{ is a solution of } \Sigma \text{ with } x(0) = x_0 \}.$$

It can be shown that Σ is span-reachable if and only if $\mathcal{V}^* = \mathbb{R}^n$, and Σ is observable, if and only if $\mathcal{W}^* = \{0\}$, see [24, 37].

Procedure (Minimization) Consider the factor space $\mathcal{V}^*/\mathcal{W}^*$; recall that the elements of this linear space are equivalence classes generated by the following equivalence relation \approx on \mathcal{V}^* : $x_1 \approx x_2$ if and only if $x_1 - x_2 \in \mathcal{W}^*$. Let $[x] = \{y \mid y \approx x\}$ be the equivalence class represented by x . The linear vector space $\mathcal{V}^*/\mathcal{W}^*$ is finite dimensional. Define the linear maps $A_q : \mathcal{V}^*/\mathcal{W}^* \ni [x] \mapsto [A_q x] \in \mathcal{V}^*/\mathcal{W}^*$, $C_q : \mathcal{V}^*/\mathcal{W}^* \ni [x] \mapsto C_q x \in \mathbb{R}^p$, $B_q : \mathbb{R}^m \ni u \mapsto [B_q u] \in \mathcal{V}^*/\mathcal{W}^*$, $q \in \mathcal{Q}$. It can be shown [22, 23] that $A_q \mathcal{V}^* \subseteq \mathcal{V}^*$, $A_q \mathcal{W}^* \subseteq \mathcal{W}^*$, $\text{Im} B_q \subseteq \mathcal{V}^*$, $\mathcal{W}^* \subseteq \ker C_q$. From this it follows that the linear maps A_q, B_q, C_q are well defined. Choose a finite basis in $\mathcal{V}^*/\mathcal{W}^*$, and let ${}^m A_q, {}^m B_q, {}^m C_q$ be the matrix representations of the linear maps A_q, B_q, C_q , in that basis. Then $\Sigma_m = \{{}^m A_q, {}^m B_q, {}^m C_q\}_{q \in \mathcal{Q}}$ is a minimal LSS which is input-output equivalent to Σ .

3.2 Existence of a Realization, Ho-Kalman Algorithm, Markov Parameters

We first define the notion of a generalized kernel representation, existence of which is a necessary condition for existence of a realization by an LSS. An input-output map f has a *generalized kernel representation*, if there exists a family of functions $\{G_v^f : \mathbb{R}_+^{|v|} \rightarrow \mathbb{R}^{p \times m}\}_{v \in \mathcal{Q}^+}$, such that for all $u \in \mathcal{U}$, $\sigma \in \mathcal{Q}$,

$$f(u, \sigma)(t) = \sum_{i=1}^k \int_0^{t_i} G_{q_1 \dots q_k}^f(t_i - s, t_{i+1}, \dots, t_k) u(s + T_{i-1}) ds$$

where $q_i \in \mathcal{Q}$, $0 < t_i \in \mathbb{R}_+$, $i = 1, \dots, k$ are such that $\sigma(s) = q_i$ for $s \in [T_{i-1}, T_i)$ for some $T_j \in \mathbb{R}_+$, $T_j < T_{j+1}$, $j \in \mathbb{N}$, $T_0 = 0$, and $t \in [T_{k-1}, T_k)$ for some $k > 0$, and $t_i = T_i - T_{i-1}$, for $i = 1, \dots, k-1$ and $t_k = t - T_{k-1}$. Moreover, $\{G_v^f\}_{v \in \mathcal{Q}^+}$ have to satisfy a number of technical conditions [22, 23]. From [22, 23] it follows that Σ of the form (1) is a realization of f , if and only if f has a generalized kernel representation and for all $q_1, \dots, q_k \in \mathcal{Q}$, $t_1, \dots, t_k \in \mathbb{R}_+$, $k > 0$,

$$G_{q_1 \dots q_k}^f(t_1, \dots, t_k) = C_{q_k} e^{A_{q_k} t_k} \dots e^{A_{q_2} t_2} e^{A_{q_1} t_1} B_{q_1}. \quad (5)$$

From the technical conditions in [22, 23] on generalized kernel representations it follows that if f has a generalized kernel representation, then there exist functions $\{S_{r,q}^f : Q^* \rightarrow \mathbb{R}^{p \times m}\}_{r,q \in Q}$ such that

$$G_{q_1 \dots q_k}^f(t_1, \dots, t_k) = \sum_{\alpha_1, \dots, \alpha_k=0}^{\infty} S_{q_k, q_1}^f(q_1^{\alpha_1} \dots q_k^{\alpha_k}) \prod_{i=1}^k \frac{t_i^{\alpha_i}}{\alpha_i!}.$$

That is, the functions $\{S_{r,q}^f\}_{r,q \in Q}$ uniquely determine $\{G_v^f\}_{v \in Q^+}$ and hence f , and conversely, the functions $\{S_{r,q}^f\}_{r,q \in Q}$ can be recovered from f . The values of $\{S_{r,q}^f\}_{r,q \in Q}$ are called the *Markov parameters* of f .

Notation For every $v \in Q^*$ and a collection of $n \times n$ matrices $\{A_q\}_{q \in Q}$ define the matrix A_v as follows: if $v = q_1 \dots q_k$ with $q_1, \dots, q_k \in Q$, $k > 0$, then $A_v = A_{q_k} A_{q_{k-1}} \dots A_{q_1}$, and if $v = \epsilon$, then $A_\epsilon = I_n$, where I_n is the $n \times n$ identity matrix.

Lemma ([22, 23]) *An LSS of the form (1) is a realization of f , if and only if f has a generalized kernel representation, and for all $v \in Q^*$, $q, q_0 \in Q$, $S_{q,q_0}^f(v) = C_q A_v B_{q_0}$.*

In order to present a Ho-Kalman-like realization algorithm and a Hankel-rank condition for existence of an LSS realization, we consider only the SISO case, i.e., $p = m = 1$, see [4, 26] for the general case, and we adapt the notion of selection from [4, 26]. We call any subset $\alpha \subset Q^* \times Q$ a *selection*. Consider selections α and β , such that α is of finite cardinality n_α and β is of finite cardinality n_β respectively. Fix an enumeration

$$\alpha = \{(u_i, q_i)\}_{i=1}^{n_\alpha}, \quad \beta = \{(v_j, \sigma_j)\}_{j=1}^{n_\beta}. \quad (6)$$

Let us now define the matrix $\mathcal{H}_{\alpha,\beta}^f \in \mathbb{R}^{n_\alpha \times n_\beta}$ as follows:

$$\left[\mathcal{H}_{\alpha,\beta}^f \right]_{i,j} = S_{q_i, \sigma_j}^f(v_j u_i) \quad i = 1, \dots, n_\alpha, \quad j = 1, \dots, n_\beta. \quad (7)$$

Intuitively, the rows of $\mathcal{H}_{\alpha,\beta}^f$ are indexed by the elements of α , and the columns by the elements of β .

Theorem (Existence [23]) *The input-output map f has a realization by an LSS, if and only if f has a generalized kernel representation, and*

$$\sup_{\alpha, \beta \subseteq Q^* \times Q, \alpha, \beta \text{ are finite}} \text{rank } \mathcal{H}_{\alpha,\beta}^f = n_m < +\infty. \quad (8)$$

If (8) holds, then n_m is the dimension of a minimal LSS realization of f . \square

The proof of Theorem 3.2 leads to the following Ho-Kalman-like algorithm. For the selections α, β from (6), define the matrices $\mathcal{H}_{q,\alpha,\beta}^f \in \mathbb{R}^{n_\alpha \times n_\beta}$, $\mathcal{H}_{\alpha,q}^f \in \mathbb{R}^{n_\alpha \times 1}$ and $\mathcal{H}_{q,\beta}^f \in \mathbb{R}^{1 \times n_\beta}$, $q \in Q$ as follows:

$$\left[\mathcal{H}_{q,\alpha,\beta}^f \right]_{i,j} = S_{q_i,\sigma_j}^f(v_j q u_i), \quad \left[\mathcal{H}_{\alpha,q}^f \right]_{i,1} = S_{q_i,q}^f(u_i), \quad \left[\mathcal{H}_{q,\beta}^f \right]_{1,j} = S_{q,\sigma_j}^f(v_j).$$

for all $i = 1, \dots, n_\alpha$, $j = 1, \dots, n_\beta$.

Procedure (Ho-Kalman algorithm) Assume that $\text{rank } \mathcal{H}_{\alpha,\beta}^f = n_m$. Consider the factorization $H_{\alpha,\beta}^f = O_{n_m} R_{n_m}$ such that $O_{n_m} \in \mathbb{R}^{n_\alpha \times n_m}$, $R_{n_m} \in \mathbb{R}^{n_m \times n_\beta}$, $\text{rank } O_{n_m} = \text{rank } R_{n_m} = n_m$. Define

$$\hat{A}_q = O_{n_m}^+ \mathcal{H}_{q,\alpha,\beta}^f R_{n_m}^+, \quad \hat{B}_q = O_{n_m}^+ \mathcal{H}_{\alpha,q}^f, \quad \hat{C}_q = \mathcal{H}_{q,\beta}^f R_{n_m}^+$$

and $O_{n_m}^+, R_{n_m}^+$ is the Moore-Penrose inverse of O_{n_m} and R_{n_m} respectively. Define the LSS $\hat{\Sigma} = \{\hat{A}_q, \hat{B}_q, \hat{C}_q\}_{q \in Q}$.

Lemma ([10, 22, 26]) *If n_m is the dimension of a minimal LSS realization of f , then the LSS $\hat{\Sigma}$ defined in Procedure 3.2 is a minimal realization of f .*

From [28] it follows that we can choose $\alpha_N = \beta_N = \{(v, q) \mid v \in Q^*, |v| \leq N, q \in Q\}$, where N is any integer not smaller than the dimension of a minimal LSS realization of f . This choice of the nice selection is not very practical, as the size of the Hankel-matrix H_{α_N, β_N}^f grows exponentially with N . Note that the Ho-Kalman algorithm described above can be used to define a smooth (analytic) manifold structure for the space of equivalence classes of minimal LSSs related by isomorphism [26].

4 Model Reduction

In model reduction, we would like to find LSSs of *smaller* dimension, input-output maps of which are *close* (but not necessarily equal) to that of the original LSS. This is in contrast to realization theory, where we were interested in finding a minimal LSS with *exactly* the same input-output map as the original one. The latter is a special case of the former. Model reduction algorithms follow the following general pattern.

Algorithm 1 Model reduction/minimization algorithm

Inputs: $\Sigma = \{A_q, B_q, C_q\}_{q \in Q}$, matrices $V \in \mathbb{R}^{n \times r_1}$, $W \in \mathbb{R}^{r_2 \times n}$.

Output: $\bar{\Sigma} = \{\bar{A}_q, \bar{B}_q, \bar{C}_q\}_{q \in Q}$.

1: Let $r = \text{rank } WV$ and let $S \in \mathbb{R}^{r \times r_2}$, $T \in \mathbb{R}^{r_1 \times r}$, $SWVT = I_r$.

2: $\bar{A}_q = SWA_qVT$, $\bar{C}_q = C_qVT$, $\bar{B}_q = SWB_q$, $q \in Q$.

3: **return** $\bar{\Sigma} = \{\bar{A}_q, \bar{B}_q, \bar{C}_q\}_{q \in Q}$.

Intuitively, Algorithm 1 restricts the system to the set $\text{Im}V$ and then merges those of its states x_1, x_2 for which $x_1 - x_2 \in \ker W$. If $\ker W = \mathcal{W}^*$ and $\text{Im}V = \mathcal{V}^*$, with $\mathcal{W}^*, \mathcal{V}^*$ from (3), then the LSS $\bar{\Sigma}$ returned by Algorithm 1 is a minimal LSS which is input-output equivalent to Σ , i.e., Algorithm 1 is just an implementation of Procedure 3.1. Algorithms for computing such matrices W, V such that $\ker W = \mathcal{W}^*$ and $\text{Im}V = \mathcal{V}^*$ are described in [22, 24].

In case of model reduction, Algorithm 1 can again be used. However, instead of applying it with matrices W and V such that $\ker W = \mathcal{W}^*$ and $\text{Im}V = \mathcal{V}^*$, we use matrices W, V such that $\mathcal{W}^* \subseteq \ker W$ and $\text{Im}V \subseteq \mathcal{V}^*$, i.e., we restrict the system to a subset of the set of reachable states, or we merge states which do not produce the same input-output behavior. The resulting LSS model will no longer be a realization of f , but its input-output map will approximate f in a suitable sense. Depending on the method we use, we will either be able to provide a global error bound on the difference between the input-output maps of the original model and the reduced one, or state that for certain switching sequences the two input-output maps coincide. We will elaborate on various methods below.

4.1 Model Reduction by Balanced Truncation

Let Σ be an LSS of the form (1), and assume that Σ is *quadratically stable*, i.e., there exists a matrix $P > 0$ such that $\forall q \in Q : A_q^T P + P A_q < 0$. In this case Σ is globally uniformly asymptotically (exponentially) stable [19] with the Lyapunov function $V(x) = x^T P x$. A matrix \mathcal{Q} will be called an *observability Gramian*, if

$$\forall q \in Q : A_q^T \mathcal{Q} + \mathcal{Q} A_q + C_q^T C_q \leq 0, \mathcal{Q} > 0. \tag{9}$$

Likewise, a matrix \mathcal{P} will be called a *controllability Gramian*, if

$$\forall q \in Q : A_q \mathcal{P} + \mathcal{P} A_q^T + B_q B_q^T \leq 0, \mathcal{P} > 0. \tag{10}$$

Note that in contrast to the linear case, controllability/observability Gramians for LSSs are not unique, since they are solutions of LMIs and not of Lyapunov equations.

The procedure for balanced truncation is as follows. We apply Algorithm 1 with the following choice of W and V . Find U such that $\mathcal{P} = U U^T$ and find an orthogonal L such that $U^T \mathcal{Q} U = L \Lambda^2 L^T$, where $\Lambda = \text{diag}(\sigma_1, \dots, \sigma_n)$ and $\sigma_1 \geq \dots \geq \sigma_n \geq 0$. Pick $r \leq n$. Define

$$W = [I_r \ 0] \Lambda^{1/2} L^T U^{-1}, \quad V = U L \Lambda^{-1/2} \begin{bmatrix} I_r \\ 0 \end{bmatrix}.$$

Then $\text{rank } W = \text{rank } V = r, \text{rank } W V = r, S = T = I_r$.

The intuition behind the procedure above is similar to that of balanced truncation for linear systems: By applying the transformation $\mathcal{S} = \Lambda^{1/2} L^T U^{-1}$ to Σ , we obtain

an LSS $\Sigma_{bal} = \{SA_q\mathcal{S}^{-1}, \mathcal{S}B_q, C_q\mathcal{S}^{-1}\}_{q \in Q}$, such that $\Lambda = \mathcal{S}^{-T} \mathcal{Q}\mathcal{S}^{-1} = \mathcal{S} \mathcal{P} \mathcal{S}^T$ is both an observability and a controllability Gramian. We obtain $\bar{\Sigma}$ from Σ_{bal} by taking the upper-left $r \times r, r \times m, p \times r$, blocks of $A_q, B_q, C_q, q \in Q$ respectively. That is, we discard those states which correspond to small values of the diagonals of Λ . The intuition behind this approach is that the discarded states are either difficult to reach (it requires high energy input to reach them) or difficult to observe (their contribution to the energy of the output is small). More precisely, let us fix an integer $r > 0$ which represents the desired state dimension of the reduced order model. Let $(\tilde{x}, u, \sigma, \tilde{y})$ be a solution of Σ_{bal} such that $\tilde{x}(0) = 0$, assume that for all $t > \tau_0, u(t) = 0$. It then can be shown [29] that

$$\begin{aligned} \sum_{i=1}^r \tilde{x}_i^2(\tau_0) \frac{1}{\sigma_i} + \sum_{i=r+1}^n \frac{1}{\sigma_i} \tilde{x}_i^2(\tau_0) &\leq \int_0^{\tau_0} \|u(s)\|_2^2 ds, \\ \sum_{i=1}^r \tilde{x}_i^2(\tau_0) \sigma_i + \sum_{i=r+1}^n \sigma_i \tilde{x}_i^2(\tau_0) &\geq \int_{\tau_0}^{\infty} \|\tilde{y}(s)\|_2^2 ds, \end{aligned} \tag{11}$$

and $\tilde{x}_i(\tau_0)$ denotes the i th component of $\tilde{x}(\tau_0)$. That is, if $\sigma_{r+1}, \dots, \sigma_n$ are small, and the energy of u is small, i.e., $\int_0^{\tau_0} \|u(s)\|_2^2 ds$ is small, then the values $\tilde{x}_{r+1}(\tau_0), \dots, \tilde{x}_n(\tau_0)$ have to be small due to the first inequality, and they contribute little to the output starting from the time instance τ_0 due to the second inequality.

The numbers $\sigma_1, \dots, \sigma_n$ are called *singular values* of the pair $(\mathcal{P}, \mathcal{Q})$ and they are the square roots of the eigenvalues of the product $\mathcal{P} \mathcal{Q}$.

Theorem ([29]) *For any $\sigma \in Q, u \in \mathcal{U}$ such that $\int_0^{\infty} \|u(s)\|_2^2 ds < +\infty$,*

$$\int_0^{\infty} \|Y_{\Sigma}(u, \sigma)(s) - Y_{\bar{\Sigma}}(u, \sigma)(s)\|_2^2 ds \leq (2 \sum_{k=r+1}^n \sigma_k)^2 \int_0^{\infty} \|u(s)\|_2^2 ds.$$

□

Further extensions The results discussed above also hold for discrete-time LSSs [29]. The assumption that \mathcal{P}, \mathcal{Q} do not depend on $q \in Q$ implies quadratic stability, which is a quite restrictive assumption. In [14] this assumption was replaced by local stability of the linear subsystems. Moreover, an error bound similar to Theorem 4.1 was derived in [14], but it holds only for switching signals with a sufficiently large dwell time. Note that [14] allows for LSSs with non-trivial linear reset maps. In [31] an alternative definition of Gramians was presented which has the advantage of having a tighter relationship with minimality.

Relationship with realization theory First, the existence of positive definite Gramians \mathcal{P}, \mathcal{Q} is a necessary (but not sufficient) condition for minimality of quadratically stable LSSs [29]: the kernels of positive semi-definite observability (resp. controllability) Gramians are contained in the set of unobservable states (resp. states which are not in the span of reachable states). Intuitively, when the Gramians are positive definite, we can bring them to a balanced form and then identify the

states corresponding to small singular values with unobservable/unreachable states. Then balanced truncation can be thought of as a numerical implementation of the minimization Procedure 3.1. Theorem 4.1 provides means to evaluate the effect of the discarded “small” singular values on the approximation error.

Realization theory is also necessary to show that balanced truncation is well posed. More precisely, the application of balanced truncation relies on the availability of a quadratically stable LSS and by Theorem 4.1, the quality of the reduced model relies on the singular values of the observability/controllability Gramians. If the original model is not quadratically stable, then one may wonder if there exist input-output equivalent quadratically stable models. Using realization theory it is shown in [29] that in order to decide if balanced truncation can be applied, it is sufficient to transform the original model to a minimal one, and then to check if the minimal model is quadratically stable. Moreover, any minimal model can be used for balanced truncation without introducing more conservativity. Indeed, in [29] it is shown that the singular values of any pair of controllability/observability Gramians for any LSS are not smaller than the singular values of some pair of Gramians of a minimal input-output equivalent LSS. Moreover, due to isomorphism, all controllability/observability Gramians of minimal input-output equivalent LSSs are related by a similarity transform and have the same singular values.

Finally, realization theory and the notion of Hankel-matrix can be used to relate singular values of Gramians to norms of a Hankel-operator [29].

4.2 Moment Matching

Consider an LSS Σ of the form (1), and let us denote its input-output map by f . For the sake of simplicity, we assume that $p = m = 1$, i.e., we deal only with the SISO case. Recall that since f is realizable by an LSS then f has to have a generalized kernel representation $\{G_v^f\}_{v \in Q^+}$. The idea of moment matching is to find a reduced order LSS $\bar{\Sigma}$ such that for certain sequences $v \in Q^+$, G_v^f is close to $G_v^{Y_{\bar{\Sigma}}}$. Intuitively, this means that the input-output map of $\bar{\Sigma}$ will be close to that f .

In Sect. 4.2.1 we present the approach [1, 4, 5], where we look for a reduced order model $\bar{\Sigma}$ such that certain Taylor-series coefficients of $G_v^{Y_{\bar{\Sigma}}}$ (Markov parameters of $Y_{\bar{\Sigma}}$) and of G_v^f (Markov parameters of f) coincide. In Sect. 4.2.2 we present the approach [15], where we consider multi-variate Laplace transforms H_v^f of G_v^f and we look for reduced order models $\bar{\Sigma}$ such that the Laplace transform $H_v^{Y_{\bar{\Sigma}}}$ of the function $G_v^{Y_{\bar{\Sigma}}}$ coincides with H_v^f for some complex values.

4.2.1 Matching Markov Parameters

Consider an LSS $\bar{\Sigma}$. Let α and β be two selections. Then $\bar{\Sigma}$ is called a (α, β) -partial realization of f , if for every $(v, q_0) \in \beta$, $(u, q) \in \alpha$,

$$S_{q,q_0}^{Y_{\bar{\Sigma}}}(vu) = S_{q,q_0}^f(vu). \tag{12}$$

That is, $\bar{\Sigma}$ is a (α, β) -partial realization of f , if those Markov parameters of f and of the input-output map $Y_{\bar{\Sigma}}$ of $\bar{\Sigma}$ which are indexed by (α, β) coincide. This means that certain high-order derivatives of f and of $Y_{\bar{\Sigma}}$ are the same. That is, a (α, β) -partial realization of f can be viewed as an LSS, input-output map of which approximates f . If $\alpha = \mathcal{Q}^* \times \mathcal{Q}$ or $\beta = \mathcal{Q}^* \times \mathcal{Q}$, then any (α, β) -partial realization of f is a realization of f .

The idea behind moment matching is then to replace an LSS Σ by a reduced order LSS $\bar{\Sigma}$ such that $\bar{\Sigma}$ is a (α, β) -partial realization of the input-output map $f = Y_{\Sigma}$ of Σ . The various algorithms differ in the way the selections α, β are chosen. The moment matching algorithms which produce (α, β) -partial realizations arise from Algorithm 1 by a suitable choice of the matrices W and V . In order to explain these choices in more detail, we introduce the following definitions. Define the subspaces

$$\mathcal{O}_{\alpha}(\Sigma) = \bigcap_{(v,q) \in \alpha} \ker C_q A_v, \quad \mathcal{R}_{\beta}(\Sigma) = \text{Span}\{A_w B_{q_0} \mid (w, q_0) \in \beta\}.$$

There are 3 choices of matrices W and V .

- (A) $\ker W = \mathcal{O}_{\alpha}(\Sigma)$ and $V = I_n$. Then Algorithm 1 returns a $(\alpha, \{\epsilon\} \times \mathcal{Q})$ -partial realization of f [4, Theorem 3].
- (B) $\text{Im} V = \mathcal{R}_{\beta}(\Sigma)$ and $W = I_n$. Then Algorithm 1 returns a $(\{\epsilon\} \times \mathcal{Q}, \beta)$ -partial realization of f , [4, Theorem 2].
- (C) $\ker W = \mathcal{O}_{\alpha}(\Sigma)$, $\text{Im} V = \mathcal{R}_{\beta}(\Sigma)$, $\text{rank } W = \text{rank } V = \text{rank } WV$. Then Algorithm 1 returns (α, β) -partial realization of f , [4, Theorem 4].

There are two strategies for choosing α, β .

The first one is to choose α (resp. β) to be of finite cardinality r such that $\dim \mathcal{O}_{\alpha}(\Sigma) = n - r$ (resp. $\dim \mathcal{R}_{\beta}(\Sigma) = r$), and in this case the reduced order model will have dimension r . In this case, $\mathcal{O}_{\alpha}(\Sigma) = \ker O_{\alpha}$ (resp. $\mathcal{R}_{\beta}(\Sigma) = \text{Im} R_{\beta}$), and

$$O_{\alpha} = [A_{v_1}^T C_{s_1}^T, \dots, A_{v_r}^T C_{s_r}^T]^T, \quad R_{\beta} = [A_{w_1} B_{q_1}, \dots, A_{w_r} B_{q_r}],$$

where $\alpha = \{(v_i, s_i)\}_{i=1}^r$, and $\beta = \{(w_i, q_i)\}_{i=1}^r$. Using these matrix representations the matrices W and V described above can easily be computed.

The second option for choosing nice selections is to choose (α, β) to be consistent with a certain set of switching signals. In this case, the dimension of the reduced order model cannot be fixed in advance, but it is known that the reduced order model will have the same input-output behavior along those switching sequences which belong to this designated set. More precisely, assume that the switching signal $\sigma \in \mathcal{Q}$ has the property that $\sigma(s) = q_i, s \in [T_{i-1}, T_i), T_0 = 0, T_i = \sum_{r=1}^i t_r$ for some $q_i \in \mathcal{Q}, 0 < t_i \in \mathbb{R}_+, 0 < i \in \mathbb{N}$. We will say that a selection (α, β) is *consistent with* σ , if for every $i > 0$, for every $\omega_i, \dots, \omega_k \in \mathbb{N}$,

$$((q_i)^{\omega_i} (q_{i+1})^{\omega_{i+1}} \dots (q_k)^{\omega_k}, q_i) \in \beta, \quad ((q_1)^{\omega_1} (q_2)^{\omega_2} \dots (q_i)^{\omega_i}, q_i) \in \alpha.$$

Theorem ([4]) *Assume that (α, β) is consistent with σ and $\bar{\Sigma}$ is an (α, β) -partial realization of f . Then $Y_{\bar{\Sigma}}(u, \sigma) = f(u, \sigma)$, for all $u \in U$. \square*

Note that for a pair of selections to be consistent with a switching signal (or a set of switching signals), the selections involved have to be infinite sets. If the prefixes of the sequences of discrete modes of the desired switching signals form a regular language, then there exist algorithms to compute matrix representations of $\mathcal{O}_\alpha(\Sigma)$, $\mathcal{R}_\beta(\Sigma)$, see [4, 5].

Further extensions. The model reduction method described above was extended to linear parameter-varying (LPV) models [3] and bilinear systems [30]. In addition, the method above was applied to LSSs arising from asynchronous sampling of linear time-invariant systems [2].

Relationship with realization theory. To begin with, the whole idea of matching Markov parameters relies on the notion of Markov parameters and partial realization, which are integral parts of realization theory. In fact, the result of the Ho-Kalman realization algorithm from Procedure 3.2 is isomorphic to the LSS returned by Algorithm 1 with the choice of the matrices W and V as described in option (C) above. That is, moment matching is just a reformulation of Ho-Kalman algorithm when the latter is applied to finite Hankel-matrices, rank of which is not maximal. The partial realization algorithm of [28] is a particular instance of this model reduction method, if $\alpha = \beta = \{v \in Q^* \mid |v| \leq N\} \times Q$ is chosen. Furthermore, Theorem 4.2.1 and its counterpart for the discrete-time case [5] can be viewed as extensions of realization theory of LSSs with constrained switching [22, 23].

4.2.2 Moment Matching in Frequency Domain

By applying multivariate Laplace transform of the functions $\{G_v^f\}_{v \in Q^+}$ we can define a sequence of functions $\{H_v^f\}_{v \in Q^+}$ of complex variables as follows:

$$H_v^f(s_1, \dots, s_k) = \int_0^\infty \cdots \int_0^\infty G_v(t_1, \dots, t_k) e^{s_1 t_1 + \cdots + s_k t_k} dt_1 \cdots dt_k \quad (13)$$

for all $\operatorname{Re}(s_i) > s_0$ for a suitable $s_0 \in \mathbb{R}$, where $k = |v|$. If f has a realization by a LSS Σ of the form (1) then $G_{q_1 \dots q_k}^f(t_1, \dots, t_k)$ satisfies (5), and hence

$$H_{q_1, q_2, \dots, q_k}^f(s_1, s_2, \dots, s_k) = C_{q_k} \Phi_{q_k}(s_k) \Phi_{q_{k-1}}(s_{k-1}) \cdots \Phi_{q_1}(s_1) B_{q_1}, \quad (14)$$

where $\Phi_q(s) = (sI_n - A_q)^{-1}$, $q_j \in Q$, $1 \leq j \leq k$. We call the functions $\{H_v^f\}_{v \in Q^+}$ the *generalized transfer functions* of the input-output map f [15].

Let Γ and Θ be finite sets of tuples so that $\Gamma, \Theta \subseteq \{(v, \underline{\mu}) \mid v \in Q^+, \underline{\mu} \in \mathbb{C}^k, k = |v|\}$. We will say that an LSS $\bar{\Sigma}$ is a (Γ, Θ) -partial realization of f if for every $(w, \underline{\mu}) \in \Gamma$, $(v, \underline{\lambda}) \in \Theta$, $H_{wv}^f(\underline{\mu}, \underline{\lambda}) = H_{wv}^{Y_{\bar{\Sigma}}}(\underline{\mu}, \underline{\lambda})$.

Our goal is to find an LSS $\bar{\Sigma}$ such that $\bar{\Sigma}$ is a (Γ, Θ) -partial realization of f , and the dimension of $\bar{\Sigma}$ is smaller than that of Σ . To this end, for any $v = q_1 \cdots q_k \in Q^+$, $q_1, \dots, q_k \in Q$, define

$$\mathbf{r}((v, \underline{\mu})) = \Phi_{q_k}(\mu_k) \cdots \Phi_{q_1}(\mu_1) B_{q_1}, \quad \mathbf{o}((v, \underline{\mu})) = C_{q_k} \Phi_{q_k}(\mu_k) \cdots \Phi_{q_1}(\mu_1),$$

for any $\underline{\mu} = (\mu_1, \dots, \mu_k) \in \mathbb{C}^k$. Assume that the cardinality of Γ and Θ are both r and consider an enumeration $\Gamma = \{(w_i, \underline{\mu}_i)\}_{i=1}^r$, $\Theta = \{(v_i, \underline{\lambda}_i)\}_{i=1}^r$ of these sets. Define the matrices

$$R = [\mathbf{r}((w_1, \underline{\mu}_1)) \cdots \mathbf{r}((w_r, \underline{\mu}_r))], \quad O = [\mathbf{o}((v_1, \underline{\lambda}_1))^T \cdots \mathbf{o}((v_r, \underline{\lambda}_r))^T]^T.$$

Assume that $\text{rank } OR = r$. We can apply Algorithm 1 with $W = O$, $V = R$ resulting in an LSS $\bar{\Sigma}$ which will have the following property.

Theorem ([15]) *With the notation and assumptions above, the LSS $\bar{\Sigma}$ is a (Γ, Θ) -partial realization of f . \square*

This method has an alternative formulation in terms of *generalized Loewner matrices* [15], thus extending the well-known Loewner matrix based model reduction method for linear systems.

Relationship with realization theory The reformulation of this method in terms of generalized Loewner matrices yields a partial realization algorithm, as it depends on data which can directly be obtained from Laplace transforms of the input-output map. In a way, this method is the first step towards a reformulation of realization theory of LSSs in frequency domain.

5 Conclusions

In this chapter we presented a brief overview of some recent results on realization theory and model reduction of linear switched systems. It is well known that for linear systems there is a deep connection between these two disciplines. We hope that this chapter convinces the reader that this remains true for hybrid systems and that it is worthwhile to do further research on this topic. There are many possible directions for future research. A particularly natural one is to extend the results of this chapter to hybrid systems with state-dependent switching, for example to piecewise linear systems. The latter can be viewed as a feedback interconnection of a linear switched system with a discrete event generator, hence we are hopeful that the results of this chapter will be useful for such an extension.

References

1. Bastug, M.: Model Reduction of Linear Switched Systems and LPV State-Space Models. Ph.D. thesis, Aalborg University (2016)
2. Bastug, M., Hetel, L., Petreczky, M.: Minimality of aperiodic sampled data systems. In: American Control Conference (ACC) (2017)
3. Bastug, M., Petreczky, M., Tóth, R., Wisniewski, R., Leth, J., Efimov, D.: Moment matching based model reduction for lpv state-space models. In: 2015 IEEE 54rd Annual Conference on Decision and Control (CDC) (2015)
4. Bastug, M., Petreczky, M., Wisniewski, R., Leth, J.: Model reduction by nice selections for linear switched systems. *IEEE Trans. Autom. Control* **61**, 3422–3437 (2016)
5. Bastug, M., Petreczky, M., Wisniewski, R., Leth, J.: Reachability and observability reduction for linear switched systems with constrained switching. *Automatica* **74**, 162–170 (2016)
6. Berstel, J., Reutenauer, C.: Rational series and Their Languages. Springer (1984)
7. Birouche, A., Guilet, J., Mourllion, B., Basset, M.: Gramian based approach to model order-reduction for discrete-time switched linear systems. In: Proceedings of the 18th Mediterranean Conference on Control and Automation, pp. 1224–1229 (2010)
8. Birouche, A., Mourllion, B., Basset, M.: Model reduction for discrete-time switched linear time-delay systems via the H_∞ stability. *Control. Intell. Syst.* **39**(1), 1–9 (2011)
9. Birouche, A., Mourllion, B., Basset, M.: Model order-reduction for discrete-time switched linear systems. *Int. J. Syst. Sci.* **43**(9), 1753–1763 (2012)
10. Cox, P., Tóth, R., Petreczky, M.: Towards efficient maximum likelihood estimation of lpv-ss models. *Automatica* **97**, 392–403 (2018)
11. Eilenberg, S.: Automata. Languages and Machines. Academic, New York, London (1974)
12. Gao, H., Lam, J., Wang, C.: Model simplification for switched hybrid systems. *Syst. Control Lett.* **55**, 1015–1021 (2006)
13. Goebel, R., Sanfelice, R.G., Teel, A.R.: Hybrid Dynamical Systems: Modeling, Stability, and Robustness. Princeton University Press, New Jersey, NJ (2012)
14. Gosea, I.V., Petreczky, M., Antoulas, A.C.: Balanced truncation for linear switched systems. *Adv. Comput. Math.* **44**(6), 1845–1886 (2018)
15. Gosea, I.V., Petreczky, M., Antoulas, A.C.: Data-driven model order reduction of linear switched systems in the Loewner framework. *SIAM J. Sci. Comput.* **40**(2), 572–610 (2018)
16. Gosea, I.V., Pontes Duff, I., Benner, P., Antoulas, A.C.: Model Order Reduction of Switched Linear Systems with Constrained Switching, IUTAM Symposium on Model Order Reduction of Coupled Systems Stuttgart, Germany, May 22–25, 2018. IUTAM Bookseries, vol. 36, pp. 41 – 53. Springer, Cham (2020)
17. Isidori, A.: Direct construction of minimal bilinear realizations from nonlinear input-output maps. *IEEE Trans. Autom. Control* 626–631 (1973)
18. Kotsalis, G., Rantzer, A.: Balanced truncation for discrete-time Markov jump linear systems. *IEEE Trans. Autom. Control* **55**(11) (2010)
19. Liberzon, D.: Switching in Systems and Control. Birkhäuser, Boston (2003)
20. Monshizadeh, N., Trentelman, H.L., Camlibel, M.K.: A simultaneous balanced truncation approach to model reduction of switched linear systems. *IEEE Trans. Autom. Control* **57**(12), 3118–3131 (2012)
21. Papadopoulos, A.V., Prandini, M.: Model reduction of switched affine systems. *Automatica* **70**, 57–65 (2016)
22. Petreczky, M.: Realization Theory of Hybrid Systems. Ph.D. thesis, Vrije Universiteit, Amsterdam (2006)
23. Petreczky, M.: Realization theory of linear and bilinear switched systems: a formal power series approach- Part I: realization theory of linear switched systems. *ESAIM Control Optim. Calc. Var.* **17**, 410–445 (2011)
24. Petreczky, M.: Realization theory of linear hybrid systems. Observation and control, Lecture Notes in Control and Information Sciences. In Hybrid Dynamical Systems. Springer (2015)

25. Petreczky, M., Bako, L., van Schuppen, J.H.: Realization theory for discrete-time linear switched systems. *Automatica* **49**(11), 3337–3344 (2013)
26. Petreczky, M., Peeters, R.: Spaces of nonlinear and hybrid systems representable by recognizable formal power series. In: *Proceedings of MTNS2010*, pp. 1051–1058 (2010)
27. Petreczky, M., van Schuppen, J.H.: Realization theory for linear hybrid systems. *IEEE Trans. Autom. Control* **55**, 2282–2297 (2010)
28. Petreczky, M., van Schuppen, J.H.: Partial-realization of linear switched systems: A formal power series approach. *Automatica* **47**(10), 2177–2184 (2011)
29. Petreczky, M., Wisniewski, R., Leth, J.: Balanced truncation for linear switched systems. *Nonlinear Anal. Hybrid Syst* **10**, 4–20 (2013)
30. Petreczky, M., Wisniewski, R., Leth, J.: Moment matching for bilinear systems with nice selections. In: *10th IFAC Symposium on Nonlinear Control Systems* (2016)
31. Pontes Duff, I., Grundel, S., Benner, P.: New Gramians for switched linear systems: reachability, observability, and model reduction. accepted in *IEEE Trans. Autom. Control* (2018). [arXiv:1806.00406](https://arxiv.org/abs/1806.00406)
32. Scarciotti, G., Astolfi, A.: Model reduction for hybrid systems with state-dependent jumps. *IFAC-PapersOnLine* **49**(18), 850–855 (2016)
33. Schulze, P., Unger, B.: Model reduction for linear systems with low-rank switching. *SIAM J. Control. Optim.* **56**(6), 4365–4384 (2018)
34. Shaker, H.R., Wisniewski, R.: Generalized gramian framework for model/controller order reduction of switched systems. *Int. J. Syst. Sci.* **42**(8), 1277–1291 (2011)
35. Shaker, H.R., Wisniewski, R.: Model reduction of switched systems based on switching generalized gramians. *Int. J. Innov. Comput. Inf. Control* **8**(7(B)), 5025–5044 (2012)
36. Sontag, E.D.: Realization theory of discrete-time nonlinear systems: Part I – the bounded case. *IEEE Trans. Circuits Syst.* **CAS-26**(4) (1979)
37. Sun, Z., Ge, S.S.: *Switched linear systems* : control and design. Springer, London (2005)
38. Zhang, L., Boukas, E., Shi, P.: μ -Dependent model reduction for uncertain discrete-time switched linear systems with average dwell time. *Int. J. Control* **82**(2), 378–388 (2009)
39. Zhang, L., Shi, P., Boukas, E., Wang, C.: H-infinity model reduction for uncertain switched linear discrete-time systems. *Automatica* **44**(8), 2944–2949 (2008)
40. Zheng-Fan, L., Chen-Xiao, C., Wen-Yong, D.: Stability analysis and H_∞ model reduction for switched discrete-time time-delay systems. *Math. Probl. Eng.* **15** (2014)

Structured Dynamical Systems

Developments in the Computation of Reduced Order Models with the Use of Dominant Spectral Zeros



Francisco Damasceno Freitas, Joost Rommes, and Nelson Martins

Abstract In this report we present a study on the computation of dominant spectral-zeros (DSZ) of a full order model (FOM) and its application to determining a passivity-preserving reduced order model (ROM). The study demonstrates that introducing a properly scaled transmission term into a Hamiltonian system allows one to more effectively compute the DSZs by using as their initial estimates in the iterative process, the poles of the original FOM. As the original dynamical system has half the dimension of the Hamiltonian system used for computing the DSZs, this strategy may speed up the overall computation by an iterative algorithm, such as the subspace accelerated dominant pole algorithm (SADPA). Additionally, we introduce alternative expressions for the input and output matrices in the Hamiltonian system which do not depend on the matrix D (avoiding its inverse) associated with the direct transmission term of the output signal and assess the DSZ computational gains achieved. Numerical experiments are shown for three test-systems, including a 4028-state model.

Keywords Spectral zeros · Reduced order model · Eigenvalue · SADPA · Hamiltonian · Power system model

1 Introduction

Preservation of passivity is a key requirement in model order reduction methods [1–3]. In [4] a new passivity preserving model order reduction (MOR) method based on spectral zero interpolation was introduced. Reference [5] then showed that the

F. D. Freitas (✉)

Department of Electrical Engineering, University of Brasilia, Brasilia-DF, Brazil

e-mail: ffreitas@ene.unb.br

J. Rommes

Siemens, Wilsonville, OR, USA

N. Martins

Electrical Energy Research Center - CEPEL, Rio de Janeiro-RJ, Brazil

e-mail: nelson.martins@ieee.org

© Springer Nature Switzerland AG 2022

C. Beattie et al. (eds.), *Realization and Model Reduction of Dynamical Systems*,

https://doi.org/10.1007/978-3-030-95157-3_12

spectral zero (SZ) interpolation problem can be formulated as an associated Hamiltonian eigenvalue problem. Results involving dominant spectral zero interpolation for large-scale systems were later obtained through the computation of the dominant modes of the Hamiltonian system [6, 7]. Other SZ solution variants were proposed for dealing with large-scale systems [8, 9]. The preservation of dissipativity in reduced order model (ROM) methods, for example, was further explored in [10].

Modal approximation methods [11, 12] are not necessarily passive and stable since they only match dominant poles, but a MOR method based on dominant SZ (DSZ) of the system ensures the preservation of passivity [4, 7].

This work investigates strategies to compute the DSZs as the dominant poles of the Hamiltonian system associated with the FOM, which may then be computed by SADPA [11]. We study three alternative Hamiltonian system formulations for the Hamiltonian systems model, one of which has its expressions for the input and output matrices independent of the direct transmission term D . Three strategies for the efficient computation of DSZs (re)using the dominant pole results for $H(s)$ obtained through SADPA iterations were compared for the test systems. The stable DSZs of the FOM are computed from the Hamiltonian system and their respective right eigenvectors are used to determine the transformation matrices employed to calculate the desired passive ROM. Experiments are described for three test-systems, including a 4028-state model. The main contributions of this work are:

- the proposal of a new Hamiltonian alternative formulation, whose input and output matrices are independent of the direct transmission D term;
- the selection of dominant poles of the original FOM as initial estimates for computing the DSZs based on one of the Hamiltonian system alternatives;
- experiments to determine the sensitivity of the SZs to the introduction of a properly scaled transmission term D and the determination of the loci of these SZs. Using this properly scaled parameter, stable DSZs are more cost-effectively computed and a passive ROM is found.

The following notation is used: the complex plane is denoted by \mathbb{C} , the open left half-plane (LHP) by \mathbb{C}^- and $j = \sqrt{-1}$. Symbols $\mathbb{R}^{n \times m}$ and $\mathbb{C}^{n \times m}$ denote $n \times m$ real and complex matrices, respectively. The matrix A^T stands for the transpose of $A \in \mathbb{R}^{n \times m}$, A^* denotes the complex conjugate and transpose of $A \in \mathbb{C}^{n \times m}$, $A^{-*} = (A^{-1})^*$, and $A^{-T} = (A^{-1})^T = (A^T)^{-1}$. The $n \times n$ identity matrix n is denoted by I_n or simply I . The term $D_\alpha = D + \alpha I_m$ is defined and applied only in the formulation of an Hamiltonian system, for a D term of the dynamical system and a scaling factor $\alpha > 0$. m is the number of inputs and outputs for a square system.

This manuscript is structured as follows. In Sect. 2 we describe how spectral zeros can be used to build passive ROMs. In Sect. 3, three alternative Hamiltonian system formulations are studied for the computation of DSZs. Numerical experiments are shown in Sects. 4–6. Section 7 concludes.

2 Spectral Zeros and Model Order Reduction

A linear time invariant (LTI) dynamic system model can be represented by the descriptor system:

$$E\dot{\mathbf{x}}(t) = A\mathbf{x}(t) + B\mathbf{u}(t) \quad (1)$$

$$\mathbf{y}(t) = C\mathbf{x}(t) + D\mathbf{u}(t), \quad (2)$$

where $\mathbf{x}(t) \in \mathbb{R}^n$ is the state vector, $\mathbf{u}(t) \in \mathbb{R}^m$ the input vector and $\mathbf{y}(t) \in \mathbb{R}^q$ the output vector; $A \in \mathbb{R}^{n \times n}$, $E \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $C \in \mathbb{R}^{q \times n}$ and $D \in \mathbb{R}^{q \times m}$ are system matrices. The transfer function or frequency response of system (1) is defined as $H(s) = C(sE - A)^{-1}B + D$.

The spectral zeros of $H(s)$ are defined in [4, 5] as the poles of the rational function

$$G(s) = [H(s) + H^*(-s)]^{-1} \quad (3)$$

where $H^*(-s)$ is defined as

$$H^*(-s) = B^*[sE^* + A^*]^{-1}(-C^*) + D^* \quad (4)$$

The desired spectral zeros of the original system $H(s)$ can be calculated as the dominant poles of $G(s)$, using an interpretation for poles of an inverse system as presented in [13].

The inverse system can be obtained by initially defining a fictitious dynamical system $H_z(s)$ which is derived from the original system model $H(s)$ [7]

$$H_z(s) = H(s) + H^*(-s) = C_z[sE_z - A_z]^{-1}B_z + D_z \quad (5)$$

where

$$A_z = \begin{bmatrix} A & 0 \\ 0 & -A^* \end{bmatrix}, B_z = \begin{bmatrix} B \\ -C^* \end{bmatrix}, E_z = \begin{bmatrix} E & 0 \\ 0 & E^* \end{bmatrix}, C_z = [C \ B^*], D_z = D + D^* \quad (6)$$

The inverse system $G(s) = H_z^{-1}(s)$ has the representation:

$$E_s\dot{\mathbf{x}}_s(t) = A_s\mathbf{x}_s(t) + B_s\mathbf{u}_s(t) \quad (7)$$

$$\mathbf{y}_s(t) = C_s\mathbf{x}_s(t) + D_s\mathbf{u}_s(t) \quad (8)$$

where $A_s = A_z - B_zD_z^{-1}C_z$, $E_s = E_z$, $B_s = B_zD_z^{-1}$, $C_s = -D_z^{-1}C_z$, $D_s = D_z^{-1}$.

Assuming D_z is nonsingular, (7)–(8) can be put as the Hamiltonian system Σ_s :

$$E_s\dot{\mathbf{x}}_s(t) = \begin{bmatrix} A - BD_z^{-1}C & -BD_z^{-1}B^* \\ C^*D_z^{-1}C & -A^* + C^*D_z^{-1}B^* \end{bmatrix} \mathbf{x}_s(t) + \begin{bmatrix} B \\ -C^* \end{bmatrix} D_z^{-1} \mathbf{u}_s(t) \quad (9)$$

$$\mathbf{y}_s(t) = -D_z^{-1} [C \ B^*] \mathbf{x}_s(t) + D_z^{-1} \mathbf{u}_s(t) \quad (10)$$

A structured (augmented) Hamiltonian system Σ_h can be defined [7] by using the Schur complement of $-D_z$:

$$\begin{bmatrix} E & 0 & 0 \\ 0 & E^* & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\mathbf{x}}_h(t) \\ \dot{\mathbf{v}}_h(t) \end{bmatrix} = \begin{bmatrix} A & 0 & B \\ 0 & -A^* & -C^* \\ C & B^* & D_z \end{bmatrix} \begin{bmatrix} \mathbf{x}_h(t) \\ \mathbf{v}_h(t) \end{bmatrix} + \begin{bmatrix} B \\ -C^* \\ 0 \end{bmatrix} D_z^{-1} \mathbf{u}_h(t)$$

Thus

$$E_h \begin{bmatrix} \dot{\mathbf{x}}_h(t) \\ \dot{\mathbf{v}}_h(t) \end{bmatrix} = A_h \begin{bmatrix} \mathbf{x}_h(t) \\ \mathbf{v}_h(t) \end{bmatrix} + B_h \mathbf{u}_h(t) \quad (11)$$

$$\mathbf{y}_h(t) = -D_z^{-1} [C \ B^* \ 0] \begin{bmatrix} \mathbf{x}_h(t) \\ \mathbf{v}_h(t) \end{bmatrix} + D_z^{-1} \mathbf{u}_h(t) = C_h \begin{bmatrix} \mathbf{x}_h(t) \\ \mathbf{v}_h(t) \end{bmatrix} + D_h \mathbf{u}_h(t) \quad (12)$$

where $D_h = D_z^{-1}$.

To compute a passive ROM for the dynamical system (1)–(2), the transformation matrices V and W are determined based on the interpolation of SZ [7]:

$$\hat{E} \hat{\dot{\mathbf{x}}}(t) = \hat{A} \hat{\mathbf{x}}(t) + \hat{B} \mathbf{u}(t) \quad (13)$$

$$\hat{\mathbf{y}}(t) = \hat{C} \mathbf{x}(t) + \hat{D} \mathbf{u}(t) \quad (14)$$

where $\hat{E} = W^* E V$, $\hat{A} = W^* A V$, $\hat{B} = W^* B$ and $\hat{C} = C V$.

In (14), the matrix $\hat{D} = D$ is invariant to the application of the transformation matrices V and W . In [7], the matrices V and W are built by using the right eigenvectors associated with the stable dominant poles of the structured Hamiltonian system (11)–(12). This involves solving the following Hamiltonian eigenvalue problem:

$$A_h R = E_h R \Lambda \quad (15)$$

where Λ is a diagonal matrix with the poles of the pencil (A_h, E_h) (spectral zeros of the dynamical system $H(s)$) and R the right eigenvector matrix associated with Λ .

The stable SZs in Λ and its associated right eigenvectors are next partitioned into stable set Λ_- , anti-stable (unstable) set, Λ_+ and infinite set Λ_∞ . The eigenvectors R are next partitioned as $R = [R_-, R_+, R_\infty]$. Considering only the eigenvalues Λ_- , their eigenvectors R_- , can also be partitioned as

$$R_- = \begin{bmatrix} X_- \\ Y_- \\ Z_- \end{bmatrix} \quad (16)$$

According to [7], if k stable DSZs and their respective right eigenvectors are selected, the matrices V and W can be computed as

$$V = X_-(:, 1 : k) \text{ and } W = Y_-(:, 1 : k) \quad (17)$$

Alternatively, [7] also suggests to build V and W as a rational Krylov subspace:

$$V = [(s_1 E - A)^{-1} B, \dots, (s_k E - A)^{-1} B] \quad (18)$$

$$W = [(-s_1^* E^* - A^*)^{-1} C^*, \dots, (-s_k^* E^* - A^*)^{-1} C^*] \quad (19)$$

The ROM $(\hat{E}, \hat{A}, \hat{B}, \hat{C}, \hat{D})$ interpolates the system (E, A, B, C, D) at the selected SZs s_i and their mirror images $-s_i^*$, $i = 1, 2, \dots, k$. The projection matrices V and W determined from (17) or by (18)–(19) ensure that the reduced system is passive [4, 5].

According to [7], the projection matrices V and W extracted from the information of SZs are sufficient to compute a reduced model as defined by (1)–(2), in case of $D = 0$ and $H(s)|_{s \rightarrow \infty} = 0$. However, when $D = 0$, but $H(s)|_{s \rightarrow \infty} = \tilde{D} \neq 0$, the matrices \hat{A}, \hat{B} and \hat{C} need corrections to include the component Z_- of the right eigenvector R calculated in (16) and also the \tilde{D} term. For this condition, the DSZ reduced model is represented by the following matrices [7]:

$$\hat{E} = W^* E V, \hat{A} = W^* A V + Z^* \tilde{D} Z, \hat{B} = W^* B - Z^* \tilde{D}, \hat{C} = C V - \tilde{D} Z, \hat{D} = \tilde{D} \quad (20)$$

For simplicity, consider a SISO system and note that (20) applies for the case when $D = 0$ but may be also extended for the case where $D \neq 0$. This is achieved by modifying the output signal in (2) as follows. First, define a new algebraic variable $z = Cx + Du$ and make $y = z$. Secondly, add the algebraic equation $0 = -z + Cx + Du$ to the original descriptor system. Therefore, an augmented dynamical system $(E, A, B, C, 0)$ is created having an additional generalized state z and another additional algebraic equation. Note that the ROM for this augmented system is also given by (20). A similar procedure is applied for a MIMO square system with $D \neq 0$, m inputs and m outputs.

3 Variant for Computing Dominant Spectral Zeros

We now describe an alternative formulation for the structured Hamiltonian system whose component matrices are independent of the direct transmission term D , as verified for the matrices B_h , C_h and D_h , an advantage over the formulation presented in (11)–(12). This development is inspired on the method for the computation of transmission zeros as described in [13]. Instead of manipulating with the transfer

matrix $H_z(s)$, in (5), we use the matrix $G(s) \in \mathbb{C}^{q \times m}$ directly. For simplicity, we will only describe the square system case where $q = m$.

Consider a MIMO system defined by $G(s)$ as in (3), such that the input/output relationship is given by $Y_w(s) = G(s)U_w(s)$:

$$Y_w(s) = [H(s) + H^*(-s)]^{-1}U_w(s) \quad (21)$$

Then, for $G(s)$ nonsingular,

$$[H(s) + H^*(-s)]Y_w(s) = U_w(s) \quad (22)$$

and

$$[C(sE - A)^{-1}B + D + B^*(-sE^* - A^*)^{-1}C^* + D^*]Y_w(s) = U_w(s) \quad (23)$$

Consider in (23) that $X_{w1}(s) = (sE - A)^{-1}BY_w(s)$ and $X_{w2}(s) = (sE^* + A^*)^{-1}(-C^*)Y_w(s)$. Therefore, in the time domain, (23) can be written as

$$0 = C\mathbf{x}_{w1}(t) + B^*\mathbf{x}_{w2}(t) + (D + D^*)\mathbf{y}_w(t) - \mathbf{u}_w(t) \quad (24)$$

The expressions for $X_{w1}(s)$ and $X_{w2}(s)$ in the time domain are:

$$E\dot{\mathbf{x}}_{w1}(t) = A\mathbf{x}_{w1}(t) + B\mathbf{y}_w(t) \quad (25)$$

$$E^*\dot{\mathbf{x}}_{w2}(t) = -A^*\mathbf{x}_{w2}(t) - C^*\mathbf{y}_w(t) \quad (26)$$

The expressions (24)–(26) can be put into the Hamiltonian system matrix form, Σ_a :

$$\begin{bmatrix} E & 0 & 0 \\ 0 & E^* & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\mathbf{x}}_{w1}(t) \\ \dot{\mathbf{x}}_{w2}(t) \\ \dot{\mathbf{y}}_w(t) \end{bmatrix} = \begin{bmatrix} A & 0 & B \\ 0 & -A^* & -C^* \\ C & B^* & D_z \end{bmatrix} \begin{bmatrix} \mathbf{x}_{w1}(t) \\ \mathbf{x}_{w2}(t) \\ \mathbf{y}_w(t) \end{bmatrix} + [0 \ 0 \ -I_m]^* \mathbf{u}_w(t)$$

$$\Rightarrow E_a \dot{\mathbf{x}}_a(t) = A_a \mathbf{x}_a(t) + B_a \mathbf{u}_w(t) \quad (27)$$

$$\mathbf{y}_w(t) = [0 \ 0 \ I_m] \begin{bmatrix} \mathbf{x}_w(t) \\ \mathbf{y}_w(t) \end{bmatrix} \Rightarrow \mathbf{y}_w(t) = C_a \mathbf{x}_a(t) \quad (28)$$

where $\mathbf{x}_w(t) = [\mathbf{x}_{w1}^*(t) \ \mathbf{x}_{w2}^*(t)]^*$, $\mathbf{x}_a(t) = [\mathbf{x}_w^*(t) \ \mathbf{y}_w^*(t)]^*$

In (27)–(28), the dependence of D is exclusively on the state matrix A_a . Furthermore, $A_a = A_h$ and $E_a = E_h$, so both system models (Σ_h and Σ_a) have the same poles. However, when using the Hamiltonian system to calculate its dominant poles (dominant spectral zeros of $H(s)$) by a method such as SADPA, there is a depen-

dence of B_h and C_h with the inverse of $(D + D^*)$, which is not possible in the case of $D=0$. This difficulty is obviated in the approach proposed in (27)–(28), since the matrices B_a and C_a are insensitive to values of D . On the other hand, even when the modeling is originally presented with a null D matrix, it is possible to obtain an equivalent modeling with D nonsingular through appropriate modification on the circuit equations and variables (refer to the example for an electrical circuit described in Sect. 5, where $H(s)$ has equivalent descriptions: depending on D or not).

The poles of the Hamiltonian system modeled by (7)–(8), (11)–(12) or (27)–(28) change when varying the term D_z defined in (6), while the poles of $H(s)$ remain unchanged. This idea which was originally explored in [7] is used here to modify D_z in order to shift the SZs towards the poles of $H(s)$. This strategy has proven effective since better initial estimates lead to fewer iterations in the SADPA computation of DSZs. Therefore, we define D_z as a function of a scaling factor $\alpha > 0$, such that $D_z = D_\alpha + D_\alpha^*$, where $D_\alpha = D + \alpha I_m$ and m is the number of inputs, which is assumed to be equal to the number of outputs (see Sects. 4 and 5 for computation of the SZs as a function of α).

A procedure to determine a ROM based on DSZs is presented in the Alg. 1. The procedure adopts the dominant poles of $H(s)$ as the initial estimates in the iterative computation of DSZs.

Algorithm 1 Procedure to compute the DSZ-ROM.

Require: Dynamical system representation (E, A, B, C, D) .

Ensure: DSZ-ROM matrices $(\hat{E}, \hat{A}, \hat{B}, \hat{C}, \hat{D})$

- 1: Start computing the N_d dominant poles s_k of the dynamical system $H(s)$ by using SADPA [11] and use them as initial estimates for computing the DSZs.
 - 2: Establish a scaling factor α , define $D_\alpha = D + \alpha I_m$, compute $D_z = D_\alpha + D_\alpha^*$ and form the Hamiltonian system, according to the alternative defined by (7)–(8), (11)–(12) or (27)–(28).
 - 3: $Z = []$; $R = []$;
 - 4: **for** $k = 1, 2, \dots, N_d$ **do**
 - 5: Set $z_0 = s_k$ and compute by SADPA the dominant SZ z_k in the neighborhood of s_k and its respective right eigenvector r_k .
 - 6: Store the DSZ z_k in $Z = [Z; z_k]$ and the right eigenvector r_k in $R = [R \ r_k]$.
 - 7: **end for**
 - 8: Compute the transformation matrices V and W from the right eigenvector matrix R , according to the procedure described in the Sect. 2.
 - 9: Compute the DSZ-ROM matrices $\hat{E}, \hat{A}, \hat{B}, \hat{C}, \hat{D}$ defined in (20).
-

The following sections details examples on the numerical computations of DSZs and DSZ-ROM when considering the approaches presented in Sects. 2 and 3.

4 The Impact of D Term Size on DSZ Numerical Computations

The structured Hamiltonian system, described by (11)–(12) or (27)–(28), is very suitable for calculating DSZs of $H(s)$ via SADPA [11]. The impact of the D term size on the SZs is investigated in this section. We point out the importance of choosing an optimum size for this term for determining better initial estimates for the DSZ computation.

This experiment illustrates the impact of α in the D_α term in driving the DSZs of $G(s)$ toward the poles of $H(s)$. The tutorial system is defined by the following matrices:

$$A = \begin{bmatrix} -1 & 4 & -1 & 1 \\ -4 & -1 & 0 & -1 \\ -1 & -1 & -2 & 8 \\ 1 & 2 & -8 & -2 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & -1 \end{bmatrix}, \quad C^T = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},$$

with $D_\alpha = D + \alpha I_2$, where α is a positive scaling factor for the I_2 term.

Figures 1 and 2 illustrate the poles and locus of the SZs as a function of the scaling factor α . The black arrows were inserted in the SZ locus plot to indicate how the spectral zeros shift as α is reduced. Figure 1 shows the SZs loci when α is varied from 1 to 0.1. The poles of the dynamical system $H(s)$ are: $-1.2426 \pm j4.06$ and $-1.7574 \pm j7.95$. The stable SZs computed for $\alpha = 1$ are $-1.4711 \pm j4.02$ and $-1.4778 \pm j7.98$, which are very much in the neighborhood of the poles of $H(s)$. On the other hand, when $\alpha = 3.5 \times 10^{-8}$ the stables SZs are $-3.9379, -6.2374, (-6.76 \pm j8.59) \times 10^3$, being therefore far away from the poles of $H(s)$. The SZs are seen to drastically shift as the scaling factor α is changed. Smaller values of α lead the stable SZ previously located near the pole of $H(s)$, $-1.7574 \pm j7.95$, to move

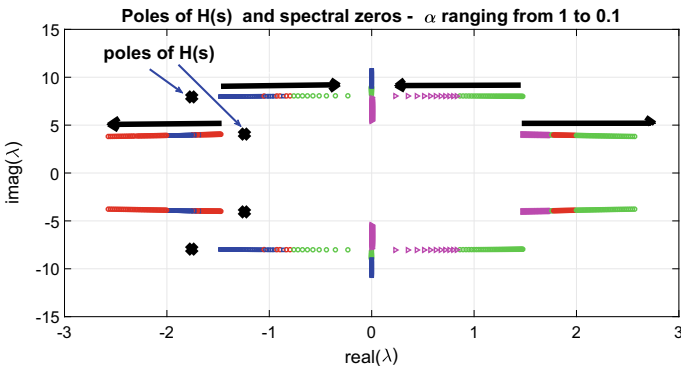


Fig. 1 Loci of poles of $H(s)$ and its spectral zeros as a function of the scaling factor α for $D = \alpha I_2$. The black arrows point in the direction of decreasing α

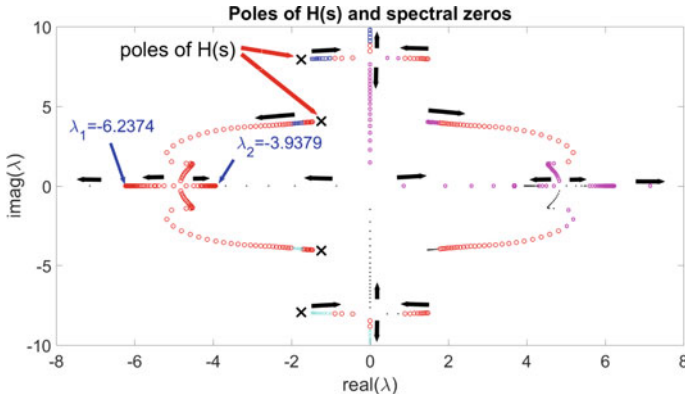


Fig. 2 Poles of $H(s)$ and spectral zeros locus as a function of the scaling factor α . The scaling factor α has value 1 in the neighborhood of the poles of $H(s)$. At this point in the plot, the SZs are $-1.4778 \pm j7.98$ and $-1.4711 \pm j4.02$. The SZs λ_1 and λ_2 in the plot occur for $\alpha = 3.5 \times 10^{-8}$

into the imaginary axis. The stable SZ near the other pole, $-1.2426 \pm j4.06$, moves further into the left hand side of the complex plane. The drastic shift experienced by the SZs when decreasing α up to very smaller values are illustrated in Fig. 2. This plot highlights stable SZs $\lambda_1 = -6.2374$ and $\lambda_2 = -3.9379$ for $\alpha = 3.5 \times 10^{-8}$, while other 4 SZs are on the imaginary axis.

Therefore, for the same dynamical system, higher values of the scaling factor α , leads to stable SZs in the vicinity of the poles of $H(s)$. On the other hand, very small values of α push the SZs away from the poles of $H(s)$.

5 Electrical Circuits

5.1 Dynamic Modeling

Figure 3 depicts an electrical circuit having only passive elements (ideal resistor, inductor and capacitor devices) and comprised of 2 RLC PI sections. Each PI section has a resistance R , an inductance L and a capacitance $C/2$ at each terminal. In the example, three nodal voltages v_1, v_2 and v_3 are highlighted. These are the same voltages across each capacitor v_{C1}, v_{C2} and v_{C3} , respectively.

The input signal is the voltage source connected to node 1, while the output signal is the current i_3 flowing through the RL load.

The equations for this circuit can be formulated using the physical laws for each device and Kirchhoff’s current and voltage laws, yielding a single-input single-output descriptor system having 6 states and 3 algebraic variables.

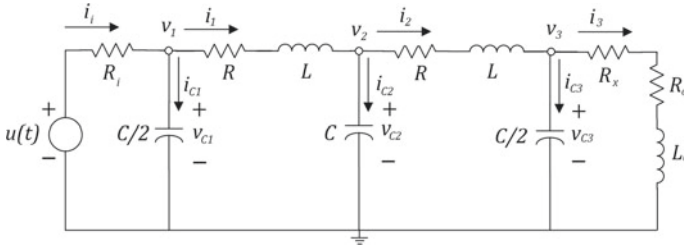


Fig. 3 3-node and 2-PI circuit with a voltage source $u(t)$ and a RL load

We show that the output variable can be expressed either as a function of a single state, without a term $Du(t)$ or as a combination of state plus the forward term $Du(t)$. In some cases, therefore it is necessary to deal with dynamical system models which have a nonzero D [3].

The vector of generalized states for this system can be defined as

$$\mathbf{x} = [i_{C1} \ i_{C2} \ i_{C3} \ i_1 \ i_2 \ i_3 \ v_1 \ v_2 \ v_3]^T \tag{29}$$

The output variable, $y = i_3$, is therefore given by

$$y = [0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0]\mathbf{x} = C_1\mathbf{x} \tag{30}$$

Alternatively, the same output y can be expressed as

$$y = [-1 \ -1 \ -1 \ 0 \ 0 \ 0 \ -\frac{1}{R_i} \ 0 \ 0]\mathbf{x} + \frac{1}{R_i}u = C_2\mathbf{x} + \frac{1}{R_i}u. \tag{31}$$

In (30) the term D is zero, while in (31) $D = 1/R_i$. Both yield numerically the same results for the transfer function $H(s) = y(s)/u(s)$. In the sequel, we present experiments for these two forms of output representations, demonstrating their equivalence for the computation of DSZ-ROM.

5.2 Results of Experiments

Experiments were initially carried out for test systems generated by increasing the number of PI sections in the circuit depicted in Fig. 3. The circuit parameters independent of the number of PI sections are: $C = 0.01$ F, $R = 0.05 \ \Omega$, $L = 0.02$ H, $R_o = 0.05 \ \Omega$, $L_o = 0.06$ H, $R_x = 0$ and $R_i = 0.05 \ \Omega$.

5.2.1 Experiments on the Circuit with 2 PI Sections

The purpose of the experiment is to evaluate the two alternative modelings for the same output variable in the electrical circuit of the Fig.3 considering the model with the term $D = 0$ as Alternative I and the model with a nonzero D term as the Alternative II.

The circuit with 2 PI sections ($N = 2$) has 9 generalized states, whose eigenvalues were computed by using a QZ routine: -3.9975×10^3 , $-1.2870 \pm j1.3759 \times 10^2$, $-1.8801 \pm j6.6316 \times 10^1$, -2.0015 , Inf , Inf , and Inf .

Alternative I has the output represented by (30) while in Alternative II, it is represented by (31). Then considering the input vector B and the output vectors C_1 (Alternative I) and C_2 (Alternative II), the spectral zeros were computed by using the QZ routine applied to the structured Hamiltonian system Σ_h , with pencil (A_h, E_h) , for both Alternatives I and II. The Hamiltonian system has dimension 19 and its finite poles, which are also the spectral zeros of the system (E, A, B, C, D) , for Alternative I are: $7.17 \times 10^4 \pm j5.26 \times 10^4$, -9.04×10^4 , $-2.85 \times 10^4 \pm j8.52 \times 10^4$, $\pm j2.07 \times 10^2$, $\pm j1.10 \times 10^2$ and $\pm j3.54 \times 10^1$. For Alternative II the SZs lying on the imaginary axis were similar to the SZs found for Alternative I. We can therefore conclude that an alternative modeling of the output variable $y = -i_{C1} - i_{C2} - i_{C3} - \frac{v_1}{R_1} + \frac{u}{R_1}$ (see (31)), with a nonzero D term in $H(s)$, does not cause significant change on the SZs, when compared with the modeling by the one state variable $y = i_3$, i.e., with $D = 0$. Moreover, we can conclude for this example that even for the smaller finite spectral zeros, they are far away from the poles of the dynamical system (E, A, B, C, D) .

The experiments on the tutorial system of Sect.4, suggested that by increasing the norm of the matrix $D_\alpha = D + \alpha I_m$ used in the Hamiltonian systems, the spectral zeros move towards the system poles. Similar results were obtained for the electrical circuit model of this section. In [7] it is proposed to raise the norm of $D_\alpha + D_\alpha^*$ by a factor δ in the range $(10^4, 10^6)$, similarly to what has been presented in Sect.4 so as to locate the SZs to near the poles of $H(s)$. An even more important criterion is to judiciously raise the D norm in order to shift the DSZs only up to a point where there is no SZ left on the imaginary axis. Both alternative system descriptions are useful for determining the appropriate D norm and initial estimates for the iterative computation of DSZs. The D norm choice is based on the control system's intuition that having MORs generated from DSZs that are close to the actual system poles may better capture the actual system dynamics.

Therefore, we now use D_α in (11)–(12) and (27)–(28), defined accordingly for the selected Alternatives I or II.

Figure4 highlights the shifting of the SZ when α is varied from 0 to 5, in steps of 0.5 for Alternative I, i.e., $D = 0$. The SZ locus is depicted only for the left-side of the complex plane. Similar mirrored movement (not shown) is observed for the right-side of the complex plane. The plot also shows how the SZ shift towards the poles of the system (E, A, B, C, D) , as α is further increased. The finite poles of the dynamical system are $\lambda_1, \lambda_1^*, \lambda_2, \lambda_2^*, \lambda_3, \lambda_4$. In the situation when $\alpha = 0$, there are 6 finite spectral zeros at the imaginary axis (see red dots on the imaginary axis). When α

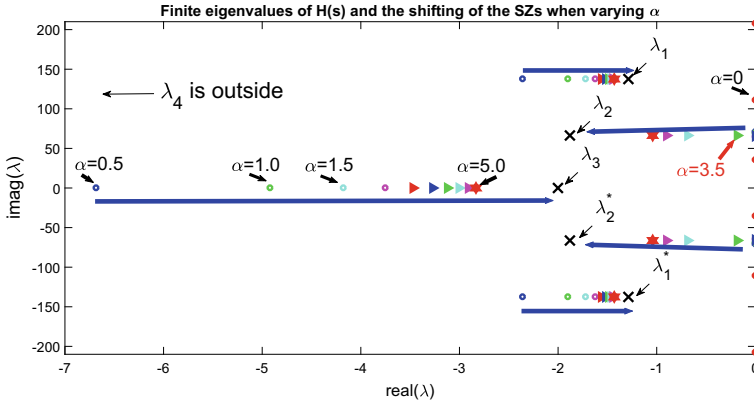


Fig. 4 Fixed poles and spectral zeros loci for values of α changing from 0 to 5, Alternative I

is increased just to 0.5, the SZ are: ± 3997 , ± 6.68 , $\pm j61.7$, $\pm j70.8$, $\pm 2.36 \pm j137.6$ and 7 other SZ at infinity. So, the four SZs $\pm j61.7$, $\pm j70.8$ were the only ones that have remained on the imaginary axis. When $\alpha = 5$ the SZs are already close to the poles of the system (E, A, B, C, D). And, higher values of α would shift the SZs to even closer to these poles. This consistent behavior among dynamic systems of different nature justifies using the poles of $H(s)$ as initial estimates to compute iteratively the dominant SZs, for example, by using SADPA [11].

The dominant poles of $H(s)$ were determined by using SADPA for two alternative equations for the output variable discussed in Subsection 5.1. Both alternatives relate to the same transfer function. Even though Alternative II has a nonzero term $\frac{1}{R_i}u(t)$, this fact does not ensure fully stable SZs. Our strategy has been to choose an adequate norm for the term D_α in the Hamiltonian system so that it only results in stable SZs. An experiment for single-input single-output model with $D_\alpha = D + \alpha$, $\alpha = 10$, was carried out and selected stable DSZs were used to compute a ROM. Both the Hamiltonian system defined by Σ_h in (11)–(12) and Σ_a in (27)–(28) were used to compute the SZs. To evaluate modal dominance on the Hamiltonian system, the SADPA selection strategy 'LR' (largest ratio $|\text{res}|/|\text{re}(\lambda)|$) was chosen. A first type of experiment uses 3 initial guess defined from the dominant poles of $H(s)$: $s_0 = \{-2.0015, -1.2870 + j1.3759 \times 10^2, -1.8801 + j6.6316 \times 10^1\}$ to determine 3 wanted DSZs. Here, no deflation is implemented by SADPA, since each time a DSZ is found, another guess is used from the set formed by s_0 . The dominant poles s_0 of $H(s)$, on the other hand, were calculated by SADPA from a single initial guess $s_0 = j100$. A second experiment was based on the computation of the dominant SZs directly from the Hamiltonian systems, starting with a initial guess $s_0 = j100$ with the purpose of determining 3 wanted DSZs. In this strategy, SADPA deflates the found pole and uses it as the guess for searching the next DSZ, proceeding this way until nwanted poles are determined.

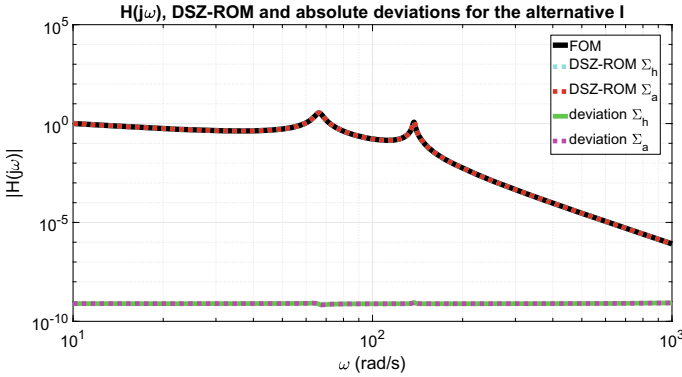


Fig. 5 Magnitude of $H(j\omega)$ of the full order model (FOM) (Alternative I), DSZ-ROMs of $H(s)$ obtained for the modeling Alternative I and calculated by using Σ_h and Σ_a . The transformation matrices associated with DSZs were computed for a scaling factor $\alpha = 10$

The reduced model computed by using the modal dominance of SZs for Alternative I is $(\hat{E}, \hat{A}, \hat{B}, \hat{C}, \hat{D})$. The term $\hat{D} = H(s)|_{s \rightarrow \infty} = 0$, giving $\hat{D} = \tilde{D} + D = 0$. Figure 5 shows the magnitude plot for the model transfer functions, for Alternative I FOM and DSZ-ROMs obtained when the Hamiltonian systems Σ_h and Σ_a are used. Also, the absolute deviations between the transfer functions of the FOM and DSZ-ROMs for both Hamiltonian systems are shown. The transformation matrices V and W were calculated from right eigenvectors associated with the DSZs obtained for the case with an scaling factor $\alpha = 10$.

Consider now the system output defined according to Alternative II, which has $D = 1/R_i$ and C_2 as in (31) to define the output signal y . Defining $D_\alpha = 1/R_i + \alpha$ and assigning $\alpha = 0$, six SZs are found at the imaginary axis. This pattern also was detected in the Alternative I. i.e., the fact of using another representation of the output y , with a nonzero D term does not change the characteristic of the SZs in relation to the case when $D = 0$. Then, in order to adopt a uniform standard for the calculation of the SZs, we propose that all output signals be represented as D null. For purpose of the SZ computation, we use $D_\alpha = D + \alpha I$. For this aim, in the situation when y is defined by an equation that has a nonzero D , as in the case of Alternative II, it is suggested to redefine y by an additional generalized state, as demonstrated in Sect. 3, adding 1 algebraic variable and 1 algebraic equation in the original dynamical system. Then, assigning $\alpha = 10$ in both the Hamiltonian systems Σ_h and Σ_a , we found similar results for the DSZ-ROMs, as found when Alternative I was selected for modeling.

SADPA was used to determine the dominant poles of the system $H(s)$ for the two alternative modelings I and II. Although the poles are only associated with the pencil (E, A) , the SADPA uses the other matrices (B, C, D) to guide the search for the dominant poles of $H(s)$. The following 5 dominant poles of $H(s)$ were found: $-1.2870 \pm j137.59$, $-1.8801 \pm j66.316$, -2.0015 . Next, only the pole located far

away from the imaginary axis was discarded, the other 4 poles being used as initial estimates for computing the DSZs. SADPA was applied sequentially to calculate one dominant spectral zero at a time, by using as initial estimate the dominant pole of the system (E, A, B, C, D) in a procedure that can be easily automated. The computation of the DSZ were carried out for the structured Hamiltonian system Σ_h , as defined in (11)–(12) and Σ_a , as defined in (27)–(28).

For Alternative I and $\alpha = 10$ the following DSZs were calculated: -2.4521 , $-1.5194 \pm j66.313$ and $-1.3607 \pm j137.59$. Their right eigenvectors R_- (see (16)) were also calculated, generating the matrices V and W according to (17). Then, with these matrices the DSZ reduced model (DSZ-ROM) for Alternative I was calculated. The DSZ-ROM obtained has the following poles -2.0015 , $-1.8801 \pm j66.316$ and $-1.2870 \pm j137.59$. This set of dominant poles to build the reduced model is approximately equal to the dominant poles of the original system $H(s)$.

5.3 Experiments on Circuits with Many PI Sections

The results in this section are related to an analogous electrical circuit model with 150 PI sections having 450 generalized states, and the structured Hamiltonian has 901 generalized states.

SADPA initially was used to determine the dominant poles of the $H(s)$ dynamical system, with $s_0 = j100$ as the initial estimate and the number of wanted poles 130 (nwanted) specified to initialize SADPA, which automatically computed 239 dominant (complex conjugate) poles for $H(s)$. These dominant poles were then used as estimates for computing the spectral zeros via SADPA again, but now applied to the structured Hamiltonian system, both for (11)–(12) and (27)–(28). A total of 239 DSZs were found in both SADPA runs. These DSZs were used to compute the transformation matrices V and W according to (18)–(19). Finally, a 239 order DSZ-ROM was calculated. Figure 6 illustrates the distribution of poles of $H(s)$ as well as DSZs and poles of the ROM computed via the DSZ approach for the Alternative I modeling. Figure 7 exhibits the frequency response for the FOM $H(s)$ and its DSZ-ROM for the alternative I, when the DSZs are computed for $\alpha = 20$. The computation of DSZs by applying both the Hamiltonian system approaches in (11)–(12) and (27)–(28) are seen to present practically identical results. Other experiments were performed for smaller unwanted parameter values such as 90 and 110. The deviation is highest for the case when unwanted is 90, the best result among these three experiments being for unwanted = 130. Evidently, higher values for unwanted improve the quality of the calculated ROM. But, this should be done up to a minimum tolerance for accepting the DSZ-ROM. These results demonstrate the efficacy of SADPA to accurately compute ROMs directly from the Hamiltonian system description.

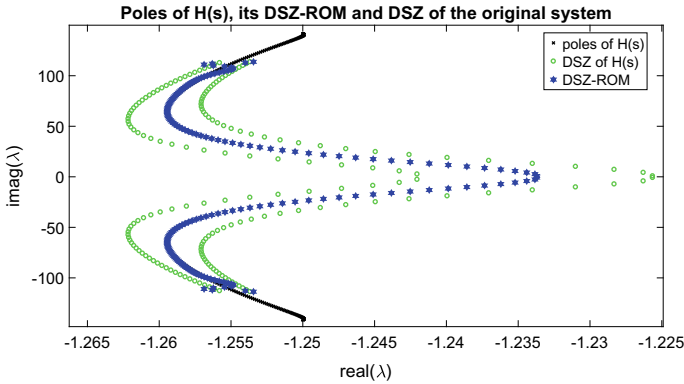


Fig. 6 Poles of $H(s)$ (FOM), poles of the DSZ-ROM and the DSZs for the circuit with 150 PI sections

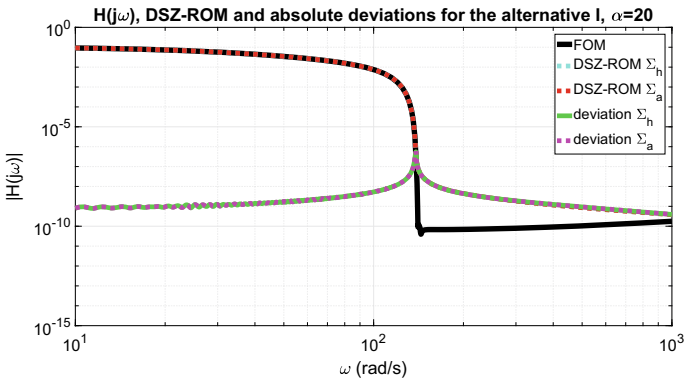


Fig. 7 Frequency responses for the 150 PI sections FOM $H(s)$ and DSZ-ROM for Alternative I, $\alpha = 20$. The Hamiltonian systems Σ_h em Σ_a were used for computing the DSZs

6 Experiments on an Actual Power System Model

This section presents results for an actual power system RLC model [14] suitable for transient studies involving frequencies as high up as 15 kHz (about 10^5 rad/s). The actual system model is stable and passive since all components are represented by lumped RLC circuits. Each transmission line is represented by a cascade of 80 PI circuits while other devices such as transformers and loads are individually represented by a RL series circuit. The complete data on this system is available at [15]. The objective in [14] was to determine reduced models by using the balanced truncation technique and employing low-rank approximations of the Gramians to compute the transformation matrices.

The full model has 4028 states and the objective is to compute its ROM by using the DSZ approaches presented in this work to preserve the properties of stability and

passivity [7]. This section reports on three strategies for computing the DSZ all of which use the basic algorithm or variants of SADPA.

The following considerations are appropriate:

- The input/output matrices of the structured Hamiltonian of a SISO $H(s)$ are built using (27)–(28) for generating a ROM1 and (11)–(12) for generating a ROM2. In both cases, the D term of the original system is zero and a $D_\alpha = \alpha$ term was utilized;
- The computations for determining the DSZs were then performed for $D_\alpha = 10$ despite the fact that in the original system $D = 0$;
- Three strategies for the efficient computation of DSZs (re)using the dominant pole results for $H(s)$ obtained through SADPA iterations were compared:
 1. Guess $H(s)$: Dominant poles of $H(s)$ are computed by SADPA and used as set of initial guesses in the DSZ computation also by SADPA, with an artificially large D_α norm used in the Hamiltonian system.
 2. Hybrid: From a guess s_0 , just one pole of $H(s)$ is computed by SADPA and the result is used to compute the DSZ nearest to this pole by a two-sided RQI algorithm, with an artificially large D_α norm used in the Hamiltonian system. After convergence to the DSZ, the pole of $H(s)$ is deflated and the next dominant pole is determined for computing the next DSZ, again by a two-sided RQI algorithm. The process continues until a wanted number of poles is computed.
 3. DSZ-deflation: DSZ computed by SADPA as dominant poles of the Hamiltonian system modified with an artificially large D norm. Initial estimates chosen by user. Basic SADPA routine deflates the already found DSZ (an Hamiltonian pole) and proceed to find the next DSZ.

Experiments were carried out for the three strategies to compute the DSZs. The following SADPA parameters (see [11] for details on parameters) were used: $K_{min} = 5$, $K_{max} = 20$, $maxrestart = 20$, tolerance of 10^{-10} , and strategy (in this work ‘LR’, $largest|residues|/|real(pole)|$ was used to characterize mode dominance).

Table 1 summarizes experiment results for the three strategies to compute DSZs for both the ROM1 and ROM2 approaches. The ROM1 approach corresponds to the structured Hamiltonian where the output/input matrices are insensitive to D_α (see (27) and (28)), proposed in this work; the ROM2 approach characterizes the study based on the method when the output/input matrices is dependent on D_α (see (11) and (12)). The number of dominant spectral zeros (#DSZ) and the computational cost in CPU time, in seconds, for obtaining a ROM are provided in Table 1.

From Table 1, we can see that all three variants produce ROMs of similar order (#DSZ). However, the best strategy in terms of computational cost is the one based on the first strategy, labeled ‘guess $H(s)$ ’ (see third column in the table), where the dominant poles of $H(s)$ were used as initial estimates for the DSZs computation. The worst run-time performance was obtained for the second variant (see fourth column, labeled ‘hybrid’).

The ROMs and FOMs were compared in the frequency domain for a frequency range from 10 to 10^5 rad/s. Figure 8 illustrates the magnitudes and absolute deviations

Table 1 Results for the MOR experiments for the 4028-state model

| Approach | | Guess $H(s)$ | Hybrid | DSZ-deflation |
|----------|--------------|--------------|--------|---------------|
| ROM1 | #DSZ | 197 | 197 | 193 |
| | Run-time (s) | 20.7 | 79.4 | 33.0 |
| ROM2 | #DSZ | 197 | 197 | 195 |
| | Run-time (s) | 17.9 | 79.1 | 30.0 |

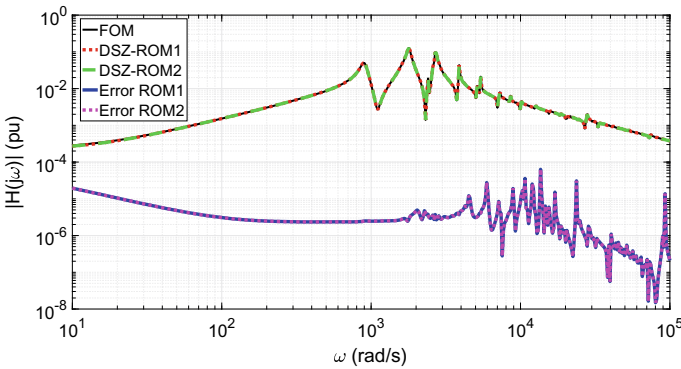


Fig. 8 Frequency response for the TF and the respective deviations between FOM and ROM, when the ‘guess $H(s)$ ’ strategy is implemented on SADPA

between the DSZ-ROM in relation to FOM for a transfer function (TF). The DSZ-ROM was calculated for $\alpha = 10$. The scalar TF relates the bus voltage (output signal) to the current injection (input signal) at the bus #21 in the industrial system model described in [14]. The TF plots for the FOM, ROM1-DSZ and ROM2-DSZ are shown in the figure. The deviations of the ROM results from the FOM are compared in error curves ‘Error ROM1’ and ‘Error ROM2’. The figure illustrates plots related to the strategy characterized by ‘guess $H(s)$ ’ in Table 1. Similar plots were obtained for the strategies ‘hybrid’ and ‘DSZ-deflation’.

When comparing the error curves computed by the three strategies (strategies ‘hybrid’ and ‘DSZ-deflation’ not shown), the first and second strategies ‘guess $H(s)$ ’ and ‘hybrid’ are seen to perform better than the strategy ‘DSZ-deflation’, if the focus is on higher frequencies. For lower frequencies and in the region of larger peaks all strategies have similar performance. We verify that for strategies ‘guess $H(s)$ ’ and ‘hybrid’ the deflation is implemented for the poles of $H(s)$, which are used as estimates for computing a DSZ. In the case of ‘DSZ-deflation’ the deflation occurs directly on the Hamiltonian system. In this case, the approaches ROM1 and ROM2 can converge for another set of DSZ or miss some DSZ. Note that from Table 1, the ROM1 and ROM2 were obtained from 193 DSZs and 195 DSZs, respectively. The other strategies required information from 197 DSZs for either ROM1 or ROM2.

These results confirm the better performance of the ‘guess $H(s)$ ’ strategy, which uses the poles of $H(s)$ as initial estimates in the computation of the DSZ set.

The strategy adopted in this work to modify D_α and select dominant poles as initial estimates for finding DSZs from the modified Hamiltonian system, is applicable for systems having the direct transmission term D equal to zero or different from zero.

In this work, all experiments were performed in Matlab in a notebook AMD Intel Core-TM i7 CPU with 2.5 GHz and 16 GB RAM.

7 Conclusion

This work presented an alternative method for computing dominant spectral zeros (DSZs) as eigenvalues of a Hamiltonian system. A ROM was built taking as foundation the stable modal part of the set of DSZs. The proposed method does not require the matrix D for computing the input and output matrices as needed in other works. The new method was implemented using SADPA and its performance was compared with other approaches whose input and output matrices do depend on D . The ROM has the characteristics of preserving the stability and passivity of the full order model. The study has demonstrated that the location of the DSZs depends heavily on the norm of the matrix D . Three strategies for selecting initial estimates when computing DSZs based on SADPA were described in detail. The efficacy of the computations was demonstrated by performing experiments for three test-systems, including a 4028-state actual power system model. For the test systems considered in this paper, the approach where a set of dominant poles of $H(s)$ is used as initial estimates for the dominant spectral zeros was found to be the most efficient.

Acknowledgements Part of this work was developed during a 3-month sabbatical leave of Prof. Francisco D. Freitas (University of Brasilia, Brazil) at the Electrical Energy Research Center - CEPEL, in 2019, through collaboration with Dr. Nelson Martins, who retired from CEPEL in early 2020, and Dr. Joost Rommes (Siemens, USA). The three above listed institutions and the development team for the CEPEL software HARMZs, for the harmonic analysis of power systems, are therefore deeply thanked by the authors for the provision of means for this research and permission to publish. The authors would like to thank the anonymous referee for the detailed and constructive reviews that helped us improve the paper. The authors would also like to express their appreciation to the editorial team for the invitation to contribute to this special publication and to Mark Embree for handling our manuscript. Last, but not least, the authors would like to thank Thanos for his breakthrough work in model order reduction and several other domains, which has been, and will be, a source of inspiration for the authors and many others.

References

1. Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems. SIAM, USA (2005)
2. Schilders, W.H., van der Vorst, H., Rommes, J.: Model Order Reduction: Theory. Springer, Research Aspects and Applications (2008)

3. Unneland, K., Van Dooren, P., Egeland, O.: A Novel Scheme for Positive Real Balanced Truncation. In: Proceedings of the American Control Conference, July 11-3, 947–952 (2007)
4. Antoulas, A.C.: A new result on passivity preserving model reduction. *Syst. Control Lett.* **54**, 361–374 (2005)
5. Sorensen, D.: Passivity preserving model reduction via interpolation of spectral zeros. *Syst. Control Lett.* **54**, 347–360 (2005)
6. Ionutiu, R.: Passivity preserving model reduction in the context of spectral zero interpolation. Master's thesis, William Marsh Rice University, Houston, TX, USA (2008)
7. Ionutiu, R., Rommes, J., Antoulas, A. C.: Passivity-preserving model reduction using dominant spectral-zero interpolation. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **27**, 2250–2263 (2008)
8. Kumar, S., Belur, M. N., Pal, D.: New results and methods in balancing/spectral-zero-interpolation based model order reduction. In: European Control Conference (ECC), Limassol, Cyprus, pp. 3227–3232 (2018)
9. Massoud, Y., Alam, M., Nieuwoudt, A.: On the Selection of Spectral Zeros for Generating Passive Reduced Order Models. In: The 6th International Workshop on System on Chip for Real Time Applications, pp. 160–164 (2006)
10. Trentelman, H. L., Minh, H. B., Rapisarda, P.: Dissipativity preserving model reduction by retention of trajectories of minimal dissipation. *Math. Control, Signals Syst.(MCSS)*, **21**, 3, 171–201 (2009)
11. Rommes, J., Martins, N.: Efficient computation of transfer function dominant poles using subspace acceleration. *IEEE Trans. Power Syst.* **21**(3), 1218–1226 (2006)
12. Rommes, J., Martins, N.: Efficient computation of multivariable transfer function dominant poles using subspace acceleration. *IEEE Trans. Power Syst.* **21**(4), 1471–1483 (2006)
13. Martins, N., Pellanda, P.C., Rommes, J.: Computation of transfer function dominant zeros with applications to oscillation damping control of large power systems. *IEEE Trans. Power Syst.* **22**, 1657–1664 (2007)
14. Freitas, F.D., Martins, N., Varricchio, S.L., Rommes, J., Veliz, F.C.: Reduced-order transfer matrices from rlc network descriptor models of electric power grids. *IEEE Trans. Power Syst.* **26**(4), 1905–1916 (2011)
15. Rommes, J.: Test systems for the dominant pole algorithm and variants (2019). <https://sites.google.com/site/rommes/software#owners>. Accessed on 19 Aug 2019

Structure-Preserving Interpolatory Model Reduction for Port-Hamiltonian Differential-Algebraic Systems



Christopher Beattie, Serkan Gugercin, and Volker Mehrmann

Dedicated to Thanos Antoulas on the occasion of his 70th birthday

Abstract We examine interpolatory model reduction methods that are particularly well-suited for treating large-scale port-Hamiltonian differential-algebraic systems. We are able to take advantage of underlying structural features of the system in a way that preserves them in the reduced model, using approaches that incorporate regularization and a prudent selection of interpolation data. We focus on linear time-invariant systems and present a systematic treatment of a variety of model classes that include combinations of index-1 and index-2 systems, describing in particular how constraints may be represented in the transfer function so that the polynomial part can be preserved with interpolatory methods. We propose an algorithm to generate effective interpolatory models and illustrate its effectiveness on a numerical example.

Keywords Port-Hamiltonian descriptor system · Model reduction · Tangential interpolation · Regularization of descriptor system · Staircase form

C. Beattie · S. Gugercin
Department of Mathematics, Virginia Tech, Blacksburg 24061-0123, USA
e-mail: beattie@vt.edu

S. Gugercin
e-mail: gugercin@vt.edu

V. Mehrmann (✉)
Institut für Mathematik, MA 4-5, Technische Universität Berlin, 10623 Berlin, Germany
e-mail: mehrmann@math.tu-berlin.de

1 Introduction

Port-Hamiltonian (pH) systems are network-based models that arise within a modeling framework in which a physical system model is decomposed into hierarchies of submodels interconnected principally through the exchange of energy. The submodels typically reflect one of a variety of core modeling paradigms that describe phenomenological aspects of the dynamics having different physical character, such as e.g. electrical, thermodynamic, or mechanical. The pH framework is able to knit together submodels featuring dramatically different physics through a disciplined focus on energy flux as a principal mode of system interconnection. pH structure is inherited via power conserving interconnection and a variety of physical properties and functional constraints (e.g., passivity and energy and momentum conservation) are encoded directly into the structure of the model equations [5, 30]. Interconnection of submodels may create further constraints on system behavior and evolution, originating as conservation laws (e.g., Kirchoff's laws or mass balance), or as position and velocity limitations in mechanical systems. As a result, system models are often naturally posed as combinations of dynamical system equations and algebraic constraint equations, i.e., as *port-Hamiltonian descriptor systems* or *port-Hamiltonian differential-algebraic equations* (pHDAE).

When a pHDAE system is linearized around a stationary solution, one obtains a linear time-invariant pHDAE with specially structured coefficient matrices, see [5]. Although the approach we develop here can be extended easily to more general settings, we narrow our focus to a particular formulation offered as one of the simpler among a variety of formulations presented in [5, 22].

Definition 1 A linear time-invariant DAE system of the form

$$\begin{aligned} \mathbf{E}\dot{\mathbf{x}} &= (\mathbf{J} - \mathbf{R})\mathbf{x} + (\mathbf{B} - \mathbf{P})\mathbf{u}, & \mathbf{x}(t_0) &= \mathbf{0}, \\ \mathbf{y} &= (\mathbf{B} + \mathbf{P})^T\mathbf{x} + (\mathbf{S} + \mathbf{N})\mathbf{u}, \end{aligned} \quad (1)$$

with $\mathbf{E}, \mathbf{J}, \mathbf{R} \in \mathbb{R}^{n \times n}$, $\mathbf{B}, \mathbf{P} \in \mathbb{R}^{n \times m}$, $\mathbf{S} = \mathbf{S}^T$, $\mathbf{N} = -\mathbf{N}^T \in \mathbb{R}^{m \times m}$, on a compact interval $\mathbb{I} \subset \mathbb{R}$, is a pHDAE system if the following properties are satisfied:

1. The differential-algebraic operator

$$\mathbf{E} \frac{d}{dt} - \mathbf{J} : C^1(\mathbb{I}, \mathbb{R}^n) \rightarrow C^0(\mathbb{I}, \mathbb{R}^n)$$

is *skew-adjoint*, i.e., $\mathbf{J}^T = -\mathbf{J}$ and $\mathbf{E} = \mathbf{E}^T$,

2. \mathbf{E} is positive semidefinite, i.e., $\mathbf{E} \geq 0$, and
3. the *passivity* matrix

$$\mathbf{W} = \begin{bmatrix} \mathbf{R} & \mathbf{P} \\ \mathbf{P}^T & \mathbf{S} \end{bmatrix} \in \mathbb{R}^{(n+m) \times (n+m)}$$

is symmetric positive semi-definite, i.e., $\mathbf{W} = \mathbf{W}^T \geq 0$.

The *Hamiltonian function* $\mathcal{H} : \mathbb{R}^n \rightarrow \mathbb{R}$ of the system is $\mathcal{H}(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T \mathbf{E}\mathbf{x}$. \mathcal{H} is a quadratic function of state describing energy storage within the system. The matrix \mathbf{E} , which is the Hessian matrix of $\mathcal{H}(\mathbf{x})$, is the *energy matrix*; \mathbf{R} is the *dissipation matrix*; \mathbf{J} is the *structure matrix* describing the energy flux among internal energy storage elements; and $\mathbf{B} \pm \mathbf{P}$ are *port matrices* describing how energy enters and leaves the system. \mathbf{S} and \mathbf{N} are matrices associated with a *direct feed-through* from input \mathbf{u} to output \mathbf{y} . If the system has a solution $\mathbf{x} \in C^1(\mathbb{I}, \mathbb{R}^n)$ in \mathbb{I} for a given input function \mathbf{u} , then the dissipation inequality

$$\frac{d}{dt}\mathcal{H}(\mathbf{x}) = \mathbf{u}^T \mathbf{y} - \begin{bmatrix} \mathbf{x} \\ \mathbf{u} \end{bmatrix}^T \mathbf{W} \begin{bmatrix} \mathbf{x} \\ \mathbf{u} \end{bmatrix} \leq \mathbf{u}^T \mathbf{y}$$

must hold. The system is both *passive* and *Lyapunov stable*, as \mathcal{H} defines a *Lyapunov function* when $\mathbf{u} = 0$, see [5].

In practice, network-based automated modeling tools such as MODELICA, MATLAB/SIMULINK, or 20-sim¹ may produce system models that are overdetermined as a consequence of redundant modeling. Thus, automated generation of a system model usually must be followed by reformulation and regularization. Evidently, these subsequent steps should preserve the pH structure and retain compatibility with standard simulation, control, and optimization tools.

pHDAE models may be very large and complex, e.g., models that arise from semi-discretized (in space) continuum models in hydrodynamics [11, 19, 29], or mechanics [25]. In such cases, model reduction techniques are necessary to apply control and optimization methods. Preservation of the pH structure, the constraint structure, and the interconnection structure is necessary to maintain model integrity. For linear time-invariant pH systems with positive-definite \mathbf{E} , such methods are well-developed, including tangential interpolation approaches [15, 16] and moment matching [26, 28, 31], as well as effort and flow constraint reduction methods [27]. Interpolatory approaches have been extended to nonlinear pH systems as well, see [3] and [9].

In the case of singular \mathbf{E} , only recently in [11, 18] have model reduction techniques been introduced that are able to preserve pH structure. Note that for unstructured DAEs, constraint-preserving reduction methods were introduced in [17, 19, 25, 29].

In this work, we discuss particular structure-preserving model reduction methods that incorporate both regularization and interpolation for linear pHDAE systems having the form (1). For different model classes, we describe how constraints are represented in the transfer function and how the polynomial part can be preserved with interpolatory model reduction methods. The key step identifies, as in [5, 24], all redundancies as well as both explicit and implicit system constraints, partitioning the system equations into redundant, algebraic, and dynamic parts. Then, only the dynamic part need be reduced in a way that preserves the structure.

In Sect. 3, we consider pHDAEs of the form (1) and note important simplifications of the general regularization procedure. Structure preserving model reduction of

¹ <https://www.modelica.org/>, <http://www.mathworks.com>, <https://www.20sim.com>.

pHDAEs using interpolatory methods is discussed in Sect. 4 for several specific model structures. We propose an algorithm to generate effective interpolation data in Sect. 5 followed by a numerical example that demonstrates its effectiveness.

2 General Differential-Algebraic Systems

In this section, we review some basic properties of linear constant coefficient DAE systems having the general form,

$$\begin{aligned} \mathbf{E}\dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, & \mathbf{x}(t_0) &= 0, \\ \mathbf{y} &= \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u}, \end{aligned} \quad (2)$$

where $\mathbf{E}, \mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times m}$, $\mathbf{C} \in \mathbb{R}^{p \times n}$, and $\mathbf{D} \in \mathbb{R}^{p \times m}$; for details, see, e.g., [7].

The matrix pencil $\lambda\mathbf{E} - \mathbf{A}$ is said to be *regular*, if $\det(\lambda\mathbf{E} - \mathbf{A}) \neq 0$ for some $\lambda \in \mathbb{C}$. For regular pencils, the *finite eigenvalues* are the values $\lambda \in \mathbb{C}$ for which $\det(\lambda\mathbf{E} - \mathbf{A}) = 0$. If the *reversed pencil* $\lambda\mathbf{A} - \mathbf{E}$ has the eigenvalue 0, then this is the *infinite eigenvalue* of $\lambda\mathbf{E} - \mathbf{A}$.

With zero initial conditions, $\mathbf{x}(t_0) = 0$ as in (2), and a regular pencil, $\lambda\mathbf{E} - \mathbf{A}$, we obtain in the frequency domain, the *transfer function*, $\mathbf{H}(s) = \mathbf{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}$. $\mathbf{H}(s)$ is a rational function mapping the Laplace transform of the input function \mathbf{u} to the Laplace transform of the output function \mathbf{y} . If the transfer function is written as

$$\mathbf{H}(s) = \mathbf{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D} = \mathbf{G}(s) + \mathbf{P}(s),$$

where $\mathbf{G}(s)$ is a proper rational matrix function and $\mathbf{P}(s)$ is a polynomial matrix function, then the finite eigenvalues are the poles of the proper rational part, $\mathbf{G}(s)$, while the infinite eigenvalues are associated with the polynomial part, $\mathbf{P}(s)$.

Regular pencils can be analyzed via the Weierstraß Canonical Form; see, e.g., [13]. The *index*, ν , of the pencil $\lambda\mathbf{E} - \mathbf{A}$ is the size of the largest block associated with the eigenvalue ∞ in the Weierstraß canonical form; if \mathbf{E} is nonsingular, then $\nu = 0$.

To analyze general DAEs and to understand the properties of the transfer function, conditions on controllability and observability are needed, see, e.g., [7, 10]. Let $S_\infty(\mathbf{E})$ and $T_\infty(\mathbf{E})$ be two matrices with orthonormal columns spanning, respectively, the right and left nullspace of \mathbf{E} . A system is called *strongly controllable* if it satisfies C1: $\text{rank}[\lambda\mathbf{E} - \mathbf{A}, \mathbf{B}] = n$ for all $\lambda \in \mathbb{C}$ and C2: $\text{rank}[\mathbf{E}, \mathbf{A}S_\infty(\mathbf{M}), \mathbf{B}] = n$.

Analogously, a system is called *strongly observable* if it satisfies

$$\text{O1: } \text{rank}[\lambda\mathbf{E}^T - \mathbf{A}^T, \mathbf{C}^T] = n \text{ for all } \lambda \in \mathbb{C} \text{ and O2: } \text{rank}[\mathbf{E}^T, \mathbf{A}^T T_\infty(\mathbf{E}), \mathbf{C}^T] = n.$$

If a system is strongly controllable and strongly observable, then it is called *minimal*.

Conditions (C1) and (C2) are preserved under non-singular equivalence transformations as well as under state and output feedback. Note, however, that regularity (or non-regularity) of the pencil, of the index, and of the polynomial part of the transfer function are generally not preserved under state or output feedback. For systems that satisfy (C2), there exists a suitable linear state feedback matrix, \mathbf{F}_1 , such that $\lambda\mathbf{E} - (\mathbf{A} + \mathbf{B}\mathbf{F}_1)$ is regular and of index at most one. Also if conditions (C2) and (O2) hold, then there exists a linear output feedback matrix, \mathbf{F}_2 , so that the pencil $\lambda\mathbf{E} - (\mathbf{A} + \mathbf{B}\mathbf{F}_2\mathbf{C})$ has this property, see [7].

3 Regularization of PHDAE Systems

In general, it cannot be guaranteed that a system, generated either from a realization procedure or an automated modeling procedure, has a regular pencil $\lambda\mathbf{E} - \mathbf{A}$. Therefore, a regularization procedure typically must be applied, see [6, 8, 20]. For pHDAEs, pH structure helps simplify these regularization procedures significantly, since the pencil $\lambda\mathbf{E} - (\mathbf{J} - \mathbf{R})$ is associated with a free *dissipative Hamiltonian* DAE system (i.e., $\mathbf{u} = 0$). Such systems have many nice properties [22]: The index ν is at most two; all eigenvalues are in the closed left half plane; and all eigenvalues on the imaginary axis are semi-simple (except for possibly the eigenvalue 0, which then may have Jordan blocks of size at most two). A singular pencil can occur only when \mathbf{E} , \mathbf{J} , and \mathbf{R} have a common nullspace [23], so if one is able to compute efficiently this common nullspace, it is possible to remove explicitly the singular part.

Lemma 1 *For the pHDAE in (1), there exists an orthogonal basis transformation matrix $\mathbf{V} \in \mathbb{R}^{n \times n}$ such that in the new variable $\tilde{\mathbf{x}} = [\tilde{\mathbf{x}}_1^T \tilde{\mathbf{x}}_2^T \tilde{\mathbf{x}}_3^T]^T = \mathbf{V}^T \mathbf{x}$, the system has the form*

$$\begin{bmatrix} \mathbf{E}_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{x}}_1 \\ \tilde{\mathbf{x}}_2 \\ \tilde{\mathbf{x}}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{J}_1 - \mathbf{R}_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{x}}_1 \\ \tilde{\mathbf{x}}_2 \\ \tilde{\mathbf{x}}_3 \end{bmatrix} + \begin{bmatrix} \mathbf{B}_1 - \mathbf{P}_1 \\ \mathbf{B}_2 \\ 0 \end{bmatrix} \mathbf{u}, \quad (3)$$

$$\mathbf{y} = [(\mathbf{B}_1 + \mathbf{P}_1)^T \ \mathbf{B}_2^T \ 0] \begin{bmatrix} \tilde{\mathbf{x}}_1 \\ \tilde{\mathbf{x}}_2 \\ \tilde{\mathbf{x}}_3 \end{bmatrix} + (\mathbf{S} + \mathbf{N})\mathbf{u},$$

where $\lambda\mathbf{E}_1 - (\mathbf{J}_1 - \mathbf{R}_1)$ is regular and \mathbf{B}_2 has full row rank. Also, the subsystem

$$\begin{bmatrix} \mathbf{E}_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{x}}_1 \\ \tilde{\mathbf{x}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{J}_1 - \mathbf{R}_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{x}}_1 \\ \tilde{\mathbf{x}}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{B}_1 - \mathbf{P}_1 \\ \mathbf{B}_2 \end{bmatrix} \mathbf{u}, \quad (4)$$

$$\mathbf{y} = [(\mathbf{B}_1 + \mathbf{P}_1)^T \ \mathbf{B}_2^T] \begin{bmatrix} \tilde{\mathbf{x}}_1 \\ \tilde{\mathbf{x}}_2 \end{bmatrix} + (\mathbf{S} + \mathbf{N})\mathbf{u},$$

obtained by removing the third equation and the variable \mathbf{x}_3 , is still a pHDAE.

Proof First determine an orthogonal matrix \mathbf{V}_1 such that

$$\mathbf{V}_1^T(\lambda\mathbf{E} - (\mathbf{J} - \mathbf{R}))\mathbf{V}_1 = \lambda \begin{bmatrix} \mathbf{E}_1 & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} \mathbf{J}_1 - \mathbf{R}_1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{V}_1^T(\mathbf{B} - \mathbf{P}) = \begin{bmatrix} \mathbf{B}_1 - \mathbf{P}_1 \\ \tilde{\mathbf{B}}_2 - \tilde{\mathbf{P}}_2 \end{bmatrix}.$$

Such \mathbf{V}_1 exists since \mathbf{E} , \mathbf{J} , \mathbf{R} have a common nullspace when the pencil $\lambda\mathbf{E} - (\mathbf{J} - \mathbf{R})$ is singular [23]. Then a row compression of $\tilde{\mathbf{B}}_2 - \tilde{\mathbf{P}}_2$ via an orthogonal matrix $\tilde{\mathbf{V}}_2$ and a congruence transformation with $\mathbf{V}_2 = \text{diag}(\mathbf{I}, \tilde{\mathbf{V}}_2)$ is performed, so that with $\mathbf{V} = \text{diag}(\mathbf{V}_1, \mathbf{V}_2)$, we obtain the zero pattern in (3). Updating the output equation accordingly and using the fact that the transformed passivity matrix

$$\tilde{\mathbf{W}} = \begin{bmatrix} \mathbf{V}^T \mathbf{R} \mathbf{V} & \mathbf{V}^T \mathbf{P} \\ \mathbf{P}^T \mathbf{V} & \mathbf{S} \end{bmatrix} \in \mathbb{R}^{(n+m) \times (n+m)}$$

is still semidefinite; it follows that $\mathbf{P}_2 = 0$ and $\mathbf{P}_3 = 0$, giving the desired form. \square

The next result presents a condensed form, showing that conditions (C2) and (O2) are equivalent and hold for the system in (4).

Lemma 2 *For the pHDAE in (4) there exists an orthogonal basis transformation $\hat{\mathbf{V}}$ in the state space and \mathbf{U} in the control space such that in the new variables $\hat{\mathbf{x}} = [\hat{\mathbf{x}}_1^T \hat{\mathbf{x}}_2^T \hat{\mathbf{x}}_3^T \hat{\mathbf{x}}_4^T \hat{\mathbf{x}}_5^T \hat{\mathbf{x}}_6^T]^T = \hat{\mathbf{V}}^T [\tilde{\mathbf{x}}_1^T \tilde{\mathbf{x}}_2^T]^T$, and $[\mathbf{u}_1^T \mathbf{u}_2^T \mathbf{u}_3^T]^T = \mathbf{U}^T \mathbf{u}$ the system has the form*

$$\begin{aligned} & \begin{bmatrix} \mathbf{E}_{11} & \mathbf{E}_{12} & 0 & 0 & 0 & 0 \\ \mathbf{E}_{21} & \mathbf{E}_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}}_1 \\ \hat{\mathbf{x}}_2 \\ \hat{\mathbf{x}}_3 \\ \hat{\mathbf{x}}_4 \\ \hat{\mathbf{x}}_5 \\ \hat{\mathbf{x}}_6 \end{bmatrix} = \begin{bmatrix} \mathbf{J}_{11} - \mathbf{R}_{11} & \mathbf{J}_{12} - \mathbf{R}_{12} & \mathbf{J}_{13} - \mathbf{R}_{13} & \mathbf{J}_{14} & \mathbf{J}_{15} & 0 \\ \mathbf{J}_{21} - \mathbf{R}_{21} & \mathbf{J}_{22} - \mathbf{R}_{22} & \mathbf{J}_{23} - \mathbf{R}_{23} & \mathbf{J}_{24} & 0 & 0 \\ \mathbf{J}_{31} - \mathbf{R}_{31} & \mathbf{J}_{32} - \mathbf{R}_{32} & \mathbf{J}_{33} - \mathbf{R}_{33} & 0 & 0 & 0 \\ \mathbf{J}_{41} & \mathbf{J}_{42} & 0 & 0 & 0 & 0 \\ \mathbf{J}_{51} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}}_1 \\ \hat{\mathbf{x}}_2 \\ \hat{\mathbf{x}}_3 \\ \hat{\mathbf{x}}_4 \\ \hat{\mathbf{x}}_5 \\ \hat{\mathbf{x}}_6 \end{bmatrix} \\ & + \begin{bmatrix} \mathbf{B}_{11} - \mathbf{P}_{11} & \mathbf{B}_{12} - \mathbf{P}_{12} & \mathbf{B}_{13} - \mathbf{P}_{13} \\ \mathbf{B}_{21} - \mathbf{P}_{21} & \mathbf{B}_{22} - \mathbf{P}_{22} & \mathbf{B}_{23} - \mathbf{P}_{23} \\ \mathbf{B}_{31} - \mathbf{P}_{31} & \mathbf{B}_{32} - \mathbf{P}_{32} & \mathbf{B}_{33} - \mathbf{P}_{33} \\ 0 & \mathbf{B}_{42} & \mathbf{B}_{43} \\ 0 & 0 & \mathbf{B}_{53} \\ 0 & 0 & \mathbf{B}_{63} \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \mathbf{u}_3 \end{bmatrix}, \quad (5) \\ & \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \end{bmatrix} = \begin{bmatrix} (\mathbf{B}_{11} + \mathbf{P}_{11})^T & (\mathbf{B}_{21} + \mathbf{P}_{21})^T & \mathbf{B}_{31} + \mathbf{P}_{31})^T & 0 & 0 & 0 \\ (\mathbf{B}_{12} + \mathbf{P}_{12})^T & (\mathbf{B}_{22} + \mathbf{P}_{22})^T & \mathbf{B}_{32} + \mathbf{P}_{32})^T & \mathbf{B}_{42}^T & 0 & 0 \\ (\mathbf{B}_{13} + \mathbf{P}_{13})^T & (\mathbf{B}_{23} + \mathbf{P}_{23})^T & \mathbf{B}_{33} + \mathbf{P}_{33})^T & \mathbf{B}_{43}^T & \mathbf{B}_{53}^T & \mathbf{B}_{63}^T \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}}_1 \\ \hat{\mathbf{x}}_2 \\ \hat{\mathbf{x}}_3 \\ \hat{\mathbf{x}}_4 \\ \hat{\mathbf{x}}_5 \\ \hat{\mathbf{x}}_6 \end{bmatrix} \\ & + \begin{bmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} & \mathbf{D}_{13} \\ \mathbf{D}_{21} & \mathbf{D}_{22} & \mathbf{D}_{23} \\ \mathbf{D}_{31} & \mathbf{D}_{32} & \mathbf{D}_{33} \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \mathbf{u}_3 \end{bmatrix}, \end{aligned}$$

where \mathbf{E}_{22} , $\mathbf{J}_{33} - \mathbf{R}_{33}$, \mathbf{J}_{15} , and \mathbf{B}_{42} and \mathbf{B}_{63} are invertible. Furthermore, the subsystem obtained by deleting the first, fifth, and sixth block row and column satisfies (C2) and equivalently (O2).

Proof The proof follows again by a sequence of orthogonal transformations. Starting from (4) in the first step one determines an orthogonal matrix \mathbf{V}_1 (via a spectral decomposition of \mathbf{E}_1) such $\tilde{\mathbf{V}}_1^T \mathbf{E}_1 \tilde{\mathbf{V}}_1 = \begin{bmatrix} \mathbf{E}_{11} & 0 \\ 0 & 0 \end{bmatrix}$ with $\mathbf{E}_{11} > 0$, and then forms a congruence transformation with $\hat{\mathbf{V}}_1 = \text{diag}(\tilde{\mathbf{V}}_1, \mathbf{I})$ yielding

$$\hat{\mathbf{V}}_1^T (\mathbf{J} - \mathbf{R}) \hat{\mathbf{V}}_1 = \begin{bmatrix} \tilde{\mathbf{J}}_{11} - \tilde{\mathbf{R}}_{11} & \tilde{\mathbf{J}}_{12} - \tilde{\mathbf{R}}_{12} \\ \tilde{\mathbf{J}}_{21} - \tilde{\mathbf{R}}_{21} & \tilde{\mathbf{J}}_{22} - \tilde{\mathbf{R}}_{22} \end{bmatrix}, \quad \hat{\mathbf{V}}_1^T (\mathbf{B} - \mathbf{P}) = \begin{bmatrix} \tilde{\mathbf{B}}_1 - \tilde{\mathbf{P}}_1 \\ \tilde{\mathbf{B}}_2 - \tilde{\mathbf{P}}_2 \end{bmatrix}.$$

Next compute a full rank decomposition $\tilde{\mathbf{V}}_2^T (\tilde{\mathbf{J}}_{22} - \tilde{\mathbf{R}}_{22}) \tilde{\mathbf{V}}_2 = \begin{bmatrix} \hat{\mathbf{J}}_{22} - \hat{\mathbf{R}}_{22} & 0 \\ 0 & 0 \end{bmatrix}$, where $\hat{\mathbf{J}}_{22} - \hat{\mathbf{R}}_{22}$ is invertible and $\hat{\mathbf{R}}_{22} \geq 0$. This exists, since $\tilde{\mathbf{J}}_{22} - \tilde{\mathbf{R}}_{22}$ has a negative semidefinite symmetric part. Then an appropriate congruence transformation with $\hat{\mathbf{V}}_2 = \text{diag}(\mathbf{I}, \tilde{\mathbf{V}}_2, \mathbf{I})$ yields

$$\hat{\mathbf{V}}_2^T \hat{\mathbf{V}}_1^T \mathbf{E} \hat{\mathbf{V}}_1 \hat{\mathbf{V}}_2 = \begin{bmatrix} \mathbf{E}_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\hat{\mathbf{V}}_2^T \hat{\mathbf{V}}_1^T (\mathbf{J} - \mathbf{R}) \hat{\mathbf{V}}_1 \hat{\mathbf{V}}_2 = \begin{bmatrix} \hat{\mathbf{J}}_{11} - \hat{\mathbf{R}}_{11} & \hat{\mathbf{J}}_{12} - \hat{\mathbf{R}}_{12} & \hat{\mathbf{J}}_{13} & 0 \\ \hat{\mathbf{J}}_{21} - \hat{\mathbf{R}}_{21} & \hat{\mathbf{J}}_{22} - \hat{\mathbf{R}}_{22} & 0 & 0 \\ \hat{\mathbf{J}}_{31} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \hat{\mathbf{V}}_2^T \hat{\mathbf{V}}_1^T (\mathbf{B} - \mathbf{P}) = \begin{bmatrix} \hat{\mathbf{B}}_1 - \hat{\mathbf{P}}_1 \\ \hat{\mathbf{B}}_2 - \hat{\mathbf{P}}_2 \\ \hat{\mathbf{B}}_3 \\ \hat{\mathbf{B}}_4 \end{bmatrix},$$

where $\hat{\mathbf{J}}_{22} - \hat{\mathbf{R}}_{22}$ is invertible and $\hat{\mathbf{B}}_4$ has full row rank.

Then one performs an orthogonal decomposition

$$\tilde{\mathbf{V}}_3^T \begin{bmatrix} \hat{\mathbf{B}}_3 \\ \hat{\mathbf{B}}_4 \end{bmatrix} \mathbf{U} = \begin{bmatrix} 0 & \mathbf{B}_{42} & \mathbf{B}_{43} \\ 0 & 0 & \mathbf{B}_{53} \\ 0 & 0 & \mathbf{B}_{63} \end{bmatrix}$$

with \mathbf{B}_{42} and \mathbf{B}_{63} square nonsingular, where the number of rows in \mathbf{B}_{63} is that of $\hat{\mathbf{B}}_4$ and applies an appropriate congruence transformation with $\hat{\mathbf{V}}_3 = \text{diag}(\mathbf{I}, \mathbf{I}, \hat{\mathbf{V}}_3, \mathbf{I})$ so that one obtains block matrices

$$\begin{bmatrix} \mathbf{E}_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} \bar{\mathbf{J}}_{11} - \bar{\mathbf{R}}_{11} & \bar{\mathbf{J}}_{12} - \bar{\mathbf{R}}_{12} & \bar{\mathbf{J}}_{13} - \bar{\mathbf{R}}_{13} & \bar{\mathbf{J}}_{14} & 0 \\ \bar{\mathbf{J}}_{21} - \bar{\mathbf{R}}_{21} & \bar{\mathbf{J}}_{22} - \bar{\mathbf{R}}_{22} & 0 & 0 & 0 \\ \bar{\mathbf{J}}_{31} & 0 & 0 & 0 & 0 \\ \bar{\mathbf{J}}_{41} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

As a final step one computes a column compression of the full row rank matrix $\bar{\mathbf{J}}_{41}$ and applies an appropriate congruence transformation. This yields the desired form.

The fact that the several \mathbf{P} blocks do not occur follows again from the semidefiniteness of the transformed passivity matrix. Since \mathbf{J}_{51} , \mathbf{B}_{42} and \mathbf{B}_{63} are invertible, it follows immediately that $\mathbf{u}_3 = 0$ and $\hat{\mathbf{x}}_1 = 0$ and that $\hat{\mathbf{x}}_5$ is uniquely determined by all the other variables. Considering the subsystem obtained by removing the first, fifth, and sixth block row and column, it follows from the symmetry structure that the condition (C2) holds if and only (O2) holds.

The procedure to compute the condensed form (3) immediately separates the dynamical part (given by the first block row), the algebraic index-1 conditions (with and without dissipation, given by the second and third block rows), and the index-2 conditions (given by the fourth block row). The last row is the singular part of the free system. However, since the conditions (C2) and (O2) hold, the system can be made regular by output feedback. Note that this is already displayed in the subsystem (4), which, for ease of notation, we denote by

$$\begin{aligned} \mathbf{E}_r \dot{\mathbf{x}}_r &= (\mathbf{J}_r - \mathbf{R}_r) \mathbf{x}_r + (\mathbf{B}_r - \mathbf{P}_r) \mathbf{u}, & \mathbf{x}_r(t_0) &= 0. \\ \mathbf{y}_r &= (\mathbf{B}_r + \mathbf{P}_r)^T \mathbf{x}_r + (\mathbf{S}_r + \mathbf{N}_r) \mathbf{u}, \end{aligned}$$

If we apply an output feedback $\mathbf{u} = -\mathbf{K}_r \mathbf{y}_r$ with $\mathbf{K}_r = \mathbf{K}_r^T > 0$ that makes the (2,2) block in the closed loop system invertible, then this corresponds to a feedback for the state-to-output map in (4) of the form $\mathbf{y}_r = (\mathbf{I} + (\mathbf{S}_r + \mathbf{N}_r) \mathbf{K}_r)^{-1} (\mathbf{B}_r + \mathbf{P}_r)^T \mathbf{x}_r$, which we can insert into the first equation to obtain

$$\begin{aligned} \mathbf{E}_r \dot{\mathbf{x}}_r &= (\mathbf{J}_r - \mathbf{R}_r) \mathbf{x}_r + (\mathbf{B}_r - \mathbf{P}_r) \mathbf{K}_r \mathbf{y}_r \\ &= ((\mathbf{J}_r - \mathbf{R}_r) - (\mathbf{B}_r - \mathbf{P}_r) (\mathbf{K}_r^{-1} + \mathbf{S}_r + \mathbf{N}_r)^{-1} (\mathbf{B}_r + \mathbf{P}_r)^T) \mathbf{x}_r. \end{aligned}$$

The system matrix is the Schur complement of

$$\begin{bmatrix} \mathbf{J}_r & \mathbf{B}_r \\ -\mathbf{B}_r^T & -\mathbf{N}_r \end{bmatrix} - \begin{bmatrix} \mathbf{R}_r & \mathbf{P}_r \\ \mathbf{P}_r^T & \mathbf{K}_r^{-1} + \mathbf{S}_r \end{bmatrix},$$

which has a symmetric part given by $-\tilde{\mathbf{W}}_r$, with $\tilde{\mathbf{W}}_r = \begin{bmatrix} \mathbf{R}_r & \mathbf{P}_r \\ \mathbf{P}_r^T & \mathbf{K}_r^{-1} + \mathbf{S}_r \end{bmatrix} \geq 0$. Hence, the closed loop system is regular and the closed loop system is still pH.

The procedures described above are computationally demanding, since they typically require large-scale singular value decompositions or spectral decompositions. Fortunately, in many practical cases the condensed form is already available directly from the modeling procedure, so that the transfer function can be formed and the model reduction method can be directly applied.

4 Interpolatory Model Reduction of pHDAEs

Given an order- n pHDAE as in (1), we want to construct an order- r reduced pHDAE, with $r \ll n$, having the same structured form

$$\begin{aligned}\widehat{\mathbf{E}}\dot{\mathbf{x}}_r &= (\widehat{\mathbf{J}} - \widehat{\mathbf{R}})\mathbf{x}_r + (\widehat{\mathbf{B}} - \widehat{\mathbf{P}})\mathbf{u}, \quad \mathbf{x}_r(t_0) = 0, \\ \mathbf{y}_r &= (\widehat{\mathbf{B}} + \widehat{\mathbf{P}})^T \mathbf{x}_r + (\widehat{\mathbf{S}} + \widehat{\mathbf{N}})\mathbf{u},\end{aligned}\quad (6)$$

such that $\widehat{\mathbf{E}}, \widehat{\mathbf{J}}, \widehat{\mathbf{R}} \in \mathbb{R}^{r \times r}$, $\widehat{\mathbf{B}}, \widehat{\mathbf{P}} \in \mathbb{R}^{r \times m}$, $\widehat{\mathbf{S}} = \widehat{\mathbf{S}}^T$, $\widehat{\mathbf{N}} = -\widehat{\mathbf{N}}^T \in \mathbb{R}^{m \times m}$ satisfy the same requirements as in Definition 1 and that the output $\mathbf{y}_r(t)$ of (6) is an accurate approximation to the original output $\mathbf{y}(t)$ over a wide range of admissible inputs $\mathbf{u}(t)$. We will enforce accuracy by constructing the reduced model (6) via interpolation.

Let $\mathbf{H}(s) = \mathcal{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathcal{B} + \mathbf{D}$ and $\widehat{\mathbf{H}}(s) = \widehat{\mathcal{C}}(s\widehat{\mathbf{E}} - \widehat{\mathbf{A}})^{-1}\widehat{\mathcal{B}} + \widehat{\mathbf{D}}$ denote the transfer functions of (1) and (6), where $\mathcal{C} = (\mathbf{B} + \mathbf{P})^T$, $\mathbf{A} = \mathbf{J} - \mathbf{R}$, $\mathcal{B} = \mathbf{B} - \mathbf{P}$, $\mathbf{D} = \mathbf{S} + \mathbf{N}$, and similarly for the reduced-order (“hat”) quantities. Given right-interpolation points $\{\sigma_1, \sigma_2, \dots, \sigma_r\} \in \mathbb{C}$ together with the corresponding right-tangent directions $\{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_r\} \in \mathbb{C}^m$ and left-interpolation points $\{\mu_1, \mu_2, \dots, \mu_r\} \in \mathbb{C}$ together with the corresponding left-tangent directions $\{\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \dots, \boldsymbol{\ell}_r\} \in \mathbb{C}^m$, we wish to construct a reduced model, $\widehat{\mathbf{H}}(s)$, that tangentially interpolates $\mathbf{H}(s)$, i.e.,

$$\mathbf{H}(\sigma_i)\mathbf{b}_i = \widehat{\mathbf{H}}(\sigma_i)\mathbf{b}_i \quad \text{and} \quad \boldsymbol{\ell}_i^T \mathbf{H}(\mu_i) = \boldsymbol{\ell}_i^T \widehat{\mathbf{H}}(\mu_i), \quad \text{for } i = 1, 2, \dots, r.$$

These tangential interpolation conditions can be enforced easily via a Petrov-Galerkin projection [1, 4, 12]. In particular, construct $\mathbf{V} \in \mathbb{C}^{n \times r}$ and $\mathbf{Z} \in \mathbb{C}^{n \times r}$ using

$$\begin{aligned}\mathbf{V} &= [(\sigma_1\mathbf{E} - \mathbf{A})^{-1}\mathcal{B}\mathbf{b}_1, \quad (\sigma_2\mathbf{E} - \mathbf{A})^{-1}\mathcal{B}\mathbf{b}_2, \quad \dots \quad (\sigma_r\mathbf{E} - \mathbf{A})^{-1}\mathcal{B}\mathbf{b}_r] \quad \text{and} \\ \mathbf{Z} &= [(\mu_1\mathbf{E} - \mathbf{A})^{-T}\mathcal{C}^T\boldsymbol{\ell}_1, \quad (\mu_2\mathbf{E} - \mathbf{A})^{-T}\mathcal{C}^T\boldsymbol{\ell}_2, \quad \dots \quad (\mu_r\mathbf{E} - \mathbf{A})^{-T}\mathcal{C}^T\boldsymbol{\ell}_r].\end{aligned}$$

Then an interpolatory reduced model can be defined via the projection:

$$\widehat{\mathbf{E}} = \mathbf{Z}^T \mathbf{E} \mathbf{V}, \quad \widehat{\mathbf{A}} = \mathbf{Z}^T \mathbf{A} \mathbf{V}, \quad \widehat{\mathbf{B}} = \mathbf{Z}^T \mathcal{B}, \quad \widehat{\mathcal{C}} = \mathcal{C} \mathbf{V}, \quad \text{and} \quad \widehat{\mathbf{D}} = \mathbf{D}. \quad (7)$$

In the setting of pHDAEs, two fundamental issues arise: First, the reduced quantities in (7) are no longer guaranteed to have pH structure. This is most easily seen by examining the reduced quantity, $\widehat{\mathcal{A}} = \mathbf{Z}^T \mathcal{A} \mathbf{V} = \mathbf{Z}^T \mathbf{J} \mathbf{V} - \mathbf{Z}^T \mathbf{R} \mathbf{V}$. If $\widehat{\mathcal{A}}$ is decomposed into its symmetric and skew-symmetric parts, it can no longer be guaranteed that the symmetric part is positive semi-definite. This could be resolved by using a Ritz-Galerkin projection, taking $\mathbf{Z} = \mathbf{V}$, but then only interpolation conditions associated with right interpolation data are satisfied. Even then, there will still be a second issue that persists. In the generic case when $r < \text{rank}(\mathbf{E})$, the reduced quantity $\widehat{\mathbf{E}}$ will generally be nonsingular; thus the reduced system will be an ODE and

the polynomial parts of $\mathbf{H}(s)$ and $\widehat{\mathbf{H}}(s)$ will not match, leading then to unbounded errors.

Structure-preserving interpolatory reduction of pH systems in the most general setting of tangential interpolation has been studied in [15, 16]. However, this work focused on the ODE case. On the other hand, [17] developed the tangential interpolation framework for reducing *unstructured* DAEs with guaranteed polynomial matching. Only recently in [11, 18], the combined problem has been investigated. We now develop a treatment of structure-preserving interpolatory model reduction problem for index-1 and index-2 pHDAEs in the general setting of tangential interpolation.

4.1 Semi-explicit Index-1 pHDAE Systems

The simplest class of pHDAEs are *semi-explicit index-1* pHDAEs of the form

$$\begin{bmatrix} \mathbf{E}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \dot{\mathbf{x}}(t) = \begin{bmatrix} \mathbf{J}_{11} - \mathbf{R}_{11} & \mathbf{J}_{12} - \mathbf{R}_{12} \\ -\mathbf{J}_{12}^T - \mathbf{R}_{12}^T & \mathbf{J}_{22} - \mathbf{R}_{22} \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} \mathbf{B}_1 - \mathbf{P}_1 \\ \mathbf{B}_2 - \mathbf{P}_2 \end{bmatrix} \mathbf{u}(t), \quad (8)$$

$$\mathbf{y}(t) = [\mathbf{B}_1^T + \mathbf{P}_1^T \quad \mathbf{B}_2^T + \mathbf{P}_2^T] \mathbf{x}(t) + (\mathbf{S} + \mathbf{N})\mathbf{u}(t),$$

where \mathbf{E}_{11} and $\mathbf{J}_{22} - \mathbf{R}_{22}$ are nonsingular. We have the following interpolation result.

Theorem 1 *Consider the pHDAE system in (8). Let the interpolation points $\{\sigma_1, \sigma_2, \dots, \sigma_r\} \in \mathbb{C}$ and the corresponding tangent directions $\{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_r\} \in \mathbb{C}^m$ be given. Construct the interpolatory model reduction basis \mathbf{V} as*

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{bmatrix} = [(\sigma_1 \mathbf{E} - \mathbf{A})^{-1} \mathcal{B} \mathbf{b}_1, \quad \dots, \quad (\sigma_r \mathbf{E} - \mathbf{A})^{-1} \mathcal{B} \mathbf{b}_r] \in \mathbb{C}^{n \times r}, \quad (9)$$

where \mathbf{V} is partitioned conformably with the system, and define the matrices

$$\mathbb{B} = [\mathbf{b}_1 \quad \mathbf{b}_2 \quad \dots \quad \mathbf{b}_r] \in \mathbb{C}^{m \times r} \text{ and } \mathcal{D} = \mathbf{D} - (\mathbf{B}_2^T + \mathbf{P}_2^T)(\mathbf{J}_{22} - \mathbf{R}_{22})^{-1}(\mathbf{B}_2 - \mathbf{P}_2) \in \mathbb{C}^{m \times m}.$$

Let $\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} = \mathbf{J} - \mathbf{R}$, partitioned accordingly to (8). Then, the transfer function $\widehat{\mathbf{H}}(s)$ of the reduced model

$$\widehat{\mathbf{E}} \dot{\mathbf{x}}_r(t) = (\widehat{\mathbf{J}} - \widehat{\mathbf{R}}) \mathbf{x}_r(t) + \widehat{\mathbf{B}} \mathbf{u}(t), \quad \mathbf{y}_r(t) = \widehat{\mathbf{C}} \mathbf{x}_r(t) + \widehat{\mathbf{D}} \mathbf{u}(t), \quad (10)$$

with

$$\begin{aligned}\widehat{\mathbf{E}} &= \mathbf{V}_1^T \mathbf{E}_{11} \mathbf{V}_1, & \widehat{\mathbf{C}} &= \mathbf{C} \mathbf{V} + (\mathbf{B}_2^T + \mathbf{P}_2^T) \mathbf{A}_{22}^{-1} (\mathbf{B}_2 - \mathbf{P}_2) \mathbb{B}, \\ \widehat{\mathbf{A}} &= \mathbf{V}^T \mathbf{A} \mathbf{V} + \mathbb{B}^T (\mathcal{D} - \mathbf{D}) \mathbb{B}, & \widehat{\mathbf{D}} &= \mathcal{D} = \mathbf{D} - (\mathbf{B}_2^T + \mathbf{P}_2^T) \mathbf{A}_{22}^{-1} (\mathbf{B}_2 - \mathbf{P}_2), \\ \widehat{\mathbf{B}} &= \mathbf{V}^T \mathbf{B} + \mathbb{B}^T (\mathbf{B}_2^T + \mathbf{P}_2^T) \mathbf{A}_{22}^{-1} (\mathbf{B}_2 - \mathbf{P}_2),\end{aligned}$$

matches the polynomial part of $\mathbf{H}(s)$ and tangentially interpolates it, i.e.,

$$\mathbf{H}(\sigma_i) \mathbf{b}_i = \widehat{\mathbf{H}}(\sigma_i) \mathbf{b}_i, \text{ for } i = 1, 2, \dots, r.$$

Define $\widehat{\mathbf{P}} = \frac{1}{2}(-\widehat{\mathbf{G}} + \widehat{\mathbf{C}})$, and decompose $\widehat{\mathbf{D}} = \widehat{\mathbf{S}} + \widehat{\mathbf{N}}$ and $\widehat{\mathbf{A}} = \widehat{\mathbf{J}} - \widehat{\mathbf{R}}$ into their symmetric and skew-symmetric parts. Then, the reduced model (10) is a pHDAE system if the reduced passivity matrix $\widehat{\mathbf{W}} = \begin{bmatrix} \widehat{\mathbf{R}} & \widehat{\mathbf{P}} \\ \widehat{\mathbf{P}}^T & \widehat{\mathbf{S}} \end{bmatrix}$ is positive semidefinite.

Proof We employ a Galerkin projection using the interpolatory model reduction basis \mathbf{V} to obtain an *intermediate* reduced model

$$\widetilde{\mathbf{E}} = \mathbf{V}_1^T \mathbf{E}_{11} \mathbf{V}_1, \quad \widetilde{\mathbf{A}} = \widetilde{\mathbf{J}} - \widetilde{\mathbf{R}} = \mathbf{V}^T \mathbf{J} \mathbf{V} - \mathbf{V}^T \mathbf{R} \mathbf{V}, \quad \widetilde{\mathbf{B}} = \mathbf{V}^T \mathbf{B}, \quad \widetilde{\mathbf{C}} = \mathbf{C} \mathbf{V}, \quad \text{and} \quad \widetilde{\mathbf{D}} = \mathbf{D}.$$

This reduced model is a pHDAE system due to the use of a one-sided Galerkin projection, and likewise it also satisfies the desired tangential interpolation conditions. However, it will *not* generally match the transfer function, $\mathbf{H}(s)$, at $s = \infty$. Indeed, $\mathbf{H}(s)$ has a polynomial part is given by

$$\lim_{s \rightarrow \infty} \mathbf{H}(s) = \mathcal{D} = \mathbf{D} - (\mathbf{B}_2^T + \mathbf{P}_2^T) \mathbf{A}_{22}^{-1} (\mathbf{B}_2 - \mathbf{P}_2) \neq \widetilde{\mathbf{D}} = \lim_{s \rightarrow \infty} \widehat{\mathbf{H}}(s)$$

A remedy to this problem, proposed in [2, 21] and employed in the general DAE setting in [17], is to modify the \mathbf{D} -term in the reduced model to match the polynomial part while at the same time *shifting* other reduced quantities appropriately so as to retain tangential interpolation. Using this approach, we obtain a modified reduced model

$$\begin{aligned}\widehat{\mathbf{E}} &= \widetilde{\mathbf{E}} = \mathbf{V}_1^T \mathbf{E}_{11} \mathbf{V}_1, \\ \widehat{\mathbf{A}} &= \widetilde{\mathbf{A}} + \mathbb{B}^T (\mathcal{D} - \mathbf{D}) \mathbb{B} = \mathbf{V}^T \mathbf{J} \mathbf{V} - \mathbf{V}^T \mathbf{R} \mathbf{V} - \mathbb{B}^T (\mathbf{B}_2^T + \mathbf{P}_2^T) \mathbf{A}_{22}^{-1} (\mathbf{B}_2 - \mathbf{P}_2) \mathbb{B}, \\ \widehat{\mathbf{B}} &= \widetilde{\mathbf{B}} + \mathbb{B}^T (\mathbf{B}_2^T + \mathbf{P}_2^T) \mathbf{A}_{22}^{-1} (\mathbf{B}_2 - \mathbf{P}_2) \\ &= \mathbf{V}_1^T (\mathbf{B}_1 - \mathbf{P}_1) + \mathbf{V}_2^T (\mathbf{B}_2 - \mathbf{P}_2) + \mathbb{B}^T (\mathbf{B}_2^T + \mathbf{P}_2^T) \mathbf{A}_{22}^{-1} (\mathbf{B}_2 - \mathbf{P}_2), \\ \widehat{\mathbf{C}} &= \widetilde{\mathbf{C}} + (\mathbf{B}_2^T + \mathbf{P}_2^T) \mathbf{A}_{22}^{-1} (\mathbf{B}_2 - \mathbf{P}_2) \mathbb{B} \\ &= (\mathbf{B}_1^T + \mathbf{P}_1^T) \mathbf{V}_1^T + (\mathbf{B}_2^T + \mathbf{P}_2^T) \mathbf{V}_2^T + (\mathbf{B}_2^T + \mathbf{P}_2^T) \mathbf{A}_{22}^{-1} (\mathbf{B}_2 - \mathbf{P}_2) \mathbb{B}, \quad \text{and} \\ \widehat{\mathbf{D}} &= \mathcal{D} = \mathbf{D} - (\mathbf{B}_2^T + \mathbf{P}_2^T) \mathbf{A}_{22}^{-1} (\mathbf{B}_2 - \mathbf{P}_2),\end{aligned}$$

which satisfies the original tangential interpolation conditions while also matching the polynomial part with a modified $\widehat{\mathbf{D}}$ -term. For this system to be pH, we need to check that the associated passivity matrix is still positive semidefinite. After rewriting

the input and output matrix in the usual way, this is exactly the condition on $\widehat{\mathbf{W}}$ in the assertion. We then have that the reduced model not only satisfies the interpolation conditions and matches the polynomial part at $s = \infty$, but also is pH. \square

Remark 1 Note that if the input does not influence the algebraic equations, i.e., if $\mathbf{B}_2 = \mathbf{P}_2 = 0$, then the shift of the constant term is not necessary and the formulas simplify significantly, i.e., $\widehat{\mathbf{A}} = \mathbf{V}^T \mathbf{A} \mathbf{V}$, $\widehat{\mathbf{B}} = \widetilde{\mathbf{B}} = \mathbf{V}^T \mathbf{B}$, $\widehat{\mathbf{C}} = \mathbf{C} \mathbf{V}$, and $\widehat{\mathbf{D}} = \mathbf{D} = \mathbf{D}$.

Another solution to preserving pH structure via interpolation can be obtained through the following theorem.

Theorem 2 Consider a full-order pHDAE system of the form (8). Let interpolation points $\{\sigma_1, \sigma_2, \dots, \sigma_r\} \in \mathbf{C}$ and the corresponding tangent directions $\{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_r\} \in \mathbf{C}^m$ be given. Construct the interpolatory model reduction basis \mathbf{V} as in (9). Then the reduced model

$$\begin{aligned} \begin{bmatrix} \mathbf{V}_1^T \mathbf{E}_{11} \mathbf{V}_1 & 0 \\ 0 & 0 \end{bmatrix} \dot{\mathbf{x}}(t) &= \begin{bmatrix} \mathbf{V}_1^T (\mathbf{J}_{11} - \mathbf{R}_{11}) \mathbf{V}_1 & \mathbf{V}_1^T (\mathbf{J}_{12} - \mathbf{R}_{12}) \\ (-\mathbf{J}_{12}^T - \mathbf{R}_{12}^T) \mathbf{V}_1 & \mathbf{J}_{22} - \mathbf{R}_{22} \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} \mathbf{V}_1^T (\mathbf{B}_1 - \mathbf{P}_1) \\ \mathbf{B}_2 - \mathbf{P}_2 \end{bmatrix} \mathbf{u}(t) \\ \mathbf{y}_r(t) &= [(\mathbf{B}_1^T + \mathbf{P}_1)^T \mathbf{V}_1 \quad \mathbf{B}_2 + \mathbf{P}_2^T] \mathbf{x}(t) + \mathbf{D} \mathbf{u}(t), \end{aligned} \quad (11)$$

retains the pH structure, tangentially interpolates the original model, and matches the polynomial part.

Proof We first note that the subspace spanned by the columns of $\mathbf{V} = [\mathbf{V}_1^T \quad \mathbf{V}_2^T]^T$ is contained in the subspace spanned by the columns of $\widehat{\mathbf{V}} := \text{diag}(\mathbf{V}_1, \mathbf{I})$. Then, the system in (11) results from reducing the original system in (8) via $\widehat{\mathbf{V}}$. Since $\text{span}(\mathbf{V}) \subseteq \text{span}(\widehat{\mathbf{V}})$, this reduced DAE automatically satisfies the interpolation conditions and since $\widehat{\mathbf{V}}$ does not alter the matrix $\mathbf{J}_{22} - \mathbf{R}_{22}$ and the matrices $\mathbf{B}_2, \mathbf{P}_2$, the polynomial part of the transfer function is $\mathbf{D} - (\mathbf{B}_2^T + \mathbf{P}_2^T)(\mathbf{J}_{22} - \mathbf{R}_{22})^{-1}(\mathbf{B}_2 - \mathbf{P}_2)$, matching that of the original model. The reduced system in (11) is pH as well since the one-sided projection that is used retains the original pH structure. \square

Remark 2 Theorem 2 appears to present an easier alternative to the more complicated Theorem 1 for structure-preserving interpolatory model reduction of index-1 pHDAEs. However, the construction of Theorem 2 may not achieve the maximal reduction possible because redundant algebraic conditions cannot be removed; see [25]. Note also that further orthogonalization of \mathbf{V}_1 may be necessary, see [11].

4.2 Semi-explicit pHDAE Systems with Index-2 Constraints

The next class of interest will be semi-explicit index-2 systems. We first consider the case that the input does not affect the algebraic equations.

Theorem 3 Consider an index-2 pHDAE system of the form

$$\begin{aligned} \begin{bmatrix} \mathbf{E}_{11} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\mathbf{x}}_1(t) \\ \dot{\mathbf{x}}_2(t) \end{bmatrix} &= \begin{bmatrix} \mathbf{J}_{11} - \mathbf{R}_{11} & \mathbf{J}_{12} \\ -\mathbf{J}_{12}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1(t) \\ \mathbf{x}_2(t) \end{bmatrix} + \begin{bmatrix} \mathbf{B}_1 - \mathbf{P}_1 \\ \mathbf{0} \end{bmatrix} \mathbf{u}(t), \\ \mathbf{y}(t) &= [\mathbf{B}_1^T + \mathbf{P}_1^T \ \mathbf{0}] \begin{bmatrix} \mathbf{x}_1(t) \\ \mathbf{x}_2(t) \end{bmatrix} + \mathbf{D}\mathbf{u}(t), \end{aligned} \quad (12)$$

with $\mathbf{E}_{11} > 0$ and set $\mathbf{A}_{11} = \mathbf{J}_{11} - \mathbf{R}_{11}$. Given interpolation points $\{\sigma_1, \sigma_2, \dots, \sigma_r\}$ and associated tangent directions $\{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_r\}$, let the vectors \mathbf{v}_i , for $i = 1, 2, \dots, r$, be the first block of the solution of

$$\begin{bmatrix} \mathbf{A}_{11} - \sigma_i \mathbf{E}_{11} & \mathbf{J}_{12} \\ -\mathbf{J}_{12}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{v}_i \\ \mathbf{z} \end{bmatrix} = \begin{bmatrix} (\mathbf{B}_1 - \mathbf{P}_1) \mathbf{b}_i \\ \mathbf{0} \end{bmatrix}. \quad (13)$$

Define $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r]$. Then, the reduced model

$$\widehat{\mathbf{E}} \dot{\mathbf{x}}_r = (\widehat{\mathbf{J}} - \widehat{\mathbf{R}}) \mathbf{x}_r + \widehat{\mathbf{B}} \mathbf{u}(t), \quad \mathbf{y}_r = \widehat{\mathbf{C}} \mathbf{x}_r + \widehat{\mathbf{D}}, \quad (14)$$

$$\begin{aligned} \text{with} \quad \widehat{\mathbf{E}} &= \mathbf{V}^T \mathbf{E}_{11} \mathbf{V}, \quad \widehat{\mathbf{J}} = \mathbf{V}^T \mathbf{J}_{11} \mathbf{V}, \quad \widehat{\mathbf{R}} = \mathbf{V}^T \mathbf{R}_{11} \mathbf{V}, \\ \widehat{\mathbf{B}} &= \mathbf{V}^T \mathbf{B}_1 - \mathbf{V}^T \mathbf{P}_1, \quad \widehat{\mathbf{C}} = \mathbf{B}_1^T \mathbf{V}^T + \mathbf{P}_1^T \mathbf{V}_1^T, \quad \text{and} \quad \widehat{\mathbf{D}} = \mathbf{D}, \end{aligned} \quad (15)$$

is still pH, matches the polynomial part of the original transfer function, and satisfies the tangential interpolation conditions,

$$\mathbf{H}(\sigma_i) \mathbf{b}_i = \widehat{\mathbf{H}}(\sigma_i) \mathbf{b}_i, \quad \text{for } i = 1, 2, \dots, r.$$

Proof Note first that the regularity of $\lambda \mathbf{E} - (\mathbf{J} - \mathbf{R})$ and the index-2 condition imply that $-\mathbf{J}_{12}^T \mathbf{E}_{11}^{-1} \mathbf{J}_{12}$ is invertible, see [5, 20]. Following [17], we write (12) as

$$\begin{aligned} \mathbf{\Pi} \mathbf{E}_{11} \mathbf{\Pi}^T \dot{\mathbf{x}}_1(t) &= \mathbf{\Pi} \mathbf{A}_{11} \mathbf{\Pi}^T \mathbf{x}_1(t) + \mathbf{\Pi} \mathbf{B}_1 \mathbf{u}(t), \\ \mathbf{y}(t) &= \mathbf{C}_1 \mathbf{\Pi}^T \mathbf{x}_1(t) + \mathbf{D} \mathbf{u}(t), \end{aligned} \quad (16)$$

and conjoin this with the algebraic equation,

$$\mathbf{x}_2(t) = -(\mathbf{J}_{12}^T \mathbf{E}_{11}^{-1} \mathbf{J}_{12})^{-1} \mathbf{J}_{12}^T \mathbf{E}_{11}^{-1} \mathbf{A}_{11} \mathbf{x}_1(t) - (\mathbf{J}_{12}^T \mathbf{E}_{11}^{-1} \mathbf{J}_{12})^{-1} \mathbf{J}_{12}^T \mathbf{E}_{11}^{-1} \mathbf{B}_1 \mathbf{u}(t).$$

The skew projector, $\mathbf{\Pi}$, in (16) is defined as $\mathbf{\Pi} = \mathbf{I} - \mathbf{E}_{11}^{-1} \mathbf{J}_{12} (\mathbf{J}_{12}^T \mathbf{E}_{11}^{-1} \mathbf{J}_{12})^{-1} \mathbf{J}_{12}^T$, and (16) is now an implicit ODE pH system that can be reduced with standard model reduction techniques. As it stands, this would appear to require computing the projector $\mathbf{\Pi}$ explicitly, see [25]. For general index-2 DAE systems one can avoid this computational step through interpolatory model reduction, see [17].

To adapt this idea to pHDAE systems, we construct \mathbf{V} using (13) and then compute the reduced-order quantities via one-sided projection as in (15). This construction of \mathbf{V} , as in [17], guarantees that the reduced model in (14) tangentially interpolates the original pHDAE system in (16) and the polynomial part of the transfer function in (12) is given by $\mathbf{D} = \mathbf{S} + \mathbf{N}$, separated with respect to its symmetric and skew-symmetric parts. Since (14) is an implicit ODE pH system with an exact \mathbf{D} -term, it matches the polynomial part of the original transfer function $\mathbf{H}(s)$.

It remains to show that (14) is pH. By construction in (15), $\widehat{\mathbf{J}}$ is skew-symmetric, $\widehat{\mathbf{R}}$ is symmetric positive semidefinite, and \mathbf{E}_{11} is symmetric positive definite. Moreover,

$$\begin{bmatrix} \mathbf{V}^T \mathbf{R}_{11} \mathbf{V} & \mathbf{V}^T \mathbf{P}_1 \\ \mathbf{P}_1^T \mathbf{V} & \mathbf{S} \end{bmatrix} = \begin{bmatrix} \mathbf{V}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{R}_{11} & \mathbf{P}_1 \\ \mathbf{P}_1^T & \mathbf{S} \end{bmatrix} \begin{bmatrix} \mathbf{V} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \geq 0,$$

since the original model is pH. Therefore, the pH-structure is retained. \square

The situation becomes more complicated when the second block in \mathcal{B} is nonzero, that is, when the system has the form

$$\begin{bmatrix} \mathbf{E}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \dot{\mathbf{x}}_1(t) \\ \dot{\mathbf{x}}_2(t) \end{bmatrix} = \begin{bmatrix} \mathbf{J}_{11} - \mathbf{R}_{11} & \mathbf{J}_{12} \\ -\mathbf{J}_{12}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1(t) \\ \mathbf{x}_2(t) \end{bmatrix} + \begin{bmatrix} \mathbf{B}_1 - \mathbf{P}_1 \\ \mathbf{B}_2 - \mathbf{P}_2 \end{bmatrix} \mathbf{u}(t), \quad (17)$$

$$\mathbf{y}(t) = \begin{bmatrix} \mathbf{B}_1^T + \mathbf{P}_1^T & \mathbf{B}_2 + \mathbf{P}_2^T \end{bmatrix} \begin{bmatrix} \mathbf{x}_1(t) \\ \mathbf{x}_2(t) \end{bmatrix} + \mathbf{D}\mathbf{u}(t).$$

Theorem 4 Consider a pHDAE system of the form (17) and define the matrices

$$\mathcal{Z} = -(\mathbf{A}_{21} \mathbf{E}_{11}^{-1} \mathbf{A}_{12})^{-1} = (\mathbf{J}_{12}^T \mathbf{E}_{11}^{-1} \mathbf{J}_{12})^{-1},$$

$$\mathcal{B} = (\mathbf{B}_1 - \mathbf{P}_1) + (\mathbf{J}_{11} - \mathbf{R}_{11}) \mathbf{E}_{11}^{-1} \mathbf{J}_{12} \mathcal{Z} (\mathbf{B}_2 - \mathbf{P}_2),$$

$$\mathcal{C} = (\mathbf{B}_1^T - \mathbf{P}_1^T) - (\mathbf{B}_2^T - \mathbf{P}_2^T) \mathcal{Z} \mathbf{J}_{12}^T \mathbf{E}_{11}^{-1} (\mathbf{J}_{11} - \mathbf{R}_{11}),$$

$$\mathcal{D}_0 = \mathbf{D} - (\mathbf{B}_2^T + \mathbf{P}_2^T) \mathcal{Z} \mathbf{J}_{12}^T \mathbf{E}_{11}^{-1} (\mathbf{B}_1 - \mathbf{P}_1), \text{ and } \mathcal{D}_1 = -(\mathbf{B}_2^T + \mathbf{P}_2^T) \mathcal{Z} (\mathbf{B}_2 - \mathbf{P}_2).$$

Given interpolation points $\{\sigma_1, \sigma_2, \dots, \sigma_r\}$ and associated tangent directions $\{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_r\}$, let the vectors \mathbf{v}_i be the first blocks of the solutions of

$$\begin{bmatrix} \mathbf{J}_{11} - \mathbf{R}_{11} - \sigma_i \mathbf{E}_{11} & \mathbf{J}_{12} \\ -\mathbf{J}_{12}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{v}_i \\ \mathbf{z} \end{bmatrix} = \mathcal{B} \mathbf{b}_i, \text{ for } i = 1, 2, \dots, r, \quad (18)$$

and set $\mathbf{V} := [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r]$, $\mathbf{u}_1 := \mathbf{u}$, $\mathbf{u}_2 := \dot{\mathbf{u}}$, and $\widehat{\mathbf{D}} := [\mathcal{D}_0 \ \mathcal{D}_1] = \widehat{\mathbf{S}} + \widehat{\mathbf{N}}$. Then, the reduced model

$$\widehat{\mathbf{E}} \dot{\mathbf{x}}_r = (\widehat{\mathbf{J}} - \widehat{\mathbf{R}}) \mathbf{x}_r + [\widehat{\mathbf{B}} - \widehat{\mathbf{P}} \mathbf{0}] \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}, \quad \mathbf{y}_r = (\widehat{\mathbf{B}} + \widehat{\mathbf{P}})^T \mathbf{x}_r + \widehat{\mathbf{D}} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}, \quad (19)$$

$$\begin{aligned}
\text{with} \quad \widehat{\mathbf{E}} &= \mathbf{V}^T \mathbf{E}_{11} \mathbf{V}, \quad \widehat{\mathbf{J}} = \mathbf{V}^T \mathbf{J}_{11} \mathbf{V}, \quad \widehat{\mathbf{R}} = \mathbf{V}^T \mathbf{R}_{11} \mathbf{V}, \\
\widehat{\mathbf{B}} &= \frac{1}{2} (\mathbf{V}^T \mathbf{B} + \mathbf{V}^T \mathbf{C}^T), \quad \text{and} \quad \widehat{\mathbf{P}} = \frac{1}{2} (\mathbf{V}^T \mathbf{C}^T - \mathbf{V}^T \mathbf{B}),
\end{aligned} \tag{20}$$

satisfies the interpolation conditions, matches the polynomial part of the transfer function, and preserves the \mathfrak{pH} structure, provided that the reduced passivity matrix $\widehat{\mathbf{W}} = \begin{bmatrix} \widehat{\mathbf{R}} & \widehat{\mathbf{P}} \\ \widehat{\mathbf{P}}^T & \widehat{\mathbf{S}} \end{bmatrix}$ is positive semidefinite.

Proof The proof is similarly to that of Theorem 3. Following [17], the state \mathbf{x}_1 can be decomposed as $\mathbf{x}_1 = \mathbf{x}_c + \mathbf{x}_g$, where $\mathbf{x}_g = \mathbf{E}_{11}^{-1} \mathbf{A}_{12} (\mathbf{J}_{12}^T \mathbf{E}_{11}^{-1} \mathbf{J}_{12})^{-1} (\mathbf{B}_2 - \mathbf{P}_2) \mathbf{u}(t)$ and $\mathbf{x}_c(t)$ satisfies $\mathbf{J}_{12}^T \mathbf{x}_c = 0$. Then, one can rewrite (17) as

$$\begin{aligned}
\Pi \mathbf{E}_{11} \Pi^T \dot{\mathbf{x}}_c(t) &= \Pi \mathbf{A}_{11} \Pi^T \mathbf{x}_c(t) + \Pi \mathbf{B}_1 \mathbf{u}(t), \\
\mathbf{y}(t) &= \mathbf{C} \Pi^T \mathbf{x}_c(t) + \mathcal{D}_0 \mathbf{u}(t) + \mathcal{D}_1 \dot{\mathbf{u}}(t).
\end{aligned} \tag{21}$$

As before, the ODE part can be reduced with usual model reduction techniques. Following [17], however, we achieve this without computing the projector Π explicitly, instead by constructing \mathbf{V} using (18) and then applying one-sided model reduction with \mathbf{V} to obtain the reduced model

$$\widehat{\mathbf{E}} \dot{\mathbf{x}}_r = (\widehat{\mathbf{J}} - \widehat{\mathbf{R}}) \mathbf{x}_r + \widetilde{\mathbf{B}} \mathbf{u}(t), \quad \mathbf{y}_r = \widetilde{\mathbf{C}}^T \mathbf{x}_r + \mathcal{D}_0 \mathbf{u}(t) + \mathcal{D}_1 \dot{\mathbf{u}}(t), \tag{22}$$

where $\widehat{\mathbf{E}} = \mathbf{V}^T \mathbf{E}_{11} \mathbf{V}$, $\widehat{\mathbf{J}} = \mathbf{V}^T \mathbf{J}_{11} \mathbf{V}$, $\widehat{\mathbf{R}} = \mathbf{V}^T \mathbf{R}_{11} \mathbf{V}$, $\widetilde{\mathbf{B}} = \mathbf{V}^T \mathbf{B}$, and $\widetilde{\mathbf{C}} = \mathbf{C} \mathbf{V}$. This reduced model, by construction, satisfies the tangential interpolation conditions. Note that the reduced model in (22) has exactly the same realization as the reduced model in (19) except for the reduced $\widetilde{\mathbf{B}}$ and $\widetilde{\mathbf{C}}$ terms. The reduced terms $\widehat{\mathbf{E}}$, $\widehat{\mathbf{J}}$, and $\widehat{\mathbf{R}}$ in (22) already have the \mathfrak{pH} structure. To recover the port symmetry, we determine matrices $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{P}}$ such that $\widetilde{\mathbf{B}} = \mathbf{V}^T \mathbf{B} = \widehat{\mathbf{B}} - \widehat{\mathbf{P}}$ and $\widetilde{\mathbf{C}} = \mathbf{C} \mathbf{V} = (\widehat{\mathbf{B}} + \widehat{\mathbf{P}})^T$ via

$$\widehat{\mathbf{B}} = \frac{1}{2} (\widetilde{\mathbf{B}} + \widetilde{\mathbf{C}}^T) = \frac{1}{2} \mathbf{V}^T (\mathbf{B} + \mathbf{C}^T) \quad \text{and} \quad \widehat{\mathbf{P}} = \frac{1}{2} (\widetilde{\mathbf{C}}^T - \widetilde{\mathbf{B}}) = \frac{1}{2} \mathbf{V}^T (\mathbf{C}^T - \mathbf{B}),$$

recovering (20). The final requirement to retain the \mathfrak{pH} structure is, then, again that $\begin{bmatrix} \widehat{\mathbf{R}} & \widehat{\mathbf{P}} \\ \widehat{\mathbf{P}}^T & \frac{1}{2} (\mathbf{D} + \mathbf{D}^T) \end{bmatrix} \geq 0$, which is the final condition in the statement of the theorem. \square

This approach has the disadvantage that $\mathbf{u}_2 = \dot{\mathbf{u}}$ is introduced as an extra input and this, in turn, may lead to difficulties when applying standard control and optimization methods. For this reason it is usually preferable to first perform index reduction via an appropriate output feedback (see Sect. 3) and then apply the results from Sect. 4.1. Note that this will change the polynomial part of the transfer function.

4.3 Semi-explicit pHDAE Systems with Index-1 and Index-2 Constraints

Finally, we consider semi-explicit index-2 systems that also have an index-1 part, see [11, 18, 25]. We will only consider the special case,

$$\begin{bmatrix} \mathbf{E}_{11} & \mathbf{E}_{22} & 0 \\ \mathbf{E}_{21} & \mathbf{E}_{22} & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\mathbf{x}}_1 \\ \dot{\mathbf{x}}_2 \\ \dot{\mathbf{x}}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{J}_{11} - \mathbf{R}_{11} & \mathbf{J}_{12} - \mathbf{R}_{12} & \mathbf{J}_{13} \\ \mathbf{J}_{21} - \mathbf{R}_{21} & \mathbf{J}_{22} - \mathbf{R}_{22} & 0 \\ \mathbf{J}_{31} & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \end{bmatrix} + \begin{bmatrix} \mathbf{B}_1 - \mathbf{P}_1 \\ \mathbf{B}_2 - \mathbf{P}_2 \\ 0 \end{bmatrix} \mathbf{u}, \quad (23)$$

$$\mathbf{y} = [(\mathbf{B}_1 + \mathbf{P}_1)^T \quad (\mathbf{B}_2 + \mathbf{P}_2)^T \quad 0] \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \end{bmatrix} + (\mathbf{S} + \mathbf{N})\mathbf{u}, \text{ where } \begin{bmatrix} \mathbf{E}_{11} & \mathbf{E}_{22} \\ \mathbf{E}_{21} & \mathbf{E}_{22} \end{bmatrix} > 0,$$

and both $\mathbf{J}_{22} - \mathbf{R}_{22}$ and \mathbf{J}_{31} are nonsingular.

Theorem 5 Consider an index-2 pHDAE system of the form (23) and construct an interpolatory model reduction basis given by

$$\mathbf{V} = [\mathbf{V}_1^T \quad \mathbf{V}_2^T \quad \mathbf{V}_3^T]^T = [(\sigma_1 \mathbf{E} - \mathbf{A})^{-1} \mathbf{B} \mathbf{b}_1, \quad \dots \quad (\sigma_r \mathbf{E} - \mathbf{A})^{-1} \mathbf{B} \mathbf{b}_r] \in \mathbb{C}^{n \times r},$$

partitioned to conform with the system. Define $\widehat{\mathbf{V}} := \text{diag}(\mathbf{I}, \mathbf{V}_2, \mathbf{I})$. Then, the reduced system

$$\widehat{\mathbf{E}} \dot{\mathbf{x}}_r = (\widehat{\mathbf{J}} - \widehat{\mathbf{R}}) \mathbf{x}_r + \widehat{\mathbf{B}} \mathbf{u}(t), \quad \mathbf{y}_r = \widehat{\mathbf{C}} \mathbf{x}_r + \widehat{\mathbf{D}}, \quad (24)$$

$$\text{with } \widehat{\mathbf{E}} = \widehat{\mathbf{V}}^T \mathbf{E} \widehat{\mathbf{V}}, \quad \widehat{\mathbf{J}} = \widehat{\mathbf{V}}^T \mathbf{J} \widehat{\mathbf{V}}, \quad \widehat{\mathbf{R}} = \widehat{\mathbf{V}}^T \mathbf{R} \widehat{\mathbf{V}},$$

$$\widehat{\mathbf{B}} = \widehat{\mathbf{V}}^T \mathbf{B} = \widehat{\mathbf{V}}^T \mathbf{B} - \widehat{\mathbf{V}}^T \mathbf{P}, \quad \widehat{\mathbf{C}} = \mathbf{C} \widehat{\mathbf{V}} = \mathbf{B}^T \widehat{\mathbf{V}}^T + \mathbf{P}^T \widehat{\mathbf{V}}^T, \quad \text{and } \widehat{\mathbf{D}} = \mathbf{D}, \quad (25)$$

is still pH, matches the polynomial part of the transfer function, and satisfies the tangential interpolation conditions, i.e., $\mathbf{H}(\sigma_i) \mathbf{b}_i = \widehat{\mathbf{H}}(\sigma_i) \mathbf{b}_i$, for $i = 1, 2, \dots, r$.

Proof It follows from the definitions of \mathbf{V} and $\widehat{\mathbf{V}}$ that $\text{span}(\mathbf{V}) \subseteq \text{span}(\widehat{\mathbf{V}})$. Therefore, the resulting reduced system automatically satisfies the interpolation conditions. Since $\widehat{\mathbf{V}}$ does not alter the algebraic constraints, the polynomial part of its transfer function is still \mathbf{D} , matching that of the original model. The reduced system in (24)-(25) is a pHDAE as this one-sided projection retains the original pH structure. \square

5 Algorithmic Considerations

The preceding analysis presumed that interpolation points and tangent directions were specified beforehand. We consider now how one might make choices that generally produce effective approximations with respect to the \mathcal{H}_2 system measure.

5.1 \mathcal{H}_2 -inspired Structure-Preserving Interpolation

The \mathcal{H}_2 distance between the full model $\mathbf{H}(s)$ and the reduced model $\mathbf{H}_r(s)$ is

$$\|\mathbf{H} - \mathbf{H}_r\|_{\mathcal{H}_2} = \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} \|\mathbf{H}(i\omega) - \mathbf{H}_r(i\omega)\|_F^2 d\omega \right)^{1/2},$$

where $i^2 = -1$ and $\|\mathbf{M}\|_F$ denote the Frobenius norm of a matrix \mathbf{M} . To have a finite \mathcal{H}_2 error norm the polynomial parts of $\mathbf{H}_r(s)$ and $\mathbf{H}(s)$ need to match. To make this precise, write $\mathbf{H}(s) = \mathbf{G}(s) + \mathbf{P}(s)$ and $\mathbf{H}_r(s) = \mathbf{G}_r(s) + \mathbf{P}_r(s)$ where $\mathbf{G}(s)$ and $\mathbf{G}_r(s)$ are strictly proper transfer functions, and $\mathbf{P}(s)$ and $\mathbf{P}_r(s)$ are the polynomial parts. Therefore, if $\mathbf{H}_r(s)$ is the \mathcal{H}_2 -optimal approximation to $\mathbf{H}(s)$, then $\mathbf{P}_r(s) = \mathbf{P}(s)$ and $\mathbf{G}_r(s)$ is the \mathcal{H}_2 -optimal approximation to $\mathbf{G}(s)$. This suggests that one decompose $\mathbf{H}(s)$ into its rational and polynomial parts $\mathbf{H}(s) = \mathbf{G}(s) + \mathbf{P}(s)$ and then apply \mathcal{H}_2 optimal reduction to $\mathbf{G}(s)$. However, this requires the explicit construction of $\mathbf{G}(s)$. This problem was resolved in [17] for *unstructured* index-1 and index-2 DAEs. On the other hand, for the ODE case, [16] proposed a pH structure preserving algorithm for minimizing the \mathcal{H}_2 norm. In this section, we aim to unify these two approaches.

First, we briefly revisit the interpolatory \mathcal{H}_2 optimality conditions; for details we refer the reader to [1, 4, 14] and the references therein. Since \mathcal{H}_2 optimality for the DAE case boils down to optimality for the ODE part, we focus on the latter. Let $\mathbf{G}_r(s) = \sum_{i=1}^r \frac{\mathbf{c}_i \mathbf{b}_i^T}{s - \lambda_i}$ be the pole-residue decomposition of $\mathbf{G}_r(s)$. For simplicity we assume simple poles. If $\mathbf{G}_r(s)$ is an \mathcal{H}_2 optimal approximation to $\mathbf{G}(s)$, then $\mathbf{G}(-\lambda_i) \mathbf{b}_i = \mathbf{G}_r(-\lambda_i) \mathbf{b}_i$, $\mathbf{c}_i^T \mathbf{G}(-\lambda_i) = \mathbf{c}_i^T \mathbf{G}_r(-\lambda_i)$, and $\mathbf{c}_i^T \mathbf{G}'(-\lambda_i) \mathbf{b}_i = \mathbf{c}_i^T \mathbf{G}'(-\lambda_i) \mathbf{b}_i$, for $i = 1, 2, \dots, r$ where $'$ denotes the derivative with respect to s . Therefore an \mathcal{H}_2 optimal reduced model is a bitangential Hermite interpolant where the interpolation points are the mirror images of its poles and the tangent directions are the residue directions. Since the optimality conditions depend on the reduced model to be computed, this requires an iterative algorithm, such as the Iterative Rational Krylov Algorithm [14]. For other approaches to \mathcal{H}_2 optimal approximation, see, e.g., [1, 4].

Following [16], to preserve the structure, we will satisfy only a subset of these conditions. We will enforce interpolation points to be the mirror images of the reduced order poles and enforce either left or right-tangential interpolation conditions ($\mathbf{Z} = \mathbf{V}$) without the derivative conditions. However, intuitively, one might expect that in the pH setting this may not cause too much deviation from true optimality, since the input-to-state matrix \mathbf{B} and the state-to-output matrix \mathbf{C} are related.

For simplicity, consider the semi-explicit index-2 case. The other scenarios follow similarly. Starting with the interpolation points $\{\sigma_1, \sigma_2, \dots, \sigma_r\}$ and the tangent directions $\{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_r\}$, construct the interpolatory reduced pHDAE $\widehat{\mathbf{H}}(s)$ as in Theorem 3. Let $\widehat{\mathbf{H}}(s) = \sum_{i=1}^r \frac{\widehat{\mathbf{e}}_i \widehat{\mathbf{b}}_i^T}{s - \lambda_i} + \mathbf{D}$ be the pole-residue decomposition. Since initially the optimality condition $\sigma_i = -\lambda_i (\widehat{\mathbf{J}} - \widehat{\mathbf{R}}, \widehat{\mathbf{E}})$ is not (generally) satisfied, choose

$-\lambda_i(\widehat{\mathbf{J}} - \widehat{\mathbf{R}}, \widehat{\mathbf{E}})$ as the next set of interpolation points and $\widehat{\mathbf{b}}_i$ as the next set of tangent directions. This is repeated until convergence upon which the reduced model $\mathbf{H}_r(s)$ is not only a structure-preserving pHDAE, but also satisfies $\sigma_i = -\lambda_i(\widehat{\mathbf{J}} - \widehat{\mathbf{R}}, \widehat{\mathbf{E}})$.

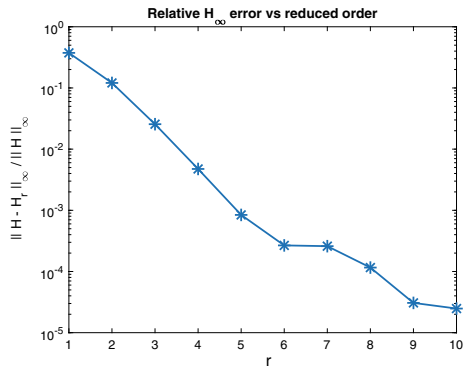
5.2 Numerical Example

We illustrate the discussed procedure with the incompressible fluid flow model of the Oseen equations, from [18, §4.1]:

$$\begin{aligned} \partial_t \mathbf{v} &= -(\mathbf{a} \cdot \nabla) \mathbf{v} + \mu \Delta \mathbf{v} - \nabla p + f & \text{in } \Omega \times (0, T], & \quad \mathbf{v} = 0, & \text{on } \partial\Omega \times (0, T], \\ 0 &= -\operatorname{div} \mathbf{v}, & \text{in } \Omega \times (0, T], & \quad \mathbf{v} = \mathbf{v}^0, & \text{in } \Omega \times 0, \end{aligned}$$

where \mathbf{v} and p are the velocity and pressure variables, $\mu = 1$ is the dynamic viscosity, $\mathbf{a} = [1 \ 1]^T$ is the convective velocity, and $\Omega = (0, 1)^2$ with boundary $\partial\Omega$. f is an externally imposed body force that, for simplicity, is assumed to be separable: $f(x, t) = b(x)u(t)$. A finite-difference discretization on a staggered rectangular grid leads to a single-input/single-output index-2 pHDAE of the form (12), see [18]. In our model, we used a uniform grid with 50 grid points yielding a descriptor system pHDAE of order $n = 7399$, of which $n_1 = 4900$ degrees of freedom are for velocity and $n_2 = 2499$ for pressure. We initialize our \mathcal{H}_2 -based approach with logarithmically spaced interpolation points in the interval $[10^{-2}, 10^4]$ and construct a reduced model of the form (14) with orders $r = 1, 2, \dots, 10$. For every r value, we compute $\|\mathbf{H} - \widehat{\mathbf{H}}\|_\infty / \|\mathbf{H}\|_\infty$ where $\|\mathbf{H}\|_\infty = \sup_{\omega \in \mathbb{R}} |\mathbf{H}(i\omega)|$. Figure 1 shows that the reduced transfer function accurately approximates the original one; for $r = 10$, the relative error is around 10^{-5} .

Fig. 1 Evolution of the model reduction error for Oseen example as r varies



6 Conclusions

We have presented interpolatory model reduction methods for several classes of large scale linear time-invariant port-Hamiltonian differential-algebraic systems. We have shown how constraints can be represented in the transfer function in such a way that the polynomial part can be preserved with interpolatory methods while still retaining important system structure. The results were illustrated with a numerical example from flow control.

Acknowledgements The works of Beattie and Gugercin were supported in parts by NSF through Grant DMS-1819110. The work of Mehrmann was supported by Deutsche Forschungsgemeinschaft through CRC 910 within project A02. Parts of this work were completed while Beattie and Gugercin visited TU Berlin, for which Gugercin acknowledges support through CRC 910 and Beattie acknowledges support through CRC TRR 154 as well as DFG Research Training Group DAEDALUS.

References

1. Antoulas, A.C., Beattie, C.A., Gugercin, S.: Interpolatory methods for model reduction. *Comput. Sci. Eng.* 21. SIAM, Philadelphia (2020)
2. Beattie, C.A., Gugercin, S.: Interpolatory projection methods for structure-preserving model reduction. *Syst. Control Lett.* **58**(3), 225–232 (2009)
3. Beattie, C.A., Gugercin, S.: Structure-preserving model reduction for nonlinear port-Hamiltonian systems. In: 50th IEEE Conference on Decision and Control and European Control Conference, pp. 6564–6569 (2011)
4. Beattie, C.A., Gugercin, S.: Model reduction by rational interpolation. In: Benner, P., Cohen, A., Ohlberger, M., Willcox, K. (eds.) *Model Reduction and Approximation: Theory and Algorithms*, CSE 15, Chap. 7, pp. 297–334. SIAM, Philadelphia (2017)
5. Beattie, C.A., Mehrmann, V., Xu, H., Zwart, H.: Linear port-Hamiltonian descriptor systems. *Math. Control, Signals, Syst.* **30**(4), 1–27 (2018)
6. Binder, A., Mehrmann, V., Miedlar, A., Schulze, P.: A `Matlab` toolbox for the regularization of descriptor systems arising from generalized realization procedures. Preprint 24–2015, Institut für Mathematik, TU Berlin (2015)
7. Bunse-Gerstner, A., Byers, R., Mehrmann, V., Nichols, N.K.: Feedback design for regularizing descriptor systems. *Linear Algebra Appl.* **299**(1–3), 119–151 (1999)
8. Campbell, S.L., Kunkel, P., Mehrmann, V.: Regularization of linear and nonlinear descriptor systems. In: Biegler, L.T., Campbell, S.L., Mehrmann, V. (eds.) *Control and Optimization with Differential-Algebraic Constraints*, *Advances in Design and Control*, vol. 23, pp. 17–36. SIAM (2012)
9. Chaturantabut, S., Beattie, C., Gugercin, S.: Structure-preserving model reduction for nonlinear port-Hamiltonian systems. *SIAM J. Sci. Comput.* **38**(5), B837–B865 (2016)
10. Dai, L.: *Singular Control Systems*. *Lecture Notes in Control and Information Sciences*, vol. 118. Springer, Berlin (1989)
11. Egger, H., Kugler, T., Liljegren-Sailer, B., Marheineke, N., Mehrmann, V.: On structure-preserving model reduction for damped wave propagation in transport networks. *SIAM J. Sci. Comput.* **40**(1), A331–A365 (2018)
12. Gallivan, K., Vandendorpe, A., Van Dooren, P.: Model reduction of MIMO systems via tangential interpolation. *SIAM J. Matrix Anal. Appl.* **26**(2), 328–349 (2005)

13. Gantmacher, F.R.: *The Theory of Matrices*, vol. II. Chelsea Publishing Company, New York, N.Y. (1959)
14. Gugercin, S., Antoulas, A.C., Beattie, C.A.: \mathcal{H}_2 Model Reduction for Large-Scale Linear Dynamical Systems. *SIAM J. Matrix Anal. Appl.* **30**(2), 609–638 (2008)
15. Gugercin, S., Polyuga, R.V., Beattie, C., van der Schaft, A.: Interpolation-based \mathcal{H}_2 model reduction for port-Hamiltonian systems. In: *Proceedings 48th IEEE Conference on Decision and Control*, pp. 5362–5369. IEEE (2009)
16. Gugercin, S., Polyuga, R.V., Beattie, C.A., van der Schaft, A.J.: Structure-preserving tangential interpolation for model reduction of port-Hamiltonian systems. *Automatica* **48**, 1963–1974 (2012)
17. Gugercin, S., Stykel, T., Wyatt, S.: Model reduction of descriptor systems by interpolatory projection methods. *SIAM J. Sci. Comput.* **35**(5), B1010–B1033 (2013)
18. Hauschild, S.-A., Marheineke, N., Mehrmann, V.: Model reduction techniques for linear constant coefficient port-Hamiltonian differential-algebraic systems. *Control and Cybernetics*, To appear (2020). [arXiv:1901.10242](https://arxiv.org/abs/1901.10242)
19. Heinkenschloss, M., Sorensen, D.C., Sun, K.: Balanced truncation model reduction for a class of descriptor systems with application to the Oseen equations. *SIAM J. Sci. Comput.* **30**(2), 1038–1063 (2008)
20. Kunkel, P., Mehrmann, V.: *Differential-Algebraic Equations*. European Mathematical Society (EMS), Zürich (2006)
21. Mayo, A.J., Antoulas, A.C.: A framework for the solution of the generalized realization problem. *Linear Algebra Appl.* **425**(2–3), 634–662 (2007)
22. Mehl, C., Mehrmann, V., Wojtylak, M.: Linear algebra properties of dissipative port-Hamiltonian descriptor systems. *SIAM J. Matrix Anal. Appl.* **39**(3), 1489–1519 (2018)
23. Mehl, C., Mehrmann, V., Wojtylak, M.: Distance problems for dissipative Hamiltonian systems and related matrix polynomials. Technical Report 01-2020, Institut für Mathematik, TU Berlin (2020)
24. Mehrmann, V., Morandin, R.: Structure-preserving discretization for port-hamiltonian descriptor systems. In: *58th IEEE Conference on Decision and Control (CDC)*, Nice, pp. 6863–6868 (2019)
25. Mehrmann, V., Stykel, T.: Balanced truncation model reduction for large-scale systems in descriptor form. In: Benner, P., Mehrmann, V., Sorensen, D.C. (eds.) *Dimension Reduction of Large-Scale Systems*. Lecture Notes in Computational Science and Engineering, vol. 45, pp. 83–115. Springer, Berlin/Heidelberg, Germany (2005)
26. Polyuga, R.V., van der Schaft, A.: Structure-preserving moment matching for port-Hamiltonian systems: Arnoldi and Lanczos. *IEEE Trans. Autom. Control* **56**(6), 1458–1462 (2011)
27. Polyuga, R.V., van der Schaft, A.: Effort- and flow-constraint reduction methods for structure-preserving model reduction of port-Hamiltonian systems. *Sys. Control Lett.* **61**(3), 412–421 (2012)
28. Polyuga, R.V., van der Schaft, A.J.: Structure preserving model reduction of port-Hamiltonian systems by moment matching at infinity. *Automatica* **46**, 665–672 (2010)
29. Stykel, T.: Balanced truncation model reduction for semidiscretized Stokes equation. *Linear Algebra Appl.* **415**(2–3), 262–289 (2006)
30. van der Schaft, A.: Port-Hamiltonian differential-algebraic systems. In: *Surveys in Differential-Algebraic Equations. I*, pp. 173–226. Springer, Heidelberg (2013)
31. Wolf, T., Lohmann, B., Eid, R., Kotyczka, P.: Passivity and structure preserving order reduction of linear port-Hamiltonian systems using Krylov subspaces. *Eur. J. Control* **16**(4), 401–406 (2010)

Data-Driven Identification of Rayleigh-Damped Second-Order Systems



Igor Pontes Duff, Pawan Goyal, and Peter Benner

Abstract In this paper, we present a data-driven approach to identify second-order systems, having internal Rayleigh damping. This means that the damping matrix is given as a linear combination of the mass and stiffness matrices. These systems typically appear when performing various engineering studies, e.g., vibrational and structural analysis. In an experimental set-up, the frequency response of a system can be measured via various approaches, for instance, by measuring the vibrations using an accelerometer. As a consequence, given frequency samples, the identification of the underlying system relies on rational approximation. To that aim, we propose an identification of the corresponding second-order system, extending the Loewner framework for this class of systems. The efficiency of the proposed method is demonstrated by means of various numerical benchmarks.

Keywords Data-driven modeling · Second-order systems · Model reduction · Loewner framework · Mechanical systems

1 Introduction

In this paper, we discuss a data-driven identification framework for a class of second-order (SO) systems of the form:

I. Pontes Duff (✉) · P. Goyal · P. Benner
Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstraße 1, 39106
Magdeburg, Germany
e-mail: pontes@mpi-magdeburg.mpg.de

P. Goyal
e-mail: goyalp@mpi-magdeburg.mpg.de

P. Benner
e-mail: benner@mpi-magdeburg.mpg.de

P. Benner
Technische Universität Chemnitz, Faculty of Mathematics, Reichenhainer Straße 41, 09126
Chemnitz, Germany

$$\Sigma_{\text{SO}} := \begin{cases} \mathbf{M}\ddot{\mathbf{x}}(t) + \mathbf{D}\dot{\mathbf{x}}(t) + \mathbf{K}\mathbf{x}(t) = \mathbf{B}\mathbf{u}(t), \\ \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t), \end{cases} \quad (1)$$

where $\mathbf{x}(t) \in \mathbb{R}^n$ is the state vector, $\mathbf{u}(t) \in \mathbb{R}^m$ are the inputs, $\mathbf{y}(t) \in \mathbb{R}^p$ are the outputs or measurements, and $\mathbf{M}, \mathbf{D}, \mathbf{K} \in \mathbb{R}^{n \times n}$ are, respectively, the mass matrix, the damping matrix and the stiffness matrix, $\mathbf{B} \in \mathbb{R}^{n \times m}$ and $\mathbf{C} \in \mathbb{R}^{p \times n}$. For simplicity, we address the problem for single-input single-output (SISO) systems, i.e., $m = p = 1$. The multiple-input multiple-output (MIMO) generalization is straightforward and can be done by following the lines of the MIMO extension of the classical Loewner framework [20] based on tangential interpolation. Such systems arise in many engineering applications, including vibration analysis [21], structural dynamics [10] and electric circuits. We denote the SO systems (1) by $\Sigma_{\text{SO}} = (\mathbf{M}, \mathbf{D}, \mathbf{K}, \mathbf{B}, \mathbf{C})$. Moreover, we assume a zero inhomogeneous condition, i.e., $\mathbf{x}(0) = \dot{\mathbf{x}}(0) = 0$. Hence, by means of the Laplace transform, the input-output behavior of the system Σ_{SO} is associated with the transfer function as follows:

$$\mathbf{H}_{\text{SO}}(s) = \mathbf{C} (s^2\mathbf{M} + s\mathbf{D} + \mathbf{K})^{-1} \mathbf{B}. \quad (2)$$

Furthermore, throughout the paper, we assume the proportional Rayleigh damping hypothesis, i.e., the damping matrix \mathbf{D} is given by a linear combination of the mass and stiffness matrices:

$$\mathbf{D} = \alpha\mathbf{M} + \beta\mathbf{K}, \quad (3)$$

for $\alpha, \beta \geq 0$. This hypothesis is often considered in engineering application, where the damper is numerically constructed in order to avoid non-dampened oscillations, see [21, Chapter 7] for more details. Recent applications of the Rayleigh damping hypothesis can be found in the fields of guided wave propagation in composite materials [16, 28], earthquake analysis of buildings [11, 12], wind turbine monopiles [6], railway catenary wire systems [24]. Moreover, the Model-Order-Reduction-Wiki (MOR Wiki) presents several large-scale benchmarks of Rayleigh damped second order systems, such as the butterfly gyroscope [25] and the recently developed artificial fishtail [33].

In the past twenty years, model order reduction of SO systems has been investigated extensively; see for instance [8, 22, 29] for balancing-type methods, and [3, 4, 7, 36] for moment matching and \mathcal{H}_2 -optimality based methods. Recently, the authors in [30] provided an extensive comparison among common methods for SO model order reduction applied to a large-scale mechanical artificial fishtail model. In all of the above-mentioned works, the authors suppose that they have access to the matrices, defining the original systems and the reduced-order systems are constructed via Petrov-Galerkin projections. Thus, the main goal is to find projection matrices $\mathbf{V}, \mathbf{W} \in \mathbb{R}^{n \times r}$, leading to the SO reduced-order system

$$\hat{\mathbf{H}}_{\text{SO}}(s) = \hat{\mathbf{C}} \left(s^2\hat{\mathbf{M}} + s\hat{\mathbf{D}} + \hat{\mathbf{K}} \right)^{-1} \hat{\mathbf{B}}, \quad (4)$$

with $\hat{\mathbf{M}} = \mathbf{W}^T \mathbf{M} \mathbf{V}$, $\hat{\mathbf{D}} = \mathbf{W}^T \mathbf{D} \mathbf{V}$, $\hat{\mathbf{K}} = \mathbf{W}^T \mathbf{K} \mathbf{V}$, $\hat{\mathbf{B}} = \mathbf{W}^T \mathbf{B}$ and $\hat{\mathbf{C}} = \mathbf{C} \mathbf{V}$.

However, it is not necessary that the realization is given or is feasible to obtain; thus, we suppose that the original system realization may not be available. Instead, we assume to have access only to frequency domain data, e.g., arising from experiments or numerical simulations. More precisely, we are interested in solving the following problem.

Problem 1 (*SO data-driven identification*) Given interpolation data

$$\{(\sigma_i, \omega_i) \mid \sigma_i \in \mathbb{C} \text{ and } \omega_i \in \mathbb{C}, i = 1, \dots, \rho\}, \quad (5)$$

construct a SO realization $\Sigma_{\text{SO}} = (\mathbf{M}, \mathbf{D}, \mathbf{K}, \mathbf{B}, \mathbf{C})$ of appropriate dimensions, satisfying the proportional Rayleigh damping hypothesis, i.e.,

$$\mathbf{D} = \alpha \mathbf{M} + \beta \mathbf{K},$$

whose transfer function $\mathbf{H}_{\text{SO}}(s) := \mathbf{C}(s^2 \mathbf{M} + s \mathbf{D} + \mathbf{K})^{-1} \mathbf{B}$ satisfies the interpolation conditions, i.e.,

$$\mathbf{H}_{\text{SO}}(\sigma_i) = \omega_i, \quad i = 1, \dots, \rho. \quad (6)$$

Problem 1 corresponds to an identification problem which aims at determining a SO realization that not only interpolates at given measurements, but also satisfies the Rayleigh damping hypothesis. A similar problem for time-delay systems was studied in [27] and [31]. Furthermore, we would like to mention that a data-driven approach for structured non-parametric systems has been studied in [32]. However, the construction of the structured reduced-order system is not a straightforward task.

The purpose of this paper is thus to extend the application domain of the Loewner framework established in [19, 20] to SO systems. With this aim, a new SO Loewner framework is developed, yielding a Rayleigh damped SO system of the form (2) that interpolates at given frequency measurements.

The rest of the paper is organized as follows. Section 2 recalls some preliminary results on the rational interpolation Loewner framework from [20]. Section 3 contains the main contribution of this paper, i.e., the extension of these results to the class of Rayleigh damped SO systems. Firstly, one assumes the knowledge of the Rayleigh damping parameters, α and β , and derives the Loewner matrices for SO systems. Then, a heuristic procedure is presented, originally proposed in [31] in the context of time-delay systems, enabling us to estimate the parameters α and β . Also, we discuss the case of more general damping strategies. Finally, Sect. 4 illustrates the proposed framework by numerical examples and Sect. 5 concludes the paper.

2 Classical Loewner Framework

In this section, we briefly recall the Loewner framework [20]. A first-order (FO) system $\Sigma_{\text{FO}} = (\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C})$ is a dynamical system of the form:

$$\Sigma_{\text{FO}} := \begin{cases} \mathbf{E}\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), & \mathbf{x}(0) = 0, \\ \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t), \end{cases} \quad (7)$$

with $\mathbf{E}, \mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times m}$ and $\mathbf{C} \in \mathbb{R}^{p \times n}$, and the leading dimension n is the order of the system. For clarity of exposition, we focus for now on the single-input single-output (SISO) case, i.e., when $m = p = 1$. The system (7) is associated with the transfer function given by

$$\mathbf{H}_{\text{FO}}(s) = \mathbf{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}. \quad (8)$$

There exist several MOR techniques for first-order systems such as explicit moment matching [35, 37], implicit moment matching using Krylov subspaces [13, 17], Sylvester equations based method [15], extensions for MIMO systems [14]. We refer the reader to the books [2, 5] for more details. However, our goal lies in the identification of linear systems using only the frequency data. Hence, the identification problem, in its SISO form, is stated as follows.

Problem 2 (*First-order data-driven model reduction*) Given interpolation data

$$\{(\sigma_i, \boldsymbol{\omega}_i) \mid \sigma_i \in \mathbb{C} \text{ and } \boldsymbol{\omega}_i \in \mathbb{C}, i = 1, \dots, \rho\} \quad (9)$$

construct a minimal-order realization $\Sigma = (\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C})$ of appropriate dimensions, whose transfer function $\mathbf{H}_{\text{FO}}(s) = \mathbf{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}$ satisfies the interpolation conditions

$$\mathbf{H}_{\text{FO}}(\sigma_i) = \boldsymbol{\omega}_i, i = 1, \dots, \rho. \quad (10)$$

A wide range of methods has been developed to solve Problem 2, e.g., vector fitting [18], the AAA algorithm [23] and the Loewner framework [20]. In this paper, we focus on the latter approach and, in what follows, we recall some of the results contained therein. Firstly, we assume that the number of interpolation data is even, i.e., $\rho = 2\ell$, and as a result, the data can be partitioned in two disjoint sets as follows:

$$\text{right interpolation set } \mathcal{P}_r : \{(\lambda_i, \mathbf{w}_i) \mid \lambda_i \in \mathbb{C} \text{ and } \mathbf{w}_i, i = 1, \dots, \ell\}, \text{ and} \quad (11a)$$

$$\text{left interpolation set } \mathcal{P}_l : \{(\mu_j, \mathbf{v}_j) \mid \mu_j \in \mathbb{C} \text{ and } \mathbf{v}_j \in \mathbb{C}, j = 1, \dots, \ell\}. \quad (11b)$$

Using this partition, we associate the following Loewner matrices.

Definition 1 (Loewner matrices [20]) Given the right \mathcal{P}_r and left \mathcal{P}_l interpolation sets, we associate them with the Loewner matrix \mathbb{L} and shifted Loewner matrix \mathbb{L}_σ given by

$$\mathbb{L} = \begin{pmatrix} \frac{\mathbf{v}_1 - \mathbf{w}_1}{\mu_1 - \lambda_1} & \dots & \frac{\mathbf{v}_1 - \mathbf{w}_\ell}{\mu_1 - \lambda_\ell} \\ \vdots & \ddots & \vdots \\ \frac{\mathbf{v}_\ell - \mathbf{w}_1}{\mu_\ell - \lambda_1} & \dots & \frac{\mathbf{v}_\ell - \mathbf{w}_\ell}{\mu_\ell - \lambda_\ell} \end{pmatrix}, \quad \mathbb{L}_\sigma = \begin{pmatrix} \frac{\mu_1 \mathbf{v}_1 - \lambda_1 \mathbf{w}_1}{\mu_1 - \lambda_1} & \dots & \frac{\mu_1 \mathbf{v}_1 - \lambda_\ell \mathbf{w}_\ell}{\mu_1 - \lambda_\ell} \\ \vdots & \ddots & \vdots \\ \frac{\mu_\ell \mathbf{v}_\ell - \lambda_1 \mathbf{w}_1}{\mu_\ell - \lambda_1} & \dots & \frac{\mu_\ell \mathbf{v}_\ell - \lambda_\ell \mathbf{w}_\ell}{\mu_\ell - \lambda_\ell} \end{pmatrix}. \quad (12)$$

Remark 1 The Loewner matrix \mathbb{L} was introduced in [1]. As shown therein, its usefulness derives from the fact that its rank is equal to the order of the minimal realization \mathbf{H}_{FO} satisfying the interpolation conditions in (10). Hence, it reveals the complexity of the reduced-order model solving Problem 2.

Next, let us introduce the following matrices associated with the interpolation problem as follows:

$$\begin{cases} \mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_\ell) \in \mathbb{C}^{\ell \times \ell} \\ \hat{\mathbf{H}}(\mathbf{\Lambda}) = [\mathbf{w}_1 \dots \mathbf{w}_\ell]^T \in \mathbb{C}^{\ell \times 1} \end{cases} \quad \text{and} \quad \begin{cases} \mathbf{M} = \text{diag}(\mu_1, \dots, \mu_\ell) \in \mathbb{C}^{\ell \times \ell} \\ \hat{\mathbf{H}}(\mathbf{M}) = [\mathbf{v}_1 \dots \mathbf{v}_\ell]^T \in \mathbb{C}^{\ell \times 1} \end{cases} \quad (13)$$

Also, let $\mathbb{1} \in \mathbb{R}^{\ell \times 1}$ be the column vector with all entries equal to one. Hence, the Loewner matrices satisfy the following Sylvester equations

$$\mathbf{M}\mathbb{L} - \mathbb{L}\mathbf{\Lambda} = \hat{\mathbf{H}}(\mathbf{M})\mathbb{1}^T - \mathbb{1}\hat{\mathbf{H}}(\mathbf{\Lambda})^T, \quad \text{and} \quad (14a)$$

$$\mathbf{M}\mathbb{L}_\sigma - \mathbb{L}_\sigma\mathbf{\Lambda} = \mathbf{M}\hat{\mathbf{H}}(\mathbf{M})\mathbb{1}^T - \mathbb{1}\hat{\mathbf{H}}(\mathbf{\Lambda})\mathbf{\Lambda}. \quad (14b)$$

An elegant solution for Problem 2 based on the Loewner pair $(\mathbb{L}, \mathbb{L}_\sigma)$ was proposed in [20]. This is summarized in the following theorem.

Theorem 1 (Loewner framework [20]) Let \mathbb{L} and \mathbb{L}_σ be the Loewner matrices associated with the partition in (13). If $(\mathbb{L}_\sigma, \mathbb{L})$ is a regular pencil with no μ_i or λ_j being an eigenvalue, then the matrices

$$\hat{\mathbf{E}} = -\mathbb{L}, \quad \hat{\mathbf{A}} = -\mathbb{L}_\sigma, \quad \hat{\mathbf{B}} = \hat{\mathbf{H}}(\mathbf{M}), \quad \hat{\mathbf{C}} = \hat{\mathbf{H}}(\mathbf{\Lambda})^T,$$

provides a realization $\hat{\Sigma}_{FO} = (\hat{\mathbf{E}}, \hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}})$ for a minimal order interpolant of Problem 2, i.e., the transfer function

$$\hat{\mathbf{H}}_{FO}(s) = \mathbf{W}(s\mathbb{L}_\sigma - \mathbb{L})^{-1}\mathbf{V}$$

satisfies the interpolation conditions in (10).

Theorem 1 allows to obtain a FO system $\hat{\mathbf{H}} = (\hat{\mathbf{E}}, \hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}})$ whose transfer function interpolates right and left data as stated in Problem 2. However, when more data than necessary are provided, then the hypothesis of Theorem 1 may not be satisfied. Hence, a singular-value decomposition (SVD) based procedure has been proposed in [20] to find an FO system interpolating the frequency data.

Next, recall that a SO system $\Sigma_{\text{SO}} = (\mathbf{M}, \mathbf{D}, \mathbf{K}, \mathbf{B}, \mathbf{C})$ can be written as a first-order realization, for instance, as follows:

$$\mathbf{H}_{\text{SO_FO}}(s) = \mathbf{C}(s\mathbf{E} - \mathcal{A})^{-1}\mathcal{B},$$

where

$$\mathcal{E} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{M} \end{bmatrix}, \quad \mathcal{A} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{K} & -\mathbf{D} \end{bmatrix}, \quad \mathcal{B} = \begin{bmatrix} \mathbf{0} \\ \mathbf{B} \end{bmatrix} \quad \text{and} \quad \mathbf{C} = [\mathbf{C} \ \mathbf{0}].$$

As a consequence, the classical Loewner framework presented in Sect. 2 can be employed to find a first-order realization. However, the intrinsic SO structure will not be preserved in the identified model. But the classical Loewner framework yields an information about the order of a SO realization fitting the data, which is outlined in the following remark.

Remark 2 (*Order of SO model*) Let us suppose that the frequency data in Problem 2 and let \mathbb{L} be a Loewner matrix given in (12) constructed with this data. Then, the order of the SO system fitting the data equals $\frac{1}{2}\text{rank}(\mathbb{L})$.

In the following section, we discuss an extension of the Loewner framework for the class of Rayleigh damped SO systems.

3 Second-Order Loewner Framework

This section contains our main contribution, which presents an extension of the Loewner framework to the class of SO Rayleigh damped systems (1). Here, we also assume that the number of interpolation data is even, i.e., $\rho = 2\ell$, and the data is partitioned into two disjoint sets as in (11a) and (11b). Moreover, the data is organized into the matrices $\mathbf{\Lambda}$, $\hat{\mathbf{H}}(\mathbf{\Lambda})$, \mathcal{M} , $\hat{\mathbf{H}}(\mathcal{M})$ as in (13). This section is divided into four parts. The first one aims at finding the matrix equations which encode the general SO systems satisfying the interpolation conditions. In the second one, we assume to have a priori knowledge of the Rayleigh damping parameters α and β and we derive the equivalents of the Loewner matrices (12) and Theorem 1 to the class of SO Rayleigh damped systems. The third part is dedicated to proposing a heuristic procedure to estimate the parameters α and β using the frequency data available. Finally, in the fourth part, we briefly discuss the case of more general dampings.

3.1 Second-Order Loewner Matrices

In what follows, we assume that Problem 1 has a minimal order r solution \mathbf{H}_{SO}^* , given by

$$\mathbf{H}_{\text{SO}}^*(s) = \mathbf{C}^* (s^2 \mathbf{M}^* + s \mathbf{D}^* + \mathbf{K}^*)^{-1} \mathbf{B}^*, \quad (15)$$

with $\mathbf{D}^* = \alpha \mathbf{M}^* + \beta \mathbf{K}^*$. Here, we also assume that the coefficients α and β from the Rayleigh damping hypothesis are known. Then, later in this section, we will show how to construct a realization equivalent to $\mathbf{H}_{\text{SO}}^*(s)$ that only depends on the frequency data. To that aim, let us first recall a result from [4] enabling projection-based structured preserving model reduction.

Theorem 2 (Structure preserving SO model reduction [4]) *Consider the SO transfer function $\mathbf{H}_{\text{SO}}(s)$ as given in (2). For given interpolation points λ_i and μ_i , $i \in \{1, \dots, \ell\}$, let the projection matrices \mathbf{V} and \mathbf{W} be as follows:*

$$\mathbf{V} = \left[(\lambda_1^2 \mathbf{M} + \lambda_1 \mathbf{D} + \mathbf{K})^{-1} \mathbf{B}, \dots, (\lambda_\ell^2 \mathbf{M} + \lambda_\ell \mathbf{D} + \mathbf{K})^{-1} \mathbf{B} \right] \quad (16a)$$

$$\mathbf{W} = \left[(\mu_1^2 \mathbf{M} + \mu_1 \mathbf{D} + \mathbf{K})^{-T} \mathbf{C}^T, \dots, (\mu_\ell^2 \mathbf{M} + \mu_\ell \mathbf{D} + \mathbf{K})^{-T} \mathbf{C}^T \right] \quad (16b)$$

Then, the reduced-order model $\hat{\mathbf{H}}_{\text{SO}}(s)$ constructed by Petrov-Galerkin projection as in (4) satisfies the interpolation conditions

$$\mathbf{H}_{\text{SO}}(\lambda_i) = \hat{\mathbf{H}}_{\text{SO}}(\lambda_i) \quad \text{and} \quad \mathbf{H}_{\text{SO}}(\mu_i) = \hat{\mathbf{H}}_{\text{SO}}(\mu_i), \quad \text{for } i = 1, \dots, \ell.$$

The above theorem allows us to construct a SO reduced-order model by interpolation. Let us apply this theorem to the SO system $\mathbf{H}_{\text{SO}}^*(s)$ (15). For this, we will construct the matrix \mathbf{V} using the interpolation points in $\mathbf{\Lambda}$, and the matrix \mathbf{W} using the interpolation points in \mathbf{M} . As a consequence, \mathbf{V} and \mathbf{W} are, respectively, the solutions of the following matrix equations

$$\mathbf{M}^* \mathbf{V} \mathbf{\Lambda}^2 + \mathbf{D}^* \mathbf{V} \mathbf{\Lambda} + \mathbf{K}^* \mathbf{V} = \mathbf{B}^* \mathbb{1}^T, \quad \text{and} \quad (17a)$$

$$\mathbf{M}^2 \mathbf{W}^T \mathbf{M}^* + \mathbf{M} \mathbf{W}^T \mathbf{D}^* + \mathbf{W}^T \mathbf{K}^* = \mathbb{1} \mathbf{C}^*, \quad (17b)$$

Multiplying Eq. (17a) on the left by \mathbf{W}^T and Eq. (17b) on the right by \mathbf{V} , respectively, one obtains

$$\begin{aligned} \mathbf{W}^T \mathbf{M}^* \mathbf{V} \mathbf{\Lambda}^2 + \mathbf{W}^T \mathbf{D}^* \mathbf{V} \mathbf{\Lambda} + \mathbf{W}^T \mathbf{K}^* \mathbf{V} &= \mathbf{W}^T \mathbf{B}^* \mathbb{1}^T, \\ \mathbf{M}^2 \mathbf{W}^T \mathbf{M}^* \mathbf{V} + \mathbf{M} \mathbf{W}^T \mathbf{D}^* \mathbf{V} + \mathbf{W}^T \mathbf{K}^* \mathbf{V} &= \mathbb{1} \mathbf{C}^* \mathbf{V}. \end{aligned}$$

If we set

$$\hat{\mathbf{M}} = \mathbf{W}^T \mathbf{M}^* \mathbf{V}, \quad \hat{\mathbf{D}} = \mathbf{W}^T \mathbf{D}^* \mathbf{V}, \quad \hat{\mathbf{K}} = \mathbf{W}^T \mathbf{K}^* \mathbf{V}, \quad (18a)$$

$$\hat{\mathbf{B}} = \mathbf{W}^T \mathbf{B}^* = \hat{\mathbf{H}}(\mathcal{M}), \quad \text{and} \quad \hat{\mathbf{C}} = \mathbf{C}^* \mathbf{V} = \hat{\mathbf{H}}(\Lambda)^T, \quad (18b)$$

then the SO system $\hat{\mathbf{H}}_{\text{SO}} = (\hat{\mathbf{M}}, \hat{\mathbf{D}}, \hat{\mathbf{K}}, \hat{\mathbf{B}}, \hat{\mathbf{C}})$ is the reduced-order model obtained by Theorem 2, satisfying the interpolation conditions from Problem 1. Hence, we can rewrite the above equations as follows:

$$\hat{\mathbf{M}}\Lambda^2 + \hat{\mathbf{D}}\Lambda + \hat{\mathbf{K}} = \hat{\mathbf{H}}(\mathcal{M})\mathbb{1}^T, \quad (19a)$$

$$\mathcal{M}^2 \hat{\mathbf{M}} + \mathcal{M} \hat{\mathbf{D}} + \hat{\mathbf{K}} = \mathbb{1} \hat{\mathbf{H}}(\Lambda)^T. \quad (19b)$$

Equations in (19) give a relationship between the matrices $\hat{\mathbf{M}}, \hat{\mathbf{D}}$ and $\hat{\mathbf{K}}$, which are the unknowns we are willing to identify, and the interpolation data encoded by $\Lambda, \hat{\mathbf{H}}(\Lambda), \hat{\mathbf{M}}$ and $\hat{\mathbf{H}}(\mathcal{M})$. However, as discussed in [32], it corresponds to an underdetermined system on the unknowns.

3.2 Identification of Rayleigh-Damped Second-Order Systems

From now on, we will assume the Rayleigh damping hypothesis, i.e., $\hat{\mathbf{D}} = \alpha \hat{\mathbf{M}} + \beta \hat{\mathbf{K}}$. As a consequence, the system of equations in (19) becomes

$$\hat{\mathbf{M}}(\Lambda^2 + \alpha\Lambda) + \hat{\mathbf{K}}(\beta\Lambda + \mathbf{I}) = \hat{\mathbf{H}}(\mathcal{M})\mathbb{1}^T, \quad (20a)$$

$$(\mathcal{M}^2 + \alpha\mathcal{M})\hat{\mathbf{M}} + (\beta\mathcal{M} + \mathbf{I})\hat{\mathbf{K}} = \mathbb{1}\hat{\mathbf{H}}(\Lambda)^T. \quad (20b)$$

Notice that the above equations can be solved for $\hat{\mathbf{M}}$ and $\hat{\mathbf{K}}$. However, in order to have an analytic expression for the matrices of the reduced-order system in a similar way as for the Loewner matrices (12), we need to introduce the following change of variables:

$$\mathbb{L}^{\text{SO}} := -(\mathbf{I} + \beta\mathcal{M})\hat{\mathbf{M}}(\mathbf{I} + \beta\Lambda), \quad \mathbb{L}_\sigma^{\text{SO}} := (\mathbf{I} + \beta\mathcal{M})\hat{\mathbf{K}}(\mathbf{I} + \beta\Lambda), \quad (21a)$$

$$\hat{\mathbf{B}}^{\text{SO}} := (\mathbf{I} + \beta\mathcal{M})\hat{\mathbf{H}}(\mathcal{M}), \quad \text{and} \quad \hat{\mathbf{C}}^{\text{SO}} := \hat{\mathbf{H}}(\Lambda)^T(\mathbf{I} + \beta\Lambda). \quad (21b)$$

Notice that the two realizations

$$\hat{\Sigma}_{\text{SO}} = (\hat{\mathbf{M}}, \alpha\hat{\mathbf{M}} + \beta\hat{\mathbf{K}}, \hat{\mathbf{K}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}) \quad \text{and} \quad \hat{\Sigma}_{\text{SO}}^{\text{Loew}} = (-\mathbb{L}^{\text{SO}}, -\alpha\mathbb{L}^{\text{SO}} + \beta\mathbb{L}_\sigma^{\text{SO}}, \mathbb{L}_\sigma^{\text{SO}}, \hat{\mathbf{B}}^{\text{SO}}, \hat{\mathbf{C}}^{\text{SO}})$$

are equivalent, i.e., they represent the same transfer function. Hence, the realization $\hat{\mathbf{H}}_{\text{SO}}^{\text{Loew}}$ also satisfies the interpolation conditions from Problem 1. Additionally, by a simple computation, we obtain that the matrices \mathbb{L}^{SO} and $\mathbb{L}_\sigma^{\text{SO}}$ satisfy the following equations

$$\begin{aligned}\mathbb{L}^{\text{SO}}\mathcal{F}(\mathbf{\Lambda}) - \mathbb{L}_{\sigma}^{\text{SO}} &= -\mathcal{D}(\mathbf{M})\hat{\mathbf{H}}(\mathbf{M})\mathbb{1}^T, \\ \mathcal{F}(\mathbf{M})\mathbb{L}^{\text{SO}} - \mathbb{L}_{\sigma}^{\text{SO}} &= -\mathbb{1}\hat{\mathbf{H}}(\mathbf{\Lambda})^T\mathcal{D}(\mathbf{\Lambda}),\end{aligned}$$

where, for a given matrix $\mathbf{\Omega}$, $\mathcal{F}(\mathbf{\Omega}) := (\mathbf{I} + \beta\mathbf{\Omega})^{-1}(\mathbf{\Omega}^2 + \alpha\mathbf{\Omega})$ and $\mathcal{D}(\mathbf{\Omega}) := (\mathbf{I} + \beta\mathbf{\Omega})$. As a consequence,

$$\mathbb{L}^{\text{SO}}\mathcal{F}(\mathbf{\Lambda}) - \mathcal{F}(\mathbf{M})\mathbb{L}^{\text{SO}} = \mathbb{1}\hat{\mathbf{H}}(\mathbf{\Lambda})^T\mathcal{D}(\mathbf{\Lambda}) - \mathcal{D}(\mathbf{M})\hat{\mathbf{H}}(\mathbf{M})\mathbb{1}^T, \quad (22a)$$

$$\mathbb{L}_{\sigma}^{\text{SO}}\mathcal{F}(\mathbf{\Lambda}) - \mathcal{F}(\mathbf{M})\mathbb{L}_{\sigma}^{\text{SO}} = \mathbb{1}\hat{\mathbf{H}}(\mathbf{\Lambda})^T\mathcal{N}(\mathbf{\Lambda}) - \mathcal{N}(\mathbf{M})\hat{\mathbf{H}}(\mathbf{M})\mathbb{1}^T, \quad (22b)$$

where, for a given matrix $\mathbf{\Omega}$, $\mathcal{N}(\mathbf{\Omega}) := (\mathbf{\Omega}^2 + \alpha\mathbf{\Omega})$. Notice that the Sylvester equations (22) are equivalent to (14a) for the case of SO systems. Hence, using those equations, one can derive analytic expressions of \mathbb{L}^{SO} and $\mathbb{L}_{\sigma}^{\text{SO}}$.

Definition 2 (SO Loewner matrices) Let us suppose α and β are known and let

$$d(s) := 1 + s\beta, \quad n(s) := s^2 + \alpha s, \quad \text{and} \quad f(s) := \frac{n(s)}{d(s)},$$

be scalar functions. Then the SO Loewner matrices, namely, the SO Loewner matrix \mathbb{L}^{SO} and the shifted Loewner matrix $\mathbb{L}_{\sigma}^{\text{SO}}$ are given by

$$\mathbb{L}^{\text{SO}} = \begin{pmatrix} \frac{d(\mu_1)\mathbf{v}_1 - d(\lambda_1)\mathbf{w}_1}{f(\mu_1) - f(\lambda_1)} & \cdots & \frac{d(\mu_1)\mathbf{v}_1 - d(\lambda_\ell)\mathbf{w}_\ell}{f(\mu_1) - f(\lambda_\ell)} \\ \vdots & \ddots & \vdots \\ \frac{d(\mu_\ell)\mathbf{v}_\ell - d(\lambda_1)\mathbf{w}_1}{f(\mu_\ell) - f(\lambda_1)} & \cdots & \frac{d(\mu_\ell)\mathbf{v}_\ell - d(\lambda_\ell)\mathbf{w}_\ell}{f(\mu_\ell) - f(\lambda_\ell)} \end{pmatrix}, \quad (23)$$

$$\mathbb{L}_{\sigma}^{\text{SO}} = \begin{pmatrix} \frac{n(\mu_1)\mathbf{v}_1 - n(\lambda_1)\mathbf{w}_1}{f(\mu_1) - f(\lambda_1)} & \cdots & \frac{n(\mu_1)\mathbf{v}_1 - n(\lambda_\ell)\mathbf{w}_\ell}{f(\mu_1) - f(\lambda_\ell)} \\ \vdots & \ddots & \vdots \\ \frac{n(\mu_\ell)\mathbf{v}_\ell - n(\lambda_1)\mathbf{w}_1}{f(\mu_\ell) - f(\lambda_1)} & \cdots & \frac{n(\mu_\ell)\mathbf{v}_\ell - n(\lambda_\ell)\mathbf{w}_\ell}{f(\mu_\ell) - f(\lambda_\ell)} \end{pmatrix}. \quad (24)$$

Moreover, by construction

$$\mathbb{L}^{\text{SO}} = -(\mathbf{I} + \beta\mathbf{M})\hat{\mathbf{M}}(\mathbf{I} + \beta\mathbf{\Lambda}) = -(\mathbf{I} + \beta\mathbf{M})\mathbf{W}^T\mathbf{M}^*\mathbf{V}(\mathbf{I} + \beta\mathbf{\Lambda}).$$

Thus, the following remark holds.

Remark 3 If we have sufficient interpolation data, then $\text{rank}(\mathbf{V}) = \text{rank}(\mathbf{W}) = r$. As a consequence, the rank of the SO Loewner matrix \mathbb{L}^{SO} gives us the order of the Rayleigh damped SO minimal realization interpolating the points, since

$\text{rank}(\mathbb{L}^{SO}) = \text{rank}(\mathbf{W}^T \mathbf{M}^* \mathbf{V}) = \text{rank}(\mathbf{M}^*) = \text{order of the minimal SO interpolant.}$

We are now able to state the analogue result to Theorem 1 for Rayleigh-damped SO systems.

Theorem 3 (SO data-driven identification) *Assume that $\mu_i \neq \lambda_j$ for all $i, j = 1, \dots, \ell$. Additionally, suppose that $(s^2 + \alpha s)\mathbb{L}^{SO} + (\beta s + 1)\mathbb{L}_\sigma^{SO}$ is invertible for all $s = \{\lambda_1, \dots, \lambda_\ell\} \cup \{\mu_1, \dots, \mu_\ell\}$. Then,*

$$\hat{\mathbf{M}} = -\mathbb{L}^{SO}, \quad \hat{\mathbf{K}} = \mathbb{L}_\sigma^{SO}, \quad \hat{\mathbf{B}}^{SO} = (\mathbf{I} + \beta \mathbf{M}) \hat{\mathbf{H}}(\mathbf{M}) \quad \hat{\mathbf{C}}^{SO} = \hat{\mathbf{H}}(\mathbf{\Lambda})^T (\mathbf{I} + \beta \mathbf{\Lambda}),$$

and $\hat{\mathbf{D}} = \alpha \hat{\mathbf{M}} + \beta \hat{\mathbf{K}}$ satisfy the interpolation conditions from Problem 1.

Theorem 3 allow us to identify a Rayleigh-damped SO from a given interpolation data set (5). As a consequence, it gives a solution for the Problem 1 whenever the constants α and β are assumed to be known.

We now consider the case where more data than necessary are provided, which is realistic for applications. In this case, the assumptions of the above theorem are not satisfied; thus, one needs to project onto the column span and the row span of a linear combination of the two Loewner matrices. More precisely, let the following assumption be satisfied:

$$\text{rank}\left(\begin{bmatrix} \mathbb{L}^{SO} & \mathbb{L}_\sigma^{SO} \end{bmatrix}\right) = \text{rank}\left(\begin{bmatrix} \mathbb{L}^{SO} \\ \mathbb{L}_\sigma^{SO} \end{bmatrix}\right) = r \quad (25)$$

Then, we consider the compact SVDs

$$\begin{bmatrix} \mathbb{L}^{SO} & \mathbb{L}_\sigma^{SO} \end{bmatrix} = \mathbf{Y}_\rho \Sigma_l \tilde{\mathbf{V}}^T \quad \text{and} \quad \begin{bmatrix} \mathbb{L}^{SO} \\ \mathbb{L}_\sigma^{SO} \end{bmatrix} = \tilde{\mathbf{W}} \Sigma_r \mathbf{X}_\rho^T. \quad (26)$$

Using the projection matrices \mathbf{V}_ρ and \mathbf{W}_ρ , we are able to remove the redundancy in the data by means of the following result.

Theorem 4 *The SO realization $\hat{\Sigma}_{SO} = (\hat{\mathbf{M}}, \hat{\mathbf{D}}, \hat{\mathbf{K}}, \hat{\mathbf{B}}, \hat{\mathbf{C}})$ of a minimal interpolant of Problem 1 is given as:*

$$\hat{\mathbf{M}} = -\mathbf{Y}_\rho^T \mathbb{L}^{SO} \mathbf{X}_\rho, \quad \hat{\mathbf{K}} = \mathbf{Y}_\rho^T \mathbb{L}_\sigma^{SO} \mathbf{X}_\rho, \quad \hat{\mathbf{D}} = \alpha \hat{\mathbf{M}} + \beta \hat{\mathbf{K}}, \quad (27a)$$

$$\hat{\mathbf{B}} = \mathbf{Y}_\rho^T \hat{\mathbf{B}}^{SO}, \quad \text{and} \quad \hat{\mathbf{C}} = \hat{\mathbf{C}}^{SO} \mathbf{X}_\rho. \quad (27b)$$

Depending on whether r in (25) is the exact or approximate rank, we obtain either an interpolant or an approximate interpolant of the data, respectively.

3.3 Optimizing Parameters

In the previous section, we have shown how to construct a SO realization for given transfer function measurements and a priori knowledge of the parameters α and β from the Rayleigh damping hypothesis. However, there are several cases, where exact values of α and β are not known but we rather can have a hint of the range for the parameters, i.e., $\alpha \in \mathcal{R}_\alpha$ and $\beta \in \mathcal{R}_\beta$. Therefore, as done for delay systems in [31], we also propose a heuristic optimization approach to obtain the parameters α and β for SO systems, satisfying the Rayleigh damping hypothesis. For this purpose, we split the data training $\mathcal{D}_{\text{training}}$ and test set $\mathcal{D}_{\text{test}}$, e.g., in the ratio 80:20. Hence, we ideally aim at solving the optimization as follows:

$$\min_{\alpha \in \mathcal{R}_\alpha, \beta \in \mathcal{R}_\beta} \mathcal{J}(\alpha, \beta) \quad (28)$$

where

$$\mathcal{J}(\alpha, \beta) := \sum_{(\sigma_k, v_k) \in \mathcal{D}_{\text{test}}} \left\| \hat{\mathbf{H}}_{\text{SO}}(\sigma_k \alpha, \beta) - v_k \right\|^2 + \sum_{(\mu_k, w_k) \in \mathcal{D}_{\text{test}}} \left\| \hat{\mathbf{H}}_{\text{SO}}(\mu_k \alpha, \beta) - w_k \right\|^2,$$

where $\hat{\mathbf{H}}_{\text{SO}}$ is constructed using only the training data. However, the optimization problem (28) is non-convex, and solving it is a challenging task. Therefore, we seek to solve a relaxed problem. For this purpose, in the paper, we make a 2-D grid for the parameters α and β in given intervals. Then, we seek to determine the parameters on the grid where the function $\mathcal{J}(\alpha, \beta)$ is minimized. Nonetheless, solving the optimization problem (28) needs future investigation and so we leave it as a possible future research problem.

3.4 A Note on More General Damping Structures

As stated before, the general SO interpolating system has its matrices $\hat{\mathbf{M}}$, $\hat{\mathbf{D}}$ and $\hat{\mathbf{K}}$ satisfying an underdetermined system given by (19). Once we assume the Rayleigh damping hypothesis, the system of equations becomes determined and an analytical solution is given by the SO Loewner matrices \mathbb{L}^{SO} and $\mathbb{L}_\sigma^{\text{SO}}$.

In a more general set-up, we may consider the damping to have a more general structure depending on the matrices $\hat{\mathbf{M}}$ and $\hat{\mathbf{K}}$. One possible way is to assume the damping matrix $\hat{\mathbf{D}}$ to be described as

$$\hat{\mathbf{D}} = \hat{\mathbf{M}}^{\frac{1}{2}} g \left(\hat{\mathbf{M}}^{-\frac{1}{2}} \hat{\mathbf{K}} \hat{\mathbf{M}}^{-\frac{1}{2}} \right) \hat{\mathbf{M}}^{\frac{1}{2}}, \quad (29)$$

where g is an analytic function on the positive real axis. It is worth noticing that the Rayleigh damping hypothesis is obtained if we consider $g(s) = \alpha + \beta s$. Another

usual choice of damping is the so-called critical damping given by

$$\hat{\mathbf{D}}_{Crit} = 2\alpha\hat{\mathbf{M}}^{\frac{1}{2}} \left(\hat{\mathbf{M}}^{-\frac{1}{2}} \hat{\mathbf{K}} \hat{\mathbf{M}}^{-\frac{1}{2}} \right)^{\frac{1}{2}} \hat{\mathbf{M}}^{\frac{1}{2}},$$

with $\alpha \in [0.02, 0.1]$, see [26, 34]. In this case, the damping is obtained with $g(s) = 2\alpha\sqrt{s}$.

Hence, if we assume the damping to be given as in (29), then the system of equations in (19) becomes

$$\hat{\mathbf{M}}\boldsymbol{\Lambda}^2 + \hat{\mathbf{M}}^{\frac{1}{2}}g \left(\hat{\mathbf{M}}^{-\frac{1}{2}}\hat{\mathbf{K}}\hat{\mathbf{M}}^{-\frac{1}{2}} \right) \hat{\mathbf{M}}^{\frac{1}{2}}\boldsymbol{\Lambda} + \hat{\mathbf{K}} = \hat{\mathbf{H}}(\mathcal{M})\mathbb{1}^T, \quad (30a)$$

$$\mathcal{M}^2\hat{\mathbf{M}} + \mathcal{M}\hat{\mathbf{M}}^{\frac{1}{2}}g \left(\hat{\mathbf{M}}^{-\frac{1}{2}}\hat{\mathbf{K}}\hat{\mathbf{M}}^{-\frac{1}{2}} \right) \hat{\mathbf{M}}^{\frac{1}{2}} + \hat{\mathbf{K}} = \mathbb{1}\hat{\mathbf{H}}(\boldsymbol{\Lambda})^T. \quad (30b)$$

Notice that in system (30), the number of unknowns matches the number of equations. However, depending on the nature of the function g , this will lead to a system of nonlinear equations and further investigation is needed to find conditions for the existence (and uniqueness) of a solution and numerical solvers for nonlinear matrix equations. Additionally, in the general case, the Petrov-Galerkin approximation of the damping model would no longer be the damping model of the Petrov-Galerkin approximation, except for the very restricted case where

$$\mathbf{W}^T g(\mathbf{Q}) \mathbf{V} = g(\mathbf{W}^T \mathbf{Q} \mathbf{V})$$

with $\mathbf{V}, \mathbf{W} \in \mathbb{R}^{n \times r}$ being the projection matrices and $\mathbf{Q} \in \mathbb{R}^{n \times n}$. Notice that this happens for the Rayleigh damping, but it is non longer true for the critical damping. Future work will be dedicated on further analysis and numerical methods for these general dampings.

4 Numerical Results

In this section, we illustrate the efficiency of the proposed methods via several numerical examples, arising in various applications. All the simulations are done on a CPU 2.6GHz intel® Core™i5, 8 GB 1600MHz DDR3, MATLAB® 9.1.0.441655 (R2016b).

4.1 Demo Example

At first, we discuss an artificial example to illustrate the proposed method. Let us consider a SO system of order $n = 2$, $\boldsymbol{\Sigma}_{SO} = (\mathbf{M}, \mathbf{D}, \mathbf{K}, \mathbf{B}, \mathbf{C})$ whose matrices are given by:

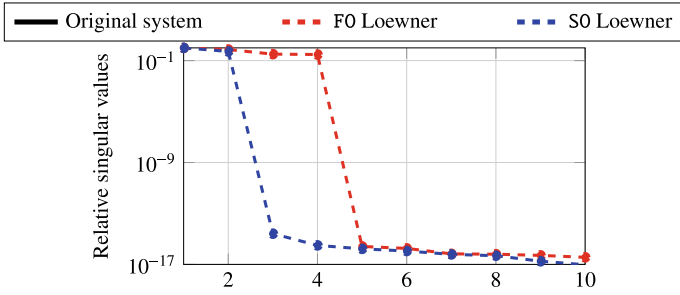


Fig. 1 Demo example: Decay of the singular values for the FO and SO Loewner matrices

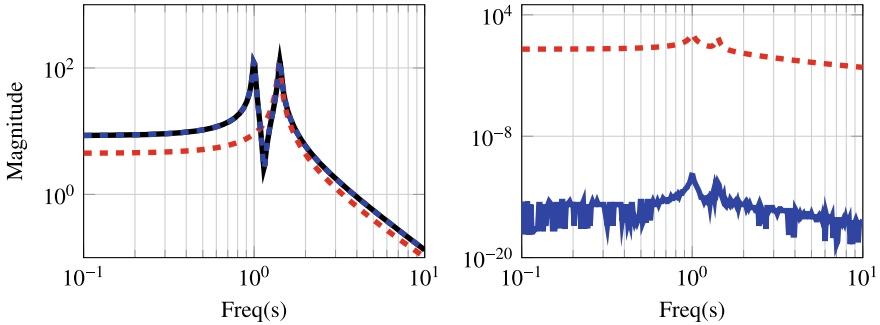


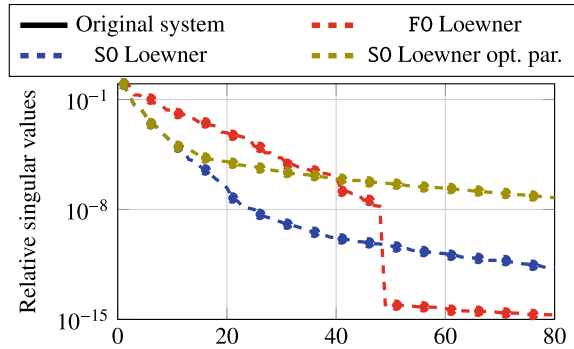
Fig. 2 Demo example: The figure on the left shows the Bode plot of the original system and the FO and SO reduced-order models. The figure on the right shows the Bode plot of the error between the original and reduced-order systems

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{K} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad \mathbf{D} = \alpha \mathbf{M} + \beta \mathbf{M}, \quad \text{and} \quad \mathbf{B}^T = \mathbf{C} = \begin{bmatrix} 2 & 3 \end{bmatrix},$$

with $\alpha = 0.01$ and $\beta = 0.02$. We collect 20 samples $(\sigma_j, \hat{\mathbf{H}}_{\text{SO}}(\sigma_j))$, for $\sigma_j \in \iota[10^{-1}, 10^1]$ logarithmically spaced. Then, we construct the FO and SO Loewner matrices in (12) and (23), receptively.

In Fig. 1, we plot the decay of the singular values of the \mathbb{L} and \mathbb{L}^{SO} matrices. It can be observed that $\text{rank}(\mathbb{L}) = 4$ and $\text{rank}(\mathbb{L}^{\text{SO}}) = 2$, as expected. Indeed, the demo system has a minimal SO realization of order 2 and a minimal FO realization of order 4. By applying the SVD procedure, we construct two reduced-order models of order 2, one for FO and the other for SO. We compare the transfer functions of the original and reduced-order systems, and the results are plotted in Fig. 2. The figure shows that the error between the original and SO reduced-order system is of the level of machine precision, which means that the SO approach has recovered an equivalent realization of the original model. Additionally, the FO reduced system of order 2 was not able to mimic the same behavior of the original system, showing that a larger order is required in this case.

Fig. 3 Build example: Decay of the singular values for the FO Loewner matrix and for SO Loewner matrices



4.2 Building Example

Let us now consider the building model from the SLICOT library [9]. It describes the displacement of a multi-storey building, for example, during an earthquake. It is a FO system of order $r = 48$, whose dynamics comes from a mechanical system. The Rayleigh damping coefficients here are $\alpha \approx 0.4947$ and $\beta \approx 0.0011$.

For this example, we collect 200 samples $\mathbf{H}(i\omega)$, with $\omega \in [10^0, 10^2]$. Then, we build the FO and SO Loewner matrices in (12) and (23), respectively. Additionally, using the heuristic procedure in Sect. 3.3, we constructed the reduced model assuming we do not know a priori the parameters α and β . After this procedure, we obtain $\alpha^* = 0.495$ and $\beta^* = 0.001$, which are fairly close to the original parameters.

In Fig. 3, we plot the decay of the singular values of the FO Loewner matrix, the SO Loewner matrix for the original parameters α and β , and the SO Loewner matrix for the estimated parameters α and β . The decay of the singular values for the SO Loewner matrix with original parameters is faster than for the FO Loewner matrix. However, for the SO Loewner matrix with estimated parameters, the decay of singular values starts fast and then becomes slower. This shows that if the parameters α and β are not well identified, a higher reduced-order will be needed to interpolate the data. By applying the SVD procedure, we construct three reduced-order models of order 16. We compare the transfer functions of the original and reduced-order systems, and the results are plotted in Fig. 4. This figure shows that for the SO Loewner approach (original parameters or with estimated parameters) outperform the classical Loewner framework.

4.3 Artificial Fishtail

As the last example, we consider the artificial fishtail model presented in [30]. This model comes from a finite-element discretization of the continuous mechanics model of an artificial fishtail. After discretization, the finite-dimensional system has a SO

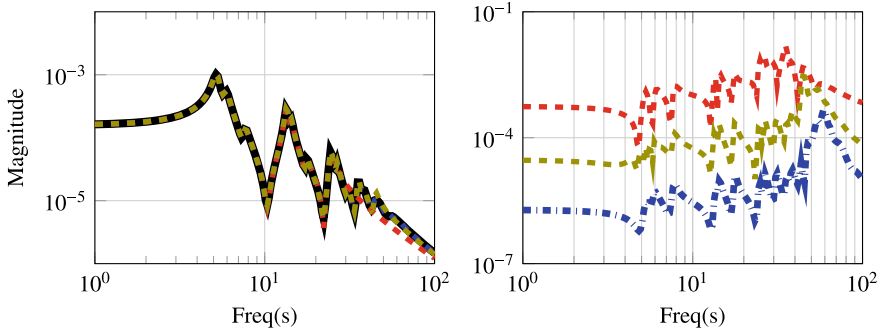
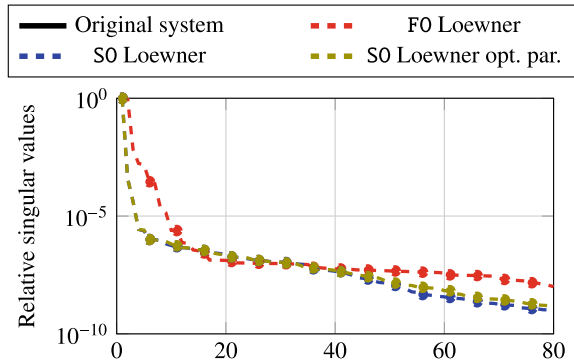


Fig. 4 Build example: The figure on the left shows the Bode plot of the original system and the FO and SO reduced-order models. The figure on the right shows the Bode plot of the error between the original and reduced-order systems

Fig. 5 Fishtail example: Decay of the singular values for the FO Loewner matrix and for SO Loewner matrices



realization of order 779, 232. For this model, the Rayleigh damping is chosen with parameters $\alpha = 1.0 \cdot 10^{-4}$, $\beta = 2 \cdot 10^{-4}$. It is a MIMO system, but for the numerical application, here we consider only the first transfer function, i.e., from u_1 to y_1 .

For this example, we collect 200 samples $\mathbf{H}(i\omega)$, with $\omega \in [10^1, 10^4]$. Then, we build FO and SO Loewner matrices in (12) and (23), respectively. Additionally, we also compute the reduced model using the heuristic procedure in Sect. 3.3, for which we obtain the estimated parameters $\alpha^* \approx 1.19 \cdot 10^{-4}$ and $\beta^* \approx 2 \cdot 10^{-4}$.

In Fig. 5, we plot the decay of the singular values of the FO Loewner matrix, the SO Loewner matrix for the original parameters α and β , and the SO Loewner matrix for the estimated parameters α^* and β^* . By applying the SVD procedure, we construct three reduced-order models of order 8. We compare the transfer functions of the original and reduced-order systems, and the results are plotted in Fig. 6. This figure shows that the SO Loewner approach with original parameters and SO Loewner with estimated parameters outperform the classical Loewner framework.

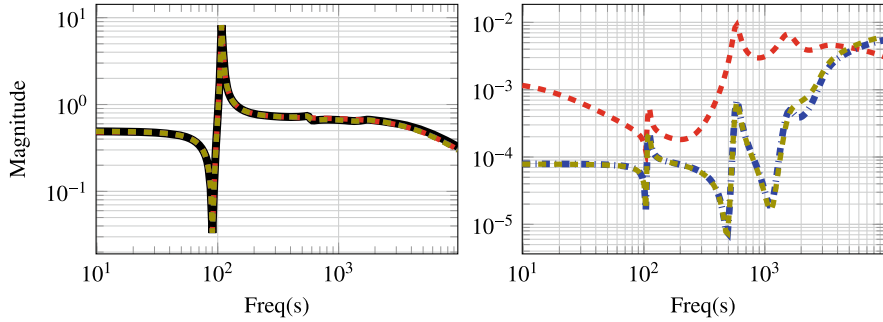


Fig. 6 Fishtail example: The figure on the left shows the Bode plot of the original system and the FO and SO reduced-order models. The figure on the right shows the Bode plot of the error between the original and reduced-order systems

5 Conclusions

In this paper, we have studied the problem of the identification of Rayleigh-damped second-order systems from frequency data. To that aim, we propose modified SO Loewner matrices which are the key tools to construct a realization interpolating the given data. Additionally, in the case of redundant data, an SVD-based scheme is presented to construct reduced-order models. Moreover, a heuristic optimization problem is sketched to estimate the damping parameters. Finally, we have illustrated the efficiency of the proposed approach in some numerical examples, and we compared the results with the classical Loewner framework.

It is important to emphasize that the current work does not intend to solve the general second-order identification problem. As stated in Sect. 3, the problem of identifying a general second-order system leads to an underdetermined system. Hence, in order to have a unique solution, some other constraints should be added to the problem. As an example, other damping structures could be considered, as discussed in Subsection 3.4. However, this case might lead to nonlinear matrix equations and further analysis and dedicated numerical solvers need to be developed. We leave this topic for future research.

Acknowledgements This work was supported by *Deutsche Forschungsgemeinschaft (DFG)*, Collaborative Research Center CRC 96 “Thermo-energetic Design of Machine Tools”. Also, we would like to thank Christopher Beattie for the comments and suggestions on a previous version of this manuscript.

References

1. Antoulas, A., Anderson, B.: On the scalar rational interpolation problem. *IMA J. Math. Control. Inf.* **3**, 61–88 (1986)
2. Antoulas, A.C.: *Approximation of Large-Scale Dynamical Systems*. SIAM Publications, Philadelphia, PA (2005)
3. Beattie, C., Benner, P.: \mathcal{H}_2 -optimality conditions for structured dynamical systems, Preprint MPIMD/14-18, Max Planck Institute Magdeburg (2014). <http://www.mpi-magdeburg.mpg.de/preprints/>
4. Beattie, C.A., Gugercin, S.: Interpolatory projection methods for structure-preserving model reduction. *Syst. Control Lett.* **58**, 225–232 (2009)
5. Benner, P., Cohen, A., Ohlberger, M., Willcox, K.: *Model Reduction and Approximation: Theory and Algorithms Computational Science & Engineering*. SIAM Publications, Philadelphia, PA (2017)
6. Carswell, W., Johansson, J., Løvholt, F., Arwade, S., Madshus, C., DeGroot, D., Myers, A.: Foundation damping and the dynamics of offshore wind turbine monopiles. *Renew. Energy* **80**, 724–736 (2015)
7. Chahlaoui, Y., Gallivan, K.A., Vandendorpe, A., Van Dooren, P.: Model reduction of second-order systems. In: *Dimension Reduction of Large-Scale Systems*, pp. 149–172. Springer (2005)
8. Chahlaoui, Y., Lemonnier, D., Vandendorpe, A., Van Dooren, P.: Second-order balanced truncation. *Linear Algebra Appl.* **415**, 373–384 (2006)
9. Chahlaoui, Y., Van Dooren, P.: A collection of benchmark examples for model reduction of linear time invariant dynamical systems, SLICOT Working Note 2002–2, University of Manchester (2002). www.slicot.org
10. Craig, R.R., Kurdila, A.J.: *Fundamentals of Structural Dynamics*. Wiley (2006)
11. Cruz, C., Miranda, E.: Evaluation of the Rayleigh damping model for buildings. *Eng. Struct.* **138**, 324–336 (2017)
12. Erduran, E.: Evaluation of Rayleigh damping and its influence on engineering demand parameter estimates. *Earthq. Eng. Struct. Dyn.* **41**, 1905–1919 (2012)
13. Gallivan, K., Grimme, E., Van Dooren, P.: A rational Lanczos algorithm for model reduction. *Numer. Algorithms* **12**, 33–63 (1996)
14. Gallivan, K., Vandendorpe, A., Van Dooren, P.: Model reduction of MIMO systems via tangential interpolation. *SIAM J. Matrix Anal. Appl.* **26**, 328–349 (2004)
15. Gallivan, K., Vandendorpe, A., Van Dooren, P.: *Sylvester equations and projection-based model reduction*. *J. Comput. Appl. Math.* **162**, 213–229 (2004)
16. Gresil, M., Giurgiutiu, V.: Prediction of attenuated guided waves propagation in carbon fiber composites using Rayleigh damping model. *J. Intell. Mater. Syst. Struct.* **26**, 2151–2169 (2015)
17. Grimme, E.J.: Krylov projection methods for model reduction, Ph.D. thesis, Univ. of Illinois at Urbana-Champaign, USA (1997)
18. Gustavsen, B., Semlyen, A.: Rational approximation of frequency domain responses by vector fitting. *IEEE Trans. Power Deliv.* **14**, 1052–1061 (1999)
19. Ionita, C.: Lagrange rational interpolation and its applications to model reduction and system identification, Ph.D. thesis, Rice University (2013)
20. Mayo, A.J., Antoulas, A.C.: A framework for the solution of the generalized realization problem, *Linear Algebra Appl.*, 425 (2007), pp. 634–662. Special Issue in honor of P.A. Fuhrmann, Edited by A. C. Antoulas, U. Helmke, J. Rosenthal, V. Vinnikov, and E. Zerz
21. Meirovitch, L.: *Principles and Techniques of Vibrations*, vol. 1, Prentice Hall, New Jersey (1997)
22. Meyer, D.G., Srinivasan, S.: Balancing and model reduction for second-order form linear systems. *IEEE Trans. Autom. Control* **41**, 1632–1644 (1996)
23. Nakatsukasa, Y., Sète, O., Trefethen, L.N.: The AAA algorithm for rational approximation. *SIAM J. Sci. Comput.* **40**, A1494–A1522 (2018)
24. Nāvīk, P., Rønquist, A., Stichel, S.: Identification of system damping in railway catenary wire systems from full-scale measurements. *Eng. Struct.* **113**, 71–78 (2016)

25. Oberwolfach Benchmark Collection: Butterfly gyroscope. hosted at MORwiki – Model Order Reduction Wiki (2004)
26. Paz, M., Leigh, W.: Structural Dynamics. Springer (1985)
27. Pontes Duff, I., Poussot-Vassal, C., Seren, C.: Realization independent single time-delay dynamical model interpolation and \mathcal{H}_2 -optimal approximation. In: 54th IEEE Conference on Decision and Control (CDC), pp. 4662–4667. IEEE (2015)
28. Ramadas, C., Balasubramaniam, K., Hood, A., Joshi, M., Krishnamurthy, C.: Modelling of attenuation of lamb waves using Rayleigh damping: numerical and experimental studies. *Compos. Struct.* **93**, 2020–2025 (2011)
29. Reis, T., Stykel, T.: Balanced truncation model reduction of second-order systems. *Math. Comput. Model. Dyn. Syst.* **14**, 391–406 (2008)
30. Saak, J., Siebelts, D., Werner, S.W.: A comparison of second-order model order reduction methods for an artificial fishtail. *at-Automatisierungstechnik* **67**, 648–667 (2019)
31. Schulze, P., Unger, B.: Data-driven interpolation of dynamical systems with delay. *Syst. Control Lett.* **97**, 125–131 (2016)
32. Schulze, P., Unger, B., Beattie, C., Gugercin, S.: Data-driven structured realization. *Linear Algebra Appl.* **537**, 250–286 (2018)
33. Siebelts, D., Kater, A., Meurer, T., Andrej, J.: Matrices for an artificial fishtail. hosted at MORwiki – Model Order Reduction Wiki (2019)
34. Truhar, N., Veselić, K.: An efficient method for estimating the optimal dampers' viscosity for linear vibrating systems using Lyapunov equation. *SIAM J. Matrix Anal. Appl.* **31**, 18–39 (2009)
35. Villetagne, D.C., Skelton, R.E.: Model reduction using a projection formulation. *Int. J. Control* **46**, 2141–2169 (1987)
36. Wyatt, S.A.: Issues in interpolatory model reduction: Inexact solves, second-order systems and DAEs, Ph.D. thesis, Virginia Tech (2012)
37. Yousuff, A., Wagie, D., Skelton, R.: Linear system approximation via covariance equivalent realizations. *J. Math. Anal. Appl.* **106**, 91–115 (1985)

Balanced Truncation Model Reduction for 3D Linear Magneto-Quasistatic Field Problems



Johanna Kerler-Back and Tatjana Stykel

Abstract We consider linear magneto-quasistatic field equations which arise in simulation of low-frequency electromagnetic devices coupled to electrical circuits. A finite element discretization of such equations on 3D domains leads to a singular system of differential-algebraic equations. First, we study the structural properties of such a system and present a new regularization approach based on projecting out the singular state components. Furthermore, we consider a Lyapunov-based balanced truncation model reduction method which preserves stability and passivity. By making use of the underlying structure of the problem, we develop an efficient model reduction algorithm. Numerical experiments demonstrate its performance on a test example.

Keywords Magneto-quasistatic equations · Differential-algebraic equations · Matrix pencils · Model order reduction · Balanced truncation · Stability · Passivity

1 Introduction

Nowadays, integrated circuits play an increasingly important role. Modelling of electromagnetic effects in high-frequency and high-speed electronic systems leads to coupled field-circuit models of high complexity. The development of efficient, fast and accurate simulation tools for such models is of great importance in the computer-aided design of electromagnetic structures offering significant savings in production cost and time.

In this paper, we consider model order reduction of linear magneto-quasistatic (MQS) systems obtained from Maxwell's equations by assuming that the contri-

J. Kerler-Back
ESG Elektroniksystem- und Logistik-GmbH, Livry-Gargan-Straße 6, 82256 Fürstentfeldbruck,
Germany
e-mail: kerler-back@esg.de

T. Stykel (✉)
Institut für Mathematik, Universität Augsburg, Universitätsstraße 12a, 86159 Augsburg, Germany
e-mail: stykel@math.uni-augsburg.de

bution of displacement current is negligible compared to the conductive currents. Such systems are commonly used for modeling of low-frequency electromagnetic devices like transformers, induction sensors and generators. Due to the presence of non-conducting subdomains, MQS models take form of partial differential-algebraic equations whose dynamics are restricted to a manifold described by algebraic constraints. A spatial discretization of MQS systems using the finite integration technique (FIT) [32] or the finite element method (FEM) [5, 19, 23] leads to differential-algebraic equations (DAEs) which are singular in the 3D case. The structural analysis and numerical treatment of singular DAEs is facing serious challenges due to the fact that the inhomogeneity has to satisfy some restricted conditions to guarantee the existence of solutions and/or that the solution space is infinite-dimensional. To overcome these difficulties, different regularization techniques have been developed for MQS systems [6, 8, 9, 15]. Here, we propose a new regularization approach which is based on a special state space transformation and withdrawal of overdetermined state components and redundant equations.

Furthermore, we exploit the special block structure of the regularized MQS system to determine the deflating subspaces of the underlying matrix pencil corresponding to zero and infinite eigenvalues. This makes it possible to extend the balanced truncation model reduction method to 3D MQS problems. Similarly to [17, 26], our approach relies on projected Lyapunov equations and preserves passivity in a reduced-order model. It should be noted that the balanced truncation method presented in [17] for 2D and 3D gauging-regularized MQS systems cannot be applied to the regularized system obtained here, since it is stable, but not asymptotically stable. To get rid of this problem, we proceed as in [26] and project out state components corresponding not only to the eigenvalue at infinity, but also to zero eigenvalues. Our method is based on computing certain subspaces of incidence matrices related to the FEM discretization which can be determined by using efficient graph-theoretic algorithms developed in [16].

2 Model Problem

We consider a system of MQS equations in vector potential formulation given by

$$\begin{aligned} \sigma \frac{\partial \mathbf{A}}{\partial t} + \nabla \times \nu \nabla \times \mathbf{A} &= \chi \iota && \text{in } \Omega \times (0, T), \\ \mathbf{A} \times n_o &= 0 && \text{on } \partial\Omega \times (0, T), \\ \mathbf{A}(\cdot, 0) &= \mathbf{A}_0 && \text{in } \Omega, \\ \int_{\Omega} \chi^T \frac{\partial \mathbf{A}}{\partial t} d\xi + R \iota &= u && \text{in } (0, T), \end{aligned} \tag{1}$$

where $\mathbf{A} : \Omega \times (0, T) \rightarrow \mathbb{R}^3$ is the magnetic vector potential, $\chi : \Omega \rightarrow \mathbb{R}^{3 \times m}$ is a divergence-free winding function, $\iota : (0, T) \rightarrow \mathbb{R}^m$ and $u : (0, T) \rightarrow \mathbb{R}^m$ are the electrical current and voltage through the stranded conductors with m terminals.

Here, $\Omega \subset \mathbb{R}^3$ is a bounded simply connected domain with a Lipschitz boundary $\partial\Omega$, and n_o is an outer unit normal vector to $\partial\Omega$. The MQS system (1) is obtained from Maxwell's equations by neglecting the contribution of the displacement currents. It is used to study the dynamical behavior of magnetic fields in low-frequency applications [14, 27]. The integral equation in (1) with a symmetric, positive definite resistance matrix $R \in \mathbb{R}^{m \times m}$ results from Faraday's induction law. This equation describes the coupling the electromagnetic devices to an external circuit [28]. Thereby, the voltage u is assumed to be given and the current ι has to be determined. In this case, the MQS system (1) can be considered as a control system with the input u , the state $[\mathbf{A}^T, \iota^T]^T$ and the output $y = \iota$.

We assume that the domain Ω is composed of the conducting and non-conducting subdomains Ω_1 and Ω_2 , respectively, such that $\overline{\Omega} = \overline{\Omega_1} \cup \overline{\Omega_2}$, $\Omega_1 \cap \Omega_2 = \emptyset$ and $\overline{\Omega_1} \subset \Omega$. Furthermore, we restrict ourselves to linear isotropic media implying that the electrical conductivity σ and the magnetic reluctivity ν are scalar functions of the spatial variable only. The electrical conductivity $\sigma : \Omega \rightarrow \mathbb{R}$ is given by

$$\sigma(\xi) = \begin{cases} \sigma_1 & \text{in } \Omega_1, \\ 0 & \text{in } \Omega_2 \end{cases}$$

with some constant $\sigma_1 > 0$, whereas the magnetic reluctivity $\nu : \Omega \rightarrow \mathbb{R}$ is bounded, measurable and uniformly positive such that $\nu(\xi) \geq \nu_0 > 0$ for a.e. in Ω . Note that since σ vanishes on the non-conducting subdomain Ω_2 , the initial condition \mathbf{A}_0 can only be prescribed in the conducting subdomain Ω_1 . Finally, for the winding function $\chi = [\chi_1, \dots, \chi_m]$, we assume that

$$\overline{\text{supp}(\chi_j)} \subset \Omega_2, \quad j = 1, \dots, m, \quad (2)$$

$$\text{supp}(\chi_i) \cap \text{supp}(\chi_j) = \emptyset \quad \text{for } i \neq j. \quad (3)$$

These conditions mean that the conductor terminals are located in Ω_2 and they do not intersect [28].

2.1 FEM Discretization

First, we present a weak formulation for the MQS system (1). For this purpose, we multiply the first equation in (1) with a test function $\phi \in H_0(\text{curl}, \Omega)$ and integrate it over the domain Ω . Using Green's formula, we obtain the variational problem

$$\begin{aligned} \frac{d}{dt} \int_{\Omega} \sigma \mathbf{A} \cdot \phi \, d\xi + \int_{\Omega} \nu (\nabla \times \mathbf{A}) \cdot (\nabla \times \phi) \, d\xi &= \int_{\Omega} (\chi \iota) \cdot \phi \, d\xi, \\ \frac{d}{dt} \int_{\Omega} \chi^T \mathbf{A} \, d\xi + R \iota &= u, \\ \mathbf{A}(\cdot, 0) &= \mathbf{A}_0. \end{aligned} \quad (4)$$

The existence, uniqueness and regularity results for this problem can be found in [25].

For a spatial discretization of (4), we use Nédélec edge and face elements as introduced in [23]. Let $\mathcal{T}_h(\Omega)$ be a regular simplicial triangulation of Ω , and let n_n , n_e and n_f denote the number of nodes, edges and facets, respectively. Furthermore, let $\Phi^e = [\phi_1^e, \dots, \phi_{n_e}^e]$ and $\Phi^f = [\phi_1^f, \dots, \phi_{n_f}^f]$ be the edge and face basis functions, respectively, which span the corresponding finite element spaces. They are related via

$$\nabla \times \Phi^e = \Phi^f C, \tag{5}$$

where $C \in \mathbb{R}^{n_f \times n_e}$ is a *discrete curl matrix* with entries

$$C_{ij} = \begin{cases} 1, & \text{if edge } j \text{ belongs to face } i \text{ and their orientations match,} \\ -1, & \text{if edge } j \text{ belongs to face } i \text{ and their orientations do not match,} \\ 0, & \text{if edge } j \text{ does not belong to face } i, \end{cases}$$

see [5, Sect. 5]. Substituting an approximation to the magnetic vector potential

$$\mathbf{A}(\xi, t) \approx \sum_{j=1}^{n_e} \alpha_j(t) \phi_j^e(\xi)$$

into the variational Eq. (4) and testing it with ϕ_i^e , we obtain a linear DAE system

$$\begin{bmatrix} M & 0 \\ X^T & 0 \end{bmatrix} \frac{d}{dt} \begin{bmatrix} a \\ \iota \end{bmatrix} = \begin{bmatrix} -K & X \\ 0 & -R \end{bmatrix} \begin{bmatrix} a \\ \iota \end{bmatrix} + \begin{bmatrix} 0 \\ I \end{bmatrix} u, \tag{6}$$

where $a = [\alpha_1, \dots, \alpha_{n_e}]^T$ and the conductivity matrix $M \in \mathbb{R}^{n_e \times n_e}$, the curl-curl matrix $K \in \mathbb{R}^{n_e \times n_e}$ and the coupling matrix $X \in \mathbb{R}^{n_e \times m}$ have entries

$$\begin{aligned} M_{ij} &= \int_{\Omega} \sigma \phi_j^e \cdot \phi_i^e \, d\xi, & i, j = 1, \dots, n_e, \\ K_{ij} &= \int_{\Omega} \nu (\nabla \times \phi_j^e) \cdot (\nabla \times \phi_i^e) \, d\xi, & i, j = 1, \dots, n_e, \\ X_{ij} &= \int_{\Omega} \chi_j \cdot \phi_i^e \, d\xi, & i = 1, \dots, n_e, j = 1, \dots, m. \end{aligned} \tag{7}$$

Note that the matrices M and K are symmetric, positive semidefinite. Using the relation (5), we can rewrite the matrix K as

$$K = \int_{\Omega} \nu (\nabla \times \Phi^e)^T (\nabla \times \Phi^e) \, d\xi = \int_{\Omega} \nu C^T (\Phi^f)^T \Phi^f C \, d\xi = C^T M_{\nu} C,$$

where the entries of the symmetric and positive definite matrix M_{ν} are given by

$$(M_v)_{ij} = \int_{\Omega} v \phi_j^f \cdot \phi_i^f d\xi, \quad i, j = 1, \dots, n_f.$$

The coupling matrix X can also be represented in a factored form using the discrete curl matrix C . This can be achieved by taking into account the divergence-free property of the winding function χ , which implies $\chi = \nabla \times \gamma$ for a certain matrix-valued function

$$\gamma = [\gamma_1, \dots, \gamma_m] : \Omega \rightarrow \mathbb{R}^{3 \times m}.$$

Using the cross product rule, Gauss's theorem as well as relations (5) and $\phi_i^e \times n_o = 0$ on $\partial\Omega$, we obtain

$$\begin{aligned} X_{ij} &= \int_{\Omega} (\nabla \times \gamma_j) \cdot \phi_i^e d\xi = \int_{\Omega} \nabla \cdot (\gamma_j \times \phi_i^e) d\xi + \int_{\Omega} \gamma_j \cdot (\nabla \times \phi_i^e) d\xi \\ &= \int_{\partial\Omega} (\gamma_j \times \phi_i^e) \cdot n_o ds + \int_{\Omega} \gamma_j \cdot \sum_{k=1}^{n_f} C_{ki} \phi_k^f d\xi \\ &= \int_{\partial\Omega} \gamma_j \cdot (\phi_i^e \times n_o) ds + \sum_{k=1}^{n_f} C_{ki} \int_{\Omega} \gamma_j \cdot \phi_k^f d\xi = \sum_{k=1}^{n_f} C_{ki} \int_{\Omega} \gamma_j \cdot \phi_k^f d\xi. \end{aligned}$$

Then the matrix X can be written as $X = C^T \Upsilon$, where the entries of $\Upsilon \in \mathbb{R}^{n_f \times m}$ are given by

$$\Upsilon_{kj} = \int_{\Omega} \gamma_j \cdot \phi_k^f d\xi, \quad k = 1, \dots, n_f, \quad j = 1, \dots, m.$$

Note that due to (3), the matrix X has full column rank. This immediately implies that Υ is also of full column rank.

3 Properties of the FEM Model

In this section, we study the structural and physical properties of the FEM model (6). We start with reordering the state vector $a = [a_1^T, a_2^T]^T$ with $a_1 \in \mathbb{R}^{n_1}$ and $a_2 \in \mathbb{R}^{n_2}$ accordingly to the conducting and non-conducting subdomains Ω_1 and Ω_2 . Then the matrices M , K , X and C can be partitioned into blocks as

$$M = \begin{bmatrix} M_{11} & 0 \\ 0 & 0 \end{bmatrix}, \quad K = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}, \quad X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \quad C = [C_1, C_2],$$

where $M_{11} \in \mathbb{R}^{n_1 \times n_1}$ is symmetric, positive definite, $K_{11} \in \mathbb{R}^{n_1 \times n_1}$, $K_{21} = K_{12}^T \in \mathbb{R}^{n_2 \times n_1}$, $K_{22} \in \mathbb{R}^{n_2 \times n_2}$, $X_1 \in \mathbb{R}^{n_1 \times m}$, $X_2 \in \mathbb{R}^{n_2 \times m}$, $C_1 \in \mathbb{R}^{n_f \times n_1}$, and $C_2 \in \mathbb{R}^{n_f \times n_2}$. Note that conditions (2) and (3) imply that $X_1 = 0$ and X_2 has full column rank. In what follows, however, we consider for completeness a general block X_1 . Solving

the second equation in (6) for $\iota = -R^{-1}X^T \frac{d}{dt}a + R^{-1}u$ and inserting this vector into the first equation in (6) yields the DAE control system

$$\begin{aligned} E \frac{d}{dt}a &= -Ka + Bu, \\ y &= -B^T \frac{d}{dt}a + R^{-1}u, \end{aligned} \quad (8)$$

with the matrices

$$\begin{aligned} E &= \begin{bmatrix} M_{11} + X_1 R^{-1} X_1^T & X_1 R^{-1} X_2^T \\ X_2 R^{-1} X_1^T & X_2 R^{-1} X_2^T \end{bmatrix} = \begin{bmatrix} I & C_1^T \gamma \\ 0 & C_2^T \gamma \end{bmatrix} \begin{bmatrix} M_{11} & 0 \\ 0 & R^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ \gamma^T C_1 & \gamma^T C_2 \end{bmatrix}, \\ K &= \begin{bmatrix} C_1^T M_v C_1 & C_1^T M_v C_2 \\ C_2^T M_v C_1 & C_2^T M_v C_2 \end{bmatrix}, \quad B = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} R^{-1} = \begin{bmatrix} C_1^T \gamma \\ C_2^T \gamma \end{bmatrix} R^{-1}. \end{aligned} \quad (9)$$

Using the block structure of the matrices E and K , we can determine their common kernel.

Theorem 1 *Assume that M_{11} , R and M_v are symmetric and positive definite. Let the columns of $Y_{C_2} \in \mathbb{R}^{n_2 \times k_2}$ form a basis of $\ker(C_2)$. Then $\ker(E) \cap \ker(K)$ is spanned by columns of the matrix $[0, Y_{C_2}^T]^T$.*

Proof Assume that $w = [w_1^T, w_2^T]^T \in \ker(E) \cap \ker(K)$. Then due to the positive definiteness of M_{11} and R , it follows from $w^T E w = 0$ with E as in (9) that

$$\begin{bmatrix} I & 0 \\ \gamma^T C_1 & \gamma^T C_2 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = 0.$$

Therefore, $w_1 = 0$ and $\gamma^T C_2 w_2 = 0$. Moreover, using the positive definiteness of M_v , we get from $w^T K w = 0$ with $w_1 = 0$ that $C_2 w_2 = 0$. This means that $w_2 \in \ker(C_2) = \text{im}(Y_{C_2})$, i.e., $w_2 = Y_{C_2} z$ for some vector z . Thus, $w = [0, Y_{C_2}^T]^T z$.

Conversely, assume that $w = [0, Y_{C_2}^T]^T z$ for some $z \in \mathbb{R}^{k_2}$. Then using (9) and $C_2 Y_{C_2} = 0$, we obtain $E w = 0$ and $K w = 0$. Thus, $w \in \ker(E) \cap \ker(K)$. \square

It follows from this theorem that if C_2 has a nontrivial kernel, then

$$\det(\lambda E + K) = 0$$

for all $\lambda \in \mathbb{C}$ implying that the pencil $\lambda E + K$ (and also the DAE system (8)) is singular. This may cause difficulties with the existence and uniqueness of the solution of (8). In the next section, we will see that the divergence-free condition of the winding function χ guarantees that (8) is solvable, but the solution is not unique. This is a consequence of nonuniqueness of the magnetic vector potential \mathbf{A} which is defined up to a gradient of an arbitrary scalar function.

3.1 Regularization

Our goal is now to regularize the singular DAE system (8). In the literature, several regularization approaches have been proposed for semidiscretized 3D MQS systems. In the context of the FIT discretization, the grad-div regularization of MQS systems has been considered in [8, 9] which is based on a spatial discretization of the Coulomb gauge equation $\nabla \cdot \mathbf{A} = 0$. For other regularization techniques, we refer to [6, 7, 15, 22]. Here, we present a new regularization method relying on a special coordinate transformation and elimination of the over- and underdetermined parts.

To this end, we consider a matrix $\hat{Y}_{C_2} \in \mathbb{R}^{n_2 \times (n_2 - k_2)}$ whose columns form a basis of $\text{im}(C_2^T)$. Then the matrix

$$T = \begin{bmatrix} I & 0 & 0 \\ 0 & \hat{Y}_{C_2} & Y_{C_2} \end{bmatrix}$$

is nonsingular. Multiplying the state equation in (8) from the left with T^T and introducing a new state vector

$$\begin{bmatrix} a_1 \\ a_{21} \\ a_{22} \end{bmatrix} = T^{-1}a, \quad (10)$$

the system matrices of the transformed system take the form

$$T^T E T = \begin{bmatrix} M_{11} + C_1^T \Upsilon R^{-1} \Upsilon^T C_1 & C_1^T \Upsilon R^{-1} \Upsilon^T C_2 \hat{Y}_{C_2} & 0 \\ \hat{Y}_{C_2}^T C_2^T \Upsilon R^{-1} \Upsilon^T C_1 & \hat{Y}_{C_2}^T C_2^T \Upsilon R^{-1} \Upsilon^T C_2 \hat{Y}_{C_2} & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$T^T K T = \begin{bmatrix} C_1^T M_v C_1 & C_1^T M_v C_2 \hat{Y}_{C_2} & 0 \\ \hat{Y}_{C_2}^T C_2^T M_v C_1 & \hat{Y}_{C_2}^T C_2^T M_v C_2 \hat{Y}_{C_2} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad T^T B = \begin{bmatrix} C_1^T \Upsilon \\ \hat{Y}_{C_2}^T C_2^T \Upsilon \\ 0 \end{bmatrix} R^{-1}.$$

This implies that the components of a_{22} are actually not involved in the transformed system and, therefore, they can be chosen freely. Moreover, the third equation $0 = 0$ is trivially satisfied showing that system (8) is solvable. Removing this equation, we obtain a regular DAE system

$$E_r \frac{d}{dt} x_r = A_r x_r + B_r u, \quad (11)$$

$$y = -B_r^T \frac{d}{dt} x_r + R^{-1} u, \quad (12)$$

with $x_r = [a_1^T, a_{21}^T]^T \in \mathbb{R}^{n_r}$, $n_r = n_1 + n_2 - k_2$, and

$$E_r = F_\sigma M_\sigma F_\sigma^T, \quad A_r = -F_v M_v F_v^T, \quad B_r = F_v \Upsilon R^{-1}, \quad (13)$$

where

$$F_\sigma = \begin{bmatrix} I & X_1 \\ 0 & \hat{Y}_{C_2}^T X_2 \end{bmatrix} = \begin{bmatrix} I & C_1^T \Upsilon \\ 0 & \hat{Y}_{C_2}^T C_2^T \Upsilon \end{bmatrix}, \quad M_\sigma = \begin{bmatrix} M_{11} & 0 \\ 0 & R^{-1} \end{bmatrix}, \quad F_\nu = \begin{bmatrix} C_1^T \\ \hat{Y}_{C_2}^T C_2^T \end{bmatrix}.$$

The regularity of $\lambda E_r - A_r$ follows from the symmetry of E_r and A_r and the fact that $\ker(E_r) \cap \ker(A_r) = \{0\}$.

3.2 Stability

Stability is an important physical property of dynamical systems characterizing the sensitivity of the solution to perturbations in the data. A pencil $\lambda E_r - A_r$ is called *stable* if all its finite eigenvalues have non-positive real part, and eigenvalues on the imaginary axis are semi-simple in the sense that they have the same algebraic and geometric multiplicity. In this case, any solution of the DAE system (11) with $u = 0$ is bounded. Furthermore, $\lambda E_r - A_r$ is called *asymptotically stable* if all its finite eigenvalues lie in the open left complex half-plane. This implies that any solution of (11) with $u = 0$ satisfies $x_r(t) \rightarrow 0$ as $t \rightarrow \infty$.

The following theorem establishes a quasi-Weierstrass canonical form for the pencil $\lambda E_r - A_r$ which immediately provides information on the finite spectrum and index of this pencil.

Theorem 2 *Let the matrices $E_r, A_r \in \mathbb{R}^{n_r \times n_r}$ be as in (13). Then there exists a nonsingular matrix $W \in \mathbb{R}^{n_r \times n_r}$ which transforms the pencil $\lambda E_r - A_r$ into the quasi-Weierstrass canonical form*

$$W^T E_r W = \begin{bmatrix} E_{11} & & \\ & I_{n_0} & \\ & & 0 \end{bmatrix}, \quad W^T A_r W = \begin{bmatrix} A_{11} & & \\ & 0 & \\ & & I_{n_\infty} \end{bmatrix}, \quad (14)$$

where $E_{11}, -A_{11} \in \mathbb{R}^{n_s \times n_s}$ are symmetric, positive definite, and $n_s + n_0 + n_\infty = n_r$. Furthermore, the pencil $\lambda E_r - A_r$ has index one and all its finite eigenvalues are real and non-positive.

Proof First, note that the existence of a nonsingular matrix W transforming $\lambda E_r - A_r$ into (14) immediately follows from the general results for Hermitian pencils [30]. However, here, we present a constructive proof to better understand the structural properties of the pencil $\lambda E_r - A_r$.

Let the columns of the matrices $Y_\sigma \in \mathbb{R}^{n_r \times n_\infty}$ and $Y_\nu \in \mathbb{R}^{n_r \times n_0}$ form bases of $\ker(F_\sigma^T)$ and $\ker(F_\nu^T)$, respectively. Then we have

$$F_\sigma^T Y_\sigma = 0, \quad F_\nu^T Y_\nu = 0. \quad (15)$$

Moreover, the matrices $Y_\nu^T E_r Y_\nu$ and $Y_\sigma^T A_r Y_\sigma$ are both nonsingular, and $[Y_\nu, Y_\sigma]$ has full column rank. These properties follow from the fact that

$$\ker(F_\sigma^T) \cap \ker(F_\nu^T) = \ker(E_r) \cap \ker(A_r) = \{0\}.$$

Consider a matrix

$$W = [W_1, Y_\nu(Y_\nu^T E_r Y_\nu)^{-1/2}, Y_\sigma(Y_\sigma^T A_r Y_\sigma)^{-1/2}], \quad (16)$$

where the columns of W_1 form a basis of $\ker([E_r Y_\nu, A_r Y_\sigma]^T)$. First, we show that this matrix is nonsingular. Assume that there exists a vector v such that $W^T v = 0$. Then $W_1^T v = 0$, $Y_\nu^T v = 0$ and $Y_\sigma^T v = 0$. Thus,

$$v \in \text{im}([E_r Y_\nu, A_r Y_\sigma]) \cap \ker(Y_\nu^T) \cap \ker(Y_\sigma^T) = \{0\},$$

and, hence, W is nonsingular.

Furthermore, using (15) and

$$W_1^T E_r Y_\nu (Y_\nu^T E_r Y_\nu)^{-1/2} = 0, \quad W_1^T A_r Y_\sigma (Y_\sigma^T A_r Y_\sigma)^{-1/2} = 0,$$

we obtain (14) with $E_{11} = W_1^T E_r W_1$ and $A_{11} = W_1^T A_r W_1$. Obviously, E_{11} and $-A_{11}$ are symmetric and positive semidefinite. For any $v_1 \in \ker(E_{11})$, we have $F_\sigma^T W_1 v_1 = 0$. This implies $W_1 v_1 \in \ker(F_\sigma^T) = \text{im}(Y_\sigma)$. Therefore, there exists a vector z such that $W_1 v_1 = Y_\sigma z$. Multiplying this equation from the left with $Y_\sigma^T E_r$, we obtain $Y_\sigma^T E_r Y_\sigma z = Y_\sigma^T E_r W_1 v_1 = 0$. Then $z = 0$ and, hence, $v_1 = 0$. Thus, E_{11} is positive definite. Analogously, we can show that $-A_{11}$ is positive definite too. This implies that all eigenvalues of the pencil $\lambda E_{11} - A_{11}$ are real and negative. Index one property immediately follows from (14). \square

As a consequence, we obtain that the DAE system (11) is stable but not asymptotically stable since the pencil $\lambda E_r - A_r$ has zero eigenvalues.

We consider now the output Eq. (12). Our goal is to transform this equation into the standard form $y = C_r x_r$ with an output matrix $C_r \in \mathbb{R}^{m \times n_r}$. For this purpose, we introduce first a reflexive inverse of E_r given by

$$E_r^- = W \begin{bmatrix} E_{11}^{-1} & & \\ & I & \\ & & 0 \end{bmatrix} W^T. \quad (17)$$

Simple calculations show that this matrix satisfies

$$E_r E_r^- E_r = E_r, \quad E_r^- E_r E_r^- = E_r^-, \quad (E_r^-)^T = E_r^-. \quad (18)$$

Next, we show that $\hat{Y}_{C_2}^T X_2$ has full column rank. Indeed, if there exists a vector v such that $\hat{Y}_{C_2}^T X_2 v = 0$, then $X_2 v \in \ker(\hat{Y}_{C_2}^T)$. On the other hand,

$$X_2 v = C_2^T \gamma v \in \text{im}(C_2^T) = \text{im}(\hat{Y}_{C_2})$$

implying $X_2 v = 0$. Since X_2 has full column rank, we get $v = 0$.

Using nonsingularity of $X_2^T \hat{Y}_{C_2} \hat{Y}_{C_2}^T X_2$, the input matrix B_r in (13) can be represented as

$$B_r = F_\sigma M_\sigma \begin{bmatrix} 0 \\ I \end{bmatrix} = F_\sigma M_\sigma \begin{bmatrix} I & 0 \\ X_1^T & X_2^T \hat{Y}_{C_2} \end{bmatrix} \begin{bmatrix} 0 \\ \hat{Y}_{C_2}^T X_2 (X_2^T \hat{Y}_{C_2} \hat{Y}_{C_2}^T X_2)^{-1} \end{bmatrix} = E_r \begin{bmatrix} 0 \\ Z \end{bmatrix} \quad (19)$$

with $Z = \hat{Y}_{C_2}^T X_2 (X_2^T \hat{Y}_{C_2} \hat{Y}_{C_2}^T X_2)^{-1}$. Then employing the first relation in (18) and the state Eq. (11), the output (12) can be written as

$$\begin{aligned} y &= -[0, Z^T] E_r \frac{d}{dt} x_r + R^{-1} u = -[0, Z^T] E_r E_r^- E_r \frac{d}{dt} x_r + R^{-1} u \\ &= -B_r^T E_r^- (A_r x_r + B_r u) + R^{-1} u = -B_r^T E_r^- A_r x_r + (R^{-1} - B_r^T E_r^- B_r) u. \end{aligned}$$

It follows from the first relation in (18) and (19) that

$$B_r^T E_r^- B_r = [0, Z^T] E_r E_r^- E_r \begin{bmatrix} 0 \\ Z \end{bmatrix} = [0, Z^T] F_\sigma M_\sigma F_\sigma^T \begin{bmatrix} 0 \\ Z \end{bmatrix} = R^{-1}.$$

Thus, the output takes the form

$$y = C_r x_r \quad (20)$$

with $C_r = -B_r^T E_r^- A_r$.

3.3 Passivity

Passivity is another crucial property of control systems especially in interconnected network design [1, 33]. The DAE control system (11), (20) is called *passive* if for all $t_f > 0$ and all inputs $u \in L_2(0, t_f)$ admissible with the initial condition $E_r x_r(0) = 0$, the output satisfies

$$\int_0^{t_f} y^T(t) u(t) dt \geq 0.$$

This inequality means that the system does not produce energy. In the frequency domain, passivity of (11), (20) is equivalent to the *positive definiteness* of its transfer function

$$H_r(s) = C_r (s E_r - A_r)^{-1} B_r$$

meaning that this function is analytic in $\mathbb{C}_+ = \{z \in \mathbb{C} : \text{Re}(z) > 0\}$ and $H_r(s) + H_r^*(s) \geq 0$ for all $s \in \mathbb{C}_+$, see [1]. Using the special structure of the system matrices in (13), we can show that the DAE system (11), (20) is passive.

Theorem 3 *The DAE system (11), (13), (20) is passive.*

Proof First, observe that the transfer function $H_r(s)$ of (11), (13), (20) is analytic in \mathbb{C}_+ . This fact immediately follows from Theorem 2. Furthermore, using the relations

$$E_r E_r^- A_r = E_r E_r^- A_r E_r^- E_r = A_r E_r^- E_r,$$

we obtain for $F(s) = (sE_r - A_r)^{-1} B_r$ and all $s \in \mathbb{C}_+$ that

$$\begin{aligned} H_r(s) + H_r^*(s) &= C_r (sE_r - A_r)^{-1} B_r + B_r^T (\bar{s}E_r - A_r)^{-1} C_r^T \\ &= -B_r^T E_r^- A_r (sE_r - A_r)^{-1} B_r - B_r^T (\bar{s}E_r - A_r)^{-1} A_r E_r^- B_r \\ &= F^*(s) (-\bar{s}E_r - A_r) E_r^- A_r - A_r E_r^- (sE_r - A_r) F(s) \\ &= 2 F^*(s) (A_r E_r^- A_r + \operatorname{Re}(s) E_r E_r^- (-A_r) E_r^- E_r) F(s) \geq 0 \end{aligned}$$

holds. In the last inequality, we utilized the property that the matrices $E_r E_r^- (-A_r) E_r^- E_r$ and $A_r E_r^- A_r$ are both symmetric and positive semidefinite. Thus, $H_r(s)$ is positive real, and, hence, system (11), (13), (20) is passive. \square

4 Balanced Truncation Model Reduction

Our goal is now to approximate the DAE system (11), (13), (20) by a reduced-order model

$$\begin{aligned} \tilde{E}_r \frac{d}{dt} \tilde{x}_r &= \tilde{A}_r \tilde{x}_r + \tilde{B}_r u, \\ \tilde{y} &= \tilde{C}_r \tilde{x}_r, \end{aligned} \quad (21)$$

where $\tilde{E}_r, \tilde{A}_r \in \mathbb{R}^{\ell \times \ell}$, $\tilde{B}_r, \tilde{C}_r^T \in \mathbb{R}^{\ell \times m}$ and $\ell \ll n_r$. This model should capture the dynamical behavior of (11). It is also important that it preserves the passivity and has a small approximation error. In order to determine the reduced-order model (21), we aim to employ a balanced truncation model reduction method [3, 20]. Unfortunately, we cannot apply this method directly to (11), (13), (20) because, as established in Sect. 3.2, this system is stable but not asymptotically stable due to the fact that the pencil $\lambda E_r - A_r$ has zero eigenvalues. Another difficulty is the presence of infinite eigenvalues due to the singularity of E_r . This may cause problems in defining the controllability and observability Gramians which play an essential role in balanced truncation.

To overcome these difficulties, we first observe that the states of the transformed system $(W^T E_r W, W^T A_r W, W^T B_r, C_r W)$ corresponding to the zero and infinite eigenvalues are uncontrollable and unobservable at the same time. This immediately follows from the representations

$$W^T B_r = [B_1^T, 0, 0]^T, \quad C_r W = [C_1, 0, 0] \quad (22)$$

with $B_1 = W_1^T B_r$ and $C_1 = -B_r^T E_r^- A_r W_1 = -B_1^T E_{11}^- A_{11}$. Therefore, these states can be removed from the system without changing its input-output behavior. Then

the standard balanced truncation approach can be applied to the remaining system. Since the system matrices of the regularized system (11), (20) have the same structure as those of RC circuit equations studied in [26], we proceed with the balanced truncation approach developed there which avoids the computation of the transformation matrix W .

For the DAE system (11), (20), we define the controllability and observability Gramians G_c and G_o as unique symmetric, positive semidefinite solutions of the projected continuous-time Lyapunov equations

$$E_r G_c A_r + A_r G_c E_r = -\Pi^T B_r B_r^T \Pi, \quad G_c = \Pi G_c \Pi^T, \quad (23)$$

$$E_r G_o A_r + A_r G_o E_r = -\Pi^T C_r^T C_r \Pi, \quad G_o = \Pi G_o \Pi^T, \quad (24)$$

where Π is the spectral projector onto the right deflating subspace of $\lambda E_r - A_r$ corresponding to the negative eigenvalues. Using the quasi-Weierstrass canonical form (14) and (16), this projector can be represented as

$$\Pi = W \begin{bmatrix} I & & \\ & 0 & \\ & & 0 \end{bmatrix} W^{-1} = W_1 \hat{W}_1^T, \quad (25)$$

where $\hat{W}_1 \in \mathbb{R}^{n_r \times n_s}$ satisfies

$$\hat{W}_1^T W_1 = I, \quad \hat{W}_1^T Y_v = 0, \quad \hat{W}_1^T Y_\sigma = 0. \quad (26)$$

Similarly to [17, Theorem 3], a relation between the controllability and observability Gramians of system (11), (13), (20) can be established.

Theorem 4 *Let G_c and G_o be the controllability and observability Gramians of system (11), (13), (20) which solve the projected Lyapunov Eqs. (23) and (24), respectively. Then*

$$E_r G_o E_r = A_r G_c A_r.$$

Proof Consider the reflexive inverse E_r^- of E_r given in (17) and the reflexive inverse of A_r given by

$$A_r^- = W \begin{bmatrix} A_{11}^{-1} & & \\ & 0 & \\ & & I \end{bmatrix} W^T.$$

Then multiplying the Lyapunov Eq. (23) (resp. (24)) from the left and right with E_r^- (resp. with A_r^-) and using the relations

$$\begin{aligned} E_r \Pi &= \Pi^T E_r, & \Pi E_r^- &= E_r^- \Pi^T, & \Pi^T E_r E_r^- &= \Pi^T A_r A_r^-, \\ A_r \Pi &= \Pi^T A_r, & \Pi A_r^- &= A_r^- \Pi^T, & E_r^- A_r A_r^- &= E_r^- \Pi^T, \end{aligned}$$

we obtain

$$A_r^-(A_r G_c A_r) E_r^- + E_r^-(A_r G_c A_r) A_r^- = -\Pi E_r^- B_r B_r^T E_r^- \Pi^T, \quad G_c = \Pi G_c \Pi^T, \quad (27)$$

$$A_r^-(E_r G_o E_r) E_r^- + E_r^-(E_r G_o E_r) A_r^- = -\Pi E_r^- B_r B_r^T E_r^- \Pi^T, \quad G_o = \Pi G_o \Pi^T. \quad (28)$$

Since E_r^- and $-A_r^-$ are symmetric and positive semidefinite and Π^T is the spectral projector onto the right deflating subspace of $\lambda E_r^- - A_r^-$ corresponding to the negative eigenvalues, the Lyapunov Eqs. (27) and (28) are uniquely solvable, and, hence, $E_r G_o E_r = A_r G_c A_r$. \square

Theorem 4 implies that we need to solve only the projected Lyapunov Eq. (23) for the Cholesky factor Z_c of $G_c = Z_c Z_c^T$. Then it follows from the relation

$$G_o = E_r^- A_r G_c A_r E_r^- = (-E_r^- A_r Z_c)(-Z_c^T A_r E_r^-)$$

that the Cholesky factor of the observability Gramian $G_o = Z_o Z_o^T$ can be calculated as $Z_o = -E_r^- A_r Z_c$. In this case, the Hankel singular values of (11), (20) can be computed from the eigenvalue decomposition

$$Z_o^T E_r Z_c = (-Z_c^T A_r E_r^-) E_r Z_c = -Z_c^T A_r Z_c = [U_1, U_2] \begin{bmatrix} \Lambda_1 & \\ & \Lambda_2 \end{bmatrix} [U_1, U_2]^T,$$

where $[U_1, U_2]$ is orthogonal, $\Lambda_1 = \text{diag}(\lambda_1, \dots, \lambda_\ell)$ and $\Lambda_2 = \text{diag}(\lambda_{\ell+1}, \dots, \lambda_{n_r})$ with $\lambda_1 \geq \dots \geq \lambda_\ell \gg \lambda_{\ell+1} \geq \dots \geq \lambda_{n_r}$. Then the reduced-order model (21) is computed by projection

$$\tilde{E}_r = U^T E_r V, \quad \tilde{A}_r = U^T A_r V, \quad \tilde{B}_r = U^T B_r, \quad \tilde{C}_r = C_r V$$

with the projection matrices $V = Z_c U_1 \Lambda_1^{-\frac{1}{2}}$ and $U = Z_o U_1 \Lambda_1^{-\frac{1}{2}} = -E_r^- A_r V$. The reduced matrices have the form

$$\begin{aligned} \tilde{E}_r &= -V^T A_r E_r^- E_r V = -\Lambda_1^{-\frac{1}{2}} U_1^T Z_c^T A_r Z_c U_1 \Lambda_1^{-\frac{1}{2}} = I, \\ \tilde{A}_r &= -V^T A_r E_r^- A_r V, \\ \tilde{B}_r &= -V^T A_r E_r^- B_r = V^T C_r^T = \tilde{C}_r^T. \end{aligned} \quad (29)$$

The balanced truncation method for the DAE system (11), (13), (20) is presented in Algorithm 1, where for numerical efficiency reasons, the Cholesky factor Z_c of the Gramian G_c is replaced by a low-rank Cholesky factor \tilde{Z}_c such that $G_c \approx \tilde{Z}_c \tilde{Z}_c^T$.

Note that the matrices \tilde{E}_r and $-\tilde{A}_r$ in (29) are both symmetric and positive definite. This implies that the reduced-order model (21), (29) is asymptotically stable. Then the transfer function $\tilde{H}_r(s) = \tilde{C}_r (s \tilde{E}_r - \tilde{A}_r)^{-1} \tilde{B}_r$ is analytic in \mathbb{C}_+ and for all $s \in \mathbb{C}_+$, it satisfies

Algorithm 1 Balanced truncation for the 3D linear MQS system

Require: $E_r, A_r \in \mathbb{R}^{n_r \times n_r}$ and $B_r \in \mathbb{R}^{n_r \times m}$

Ensure: a reduced-order system $(\tilde{E}_r, \tilde{A}_r, \tilde{B}_r, \tilde{C}_r)$.

- 1: Solve the projected Lyapunov Eq. (23) for a low-rank Cholesky factor $\tilde{Z}_c \in \mathbb{R}^{n_r \times n_c}$ of the controllability Gramian $G_c \approx \tilde{Z}_c \tilde{Z}_c^T$.
- 2: Compute the eigenvalue decomposition

$$-\tilde{Z}_c^T A_r \tilde{Z}_c = [U_1, U_2] \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} [U_1, U_2]^T,$$

where $[U_1, U_2]$ is orthogonal, $\Lambda_1 = \text{diag}(\lambda_1, \dots, \lambda_\ell)$ and $\Lambda_2 = \text{diag}(\lambda_{\ell+1}, \dots, \lambda_{n_c})$.

- 3: Compute the reduced matrices

$$\tilde{E}_r = I, \quad \tilde{A}_r = -V^T A_r E_r^- A_r V, \quad \tilde{B}_r = -V^T A_r E_r^- B_r, \quad \tilde{C}_r = \tilde{B}_r^T$$

with the projection matrix $V = \tilde{Z}_c U_1 \Lambda_1^{-\frac{1}{2}}$.

$$\begin{aligned} \tilde{H}_r(s) + \tilde{H}_r^*(s) &= \tilde{B}_r^T (s\tilde{E}_r - \tilde{A}_r)^{-1} \tilde{B}_r + \tilde{B}_r^T (\bar{s}\tilde{E}_r - \tilde{A}_r)^{-1} \tilde{B}_r \\ &= 2\tilde{B}_r^T (\bar{s}\tilde{E}_r - \tilde{A}_r)^{-1} (\text{Re}(s)\tilde{E}_r - \tilde{A}_r) (s\tilde{E}_r - \tilde{A}_r)^{-1} \tilde{B}_r \geq 0. \end{aligned}$$

Thus, $\tilde{H}_r(s)$ is positive real and, hence, the reduced-order model (21) is passive. Moreover, taking into account that the controllability and observability Gramians \tilde{G}_c and \tilde{G}_o of (21) satisfy $\tilde{G}_c = \tilde{G}_o = \Lambda_1 > 0$, we conclude that (21) is balanced and minimal. Finally, we obtain the following bound on the \mathcal{H}_∞ -norm of the approximation error

$$\|H_r - \tilde{H}_r\|_{\mathcal{H}_\infty} := \sup_{\omega \in \mathbb{R}} \|H_r(i\omega) - \tilde{H}_r(i\omega)\| \leq 2(\lambda_{\ell+1} + \dots + \lambda_{n_r}), \quad (30)$$

which can be proved analogously to [11, 12]. Here, $\|\cdot\|$ denotes the spectral matrix norm. Using (14) and (22), the error system can be written as

$$\begin{aligned} H_r(s) - \tilde{H}_r(s) &= C_r (sE_r - A_r)^{-1} B_r - \tilde{C}_r (s\tilde{E}_r - \tilde{A}_r)^{-1} \tilde{B}_r \\ &= B_1^T (sE_{11} - A_{11}^{-1}) E_{11} - (-E_{11})^{-1} B_1 - \tilde{B}_r^T (s\tilde{E}_r - \tilde{A}_r)^{-1} \tilde{B}_r \\ &= C_e (sE_e - A_e)^{-1} B_e \end{aligned}$$

with

$$E_e = \begin{bmatrix} -E_{11} A_{11}^{-1} E_{11} & \\ & \tilde{E}_r \end{bmatrix}, \quad A_e = \begin{bmatrix} -E_{11} & \\ & \tilde{A}_r \end{bmatrix}, \quad B_e = \begin{bmatrix} B_1 \\ \tilde{B}_r \end{bmatrix} = C_e^T.$$

Since E_e and $-A_e$ are both symmetric, positive definite and $B_e = C_e^T$, it follows from [26, Theorem 4.1(iv)] that

$$\|H_r - \tilde{H}_r\|_{\mathcal{H}_\infty} = \|H_r(0) - \tilde{H}_r(0)\|. \quad (31)$$

Using the output Eq. (12) instead of (20), the transfer function $H_r(s)$ can also be written as

$$H_r(s) = -sB_r^T(sE_r - A_r)^{-1}B_r + R^{-1}.$$

Then the computation of the \mathcal{H}_∞ -error is simplified to

$$\|H_r - \tilde{H}_r\|_{\mathcal{H}_\infty} = \|R^{-1} + \tilde{B}_r^T \tilde{A}_r^{-1} \tilde{B}_r\|. \quad (32)$$

We will use this relation in numerical experiments to verify the efficiency of the error bound (30).

Note that the presented model reduction method for the DAE system (11), (13), (20) is not balanced truncation applied to the frequency-inverted system with the transfer function

$$\begin{aligned} H_r\left(\frac{1}{s}\right) &= -\frac{1}{s}B_r^T\left(\frac{1}{s}E_r - A_r\right)^{-1}B_r + R^{-1} \\ &= B_r^T(sA_r - E_r)^{-1}B_r + R^{-1} = sB_r^T E_r^-(sE_r^- - A_r^-)^{-1}E_r^- B_r, \end{aligned}$$

as it might be presumed at first glance. Our method can rather be interpreted as balanced truncation applied to the transformed system obtained by multiplication the state Eq. (11) from the left with the nonsingular transformation matrix $T_r = (-A_r + E_r\Pi_0)(E_r + A_r\Pi_\infty)^{-1}$, where

$$\begin{aligned} \Pi_0 &= Y_v(Y_v^T E_r Y_v)^{-1}Y_v^T E_r = W \begin{bmatrix} 0 & \\ & I \\ & & 0 \end{bmatrix} W^{-1}, \\ \Pi_\infty &= Y_\sigma(Y_\sigma^T A_r Y_\sigma)^{-1}Y_\sigma^T A_r = W \begin{bmatrix} 0 & \\ & 0 \\ & & I \end{bmatrix} W^{-1} \end{aligned} \quad (33)$$

are the spectral projectors onto the right deflating subspaces of $\lambda E_r - A_r$ corresponding to the zero and infinite eigenvalues, respectively. Observe that the transformed system with the system matrices

$$\begin{aligned}\hat{E} &= T_r E_r = W^{-T} \begin{bmatrix} -A_{11} & & \\ & I & \\ & & 0 \end{bmatrix} W^{-1}, \\ \hat{A} &= T_r A_r = W^{-T} \begin{bmatrix} -A_{11} E_{11}^{-1} A_{11} & & \\ & 0 & \\ & & -I \end{bmatrix} W^{-1}, \\ \hat{B} &= T_r B_r = W^{-T} [-B_1^T E_{11}^{-1} A_{11}, 0, 0]^T, \\ \hat{C} &= C_r = [-B_1^T E_{11}^{-1} A_{11}, 0, 0] W^{-1}\end{aligned}$$

has the same transfer function as (11), (20) and is symmetric in the sense that \hat{E} and \hat{A} are both symmetric and $\hat{B} = \hat{C}^T$. Then projecting this system with the projection matrix $V = \Pi V$, we obtain

$$\begin{aligned}V^T \hat{E} V &= -V^T A_r E_r^- E_r V = \tilde{E}_r, \\ V^T \hat{A} V &= -V^T A_r E_r^- A_r V = \tilde{A}_r, \\ V^T \hat{B} &= -V^T A_r E_r^- B_r = \tilde{B}_r = \tilde{C}_r^T.\end{aligned}$$

Consequently, the model reduction method in Algorithm 1 inherits the properties of the balanced truncation method for symmetric systems [18, 26]. In particular, it provides a symmetric reduced-order model which is exact at the frequency $s = \infty$ and, as follows from (31), achieves the maximal error at $s = 0$.

5 Computational Aspects

In this section, we discuss the computational aspects of Algorithm 1. This includes solving the projected Lyapunov Eq. (23) and computing the basis matrices for certain subspaces.

For the numerical solution of the projected Lyapunov Eq. (23) in Step 1 of Algorithm 1, we apply the low-rank alternating directions implicit (LR-ADI) method as presented in [29] with appropriate modifications proposed in [4] for cheap evaluation of the Lyapunov residuals. First, note that due to (22) the input matrix satisfies $\Pi^T B_r = B_r$. Then setting

$$\begin{aligned}F_1 &= (\tau_1 E_r + A_r)^{-1} B_r, \\ R_1 &= B_r - 2\tau_1 E_r F_1, \\ Z_1 &= \sqrt{-\tau_1} F_1,\end{aligned}$$

the LR-ADI iteration is given by

$$\begin{aligned}
F_k &= (\tau_k E_r + A_r)^{-1} R_{k-1}, \\
R_k &= R_{k-1} - 2\tau_1 E_r F_k, \\
Z_k &= [Z_{k-1}, \sqrt{-\tau_k} F_k],
\end{aligned} \tag{34}$$

with negative shift parameters τ_k which strongly influence the convergence of this iteration. Note that they can be chosen to be real, since the pencil $\lambda E_r - A_r$ has real finite eigenvalues. This also enables to determine the optimal ADI shift parameters by the Wachspress method [31] ones the spectral bounds $a = -\lambda_{\max}(E_r, A_r)$ and $b = -\lambda_{\min}(E_r, A_r)$ are available. Here, $\lambda_{\max}(E_r, A_r)$ and $\lambda_{\min}(E_r, A_r)$ denote the largest and smallest nonzero eigenvalues of $\lambda E_r - A_r$. They can be computed simultaneously by applying the Lanczos procedure to $E_r^- A_r$ and $v = \Pi v$, see [13, Sect. 10.1]. As a starting vector v , we can take, for example, one of the columns of the matrix $E_r^- B_r$. In the Lanczos procedure and also in Step 3 of Algorithm 1, it is required to compute the products $E_r^- A_r \Pi v$. Of course, we never compute and store the reflexive inverse E_r^- explicitly. Instead, we can use the following lemma to calculate such products in a numerically efficient way.

Lemma 1 *Let E_r and A_r be given as in (13), $Z = \hat{Y}_{C_2}^T X_2 (X_2^T \hat{Y}_{C_2} \hat{Y}_{C_2}^T X_2)^{-1}$, and $v \in \mathbb{R}^{n_r}$. Then the vector $z = E_r^- A_r \Pi v$ can be determined as*

$$z = (I - \Pi_\infty) \hat{Y}_\sigma (\hat{Y}_\sigma^T E_r \hat{Y}_\sigma)^{-1} \hat{Y}_\sigma^T A_r \Pi v, \tag{35}$$

where Π_∞ is the spectral projector as in (38), and

$$\hat{Y}_\sigma = \begin{bmatrix} I & 0 \\ 0 & Z \end{bmatrix} \tag{36}$$

is a basis matrix for $\text{im}(F_\sigma)$.

Proof We show first that the full column matrix \hat{Y}_σ in (36) satisfies the equation $\text{im}(\hat{Y}_\sigma) = \text{im}(F_\sigma)$. This property immediately follows from the relation

$$F_\sigma = \begin{bmatrix} I & X_1 \\ 0 & \hat{Y}_{C_2}^T X_2 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & Z \end{bmatrix} \begin{bmatrix} I & X_1 \\ 0 & X_2^T \hat{Y}_{C_2} \hat{Y}_{C_2}^T X_2 \end{bmatrix}.$$

Since $F_\sigma^T \hat{Y}_\sigma$ has full column rank, the matrix $\hat{Y}_\sigma^T E_r \hat{Y}_\sigma = \hat{Y}_\sigma^T F_\sigma F_\sigma^T \hat{Y}_\sigma$ is nonsingular, i.e., z in (35) is well-defined. Obviously, this vector fulfills $\Pi_\infty z = 0$. Furthermore, we have

$$E_r z = E_r (I - \Pi_\infty) \hat{Y}_\sigma (\hat{Y}_\sigma^T E_r \hat{Y}_\sigma)^{-1} \hat{Y}_\sigma^T A_r \Pi v = E_r \hat{Y}_\sigma (\hat{Y}_\sigma^T E_r \hat{Y}_\sigma)^{-1} \hat{Y}_\sigma^T A_r \Pi v.$$

Then

$$\begin{aligned}
\hat{Y}_\sigma^T E_r z &= \hat{Y}_\sigma^T A_r \Pi v, \\
Y_\sigma^T E_r z &= 0 = Y_\sigma^T (I - \Pi_\infty) A_r \Pi v = Y_\sigma^T A_r \Pi v.
\end{aligned}$$

Since $[\hat{Y}_\sigma, Y_\sigma]$ is nonsingular, these equations imply $E_r z = A_r \Pi v$. Multiplying this equation from the left with E_r^- , we get

$$z = (I - \Pi_\infty)z = E_r^- E_r z = E_r^- A_r \Pi v.$$

This completes the proof. \square

Using (36), we find by simple calculations that

$$\hat{Y}_\sigma (\hat{Y}_\sigma^T E_r \hat{Y}_\sigma)^{-1} \hat{Y}_\sigma^T = \begin{bmatrix} M_{11}^{-1} & -M_{11}^{-1} X_1 Z^T \\ -Z X_1^T M_{11}^{-1} & Z(X_1^T M_{11}^{-1} X_1 + R) Z^T \end{bmatrix}.$$

Next, we discuss the computation of $Y_\sigma (Y_\sigma^T A_r Y_\sigma)^{-1} Y_\sigma^T v$ for a vector v . By taking $v = A_r w$, this enables to calculate the product $\Pi_\infty w = Y_\sigma (Y_\sigma^T A_r Y_\sigma)^{-1} Y_\sigma^T A_r w$ required in (35).

Lemma 2 *Let A_r be as in (13) and let Y_σ be a basis of $\ker(F_\sigma^T)$. Then for the vector $v = [v_1^T, v_2^T]^T \in \mathbb{R}^{n_r}$, the product*

$$z = Y_\sigma (Y_\sigma^T A_r Y_\sigma)^{-1} Y_\sigma^T v \tag{37}$$

can be determined as $z = [0, z_2^T]^T$, where z_2 satisfies the linear system

$$\begin{bmatrix} -\hat{Y}_{C_2}^T K_{22} \hat{Y}_{C_2} & \hat{Y}_{C_2}^T X_2 \\ X_2^T \hat{Y}_{C_2} & 0 \end{bmatrix} \begin{bmatrix} z_2 \\ \hat{z}_2 \end{bmatrix} = \begin{bmatrix} v_2 \\ 0 \end{bmatrix}. \tag{38}$$

Proof We first show that $z = Y_\sigma (Y_\sigma^T A_r Y_\sigma)^{-1} Y_\sigma^T v$ if and only if

$$\begin{bmatrix} A_r & \hat{Y}_\sigma \\ \hat{Y}_\sigma^T & 0 \end{bmatrix} \begin{bmatrix} z \\ \hat{z} \end{bmatrix} = \begin{bmatrix} v \\ 0 \end{bmatrix}, \tag{39}$$

where \hat{Y}_σ is as in (36). Let $[z^T, \hat{z}^T]^T$ solves Eq. (39). Then $\hat{Y}_\sigma^T z = 0$ and, hence, $z \in \ker(\hat{Y}_\sigma^T) = \text{im}(Y_\sigma)$. This means that there exists a vector \hat{w} such that $z = Y_\sigma \hat{w}$. Inserting this vector into the first equation in (39), we obtain $A_r Y_\sigma \hat{w} + \hat{Y}_\sigma \hat{z} = v$. Multiplying this equation from the left with Y_σ^T and solving it for \hat{w} , we get $z = Y_\sigma (Y_\sigma^T A_r Y_\sigma)^{-1} Y_\sigma^T v$.

Conversely, for z as in (37) and $\hat{z} = (\hat{Y}_\sigma^T \hat{Y}_\sigma)^{-1} \hat{Y}_\sigma^T (v - A_r z)$, we have $\hat{Y}_\sigma^T z = 0$ and

$$\begin{aligned} A_r z + \hat{Y}_\sigma \hat{z} &= A_r z + \hat{Y}_\sigma (\hat{Y}_\sigma^T \hat{Y}_\sigma)^{-1} \hat{Y}_\sigma^T (v - A_r z) \\ &= (I - \hat{Y}_\sigma (\hat{Y}_\sigma^T \hat{Y}_\sigma)^{-1} \hat{Y}_\sigma^T) A_r z + \hat{Y}_\sigma (\hat{Y}_\sigma^T \hat{Y}_\sigma)^{-1} \hat{Y}_\sigma^T v. \end{aligned}$$

Using $\hat{Y}_\sigma (\hat{Y}_\sigma^T \hat{Y}_\sigma)^{-1} \hat{Y}_\sigma^T + Y_\sigma (Y_\sigma^T Y_\sigma)^{-1} Y_\sigma^T = I$ twice, we obtain

$$\begin{aligned} A_r z + \hat{Y}_\sigma \hat{z} &= Y_\sigma (Y_\sigma^T Y_\sigma)^{-1} Y_\sigma^T A_r z + \hat{Y}_\sigma (\hat{Y}_\sigma^T \hat{Y}_\sigma)^{-1} \hat{Y}_\sigma^T v \\ &= Y_\sigma (Y_\sigma^T Y_\sigma)^{-1} Y_\sigma^T A_r Y_\sigma (Y_\sigma^T A_r Y_\sigma)^{-1} Y_\sigma^T v + \hat{Y}_\sigma (\hat{Y}_\sigma^T \hat{Y}_\sigma)^{-1} \hat{Y}_\sigma^T v = v. \end{aligned}$$

Thus, $[z^T, \hat{z}^T]^T$ satisfies Eq. (39).

Equation (39) can be written as

$$\begin{bmatrix} -K_{11} & -K_{12} \hat{Y}_{C_2} & I & 0 \\ -\hat{Y}_{C_2}^T K_{21} & -\hat{Y}_{C_2}^T K_{22} \hat{Y}_{C_2} & 0 & Z \\ \hat{I} & 0 & 0 & 0 \\ 0 & Z^T & 0 & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ 0 \\ 0 \end{bmatrix}, \quad (40)$$

with $z = [z_1^T, z_2^T]^T$, $\hat{z} = [z_3^T, z_4^T]^T$ and $v = [v_1^T, v_2^T]^T$. The third equation in (40) yields $z_1 = 0$. Furthermore, multiplying the fourth equation in (40) from the left with $X_2^T \hat{Y}_{C_2} \hat{Y}_{C_2}^T X_2$ and introducing a new variable $\hat{z}_2 = (X_2^T \hat{Y}_{C_2} \hat{Y}_{C_2}^T X_2)^{-1} z_4$, we obtain Eq. (38) which is uniquely solvable since $\hat{Y}_{C_2}^T K_{22} \hat{Y}_{C_2}$ is symmetric, positive definite and $\hat{Y}_{C_2}^T X_2$ has full column rank. Thus, $z = [0, z_2^T]^T$ with z_2 satisfying (38). \square

We summarize the computation of $z = E_r^- A_r v$ with $v = \Pi v$ in Algorithm 2.

Algorithm 2 Computation of $E_r^- A_r v$

Require: $M_{11}, K_{11}, K_{12}, K_{21}, K_{22}, X_1, X_2, R, \hat{Y}_{C_2}$, and $v = \Pi v = [v_1^T, v_2^T]^T$.

Ensure: $z = E_r^- A_r v$ with E_r and A_r as in (13).

- 1: Compute $\begin{bmatrix} \hat{v}_1 \\ \hat{v}_2 \end{bmatrix} = \begin{bmatrix} -K_{11} v_1 - K_{12} \hat{Y}_{C_2} v_2 \\ -\hat{Y}_{C_2}^T K_{21} v_1 - \hat{Y}_{C_2}^T K_{22} \hat{Y}_{C_2} v_2 \end{bmatrix}$.
 - 2: Compute $Z = \hat{Y}_{C_2}^T X_2 (X_2^T \hat{Y}_{C_2} \hat{Y}_{C_2}^T X_2)^{-1}$.
 - 3: Compute $\hat{w}_2 = Z^T \hat{v}_2$.
 - 4: Solve $M_{11} w_1 = \hat{v}_1 - X_1 \hat{w}_2$ for w_1 .
 - 5: Compute $w_2 = -Z (X_1^T w_1 - R \hat{w}_2)$.
 - 6: Solve $\begin{bmatrix} -\hat{Y}_{C_2}^T K_{22} \hat{Y}_{C_2} & \hat{Y}_{C_2}^T X_2 \\ X_2^T \hat{Y}_{C_2} & 0 \end{bmatrix} \begin{bmatrix} z_2 \\ \hat{z}_2 \end{bmatrix} = \begin{bmatrix} -\hat{Y}_{C_2}^T K_{21} w_1 - \hat{Y}_{C_2}^T K_{22} \hat{Y}_{C_2} w_2 \\ 0 \end{bmatrix}$ for z_2 .
 - 7: Compute $z = \begin{bmatrix} w_1 \\ w_2 - z_2 \end{bmatrix}$.
-

The major computational effort in the LR-ADI method (34) is the computation of $(\tau_k E_r + A_r)^{-1} w$ for some vector w . If $\tau_k E_r + A_r$ remains sparse, we just solve the linear system $(\tau_k E_r + A_r) z = w$ of dimension n_r . If $\tau_k E_r + A_r$ gets fill-in due to the multiplication with \hat{Y}_{C_2} , then we can use the following lemma to compute $z = (\tau_k E_r + A_r)^{-1} w$.

Lemma 3 *Let E_r and A_r be as in (13), $w = [w_1^T, w_2^T]^T \in \mathbb{R}^{n_r}$, and $\tau < 0$. Then the vector $z = (\tau E_r + A_r)^{-1} w$ can be determined as*

$$z = \begin{bmatrix} z_1 \\ (\hat{Y}_{C_2}^T \hat{Y}_{C_2})^{-1} \hat{Y}_{C_2}^T z_2 \end{bmatrix},$$

where z_1 and z_2 satisfy the linear system

$$\begin{bmatrix} \tau M_{11} - K_{11} & -K_{12} & X_1 & 0 \\ -K_{21} & -K_{22} & X_2 & Y_{C_2} \\ \tau X_1^T & \tau X_2^T & -R & 0 \\ 0 & Y_{C_2}^T & 0 & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} = \begin{bmatrix} w_1 \\ \hat{Y}_{C_2} (\hat{Y}_{C_2}^T \hat{Y}_{C_2})^{-1} w_2 \\ 0 \\ 0 \end{bmatrix} \quad (41)$$

of dimension $n + m + k_2$.

Proof First, note that due to the choice of Y_{C_2} the coefficient matrix in system (41) is nonsingular. This system can be written as

$$(\tau M_{11} - K_{11})z_1 - K_{12}z_2 + X_1z_3 = w_1, \quad (42a)$$

$$-K_{21}z_1 - K_{22}z_2 + X_2z_3 + Y_{C_2}z_4 = \hat{Y}_{C_2}^T (\hat{Y}_{C_2}^T \hat{Y}_{C_2})^{-1} w_2, \quad (42b)$$

$$\tau X_1^T z_1 + \tau X_2^T z_2 - Rz_3 = 0, \quad (42c)$$

$$Y_{C_2}^T z_2 = 0. \quad (42d)$$

It follows from (42d) that $z_2 \in \ker(Y_{C_2}^T) = \text{im}(\hat{Y}_{C_2})$. Then there exists \hat{z}_2 such that $z_2 = \hat{Y}_{C_2} \hat{z}_2$. Since \hat{Y}_{C_2} has full column rank, it holds

$$\hat{z}_2 = (\hat{Y}_{C_2}^T \hat{Y}_{C_2})^{-1} \hat{Y}_{C_2}^T z_2. \quad (43)$$

Further, from Eq. (42c) we obtain $z_3 = \tau R^{-1} X_1^T z_1 + \tau R^{-1} X_2^T z_2$. Substituting z_2 and z_3 into (42a) and (42b) and multiplying Eq. (42b) from the left with $\hat{Y}_{C_2}^T$ yields

$$(\tau E_r + A_r) \begin{bmatrix} z_1 \\ \hat{z}_2 \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}.$$

This equation together with (43) implies that

$$\begin{bmatrix} z_1 \\ (\hat{Y}_{C_2}^T \hat{Y}_{C_2})^{-1} \hat{Y}_{C_2}^T z_2 \end{bmatrix} = (\tau E_r + A_r)^{-1} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

that completes the proof. \square

Finally, we discuss the computation of the basis matrices Y_{C_2} and \hat{Y}_{C_2} required in Algorithm 2 and the LR-ADI iteration. To this end, we introduce a *discrete gradient matrix* $G_0 \in \mathbb{R}^{n_e \times n_n}$ whose entries are defined as

$$(G_0)_{ij} = \begin{cases} 1, & \text{if edge } i \text{ leaves node } j, \\ -1, & \text{if edge } i \text{ enters node } j, \\ 0, & \text{else.} \end{cases}$$

Note that the discrete curl and gradient matrices C and G_0 satisfy the relations $\text{rank}(C) = n_e - n_n + 1$, $\text{rank}(G_0) = n_n - 1$ and $CG_0 = 0$, see [5]. Then by removing one column of G_0 , we get the reduced discrete gradient matrix G whose columns form a basis of $\ker(C)$. The matrices C and G^T can be considered as the loop and incidence matrices, respectively, of a directed graph whose nodes and branches correspond to the nodes and edges of the triangulation $\mathcal{T}_h(\Omega)$, see [10]. Then the basis matrices Y_{C_2} and \hat{Y}_{C_2} can be determined by using the graph-theoretic algorithms as presented in [16].

Let the reduced gradient matrix $G = [G_1^T \ G_2^T]^T$ be partitioned into blocks according to $C = [C_1 \ C_2]$. It follows from [16, Theorem 9] that

$$\ker(C_2) = \text{im}(G_2 Z_1),$$

where the columns of the matrix Z_1 form a basis of $\ker(G_1)$. Then \hat{Y}_{C_2} can be determined as $\hat{Y}_{C_2} = \text{kernelAk}(Z_1^T G_2^T)$ with the function `kernelAk` from [16, Sect. 4.2], where the basis Z_1 is computed by applying the function `kernelAT` from [16, Sect. 3] to G_1^T .

6 Numerical Results

In this section, we present some results of numerical experiments demonstrating the balanced truncation model reduction method for 3D linear MQS systems. For the FEM discretization with Nédélec elements, the 3D tetrahedral mesh generator NETGEN¹ and the MATLAB toolbox² from [2] were used as described in [21]. All computations were done with MATLAB R2018a.

As a test model, we consider a coil wound round a conducting tube surrounded by air. Such a model was studied in [24] in the context of optimal control. A bounded domain

$$\Omega = (-c_1, c_1) \times (-c_2, c_2) \times (-c_3, c_3) \subset \mathbb{R}^3$$

consists of the conducting subdomain $\Omega_1 = \Omega_{\text{iron}}$ of the iron tube and the non-conducting subdomain $\Omega_2 = \Omega_{\text{coil}} \cup \Omega_{\text{air}}$, where

$$\begin{aligned} \Omega_{\text{iron}} &= \{ \xi \in \mathbb{R}^3 : 0 < r_1 < \xi_1^2 + \xi_2^2 < r_2, \ z_1 < \xi_3 < z_2 \}, \\ \Omega_{\text{coil}} &= \{ \xi \in \mathbb{R}^3 : 0 < r_3 < \xi_1^2 + \xi_2^2 < r_4, \ z_3 < \xi_3 < z_4 \} \end{aligned}$$

¹ <https://sourceforge.net/projects/netgen-mesher/>.

² <http://www.mathworks.com/matlabcentral/fileexchange/46635>.

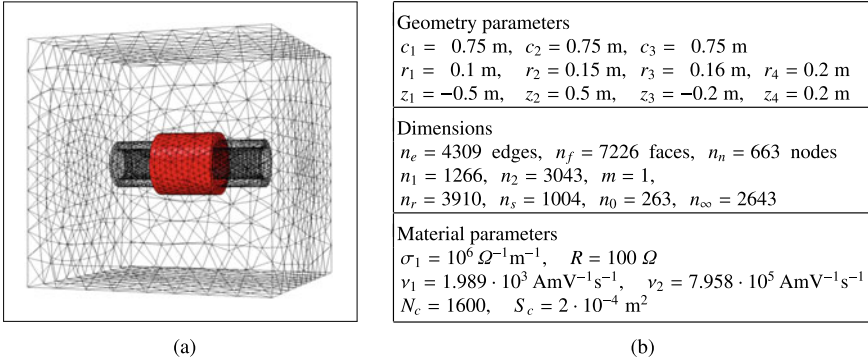


Fig. 1 Coil-tube model: **a** geometry; **b** dimensions and model parameters

with $r_1 < r_2 < r_3 < r_4$ and $z_1 < z_3 < z_4 < z_2$, see Fig. 1(a). The divergence-free winding function $\chi : \Omega \rightarrow \mathbb{R}^3$ is defined by

$$\chi(\xi) = \begin{cases} \frac{N_c}{S_c \sqrt{\xi_1^2 + \xi_2^2}} \begin{bmatrix} -\xi_2 \\ \xi_1 \\ 0 \end{bmatrix}, & \xi \in \Omega_{\text{coil}}, \\ 0, & \xi \in \Omega \setminus \Omega_{\text{coil}}, \end{cases}$$

where N_c is the number of coil turns and S_c is the cross section area of the coil. The dimensions, geometry and material parameters are given in Fig. 1(b).

The DAE system (8) has $n_1 + m = 1267$ differential variables, $n_2 - m - k_2 = 2643$ algebraic variables and $k_2 = 399$ singular variables. The regularized pencil $\lambda E_r - A_r$ has $n_s = 1004$ negative eigenvalues and $n_0 = 263$ zero eigenvalues. It seems that the interface conditions between the conducting and non-conducting subdomains are responsible for the zero eigenvalues. This follows from the fact that the number of zero eigenvalues is equal to $n_{n,\text{iron/air}} - 1$, where $n_{n,\text{iron/air}}$ is the number of nodes on the interface boundary between the iron tube and the surrounding air.

The controllability Gramian was approximated by a low-rank matrix $G_c \approx Z_{n_c} Z_{n_c}^T$ with $Z_{n_c} \in \mathbb{R}^{n_r \times n_c}$ with $n_c = 24$. The normalized residual norm

$$\frac{\|E_r Z_k Z_k^T A_r + A_r Z_k Z_k^T E_r + B_r B_r^T\|_F}{\|B_r B_r^T\|_F} = \frac{\|R_k R_k^T\|_F}{\|B_r B_r^T\|_F} = \frac{\|R_k^T R_k\|_F}{\|B_r^T B_r\|_F}$$

for the LR-ADI iteration (34) is presented in Fig. 2a. Here, $\|\cdot\|_F$ denotes the Frobenius matrix norm. Figure 2b shows the Hankel singular values $\lambda_1, \dots, \lambda_{n_c}$. We approximate the regularized MQS system (11), (12) of dimension $n_r = 3910$ by a reduced model of dimension $\ell = 5$. In Fig. 3a, we present the absolute values of the frequency responses $|H_r(i\omega)|$ and $|\tilde{H}_r(i\omega)|$ of the full and reduced-order models for the frequency range $\omega \in [10^{-4}, 10^6]$. The absolute error $|H_r(i\omega) - \tilde{H}_r(i\omega)|$ and the error bound computed as

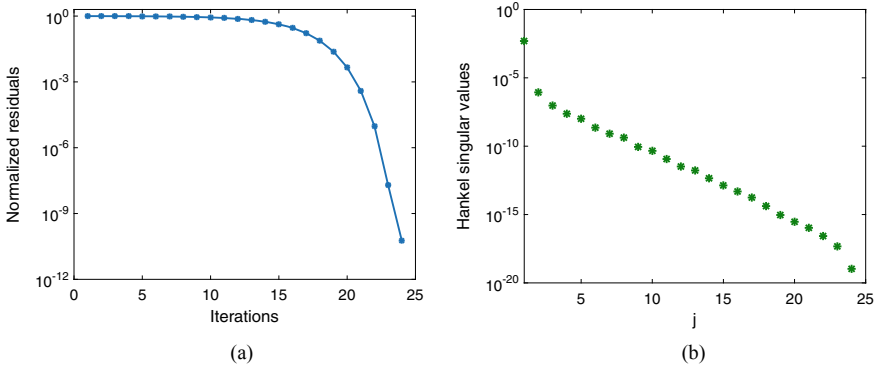


Fig. 2 **a** Convergence history for the LR-ADI method; **b** Hankel singular values

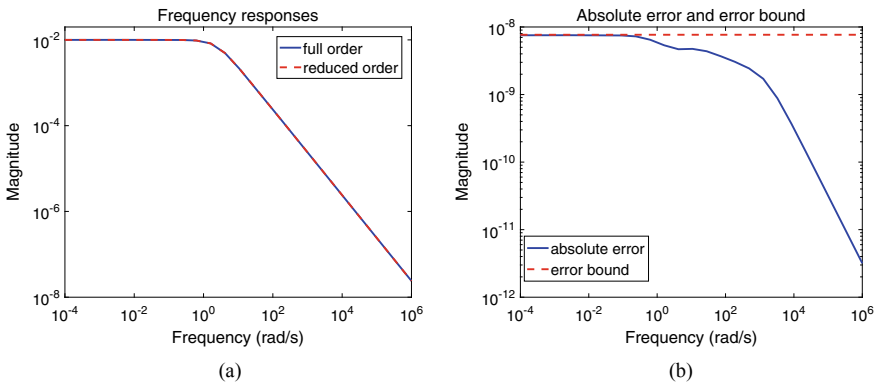


Fig. 3 **a** Frequency responses of the full-order and reduced-order systems; **b** Absolute error and error bound

$$2(\lambda_{\ell+1} + \dots + \lambda_{n_c-1} + (n_s - \ell + 1)\lambda_{n_c}) = 7.6714 \cdot 10^{-9}$$

are given in Fig. 3b. Furthermore, using (32) we compute the error

$$\|H_r - \tilde{H}_r\|_{\mathcal{H}_\infty} = 7.5385 \cdot 10^{-9}$$

showing that the error bound is very tight.

In Fig. 4a, we present the outputs $y(t)$ and $\tilde{y}(t)$ of the full and reduced-order systems on the time interval $[0, 0.08]s$ computed for the input $u(t) = 5 \cdot 10^4 \sin(300\pi t)$ and zero initial condition using the implicit Euler method with 300 time steps. The relative error

$$\frac{|y(t) - \tilde{y}(t)|}{\max_{t \in [0, 0.08]} |y(t)|}$$

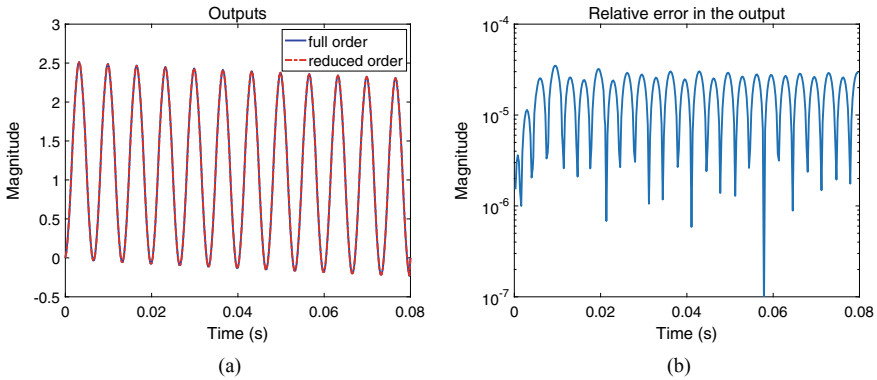


Fig. 4 **a** Outputs of the full-order and reduced-order systems; **b** Relative error in the output

is given in Fig. 4b. One can see that the reduced-order model approximates well the original system in both time and frequency domain.

Acknowledgements The authors would like to thank Hanko Ipach for providing the MATLAB functions for computing the kernels and ranges of incidence matrices and Inga Muetzfeldt for providing the semidiscretized MQS model.

References

1. Anderson, B., Vongpanitlerd, S.: Network Analysis and Synthesis. Prentice Hall, Englewood Cliffs, NJ (1973)
2. Anjam, I., Valdman, J.: Fast MATLAB assembly of FEM matrices in 2D and 3D: edge elements. *Appl. Math. Comput.* **267**, 252–263 (2015)
3. Antoulas, A.: Approximation of Large-Scale Dynamical Systems. SIAM, Philadelphia, PA (2005)
4. Benner, P., Kürschner, P., Saak, J.: An improved numerical method for balanced truncation for symmetric second order systems. *Math. Comput. Model. Dyn. Systems* **19**(6), 593–615 (2013)
5. Bossavit, A.: Computational Electromagnetism. Academic Press, San Diego (1998)
6. Bossavit, A.: “Stiff” problems in eddy-current theory and the regularization of Maxwell’s equations. *IEEE Trans. Magn.* **37**(5), 3542–3545 (2001)
7. Cendes, Z., Manges, J.: A generalized tree-cotree gauge for magnetic field computation. *IEEE Trans. Magn.* **31**(3), 1342–1347 (1995)
8. Clemens, M., Schöps, S., Gersem, H.D., Bartel, A.: Decomposition and regularization of non-linear anisotropic curl-curl DAEs. *COMPEL* **30**(6), 1701–1714 (2011)
9. Clemens, M., Weiland, T.: Regularization of eddy-current formulations using discrete grad-div operators. *IEEE Trans. Magn.* **38**(2), 569–572 (2002)
10. Deo, N.: Graph Theory with Applications to Engineering and Computer Science. Prentice-Hall, Englewood Cliffs, N.J. (1974)
11. Enns, D.: Model reduction with balanced realization: an error bound and a frequency weighted generalization. In: Proceedings of the 23rd IEEE Conference on Decision and Control, pp. 127–132. Las Vegas (1984)

12. Glover, K.: All optimal hankel-norm approximations of linear multivariable systems and their L^∞ -error bounds. *Internat. J. Control* **39**, 1115–1193 (1984)
13. Golub, G., Van Loan, C.: *Matrix Computations*, 4th edn. The Johns Hopkins University Press, Baltimore (2013)
14. Haus, H., Melcher, J.: *Electromagnetic Fields and Energy*. Prentice Hall, Englewood Cliffs (1989)
15. Hiptmair, R.: Multilevel gauging for edge elements. *Computing* **64**(2), 97–122 (2000)
16. Ipach, H.: *Grafentheoretische Anwendung in der Analyse elektrischer Schaltkreise*. Bachelor thesis, Universität Hamburg (2013)
17. Kerler-Back, J., Stykel, T.: Model reduction for linear and nonlinear magneto-quasistatic equations. *Int. J. Numer. Meth. Eng.* **111**(13), 1274–1299 (2017)
18. Liu, W., Sreeram, V., Teo, K.: Model reduction for state-space symmetric systems. *Syst. Control Lett.* **34**(4), 209–215 (1998)
19. Monk, P.: *Finite Element Methods for Maxwell's Equations*. Numerical Mathematics and Scientific Computation, Oxford University Press (2003)
20. Moore, B.: Principal component analysis in linear systems: controllability, observability, and model reduction. *IEEE Trans. Automat. Control* **AC-26**(1), 17–32 (1981)
21. Muetzelfeld, I.: *Model order reduction of magneto-quasistatic equations in 3D domains*. Master thesis, Universität Augsburg (2017)
22. Munteanu, I.: Tree-cotree condensation properties. *ICS Newsletter (International Compumag Society)* **9**, 10–14 (2002)
23. Nédélec, J.: Mixed finite elements in \mathbb{R}^3 . *Numerische Mathematik* **35**(3), 315–341 (1980)
24. Nicaise, S., Stingelin, S., Tröltzsch, F.: On two optimal control problems for magnetic fields. *Comput. Methods Appl. Math.* **14**(4), 555–573 (2014)
25. Nicaise, S., Tröltzsch, F.: A coupled Maxwell integrodifferential model for magnetization processes. *Mathematische Nachrichten* **287**(4), 432–452 (2013)
26. Reis, T., Stykel, T.: Lyapunov balancing for passivity-preserving model reduction of RC circuits. *SIAM J. Appl. Dyn. Syst.* **10**(1), 1–34 (2011)
27. Rodriguez, A., Valli, A.: *Eddy Current Approximation of Maxwell Equations: Theory Algorithms and Applications*. Springer, Mailand (2010)
28. Schöps, S., Gersem, H.D., Weiland, T.: Winding functions in transient magnetoquasistatic field-circuit coupled simulations. *COMPEL* **32**(6), 2063–2083 (2013)
29. Stykel, T.: Low-rank iterative methods for projected generalized Lyapunov equations. *Electron. Trans. Numer. Anal.* **30**, 187–202 (2008)
30. Thompson, R.: The characteristic polynomial of a principal subpencil of a Hermitian matrix pencil. *Linear Algebra Appl.* **14**, 135–177 (1976)
31. Wachspress, E.: *The ADI Model Problem*. Springer, New York (2013)
32. Weiland, T.: A discretization method for the solution of Maxwell's equations for six-component fields. *Electron. Commun.* **31**(3), 116–120 (1977)
33. Willems, J., Takaba, K.: Dissipativity and stability of interconnections. *Int. J. Robust Nonlinear Control* **17**, 563–586 (2007)

Structure-Preserving Model Reduction of Physical Network Systems



Arjan van der Schaft

Abstract This paper considers physical network systems where the energy storage is naturally associated to the nodes of the graph, while the edges of the graph correspond to static couplings. The first sections deal with the linear case, covering examples such as mass-damper and hydraulic systems, which have a structure that is similar to symmetric consensus dynamics. The last section is concerned with a specific class of nonlinear physical network systems; namely detailed-balanced chemical reaction networks governed by mass action kinetics. In both cases, linear and nonlinear, the structure of the dynamics is similar, and is based on a weighted Laplacian matrix, together with an energy function capturing the energy storage at the nodes. We discuss two methods for structure-preserving model reduction. The first one is clustering; aggregating the nodes of the underlying graph to obtain a reduced graph. The second approach is based on neglecting the energy storage at some of the nodes, and subsequently eliminating those nodes (called Kron reduction).

Keywords Model reduction · Structure preservation · Clustering · Kron reduction · Chemical reaction networks

1 Introduction

Large-scale network dynamical systems arise in many areas of science and engineering: from power networks to metabolic reaction networks in living cells. From an analysis, simulation, and control point of view there is a clear need to approximate these systems by lower-dimensional ones of lesser complexity; i.e., *model reduction*. In this paper we study *structure-preserving* model reduction of such physical network systems, where we want to approximate the original, complex, physical network system by a physical network system of lesser complexity, but *within the same class* of physical network systems. We focus on linear physical network sys-

A. van der Schaft (✉)
Bernoulli Institute and Jan C. Willems Center for Systems and Control, University of Groningen,
Groningen, The Netherlands
e-mail: a.j.van.der.schaft@rug.nl

tems with quadratic energy storage at the nodes and linear static couplings at the edges, as well as on detailed-balanced chemical reaction networks governed by mass action kinetics. Although the latter class consists of *nonlinear* network systems, the underlying graph structure is similar. As a result in both cases structure-preserving reduction can be performed using the same two approaches; namely *clustering* and *Kron reduction*.

2 A Class of Linear Physical Network Systems

Consider a directed graph, with n nodes and k edges. The graph is specified by its *incidence matrix* D , which is an $n \times k$ matrix where every column corresponds to an edge of the graph, and contains exactly one -1 at the row corresponding to the tail node of the edge and one $+1$ at the row corresponding to its head node, while the other elements are 0. Clearly $\mathbb{1}^T D = 0$, where $\mathbb{1}$ denotes the vector of all ones. We will throughout assume that the graph is *connected*, or equivalently $\ker D^T = \text{span } \mathbb{1}$. (Otherwise we may consider the connected components of the graph separately.) We consider the following type of dynamics on the graph

$$\begin{aligned} \dot{x} &= -DRD^T M^{-1}x + Eu, \quad x \in \mathbb{R}^n, u \in \mathbb{R}^m \\ y &= E^T M^{-1}x \quad \quad \quad y \in \mathbb{R}^m, \end{aligned} \quad (1)$$

where R and M are positive diagonal matrices. Here, E is an $n \times m$ input matrix whose columns are basis vectors, defining which of the nodes are directly controlled. The matrix $L := DRD^T$ is a *weighted Laplacian matrix*, with weights given by the diagonal elements r_1, \dots, r_k of R .

The energy $\frac{1}{2}x^T M^{-1}x$ of the system satisfies

$$\frac{d}{dt} \frac{1}{2} x^T M^{-1} x = -x^T M^{-1} DRD^T M^{-1} x + x^T M^{-1} E u \leq y^T u, \quad (2)$$

showing *passivity* of the system. For $u = 0$ the set of equilibria is given by the 1-dimensional subspace $\ker D^T M^{-1}$, i.e., as all states x for which the components of $v := M^{-1}x$ are identical. Furthermore, $\mathbb{1}^T x$ is a conserved quantity for all u such that $\mathbb{1}^T E u = 0$. It follows that for every initial condition x_0 there exists exactly one equilibrium x_∞ with $\mathbb{1}^T x_0 = \mathbb{1}^T x_\infty$, to which the system converges exponentially for $u = 0$; see e.g. [1] for further details.

A variety of linear physical network systems is of this form, as illustrated by the following examples; see [2, 3].

Example 1 (*Mass-damper systems*) Consider a mass-damper system on \mathbb{R} with n masses with masses m_1, \dots, m_n and k linear dampers with damping coefficients r_1, \dots, r_k . Define the diagonal matrices $M = \text{diag}(m_1, \dots, m_n)$, $R = \text{diag}(r_1, \dots, r_k)$ and the state x as the vector of momenta of the masses. Hence

$v = M^{-1}x$ are the velocities of the masses, and $\mathbb{1}^T x$ is the total momentum of the system. The system for $u = 0$ will converge to a situation where all the masses move with equal velocity, while the total momentum remains the same. The non-zero rows of E correspond to masses which are actuated by external forces u .

Example 2 (*RC electrical networks*) Consider an RC-circuit with n grounded capacitors with capacitances C_1, \dots, C_n and k linear resistors, with conductances (reciprocal of resistance) g_1, \dots, g_k . Define the diagonal matrices $M = \text{diag}(C_1, \dots, C_n)$, $R = \text{diag}(g_1, \dots, g_k)$, and the state x as the vector of charges Q of the capacitors. The total charge $\mathbb{1}^T Q$ is a conserved quantity, and for $u = 0$ the system will converge to an equilibrium where the voltage potentials $\frac{Q_i}{C_i}$ at the nodes are all equal. The non-zero rows of E correspond to capacitors with current sources u .

Example 3 (*Hydraulic networks*) Consider a hydraulic network with n fluid reservoirs whose storage is described by the elements of a vector x . Mass balance corresponds to $\dot{x} = Df$ where $f \in \mathbb{R}^k$ is the flow through the k pipes linking the reservoirs. Let each storage variable x_i determine a pressure $\frac{x_i}{m_i}$ for a certain constant m_i . Assuming that the flow f_k is proportional to the difference between the pressure of the head reservoir and the pressure of the tail reservoir this leads to the model (1), where Eu denotes the direct in-flow or out-flow at the reservoirs corresponding to non-zero rows of E . Clearly, for $u = 0$ the system converges to a unique equilibrium where the pressures at all nodes are equal.

Example 4 (*Single-species chemical reaction networks*) Consider a chemical reaction network governed by mass action kinetics with *single-species reactions* (that is, reactions that involve only one chemical species as substrate and one chemical species as product). In case the network is *detailed-balanced* (there exists an equilibrium for which all the forward reactions are balanced by the reverse reactions) the dynamics of the vector x of concentrations can be brought into the form (see Sect. 6 and [4] for further details) $\dot{x} = -DRD^T \frac{x}{x^*}$, where D is the incidence matrix of the network, x^* is the assumed equilibrium, and $\frac{x}{x^*}$ denotes element-wise division. The j -th element of the diagonal matrix R is specified by the forward and backward reaction constants of the j -th reaction and x^* (see [4] for details). Defining the diagonal matrix $M := \text{diag}(x_1^*, \dots, x_m^*)$ this takes the form (1), where Eu denote additional in-flow or out-flow of the chemical species corresponding to non-zero rows of E .

A 'non-physical' example of (1) is the standard symmetric *consensus algorithm* in continuous time, where $M = I_n$ and Eu corresponds to the leaders of the multi-agent system (see e.g. [5]).

3 Structure-Preserving Model Reduction by Clustering

The general problem of *structure-preserving* model reduction of physical network systems is the following. Given a large-scale physical network system as in (1) (e.g.,

a high-dimensional mass-damper system). How to find to a physical network system of the same type (again a mass-damper system), but of lesser complexity, which is offering a 'suitable' approximation of the system?

The first approach, discussed in this section, is to consider reduction of the physical network system by *clustering* of the nodes, so as to obtain a graph with fewer nodes [2]. This idea is quite natural, and appears e.g. in [6–9], and in the context of symmetric consensus dynamics in [10].

Consider a *partition* of the node set of the directed graph into \hat{n} disjoint cells $C_1, C_2, \dots, C_{\hat{n}}$, together with a corresponding $n \times \hat{n}$ *characteristic matrix* P . The columns of P equal the characteristic vectors of the cells; the characteristic vector of a cell C_i being defined as the vector with 1 at the place of every node contained in the cell C_i , and 0 elsewhere. With some abuse of notation we will denote the partition simply by its characteristic matrix P .

Based on a partition P we reduce the system (1) to

$$\begin{aligned}\dot{\hat{x}} &= -(P^T D R D^T P)(P^T M P)^{-1} \hat{x} + P^T E \hat{u} \\ \hat{y} &= E^T P (P^T M P)^{-1} \hat{x}\end{aligned}\tag{3}$$

where $\hat{x} := P^T x \in \mathbb{R}^{\hat{n}}$ is the clustered state vector. This is a *Petrov-Galerkin* reduced model with projection matrices P and $M P (P^T M P)^{-1}$, cf. [2, 11].

The reduced system (3) is again of the form (1). In fact, after permutation of the edges we may write $D = [D_p \ D_c]$, with D_c corresponding to the edges between nodes of a same cell, and D_p corresponding to edges in between different cells. Then

$$P^T D = [P^T D_p \ P^T D_c] = [\hat{D} \ 0],\tag{4}$$

where the $\hat{n} \times \hat{k}$ matrix \hat{D} is the incidence matrix of the *reduced graph*, with nodes being the cells of the original graph, and with edge set containing all the edges between nodes in different cells. Correspondingly, define \hat{R} as the $\hat{k} \times \hat{k}$ diagonal matrix obtained from R by leaving out the rows and columns corresponding to the edges between nodes in a same cell, and define the $\hat{n} \times \hat{n}$ diagonal matrix $\hat{M} := P^T M P$, and $\hat{E} := P^T E$. It follows that the reduced system (3) is given as

$$\begin{aligned}\dot{\hat{x}} &= -\hat{D} \hat{R} \hat{D}^T \hat{M}^{-1} \hat{x} + \hat{E} \hat{u}, \quad \hat{x} \in \mathbb{R}^{\hat{n}} \\ \hat{y} &= \hat{E}^T \hat{M}^{-1} \hat{x}\end{aligned}\tag{5}$$

In the case of a mass-damper system, x is the vector of momenta p of the masses, and the i -th component of \hat{x} denotes the total momentum $\hat{p}_i = \sum p_j$ of the masses contained in cell C_i , with summation ranging over the indices of the nodes in cell C_i . Furthermore the i -th diagonal element of \hat{M} is the total mass $\hat{m}_i = \sum m_j$ contained in cell C_i . The velocity \hat{v}_i of cell C_i is defined as $\hat{v}_i = \sum p_j / \sum m_j$. Note that if the cell C_i is connected then the masses within C_i when separated from the other cells and the external forces will converge to this common velocity.

The comparison between the full-order model (1) and the reduced-order clustered model (3) should be based on subtracting from the dynamics of the full-order state vector $v := M^{-1}x$ (vector of velocities in the mass-damper case), the dynamics of the vector $P\hat{v} = P\hat{M}^{-1}\hat{x}$, representing the averaged full-order state vector. This leads to the consideration of the *error dynamics*

$$\dot{v} - P\dot{\hat{v}} = M^{-1}DRD^T v - P\hat{M}^{-1}\hat{D}\hat{R}\hat{D}^T \hat{v} + (M^{-1}E - P\hat{M}^{-1}\hat{E})u \quad (6)$$

Of course, a main question is *which* nodes should be clustered in order to obtain a good approximation of the full-order system. The approach as advocated in [12] (in a more general setting) is to cluster those nodes for which the ‘activity’ at the edges in between these nodes is ‘low’. This can be formalized as follows. Consider the full-order physical network system (1) as before, where we additionally assume that E can be factorized as $E = DF$ for some matrix F . This is equivalent to the assumption $\mathbb{1}^T E = 0$, and can be also interpreted in the sense that the inputs u control the *flows* along part of the edges.

Then define the vector of *edge variables* $e = D^T M^{-1}x (= D^T v) \in \mathbb{R}^k$. Their dynamics is given as

$$\begin{aligned} \dot{e} &= D^T M^{-1}\dot{x} = -L_e R e + L_e F u_e, & L_e &:= D^T M^{-1} D \\ y_e &= F^T e \end{aligned} \quad (7)$$

where $L_e := D^T M^{-1} D$ is called the *edge Laplacian*. Note that L_e is invertible iff the graph has no *cycles* ($\ker D = 0$); in which case the edge dynamics defines a *gradient system*, with inner product given by L_e^{-1} . By using *generalized balancing* of the edge dynamics (7) one now can identify the edges which are *least important*; thus determining the characteristic matrix P of an appropriate clustering [12]. We refer to [12] for error bounds; cf. also [6, 7].

Another approach to clustering and error bounds can be found in [9]. For *almost equitable* partitions error bounds were derived in [13].

4 Structure-Preserving Model Reduction by Kron Reduction

A second method for structure-preserving model reduction of physical network systems (1) is *Kron reduction*. This amounts to splitting the n nodes into two distinct subsets K and J , and to partitioning the state vectors x , $v = M^{-1}x$, the Laplacian matrix $L = DRD^T$, and the input matrix E accordingly, i.e.

$$x = \begin{bmatrix} x_K \\ x_J \end{bmatrix}, \quad v = \begin{bmatrix} v_K \\ v_J \end{bmatrix}, \quad L = \begin{bmatrix} L_{KK} & L_{KJ} \\ L_{JK} & L_{JJ} \end{bmatrix}, \quad E = \begin{bmatrix} E_K \\ E_J \end{bmatrix} \quad (8)$$

Then setting \dot{x}_J zero results in the reduced model¹

$$\begin{aligned}\dot{x}_K &= -(L_{KK} - L_{KJ}L_{JJ}^{-1}L_{JK})v_K + (E_K - L_{KJ}L_{JJ}^{-1}E_J)\hat{u}_K \\ \hat{y}_K &= (E_K - L_{KJ}L_{JJ}^{-1}E_J)^T v_K + E_J^T L_{JJ}^{-1} E_J u,\end{aligned}\quad (9)$$

where, since M is diagonal, $v_K = M_{KK}^{-1}x_K$.

Furthermore, as shown in [14], the Schur complement $L_{KK} - L_{KJ}L_{JJ}^{-1}L_{JK}$ of a Laplacian matrix is again a Laplacian matrix; and thus of the form $D_K R_K D_K^T$, for some incidence matrix D_K of a directed graph with set of nodes given by K , and a positive diagonal matrix R_K . Hence the reduced model is again of the form (1) with Laplacian matrix $\hat{L} := L_{KK} - L_{KJ}L_{JJ}^{-1}L_{JK}$ and ‘mass’ matrix M_{KK} ; although with the addition of a feedthrough term $E_J^T L_{JJ}^{-1} E_J u$. We call (9) the *Kron reduced model*.

Note that $E_J^T L_{JJ}^{-1} E_J \geq 0$, implying that the Kron reduced model satisfies

$$\begin{aligned}\frac{d}{dt} \frac{1}{2} x_K^T M_{KK}^{-1} x_K &\leq x_K^T M_{KK}^{-1} (E_K - L_{KJ}L_{JJ}^{-1}E_J) \hat{u}_K \\ &= \hat{y}_K^T \hat{u}_K - \hat{u}_K^T E_J^T L_{JJ}^{-1} E_J \hat{u}_K \leq \hat{y}_K^T \hat{u}_K,\end{aligned}\quad (10)$$

thus showing passivity with respect to the reduced storage function $\frac{1}{2} x_K^T M_{KK}^{-1} x_K$.

Kron reduction may lead to a drastic reduction of the state vector. On the other hand, the *number of edges* in the reduced network system may easily *increase*. This is due to the fact that the Schur complement of a sparse Laplacian matrix often gives rise to a *dense* reduced Laplacian matrix.

The interpretation of the Kron reduced model is somewhat dual to the interpretation of the reduced model resulting from clustering. In case of a mass-damper system Kron reduction amounts to setting the masses m_i of the nodes in the subset J equal to zero, thus leading to a zero total force at the nodes in J and $\dot{x}_J = 0$. More generally, in Kron reduction the energy storage at the nodes in the index subset J is neglected.

5 Steady State Equivalence of the Full and Reduced Order Model

Consider the full-order model (1). The steady state response to a constant input $u = \bar{u}$ is given by a state \bar{x} satisfying

$$DRD^T M^{-1} \bar{x} = E\bar{u}, \quad (11)$$

with resulting steady state output $\bar{y} = E^T M^{-1} \bar{x}$. Notice that such a steady state \bar{x} exists if and only if $\mathbb{1}^T E\bar{u} = 0$, and that there exists a steady state for *all* \bar{u} if

¹ Note that by the assumption of connectedness any non-trivial leading submatrix L_{JJ} of L is invertible.

and only if $\mathbb{1}^T E = 0$. Furthermore given the existence of a steady state \bar{x} all other states x such that $v = M^{-1}\bar{x} + c\mathbb{1}$, $c \in \mathbb{R}$ also qualify as steady states, leading to steady state outputs $y = E^T (M^{-1}\bar{x} + c\mathbb{1})$, which are all equal to $\bar{y} = E^T M^{-1}\bar{x}$ if $\mathbb{1}^T E = 0$. In the mass-damper case (or similar cases) these conditions have an immediate interpretation; e.g., $\mathbb{1}^T E \bar{u} = 0$ means that the sum of the applied external forces is zero.

Now let us compare the steady state response of the full-order model with that of the reduced-order model (3) resulting from clustering. Consider a constant input $\hat{u} = \bar{u}$ for which there exists \hat{x} such that $\hat{D}\hat{R}\hat{D}^T \hat{M}^{-1}\hat{x} = E\bar{u}$, with accompanying steady state output $\hat{y} = \hat{E}^T (\hat{M}^{-1}\hat{x} + c\mathbb{1})$. It can be seen that the steady state response of the reduced order model (3) equals that of the full order model (1) if and only if the constant \bar{u} and the steady state \bar{x} are such that

$$M^{-1}\bar{x} = P(P^T M P)^{-1}\bar{\hat{x}}, \quad (12)$$

In turn, this is the case if and only if $\bar{v} = M^{-1}\bar{x}$ is contained in $\text{im } P$. We have thus arrived at the following proposition.

Proposition 1 *Consider a network system (1), a characteristic matrix P , and the corresponding reduced order model (3). Let \bar{u} satisfy $\mathbb{1}^T E \bar{u} = 0$, and the corresponding steady state \bar{x} of (1) be such that $\bar{v} = M^{-1}\bar{x} \in \text{im } P$. Then the steady state response \bar{y} of the full order system is the same as the steady state response \hat{y} of the reduced order system for $\hat{u} = \bar{u}$. The steady state response of the full- and reduced-order system are equal for any $\hat{u} = \bar{u}$ if and only if $\text{im } E \subset DRD^T (\text{im } P)$.*

In the mass-damper case the condition $\bar{v} = M^{-1}\bar{x} \in \text{im } P$ means that the steady state \bar{x} and the corresponding vector of steady state velocities $M^{-1}\bar{x}$ are such that the steady state velocities in each cell are equal, while $\text{im } E \subset DRD^T (\text{im } P)$ means that all external forces can be compensated by damping forces generated by the dampers linking the different cells (without involving the damping forces due to dampers *within* the cell).

Steady state equivalence between full-order and reduced-order model is even more simple in the Kron reduction case. Indeed, by definition of Kron reduction, the steady state response of the Kron reduced order model (9) is *always* the same as that of the full order model (1) (whenever defined, i.e., $\mathbb{1}^T E \bar{u} = 0$).

Finally, let us look more closely at the properties of the steady state response of the full- and reduced-order model. The steady state \bar{x} corresponding to a constant input \bar{u} for the full order system satisfies

$$DRD^T \bar{v} = E\bar{u}, \quad \bar{v} = M^{-1}\bar{x} \quad (13)$$

Now assume that all columns of E are basis vectors (i.e., in the mass-damper case, part of the masses are individually actuated, and the rest is not). Then, by permutation of the nodes, we may assume without loss of generality that $E = \begin{bmatrix} I_m \\ 0 \end{bmatrix}$. Partitioning

accordingly $\bar{v} = \begin{bmatrix} \bar{v}_1 \\ \bar{v}_2 \end{bmatrix}$ this amounts to

$$DRD^T \begin{bmatrix} \bar{v}_1 \\ \bar{v}_2 \end{bmatrix} = \begin{bmatrix} \bar{u} \\ 0 \end{bmatrix} \quad (14)$$

By the *maximum modulus principle* for graphs [15] this implies

$$\max \bar{v}_2 \leq \max \bar{v}_1, \quad \min \bar{v}_2 \geq \min \bar{v}_1, \quad (15)$$

where $\max z$ means the maximum of the elements in the vector z , and $\min z$ the minimum of the elements in the vector z . Hence, the elements of \bar{v}_2 are ‘squeezed’ in between the minimum and the maximum of the elements of $\bar{y} = \bar{v}_1$; the generalized velocities corresponding to the generalized external forces \bar{u} .

6 Structure-Preserving Model Reduction of Chemical Reaction Networks

In this section we will focus on structure-preserving model reduction of a specific class of physical network systems, namely *chemical reaction networks*. That is, we want the approximating model to be again in the form of a chemical reaction network. A key difference with the physical network systems as discussed in the previous sections is the fact that most chemical reaction networks are intrinsically *nonlinear*.

This will be done for chemical reaction networks that are described by *mass action kinetics*, the most basic type of chemical kinetics giving rise to *polynomial* reaction rates. Furthermore, we will concentrate on *isothermal* reaction networks. Finally, we will make the assumption (natural from a thermodynamics perspective) that the mass action kinetics chemical reaction network is *detailed-balanced*, in the sense of having an equilibrium where all forward reaction rates are balanced by backward reaction rates. Applications of this theory can be e.g. found in systems biology, where metabolic reaction networks with e.g. 500 chemical species and reactions are not uncommon. The resulting structure-preserving model reduction theory is very much in line with the model reduction theory exposed in the previous sections (and also based either on clustering or on Kron reduction), but is technically different because the mass action kinetics dynamics of multiple species chemical reactions is nonlinear.

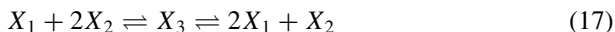
6.1 The Canonical Representation of Detailed-Balanced Mass Action Kinetics Reaction Networks

First recall from [4] the representation of detailed-balanced mass action kinetics reaction networks as nonlinear physical network systems, which share the same underlying network structure with the linear physical network systems as treated before. This representation will form the basis of the structure-preserving model reduction approaches based on clustering and Kron reduction, to be treated in the next two subsections.

Consider an isothermal chemical reaction network involving m chemical species (metabolites), among which r chemical reactions take place. The basic structure underlying the dynamics of the vector $x \in \mathbb{R}_+^m$ of concentrations $x_i, i = 1, \dots, m$, of the chemical species is given by the *balance laws*

$$\dot{x} = Sv, \quad (16)$$

where S is an $m \times r$ matrix, called the *stoichiometric matrix*, consisting of (positive and negative) integer elements. The elements of the vector $v \in \mathbb{R}^r$ are called the *reaction rates*. For example, the stoichiometric matrix of the two coupled reactions involving the chemical species X_1, X_2, X_3 given as



is

$$S = \begin{bmatrix} -1 & 2 \\ -2 & 1 \\ 1 & -1 \end{bmatrix} \quad (18)$$

Chemical reaction network theory, as originating from the fundamental papers [16–18], identifies the nodes of the network with the *complexes* of the chemical reactions, i.e., all the different left- and right-hand sides (substrates and products) of the reactions in the network. For example, the complexes of the network (17) are $X_1 + 2X_2, X_3$, and $2X_1 + X_2$. The complexes on the left-hand side of the reactions are called the *substrate* complexes and those on the right-hand side of the reactions the *product* complexes.

Denoting the number of complexes by c , the expression of the complexes in terms of the chemical species is formalized by an $m \times c$ matrix Z , called the *complex composition matrix*, whose ρ -th column captures the expression of the ρ -th complex in the m chemical species. For the network (17)

$$Z = \begin{bmatrix} 1 & 0 & 2 \\ 2 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad (19)$$

The complexes can be associated with the nodes of a *directed graph*, with edges corresponding to the reactions. The resulting graph, called the *graph of complexes*, is characterized by its incidence matrix D . The relation between S , Z and D is simply given as $S = ZD$.

The most basic model for specifying the reaction rates is *mass action kinetics*, defined as follows. Consider the first reaction in (17) $X_1 + 2X_2 \rightleftharpoons X_3$, involving the three chemical species X_1, X_2, X_3 with concentrations x_1, x_2, x_3 . The reaction is considered to be a combination of the *forward reaction* $X_1 + 2X_2 \rightarrow X_3$ with $v^+(x_1, x_2) = k^+ x_1 x_2^2$ and the *backward reaction* $X_1 + 2X_2 \leftarrow X_3$, with $v^-(x_3) = k^- x_3$. The constants k^+, k^- are called respectively the *forward* and *backward reaction constants*. The net reaction rate is thus

$$v(x_1, x_2, x_3) = v^+(x_1, x_2) - v^-(x_3) = k^+ x_1 x_2^2 - k^- x_3 \quad (20)$$

In general, the mass action reaction rate of the j -th reaction of a chemical reaction network, from a substrate complex \mathcal{S}_j to a product complex \mathcal{P}_j , is given as

$$v_j(x) = k_j^+ \prod_{i=1}^m x_i^{Z_{i\mathcal{S}_j}} - k_j^- \prod_{i=1}^m x_i^{Z_{i\mathcal{P}_j}}, \quad (21)$$

where $Z_{i\rho}$ is the (i, ρ) -th element of the complex composition matrix Z , and $k_j^+, k_j^- \geq 0$, are the forward/backward reaction constants of the j -th reaction, respectively. Let $Z_{\mathcal{S}_j}$ and $Z_{\mathcal{P}_j}$ denote the columns of the complex composition matrix Z corresponding to the substrate complex \mathcal{S}_j and the product complex \mathcal{P}_j of the j -th reaction. Defining the mapping $\text{Ln} : \mathbb{R}_+^c \rightarrow \mathbb{R}^c$ to be the elementwise logarithm, the mass action reaction Eq. (21) for the j -th reaction can be rewritten as

$$v_j(x) = k_j^+ \exp(Z_{\mathcal{S}_j}^T \text{Ln}(x)) - k_j^- \exp(Z_{\mathcal{P}_j}^T \text{Ln}(x)) \quad (22)$$

Definition 1 A vector of concentrations $x^* \in \mathbb{R}_+^m$ is called a *thermodynamic equilibrium* if $v(x^*) = 0$. A chemical reaction network $\dot{x} = Sv(x)$ is called *detailed-balanced* if there exists a thermodynamic equilibrium $x^* \in \mathbb{R}_+^m$.

Let $x^* \in \mathbb{R}_+^m$ be any thermodynamic equilibrium. Define the 'conductance' of the j -th reaction as

$$\kappa_j(x^*) := k_j^+ \exp\left(Z_{\mathcal{S}_j}^T \text{Ln}(x^*)\right) = k_j^- \exp\left(Z_{\mathcal{P}_j}^T \text{Ln}(x^*)\right) \quad (23)$$

Then the reaction rate (22) can be rewritten as

$$v_j(x) = \kappa_j(x^*) \left[\exp\left(Z_{\mathcal{S}_j}^T \text{Ln}\left(\frac{x}{x^*}\right)\right) - \exp\left(Z_{\mathcal{P}_j}^T \text{Ln}\left(\frac{x}{x^*}\right)\right) \right] \quad (24)$$

Defining the $r \times r$ diagonal matrix of conductances as

$$\mathcal{K} := \text{diag}(\kappa_1(x^*), \dots, \kappa_r(x^*)) \quad (25)$$

it follows that the vector of mass action reaction rates of a detailed-balanced reaction network can be written as

$$v(x) = -\mathcal{K}D^T \text{Exp} \left(Z^T \text{Ln} \left(\frac{x}{x^*} \right) \right), \quad (26)$$

where Exp denotes the elementwise exponential mapping. Hence the dynamics of the network takes the form, cf. [4],

$$\dot{x} = -ZDKD^T \text{Exp} \left(Z^T \text{Ln} \left(\frac{x}{x^*} \right) \right), \quad \mathcal{K} > 0 \quad (27)$$

The matrix $L := DKD^T$ defines a *weighted Laplacian* matrix for the graph of complexes.

Note the similarity, as well as the difference, with the dynamics (1) as studied so far. The similarity is the underlying directed graph and the weighted Laplacian matrix. The differences are the presence of the complex composition matrix Z , which defines a *representation* [15] of the graph in the space \mathbb{R}^m of chemical species, and the nonlinearity of the expression $\text{Exp} \left(Z^T \text{Ln} \left(\frac{x}{x^*} \right) \right)$, instead of the linear expression $M^{-1}x$ as encountered before. Note that in case $Z = I$ (single-species chemical reactions) (27) reduces to (1), with M the diagonal matrix with diagonal elements given by the components of x^* ; cf. Example 4.

As shown in [4], the dynamics (27) can be further rewritten in the following form. Define the vector of *chemical potentials* μ and the reference chemical potential vector μ^0 as

$$\mu(x) := RT \text{Ln} \left(\frac{x}{x^*} \right), \quad \mu^0 := -RT \text{Ln}(x^*), \quad (28)$$

with T the constant temperature and R the universal gas constant. Furthermore, define the Gibbs' energy as

$$G(x) = RT x^T \text{Ln} \left(\frac{x}{x^*} \right) + RT (x^* - x)^T \mathbb{1}_m \quad (29)$$

It is immediately checked that $\frac{\partial G}{\partial x}(x) = RT \text{Ln} \left(\frac{x}{x^*} \right) = \mu(x)$. Hence (27) can be equivalently written as

$$\dot{x} = -ZL \text{Exp} \left(\frac{1}{RT} Z^T \frac{\partial G}{\partial x}(x) \right), \quad (30)$$

Equation (30) forms the basis of the stability analysis performed in [4], re-deriving classical results obtained in e.g. [16–18] in a very simple and insightful manner. Indeed, the function G can be used as a Lyapunov function, using the fundamental inequality

$$\gamma^T L \text{Exp } \gamma \geq 0 \text{ for all } \gamma, \quad \gamma^T L \text{Exp } \gamma = 0 \text{ iff } D^T \gamma = 0 \quad (31)$$

Furthermore, the framework is immediately extended to chemical reaction networks subject to external in/outflows of chemical species. This leads to the input-state-output system

$$\begin{aligned} \dot{x} &= -ZDKD^T \text{Exp} \left(Z^T \text{Ln} \left(\frac{x}{x^*} \right) \right) + S_b u \\ y &= S_b RT \text{Ln} \left(\frac{x}{x^*} \right), \end{aligned} \quad (32)$$

for a certain matrix S_b (typically composed of some basis vectors), with y the vector of *boundary chemical potentials* (the chemical potentials of the species whose in/outflow is controlled by u). This leads to the passivity property

$$\frac{d}{dt} G \leq y^T u \quad (33)$$

Remark 1 We mention that the above theory can be further generalized to *complex-balanced* mass action reaction networks, cf. [19], preserving the main stability results. In this case, the matrix $L = DKD^T$ in (27) is replaced by a, possibly non-symmetric, Laplacian matrix which is *balanced*, i.e., column and row sums are zero.

6.2 Structure-Preserving Model Reduction by Clustering

Consider a detailed-balanced chemical reaction network as described in the canonical form (27). Structure-preserving model reduction by clustering amounts to clustering the *complexes* of the reaction network, corresponding to a characteristic matrix P of the *graph of complexes*. As before, this leads to a *reduced Laplacian matrix*

$$\hat{L} := \hat{D} \hat{K} \hat{D}^T, \quad P^T D = [\hat{D} \ 0] \quad (34)$$

Furthermore, the *reduced complex composition matrix* is defined as

$$\hat{Z} := ZP, \quad (35)$$

which corresponds to a smaller set of new complexes, which are obtained by joining the complexes in each cell. Taken together, this defines the reduced dynamics

$$\dot{x} = -\hat{Z} \hat{D} \hat{K} \hat{D}^T \text{Exp} \left(\hat{Z}^T \text{Ln} \left(\frac{x}{x^*} \right) \right) + S_b u, \quad (36)$$

which is again in the form (27), with reduced stoichiometric matrix $\hat{S} := \hat{Z} \hat{D}$.

A key difference with the reduced linear dynamics (3) is the fact that the state vector x remains the same; instead of being replaced by a lower-dimensional state

vector \hat{x} as in (3). Thus the dimension of the dynamics has not been reduced, but instead the dynamics has been simplified by reducing the number of complexes and reactions.²

Application of this clustering method to the special case of single-species chemical reaction networks ($Z = I$) results in the dynamics (take for simplicity $u = 0$)

$$\dot{x} = -P\hat{D}\hat{K}\hat{D}^T \text{Exp}\left(P^T \text{Ln}\left(\frac{x}{x^*}\right)\right) \quad (37)$$

Interestingly, this reduced model is *different* from the reduced model resulting from applying the linear clustering reduction as proposed in the previous section (cf. Example 4) to

$$\dot{x} = -IDKD^T \text{Exp}\left(I^T \text{Ln}\left(\frac{x}{x^*}\right)\right) = -DKD^T \frac{x}{x^*} \quad (38)$$

Indeed, in this case the reduced model is (with $[x^*]$ denoting the diagonal matrix with diagonal x^*)

$$\dot{\hat{x}} = -\hat{D}\hat{K}\hat{D}^T (P^T [x^*] P)^{-1} \hat{x} \quad (39)$$

This reduced model is *linear* in the reduced vector of chemical species $\hat{x} = P^T x$, while on the other hand the reduced model (36) is *nonlinear* in the full state vector x . This is due to an intrinsic difference: the structure-preserving reduction method as discussed in the present section is based on *clustering the complexes*, and then applying mass action kinetics to the clustered complexes; thus giving rise to nonlinear dynamics. On a more fundamental level, the energy function of the linear system (1) is taken to be the quadratic function $\frac{1}{2}x^T M^{-1}x$, while the *physical* energy function of chemical reaction networks, even for the single-species case, is instead the Gibbs' energy $G(x) = RTx^T \text{Ln}\left(\frac{x}{x^*}\right) + RT(x^* - x)^T \mathbf{1}_m$.

6.3 Structure Preserving Model Reduction by Kron Reduction

Kron reduction, as treated before, can be applied to detailed-balanced chemical reaction networks described in the canonical form (27) as follows. Similarly as before, the c nodes of the graph of complexes are split into two distinct subsets K and J , and by deleting the complexes in the subset J we obtain a *reduced graph of complexes*, with weighted Laplacian $\hat{L} = \hat{D}\hat{K}\hat{D}^T$. For simplicity we will assume that the rows of S_b corresponding to the index subset J are zero, and the remaining matrix without these zero rows will be denoted by \hat{S}_b .

Furthermore, by leaving out the columns of the complex stoichiometric matrix Z corresponding to the index set J one obtains a *reduced complex composition*

² In some sense, the reduced dynamics (36) is similar to the dynamics of $P\hat{v}$ as in (6).

matrix \hat{Z} (with as many columns as the number of complexes in the reduced graph of complexes; i.e., the cardinality of K), leading to the reduced chemical reaction network

$$\dot{x} = -\hat{Z}\hat{D}\hat{K}\hat{D}^T \text{Exp}\left(\hat{Z}^T \text{Ln}\left(\frac{x}{x^*}\right)\right) + \hat{S}_b u, \quad \hat{K} > 0 \quad (40)$$

This again defines a *balanced chemical reaction network* governed by mass action kinetics in the canonical form (27), with a reduced number of complexes and stoichiometric matrix $\hat{S} := \hat{Z}\hat{D}$. Note that this reduced stoichiometric matrix is fundamentally different from the reduced stoichiometric matrix \hat{S} as obtained by clustering, both in graph topology captured by \hat{D} , and in the definition of the complexes encoded in the complex stoichiometric matrix \hat{Z} .

The dynamical interpretation of this Kron reduction procedure for chemical reaction networks is similar as before. Write out the full-order chemical reaction network as

$$\dot{x} = -\begin{bmatrix} Z_K & Z_J \end{bmatrix} \begin{bmatrix} L_{KK} & L_{KJ} \\ L_{JK} & L_{JJ} \end{bmatrix} \begin{bmatrix} \text{Exp}\left(Z_K^T \text{Ln}\left(\frac{x}{x^*}\right)\right) \\ \text{Exp}\left(Z_J^T \text{Ln}\left(\frac{x}{x^*}\right)\right) \end{bmatrix} \quad (41)$$

Consider now the *auxiliary* dynamical system

$$\begin{bmatrix} \dot{y}_K \\ \dot{y}_J \end{bmatrix} = -\begin{bmatrix} L_{KK} & L_{KJ} \\ L_{JK} & L_{JJ} \end{bmatrix} \begin{bmatrix} w_K \\ w_J \end{bmatrix}, \quad (42)$$

where we impose the constraint $\dot{y}_J = 0$. It follows that $w_J = -L_{JJ}^{-1}L_{JK}w_K$, leading to the reduced dynamics $\dot{y}_K = -(L_{KK} - L_{KJ}L_{JJ}^{-1}L_{JK})w_K = -\hat{L}w_K$. Substituting $w_K = \text{Exp}(\hat{Z}_K^T \text{Ln}(\frac{x}{x^*}))$, making use of $\dot{x} = Z_K \dot{y}_K + Z_J \dot{y}_J = Z_K \dot{y}_K = \hat{Z} \dot{y}_K$, we then obtain the reduced network given by (40).

We conclude that Kron reduction of chemical reaction networks rests on the assumption that the deleted complexes in the subset J are (approximately) constant in time; or, equivalently, the total flows into the deleted complexes are (almost) zero. Applications of the Kron reduction method to a number of detailed-balanced chemical reaction networks can be found in [4, 20].

7 Conclusions and Outlook

In this paper, continuing upon our previous work [2, 4, 11, 13, 20], we have elaborated on two approaches to structure-preserving model reduction for physical network systems, namely *clustering* and *Kron reduction*. Although both methods give rise to a natural network reduction, the important question of obtaining tight error bounds is still largely open. Another problem concerns the extension of these two methods to physical network systems that also contain energy storage associated to the *edges* of the underlying graph; like mass-spring-damper systems (where spring energy is associated to part of the edges [1]). A final open problem concerns the reduced-order

network modeling of physical systems based on *data*; that is the determination of an underlying graph, together with energy storage and static couplings, resulting in a physical network system (1) that is explaining the data in an (approximate) way.

Acknowledgements This paper is dedicated to Thanos Antoulas at the occasion of his 70th birthday, for his fundamental contributions to model reduction theory, and for his scientific leadership, enthusiasm and collegiality. I have tremendously enjoyed our scientific contacts, and do hope for more. Happy birthday Thanos!

References

1. van der Schaft, A.J., Maschke, B.M.: Port-Hamiltonian systems on graphs. *SIAM J. Control. Optim.* **51**(2), 906–937 (2013)
2. van der Schaft, A.J.: On model reduction of physical network systems. In: Proceedings of the 21st International Symposium on Mathematical Theory of Networks and Systems (MTNS2014), pp. 1419–1425. 7–14 July 2014. Groningen, the Netherlands
3. van der Schaft, A.J.: Modeling of physical network systems. *Syst. Control Lett.* **101**, 1–27 (2017)
4. van der Schaft, A.J., Rao, S., Jayawardhana, B.: On the mathematical structure of balanced chemical reaction networks governed by mass action kinetics. *SIAM J. Appl. Math.* **73**(2), 953–973 (2013)
5. Mesbahi, M., Egerstedt, M.: Graph theoretic methods in multiagent networks. Princeton University Press (2010)
6. Imura, J.-I.: Clustered model reduction of large-scale complex networks. In: Proceedings of the 20th international symposium on mathematical theory of networks and systems (MTNS), Melbourne, Australia (2012)
7. Ishizaki, T., Kashida, K., Imura, J.-I., Aihara, K.: Network clustering for SISO linear dynamical networks via reaction-diffusion transformation. In: Proceedings of the 18th IFAC world congress, Milan, Italy, pp. 5639–5644 (2011)
8. Sandberg, H., Murray, R.M.: Model reduction of interconnected linear systems. *Optim. Control Appl. Methods* **30**(3), 225–245 (2009)
9. Chen, X., Kawano, Y., Scherpen, J.M.A.: Model reduction of multi-agent systems using dissimilarity-based clustering. *IEEE Trans. Autom. Control* **64**(4), 1663–1670 (2019)
10. Monshizadeh, N., Camlibel, M.K., Trentelman, H.L.: Projection based model reduction of multi-agent systems using graph partitions. *IEEE Trans. Control. Netw. Syst.* **1**(2), 145–154 (2014)
11. van der Schaft, A.J.: Physical network systems and model reduction. In: Belur, M.N., Camlibel, M.K., Rapisarda, P., Scherpen, J.M.A. (eds.) *Mathematical Control Theory II, Behavioral Systems and Robust Control*, eds., Springer Lecture Notes in Control and Information Sciences, vol. 462, pp. 199–210 (2015)
12. Besselink, B., Sandberg, H., Johansson, K.H.: Clustering-based model reduction of networked passive systems. *IEEE Trans. Autom. Control* **61**(10), 2958–2973 (2016)
13. Monshizadeh, N., van der Schaft, A.J.: Structure-preserving model reduction of physical network systems by clustering. In: Proceedings 53rd IEEE Conference on Decision and Control, Los Angeles, CA, USA (2014)
14. van der Schaft, A.J.: Characterization and partial synthesis of the behavior of resistive circuits at their terminals. *Syst. Control Lett.* **59**, 423–428 (2010)
15. Bollobas, B.: *Modern Graph Theory*, Graduate Texts in Mathematics, vol. 184. Springer, New York (1998)
16. Horn, F.J.M., Jackson, R.: General mass action kinetics. *Arch. Rational Mech. Anal.* **47**, 81–116 (1972)

17. Horn, F.J.M.: Necessary and sufficient conditions for complex balancing in chemical kinetics. *Arch. Rational Mech. Anal.* **49**, 172–186 (1972)
18. Feinberg, M.: The existence and uniqueness of steady states for a class of chemical reaction networks. *Arch. Rational Mech. Anal.* **132**, 311–370 (1995)
19. Rao, S., van der Schaft, A.J., Jayawardhana, B.: A graph-theoretical approach for the analysis and model reduction of complex-balanced chemical reaction networks. *J. Math. Chem.* **51**, 2401–2422 (2013)
20. Rao, S., van der Schaft, A.J., van Eunen, K., Bakker, B.M., Jayawardhana, B.: A model reduction method for biochemical reaction networks. *BMC Syst. Biol.* **8**, 52 (2014)

Model Reduction for Control

\mathcal{H}_2 -gap Model Reduction for Stabilizable and Detectable Systems



Tobias Breiten, Christopher Beattie, and Serkan Gugercin

Abstract We formulate here an approach to model reduction that is well-suited for linear time-invariant control systems that are stabilizable and detectable but may otherwise be unstable. We introduce a modified \mathcal{H}_2 -error metric, the \mathcal{H}_2 -gap, that provides an effective measure of model fidelity in this setting. While the direct evaluation of the \mathcal{H}_2 -gap requires the solutions of a pair of algebraic Riccati equations associated with related closed-loop systems, we are able to work entirely within an interpolatory framework, developing algorithms and supporting analysis that do not reference full-order closed-loop Gramians. This leads to a computationally effective strategy yielding reduced models designed so that the corresponding reduced closed-loop systems will interpolate the full-order closed-loop system at specially adapted interpolation points, without requiring evaluation of the full-order closed-loop system nor even computation of the feedback law that determines it. The analytical framework and computational algorithm presented here provides an effective new approach toward constructing reduced-order models for unstable systems. Numerical examples for an unstable convection diffusion equation and a linearized incompressible Navier-Stokes equation illustrate the effectiveness of this approach.

Keywords \mathcal{H}_2 optimality · Unstable systems · Interpolation · Riccati equations

T. Breiten (✉)

Institute of Mathematics, Technical University of Berlin, 10623 Berlin, Germany
e-mail: tobias.breiten@tu-berlin.de

C. Beattie · S. Gugercin

Department of Mathematics, Virginia Tech, Blacksburg, VA 24061-0123, USA
e-mail: beattie@vt.edu

S. Gugercin

e-mail: gugercin@vt.edu

© Springer Nature Switzerland AG 2022

C. Beattie et al. (eds.), *Realization and Model Reduction of Dynamical Systems*,
https://doi.org/10.1007/978-3-030-95157-3_17

1 Introduction

Consider a linear time-invariant (LTI) control system

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), & \mathbf{x}(0) &= \mathbf{x}_0, \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t),\end{aligned}\tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{B} \in \mathbb{R}^{n \times m}$ and $\mathbf{C} \in \mathbb{R}^{p \times n}$. The quantities $\mathbf{x}(t)$, $\mathbf{u}(t)$ and $\mathbf{y}(t)$ denote respectively the state, control and output of the system, viewed as vector-valued functions of time. In practical applications, for example when (1) is obtained by a (method-of-lines) semidiscretization of a partial differential equation, the dimension n of the state space can remain very large; indeed, often large enough to impede subsequent analysis. In this case, *model reduction* provides useful tools for constructing simpler surrogates or *reduced-order models* (ROMs) that produce dynamics analogous to (1):

$$\begin{aligned}\hat{\mathbf{x}}(t) &= \hat{\mathbf{A}}\hat{\mathbf{x}}(t) + \hat{\mathbf{B}}\mathbf{u}(t), & \hat{\mathbf{x}}(0) &= \hat{\mathbf{x}}_0, \\ \hat{\mathbf{y}}(t) &= \hat{\mathbf{C}}\hat{\mathbf{x}}(t),\end{aligned}\tag{2}$$

where $\hat{\mathbf{A}} \in \mathbb{R}^{r \times r}$ and $\hat{\mathbf{B}} \in \mathbb{R}^{r \times m}$ and $\hat{\mathbf{C}} \in \mathbb{R}^{p \times r}$ are to be determined so that $r \ll n$ and $\hat{\mathbf{y}}(t) \approx \mathbf{y}(t)$ for all $t \geq 0$ and all $\mathbf{u} \in \mathcal{U}$, where, e.g., $\mathcal{U} = L^2(0, \infty; \mathbb{R}^m)$. Assuming null initial conditions for (1) and (2), we may transform (1) and (2) into the frequency domain and introduce the corresponding transfer functions, $\mathbf{G}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$ and $\hat{\mathbf{G}}(s) = \hat{\mathbf{C}}(s\mathbf{I} - \hat{\mathbf{A}})^{-1}\hat{\mathbf{B}}$, defined for $s \in \mathbb{C}$ with $\operatorname{Re}(s) > \alpha$ for $\alpha \in \mathbb{R}$ sufficiently large. Observing that \mathbf{G} and $\hat{\mathbf{G}}$ are each rational functions of s , the model reduction problem may be restated as a rational approximation problem

$$\tilde{\mathbf{G}} = \arg \min_{\dim(\tilde{\mathbf{G}})=r} \|\mathbf{G} - \tilde{\mathbf{G}}\|\tag{3}$$

with respect to an appropriately chosen norm. For example, in case that \mathbf{A} is asymptotically stable (i.e., eigenvalues of \mathbf{A} lie in the open left half-plane, \mathbb{C}_-), the most prominent choices are the \mathcal{H}_∞ and the \mathcal{H}_2 -norms. The \mathcal{H}_∞ and the \mathcal{H}_2 spaces with the associated norms are given by

$$\begin{aligned}\mathcal{H}_2 &= \left\{ \mathbf{F}: \mathbb{C}^+ \rightarrow \mathbb{C}^{p \times m} \mid \mathbf{F} \text{ is analytic, } \|\mathbf{F}\|_{\mathcal{H}_2} := \left(\sup_{\sigma > 0} \int_{-\infty}^{\infty} \|\mathbf{F}(\sigma + i\omega)\|_F^2 d\omega \right)^{\frac{1}{2}} < \infty \right\}, \\ \mathcal{H}_\infty &= \left\{ \mathbf{F}: \mathbb{C}^+ \rightarrow \mathbb{C}^{p \times m} \mid \mathbf{F} \text{ is analytic, } \|\mathbf{F}\|_{\mathcal{H}_\infty} := \sup_{z \in \mathbb{C}^+} \|\mathbf{F}(z)\|_2 < \infty \right\}.\end{aligned}$$

We focus primarily on cases where \mathbf{A} is *unstable*, i.e., having at least one eigenvalue in \mathbb{C}_+ . The complementary case where \mathbf{A} is asymptotically stable is a standard setting that assures whenever $\mathbf{u} \in L^2(0, \infty; \mathbb{R}^m)$, then also $\mathbf{x} \in L^2(0, \infty; \mathbb{R}^n)$ and $\mathbf{y} \in L^2(0, \infty; \mathbb{R}^p)$. When \mathbf{A} is unstable, the potential for explosive growth of system

state and output may reflect features that are fundamental to the modeled dynamics, either naturally so (e.g., linearizations of energetic gyres in ocean circulation) or by design (e.g., agile flight vehicles that depend on active control strategies to survive high-speed maneuvers).

For such systems a variety of model reduction strategies have been proposed. One possibility is to consider the model reduction problem on a finite time horizon $[0, T_{\max}]$; see for example, [14, 16, 21, 28]. In cases that $\sigma(\mathbf{A}) \cap i\mathbb{R} = \emptyset$, an alternative approach is to decouple the (unstable) system as $\mathbf{G} = \mathbf{G}_s + \mathbf{G}_u$, into a purely stable and a purely anti-stable part, then perform model reduction on each of these two subsystems, as was done e.g., using \mathcal{H}_2 -optimal interpolation techniques in [23]. Yet another approach extends balanced truncation to unstable systems [5, 9, 31] using frequency domain definitions of the system Gramians.

In this work, we also focus on potentially unstable systems, but we do assume that (\mathbf{A}, \mathbf{B}) is stabilizable and (\mathbf{A}, \mathbf{C}) is detectable, a circumstance that commonly occurs. It is well known ([25]), that in this case it is possible to factorize the transfer function as $\mathbf{G} = \mathbf{M}^{-1}\mathbf{N}$, with transfer functions $\mathbf{M}, \mathbf{N} \in \mathcal{H}_\infty$. By introducing a similar representation for the reduced transfer function, $\widehat{\mathbf{G}} = \widehat{\mathbf{M}}^{-1}\widehat{\mathbf{N}}$, and a particular (unique) left-coprime factorization, we consider an \mathcal{H}_2 -type best-approximation problem having the form

$$[\widetilde{\mathbf{M}}, \widetilde{\mathbf{N}}] = \arg \min_{\substack{\dim(\widetilde{\mathbf{M}})=r \\ \dim(\widetilde{\mathbf{N}})=r}} \left\| [\mathbf{M}, \mathbf{N}] - [\widehat{\mathbf{M}}, \widehat{\mathbf{N}}] \right\|_{\mathcal{H}_2}, \quad (4)$$

which can be interpreted as seeking reduced order *factors* that are as close as possible to the corresponding factors of the original system. We refer to the error measure in (4) as the \mathcal{H}_2 -gap. Our approach is motivated by the method of linear quadratic Gaussian (LQG) balancing ([12, 20, 24]) which introduces a similar approximation problem, but using the \mathcal{H}_∞ -norm instead of the \mathcal{H}_2 -norm as we have posed it here. Besides the evident applicability to unstable systems that is our focus, the \mathcal{H}_∞ -induced metric also carries particular significance for associated closed-loop behavior, see, e.g., [29].

Starting in Sect. 2, we will introduce our approximation problem in more detail and review some known results on left-coprime factorizations for stabilizable and detectable LTI systems. We will show that our approximation problem has a natural connection to an $\mathcal{L}_2(i\mathbb{R})$ -model reduction problem. Section 3 provides a pole-residue expansion for the \mathcal{H}_2 -gap and analyzes individual error terms. We suggest a modification of the iterative rational Krylov algorithm (IRKA) in order to eliminate portions of this error expression. In Sect. 4, we consider some numerical examples that suggest the competitiveness of our approach relative to existing methods. Conclusions with perspectives for future research are given in Sect. 5.

2 Left-Coprime Factorizations and the Gap Metric

In this section, we provide additional details for the specific error measure described in (4). Since we do not assume the system matrix \mathbf{A} to be asymptotically stable, the \mathcal{H}_2 -norm of $\mathbf{G}(\cdot)$, for $\mathbf{G}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$, may not be finite. However, for a stabilizable and detectable LTI system, we can select any matrix \mathbf{F} so that $\mathbf{A}_F := \mathbf{A} - \mathbf{F}\mathbf{C}$ is asymptotically stable (see e.g., [29, Lemma 6.1]), and then

$$\mathbf{M}(s) = \mathbf{I} - \mathbf{C}(s\mathbf{I} - \mathbf{A}_F)^{-1}\mathbf{F} \quad \mathbf{N}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A}_F)^{-1}\mathbf{B} \quad (5)$$

defines a left-coprime factorization of $\mathbf{G}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$. In particular,

$$\mathbf{G}(s) = [\mathbf{M}(s)]^{-1}\mathbf{N}(s), \quad \text{where} \quad [\mathbf{M}(s)]^{-1} = \mathbf{I} + \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{F}. \quad (6)$$

In the context of LQG-balanced truncation, e.g., in [24], the stabilizing matrix \mathbf{F} is been chosen as $\mathbf{F} = \mathbf{P}\mathbf{C}^T$ where $\mathbf{P} = \mathbf{P}^T \geq 0$ denotes the solution of the algebraic Riccati equation

$$\mathbf{A}\mathbf{P} + \mathbf{P}\mathbf{A}^T - \mathbf{P}\mathbf{C}^T\mathbf{C}\mathbf{P} + \mathbf{B}\mathbf{B}^T = 0. \quad (7)$$

Note that asymptotic stability of the system matrix \mathbf{A}_F is ensured by the stabilizability and detectability properties of (\mathbf{A}, \mathbf{B}) and (\mathbf{A}, \mathbf{C}) , respectively. Throughout the rest of this work, we assume that $\mathbf{F} = \mathbf{P}\mathbf{C}^T$ and refer to $\mathbf{A}_F = \mathbf{A} - \mathbf{P}\mathbf{C}^T\mathbf{C}$ as the *closed-loop* system matrix.

Similarly, for a stabilizable and detectable reduced system $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\mathbf{C}})$, we denote by $\widehat{\mathbf{F}}, \widehat{\mathbf{M}}, \widehat{\mathbf{N}}, \widehat{\mathbf{P}}$ the equivalent reduced-order expressions; in particular,

$$\widehat{\mathbf{G}}(s) = [\widehat{\mathbf{M}}(s)]^{-1}\widehat{\mathbf{N}}(s), \quad \widehat{\mathbf{M}}(s) = \mathbf{I} - \widehat{\mathbf{C}}(s\mathbf{I} - \widehat{\mathbf{A}}_F)^{-1}\widehat{\mathbf{F}}, \quad \text{and} \quad \widehat{\mathbf{N}}(s) = \widehat{\mathbf{C}}(s\mathbf{I} - \widehat{\mathbf{A}}_F)^{-1}\widehat{\mathbf{B}}. \quad (8)$$

Moreover, it holds that $\mathbf{M}, \widehat{\mathbf{M}} \in \mathcal{H}_\infty$ and $\mathbf{N}, \widehat{\mathbf{N}} \in \mathcal{H}_\infty \cap \mathcal{H}_2$. Define the two new systems

$$\mathbf{G}_F(s) := [\mathbf{M}(s), \mathbf{N}(s)], \quad \widehat{\mathbf{G}}_F(s) := [\widehat{\mathbf{M}}(s), \widehat{\mathbf{N}}(s)]. \quad (9)$$

Using (5) and (9), we can compute a state-space realization for \mathbf{G}_F :

$$\begin{aligned} \mathbf{G}_F(s) &= [\mathbf{M}(s), \mathbf{N}(s)] = [\mathbf{I} - \mathbf{C}(s\mathbf{I} - \mathbf{A}_F)^{-1}\mathbf{F}, \mathbf{C}(s\mathbf{I} - \mathbf{A}_F)^{-1}\mathbf{B}] \\ &= [\mathbf{I}, 0] + \mathbf{C}(s\mathbf{I} - \mathbf{A}_F)^{-1}[-\mathbf{F}, \mathbf{B}]. \end{aligned} \quad (10)$$

A state-space realization for $\widehat{\mathbf{G}}_F$ can be obtained similarly:

$$\widehat{\mathbf{G}}_F(s) = [\mathbf{I}, 0] + \widehat{\mathbf{C}}(s\mathbf{I} - \widehat{\mathbf{A}}_F)^{-1}[-\widehat{\mathbf{F}}, \widehat{\mathbf{B}}]. \quad (11)$$

Note that \mathbf{G}_F and $\widehat{\mathbf{G}}_F$ are dynamical systems with $m + p$ inputs and p outputs. LQG balanced truncation exploits that the Gramians of \mathbf{G}_F and $\widehat{\mathbf{G}}_F$ are closely related to \mathbf{P} , $\widehat{\mathbf{P}}$ and their dual counterparts \mathbf{Q} , $\widehat{\mathbf{Q}}$, which satisfy

$$\begin{aligned} \mathbf{A}^T \mathbf{Q} + \mathbf{Q} \mathbf{A} - \mathbf{Q} \mathbf{B} \mathbf{B}^T \mathbf{Q} + \mathbf{C}^T \mathbf{C} &= 0 \quad \text{and} \\ \widehat{\mathbf{A}}^T \widehat{\mathbf{Q}} + \widehat{\mathbf{Q}} \widehat{\mathbf{A}} - \widehat{\mathbf{Q}} \widehat{\mathbf{B}} \widehat{\mathbf{B}}^T \widehat{\mathbf{Q}} + \widehat{\mathbf{C}}^T \widehat{\mathbf{C}} &= 0. \end{aligned}$$

In fact, as shown in [12], the controllability and observability Gramians \mathcal{L}_c and \mathcal{L}_o of \mathbf{G}_F are given by $\mathcal{L}_c = \mathbf{P}$ and $\mathcal{L}_o = \mathbf{Q}(\mathbf{I} + \mathbf{P}\mathbf{Q})^{-1}$. Based on these relations, balancing and truncation with respect to \mathbf{P} and \mathbf{Q} allows to construct a reduced-order model that satisfies an error bound of the following type

$$\|[\mathbf{M}, \mathbf{N}] - [\widehat{\mathbf{M}}, \widehat{\mathbf{N}}]\|_{\mathcal{H}_\infty} \leq 2 \sum_{i=r+1}^n \sigma_i, \quad (12)$$

where σ_i are the Hankel singular values of \mathbf{G}_F .

While balanced truncation is primarily related to the \mathcal{H}_∞ -norm, we are interested in the case when the deviation between \mathbf{G}_F and $\widehat{\mathbf{G}}_F$ is measured using the \mathcal{H}_2 -norm. For this purpose, assume that left-coprime factorizations $\mathbf{G}_F = [\mathbf{M}, \mathbf{N}]$, $\widehat{\mathbf{G}}_F = [\widehat{\mathbf{M}}, \widehat{\mathbf{N}}]$ with $\mathbf{F} = \mathbf{P}\mathbf{C}^T$ and $\widehat{\mathbf{F}} = \widehat{\mathbf{P}}\widehat{\mathbf{C}}^T$ are given. We then define

$$\|\mathbf{G} - \widehat{\mathbf{G}}\|_{\mathcal{H}_2\text{-gap}} := \|[\mathbf{M}, \mathbf{N}] - [\widehat{\mathbf{M}}, \widehat{\mathbf{N}}]\|_{\mathcal{H}_2}. \quad (13)$$

Note that even though \mathbf{M} , $\widehat{\mathbf{M}} \notin \mathcal{H}_2$, the previous expression is well-defined since

$$\mathbf{M}(\cdot) - \widehat{\mathbf{M}}(\cdot) = \widehat{\mathbf{C}}(\mathbf{I} - \widehat{\mathbf{A}}_F)^{-1} \widehat{\mathbf{F}} - \mathbf{C}(\mathbf{I} - \mathbf{A}_F)^{-1} \mathbf{F} \in \mathcal{H}_2.$$

If \mathbf{G} and $\widehat{\mathbf{G}}$ have no poles on the imaginary axis, we show below that the \mathcal{H}_2 -gap, (13), provides a bound to the $\mathcal{L}_2(i\mathbb{R})$ -error. For this purpose, as in [4, Sect. 5] we define

$$\begin{aligned} \mathcal{L}_2(i\mathbb{R}) &:= \left\{ \mathbf{H}: \mathbb{C} \rightarrow \mathbb{C}^{p \times m} \mid \mathbf{H} \text{ is meromorphic, } \|\mathbf{H}\|_{\mathcal{L}_2} := \left(\int_{-\infty}^{\infty} \|\mathbf{H}(i\omega)\|_F^2 d\omega \right)^{\frac{1}{2}} < \infty \right\}, \\ \mathcal{L}_\infty(i\mathbb{R}) &:= \left\{ \mathbf{H}: \mathbb{C} \rightarrow \mathbb{C}^{p \times m} \mid \mathbf{H} \text{ is meromorphic, } \|\mathbf{H}\|_{\mathcal{L}_\infty} := \sup_{\omega \in \mathbb{R}} \sigma_{\max}(\mathbf{H}(i\omega)) < \infty \right\}. \end{aligned}$$

In particular, we have the orthogonal decomposition $\mathcal{L}_2(i\mathbb{R}) = \mathcal{H}_2(\mathbb{C}^-) \oplus \mathcal{H}_2(\mathbb{C}^+)$.

Next, we prove a similar result as in [26, Lemma 2.2].

Proposition 1 *Let $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ and $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\mathbf{C}})$ be two stabilizable and detectable LTI systems with $\sigma(\mathbf{A}) \cap i\mathbb{R} = \emptyset = \sigma(\widehat{\mathbf{A}}) \cap i\mathbb{R}$. Let further $\mathbf{G}_F = [\mathbf{M}, \mathbf{N}]$, $\widehat{\mathbf{G}}_F = [\widehat{\mathbf{M}}, \widehat{\mathbf{N}}]$ be left-coprime factorizations with $\mathbf{F} = \mathbf{P}\mathbf{C}^T$ and $\widehat{\mathbf{F}} = \widehat{\mathbf{P}}\widehat{\mathbf{C}}^T$. Then,*

$$\|\mathbf{G} - \widehat{\mathbf{G}}\|_{\mathcal{L}_2} \leq \|\widehat{\mathbf{M}}^{-1}\|_{\mathcal{L}_\infty} (\|\mathbf{G}\|_{\mathcal{L}_\infty} \|\mathbf{M} - \widehat{\mathbf{M}}\|_{\mathcal{H}_2} + \|\mathbf{N} - \widehat{\mathbf{N}}\|_{\mathcal{H}_2}), \quad (14)$$

and, thus

$$\|\mathbf{G} - \widehat{\mathbf{G}}\|_{\mathcal{L}_2} \leq \|\widehat{\mathbf{M}}^{-1}\|_{\mathcal{L}_\infty} (1 + \|\mathbf{G}\|_{\mathcal{L}_\infty}) \|\mathbf{G} - \widehat{\mathbf{G}}\|_{\mathcal{H}_2\text{-gap}}. \quad (15)$$

Proof By using the left-coprime factorizations of \mathbf{G} and $\widehat{\mathbf{G}}$, we obtain

$$\begin{aligned} \mathbf{G} - \widehat{\mathbf{G}} &= \mathbf{G} - \widehat{\mathbf{M}}^{-1}\widehat{\mathbf{N}} = \widehat{\mathbf{M}}^{-1}(\widehat{\mathbf{M}}\mathbf{G} - \widehat{\mathbf{N}}) \\ &= \widehat{\mathbf{M}}^{-1}(\widehat{\mathbf{M}}\mathbf{G} - \mathbf{M}\mathbf{M}^{-1}\mathbf{N} + \mathbf{N} - \widehat{\mathbf{N}}) = \widehat{\mathbf{M}}^{-1}((\widehat{\mathbf{M}} - \mathbf{M})\mathbf{G} + (\mathbf{N} - \widehat{\mathbf{N}})). \end{aligned}$$

Since \mathbf{A} and $\widehat{\mathbf{A}}$ are assumed to have no purely imaginary eigenvalues, we know that $\mathbf{G} \in \mathcal{L}_\infty(i\mathbb{R})$ and $\widehat{\mathbf{M}}^{-1}(\cdot) = \mathbf{I} + \widehat{\mathbf{C}}(\cdot - \widehat{\mathbf{A}})^{-1}\widehat{\mathbf{F}} \in \mathcal{L}_\infty(i\mathbb{R})$. The assertion (14) then follows from the fact that $\|\mathbf{G}\mathbf{H}\|_{\mathcal{L}_2} \leq \|\mathbf{G}\|_{\mathcal{L}_\infty}\|\mathbf{H}\|_{\mathcal{L}_2}$ for all $\mathbf{G} \in \mathcal{L}_\infty$ and $\mathbf{H} \in \mathcal{L}_2$. In particular, note that $\mathbf{M} - \widehat{\mathbf{M}} \in \mathcal{H}_2$ and $\mathbf{N} - \widehat{\mathbf{N}} \in \mathcal{H}_2$, which implies $\|\mathbf{M} - \widehat{\mathbf{M}}\|_{\mathcal{L}_2} = \|\mathbf{M} - \widehat{\mathbf{M}}\|_{\mathcal{H}_2}$, $\|\mathbf{N} - \widehat{\mathbf{N}}\|_{\mathcal{L}_2} = \|\mathbf{N} - \widehat{\mathbf{N}}\|_{\mathcal{H}_2}$. Finally, the assertion (15) directly follows from (14) and the definition of the \mathcal{H}_2 -gap in (13).

Remark 1 Note that if we split $\mathbf{G} = \mathbf{G}_s + \mathbf{G}_u$ and $\widehat{\mathbf{G}} = \widehat{\mathbf{G}}_s + \widehat{\mathbf{G}}_u$ into their stable and anti-stable parts and use the orthogonal decomposition $\mathcal{L}_2(i\mathbb{R}) = \mathcal{H}_2(\mathbb{C}_-) \oplus \mathcal{H}_2(\mathbb{C}_+)$, we also obtain bounds for $\|\mathbf{G}_s - \widehat{\mathbf{G}}_s\|_{\mathcal{H}_2(\mathbb{C}_+)}$ and $\|\mathbf{G}_u - \widehat{\mathbf{G}}_u\|_{\mathcal{H}_2(\mathbb{C}_-)}$:

$$\begin{aligned} &\max(\|\mathbf{G}_s - \widehat{\mathbf{G}}_s\|_{\mathcal{H}_2(\mathbb{C}_+)}, \|\mathbf{G}_u - \widehat{\mathbf{G}}_u\|_{\mathcal{H}_2(\mathbb{C}_-)}) \\ &\leq \|\widehat{\mathbf{M}}^{-1}\|_{\mathcal{L}_\infty} (1 + \|\mathbf{G}\|_{\mathcal{L}_\infty}) \|\mathbf{G} - \widehat{\mathbf{G}}\|_{\mathcal{H}_2\text{-gap}}. \end{aligned}$$

Proposition 1 bounds the \mathcal{L}_2 distance between the full model $\mathbf{G}(s)$ and the reduced model $\widehat{\mathbf{G}}(s)$ with the \mathcal{H}_2 -gap distance between full closed-loop system $\mathbf{G}_F(s)$ and the reduced one $\widehat{\mathbf{G}}_F(s)$. This immediately motivates a model reduction approach in which one tries to minimize the \mathcal{H}_2 -gap. This is what we investigate next.

3 \mathcal{H}_2 -gap Model Reduction

In this section, we analyze the \mathcal{H}_2 -gap in more detail. We begin with the derivation of a pole-residue formula that extends the one discussed for the standard case in, e.g., [17]. Subsequently, we discuss the individual error terms from a rational interpolation-based perspective which suggests the use of an iterative algorithm generalizing IRKA.

3.1 Pole-Residue Formulae for the \mathcal{H}_2 -gap Measure

Recall the state-space representation of $\mathbf{G}_F(s)$:

$$\mathbf{G}_F(s) = [\mathbf{I}, 0] + \mathbf{C}(s\mathbf{I} - \mathbf{A}_F)^{-1}[-\mathbf{F}, \mathbf{B}] \quad (10).$$

Let \mathbf{w}_i denote the left eigenvector of \mathbf{A}_F associated with the eigenvalue λ_i , i.e., $\mathbf{w}_i^T \mathbf{A}_F = \lambda_i \mathbf{w}_i^T$.¹ For simplicity of presentation, assume the poles λ_i are semi-simple and write $\mathbf{W}^T \mathbf{A}_F = \Lambda \mathbf{W}^T$ where $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n] \in \mathbb{C}^{n \times n}$ and $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n) \in \mathbb{C}^{n \times n}$. Define $\mathbf{V} = \mathbf{W}^{-T}$. Then, a state-space transformation by \mathbf{W}^T in (10) yields the pole-residue representation of $\mathbf{G}_F(s)$:

$$\mathbf{G}_F(s) = [\mathbf{I}, 0] + \sum_{i=1}^n \frac{\mathbf{c}_i [\mathbf{f}_i^T, \mathbf{b}_i^T]}{s - \lambda_i}, \tag{16}$$

where

$$\mathbf{b}_i = \mathbf{B}^T \mathbf{w}_i, \quad \mathbf{f}_i = -\mathbf{F}^T \mathbf{w}_i = -\mathbf{C} \mathbf{P} \mathbf{w}_i, \quad \text{and} \quad \mathbf{c}_i = \mathbf{C} \mathbf{v}_i. \tag{17}$$

The same line of arguments yield the pole-residue representation of $\widehat{\mathbf{G}}_F$: Let $\widehat{\mathbf{w}}_j$ be the left eigenvector of $\widehat{\mathbf{A}}_F$ associated with the eigenvalue $\widehat{\lambda}_j$. Assume the poles $\widehat{\lambda}_j$ are semi-simple, and define $\widehat{\mathbf{W}} = [\widehat{\mathbf{w}}_1, \widehat{\mathbf{w}}_2, \dots, \widehat{\mathbf{w}}_r] \in \mathbb{C}^{r \times r}$ and $\widehat{\mathbf{V}} = \widehat{\mathbf{W}}^{-T}$. Then,

$$\widehat{\mathbf{G}}_F(s) = [\mathbf{I}, 0] + \sum_{j=1}^r \frac{\widehat{\mathbf{c}}_j [\widehat{\mathbf{f}}_j^T, \widehat{\mathbf{b}}_j^T]}{s - \widehat{\lambda}_j}, \tag{18}$$

where

$$\widehat{\mathbf{b}}_j = \widehat{\mathbf{B}}^T \widehat{\mathbf{w}}_j, \quad \widehat{\mathbf{f}}_j = -\widehat{\mathbf{F}}^T \widehat{\mathbf{w}}_j = -\widehat{\mathbf{C}} \widehat{\mathbf{P}} \widehat{\mathbf{w}}_j, \quad \text{and} \quad \widehat{\mathbf{c}}_j = \widehat{\mathbf{C}} \widehat{\mathbf{v}}_j. \tag{19}$$

Now, using the representations (16) and (18), we extend the pole-residue based formula for the \mathcal{H}_2 -norm to the \mathcal{H}_2 -gap norm.

Proposition 2 *Let \mathbf{G} and $\widehat{\mathbf{G}}$ be stabilizable detectable with coprime factorizations in (6) and (8). Further, let \mathbf{G}_F and $\widehat{\mathbf{G}}_F$ have the pole-residue representations in (16) and (18). Then,*

$$\begin{aligned} \|\mathbf{G} - \widehat{\mathbf{G}}\|_{\mathcal{H}_{2\text{-gap}}}^2 &= \sum_{i=1}^n \mathbf{c}_i^T (\mathbf{G}_F(-\lambda_i) - \widehat{\mathbf{G}}_F(-\lambda_i)) \begin{bmatrix} \mathbf{f}_i \\ \mathbf{b}_i \end{bmatrix} \\ &\quad + \sum_{j=1}^r \widehat{\mathbf{c}}_j^T (\widehat{\mathbf{G}}_F(-\widehat{\lambda}_j) - \mathbf{G}_F(-\widehat{\lambda}_j)) \begin{bmatrix} \widehat{\mathbf{f}}_j \\ \widehat{\mathbf{b}}_j \end{bmatrix}. \end{aligned} \tag{20}$$

Proof First, we rewrite the error as

$$\|\mathbf{G} - \widehat{\mathbf{G}}\|_{\mathcal{H}_{2\text{-gap}}}^2 = \|\mathbf{G}_F - \widehat{\mathbf{G}}_F\|_{\mathcal{H}_2}^2 = \|(\mathbf{G}_F - [\mathbf{I}, 0]) + ([\mathbf{I}, 0] - \widehat{\mathbf{G}}_F)\|_{\mathcal{H}_2}^2.$$

Note that $(\mathbf{G}_F - [\mathbf{I}, 0]) \in \mathcal{H}_2$ and $([\mathbf{I}, 0] - \widehat{\mathbf{G}}_F) \in \mathcal{H}_2$. For a transfer function $\mathbf{H} \in \mathcal{H}_2$ with pole-residue representation $\mathbf{H}(s) = \sum_{k=1}^n \frac{\mathbf{h}_k \mathbf{g}_k^T}{s - \mu_k}$, the \mathcal{H}_2 norm satisfies

¹ We define the left eigenpair via the relationship $\mathbf{w}_i^T \mathbf{A}_F = \lambda_i \mathbf{w}_i^T$ as opposed to the more usual definition: $\mathbf{w}_i^* \mathbf{A}_F = \lambda_i \mathbf{w}_i^*$. Even though \mathbf{w}_i is potentially a complex vector, the version we adopt eliminates the need to use $\bar{\mathbf{w}}_i$ in many definitions and equations that follow.

$\|\mathbf{H}\|_{\mathcal{H}_2}^2 = \sum_{k=1}^n \mathbf{h}_k^T \mathbf{H}(-\mu_k) \mathbf{g}_k$; see, e.g., [17, Lemma 2.4] for the SISO version and [2, Lemma 1.1] for the MIMO version. Then, the result (20) follows from applying this \mathcal{H}_2 norm formula to the pole-residue representation of $\mathbf{G}_F - \widehat{\mathbf{G}}_F$, which can be obtained from (16) and (18) by eliminating the leading (constant) terms.

3.2 \mathcal{H}_2 -gap Formula and Interpolation.

Proposition 2 reveals two components that contribute to the \mathcal{H}_2 -gap in a way that is similar to the standard \mathcal{H}_2 -error measure. The first component is due to the mismatch of the transfer functions \mathbf{G}_F and $\widehat{\mathbf{G}}_F$ at the mirror images of the *full-order closed-loop poles*, λ_i , and the second component reflects the mismatch at the mirror images of the *reduced-order closed-loop poles*, $\widehat{\lambda}_i$. In order to reduce the \mathcal{H}_2 -gap, a reasonable approach might be to eliminate terms from these two components. For example, if one enforces $(\mathbf{G}_F(-\lambda_i) - \widehat{\mathbf{G}}_F(-\lambda_i)) \begin{bmatrix} \mathbf{f}_i \\ \mathbf{b}_i \end{bmatrix} = 0$, then the i th term from the first term will be eliminated. This condition is referred to as *right-tangential interpolation*; more specifically, we state $\widehat{\mathbf{G}}_F(s)$ tangentially interpolates $\mathbf{G}_F(s)$ at the interpolation point $-\lambda_i$ along the right-tangential direction $\begin{bmatrix} \mathbf{f}_i \\ \mathbf{b}_i \end{bmatrix}$; see [2, 3] for further details. We can eliminate the i th term in the first sum by enforcing $\mathbf{c}_i^T (\mathbf{G}_F(-\lambda_i) - \widehat{\mathbf{G}}_F(-\lambda_i)) = 0$ as well. This is referred to as *left-tangential interpolation*. The terms in the second sum can be similarly eliminated. This interpretation of the \mathcal{H}_2 -error norm and elimination of the error terms via interpolation have been proposed in the regular \mathcal{H}_2 -error measure [15]. Since the second-term depends on the reduced-model to-be-computed and are not known a priori, [15] proposed eliminating the dominant terms from the first term. Even though this is not an optimal reduction strategy, this approach has worked well in various examples. The situation is rather different here.

In order to minimize the \mathcal{H}_2 -gap, we construct a reduced model $\widehat{\mathbf{G}}$ from \mathbf{G} . Yet Proposition 2 shows that the error depends on \mathbf{G}_F and $\widehat{\mathbf{G}}_F$, which we do not have direct access to. Consider the \mathcal{H}_2 -gap once again: $\|\mathbf{G} - \widehat{\mathbf{G}}\|_{\mathcal{H}_2\text{-gap}} = \|\mathbf{G}_F - \widehat{\mathbf{G}}_F\|_{\mathcal{H}_2}$. In order to minimize the \mathcal{H}_2 -gap, suppose we perform an optimal \mathcal{H}_2 reduction on \mathbf{G}_F , and let $\mathbf{V} \in \mathbb{R}^{n \times r}$ and $\mathbf{W} \in \mathbb{R}^{n \times r}$ be the corresponding optimal model reduction bases with $\mathbf{W}^T \mathbf{V} = \mathbf{I}$. The state-space representation for the reduced $\widehat{\mathbf{G}}$ is given by

$$\begin{aligned} \widehat{\mathbf{G}}_F(s) &= [\mathbf{I}, 0] + \widehat{\mathbf{C}}(s\mathbf{I} - \widehat{\mathbf{A}}_F)^{-1}[-\widehat{\mathbf{F}}, \widehat{\mathbf{B}}] \\ &= [\mathbf{I}, 0] + \mathbf{C}\mathbf{V}(s\mathbf{I} - \mathbf{W}^T \mathbf{A}\mathbf{V} - \mathbf{W}^T \mathbf{F}\mathbf{C}\mathbf{V})^{-1}[-\mathbf{W}^T \mathbf{F}, \mathbf{W}^T \mathbf{B}]. \end{aligned}$$

We need to extract the corresponding reduced model $\widehat{\mathbf{G}}$ from this reduced closed-loop model $\widehat{\mathbf{G}}_F(s)$. One might reasonably assume that $\widehat{\mathbf{C}} = \mathbf{C}\mathbf{V}$, $\widehat{\mathbf{B}} = \mathbf{W}^T \mathbf{B}$, $\widehat{\mathbf{F}} = \mathbf{W}^T \mathbf{F}$, and $\widehat{\mathbf{A}} = \mathbf{W}^T \mathbf{A}\mathbf{V}$. However, these reduced quantities need also to satisfy $\widehat{\mathbf{F}} = \mathbf{W}^T \mathbf{F} = \mathbf{W}^T \mathbf{P}\mathbf{C} = \widehat{\mathbf{P}}\widehat{\mathbf{C}}^T$ where $\widehat{\mathbf{P}}$ solves $\widehat{\mathbf{A}}\widehat{\mathbf{P}} + \widehat{\mathbf{P}}\widehat{\mathbf{A}}^T - \widehat{\mathbf{P}}\widehat{\mathbf{C}}^T \widehat{\mathbf{C}}\widehat{\mathbf{P}} + \widehat{\mathbf{B}}\widehat{\mathbf{B}}^T = 0$. Clearly, this is not true in general and we cannot expect to extract a reduced system $\widehat{\mathbf{G}}$ that

would have created the reduced closed-loop model $\widehat{\mathbf{G}}_{\mathbf{F}}(s)$. A similar issue arises in the weighted \mathcal{H}_2 model reduction problem, where, given a weighting functions $\mathbf{W}_o(s)$, one tries to minimize the weighted error $\|\mathbf{W}_o(\mathbf{G} - \widehat{\mathbf{G}})\|_{\mathcal{H}_2}$. As [1, 10] prove, in the weighted- \mathcal{H}_2 problem, the error (and optimality) requires that a function of \mathbf{G} interpolates a function of $\widehat{\mathbf{G}}$, leading to the same issue that we encounter here.

To address this issue at least partially, Lemmas 1–2 enable us to rewrite the \mathcal{H}_2 -gap formula in (20) as a function of \mathbf{G} and $\widehat{\mathbf{G}}$. The result in Theorem 1 will, then, form the foundation of the proposed method outlined in Algorithm 1.

Lemma 1 *Let \mathbf{G} be a stabilizable and detectable linear system with coprime factorization in (6). Further, let $\mathbf{G}_{\mathbf{F}} = [\mathbf{M} \ \mathbf{N}]$ have the pole-residue representation $\mathbf{G}_{\mathbf{F}}(s) = [\mathbf{I}, 0] + \sum_{i=1}^n \frac{\mathbf{c}_i[\mathbf{f}_i^T, \mathbf{b}_i^T]}{s-\lambda_i}$ as in (16). Assume that $\sigma(\mathbf{A}) \cap \sigma(-\mathbf{A}_{\mathbf{F}}) = \emptyset$. Then,*

$$\mathbf{G}(-\lambda_i)\mathbf{b}_i = -\mathbf{f}_i \quad \text{and} \quad \mathbf{G}_{\mathbf{F}}(-\lambda_i) \begin{bmatrix} \mathbf{f}_i \\ \mathbf{b}_i \end{bmatrix} = 0, \quad \text{for } i = 1, \dots, n. \quad (21)$$

Proof Let \mathbf{w}_i denote the left eigenvector of $\mathbf{A}_{\mathbf{F}}$ associated with the eigenvalue λ_i , i.e., $\mathbf{w}_i^T \mathbf{A}_{\mathbf{F}} = \lambda_i \mathbf{w}_i^T$. Definition of \mathbf{b}_i in (17) yields

$$\mathbf{G}(-\lambda_i)\mathbf{b}_i = \mathbf{C}(-\lambda_i \mathbf{I} - \mathbf{A})^{-1} \mathbf{B} \mathbf{B}^T \mathbf{w}_i. \quad (22)$$

Using the fact that $\mathbf{A}_{\mathbf{F}} = \mathbf{A} - \mathbf{P} \mathbf{C}^T \mathbf{C}$, the Riccati equation (7) can be rewritten as the Sylvester equation $\mathbf{A} \mathbf{P} + \mathbf{P} \mathbf{A}_{\mathbf{F}}^T + \mathbf{B} \mathbf{B}^T = 0$. Postmultiplication with \mathbf{w}_i then yields

$$\mathbf{P} \mathbf{w}_i = (-\lambda_i \mathbf{I} - \mathbf{A})^{-1} \mathbf{B} \mathbf{B}^T \mathbf{w}_i.$$

Inserting this last expression into (22) leads to

$$\mathbf{G}(-\lambda_i)\mathbf{b}_i = \mathbf{C} \mathbf{P} \mathbf{w}_i = -\mathbf{f}_i,$$

which proves the first assertion in (21). To prove the second assertion, we compute

$$\begin{aligned} \widehat{\mathbf{G}}_{\mathbf{F}}(-\lambda_i) \begin{bmatrix} \mathbf{f}_i \\ \mathbf{b}_i \end{bmatrix} &= [\mathbf{M}(-\lambda_i) \ \mathbf{N}(-\lambda_i)] \begin{bmatrix} \mathbf{f}_i \\ \mathbf{b}_i \end{bmatrix} = \mathbf{M}(-\lambda_i)\mathbf{f}_i + \mathbf{N}(-\lambda_i)\mathbf{b}_i \\ &= \mathbf{M}(-\lambda_i)(\mathbf{f}_i + [\mathbf{M}(-\lambda_i)]^{-1} \mathbf{N}(-\lambda_i)\mathbf{b}_i) \\ &= \mathbf{M}(-\lambda_i)(\mathbf{f}_i + \mathbf{G}(-\lambda_i)\mathbf{b}_i) = 0, \end{aligned}$$

where, in the last step we used the just-proven first assertion in (21).

Lemma 2 *Let \mathbf{G} be a stabilizable and detectable linear system with the closed-loop transfer function $\mathbf{G}_{\mathbf{F}}(s) = [\mathbf{M} \ \mathbf{N}] = [\mathbf{I}, 0] + \sum_{i=1}^n \frac{\mathbf{c}_i[\mathbf{f}_i^T, \mathbf{b}_i^T]}{s-\lambda_i}$ as in (16). Let $\widehat{\mathbf{G}}_{\mathbf{F}} = [\widehat{\mathbf{M}} \ \widehat{\mathbf{N}}]$ denote the closed-loop transfer function of a stabilizable and detectable system reduced model $\widehat{\mathbf{G}}$. If $\sigma(\mathbf{A}) \cap \sigma(-\mathbf{A}_{\mathbf{F}}) = \emptyset$, then*

$$\widehat{\mathbf{G}}_{\mathbf{F}}(-\lambda_i) \begin{bmatrix} \mathbf{f}_i \\ \mathbf{b}_i \end{bmatrix} = -\widehat{\mathbf{M}}(-\lambda_i) (\mathbf{G}(-\lambda_i) - \widehat{\mathbf{G}}(-\lambda_i)) \mathbf{b}_i, \text{ for } i = 1, \dots, n. \quad (23)$$

Proof Using $\widehat{\mathbf{G}}(s) = [\widehat{\mathbf{M}}(s)]^{-1} \widehat{\mathbf{N}}(s)$, we evaluate

$$\begin{aligned} \widehat{\mathbf{G}}_{\mathbf{F}}(-\lambda_i) \begin{bmatrix} \mathbf{f}_i \\ \mathbf{b}_i \end{bmatrix} &= \widehat{\mathbf{M}}(-\lambda_i) \mathbf{f}_i + \widehat{\mathbf{N}}(-\lambda_i) \mathbf{b}_i = \widehat{\mathbf{M}}(-\lambda_i) (\mathbf{f}_i + [\widehat{\mathbf{M}}(-\lambda_i)]^{-1} \widehat{\mathbf{N}}(-\lambda_i) \mathbf{b}_i) \\ &= \widehat{\mathbf{M}}(-\lambda_i) (-\mathbf{G}(-\lambda_i) \mathbf{b}_i + \widehat{\mathbf{G}}(-\lambda_i) \mathbf{b}_i), \end{aligned}$$

where in the last step we used the first assertion in (21).

As a direct consequence of Proposition 2, and Lemmas 1 and 2, we obtain an alternative representation of the \mathcal{H}_2 -gap error, one of our main results.

Theorem 1 *Let \mathbf{G} and $\widehat{\mathbf{G}}$ be stabilizable and detectable with coprime factorizations in (6) and (8). Further, let $\mathbf{G}_{\mathbf{F}}$ and $\widehat{\mathbf{G}}_{\mathbf{F}}$ have the pole-residue representations in (16) and (18). Then,*

$$\begin{aligned} \|\mathbf{G} - \widehat{\mathbf{G}}\|_{\mathcal{H}_2\text{-gap}}^2 &= \sum_{i=1}^n \mathbf{c}_i^T \widehat{\mathbf{M}}(-\lambda_i) (\mathbf{G}(-\lambda_i) - \widehat{\mathbf{G}}(-\lambda_i)) \mathbf{b}_i \\ &\quad + \sum_{j=1}^r \widehat{\mathbf{c}}_j^T \mathbf{M}(-\widehat{\lambda}_j) (\widehat{\mathbf{G}}(-\widehat{\lambda}_j) - \mathbf{G}(-\widehat{\lambda}_j)) \widehat{\mathbf{b}}_j. \end{aligned} \quad (24)$$

Proof Consider the first sum (20), i.e., $\sum_{i=1}^n \mathbf{c}_i^T (\mathbf{G}_{\mathbf{F}}(-\lambda_i) - \widehat{\mathbf{G}}_{\mathbf{F}}(-\lambda_i)) \begin{bmatrix} \mathbf{f}_i \\ \mathbf{b}_i \end{bmatrix}$. First using the second assertion of Lemma 1 and then using Lemma 2 yield

$$\sum_{i=1}^n \mathbf{c}_i^T (\mathbf{G}_{\mathbf{F}}(-\lambda_i) - \widehat{\mathbf{G}}_{\mathbf{F}}(-\lambda_i)) \begin{bmatrix} \mathbf{f}_i \\ \mathbf{b}_i \end{bmatrix} = -\sum_{i=1}^n \mathbf{c}_i^T \widehat{\mathbf{G}}_{\mathbf{F}}(-\lambda_i) \begin{bmatrix} \mathbf{f}_i \\ \mathbf{b}_i \end{bmatrix} = \sum_{i=1}^n \mathbf{c}_i^T \widehat{\mathbf{M}}(-\lambda_i) (\mathbf{G}(-\lambda_i) - \widehat{\mathbf{G}}(-\lambda_i)) \mathbf{b}_i,$$

which is the first sum, indexed by i , in (24). The second part of (24) follows similarly by interchanging the roles of $\mathbf{G}_{\mathbf{F}}$ and $\widehat{\mathbf{G}}_{\mathbf{F}}$ in Lemmas 1 and 2. Theorem 1 achieves what we were seeking to accomplish: representation of the \mathcal{H}_2 -gap in terms of the model to be reduced, \mathbf{G} , and the reduced model itself, $\widehat{\mathbf{G}}$. Now, we can reduce \mathbf{G} , e.g., via interpolatory projection-based methods, to construct the reduced model $\widehat{\mathbf{G}}$ that tangentially interpolates \mathbf{G} and then we eliminate the selected terms from the error formula in (24). We will make this interpolation aspect more concrete next.

The following result is a direct consequence of Lemmas 1 and 2.

Corollary 1 *Assume the same set-up in Lemmas 1 and 2. Then, for $i = 1, \dots, n$,*

$$\text{If } \widehat{\mathbf{G}}(-\lambda_i) \mathbf{b}_i = \mathbf{G}(-\lambda_i) \mathbf{b}_i, \text{ then } \widehat{\mathbf{G}}_{\mathbf{F}}(-\lambda_i) \begin{bmatrix} \mathbf{f}_i \\ \mathbf{b}_i \end{bmatrix} = \mathbf{G}_{\mathbf{F}}(-\lambda_i) \begin{bmatrix} \mathbf{f}_i \\ \mathbf{b}_i \end{bmatrix} = 0. \quad (25)$$

Moreover, for $j = 1, \dots, n$,

$$\text{If } \widehat{\mathbf{G}}(-\widehat{\lambda}_j)\widehat{\mathbf{b}}_j = \mathbf{G}(-\widehat{\lambda}_j)\widehat{\mathbf{b}}_j, \text{ then } \widehat{\mathbf{G}}_{\mathbf{F}}(-\widehat{\lambda}_j) \begin{bmatrix} \widehat{\mathbf{f}}_j \\ \widehat{\mathbf{b}}_j \end{bmatrix} = \mathbf{G}_{\mathbf{F}}(-\widehat{\lambda}_j) \begin{bmatrix} \widehat{\mathbf{f}}_j \\ \widehat{\mathbf{b}}_j \end{bmatrix} = 0. \quad (26)$$

Corollary 1 first reveals that we can enforce the closed-loop systems $\mathbf{G}_{\mathbf{F}}$ and $\widehat{\mathbf{G}}_{\mathbf{F}}$ to tangentially interpolate each other by forcing interpolation of \mathbf{G} and $\widehat{\mathbf{G}}$. The resulting interpolation is occurring at specially adapted points, namely at the mirror images of the full- or reduced-order closed-loop poles. Moreover, the interpolated value is zero. It is worth mentioning that reduced-order closed-loop poles have also been studied in the context of rational Krylov subspace methods for solving the algebraic Riccati equation in [22]. In that work, the authors showed that the rational Krylov subspace method coincides with a subspace iteration if the shifts are chosen as the mirrored reduced-order closed-loop poles.

Corollary 1 also reveals that we can enforce interpolation of the closed-loop model $\mathbf{G}_{\mathbf{F}}$ *without ever constructing* $\mathbf{G}_{\mathbf{F}}$, i.e., without needing to solve the (large-scale) Riccati equation (7) to compute \mathbf{P} . This will have substantial numerical advantages in the large-scale settings, because unlike most methods used for model reduction using the gap measure, one does not need to solve for the Gramian \mathbf{P} . Next, we will discuss the numerical framework to enforce these desired interpolation conditions.

3.3 Model Reduction with Respect to the \mathcal{H}_2 -gap Measure

We review briefly the projection-based tangential interpolation framework. For details, we refer the reader to, e.g., [2, 3, 13, 18]. Let $\mathbf{G}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$ denote the transfer function of the full-model with m -inputs and p -outputs. Suppose the left-interpolation points $\{\mu_1, \mu_2, \dots, \mu_r\} \in \mathbb{C}$ are chosen together with non-trivial left-directions $\{\ell_1, \ell_2, \dots, \ell_r\} \in \mathbb{C}^p$. Also suppose the right-interpolation points $\{\sigma_1, \sigma_2, \dots, \sigma_r\} \in \mathbb{C}$ are chosen together with non-trivial right-directions $\{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_r\} \in \mathbb{C}^m$. Construct model reduction bases $\mathbf{V}_r \in \mathbb{C}^{n \times r}$ and $\mathbf{W}_r \in \mathbb{C}^{n \times r}$:

$$\begin{aligned} \mathbf{V}_r &= [(\sigma_1\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{r}_1, (\sigma_2\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{r}_2, \dots, (\sigma_r\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{r}_r] \text{ and} \\ \mathbf{W}_r &= [(\mu_1\mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{C}^T\ell_1, (\mu_2\mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{C}^T\ell_2, \dots, (\mu_r\mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{C}^T\ell_r]. \end{aligned}$$

Assume, without loss of generality, that a basis transformation is performed so that $\mathbf{W}_r^T\mathbf{V}_r = \mathbf{I}_r$. Construct the reduced model $\widehat{\mathbf{G}}(s) = \widehat{\mathbf{C}}(s\mathbf{I} - \widehat{\mathbf{A}})^{-1}\widehat{\mathbf{B}}$ via Petrov-Galerkin projection, i.e.,

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

Then, the reduced model $\widehat{\mathbf{G}}$ tangentially interpolates \mathbf{G} in the sense that

$$\mathbf{G}(\sigma_j)\mathbf{r}_j = \widehat{\mathbf{G}}(\sigma_j)\mathbf{r}_j, \quad \ell_j^T\mathbf{G}(\mu_j) = \ell_j^T\widehat{\mathbf{G}}(\mu_j), \quad \ell_j^T\mathbf{G}(\sigma_j)\mathbf{r}_j = \ell_j^T\widehat{\mathbf{G}}(\sigma_j)\mathbf{r}_j,$$

for $j = 1, 2, \dots, r$. Moreover, if $\sigma_k = \mu_k$, then one additionally satisfies a tangential Hermite interpolation, namely $\ell_k^T \mathbf{G}'(\sigma_k) \mathbf{r}_k = \ell_k^T \widehat{\mathbf{G}}'(\sigma_k) \mathbf{r}_k$ where “ $'$ ” denotes the derivative with respect to s . Therefore, if the interpolation points and directions are specified, then constructing a reduced interpolatory transfer function can be easily constructed as described, with the main cost of solving the shifted linear systems in computing \mathbf{V} and \mathbf{W} . Then, the natural question to ask is how to choose the interpolation points and directions to minimize an error measure. This question has been answered using the regular \mathcal{H}_2 error measure. Let $\widehat{\mathbf{G}}(s) = \sum_{j=1}^r \frac{\ell_j \mathbf{r}_j^T}{s + \sigma_j}$ be the

pole-residue decomposition. If $\widehat{\mathbf{G}}(s)$ is the \mathcal{H}_2 -optimal approximation to $\mathbf{G}(s)$ in the \mathcal{H}_2 norm, then, $\mathbf{G}(\sigma_j) \mathbf{r}_j = \widehat{\mathbf{G}}(\sigma_j) \mathbf{r}_j$, $\ell_j^T \mathbf{G}(\sigma_j) = \ell_j^T \widehat{\mathbf{G}}(\sigma_j)$, and $\ell_j^T \mathbf{G}'(\sigma_j) \mathbf{r}_j = \ell_j^T \widehat{\mathbf{G}}'(\sigma_j) \mathbf{r}_j$, for $j = 1, 2, \dots, r$. Therefore, tangential Hermite interpolation is a necessary condition for \mathcal{H}_2 optimality. Note that optimal interpolation points are $\{\sigma_j\}$, the mirror images of the poles of the reduced model $\widehat{\mathbf{G}}(s)$, and the optimal tangential directions are based on the residues of $\widehat{\mathbf{G}}(s)$; neither known a priori. The Iterative Rational Krylov Algorithm (IRKA) of [17] and its variants [6, 7, 11, 30] resolve this issue by iteratively correcting the interpolation points and directions until the desired optimality conditions are met. For details, we refer the reader to [2, 3, 17, 18] and the references therein.

The situation is similar in the \mathcal{H}_2 -gap problem we consider here. First, leave the question of optimality aside and focus on reasonable/well-informed interpolation points and directions selection. Recall the \mathcal{H}_2 -gap error formula in (24). We can eliminate the i th term in the first sum by choosing $\sigma_i = -\lambda_i$ as an interpolation point and $\mathbf{r}_i = \mathbf{b}_i$ as an interpolation direction. Clearly, the first sum has n components and one can only eliminate r conditions from there using r interpolation points. One can choose the poles λ_i with *dominant* residue terms $\mathbf{c}_i \mathbf{b}_i^T$, for example. However, this requires that we compute the full-order closed-loop poles λ_i by solving for the Gramian \mathbf{P} . Also, as discussed above, the regular \mathcal{H}_2 minimization via interpolation reveals that the optimal interpolation points are determined by the reduced-order poles, not the full-order ones.

To eliminate the j th term from the second sum in (24), we can enforce $\widehat{\mathbf{G}}(-\widehat{\lambda}_j) \widehat{\mathbf{b}}_j = \mathbf{G}(-\widehat{\lambda}_j) \widehat{\mathbf{b}}_j$. This puts us in the framework of the regular \mathcal{H}_2 -optimality problem. The interpolation points $\sigma_j = -\widehat{\lambda}_j$ and the tangential directions $\mathbf{r}_j = \widehat{\mathbf{b}}_j$, for $j = 1, 2, \dots, r$, depend on the reduced-model $\widehat{\mathbf{G}}$ (or more precisely $\widehat{\mathbf{G}}_{\mathbf{F}}$) we want to compute; note that the interpolation data is not known a priori. As in IRKA, this requires an iterative algorithm that adaptively corrects the interpolation data. The major advantage compared to eliminating terms from the first sum in (24) is that this adaptive correction process does not require computing full-order closed-loop poles, i.e., computing \mathbf{P} . Yet, as Corollary 1 illustrates, we are still able to interpolate the full-order closed-loop model.

Algorithm 1 gives a sketch of the proposed numerical scheme. Starting with an initial selection of interpolation data, the algorithm computes an interpolatory reduced model $\widehat{\mathbf{G}}$ (Lines 2–4) and then computes the pole-residue representation of the reduced-order closed-loop model $\widehat{\mathbf{G}}$ (Lines 5–6). Note that these compu-

tations are trivial since it is performed at the reduced-order dimension. Line 7 updates the interpolation data so that upon convergence of Algorithm 1, we have $\sigma_j = -\hat{\lambda}_j$ and $\mathbf{r}_j = \hat{\mathbf{b}}_j$, for $j = 1, 2, \dots, r$, as we wanted to accomplish. Upon convergence of Algorithm 1, we enforce $\hat{\mathbf{G}}(-\hat{\lambda}_j)\hat{\mathbf{b}}_j = \mathbf{G}(-\hat{\lambda}_j)\hat{\mathbf{b}}_j$ and the second sum in (24) is completely eliminated; thus leading to the eventual \mathcal{H}_2 -gap error

$$\|\mathbf{G} - \hat{\mathbf{G}}\|_{\mathcal{H}_2\text{-gap}}^2 = \sum_{i=1}^n \mathbf{c}_i^T \hat{\mathbf{M}}(-\lambda_i) (\mathbf{G}(-\lambda_i) - \hat{\mathbf{G}}(-\lambda_i)) \mathbf{b}_i.$$

Algorithm 1 gap-IRKA

Input: $\{\sigma_1, \dots, \sigma_r\}$, $\{\mathbf{r}_1, \dots, \mathbf{r}_r\}$ and $\{\ell_1, \dots, \ell_r\}$.

Output: $\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}$

1: **while** relative change in $\{\sigma_j\} > \text{tol}$ **do**

2: Compute \mathbf{V}_r and \mathbf{W}_r from

$$\begin{aligned} \mathbf{V}_r &= [(\sigma_1 \mathbf{I} - \mathbf{A})^{-1} \mathbf{B} \mathbf{r}_1, \dots, (\sigma_r \mathbf{I} - \mathbf{A})^{-1} \mathbf{B} \mathbf{r}_r], \\ \mathbf{W}_r &= [(\sigma_1 \mathbf{I} - \mathbf{A}^T)^{-1} \mathbf{C}^T \ell_1, \dots, (\sigma_r \mathbf{I} - \mathbf{A}^T)^{-1} \mathbf{C}^T \ell_r]. \end{aligned}$$

3: Perform basis change $\mathbf{W}_r \leftarrow \mathbf{W}_r (\mathbf{V}_r^T \mathbf{W}_r)^{-1}$ so that $\mathbf{W}_r^T \mathbf{V}_r = \mathbf{I}_r$.

4: Update ROM: $\hat{\mathbf{A}} = \mathbf{W}_r^T \mathbf{A} \mathbf{W}_r$, $\hat{\mathbf{B}} = \mathbf{W}_r^T \mathbf{B}$, $\hat{\mathbf{C}} = \mathbf{C} \mathbf{W}_r$.

5: Solve $\hat{\mathbf{A}} \hat{\mathbf{P}} + \hat{\mathbf{P}} \hat{\mathbf{A}}^T - \hat{\mathbf{P}} \hat{\mathbf{C}}^T \hat{\mathbf{C}} \hat{\mathbf{P}} + \hat{\mathbf{B}} \hat{\mathbf{B}}^T = 0$.

6: Compute $\hat{\mathbf{G}}_{\mathbf{F}}(s) = [\mathbf{I}, 0] + \sum_{j=1}^r \frac{\hat{\mathbf{c}}_j [\hat{\mathbf{f}}_j^T, \hat{\mathbf{b}}_j^T]}{s - \hat{\lambda}_j}$.

7: $\sigma_j \leftarrow -\hat{\lambda}_j$, $\mathbf{r}_j \leftarrow \hat{\mathbf{b}}_j$, and $\ell_j \leftarrow \hat{\mathbf{c}}_j$ for $j = 1, 2, \dots, r$.

8: **end while**

3.4 Algorithm 1 and \mathcal{H}_2 -gap Optimality.

So far we have motivated Algorithm 1 as a way to eliminate the contribution from the second-sum in the error expression (24). However, the reduced model $\hat{\mathbf{G}}$ from Algorithm 1 achieves more. First, recall that $\|\mathbf{G} - \hat{\mathbf{G}}\|_{\mathcal{H}_2\text{-gap}} = \|\mathbf{G}_{\mathbf{F}} - \hat{\mathbf{G}}_{\mathbf{F}}\|_{\mathcal{H}_2}$. Therefore, if we interpret the problem as the \mathcal{H}_2 optimal model reduction of the closed-loop model $\mathbf{G}_{\mathbf{F}}$, the \mathcal{H}_2 optimal reduced closed-loop model (or equivalently \mathcal{H}_2 -gap optimal closed-loop model) $\hat{\mathbf{G}}_{\mathbf{F}}(s) = [\mathbf{I}, 0] + \sum_{j=1}^r \frac{\hat{\mathbf{c}}_j [\hat{\mathbf{f}}_j^T, \hat{\mathbf{b}}_j^T]}{s - \hat{\lambda}_j}$ satisfies

$$\begin{aligned} (\mathbf{G}_{\mathbf{F}}(-\hat{\lambda}_j) - \hat{\mathbf{G}}_{\mathbf{F}}(-\hat{\lambda}_j)) \begin{bmatrix} \hat{\mathbf{f}}_j \\ \hat{\mathbf{b}}_j \end{bmatrix} &= \mathbf{0}, \quad \hat{\mathbf{c}}_j^T (\mathbf{G}_{\mathbf{F}}(-\hat{\lambda}_j) - \hat{\mathbf{G}}_{\mathbf{F}}(-\hat{\lambda}_j)) = 0, \\ \text{and } \hat{\mathbf{c}}_j^T (\mathbf{G}'_{\mathbf{F}}(-\hat{\lambda}_j) - \hat{\mathbf{G}}'_{\mathbf{F}}(-\hat{\lambda}_j)) \begin{bmatrix} \hat{\mathbf{f}}_j \\ \hat{\mathbf{b}}_j \end{bmatrix} &= 0, \quad \text{for } j = 1, 2, \dots, r. \end{aligned} \quad (27)$$

Therefore, upon convergence, Algorithm 1 enforces the first set of necessary conditions in (27), namely the right-tangential interpolation conditions.

One can also think about necessary and sufficient optimality conditions for the restricted setting. Assume that the reduced closed-loop poles $\{\widehat{\lambda}_j\}$ and the reduced closed-loop left residue-directions $\{\widehat{\mathbf{c}}_j\}$ are fixed, so that the only free variables in $\widehat{\mathbf{G}}_F$ are the right residue-directions, $\{[\widehat{\mathbf{f}}_j^T \ \widehat{\mathbf{b}}_j^T]\}$. As [7] showed, for fixed reduced poles and left residue-directions, $\widehat{\mathbf{G}}_F$ minimizes $\|\mathbf{G}_F - \widehat{\mathbf{G}}_F\|_{\mathcal{H}_2} = \|\mathbf{G} - \widehat{\mathbf{G}}\|_{\mathcal{H}_2\text{-gap}}$ if and only if $(\mathbf{G}_F(-\widehat{\lambda}_j) - \widehat{\mathbf{G}}_F(-\widehat{\lambda}_j)) \begin{bmatrix} \widehat{\mathbf{f}}_j \\ \widehat{\mathbf{b}}_j \end{bmatrix} = \mathbf{0}$. That is, right tangential interpolation at the mirror images of the reduced poles becomes a necessary and sufficient condition for optimality. Therefore, for the converged reduced closed-loop poles and left residue-directions, Algorithm 1 gives the global minimizer.

We end this discussion with a warning. The optimality conditions that we argue Algorithm 1 satisfies view $\widehat{\mathbf{G}}_F$ as the variable with which to minimize the error. Corollary 1 reveals that we can enforce these optimality conditions via choosing $\widehat{\mathbf{G}}$ appropriately. However, the \mathcal{H}_2 -gap optimality conditions with respect to $\widehat{\mathbf{G}}$ will be different than those in (27). Those conditions, together with an algorithm to satisfy them, will be the focus of future work.

4 Numerical Examples

We present two numerical examples resulting from spatial semidiscretizations of partial differential equations. We compare Algorithm 1 with the method of LQG balanced truncation, as well as the standard version of IRKA. For the implementation of LQG balanced truncation, we rely on the built-in MATLAB routine `care` for solving the required Riccati equations. Note that both examples result in unstable dynamical systems—so the application of IRKA needs further explanation. IRKA is a method for optimal \mathcal{H}_2 model reduction of asymptotically stable dynamical systems. From a computational perspective, its implementation does not prevent one from using it to reduce unstable systems having no poles on the imaginary axis. This has been studied extensively in [27] where it was shown that IRKA applied to unstable systems having a modest number of unstable poles produces accurate approximations. We have chosen this formulation of IRKA as opposed to the modified version in [23] for unstable systems since the latter requires a full stable/anti-stable decomposition of the full model. All simulations were generated on an AMD Ryzen 7 1800X @ 3.68 GHz \times 16, 64 GB RAM, MATLAB Version 9.2.0.538062 (R2017a).

4.1 An Unstable Convection Diffusion Equation

The first example is a (scalable) finite-difference discretization of the following controlled convection diffusion equation on the unit square

$$\begin{aligned}
 v_t &= \Delta v - 20 \cdot \sin(x)v_x + 50 \cdot v + \chi_\omega \cdot u(t) && \text{in } (0, 1)^2 \times (0, T), \\
 v(x, 0, t) &= v(0, y, t) = v(x, 1, t) = v(1, y, t) = 0 && \text{in } (0, 1) \times (0, T), \\
 v(x, y, 0) &= v_0(x, y) = 0 && \text{in } (0, 1)^2.
 \end{aligned}$$

Here, by χ we denote the characteristic function on the control domain $\omega = [0.2, 0.3] \times [0.2, 0.3]$. We augment the system by an output variable $y(t)$ corresponding to the mean value in an observable domain $y(t) = \int_{0.5}^{0.7} \int_{0.7}^{0.9} v(x, y, t) dx dy$. We present the results for a system of dimension $n = 400$ corresponding to a uniform 20×20 grid. The discretized system matrix \mathbf{A} has 12 eigenvalues in the right half plane but the pairs (\mathbf{A}, \mathbf{B}) and (\mathbf{A}, \mathbf{C}) satisfy the required stabilizability assumptions.

Table 1 shows the error of different reduced-order systems with respect to the gap topology as well as the newly introduced \mathcal{H}_2 -gap. Note that in all cases, the reduced-order systems computed via the proposed method gap-IRKA yield smaller \mathcal{H}_2 -gap errors than LQG balanced truncation as well as IRKA. The results are missing for $r = 1$ for IRKA since it did not converge. Even for the \mathcal{H}_∞ -gap, in all but three cases, gap-IRKA leads to better results than the other methods. Improving upon results from LQG balanced truncation with respect to the \mathcal{H}_∞ -gap without ever computing large-scale Riccati solutions is a remarkable demonstration of the effectiveness of the proposed interpolatory framework in reducing unstable systems.

Table 1 Approximation error $\|\mathbf{G}_F - \widehat{\mathbf{G}}_F\|_{\mathcal{H}_2}$ (left) and $\|\mathbf{G}_F - \widehat{\mathbf{G}}_F\|_{\mathcal{H}_\infty}$ (right)

| r | $\ \mathbf{G}_F - \widehat{\mathbf{G}}_F\ _{\mathcal{H}_2}$ | | | $\ \mathbf{G}_F - \widehat{\mathbf{G}}_F\ _{\mathcal{H}_\infty}$ | | |
|-----|---|----------------------|--|--|--|--|
| | IRKA | LQG-BT | gap-IRKA | IRKA | LQG-BT | gap-IRKA |
| 1 | - | $5.34 \cdot 10^{-1}$ | $5.34 \cdot 10^{-1}$ | - | $2.29 \cdot 10^{-1}$ | $2.28 \cdot 10^{-1}$ |
| 2 | $3.09 \cdot 10^{-1}$ | $1.03 \cdot 10^{-1}$ | $1.03 \cdot 10^{-1}$ | $5.66 \cdot 10^{-2}$ | $1.87 \cdot 10^{-2}$ | $1.86 \cdot 10^{-2}$ |
| 3 | $1.02 \cdot 10^{-1}$ | $1.51 \cdot 10^{-2}$ | $1.47 \cdot 10^{-2}$ | $4.80 \cdot 10^{-2}$ | $4.75 \cdot 10^{-3}$ | $4.28 \cdot 10^{-3}$ |
| 4 | $1.09 \cdot 10^{-2}$ | $7.00 \cdot 10^{-3}$ | $5.89 \cdot 10^{-3}$ | $1.12 \cdot 10^{-3}$ | $7.87 \cdot 10^{-4}$ | $2.38 \cdot 10^{-3}$ |
| 5 | $1.38 \cdot 10^{-3}$ | $9.16 \cdot 10^{-4}$ | $9.04 \cdot 10^{-4}$ | $1.57 \cdot 10^{-4}$ | $9.21 \cdot 10^{-5}$ | $7.54 \cdot 10^{-5}$ |
| 6 | $2.23 \cdot 10^{-4}$ | $6.91 \cdot 10^{-5}$ | $6.83 \cdot 10^{-5}$ | $3.70 \cdot 10^{-5}$ | $7.59 \cdot 10^{-6}$ | $8.35 \cdot 10^{-6}$ |
| 7 | $6.01 \cdot 10^{-5}$ | $5.88 \cdot 10^{-5}$ | $5.78 \cdot 10^{-5}$ | $3.56 \cdot 10^{-6}$ | $3.58 \cdot 10^{-6}$ | $3.45 \cdot 10^{-6}$ |
| 8 | $6.24 \cdot 10^{-6}$ | $5.23 \cdot 10^{-6}$ | $5.18 \cdot 10^{-6}$ | $4.31 \cdot 10^{-7}$ | $3.48 \cdot 10^{-7}$ | $3.34 \cdot 10^{-7}$ |
| 9 | $7.78 \cdot 10^{-7}$ | $4.08 \cdot 10^{-7}$ | $3.98 \cdot 10^{-7}$ | $1.01 \cdot 10^{-7}$ | $3.99 \cdot 10^{-8}$ | $4.49 \cdot 10^{-8}$ |
| 10 | $3.11 \cdot 10^{-7}$ | $3.06 \cdot 10^{-7}$ | $3.02 \cdot 10^{-7}$ | $1.23 \cdot 10^{-8}$ | $1.34 \cdot 10^{-8}$ | $1.21 \cdot 10^{-8}$ |

4.2 Linearized Navier-Stokes Equations

We consider a linearization of the incompressible Navier-Stokes equations around an unsteady flow profile, giving rise to the Stokes-Oseen system

$$\begin{aligned}
 v_t &= \nu \Delta v - (v \cdot \nabla)z - (z \cdot \nabla)v - \nabla p + Bu && \text{in } \Omega \times (0, T), \\
 \operatorname{div} v &= 0 && \text{in } \Omega \times (0, T), \\
 v &= 0 && \text{on } \Gamma \times (0, T), \\
 v(0) &= 0, && \text{in } \Omega,
 \end{aligned}
 \tag{28}$$

where $\nu = \frac{1}{\operatorname{Re}} = \frac{1}{90}$ and the geometry $\Omega = (0, 2.2) \times (0.41)$ describes the flow around a cylindrical obstacle. The precise setup together with a description of the control and observation operators B and C , respectively, is given in [8], to which we also refer for further details. We used the semi-discretized model from [8] corresponding to a Taylor-Hood finite element discretization of the Navier-Stokes equations with $n = n_v + n_p = 9356 + 1289$ degrees of freedom. The original system results in a differential algebraic system; the pressure term is explicitly eliminated with an algebraic approach described in [19]. Note that the system matrices of the transformed

Table 2 Approximation error $\|G_F - \widehat{G}_F\|_{\mathcal{H}_2}$ (left) and $\|G_F - \widehat{G}_F\|_{\mathcal{H}_\infty}$ (right)

| r | $\ G_F - \widehat{G}_F\ _{\mathcal{H}_2}$ | | | $\ G_F - \widehat{G}_F\ _{\mathcal{H}_\infty}$ | | |
|-----|---|----------------------|--|--|--|--|
| | IRKA | LQG-BT | gap-IRKA | IRKA | LQG-BT | gap-IRKA |
| 2 | - | $5.69 \cdot 10^{-1}$ | $5.67 \cdot 10^{-1}$ | - | $3.76 \cdot 10^{-1}$ | $3.70 \cdot 10^{-1}$ |
| 4 | $2.42 \cdot 10^{-1}$ | $1.99 \cdot 10^{-1}$ | $1.93 \cdot 10^{-1}$ | $9.84 \cdot 10^{-2}$ | $8.26 \cdot 10^{-2}$ | $7.52 \cdot 10^{-2}$ |
| 6 | $1.08 \cdot 10^{-1}$ | $1.84 \cdot 10^{-1}$ | $1.01 \cdot 10^{-1}$ | $5.11 \cdot 10^{-2}$ | $9.02 \cdot 10^{-2}$ | $5.11 \cdot 10^{-2}$ |
| 8 | $8.56 \cdot 10^{-2}$ | $9.33 \cdot 10^{-2}$ | $8.28 \cdot 10^{-2}$ | $5.10 \cdot 10^{-2}$ | $4.01 \cdot 10^{-2}$ | $4.86 \cdot 10^{-2}$ |
| 10 | $5.78 \cdot 10^{-2}$ | $7.02 \cdot 10^{-2}$ | $5.63 \cdot 10^{-2}$ | $3.01 \cdot 10^{-2}$ | $2.47 \cdot 10^{-2}$ | $3.01 \cdot 10^{-2}$ |
| 12 | $3.25 \cdot 10^{-2}$ | $3.38 \cdot 10^{-2}$ | $2.98 \cdot 10^{-2}$ | $9.68 \cdot 10^{-3}$ | $1.15 \cdot 10^{-2}$ | $9.63 \cdot 10^{-3}$ |
| 14 | $1.79 \cdot 10^{-2}$ | $1.96 \cdot 10^{-2}$ | $1.71 \cdot 10^{-2}$ | $7.16 \cdot 10^{-3}$ | $7.45 \cdot 10^{-3}$ | $6.62 \cdot 10^{-3}$ |
| 16 | $1.12 \cdot 10^{-2}$ | $1.34 \cdot 10^{-2}$ | $1.13 \cdot 10^{-2}$ | $5.15 \cdot 10^{-3}$ | $3.73 \cdot 10^{-3}$ | $4.70 \cdot 10^{-3}$ |
| 18 | $7.84 \cdot 10^{-3}$ | $9.62 \cdot 10^{-3}$ | $7.19 \cdot 10^{-3}$ | $4.78 \cdot 10^{-3}$ | $3.14 \cdot 10^{-3}$ | $4.44 \cdot 10^{-3}$ |
| 20 | $4.79 \cdot 10^{-3}$ | $5.25 \cdot 10^{-3}$ | $4.66 \cdot 10^{-3}$ | $2.44 \cdot 10^{-3}$ | $1.28 \cdot 10^{-3}$ | $2.38 \cdot 10^{-3}$ |
| 22 | $3.52 \cdot 10^{-3}$ | $4.29 \cdot 10^{-3}$ | $3.44 \cdot 10^{-3}$ | $1.12 \cdot 10^{-3}$ | $9.55 \cdot 10^{-4}$ | $1.06 \cdot 10^{-3}$ |
| 24 | $2.49 \cdot 10^{-3}$ | $2.69 \cdot 10^{-3}$ | $2.38 \cdot 10^{-3}$ | $5.81 \cdot 10^{-4}$ | $7.38 \cdot 10^{-4}$ | $5.67 \cdot 10^{-4}$ |
| 26 | $1.89 \cdot 10^{-3}$ | $2.51 \cdot 10^{-3}$ | $1.87 \cdot 10^{-3}$ | $5.52 \cdot 10^{-4}$ | $8.33 \cdot 10^{-4}$ | $5.58 \cdot 10^{-4}$ |
| 28 | $1.39 \cdot 10^{-3}$ | $1.89 \cdot 10^{-3}$ | $1.38 \cdot 10^{-3}$ | $4.41 \cdot 10^{-4}$ | $4.39 \cdot 10^{-4}$ | $4.44 \cdot 10^{-4}$ |
| 30 | $1.22 \cdot 10^{-3}$ | $1.41 \cdot 10^{-3}$ | $1.21 \cdot 10^{-3}$ | $4.39 \cdot 10^{-4}$ | $2.91 \cdot 10^{-4}$ | $4.32 \cdot 10^{-4}$ |
| 32 | $8.94 \cdot 10^{-4}$ | $1.01 \cdot 10^{-3}$ | $8.78 \cdot 10^{-4}$ | $2.63 \cdot 10^{-4}$ | $2.02 \cdot 10^{-4}$ | $2.60 \cdot 10^{-4}$ |
| 34 | $7.01 \cdot 10^{-4}$ | $8.52 \cdot 10^{-4}$ | $6.89 \cdot 10^{-4}$ | $2.47 \cdot 10^{-4}$ | $1.62 \cdot 10^{-4}$ | $2.47 \cdot 10^{-4}$ |
| 36 | $5.31 \cdot 10^{-4}$ | $6.08 \cdot 10^{-4}$ | $5.17 \cdot 10^{-4}$ | $1.22 \cdot 10^{-4}$ | $1.35 \cdot 10^{-4}$ | $1.21 \cdot 10^{-4}$ |
| 38 | $3.60 \cdot 10^{-4}$ | $4.02 \cdot 10^{-4}$ | $3.53 \cdot 10^{-4}$ | $7.54 \cdot 10^{-5}$ | $6.71 \cdot 10^{-5}$ | $7.53 \cdot 10^{-5}$ |
| 40 | $2.51 \cdot 10^{-4}$ | $2.85 \cdot 10^{-4}$ | $2.45 \cdot 10^{-4}$ | $5.30 \cdot 10^{-5}$ | $6.20 \cdot 10^{-5}$ | $5.12 \cdot 10^{-5}$ |

ODE are dense and an explicit computation generally should be avoided. In our case, the dimension of the ODE is $n = n_v - n_p = 8067$ and can still be handled by direct solvers in MATLAB; we refrain from a more sophisticated approach here.

We repeat similar experiments as described in Sect. 4.1 and compare the proposed method to IRKA and LQG balanced truncation. The results are depicted in Table 2; the missing data for $r = 2$ for IRKA is due to non-convergence. As Table 2 illustrates, the reduced systems generated by Algorithm 1 in all cases save one, yield the smallest \mathcal{H}_2 -gap error. Moreover, in eight out of the twenty cases tested, the proposed method outperforms the LQG balanced truncation in terms of the \mathcal{H}_∞ -gap as well, without computing a large-scale Riccati-based Gramians and despite not being developed for this measure.

5 Conclusion

We have presented a new approach for model reduction of linear stabilizable and detectable control systems. Based on the theory of left-coprime factorizations and a newly introduced \mathcal{H}_2 -gap, we have derived pole-residue formulae that suggest tangentially interpolating the original transfer function at the mirrored closed-loop reduced system poles. Since these are not known a priori, we modified the iterative rational Krylov algorithm accordingly. Two numerical examples associated with (unstable) partial differential equations illustrate the applicability and good performance of the new approach.

Acknowledgements Parts of this work were completed while the first author visited Virginia Tech; the kind hospitality was greatly appreciated. The work of Beattie was supported in parts by NSF through Grant DMS-1819110. The work of Gugercin was supported in parts by NSF through Grants DMS-1720257 and DMS-1819110.

References

1. Anić, B., Beattie, C.A., Gugercin, S., Antoulas, A.C.: Interpolatory weighted- \mathcal{H}_2 model reduction. *Automatica* **49**(5), 1275–1280 (2013)
2. Antoulas, A.C., Beattie, C.A., Gugercin, S.: Interpolatory model reduction of large-scale dynamical systems. In: *Efficient Modeling and Control of Large-Scale Systems*, pp. 3–58. Springer (2010)
3. Antoulas, A.C., Beattie, C.A., Gugercin, S.: *Interpolatory methods for model reduction*. Computational Science and Engineering 21. SIAM, Philadelphia (2020)
4. Antoulas, A.C.: *Approximation of Large-Scale Dynamical Systems*. SIAM (2005)
5. Barrachina, S., Benner, P., Quintana-Ortí, E.S., Quintana-Ortí, G.: Parallel algorithms for balanced truncation of large-scale unstable systems. In: *Proceedings of the 44th IEEE Conference on Decision and Control*, pp. 2248–2253. IEEE (2005)
6. Beattie, C.A., Gugercin, S.: Krylov-based minimization for optimal \mathcal{H}_2 model reduction. In: *Proceedings of 46th IEEE Conference on Decision and Control*, pp. 4385–4390 (2007)
7. Beattie, C.A., Gugercin, S.: Realization-independent \mathcal{H}_2 -approximation. In: *Proceedings of 51st IEEE Conference on Decision and Control*, pp. 4953–4958 (2012)

8. Behr, M., Benner, P., Heiland, J.: Example setups of Navier-Stokes equations with control and observation: spatial discretization and representation via linear-quadratic matrix coefficients. Technical report, Max Planck Institute for Complex Dynamical Systems (2017). <https://arxiv.org/abs/1707.08711>
9. Benner, P., Saak, J., Uddin, M.M.: Balancing based model reduction for structured index-2 unstable descriptor systems with application to flow control. *Numer. Algebr., Control Optim.* **6**, 1–20 (2016)
10. Breiten, T., Beattie, C.A., Gugercin, S.: Near-optimal frequency-weighted interpolatory model reduction. *Syst. Control Lett.* **78**, 8–18 (2015)
11. Bunse-Gerstner, A., Kubalinska, D., Vossen, G., Wilczek, D.: \mathcal{H}_2 -norm optimal model reduction for large scale discrete dynamical MIMO systems. *J. Comput. Appl. Math.* **233**(5), 1202–1216 (2010)
12. Curtain, R.F.: Model reduction for control design for distributed parameter systems. In: Smith, R.C., Demetriou, M. (eds.) *Research Directions in Distributed Parameter Systems*, pp. 95–121. SIAM (2003)
13. Gallivan, K., Vandendorpe, A., Van Dooren, P.: Model reduction of MIMO systems via tangential interpolation. *SIAM J. Matrix Anal. Appl.* **26**(2), 328–349 (2005)
14. Gawronski, W., Juang, J.-N.: Model reduction in limited time and frequency intervals. *Int. J. Syst. Sci.* **21**(2), 349–376 (1990)
15. Gugercin, S.: Projection methods for model reduction of large-scale dynamical systems. Ph.D. thesis, Rice University (2003)
16. Gugercin S, Antoulas, A.C.: A time-limited balanced reduction method. In: 42nd IEEE International Conference on Decision and Control, vol. 5, pp. 5250–5253 (2003)
17. Gugercin, S., Antoulas, A.C., Beattie, C.A.: \mathcal{H}_2 model reduction for large-scale linear dynamical systems. *SIAM J. Matrix Anal. Appl.* **30**(2), 609–638 (2008)
18. Gugercin, S., Beattie, C.A.: Model reduction by rational interpolation. In: *Model Reduction and Approximation*, pp. 297–334. SIAM, Philadelphia (2017)
19. Heinkenschloss, M., Sorensen, D., Sun, K.: Balanced truncation model reduction for a class of descriptor systems with application to the Oseen equations. *SIAM J. Sci. Comput.* **30**(2), 1038–1063 (2008)
20. Jonckheere, E., Silverman, L.: A new set of invariants for linear systems—application to reduced order compensator design. *IEEE Trans. Autom. Control* **28**(10), 953–964 (1983)
21. Kürschner, P.: Balanced truncation model order reduction in limited time intervals for large systems. *Adv. Comput. Math.* **44**(6), 1821–1844 (2018)
22. Lin, Y., Simoncini, V.: A new subspace iteration method for the algebraic Riccati equation. *Numer. Linear Algebr. Appl.* **22**(1), 26–47 (2015)
23. Magruder, C., Beattie, C., Gugercin, S.: Rational Krylov methods for optimal \mathcal{L}_2 model reduction. In: 49th IEEE Conference on Decision and Control (CDC), pp. 6797–6802 (2010)
24. Meyer, D.G.: Fractional balanced reduction: model reduction via fractional representation. *IEEE Trans. Autom. Control* **35**(12), 1341–1345 (1990)
25. Nett, C., Jacobson, C., Balas, M.: A connection between state-space and doubly coprime fractional representations. *IEEE Trans. Autom. Control* **29**(9), 831–832 (1984)
26. Partington, J.R.: Approximation of unstable infinite-dimensional systems using coprime factors. *Syst. Control Lett.* **16**, 89–96 (1991)
27. Sinani, K.: Iterative rational Krylov algorithm for unstable dynamical systems and generalized coprime factorizations. Master’s thesis, Virginia Tech (2016)
28. Sinani, K., Gugercin, S.: $\mathcal{H}_2(t_f)$ optimality conditions for a finite-time horizon. *Automatica* **110**, 108604 (2019)
29. Vidyasagar, M.: The graph metric for unstable plants and robustness estimates for feedback stability. *IEEE Trans. Autom. Control* **29**(5), 403–418 (1984)
30. Xu, Y., Zeng, T.: Optimal \mathcal{H}_2 model reduction for large scale MIMO systems via tangential interpolation. *Int. J. Numer. Anal. Model.* **8**(1), 174–188 (2011)
31. Zhou, K., Salomon, G., Wu, E.: Balanced realization and model reduction for unstable systems. *Int. J. Robust Nonlinear Control* **9**(3), 183–198 (1999)

Reduced Order Model Hessian Approximations in Newton Methods for Optimal Control



Matthias Heinkenschloss and Caleb Magruder

Abstract This paper introduces reduced order model (ROM) based Hessian approximations for use in inexact Newton methods for the solution of optimization problems implicitly constrained by a large-scale system, typically a discretization of a partial differential equation (PDE). The direct application of an inexact Newton method to this problem requires the solution of many PDEs per optimization iteration. To reduce the computational complexity, a ROM Hessian approximation is proposed. Since only the Hessian is approximated, but the original objective function and its gradient is used, the resulting inexact Newton method maintains the first-order global convergence property, under suitable assumptions. Thus even computationally inexpensive lower fidelity ROMs can be used, which is different from ROM approaches that replace the original optimization problem by a sequence of ROM optimization problem and typically need to accurately approximate function and gradient information of the original problem. In the proposed approach, the quality of the ROM Hessian approximation determines the rate of convergence, but not whether the method converges. The projection based ROM is constructed from state and adjoint snapshots, and is relatively inexpensive to compute. Numerical examples on semilinear parabolic optimal control problems demonstrate that the proposed approach can lead to substantial savings in terms of overall PDE solves required.

Keywords Model reduction · Optimization · Hessian approximation · Newton method

M. Heinkenschloss (✉)
Department of Computational and Applied Mathematics, Rice University, Houston, TX, USA
e-mail: heinken@rice.edu

C. Magruder
MathWorks, Natick, MA, USA
e-mail: cmagrude@mathworks.com

1 Introduction

We introduce reduced order model (ROM) based Hessian approximations for use in inexact Newton methods for the solution of large-scale smooth optimization problems

$$\min_{\mathbf{u} \in \mathbb{R}^m} \widehat{J}(\mathbf{u}) = J(\mathbf{y}(\mathbf{u}), \mathbf{u}), \quad (1)$$

where for given $\mathbf{u} \in \mathbb{R}^m$ the vector $\mathbf{y}(\mathbf{u}) \in \mathbb{R}^n$ is the solution of

$$\mathbf{c}(\mathbf{y}(\mathbf{u}), \mathbf{u}) = \mathbf{0}. \quad (2)$$

The optimization problem (1, 2) often comes from a discretization of optimal control problems and in this case $\mathbf{u} \in \mathbb{R}^m$ is the discretized control, $\mathbf{y} \in \mathbb{R}^n$ is the discretized state, and (2) is the discretized state equation. Our approach can be easily extended to a setting where the control space \mathcal{U} and the state space \mathcal{Y} are Hilbert spaces, but because of space restrictions we limit ourselves to the finite dimensional case. In our applications, the state Eq. (2) is a discretized parabolic partial differential equation (PDE). In this case, each objective function (1) evaluation requires the solution of a discretized PDE, the evaluation of the objective function gradient requires the solution of the (linear) adjoint PDE, and the evaluation of a Hessian-times-vector multiplication requires the additional solution of two (linear) PDE. This is computationally expensive. In addition, the additional PDEs that have to be solved for gradient and Hessian-times-vector computations depend on the solution of $\mathbf{y}(\mathbf{u}) \in \mathbb{R}^n$ of the state equation, which for time dependent PDEs is also memory intensive. ROMs can be used to reduce the computation time and memory requirements.

Previous approaches of using ROMs in optimization have approximated the mapping $\mathbf{u} \mapsto \mathbf{y}(\mathbf{u}) \in \mathbb{R}^n$ using a ROM applied to the state Eq. (2). Typically, the state Eq. (2) is approximated by a projection based ROM

$$\mathbf{V}^T \mathbf{c}(\mathbf{V} \widehat{\mathbf{y}}(\mathbf{u}), \mathbf{u}) = \mathbf{0}, \quad (3)$$

where $\mathbf{V} \in \mathbb{R}^{n \times r}$ is a matrix with rank $r \ll n$. The ROM state solution is then used to approximate (1) by

$$\min_{\mathbf{u} \in \mathbb{R}^m} J(\mathbf{V} \widehat{\mathbf{y}}(\mathbf{u}), \mathbf{u}). \quad (4)$$

For many problems, including example problems in Sect. 4 of this paper, the ROM approximation (4, 3) of the states implies via the optimality conditions that the controls \mathbf{u} are also contained in a low dimensional subspace related to \mathbf{V} .

In special cases, one can extend ROM approaches for dynamical systems [5] to find one ROM such that the solution of (4, 3) is a good approximation of the solution of (1, 2). See, e.g., [2, 3], and the survey [7]. For general problems, however, the ROM is only valid in a potentially small neighborhood of the current control \mathbf{u}_c and corresponding state $\mathbf{y}_c = \mathbf{y}(\mathbf{u}_c)$. In this case trust-region based model management approaches have been proposed. See [1, 10, 13, 16, 19, 20], and the surveys [7, 17].

While in principle, one can use trust-region based model management approaches that generate a new ROM at each new iterate, in practice their use is limited by the computational cost of ROM generation versus the computational savings resulting from that ROM.

Therefore, instead of generating ROMs for the optimization problem, we generate ROM approximations of the Hessians. These ROMs are computationally cheaper than a ROM that also has to approximate the objective function (1) and its gradient over a range of controls. Ideally the ROM Hessians still generate optimization steps that are close to the Newton steps, and therefore lead to fast local convergence of the optimization algorithm at a fraction of the cost of a corresponding Newton-type methods applied to the full order model (FOM) problem (1, 2).

In Sect. 2 we will outline our new approach and in Sect. 3 we will specify it further for a discretized parabolic optimal control problem. We demonstrate our approach on semilinear parabolic optimal control problems in Sect. 4.

2 Inexact Newton Methods and ROM Hessian Approximations

This section provides a general outline of our proposed approach. We begin with a review of gradient and Hessian computation and the line-search Newton- Conjugate-Gradient (Newton-CG) method. Then we will present the general structure of our ROM Hessian approximation, and a heuristic for choosing the ROM subspace within this approximation.

Inexact Newton Method and ROM Hessian Approximation. To make the definition of the objective function \widehat{J} and its gradient and Hessian calculation rigorous, we make the following assumptions throughout.

- (A1) There is an open set $D_u \subset \mathbb{R}^m$ such that for all $\mathbf{u} \in D_u$ the equation $\mathbf{c}(\mathbf{y}, \mathbf{u}) = 0$ has a unique solution $\mathbf{y} \in \mathbb{R}^n$.
- (A2) There exists an open set $D_y \subset \mathbb{R}^n$ such that $\{\mathbf{y}(\mathbf{u}) : \mathbf{u} \in D_u\} \subset D_y$, and the functions J and \mathbf{c} are twice continuously differentiable on $D_u \times D_y$.
- (A3) The inverse $\mathbf{c}_y(\mathbf{y}, \mathbf{u})^{-1}$ exists for all $(\mathbf{y}, \mathbf{u}) \in \{(\mathbf{y}, \mathbf{u}) \in D_u \times D_y : \mathbf{c}(\mathbf{y}, \mathbf{u}) = \mathbf{0}\}$.

Here we use $\nabla_y J(\mathbf{y}, \mathbf{u}) \in \mathbb{R}^n$, $\nabla_u J(\mathbf{y}, \mathbf{u}) \in \mathbb{R}^m$ to denote the partial gradients and $\mathbf{c}_y(\mathbf{y}, \mathbf{u}) \in \mathbb{R}^{n \times n}$, $\mathbf{c}_u(\mathbf{y}, \mathbf{u}) \in \mathbb{R}^{n \times m}$ to denote the partial Jacobians. Similar notation is used for second derivatives.

The gradient of the objective \widehat{J} in (1, 2) can be computed using the so-called adjoint equation approach. See, e.g., [12, Sect. 1.6]. Define the Lagrangian

$$L(\mathbf{y}, \mathbf{u}, \boldsymbol{\lambda}) := J(\mathbf{y}, \mathbf{u}) + \boldsymbol{\lambda}^T \mathbf{c}(\mathbf{y}, \mathbf{u})$$

corresponding to (1, 2). Given a control \mathbf{u} and corresponding state $\mathbf{y}(\mathbf{u})$ the gradient of the objective \widehat{J} is computed by first solving the adjoint equation

$$\mathbf{c}_y(\mathbf{y}(\mathbf{u}), \mathbf{u})^T \boldsymbol{\lambda} = -\nabla_y J(\mathbf{y}(\mathbf{u}), \mathbf{u}) \quad (5a)$$

for $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\mathbf{u})$ and then setting

$$\nabla \widehat{J}(\mathbf{u}) = \nabla_u J(\mathbf{y}(\mathbf{u}), \mathbf{u}) + \mathbf{c}_u(\mathbf{y}(\mathbf{u}), \mathbf{u})^T \boldsymbol{\lambda}(\mathbf{u}). \quad (5b)$$

Given a control \mathbf{u} , the corresponding state $\mathbf{y}(\mathbf{u})$, and the corresponding adjoint $\boldsymbol{\lambda}(\mathbf{u})$, the Hessian-times-vector product $\nabla^2 \widehat{J}(\mathbf{u}) \mathbf{d}$ is computed by solving the linearized state equation

$$\mathbf{c}_y(\mathbf{y}(\mathbf{u}), \mathbf{u}) \mathbf{w} = \mathbf{c}_u(\mathbf{y}(\mathbf{u}), \mathbf{u}) \mathbf{d} \quad (6a)$$

for \mathbf{w} , then the second order adjoint equation

$$\mathbf{c}_y(\mathbf{y}(\mathbf{u}), \mathbf{u})^T \mathbf{p} = \nabla_{yy} L(\mathbf{y}(\mathbf{u}), \mathbf{u}, \boldsymbol{\lambda}(\mathbf{u})) \mathbf{w} - \nabla_{yu} L(\mathbf{y}(\mathbf{u}), \mathbf{u}, \boldsymbol{\lambda}(\mathbf{u})) \mathbf{d} \quad (6b)$$

for \mathbf{p} , and then computing

$$\begin{aligned} \nabla^2 \widehat{J}(\mathbf{u}) \mathbf{d} = & \mathbf{c}_u(\mathbf{y}(\mathbf{u}), \mathbf{u})^T \mathbf{p} - \nabla_{uy} L(\mathbf{y}(\mathbf{u}), \mathbf{u}, \boldsymbol{\lambda}(\mathbf{u})) \mathbf{w} \\ & + \nabla_{uu} L(\mathbf{y}(\mathbf{u}), \mathbf{u}, \boldsymbol{\lambda}(\mathbf{u})) \mathbf{d}. \end{aligned} \quad (6c)$$

For optimal control problems, the computation of $\mathbf{y}(\mathbf{u})$ requires the solution of a nonlinear discretized PDE. The gradient computation requires the solution of a linear discretized adjoint PDE to compute $\boldsymbol{\lambda}(\mathbf{u})$. Each Hessian-times-vector operation $\nabla^2 \widehat{J}(\mathbf{u}) \mathbf{d}$ requires the solution of two linear discretized PDEs to compute \mathbf{w} and \mathbf{p} , respectively.

The problem (1) is solved using a line-search Newton-CG Method. However, a trust-region method could be used as well, and the proposed ROM Hessian approximation provides the same computational benefits in a trust-region setting.

Given a current iterate \mathbf{u}_c the line-search Newton-CG Method [15, Algorithm 7.1] applies the truncated CG method to approximately solve the Newton subproblem

$$\nabla^2 \widehat{J}(\mathbf{u}_c) \mathbf{d} = -\nabla \widehat{J}(\mathbf{u}_c). \quad (7)$$

Then it computes a suitable step-size $\alpha_c \in (0, 1]$ such that the sufficient decrease condition

$$\widehat{J}(\mathbf{u}_c + \alpha_c \mathbf{d}) \leq \widehat{J}(\mathbf{u}_c) + 10^{-4} \alpha_c \mathbf{d}^T \nabla \widehat{J}(\mathbf{u}_c) \quad (8)$$

is satisfied. The new iterate is given by $\mathbf{u}_+ = \mathbf{u}_c + \alpha_c \mathbf{d}$.

The ROM Hessian approximation is computed by applying a projection ROM to the Eq. (6a, b). Let $\mathbf{V} \in \mathbb{R}^{n \times r}$, $r \ll n$, with linearly independent columns such that $\mathbf{V}^T \mathbf{c}_y(\mathbf{y}(\mathbf{u}), \mathbf{u}) \mathbf{V}$ is invertible.

The ROM Hessian-times-vector product is computed by first solving the projected linearized state equation

$$\mathbf{V}^T \mathbf{c}_y(\mathbf{y}(\mathbf{u}), \mathbf{u}) \mathbf{V} \widehat{\mathbf{w}} = \mathbf{V}^T \mathbf{c}_u(\mathbf{y}(\mathbf{u}), \mathbf{u}) \mathbf{d} \quad (9a)$$

for $\widehat{\mathbf{w}}$, then solving the projected second order adjoint equation

$$\mathbf{V}^T \mathbf{c}_y(\mathbf{y}(\mathbf{u}), \mathbf{u})^T \mathbf{V} \widehat{\mathbf{p}} = \mathbf{V}^T \nabla_{yy} L(\mathbf{y}(\mathbf{u}), \mathbf{u}, \lambda(\mathbf{u})) \mathbf{V} \widehat{\mathbf{w}} - \mathbf{V}^T \nabla_{yu} L(\mathbf{y}(\mathbf{u}), \mathbf{u}, \lambda(\mathbf{u})) \mathbf{d} \quad (9b)$$

for $\widehat{\mathbf{p}}$, and then computing

$$\begin{aligned} \widetilde{\nabla^2 \widehat{J}}(\mathbf{u}) \mathbf{d} &= \mathbf{c}_u(\mathbf{y}(\mathbf{u}), \mathbf{u})^T \mathbf{V} \widehat{\mathbf{p}} - \nabla_{uy} L(\mathbf{y}(\mathbf{u}), \mathbf{u}, \lambda(\mathbf{u})) \mathbf{V} \widehat{\mathbf{w}} \\ &\quad + \nabla_{uu} L(\mathbf{y}(\mathbf{u}), \mathbf{u}, \lambda(\mathbf{u})) \mathbf{d}. \end{aligned} \quad (9c)$$

The ROM Hessian approximation (9) is used to compute the direction in the inexact Newton method. Instead of solving (7), given a current iterate \mathbf{u}_c the direction $\widetilde{\mathbf{d}}$ is computed by using the truncated CG method to approximately solve

$$\widetilde{\nabla^2 \widehat{J}}(\mathbf{u}_c) \widetilde{\mathbf{d}} = -\nabla \widehat{J}(\mathbf{u}_c). \quad (10)$$

The step size α_c is computed as before using (8) with \mathbf{d} replaced by $\widetilde{\mathbf{d}}$. The new iterate is given by $\mathbf{u}_+ = \mathbf{u}_c + \alpha_c \widetilde{\mathbf{d}}$.

The global convergence of the original line-search Newton method and the line-search Newton method with ROM Hessian approximation can be proven using standard arguments. See, e.g., [15, Theorem 3.2]. While first order global convergence can be ensured under standard conditions if the ROM Hessian approximation is used, the speed with which the inexact Newton method converges in this case depends on the quality of the ROM Hessian approximation, and in particular on \mathbf{V} . Next we motivate our choice of \mathbf{V} .

Basic Properties of the ROM Hessian. Using (6) it can be seen that the Hessian of \widehat{J} is given by

$$\nabla^2 \widehat{J}(\mathbf{u}) = \begin{pmatrix} \mathbf{c}_y(\mathbf{y}, \mathbf{u})^{-1} \mathbf{c}_u(\mathbf{y}, \mathbf{u}) \\ \mathbf{I} \end{pmatrix}^T \begin{pmatrix} \nabla_{yy} L(\mathbf{y}, \mathbf{u}, \lambda) & \nabla_{yu} L(\mathbf{y}, \mathbf{u}, \lambda) \\ \nabla_{uy} L(\mathbf{y}, \mathbf{u}, \lambda) & \nabla_{uu} L(\mathbf{y}, \mathbf{u}, \lambda) \end{pmatrix} \begin{pmatrix} \mathbf{c}_y(\mathbf{y}, \mathbf{u})^{-1} \mathbf{c}_u(\mathbf{y}, \mathbf{u}) \\ \mathbf{I} \end{pmatrix}, \quad (11)$$

where we have set $\mathbf{y} = \mathbf{y}(\mathbf{u})$ and $\lambda = \lambda(\mathbf{u})$ to simplify notation. We will use this simplified notation frequently during the remainder of this section.

Applying the Implicit Function Theorem to the state Eq. (2) shows that the sensitivity of $\mathbf{u} \mapsto \mathbf{y}(\mathbf{u})$ is given by

$$\mathbf{y}_u(\mathbf{u}) = \mathbf{c}_y(\mathbf{y}, \mathbf{u})^{-1} \mathbf{c}_u(\mathbf{y}, \mathbf{u}) \in \mathbb{R}^{n \times m}. \quad (12)$$

Similarly, applying the Implicit Function Theorem to the adjoint Eq. (5a) shows that the derivative of $\mathbf{u} \mapsto \lambda(\mathbf{u})$ is given by

$$\begin{aligned}\lambda_{\mathbf{u}}(\mathbf{u}) &= \mathbf{c}_{\mathbf{y}}(\mathbf{y}, \mathbf{u})^{-T} \left(\nabla_{\mathbf{y}\mathbf{y}}L(\mathbf{y}, \mathbf{u}, \boldsymbol{\lambda}), \nabla_{\mathbf{y}\mathbf{u}}L(\mathbf{y}, \mathbf{u}, \boldsymbol{\lambda}) \right) \begin{pmatrix} \mathbf{c}_{\mathbf{y}}(\mathbf{y}, \mathbf{u})^{-1} \mathbf{c}_{\mathbf{u}}(\mathbf{y}, \mathbf{u}) \\ \mathbf{I} \end{pmatrix} \\ &= \mathbf{c}_{\mathbf{y}}(\mathbf{y}, \mathbf{u})^{-T} \left(\nabla_{\mathbf{y}\mathbf{y}}L(\mathbf{y}, \mathbf{u}, \boldsymbol{\lambda}), \nabla_{\mathbf{y}\mathbf{u}}L(\mathbf{y}, \mathbf{u}, \boldsymbol{\lambda}) \right) \begin{pmatrix} \mathbf{y}_{\mathbf{u}}(\mathbf{u}) \\ \mathbf{I} \end{pmatrix} \in \mathbb{R}^{n \times m}.\end{aligned}\quad (13)$$

Using these derivatives, the Hessian can be written as

$$\nabla^2 \widehat{J}(\mathbf{u}) = \mathbf{c}_{\mathbf{u}}(\mathbf{y}, \mathbf{u})^T \lambda_{\mathbf{u}}(\mathbf{u}) + \left(\nabla_{\mathbf{u}\mathbf{y}}L(\mathbf{y}, \mathbf{u}, \boldsymbol{\lambda}), \nabla_{\mathbf{u}\mathbf{u}}L(\mathbf{y}, \mathbf{u}, \boldsymbol{\lambda}) \right) \begin{pmatrix} \mathbf{y}_{\mathbf{u}}(\mathbf{u}) \\ \mathbf{I} \end{pmatrix}.\quad (14)$$

Using (9) it can be seen that the ROM Hessian approximation is given by

$$\begin{aligned}\widetilde{\nabla^2 \widehat{J}}(\mathbf{u}) &= \left(\mathbf{V} \begin{pmatrix} \mathbf{V}^T \mathbf{c}_{\mathbf{y}}(\mathbf{y}, \mathbf{u}) \mathbf{V} \\ \mathbf{I} \end{pmatrix}^{-1} \mathbf{V}^T \mathbf{c}_{\mathbf{u}}(\mathbf{y}, \mathbf{u}) \right)^T \begin{pmatrix} \nabla_{\mathbf{y}\mathbf{y}}L(\mathbf{y}, \mathbf{u}, \boldsymbol{\lambda}) & \nabla_{\mathbf{y}\mathbf{u}}L(\mathbf{y}, \mathbf{u}, \boldsymbol{\lambda}) \\ \nabla_{\mathbf{u}\mathbf{y}}L(\mathbf{y}, \mathbf{u}, \boldsymbol{\lambda}) & \nabla_{\mathbf{u}\mathbf{u}}L(\mathbf{y}, \mathbf{u}, \boldsymbol{\lambda}) \end{pmatrix} \\ &\quad \times \begin{pmatrix} \mathbf{V} \begin{pmatrix} \mathbf{V}^T \mathbf{c}_{\mathbf{y}}(\mathbf{y}, \mathbf{u}) \mathbf{V} \\ \mathbf{I} \end{pmatrix}^{-1} \mathbf{V}^T \mathbf{c}_{\mathbf{u}}(\mathbf{y}, \mathbf{u}) \end{pmatrix}.\end{aligned}\quad (15)$$

If we define

$$\widetilde{\mathbf{y}}_{\mathbf{u}}(\mathbf{u}) = \mathbf{V} \begin{pmatrix} \mathbf{V}^T \mathbf{c}_{\mathbf{y}}(\mathbf{y}, \mathbf{u}) \mathbf{V} \\ \mathbf{I} \end{pmatrix}^{-1} \mathbf{V}^T \mathbf{c}_{\mathbf{u}}(\mathbf{y}, \mathbf{u}) \in \mathbb{R}^{n \times m}\quad (16)$$

and

$$\widetilde{\lambda}_{\mathbf{u}}(\mathbf{u}) = \mathbf{V} \begin{pmatrix} \mathbf{V}^T \mathbf{c}_{\mathbf{y}}(\mathbf{y}, \mathbf{u}) \mathbf{V} \\ \mathbf{I} \end{pmatrix}^{-T} \mathbf{V}^T \left(\nabla_{\mathbf{y}\mathbf{y}}L(\mathbf{y}, \mathbf{u}, \boldsymbol{\lambda}), \nabla_{\mathbf{y}\mathbf{u}}L(\mathbf{y}, \mathbf{u}, \boldsymbol{\lambda}) \right) \begin{pmatrix} \widetilde{\mathbf{y}}_{\mathbf{u}}(\mathbf{u}) \\ \mathbf{I} \end{pmatrix} \in \mathbb{R}^{n \times m},\quad (17)$$

then the ROM Hessian approximation can be written as

$$\widetilde{\nabla^2 \widehat{J}}(\mathbf{u}) = \mathbf{c}_{\mathbf{u}}(\mathbf{y}, \mathbf{u})^T \widetilde{\lambda}_{\mathbf{u}}(\mathbf{u}) + \left(\nabla_{\mathbf{u}\mathbf{y}}L(\mathbf{y}, \mathbf{u}, \boldsymbol{\lambda}), \nabla_{\mathbf{u}\mathbf{u}}L(\mathbf{y}, \mathbf{u}, \boldsymbol{\lambda}) \right) \begin{pmatrix} \widetilde{\mathbf{y}}_{\mathbf{u}}(\mathbf{u}) \\ \mathbf{I} \end{pmatrix}.\quad (18)$$

The next result ensures the positive definiteness of the Hessian and its ROM approximation in situations that are often encountered in classes of applications, such as those in Sect. 4.

Proposition 1 *Let $\mathbf{V} \in \mathbb{R}^{n \times r}$ have rank $r < n$ and satisfy that $\mathbf{V}^T \mathbf{c}_{\mathbf{y}}(\mathbf{y}(\mathbf{u}), \mathbf{u}) \mathbf{V}$ is invertible. If $\nabla_{\mathbf{u}\mathbf{u}}L(\mathbf{y}, \mathbf{u}, \boldsymbol{\lambda})$ is positive definite, $\nabla_{\mathbf{y}\mathbf{y}}L(\mathbf{y}, \mathbf{u}, \boldsymbol{\lambda})$ is positive semi-definite, and $\nabla_{\mathbf{y}\mathbf{u}}L(\mathbf{y}, \mathbf{u}, \boldsymbol{\lambda}) = \mathbf{0}$, then $\nabla^2 \widehat{J}(\mathbf{u})$ and $\widetilde{\nabla^2 \widehat{J}}(\mathbf{u})$ are positive definite with smallest eigenvalue bounded from below by the smallest eigenvalue of $\nabla_{\mathbf{u}\mathbf{u}}L(\mathbf{y}, \mathbf{u}, \boldsymbol{\lambda})$.*

Proof The statement follows immediately from (11) and (15).

Comparing (14) with (18) shows that $\nabla^2 \widehat{J}(\mathbf{u}) \approx \nabla^2 \widetilde{J}(\mathbf{u})$ if $\mathbf{y}_u(\mathbf{u}) \approx \widetilde{\mathbf{y}}_u(\mathbf{u})$ and $\boldsymbol{\lambda}_u(\mathbf{u}) \approx \widetilde{\boldsymbol{\lambda}}_u(\mathbf{u})$. The latter two approximation conditions motivate our choice for the ROM matrix \mathbf{V} .

Proposition 2 *Let $\mathbf{V} \in \mathbb{R}^{n \times r}$ have rank $r < n$ and satisfy that $\mathbf{V}^T \mathbf{c}_y(\mathbf{y}(\mathbf{u}), \mathbf{u}) \mathbf{V}$ is invertible. If $\mathbf{y}(\mathbf{u} + \mathbf{d}) \in \mathcal{R}(\mathbf{V})$ for all sufficiently small \mathbf{d} , then*

$$\widetilde{\mathbf{y}}_u(\mathbf{u}) = \mathbf{y}_u(\mathbf{u}). \quad (19)$$

If, in addition, $\boldsymbol{\lambda}(\mathbf{u} + \mathbf{d}) \in \mathcal{R}(\mathbf{V})$ for all sufficiently small \mathbf{d} , then

$$\widetilde{\boldsymbol{\lambda}}_u(\mathbf{u}) = \boldsymbol{\lambda}_u(\mathbf{u}) \quad \text{and} \quad \nabla^2 \widetilde{J}(\mathbf{u}) = \nabla^2 \widehat{J}(\mathbf{u}). \quad (20)$$

Proof If $\mathbf{y}(\mathbf{u} + \mathbf{d}) \in \mathcal{R}(\mathbf{V})$ for all sufficiently small \mathbf{d} . Then $\mathbf{y}_u(\mathbf{u}) \mathbf{d} \in \mathcal{R}(\mathbf{V})$ for all \mathbf{d} . Furthermore, for every $\mathbf{u} + \mathbf{d}$ with \mathbf{d} sufficiently small there exists $\widetilde{\mathbf{y}}(\mathbf{u} + \mathbf{d})$ such that $\mathbf{y}(\mathbf{u} + \mathbf{d}) = \mathbf{V} \widetilde{\mathbf{y}}(\mathbf{u} + \mathbf{d})$. Inserting this into the state Eq. (2) implies

$$\mathbf{V}^T \mathbf{c}(\mathbf{V} \widetilde{\mathbf{y}}(\mathbf{u} + \mathbf{d}), \mathbf{u} + \mathbf{d}) = \mathbf{0} \quad \text{for all sufficiently small } \mathbf{d}.$$

Applying the Implicit Function Theorem to this equation shows that

$$\widetilde{\mathbf{y}}_u(\mathbf{u}) = \left(\mathbf{V}^T \mathbf{c}_y(\mathbf{y}(\mathbf{u}), \mathbf{u}) \mathbf{V} \right)^{-1} \mathbf{V}^T \mathbf{c}_u(\mathbf{y}(\mathbf{u}), \mathbf{u})$$

and (16) can be written as $\widetilde{\mathbf{y}}_u(\mathbf{u}) = \mathbf{V} \widetilde{\mathbf{y}}_u(\mathbf{u})$. Since $\mathbf{y}(\mathbf{u} + \mathbf{d}) = \mathbf{V} \widetilde{\mathbf{y}}(\mathbf{u} + \mathbf{d})$ for all sufficiently small \mathbf{d} ,

$$\widetilde{\mathbf{y}}_u(\mathbf{u}) = \mathbf{V} \widetilde{\mathbf{y}}_u(\mathbf{u}) = \mathbf{y}_u(\mathbf{u}),$$

which is (19).

If, in addition, $\boldsymbol{\lambda}(\mathbf{u} + \mathbf{d}) \in \mathcal{R}(\mathbf{V})$ for all sufficiently small \mathbf{d} , then we can use similar argument to show the first equality in (20).

The second identity in (20) follows immediately from (14), (18), (19), and first identity (20).

In general there is no matrix $\mathbf{V} \in \mathbb{R}^{n \times r}$ with small rank $r \ll n$, such that the conditions in Proposition 2 hold. However, if there exists $\mathbf{V} \in \mathbb{R}^{n \times r}$, $r \ll n$, with orthogonal columns such that the projections of $\mathbf{y}(\mathbf{u} + \mathbf{d})$ and $\boldsymbol{\lambda}(\mathbf{u} + \mathbf{d})$ onto $\mathcal{R}(\mathbf{V})$ are close to $\mathbf{y}(\mathbf{u} + \mathbf{d})$ and $\boldsymbol{\lambda}(\mathbf{u} + \mathbf{d})$, respectively, or equivalently such that

$$\|(\mathbf{I} - \mathbf{V} \mathbf{V}^T) \mathbf{y}(\mathbf{u} + \mathbf{d})\| \leq \text{tol} \quad \text{and} \quad \|(\mathbf{I} - \mathbf{V} \mathbf{V}^T) \boldsymbol{\lambda}(\mathbf{u} + \mathbf{d})\| \leq \text{tol} \quad (21)$$

for a small tolerance $\text{tol} \ll 1$, then we expect $\mathbf{y}_u(\mathbf{u}) \approx \widetilde{\mathbf{y}}_u(\mathbf{u})$, $\boldsymbol{\lambda}_u(\mathbf{u}) \approx \widetilde{\boldsymbol{\lambda}}_u(\mathbf{u})$, and consequently that the ROM Hessian $\nabla^2 \widetilde{J}(\mathbf{u})$ is a good approximation of the original FOM Hessian $\nabla^2 \widehat{J}(\mathbf{u})$.

However, finding $\mathbf{V} \in \mathbb{R}^{n \times r}$ such that (21) is guaranteed to hold for *all* sufficiently small \mathbf{d} would be computationally expensive. In practice we compute $\mathbf{V} \in \mathbb{R}^{n \times r}$ using proper orthogonal decomposition (POD) applied to the state $\mathbf{y}(\mathbf{u})$ and the adjoint $\boldsymbol{\lambda}(\mathbf{u})$ at the current control.

ROM Hessians versus ROM Optimization Proposition 2 and even (21) describe optimistic scenarios. If we were able to find a \mathbf{V} such that (21) holds at the current control $\mathbf{u} = \mathbf{u}_c$, then it could be used to approximate the original problem (1, 2) in a trust-region based model management framework and approximately solve (4, 3) over all \mathbf{u} in a neighborhood (trust-region) around the current control \mathbf{u}_c . Rather than using the ROM model $\mathbf{u} \mapsto J(\mathbf{V}\widehat{\mathbf{y}}(\mathbf{u}), \mathbf{u})$ in a neighborhood of the current control \mathbf{u}_c , our approach uses the approximate quadratic Taylor expansion $\mathbf{d} \mapsto \widehat{J}(\mathbf{u}_c) + \nabla \widehat{J}(\mathbf{u}_c)^T \mathbf{d} + \mathbf{d}^T \nabla^2 \widehat{J}(\mathbf{u}_c) \mathbf{d}$ as a model at controls $\mathbf{u} = \mathbf{u}_c + \mathbf{d}$. The true function $\widehat{J}(\mathbf{u})$ and this Taylor model have the same function and gradient value at \mathbf{u}_c , no matter how good \mathbf{V} is. If $\nabla^2 \widehat{J}(\mathbf{u}_c)$ well approximates $\nabla^2 \widehat{J}(\mathbf{u}_c)$, we will essentially compute a Newton step. If the resulting \mathbf{V} is less good, our approach can still generate a descent direction that is much better than a simple gradient step. Computationally inexpensive, but possibly less accurate ROMs can still be used to accelerate the overall optimization.

3 ROM Hessian Approximations for Discrete Time Optimal Control Problems

In this section we apply the ROM Hessian approach to a minimization problem that arises, e.g., from a discretized parabolic optimal control problem.

Model Problem. Given functions $\ell : \mathbb{R}^{n_y} \rightarrow \mathbb{R}$, $\sigma : \mathbb{R}^{n_u} \rightarrow \mathbb{R}$, $\mathbf{F} : \mathbb{R}^{n_y} \rightarrow \mathbb{R}^{n_y}$, and $\mathbf{G} : \mathbb{R}^{n_u} \rightarrow \mathbb{R}^{n_y}$, and given $\mathbf{y}_0 \in \mathbb{R}^{n_y}$ consider the following minimization problem in the variables $\mathbf{u} = (\mathbf{u}_1^T, \dots, \mathbf{u}_{n_t}^T)^T \in \mathbb{R}^{n_u n_t}$, $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_{n_t}^T)^T \in \mathbb{R}^{n_y n_t}$.

$$\text{Minimize } \sum_{k=1}^{n_t} \ell(\mathbf{y}_k) + \sigma(\mathbf{u}_k), \quad (22a)$$

where \mathbf{y} and \mathbf{u} satisfy an implicit constraint,

$$\mathbf{M} \left(\frac{\mathbf{y}_{k+1} - \mathbf{y}_k}{\Delta t} \right) + \mathbf{F}(\mathbf{y}_{k+1}) = \mathbf{G}(\mathbf{u}_{k+1}), \quad k = 0, \dots, n_t - 1, \quad \mathbf{y}_0 \text{ given.} \quad (22b)$$

It is easily possible to extent our ROM Hessian approach to generalizations of (22), e.g., replace $\ell(\mathbf{y}_k)$ and $\mathbf{G}(\mathbf{u}_k)$ by $\ell(\mathbf{y}_k, \mathbf{u}_k)$ and $\mathbf{G}(\mathbf{y}_k, \mathbf{u}_k)$. We consider (22) to simplify notation.

Throughout this section we make the following assumptions, which are adaptations of the assumptions (A1)–(A3) to the problem (22).

- (A1') There exists an open set $D_u \subset \mathbb{R}^{n_u n_t}$ such that for every $\mathbf{u} = (\mathbf{u}_1^T, \dots, \mathbf{u}_{n_t}^T)^T \in D$ the state Eq. (22b) has a unique solution $\mathbf{y}(\mathbf{u}) = (\mathbf{y}_1(\mathbf{u})^T, \dots, \mathbf{y}_{n_t}(\mathbf{u})^T)^T$.
- (A2') There exists an open set $D_y \subset \mathbb{R}^{n_y n_t}$ such that $\{\mathbf{y}(\mathbf{u}) : \mathbf{u} \in D_u\} \subset D_y$, and the functions $\ell : D_y \rightarrow \mathbb{R}$, $\sigma : D_u \rightarrow \mathbb{R}$, $\mathbf{F} : D_y \rightarrow \mathbb{R}^{n_y}$, and $\mathbf{G} : D_u \rightarrow \mathbb{R}^{n_y}$ are twice continuously differentiable.
- (A3') For every $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_{n_t}^T)^T \in \{\mathbf{y}(\mathbf{u}) : \mathbf{u} \in D_u\}$ and every $\mathbf{r}_1, \dots, \mathbf{r}_{n_t} \in \mathbb{R}^{n_y}$ there exists a unique solution $\mathbf{w}_1, \dots, \mathbf{w}_{n_t} \in \mathbb{R}^{n_y}$ of the equations

$$\mathbf{w}_0 = \mathbf{0}, \quad \mathbf{w}_{k+1} - \mathbf{w}_k = \Delta t \mathbf{F}_y(\mathbf{y}_{k+1}) \mathbf{w}_{k+1} + \mathbf{r}_{k+1}, \quad k = 0, \dots, n_t - 1.$$

Under the assumptions (A1')–(A3') we can define the function $\hat{J} : D_u \rightarrow \mathbb{R}$, $\hat{J}(\mathbf{u}) = \sum_{k=1}^K \ell(\mathbf{y}_k(\mathbf{u})) + \sigma(\mathbf{u}_k)$, so that (22), is a special case of (1, 2).

Gradient and Hessian Computation. Under the assumptions (A1')–(A3') the function $\hat{J} : D_u \rightarrow \mathbb{R}$ is twice continuously differentiable. Its gradient can be computed using the adjoint equation approach in Algorithm 1. Hessian-times-vector operations are computed using Algorithm 2.

Algorithm 1: Gradient Computation

- 1: Given $\mathbf{u} = (\mathbf{u}_1^T, \dots, \mathbf{u}_{n_t}^T)^T$ let $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_{n_t}^T)^T$ solve the state Eq. (22b).
- 2: Solve the adjoint equations for $\lambda_{n_t}, \dots, \lambda_1$:

$$[\mathbf{M} + \Delta t \mathbf{F}_y(\mathbf{y}_{n_t})]^T \lambda_{n_t} = -\nabla_y \ell(\mathbf{y}_{n_t}), \quad (23a)$$

$$[\mathbf{M} + \Delta t \mathbf{F}_y(\mathbf{y}_k)]^T \lambda_k = \mathbf{M} \lambda_{k+1} - \nabla_y \ell(\mathbf{y}_k), \quad k = n_t - 1, \dots, 1. \quad (23b)$$

- 3: Compute the gradient

$$\nabla J(\mathbf{u}) = \begin{pmatrix} \nabla_u \sigma(\mathbf{u}_1) - \Delta t \mathbf{G}_u(\mathbf{u}_1)^T \lambda_1 \\ \vdots \\ \nabla_u \sigma(\mathbf{u}_{n_t}) - \Delta t \mathbf{G}_u(\mathbf{u}_{n_t})^T \lambda_{n_t} \end{pmatrix}.$$

Note that since the objective function and the implicit constraints in the model problem (22) are separable, $\nabla_{\mathbf{u}\mathbf{y}} L(\mathbf{y}, \mathbf{u}, \lambda) = \mathbf{0}$.

Hessian Approximation by Model Order Reduction. As mentioned towards the end of Sect. 2, we compute a ROM from state \mathbf{y} and adjoint λ information. In this example the state \mathbf{y} and the adjoint λ are vectors of vectors $\mathbf{y}_k \in \mathbb{R}^{n_y}$ and $\lambda_k \in \mathbb{R}^{n_y}$ respectively. We compute a matrix $\mathbf{V} \in \mathbb{R}^{n_y \times r}$ such that the components \mathbf{y}_k and λ_k are approximately contained in the range of \mathbf{V} (this \mathbf{V} is a component of the projection matrix in Sect. 2) The projection matrix for \mathbf{y} and λ is the $n_t n_y \times n_t r$ block diagonal matrix with identical diagonal blocks given by \mathbf{V} .

Algorithm 2: Hessian-Times-Vector Computation – $\nabla^2 \widehat{J}(\mathbf{u})\mathbf{v}$

- 1: Given $\mathbf{u} = (\mathbf{u}_1^T, \dots, \mathbf{u}_{n_t}^T)^T$ let $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_{n_t}^T)^T$ solve the state Eq. (22b) and let $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^T, \dots, \boldsymbol{\lambda}_{n_t}^T)^T$ solve the adjoint Eq. (23).
- 2: Set $\mathbf{w}_0 = \mathbf{0}$ and solve the linearized state equation $\mathbf{w}_1, \dots, \mathbf{w}_{n_t}$:

$$[\mathbf{M} + \Delta t \mathbf{F}_y(\mathbf{y}_{k+1})]\mathbf{w}_{k+1} = \mathbf{M}\mathbf{w}_k + \Delta t \mathbf{G}_u(\mathbf{u}_{k+1})\mathbf{v}_{k+1}, \quad k = 0, \dots, n_t - 1.$$

- 3: Solve the second order adjoint equations for $\mathbf{p}_{n_t}, \dots, \mathbf{p}_1$.

$$[\mathbf{M} + \Delta t \mathbf{F}_y(\mathbf{y}_{n-t})]^T \mathbf{p}_{n_t} = -[\Delta t (\boldsymbol{\lambda}_{n_t}^T \mathbf{F}(\mathbf{y}_{n_t}))_{yy} + \ell_{yy}(\mathbf{y}_{n_t})]\mathbf{w}_{n_t},$$

$$[\mathbf{M} + \Delta t \mathbf{F}_y(\mathbf{y}_k)]^T \mathbf{p}_k = \mathbf{M}\mathbf{p}_{k+1} - [\Delta t (\boldsymbol{\lambda}_k^T \mathbf{F}(\mathbf{y}_k))_{yy} + \ell_{yy}(\mathbf{y}_k)]\mathbf{w}_k, \quad k = n_t - 1, \dots, 1.$$

- 4: Compute the action of the Hessian

$$\nabla^2 \widehat{J}(\mathbf{u})\mathbf{v} = \begin{pmatrix} -\Delta t \mathbf{G}_u(\mathbf{u}_1)^T \mathbf{p}_1 + [\sigma_{uu}(\mathbf{u}_1) - \Delta t (\boldsymbol{\lambda}_1^T \mathbf{G}(\mathbf{u}_1))_{uu}]\mathbf{v}_1 \\ \vdots \\ -\Delta t \mathbf{G}_u(\mathbf{u}_{n_t})^T \mathbf{p}_{n_t} + [\sigma_{uu}(\mathbf{u}_{n_t}) - \Delta t (\boldsymbol{\lambda}_{n_t}^T \mathbf{G}(\mathbf{u}_{n_t}))_{uu}]\mathbf{v}_{n_t} \end{pmatrix}.$$

We apply POD to the computed state and adjoint to compute $\mathbf{V} \in \mathbb{R}^{n_y \times r}$. Since states and adjoint typically have very different scales, we do not apply POD to the combined snapshots, but individually, as stated in Algorithm 3.

Algorithm 3: Construction of POD Subspace

- 1: Collate the solution $\mathbf{y}_1, \dots, \mathbf{y}_{n_t}$ to the state equation and the solution $\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_{n_t}$ to the adjoint equation into snapshot matrices,

$$\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_{n_t}], \quad \boldsymbol{\Lambda} = [\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_{n_t}].$$

- 2: Construct by truncated SVD matrices $\mathbf{V}_y \in \mathbb{R}^{n_y \times r_y}$ and $\mathbf{V}_\lambda \in \mathbb{R}^{n_y \times r_\lambda}$ so that

$$\frac{\|(\mathbf{I} - \mathbf{V}_y \mathbf{V}_y^T) \mathbf{Y}\|}{\|\mathbf{Y}\|} < \text{tol}_{\text{POD}} \quad \text{and} \quad \frac{\|(\mathbf{I} - \mathbf{V}_\lambda \mathbf{V}_\lambda^T) \boldsymbol{\Lambda}\|}{\|\boldsymbol{\Lambda}\|} < \text{tol}_{\text{POD}}.$$

- 3: Orthogonalize. $\text{orth}([\mathbf{V}_y \ \mathbf{V}_\lambda]) \mapsto \mathbf{V}$.
-

The ROM Hessian-time-vector computation is specified in Algorithm 4. The subspace matrix $\mathbf{V} \in \mathbb{R}^{n_y \times n_r}$ used in steps 2 and 3 is computed via Algorithm 3.

Computational Efficiency of ROM Hessian. While matrices like $\mathbf{V}^T \mathbf{F}_y(\mathbf{y}_k) \mathbf{V}$ in the ROM Hessian computation are smaller than $\mathbf{F}_y(\mathbf{y}_k)$, the dependence of \mathbf{F}_y on \mathbf{y}_k , which changes with time step k makes the computation of $\mathbf{V}^T \mathbf{F}_y(\mathbf{y}_k) \mathbf{V}$ expensive. This well-known issue [4, 6, 9, 11, 18] is addressed via hyperreduction. Specifically, we use a so-called unassembled form of the Discrete Empirical Interpolation Method (DEIM), which originates from the (D)EIM [6, 9] and is described in more detail

Algorithm 4: ROM Hessian-Times-Vector Computation— $\widetilde{\nabla^2 \widehat{J}(\mathbf{u})}$

- 1: Given $\mathbf{u} = (\mathbf{u}_1^T, \dots, \mathbf{u}_{n_t}^T)^T$ let $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_{n_t}^T)^T$ be the solution of the state Eq. (22b) and let $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^T, \dots, \boldsymbol{\lambda}_{n_t}^T)^T$ be the solution of the adjoint Eq. (23).
 2: Set $\widehat{\mathbf{w}}_0 = \mathbf{0}$ and solve the POD linearized state equation for $\widehat{\mathbf{w}}_k$,

$$[\mathbf{I} + \Delta t \mathbf{V}^T \mathbf{F}_y(\mathbf{y}_{k+1}) \mathbf{V}] \widehat{\mathbf{w}}_{k+1} = \widehat{\mathbf{w}}_k + \Delta t \mathbf{V}^T \mathbf{G}_u(\mathbf{u}_{k+1}) \mathbf{v}_{k+1}, \quad k = 0, \dots, n_t - 1.$$

- 3: Solve the POD second order adjoint equation for $\widehat{\mathbf{p}}_k$,

$$\begin{aligned} [\mathbf{I} + \Delta t \mathbf{V}^T \mathbf{F}_y(\mathbf{y}_{n_t}) \mathbf{V}]^T \widehat{\mathbf{p}}_{n_t} &= -\mathbf{V}^T [\Delta t (\boldsymbol{\lambda}_{n_t}^T \mathbf{F}(\mathbf{y}_{n_t}))_{yy} + \ell_{yy}(\mathbf{y}_{n_t})] \mathbf{V} \widehat{\mathbf{w}}_{n_t}, \\ [\mathbf{I} + \Delta t \mathbf{V}^T \mathbf{F}_y(\mathbf{y}_k) \mathbf{V}]^T \widehat{\mathbf{p}}_k &= \widehat{\mathbf{p}}_{k+1} - \mathbf{V}^T [\Delta t (\boldsymbol{\lambda}_k^T \mathbf{F}(\mathbf{y}_k))_{yy} + \ell_{yy}(\mathbf{y}_k)] \mathbf{V} \widehat{\mathbf{w}}_k, \\ & \quad k = n_t - 1, \dots, 1. \end{aligned}$$

- 4: Compute the approximation to the action of the Hessian,

$$\widetilde{\nabla^2 \widehat{J}(\mathbf{u})} \mathbf{v} = \begin{pmatrix} -\Delta t \mathbf{G}_u(\mathbf{u}_1)^T \mathbf{V} \widehat{\mathbf{p}}_1 + [\sigma_{uu}(\mathbf{u}_1) - \Delta t (\boldsymbol{\lambda}_1^T \mathbf{G}(\mathbf{u}_1))_{uu}] \mathbf{v}_1 \\ \vdots \\ -\Delta t \mathbf{G}_u(\mathbf{u}_{n_t})^T \mathbf{V} \widehat{\mathbf{p}}_{n_t} + [\sigma_{uu}(\mathbf{u}_{n_t}) - \Delta t (\boldsymbol{\lambda}_1^T \mathbf{G}(\mathbf{u}_1))_{uu}] \mathbf{v}_{n_t} \end{pmatrix}$$

in [4, 18]. This leads to a further approximation of the ROM Hessian computed by Algorithm 4. The error in this additional approximation is controlled by the tolerance tol_{DEIM} used in DEIM. We refer to [14, Sects. 5.4, 6.2] for further details. We will use $\widetilde{\nabla^2 \widehat{J}(\mathbf{u})}$ to denote the final ROM Hessian approximation. We can replace $\mathbf{y}_k \approx \mathbf{V} \widehat{\mathbf{y}}_k$, where $\widehat{\mathbf{y}}_k = \mathbf{V}^T \mathbf{y}_k \in \mathbb{R}^r$, and $\boldsymbol{\lambda}_k \approx \mathbf{V} \widehat{\boldsymbol{\lambda}}_k$, where $\widehat{\boldsymbol{\lambda}}_k = \mathbf{V}^T \boldsymbol{\lambda}_k \in \mathbb{R}^r$ to reduce storage requirements.

4 Numerical Results

We apply the ROM Hessian approximation to semi-linear parabolic optimal control problems of the form

$$\min_{u \in L^2(\Omega \times (0,1))} \frac{1}{2} \int_0^1 \int_{\Omega} (y(\mathbf{x}, t; u) - y_d(\mathbf{x}, t))^2 d\mathbf{x} dt + \frac{\alpha}{2} \int_0^1 \int_{\Omega} u(\mathbf{x}, t)^2 d\mathbf{x} dt, \quad (25a)$$

where for given function $u \in L^2(\Omega \times (0, 1))$ the function $y(\cdot, \cdot; u)$ is the solution of

$$y_t(\mathbf{x}, t) - \Delta y(\mathbf{x}, t) + f(y(\mathbf{x}, t)) = u(\mathbf{x}, t) \quad (\mathbf{x}, t) \in \Omega \times (0, T), \quad (25b)$$

$$y(\mathbf{x}, t) = 0 \quad (\mathbf{x}, t) \in \partial\Omega \times (0, T), \quad (25c)$$

$$y(\mathbf{x}, 0) = y_0(\mathbf{x}) \quad \mathbf{x} \in \Omega, \quad (25d)$$

Table 1 Example 1. Optimization using FOM Hessian (left) and ROM Hessian $\text{tol}_{\text{POD}} = 10^{-3}$, $\text{tol}_{\text{DEIM}} = 10^{-3}$ (right). Stopping criteria in both approaches $\|\nabla \widehat{J}(\mathbf{u})\| < 10^{-7}$. Both optimization approaches show nearly identical convergence behavior. Iteration i , objective function value ($\|\widehat{J}(\mathbf{u})\|$), gradient norm ($\|\nabla \widehat{J}(\mathbf{u})\|$), step-size (α), and number of CG iterations (CG) are shown

| i | $\widehat{J}(\mathbf{u}_i)$ | $\ \nabla \widehat{J}(\mathbf{u}_i)\ $ | α_i | CG |
|-----|-----------------------------|--|------------|----|
| 0 | 4.397932 | 2.7878e-04 | 1.00 | 18 |
| 1 | 1.665062 | 2.5098e-05 | 1.00 | 20 |
| 2 | 1.598802 | 4.7573e-06 | 1.00 | 25 |
| 3 | 1.595127 | 2.9272e-07 | 1.00 | 35 |
| 4 | 1.595110 | 1.4527e-09 | | |
| i | $\widehat{J}(\mathbf{u}_i)$ | $\ \nabla \widehat{J}(\mathbf{u}_i)\ $ | α_i | CG |
| 0 | 4.397932 | 2.7878e-04 | 1.00 | 18 |
| 1 | 1.665253 | 2.5120e-05 | 1.00 | 20 |
| 2 | 1.598817 | 4.7625e-06 | 1.00 | 25 |
| 3 | 1.595127 | 2.9294e-07 | 1.00 | 39 |
| 4 | 1.595110 | 1.7659e-08 | | |

and $\Omega = (0, 1)^2$. The problem is discretized in space using piecewise linear finite elements on a triangulation obtained by dividing $\Omega = (0, 1)^2$ into 40×40 squares and then dividing each square into two triangles. This results in a semi-discretization with $n_u = 1681$ and $n_y = 1521$ degrees of freedom in the control and state respectively. The resulting semi-discretization is then discretized in time using the backward Euler method using $n_t = 100$ time steps. This results in a problem of the type (22) with $\mathbf{G}(\mathbf{u}_k) = \mathbf{M}_u \mathbf{u}_k$ and $\sigma(\mathbf{u}_k) = (\alpha/2) \mathbf{u}_k^T \mathbf{M}_u \mathbf{u}_k$, where \mathbf{M}_u is the mass matrix, and $\ell(\mathbf{y}_k) = (\mathbf{y}_k - \mathbf{y}_{k,d})^T \mathbf{M}(\mathbf{y}_k - \mathbf{y}_{k,d})$, where \mathbf{M} is the mass matrix that also appears in (22b) (\mathbf{M}_u and \mathbf{M} differ in size because of boundary conditions for y) and $\mathbf{y}_{k,d}$ comes from a discretization of the desired state y_d in (25a).

All optimization runs are initialized at $\mathbf{u} = \mathbf{0}$. The truncated CG is stopped when negative curvature is detected or when the CG residual satisfies

$$\|\mathbf{H}_i \mathbf{d} + \nabla \widehat{J}(\mathbf{u}_i)\| \leq \min\{\|\nabla \widehat{J}(\mathbf{u}_i)\|^2, 0.01 \|\nabla J(\mathbf{u}_i)\|\},$$

$\mathbf{H}_i = \nabla^2 \widehat{J}(\mathbf{u}_i)$ or $\widetilde{\nabla^2 \widehat{J}(\mathbf{u}_i)}$ depending on whether a FOM or ROM Hessian is used.

Example 1: Cubic Reaction. In the first example the nonlinearity in (25b) is cubic, $f(y) = y^3$. The desired state is $y_d(\mathbf{x}, t) = 2e^t + 2\mathbf{x}_1(\mathbf{x}_1 - 1) + 2\mathbf{x}_2(\mathbf{x}_2 - 1)$, the initial state is $y_0(\mathbf{x}, t) = \sin(2\pi \mathbf{x}_1)$, and the control penalty is $\alpha = 10^{-4}$.

Both optimization with FOM Hessians and with ROM Hessians converged to virtually the same solution. The computed optimal control u at $t = 0.1, 0.5, 1$ is shown in Fig. 2. Table 1 shows that the optimization histories for the two approaches are nearly the same.

While the optimization histories for the two approaches are nearly the same, the ROM Hessian approach is much faster as shown in Table 2. The timing reported

Table 2 Example 1. The FOM Hessian approach requires $2 \cdot 98$ PDE solves, which are replaced by $2 \cdot 102$ ROM PDE solves in the ROM Hessian approach, resulting in an overall speedup of 5.5

| | FOM | ROM |
|----------------|--------|-------|
| State solves | 5 | 5 |
| Adjoint solves | 5 | 5 |
| Hess-Vec mult | 98 | 102 |
| Total time (s) | 410.42 | 74.13 |

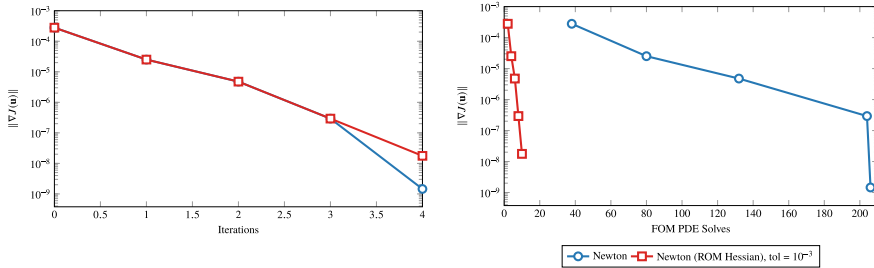


Fig. 1 Convergence history for Newton-CG with and without POD+DEIM approximations in the Hessian-vector computation

in Table 2 is for computations using MATLAB running on a MacBook Pro (13-inch, Retina, Mid 2014). The difference in computing time is due to the cost of Hessian-times-vector multiplications. In this example, using FOM Hessian requires 98 Hessian-times-vector multiplications and therefore 196 linear discretized PDE solves. In contrast, using ROM Hessian requires 102 Hessian-times-vector multiplications and therefore 204 linear ROM PDE solves.

Figure 1 is another presentation of the convergence results in Tables 1 and 2.

The left plot shows that both approaches have nearly the same iteration history. However, when convergence history is plotted against computational work performed, measured in terms of PDE solves, the ROM Hessian approach is much faster. Note that here we do not differentiate between the nonlinear state PDE solve and the linear PDE solves needed for gradient, adjoint and FOM Hessian-times-vector computations. While the discretized state Eq. (22b) is nonlinear and computing y_{k+1} is performed via Newton’s method, only 1–2 Newton iterations per time step in the state computation are needed in this example. Therefore the difference between a nonlinear state PDE solve and linear PDE solve is small.

Example 2: Solid Fuel Ignition Model. This example is modeled after [8]. The nonlinearity in (25b) is $f(y) = -\delta e^y$ with $\delta = 5$. The desired state is $y_d(\mathbf{x}, t) = \pi^{-2} \sin(\pi \mathbf{x}_1) \sin(\pi \mathbf{x}_2)$, the initial state is $y_0 \equiv \mathbf{0}$, and the penalty is $\alpha = 5 \cdot 10^{-3}$.

The numerical results for this example mirror those of the previous example. Optimization with FOM Hessians and with ROM Hessians converged to virtually the same solution, and the optimization histories for the two approaches are nearly

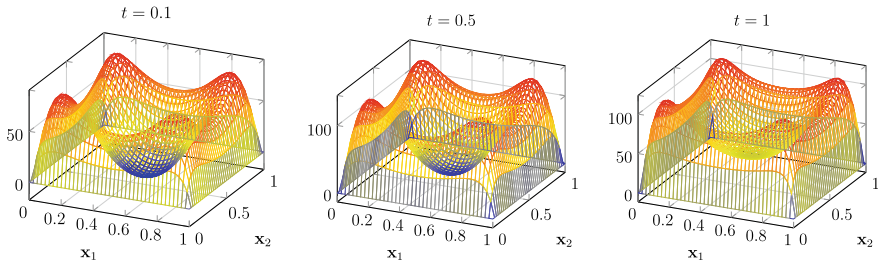


Fig. 2 Optimal control u at $t = 0.1, 0.5, 1$ for Example 1

Table 3 Example 2. Optimization using FOM Hessian (left) and ROM Hessian $\text{tol}_{\text{POD}} = 10^{-3}$, $\text{tol}_{\text{DEIM}} = 10^{-3}$ (right). Stopping criteria in both approaches $\|\nabla \widehat{J}(\mathbf{u})\| < 10^{-8}$. Both optimization approaches show nearly identical convergence behavior. Iteration i , objective function value ($\|\widehat{J}(\mathbf{u})\|$), gradient norm ($\|\nabla \widehat{J}(\mathbf{u})\|$), step-size (α), and number of CG iterations (CG) are shown

| i | $\widehat{J}(\mathbf{u}_i)$ | $\ \nabla \widehat{J}(\mathbf{u}_i)\ $ | α_i | CG |
|-----|-----------------------------|--|------------|----|
| 0 | 2.6981e-02 | 4.5436e-05 | 1.00 | 9 |
| 1 | 1.3028e-02 | 3.8587e-06 | 1.00 | 12 |
| 2 | 1.2917e-02 | 2.6076e-08 | 1.00 | 27 |
| 3 | 1.2917e-02 | 1.2552e-12 | | |
| i | $\widehat{J}(\mathbf{u}_i)$ | $\ \nabla \widehat{J}(\mathbf{u}_i)\ $ | α_i | CG |
| 0 | 2.6981e-02 | 4.5436e-05 | 1.00 | 9 |
| 1 | 1.3028e-02 | 3.8601e-06 | 1.00 | 12 |
| 2 | 1.2917e-02 | 2.6116e-08 | 1.00 | 28 |
| 3 | 1.2917e-02 | 2.4637e-12 | | |

Table 4 Example 2. The FOM Hessian approach requires 2×48 PDE solves, which are replaced by 2×49 ROM PDE solves in the ROM Hessian approach, resulting in an overall speedup of 4

| | FOM | ROM |
|----------------|--------|-------|
| State solves | 4 | 4 |
| Adjoint solves | 4 | 4 |
| Hess-Vec mult | 48 | 49 |
| Total time (s) | 234.78 | 56.73 |

the same, as shown in Table 3. The computed optimal control u at $t = 0.1, 0.5, 1$ is shown in Fig. 4.

The ROM Hessian approach is much faster as shown in Table 4. Again, the timing reported in Table 4 is for computations using MATLAB running on a MacBook Pro (13-inch, Retina, Mid 2014) and the reason for the difference in computing time is the cost of Hessian-times-vector multiplications. The ROM Hessian requires 96 ROM PDE solves for ROM Hessian-vector operations in contrast to the 96 FOM PDE solves in the FOM Hessian approach.

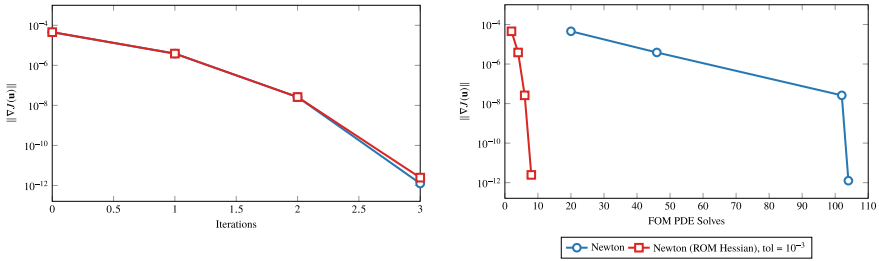


Fig. 3 Convergence history for Newton-CG with and without POD+DEIM approximations in the Hessian-vector computation

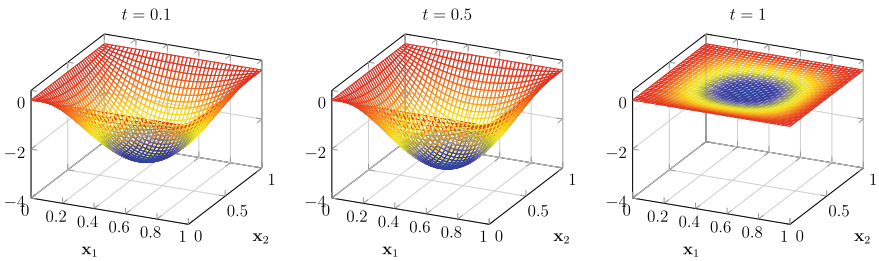


Fig. 4 Optimal control u at $t = 0.1, 0.5, 1$ for Example 2

Figure 3 shows that when convergence history is plotted not against iteration count, but against computational work performed, measured in terms of PDE solves, the ROM Hessian approach is much faster.

5 Conclusions

We have introduced ROM Hessian approximations for use in inexact Newton methods for large-scale smooth optimization problems obtained from discretizations of optimal control problems governed by parabolic PDEs. In contrast to other ROM approaches, which approximate the optimization problem, our approach retains the original FOM objective function and gradient. The Hessian ROM is computed by applying POD to state and adjoint snapshots that have to be computed anyway, and therefore ROM computation is relatively inexpensive. Since original objective functions and gradients are used, the qualitative global convergence properties of the original line-search Newton method and the line-search Newton method with ROM Hessian approximation are the same. However, since Hessian approximations are significantly cheaper, the ROM Hessian approach can lead to substantial savings. This is confirmed by numerical experiments with two semilinear parabolic optimal control problems. The two optimization approaches had essentially the same convergence behavior, but the ROM Hessian approach had a factor 5–8 speedup because

of computational savings due to the Hessian approximations. However, it is possible to construct examples, where the ROM Hessian approach does not lead to computational savings. When the ROM Hessian too poorly approximates the true one, the ROM Hessian approach can require substantially more optimization iterations, which erase savings enjoyed within an optimization iteration. Additional analysis and numerical tests are part of future work.

Acknowledgements The research in this paper was performed before Caleb Magruder joined MathWorks. This research was funded in part through a sponsored research agreement with the ExxonMobil Upstream Research Company and by NSF grants CCF-1816219 and DMS-1819144.

References

1. Afanasiev, K., Hinze, M.: Adaptive control of a wake flow using proper orthogonal decomposition. In: *Shape Optimization and Optimal Design* (Cambridge, 1999), Lecture Notes in Pure and Appl. Math., vol. 216, pp. 317–332. Dekker, New York (2001)
2. Antil, H., Heinkenschloss, M., Hoppe, R.H.W.: Domain decomposition and balanced truncation model reduction for shape optimization of the Stokes system. *Optim. Methods Softw.* **26**(4–5), 643–669 (2011). <http://dx.doi.org/10.1080/10556781003767904>
3. Antil, H., Heinkenschloss, M., Hoppe, R.H.W., Sorensen, D.C.: Domain decomposition and model reduction for the numerical solution of PDE constrained optimization problems with localized optimization variables. *Comput. Vis. Sci.* **13**, 249–264 (2010). <https://doi.org/10.1007/s00791-010-0142-4>
4. Antil, H., Heinkenschloss, M., Sorensen, D.C.: Application of the discrete empirical interpolation method to reduced order modeling of nonlinear and parametric systems. In: Quarteroni, A., Rozza, G. (eds.) *Reduced Order Methods for Modeling and Computational Reduction*, MS&A. Model. Simul. Appl., vol. 9, pp. 101–136. Springer Italia, Milan (2014). https://doi.org/10.1007/978-3-319-02090-7_4
5. Antoulas, A.C.: *Approximation of Large-Scale Dynamical Systems*. Advances in Design and Control, vol. 6. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (2005). <https://doi.org/10.1137/1.9780898718713>
6. Barrault, M., Maday, Y., Nguyen, N.D., Patera, A.T.: An ‘empirical interpolation’ method: application to efficient reduced-basis discretization of partial differential equations. *C. R. Math. Acad. Sci. Paris* **339**(9), 667–672 (2004). <https://doi.org/10.1016/j.crma.2004.08.006>
7. Benner, P., Sachs, E., Volkwein, S.: Model order reduction for PDE constrained optimization. In: Leugering, G., Benner, P., Engell, S., Griewank, A., Harbrecht, H., Hinze, M., Rannacher, R., Ulbrich, S. (eds.) *Trends in PDE Constrained Optimization*, Internat. Ser. Numer. Math., vol. 165, pp. 303–326. Birkhäuser/Springer, Cham (2014). https://doi.org/10.1007/978-3-319-05083-6_19
8. Borzi, A., Kunisch, K.: The numerical solution of the steady state solid fuel ignition model and its optimal control. *SIAM J. Sci. Comput.* **22**(1), 263–284 (electronic) (2000). <https://doi.org/10.1137/S1064827599360194>
9. Chaturantabut, S., Sorensen, D.C.: Nonlinear model reduction via discrete empirical interpolation. *SIAM J. Sci. Comput.* **32**(5), 2737–2764 (2010). <https://doi.org/10.1137/090766498>
10. Fahl, M., Sachs, E.: Reduced order modelling approaches to PDE–constrained optimization based on proper orthogonal decomposition. In: Biegler, L.T., Ghattas, O., Heinkenschloss, M., van Bloemen Waanders, B. (eds.) *Large-Scale PDE–Constrained Optimization*, Lecture Notes in Computational Science and Engineering, vol. 30, pp. 268–280. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-642-55508-4_16

11. Farhat, C., Chapman, T., Avery, P.: Structure-preserving, stability, and accuracy properties of the energy-conserving sampling and weighting method for the hyper reduction of nonlinear finite element dynamic models. *Int. J. Numer. Methods Eng.* **102**(5), 1077–1110 (2015). <https://doi.org/10.1002/nme.4820>
12. Hinze, M., Pinnau, R., Ulbrich, M., Ulbrich, S.: *Optimization with PDE Constraints. Mathematical Modelling, Theory and Applications*, vol. 23. Springer, Heidelberg, New York, Berlin (2009). <https://doi.org/10.1007/978-1-4020-8839-1>
13. Kouri, D.P., Heinkenschloss, M., Ridzal, D., van Bloemen Waanders, B.G.: Inexact objective function evaluations in a trust-region algorithm for PDE-constrained optimization under uncertainty. *SIAM J. Sci. Comput.* **36**(6), A3011–A3029 (2014). <https://doi.org/10.1137/140955665>
14. Magruder, C.: *Projection-based model reduction in the context of optimization with implicit PDE constraints*. Ph.D. thesis, Department of Computational and Applied Mathematics, Rice University, Houston, TX (2017)
15. Nocedal, J., Wright, S.J.: *Numerical Optimization*, 2nd edn. Springer Verlag, Berlin, Heidelberg, New York (2006). <https://doi.org/10.1007/978-0-387-40065-5>
16. Qian, E., Grepl, M.A., Veroy, K., Wilcox, K.: A certified trust region reduced basis approach to PDE-constrained optimization. *SIAM J. Sci. Comput.* **39**(5), S434–S460 (2017). <https://doi.org/10.1137/16M1081981>
17. Sachs, E.W., Volkwein, S.: POD-Galerkin approximations in PDE-constrained optimization. *GAMM-Mitteilungen* **33**(2), 194–208 (2010). <https://doi.org/10.1002/gamm.201010015>
18. Tiso, P., Rixen, D.J.: Discrete empirical interpolation method for finite element structural dynamics. In: Kerschen, G., Adams, D., Carrella, A. (eds.) *Topics in Nonlinear Dynamics*, vol. 1, Conference Proceedings of the Society for Experimental Mechanics Series, vol. 35, pp. 203–212. Springer New York (2013). https://doi.org/10.1007/978-1-4614-6570-6_18
19. Yue, Y., Meerbergen, K.: Accelerating optimization of parametric linear systems by model order reduction. *SIAM J. Optim.* **23**(2), 1344–1370 (2013). <https://doi.org/10.1137/120869171>
20. Zou, Z., Kouri, D.P., Aquino, W.: An adaptive local reduced basis method for solving PDEs with uncertain inputs and evaluating risk. *Comput. Methods Appl. Mech. Engrg.* **345**, 302–322 (2019). <https://doi.org/10.1016/j.cma.2018.10.028>

Interpolation-Based Irrational Model Control Design and Stability Analysis



Charles Poussot-Vassal, Pauline Kergus, and Pierre Vuillemin

Abstract The versatility of data-driven approximation by interpolatory methods, originally settled for model approximation purpose, is illustrated in the context of linear controller design and stability analysis of irrational models. To this aim, following an academic driving example described by a linear partial differential equation, it is shown how the Loewner-based interpolation may be an essential ingredient for control design and stability analysis. More specifically, the interpolatory framework is first used to approximate the irrational model by a rational one that can be used for model-based control, and secondly, it is used for direct data-driven control design, showing equivalent results. Finally, this interpolation framework is employed for estimating the stability of the interconnection of the irrational model with a rational controller.

Keywords Interpolation · Loewner · Data-driven · Control · Stability

1 Introduction and Problem Statement

1.1 Motivations for Interpolation as a Pivotal Tool and Driving Example

Modelling, simulation, control and analysis of irrational models such as those described by linear Partial Differential Equations (**PDE**), are challenging tasks for many practitioners. Indeed, standard numerical tools developed for the rational function case are not tailored in the irrational setting and require dedicated attention (e.g.

C. Poussot-Vassal (✉) · P. Kergus · P. Vuillemin
ONERA/DTIS, Université de Toulouse, 31055 Toulouse, France
e-mail: Charles.Poussot-Vassal@onera.fr

P. Kergus
e-mail: pauline.kergus@cnrs.fr

P. Vuillemin
e-mail: Pierre.Vuillemin@onera.fr

eigenvalues, time-domain simulation ...). In practice, engineers are often requested to discretise the irrational (infinite-dimensional) model before deploying all the numerical tools they dispose of. Beside being time consuming, this step may introduce numerical errors and lead to an iterative procedure between the control design and the model construction teams. Indeed, in an industrial context, the modelling—finite element approximation—and the control design tasks may be split in different teams and iterations to choose the “good” level of modelling might become an issue.

In this chapter, the control synthesis for irrational (infinite-dimensional) linear dynamical models is firstly considered. More specifically, the problem of synthesising a rational control law achieving some performances is addressed through the lens of model interpolatory features. Second, the stability estimation of the closed-loop interconnection of the original linear dynamical irrational model with the synthesised linear rational controller is also done using interpolatory-like methods and approximation-oriented arguments. These arguments (tailored to linear systems only) are illustrated through a single driving example involving a transport equation controlled at a boundary. It is modelled by (1), a linear **PDE** with constant coefficients (such equation set representing a first order linear transport equation, may be used to represent a simplified one dimensional wave equation in telecommunication, traffic jam, *etc.*),

$$\begin{aligned}
 \frac{\partial \tilde{\mathbf{y}}(x,t)}{\partial x} + 2x \frac{\partial \tilde{\mathbf{y}}(x,t)}{\partial t} &= 0 && \text{(transport equation)} \\
 \tilde{\mathbf{y}}(x, 0) &= 0 && \text{(initial condition)} \\
 \tilde{\mathbf{y}}(0, t) &= \frac{1}{\sqrt{t}} * \tilde{\mathbf{u}}_f(0, t) && \text{(control input)} \\
 \frac{\omega_0^2}{s^2 + m\omega_0 s + \omega_0^2} \mathbf{u}(0, s) &= \mathbf{u}_f(0, s) && \text{(actuator model),}
 \end{aligned} \tag{1}$$

where $x \in [0 L]$, $L = 3$ is the space variable and $\omega_0 = 3$ and $m = 0.5$ are the input filter parameters. The scalar input of the model is $\tilde{\mathbf{u}}(0, t)$ (or $\mathbf{u}(0, s)$), the vertical force applied at the left boundary. Applying the Laplace transform, one obtains

$$\frac{\partial \mathbf{y}(x, s)}{\partial x} + 2x (s\mathbf{y}(x, s) - \tilde{\mathbf{y}}(x, 0)) = 0, \tag{2}$$

which solution can be given as $\mathbf{y}(x, s) = a(s)e^{\int -2x s dx} = a(s)e^{-x^2 s}$. Due to boundary condition $\tilde{\mathbf{y}}(0, t) = \frac{1}{\sqrt{t}} * \tilde{\mathbf{u}}_f(t)$, we have $\mathbf{y}(0, s) = \frac{\sqrt{\pi}}{\sqrt{s}} \mathbf{u}_f(s)$, and consequently $a(s) = \frac{\sqrt{\pi}}{\sqrt{s}} \mathbf{u}_f(s)$. The transfer function from the input $\mathbf{u}(0, s)$ to the output $\mathbf{y}(x, s)$ reads

$$\mathbf{y}(x, s) = \frac{\sqrt{\pi}}{\sqrt{s}} e^{-x^2 s} \frac{\omega_0^2}{s^2 + m\omega_0 s + \omega_0^2} \mathbf{u}(s) = \mathbf{H}(x, s) \mathbf{u}(0, s). \tag{3}$$

Relation (3) links the (left boundary) input to the output through an irrational transfer function $\mathbf{H}(x, s)$ for any x value.¹ In addition, for the control design purpose,

¹ The exact time-domain solution of (1) along x is given by $\tilde{\mathbf{y}}(x, t) = \tilde{\mathbf{u}}_f^{t-x^2} / \sqrt{t}$, where $\tilde{\mathbf{u}}_f$ is the output of the second order actuator transfer function, in response to \mathbf{u} .

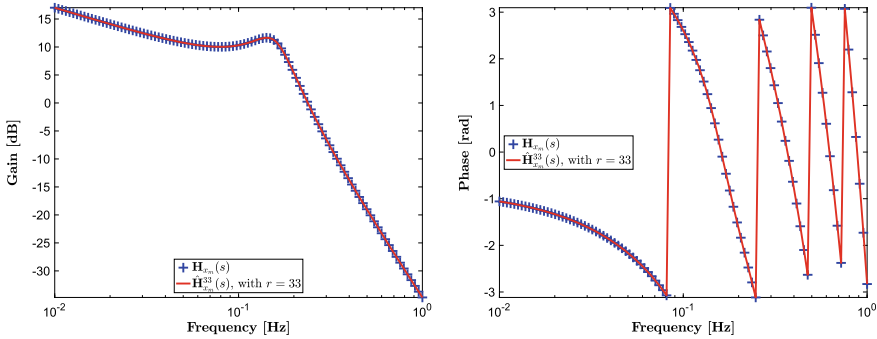


Fig. 1 Frequency response gain (left) and phase (right) of (4) (blue +) and its rational approximation of order $r = 33$ (solid red line)

let us now consider that one single sensor is available, and is located at $x_m = 1.9592$ along the x -axis.² The transfer from the same input $\mathbf{u}(0, s)$ to $\mathbf{y}_{x_m}(s) = \mathbf{y}(x_m, s)$ then reads:

$$\mathbf{y}_{x_m}(s) = \mathbf{H}_{x_m}(s)\mathbf{u}(0, s) \tag{4}$$

As an illustration of the above transfer function, Fig. 1 shows the frequency and phase responses at point $x = x_m$ (blue +). In addition, the rational approximation obtained by Loewner interpolation is also reported (solid red line).

1.2 Problem Statement, Contribution and Organisation

Given a meromorphic (rational or irrational) transfer function \mathbf{H} , given as in (3), the objective is to design a (low order) rational feedback controller \mathbf{K} , such that the $\{\mathbf{H}, \mathbf{K}\}$ interconnection leads to a stable closed-loop and achieves some frequency-oriented performances. The challenge is to suggest an approach that fits to transfer function \mathbf{H} being either rational or irrational.

In this chapter, we aim at illustrating how the Loewner framework may be a pivotal tool for solving approximation problems, but also control and stability issues.

The remainder of the chapter is organised as follows: Sect. 2 recalls some generalities on the Loewner framework as an interpolation tool. The two proposed Loewner-driven control design methodologies are then gathered in Sect. 3. The latter presents

² In the rest of the chapter, x will be discretised with 50 points from 0 to $L = 3$, and x_m has been chosen to be located at $x(\lfloor 50 \times 2/3 \rfloor) = x(33)$, the 33-th element of x .

both the standard approximate and control approach and the data-driven control approach rooted on the Loewner framework. Both are shown to be comparable and even equivalent. Finally, in Sect. 4, the stability of irrational models is addressed through the lens of the Loewner tool. It is applied on the interconnection of the controllers obtained in Sect. 3 and the irrational model (3).

1.3 Notations and Preliminaries

Along the chapter, the following notations are employed: we denote \mathcal{H}_2 (resp. \mathcal{H}_∞), the open subspace of \mathcal{L}_2 (resp. \mathcal{L}_∞) with matrix-valued function $\mathbf{H}(s)$ with n_y outputs, n_u inputs, $\forall s \in \mathbb{C}$, which are analytic in $\mathbf{Re}(s) > 0$ (resp. $\mathbf{Re}(s) \geq 0$). Mathematically, the \mathcal{L}_2 space is a vector-space of matrix valued functions defined on the imaginary axis satisfying

$$\int_{\mathbb{R}} \text{tr}(\overline{\mathbf{H}(j\omega)} \mathbf{H}(j\omega)^T) d\omega < \infty.$$

The \mathcal{L}_∞ one considers functions defined over \mathbb{C}_+ satisfying

$$\sup_{\omega} \|\mathbf{H}(j\omega)\|_2 < \infty.$$

Moreover \mathcal{H}_2 and \mathcal{H}_∞ spaces consider analytic functions over the right half plane. The rational functions of the \mathcal{H}_∞ space are denoted \mathcal{RH}_∞ . A more detailed definition is given in the books [2, 14].

Continuous **MIMO LTI** dynamical model (or system) Σ is defined as an “input-output” map associating an input signal \mathbf{u} to an output one \mathbf{y} by means of the convolution operation, defined as $\mathbf{y}(t) = \int_{-\infty}^{\infty} \mathbf{h}(t - \tau) \mathbf{u}(\tau) d\tau = \mathbf{h}(t) * \mathbf{u}(t)$, where $\mathbf{h}(t)$ is the impulse response of the system Σ . It is (strictly) causal if and only if $\mathbf{h}(t) = 0$ for $(t \leq 0) t < 0$. Then, by taking the Laplace transform of the causal convolution product above defined, one obtains

$$\mathbf{y}(s) = \mathbf{H}(s) \mathbf{u}(s), \tag{5}$$

where $\mathbf{u}(s)$ and $\mathbf{y}(s)$ are the Laplace transform of $\mathbf{u}(t)$ and $\mathbf{y}(t)$. The $n_y \times n_u$ complex-valued matrix function $\mathbf{H}(s)$ is the transfer function of the **LTI** model. An **LTI** system Σ is said to be stable if and only if its transfer function \mathbf{H} is bounded and analytic on \mathbb{C}_+ , *i.e.* it has no singularities on the closed right half-plane. Conversely, it is said to be anti-stable if and only if its transfer function is bounded and analytic on \mathbb{C}_- (see also [7] or Chap. 2 of [15] for more details).

In the case where \mathbf{H} is rational, it has a finite number of singularities and can be represented by a first order descriptor realisation $\mathcal{S} : (E, A, B, C, D)$ with n_u inputs, n_y outputs and n internal variables. The model is then given by:

$$E\dot{\mathbf{x}}(t) = A\mathbf{x}(t) + B\mathbf{u}(t), \mathbf{y}(t) = C\mathbf{x}(t) + D\mathbf{u}(t) \quad (6)$$

where, $\mathbf{x}(t) \in \mathbb{R}^n$ denotes the internal variables (the state variables if E is invertible), and $\mathbf{u}(t) \in \mathbb{R}^{n_u}$ and $\mathbf{y}(t) \in \mathbb{R}^{n_y}$ are the input, output functions, respectively, while $E, A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times n_u}$, $C \in \mathbb{R}^{n_y \times n}$ and $D \in \mathbb{R}^{n_y \times n_u}$, are constant matrices. If the matrix pencil (E, A) is regular, $\mathbf{H}(s) = C(sE - A)^{-1}B + D$, is called the transfer function associated to the realisation \mathcal{S} of the system Σ .

2 Background on Data-Driven LTI Model Approximation

Let us recall the main tool involved in this chapter, namely, the Loewner matrices. First, we define the connection between model-based and input-output data-driven approximation, then, the Loewner framework, as detailed in [1, 13] is briefly recalled.

2.1 LTI Dynamical Models and Input-Output Data

Given the complex-valued (rational or irrational) transfer function matrix \mathbf{H} mapping the n_u inputs \mathbf{u} to the n_y outputs \mathbf{y} (as in (5)) or the input-output data collection $\{z_i, \Phi_i\}_{i=1}^N = \{z_i, \mathbf{H}(z_i)\}_{i=1}^N$ (where $z_i \in \mathbb{C}$ and $\Phi_i \in \mathbb{C}^{n_y \times n_u}$) defined as,

$$\mathbf{y}(z_i) = \mathbf{H}(z_i) = \Phi_i \mathbf{u}(z_i), \quad (7)$$

the approximation problem aims at constructing the (approximate) rational transfer function matrix $\hat{\mathbf{H}}$ mapping inputs \mathbf{u} to the approximate outputs $\hat{\mathbf{y}}$ such that

$$\hat{\mathbf{y}}(s) = \hat{\mathbf{H}}(s)\mathbf{u}(s). \quad (8)$$

Obviously, some objective are that (i) the reduced inputs to outputs map should be “close” to the original, (ii) the critical system features and structure should be preserved, and, (iii) the strategies for computing the reduced system should be numerically robust and stable. Approximating (5) with (8) is a *model-based* approximation, while, approximating (7) with (8) belongs to the *data-driven* family (see [2, 17] for examples). In this chapter, we follow the *data-driven* philosophy, and more specifically the interpolation-based approach using the Loewner framework.

2.2 Data-Driven Approximation and Loewner Framework at a Glance

The main elements of the Loewner framework are recalled here in the single-input single-output (SISO) case and readers may refer to [13] for a complete description and extension to the MIMO one. The Loewner approach is a data-driven method building a (reduced) rational descriptor LTI dynamical model \mathbf{H}^m ($\hat{\mathbf{H}}^r$) of dimension m ($r < m$) of the same form as (6), which interpolates frequency-domain data given as (7). More specifically, let us consider a set of distinct interpolation points $\{z_i\}_{i=1}^{2m} \subset \mathbb{C}$ which is split in two subsets of equal length as $\{z_i\}_{i=1}^{2m} = \{\mu_i\}_{i=1}^m \cup \{\lambda_i\}_{i=1}^m$. The method consists in building the Loewner and shifted Loewner matrices as,

$$[\mathbb{L}]_{ij} = \frac{\mathbf{H}(\mu_i) - \mathbf{H}(\lambda_j)}{\mu_i - \lambda_j} \text{ and } [\mathbb{L}_s]_{ij} = \frac{\mu_i \mathbf{H}(\mu_i) - \lambda_j \mathbf{H}(\lambda_j)}{\mu_i - \lambda_j}. \quad (9)$$

The model \mathbf{H}^m that interpolates \mathbf{H} is given by the following descriptor realisation,

$$E^m \dot{\mathbf{x}}(t) = A^m \mathbf{x}(t) + B^m \mathbf{u}(t), \mathbf{y}(t) = C^m \mathbf{x}(t) \quad (10)$$

where $E^m = -\mathbb{L}$, $A^m = -\mathbb{L}_s$, $[B^m]_i = \mathbf{H}(\mu_i)$ and $[C^m]_i = \mathbf{H}(\lambda_i)$ ($i = 1, \dots, m$). Assuming that the number $2m$ of available data is large enough, then it has been shown in [13] that a minimal model $\hat{\mathbf{H}}^r$ of dimension $r < m$ that still interpolates the data³ can be built with a projection of (10) provided that, for $i = 1, \dots, 2m$ (note that to avoid complex arithmetic, m points and their conjugate are selected),

$$\text{rank}(z_i \mathbb{L} - \mathbb{L}_s) = \text{rank}([\mathbb{L} \ \mathbb{L}_s]) = \text{rank}([\mathbb{L}^T \ \mathbb{L}_s^T]^T) = r. \quad (11)$$

In that case, denoting $Y \in \mathbb{C}^{m \times r}$ and (resp. $X \in \mathbb{C}^{m \times r}$) the matrix containing the first r left (resp. right) singular vectors of $[\mathbb{L} \ \mathbb{L}_s]$ (resp. $[\mathbb{L}^T \ \mathbb{L}_s^T]^T$). Then, $\hat{E}^r = Y^H E^m X$, $\hat{A}^r = Y^H A^m X$, $\hat{B}^r = Y^H B^m$, and $\hat{C}^r = C^m X$, is a realisation of this model $\hat{\mathbf{H}}^r$ with a McMillan degree equal to $\text{rank}(\mathbb{L})$. Note that if r in (11) is superior to $\text{rank}(\mathbb{L})$ then $\hat{\mathbf{H}}^r$ can either have a direct feed-through $\hat{D} \neq 0$ or a polynomial part. In the rest of the paper, one assumes that no polynomial term is present in the state-space realisation (10).

3 Model- and Data-Driven Control Synthesis Paradigm

Based on the above tools, let us now jump in the first contribution of the chapter, namely, the design of a feedback controller for irrational models. This objective, presented in Fig. 2, aims at seeking for a controller $\mathbf{K} \in \mathcal{RH}_\infty$ such that the intercon-

³ The model $\hat{\mathbf{H}}^r$ interpolates the original model \mathbf{H} at the z_i points. This is the reason why this method is also known as an interpolation one.

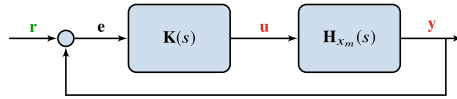


Fig. 2 Feedback loop scheme. The objective is to find $\mathbf{K}(s)$ stabilising $\mathbf{H}_{x_m}(s)$ and achieving some performances

nection closed-loop $\{\mathbf{H}, \mathbf{K}\}$ is stable and achieves some performance *e.g.* minimises some \mathcal{H}_∞ -norm or track some closed-loop performances.

In the rest of this section, two control design approaches are developed. The first one is a - standard - approximate and model-driven method, while the second is a direct data-driven one. The model-driven control design is based on a rational approximation of model (3) using the Loewner framework, followed with a structured \mathcal{H}_∞ -norm oriented control design step (note that any other approach can be used), while the data-driven relies on the Loewner framework to directly identify the controller, on the basis of the data only.

3.1 Model-Driven Approximation and Control

On Fig. 1, the transfer from the boundary input \mathbf{u} to the measurement point \mathbf{y}_{x_m} of both the original model \mathbf{H}_{x_m} and of the rational approximation one $\hat{\mathbf{H}}_{x_m}^{33}$, were illustrated. $\hat{\mathbf{H}}_{x_m}^{33}$ (of dimension $r = 33$) has been obtained by sampling \mathbf{H}_{x_m} as $\{\mathbf{H}_{x_m}(z_i)\}_{i=1}^{2m}$ where $z_i = \iota\omega_i$ such that z_i are closed under conjugation and for ω_i logarithmically spaced between $2\pi 10^{-2}$ and 2π ($m = 200$). At this point, thanks to the Loewner approach, it is both possible to simulate the equation using a standard **ODE** solver and to design a control law using rational model-based control design methods.

Using the rational approximation $\hat{\mathbf{H}}_{x_m}^{33}$ of the irrational model \mathbf{H}_{x_m} , standard feedback synthesis methods can be applied (*i.e.* design a feedback loop as on Fig. 2, but involving $\hat{\mathbf{H}}_{x_m}^{33}$ instead of \mathbf{H}_{x_m}). In this examples, the **hinfstruct** function embedded the MATLAB Robust Control Toolbox has been used [3]; it allows designing fixed structure controllers while minimising some \mathcal{H}_∞ -norm oriented performance transfer. Starting from $\hat{\mathbf{H}}_{x_m}^{33}$, let us first define the following generalised plant $\mathbf{T} = \hat{\mathbf{H}}_{x_m}^{33} \mathbf{W}_o$, where \mathbf{W}_o is the weighting filter defining the output signals on which the \mathcal{H}_∞ -norm optimisation will be done. \mathbf{W}_o is constructed to define the desired closed-loop performances attenuation and its bandwidth. The resulting state-space realisation of the generalised plant \mathbf{T} is then given by

$$\begin{cases} \dot{\xi}(t) = A_\xi \xi(t) + B_1 \mathbf{r}(t) + B_2 \mathbf{u}(t) \\ \mathbf{z}(t) = C_1 \xi(t) + D_{11} \mathbf{r}(t) + D_{12} \mathbf{u}(t) \\ \mathbf{e}(t) = \mathbf{r}(t) - \mathbf{y} \end{cases} \quad (12)$$

where $\boldsymbol{\xi}(t) \in \mathbb{R}^N$ and $\mathbf{z}(t) \in \mathbb{R}^{n_z}$ are the states (model plus weight state variables) and performance output signals, respectively. The performance transfer from \mathbf{r} to \mathbf{z} , is defined as $\mathbf{T}_{\mathbf{r}\mathbf{z}} = \hat{\mathbf{H}}_{x_m}^{33} \mathbf{W}_o$. In the considered case, one aims at tracking the reference signal \mathbf{r} and limiting the control action \mathbf{u} . One can then construct $\mathbf{W}_o = \mathbf{blkdiag}(W_e, W_u) = \mathbf{blkdiag}(10 \frac{s+1}{s}, \frac{s+10}{s+1000})$ describing performance output $\mathbf{z} = \mathbf{blkdiag}(W_o \mathbf{e}, W_u \text{spacerealisationofthe})$. The W_o weighting filter has been chosen to weight the sensitivity function and guarantee no steady-state error (e.g. roll-off in low frequency) and a bandwidth around 10^{-1}rad/s . The W_u one is instead used to weight actuator action in high frequencies (here the actuator will roll-off above 10rad/s). It can be mentioned that it is also a fairly standard way of weight selection. The \mathcal{H}_∞ control design consists in finding the controller \mathbf{K} , mapping \mathbf{e} to *spacerealisationofthe*, such that,

$$\mathbf{K} := \arg \min_{\tilde{\mathbf{K}} \in \mathcal{K}} \|\mathcal{F}_l(\mathbf{T}_{\mathbf{r}\mathbf{z}}, \tilde{\mathbf{K}})\|_{\mathcal{H}_\infty} \quad (13)$$

where $\mathcal{F}_l(\cdot, \cdot)$ is the lower fractional operator defined as (for appropriate partitions of M and K) by $\mathcal{F}_l(M, K) = M_{11} + M_{12}K(I - M_{22}K)^{-1}M_{21}$ [12]. With reference to (13), it is possible to define the class \mathcal{K} of \mathbf{K} to be restricted to the Proportional Integral one, meaning that one is seeking for \mathbf{K} with the following form, $\mathbf{K}(s) = k_p + k_i \frac{1}{s}$, where $k_p, k_i \in \mathbb{R}$. After optimisation, one obtains $k_p = 0.191$ and $k_i = 0.0252$ (note also that in this case, the optimal attenuation reached is $\gamma_\infty = 66.954$).⁴ Figure 3 then shows the sensitivity function \mathbf{S} (transfer from \mathbf{r} to \mathbf{e}) applied both on the rational approximation $\hat{\mathbf{H}}_{x_m}^{33}$ and the original model \mathbf{H}_{x_m} . It shows that good tracking in low frequencies is ensured, as well as some margin properties. In addition, the complementary sensitivity function, $\mathbf{M} = 1 - \mathbf{S}$, is reported on Fig. 4, illustrates the closed-loop transfer from the reference \mathbf{r} to the model output \mathbf{y} or $\hat{\mathbf{y}}$.

The resulting control law gives similar results for both approximated (rational) and original (irrational) models thus validating the approach.

3.2 Data-Driven Control

So far, the control design has been done in a fairly standard way, involving the approximated rational model. Instead of designing a controller on the basis of a rational reduced-order model, as the true system's behaviour is known through (5), a data-driven control strategy, based on the approach presented in [10], is followed. Authors stress that the main contribution of this section with respect to [10] stands in the comparison of the data-driven approach with the model-based one, resulting, as we will see later in the chapter, in exactly similar results.

⁴ The optimisation is done using the `hinfstruct` routine, allowing minimising the closed-loop interconnection of $\mathbf{T}_{\mathbf{r}\mathbf{z}}$ with $\tilde{\mathbf{K}}$. In general, one seek for $\|\mathcal{F}_l(\mathbf{T}_{\mathbf{r}\mathbf{z}}, \mathbf{K})\|_{\mathcal{H}_\infty} = \gamma_\infty \leq 1$. Here we simply aim at reaching stability and tracking performances.

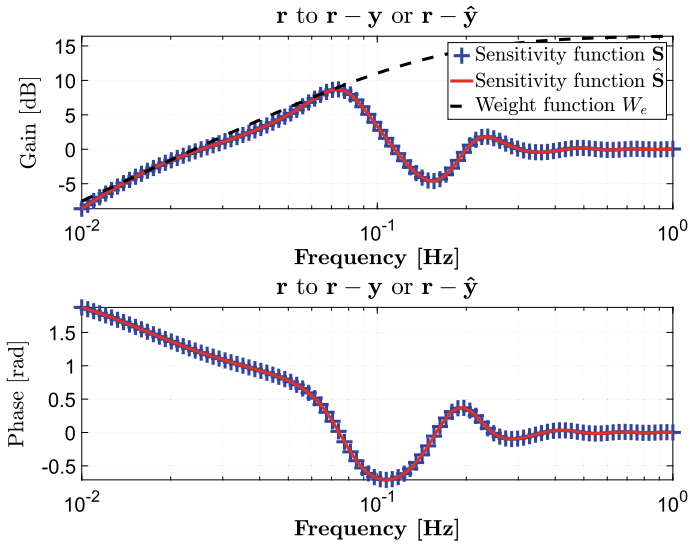


Fig. 3 Sensitivity function S (blue +) and \hat{S} (solid red). Weighting function W_e used in the control design (solid black)

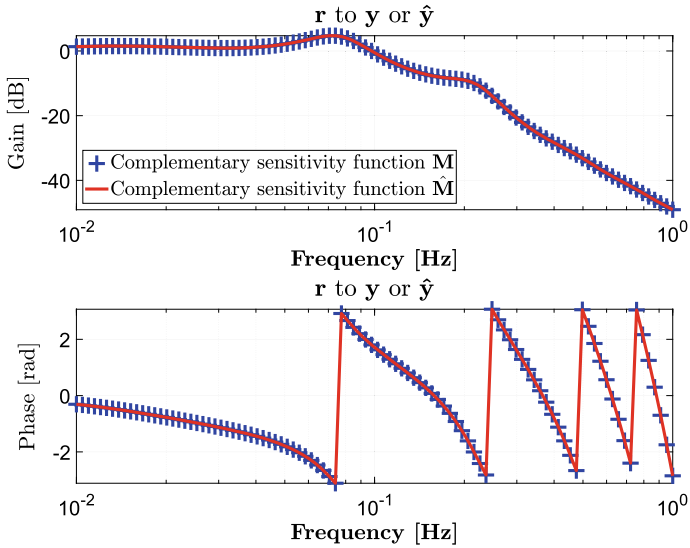
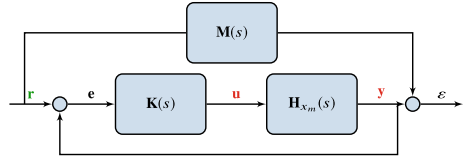


Fig. 4 Complementary sensitivity function M and \hat{M} , linking r to y (blue +) or \hat{y} (solid red)

Fig. 5 Data-driven control problem formulation: \mathbf{M} is the reference model (objective) and \mathbf{K} the controller to be designed



3.2.1 Introducing the Ideal Controller

Data-driven control consists in recasting the control design problem as an identification one. The main advantage of this strategy is that it provides a controller tailored to the actual system. This change of paradigm shift the identification/simplification process of model to the controller directly. Different techniques have been proposed, see references in [10], considering a set of structured controllers. Recently, [10] pushed the interpolatory framework, and especially the Loewner one, into the process, enabling **MIMO** controllers design without a-priori structure selection.

The objective is to find a controller \mathbf{K} minimising the difference between the resulting closed-loop and a given reference model \mathbf{M} (see Fig. 5). This is made possible through the definition of the ideal controller \mathbf{K}^* , being the **LTI** controller that would have given the desired reference model behaviour if inserted in the closed-loop. It is then defined as follows:

$$\mathbf{K}^* = \mathbf{H}_{x_m}^{-1} \mathbf{M} (\mathbf{I} - \mathbf{M})^{-1}. \quad (14)$$

Given a reference model \mathbf{M} and provided frequency-domain data $\{\omega_i, \Phi_i\}_{i=1}^{2m}$ from the plant \mathbf{H}_{x_m} , it is possible to evaluate the frequency-response of the ideal controller \mathbf{K}^* at these very same frequencies. The main idea of the Loewner Data-Driven Control (**LDDC**) algorithm introduced in [10] is to interpolate the frequency-response of the ideal controller \mathbf{K}^* and to reduce it to an acceptable order. For the present example, $m = 200$ samples of the frequency-response, logarithmically spaced between $2\pi \cdot 10^{-2}$ and 2π rad.s⁻¹ are considered (similar to the one used for the computation of a rational reduced-order model in Sect. 2).

3.2.2 Data-Driven Control Design and Model-Free Stability Analysis

As explained in [9], the reference model choice is a key factor for the **LDDC** success, as for any other model reference control techniques. Indeed, this latter should not only represent a desirable closed-loop behaviour, but also an achievable dynamic of the considered system (i.e. the ideal controller should not internally destabilise the plant and imply terrible dynamics). A reference model is said to be achievable by the plant if the corresponding ideal controller internally stabilises the plant.

The closed-loop performances are limited by the system's right hand side poles $\{p_j\}_{j=1}^{n_p}$ and zeros $\{z_i\}_{i=1}^{n_z}$ (and their respective output directions in the **MIMO** case)

defined as $\mathbf{H}_{x_m}(z_i) = 0$ and $\mathbf{H}_{x_m}(p_j) = \infty$. Finally, the class \mathcal{M} of achievable reference models is defined as follows:

$$\mathcal{M} = \{ \mathbf{M} \in \mathcal{H}_\infty : \mathbf{M}(z_i) = 0 \text{ and } \mathbf{M}(p_j) = 1 \}. \tag{15}$$

In the general case, the right half plane poles and zeros of the system are estimated in [9] in order to build an achievable reference model on the basis of the initial specifications given by the user. This is made possible through a data-driven stability analysis introduced in [6] and the associated estimation of instabilities presented in [5].

Here, the PDE describing the system’s dynamic is known and allows to determine the performance limitations without applying the aforementioned data-driven stability analysis. The instabilities are $\mathbf{H}_{x_m}(+\infty) = 0$ and $\mathbf{H}_{x_m}(0) = \infty$, implying that the reference model should satisfy $\mathbf{M}(+\infty) = 0$ and $\mathbf{M}(0) = 1$.

As shown in [8], once the set \mathcal{M} of achievable reference models is determined, the one chosen in this set influences a lot the control design. To illustrate this point, the LDDC algorithm is applied on the proposed example using two reference models \mathbf{M}_1 and \mathbf{M}_2 . \mathbf{M}_1 is chosen as a perfectly damped second-order model with a natural frequency $\omega_0 = 0.5 \text{ rad.s}^{-1}$ and reads

$$\mathbf{M}_1(s) = \frac{1}{s^2/\omega_0^2 + 2s/\omega_0 + 1},$$

satisfying the performance limitations of the system. The second reference model \mathbf{M}_2 is the closed-loop obtained in the model-based control design obtained in Sect. 3.1 ($\mathbf{M}_2 = \hat{\mathbf{M}}$, see Fig. 4).⁵

Once the reference models \mathbf{M}_1 and \mathbf{M}_2 are chosen, by following (14), it is possible to compute the frequency-responses of the associated ideal controllers, denoted \mathbf{K}_1^* and \mathbf{K}_2^* respectively, at the frequencies where data from the plant are available. In order to obtain a controller model \mathbf{K} , the Loewner framework is then applied considering the following interpolatory conditions: $\forall i = 1 \dots 2m, \mathbf{K}(i\omega_i) = \mathbf{K}^*(i\omega_i)$.

In the present case, minimal realisations of \mathbf{K}_1^* and \mathbf{K}_2^* of order $n_1 = 34$ and $n_2 = 40$ respectively are obtained. In order to compare the results of the model-based approach with the data-driven strategy, the ideal controllers \mathbf{K}_1^* and \mathbf{K}_2^* are reduced up to a first order (using the rank revealing factorisation embedded in the Loewner framework), giving two controller models denoted \mathbf{K}_1 and \mathbf{K}_2 , which transfer functions are given in (16).

$$\mathbf{K}_1(s) = \frac{0.02277}{(s + 0.0382)} \text{ and } \mathbf{K}_2(s) = \frac{0.1914(s + 0.1315)}{s} = 0.1914 + \frac{0.0252}{s}. \tag{16}$$

⁵ Since the procedure used to get $\hat{\mathbf{M}}$ preserves internal stability, the obtained closed-loop is necessarily achievable by the plant \mathbf{H}_{x_m} .

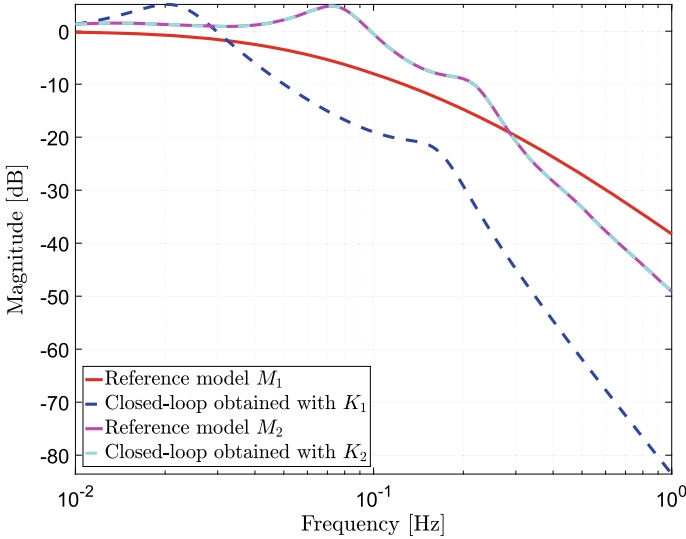


Fig. 6 Resulting closed-loop transfers involving \mathbf{K}_1 and \mathbf{K}_2 , obtained using \mathbf{M}_1 and \mathbf{M}_2 reference models

When using \mathbf{M}_2 as reference model, the first order controller \mathbf{K}_2 has exactly the same expression than the one obtained in the model-based approach solving (13). The frequency responses of the resulting closed-loops obtained with \mathbf{K}_1 and \mathbf{K}_2 are visible on Fig. 6.

Interestingly, with reference to Fig. 6, \mathbf{K}_2 perfectly recovers the requested performance of \mathbf{M}_2 with a controller of rational order one (indeed, we expected to observe this result since we knew from the model-based approach presented in Sect. 3.1 that a rational control of order leading to this performance was achievable). Conversely, \mathbf{K}_1 (reduced to a rational form of dimension one) is not able to recover the performances of \mathbf{M}_1 , and a higher degree would be expected (as it is not the topic of the chapter, this point is not detailed further).

One of the major challenges in this data-driven control strategy is to preserve internal stability. While the ideal controller is known to be stabilising thanks to the choice of an achievable reference model, see [9], there is no guarantee regarding the reduced-order controllers. Therefore, it is necessary to analyse the internal stability during the controller reduction step. To that extent, the resulting closed-loop is written as on Fig. 7. This scheme makes the controller error appear as a perturbation.

It is then possible to apply the small-gain theorem: the interconnected system shown on Fig. 7 is well-posed and internally stable for all stable $\Delta = \mathbf{K} - \mathbf{K}^*$ with $\|\Delta\|_\infty < \frac{1}{\gamma}$ if and only if $\|\mathbf{H}_{x_m}(1 - \mathbf{M})\|_\infty \leq \gamma$. The bound γ on the controller modelling error can then be estimated using the data only as:

$$\tilde{\gamma} = \max_{i=1\dots m} |\mathbf{H}_{x_m}(i\omega_i)(1 - \mathbf{M}(i\omega_i))|. \quad (17)$$

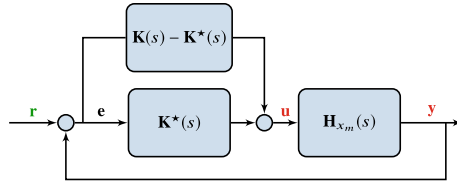


Fig. 7 Stability analysis of the closed-loop obtained with a reduced-order controller \mathbf{K} : the closed-loop is reformulated according to the controller modelling error $\mathbf{K} - \mathbf{K}^*$

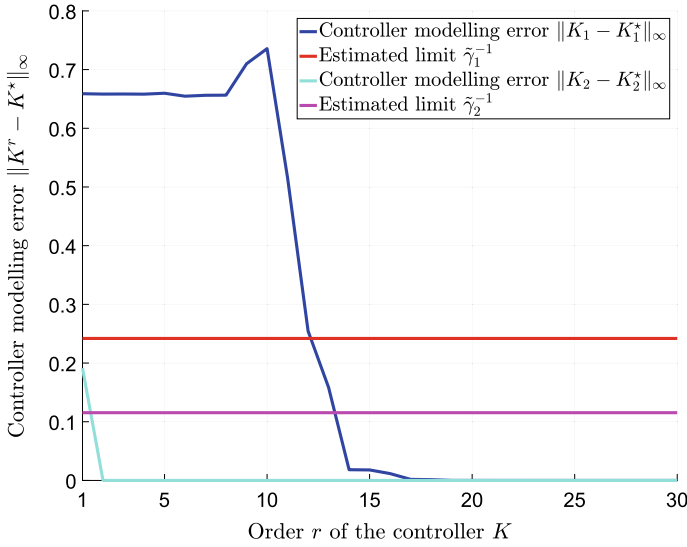


Fig. 8 Evolution of the controller modelling error as a function of the controller reduction step

The evolution of the controller modelling error $\|\mathbf{K} - \mathbf{K}^*\|_\infty$ according to the order of the reduced controller \mathbf{K} is represented on Fig. 8 for the two considered reference models, \mathbf{M}_1 and \mathbf{M}_2 . According to Fig. 8, to ensure that the resulting closed-loop is internally stable, one may reduce \mathbf{K}_1^* up to $r = 13$ and \mathbf{K}_2^* up to $r = 2$.

However, this stability test is too conservative: it is not possible to conclude anything regarding internal stability for controllers \mathbf{K} for which $\|\mathbf{K} - \mathbf{K}^*\|_\infty > \tilde{\gamma}$. In the present case, both order one controllers given in (16) stabilise internally the rational model. Another solution would be to use the model-free stability analysis presented in [6] but it may be complicated for the user to conclude regarding internal stability. To that extent, another approach to analyse stability in a data-driven framework is introduced in Sect. 4.

3.3 Model- Vs. Data-Driven Control Design Remarks

Throughout this section, it has been demonstrated how central the Loewner tool can be, either for model-driven and data-driven control. Interestingly, by choosing the closed-loop performances $\hat{\mathbf{M}} = \mathbf{M}_2$ obtained with the first approach, the second controller \mathbf{K}_2 is able to recover exactly the same controller properties, while avoiding the complex model construction step. This property reduces the time consuming model construction step and allows a quick design of the controller. However, this main advantage is balanced by the fact that in the model-based approach, the stability assessment is usually carried out using the approximated model, here $\hat{\mathbf{H}}_{x_m}^{33}$. This latter being very accurate, the eigenvalues computation is traditionally enough for concluding of the stability, robustness ... On the contrary, in the second data-driven approach, the stability cannot be analysed as is and robustness (conservative) bounds are used instead.

Both approaches can be viewed as equivalent since they lead to the same controller. Moreover, in both cases, the interpolatory framework offered by the Loewner matrices is the major ingredient for the success of the design. One may consider these approaches as complementary: the model-based approach may be privileged for critical systems where model understanding is of major importance and for which engineering time can be spent, while the data-driven one should be the best solution for fast computation, preliminary design, for which neither safety nor critical issues are in the scope.

4 Stability Assessment of \mathcal{L}_2 Meromorphic Functions

Independently of the chosen control design approach, both methodology rely on a rational model and the stability and performances obtained by the controller on the irrational model cannot be guaranteed. This is why, in this section, the stability involving the original irrational transfer \mathbf{H}_{x_m} , is addressed in a nonstandard manner and involving once again the Loewner matrices.

4.1 The \mathcal{L}_∞ Procedure

Being given a rational controller \mathbf{K} and the irrational model defined by the meromorphic function \mathbf{H}_{x_m} , one important challenge is to assess the stability of the closed-loop and e.g. to evaluate its stability when delay enters in the loop (delay margin of the interconnection). These questions are gathered on Fig. 9, for which the corresponding closed-loop naturally reads

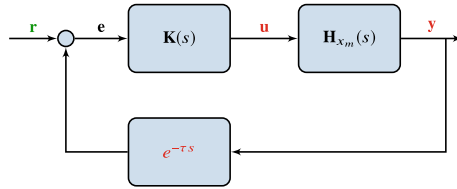


Fig. 9 Feedback loop scheme for stability and margin computation. The objective is to assess the stability of the interconnection $\mathbf{K}(s)$ using $\mathbf{H}_{x_m}(s)$ and a fixed delay τ in the loop

$$\mathbf{M}_\tau(s) = \frac{\mathbf{H}_{x_m}(s)\mathbf{K}(s)}{1 + \mathbf{H}_{x_m}(s)\mathbf{K}(s)e^{-\tau s}}, \tag{18}$$

where $\tau \in \mathbb{R}_+$ is the delay value affecting the loop. On the basis of [16] and on Chap. 5 of [17], let us now propose a numerical procedure for the stability approximation of infinite \mathcal{L}_∞ meromorphic functions. Note that the proposed version extends the one presented first in [16] and second in [17] by providing a much more numerically robust version, and now considers \mathcal{L}_∞ functions. The \mathcal{L}_∞ -MFSA procedure given in Algorithm 1 is first proposed.

Algorithm 1 \mathcal{L}_∞ -MFSA - \mathcal{L}_∞ Meromorphic Function Stability Approximation

Require: $\mathbf{H} \in \mathcal{L}_\infty$, $\{\omega_i\}_{i=1}^N \in \mathbb{R}_+$, $N \in \mathbb{N}$ and $\epsilon \in \mathbb{R}_+$ (typically twice machine precision)

- 1: Sample \mathbf{H} and obtain $\{t\omega_i, \mathbf{H}(t\omega_i)\}_{i=1}^N$
 - 2: Construct an exact Loewner interpolant and obtain $\hat{\mathbf{H}}^r$, ensuring interpolatory conditions
 - 3: Compute $\hat{\mathbf{H}}^r_+$, the best stable approximation of $\hat{\mathbf{H}}^r$ (e.g. using [11])
 - 4: Compute the stability index as $\text{stabTag} = \|\hat{\mathbf{H}}^r_+ - \hat{\mathbf{H}}^r\|_{\mathcal{L}_\infty}$
 - 5: If $\text{stabTag} < \epsilon$, then \mathbf{H} is stable, otherwise, \mathbf{H} is unstable
-

Algorithm 1 embeds a relative simple procedure, which will be shown to be actually quite effective, fast and reliable. The idea consists in exactly matching the original input-output model by a rational model $\hat{\mathbf{H}}^r$, by guaranteeing interpolatory conditions. Then, to seek for the best stable approximation $\hat{\mathbf{H}}^r_+$ of the obtained model $\hat{\mathbf{H}}^r$. The \mathcal{L}_∞ distance between the interpolated $\hat{\mathbf{H}}^r$ and stable $\hat{\mathbf{H}}^r_+$ models is then computed. If this latter is smaller than a given threshold $\epsilon > 0$, then we conclude that \mathbf{H} is stable, and unstable otherwise. By applying the procedure to (18), including the irrational model (4), for varying frozen values of delay τ_j (20 linearly space between 4.6 and 5.5 has been chosen, surrounding the delay instability margin), leads to the following results of the stabTag : $[0, 0.8412 \times 10^{-11}, 0.2496 \times 10^{-11}, 0.4084 \times 10^{-11}, 0, 0, 0, 0, 0, 0, 1.7719 \times 10^7, 1.6346 \times 10^6, 50.8770, 34.8265, 26.4887, 21.3817, 17.9326, 15.4471, 13.5709, 12.1046]$. These values indicate that the closed-loop is stable up to the destabilising delay value $\tau_j \approx 5.0737\text{s}$.

To assess this approach on such a simple SISO case, the stability can also be checked using the Nyquist graph of $\mathbf{L} = \mathbf{H}_{x_m}\mathbf{K}$. Figure 10 illustrates the Nyquist

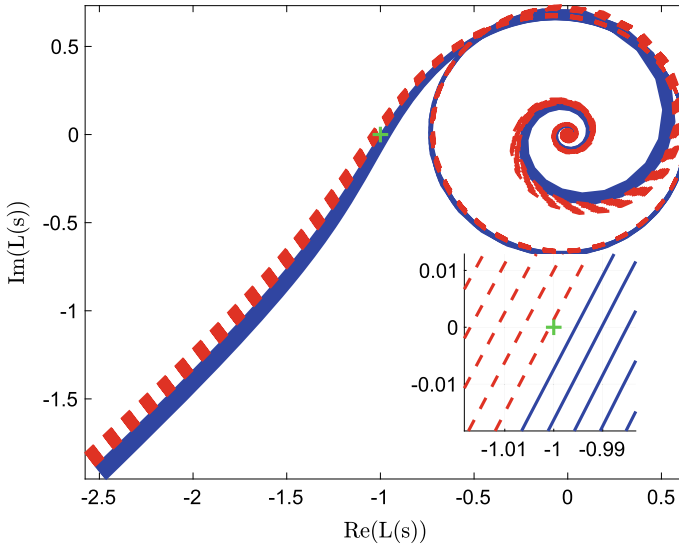


Fig. 10 Nyquist diagram of $L(s)$ for varying values of τ . Solid blue curves are Nyquist for which the stability tag `stabTag` is below 10^{-10} (stable configuration) and dashed red curves for which tag is above 10^{-10} . Bottom right: zoom around the Nyquist point (green +)

curve bundle as a function of the τ and shows that the proposed approach leads to a good delay stability approximation.

4.2 Approximation-Driven Arguments for the \mathcal{L}_∞ -MFSA

Algorithm 1 is rather simple and deserves some comments arguing of its viability. Let us first refer to [16] where arguments and a similar procedure, involving the **TF-IRKA** algorithm [4] have been suggested (this procedure ensures \mathcal{H}_2 -optimal bi-tangential interpolatory conditions). This later provides good results but lacks in determining the approximation order r . Additionally, as **TF-IRKA** is an \mathcal{H}_2 -oriented procedure its validity in the $\mathcal{L}_2(i\mathbb{R})$ function space is limited. Here, the path presented in Fig. 11 is used to construct the \mathcal{L}_∞ -MFSA procedure.

With reference to Fig. 11 (top left three blocks), the **TF-IRKA** interpolatory conditions are released and the Loewner framework is used instead. First, one major benefit of such a trade stands in the automatic selection of the approximating order r , done by a rank revealing factorisation where machine precision is expected. Second, it interpolates the data without any stability constraint. One obtains $\hat{\mathbf{H}}^r$ which tangentially interpolates the data and which, by increasing the numbers of samples z_i in (7) hopefully converges to the same model \mathbf{H}^r (e.g. the rational approximation

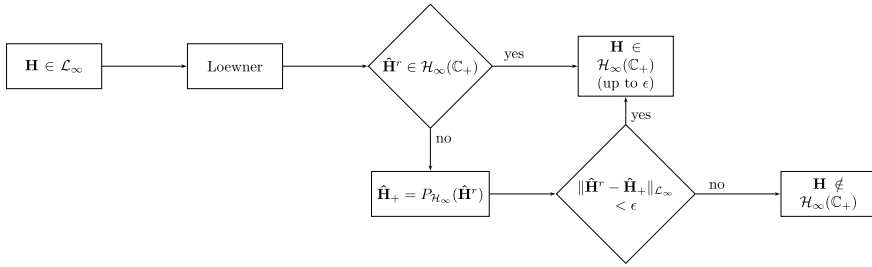


Fig. 11 Graphical illustration of the \mathcal{L}_∞ -MFSA procedure (starting point, at top left)

obtained by Loewner matrices is not affected by the number of interpolation points). At this point, the resulting model $\hat{\mathbf{H}}^r$ may be stable or unstable.

- If $\hat{\mathbf{H}}^r$ is stable, then one concludes on the stability of \mathbf{H} , up to some tolerance $\epsilon > 0$. Indeed, one could always add $\epsilon \mathbf{H}_a \in \mathcal{H}_\infty$ to the data without being able to see the effects on the interpolation conditions achieved by $\hat{\mathbf{H}}^r$, due to numerical accuracy.
- If $\hat{\mathbf{H}}^r$ is unstable, as it is always possible to find an unstable approximant to a stable model in the \mathcal{L}_2 sense [16], one may use the projection onto a stable subspace, here the \mathcal{H}_∞ one, denoted $\hat{\mathbf{H}}_+ = P_{\mathcal{H}_\infty}(\hat{\mathbf{H}}^r)$, to emphasise the importance of the unstable part of $\hat{\mathbf{H}}^r$ on the interpolation conditions. If the unstable part plays a negligible role (i.e. $\|\hat{\mathbf{H}}^r - \hat{\mathbf{H}}_+\|_{\mathcal{L}_\infty} < \epsilon$), then a stable interpolating model has been found for \mathbf{H} which is therefore likely to be stable. Otherwise, if the unstable part cannot be removed, then it is likely that \mathbf{H} is unstable.

The Loewner framework allows to find a rational model $\hat{\mathbf{H}} \in \mathcal{RL}_\infty$ that interpolates $\mathbf{H} \in \mathcal{L}_\infty$ at an arbitrary number of frequencies. The suggestion we claim is in twofold. One is always able to find a rational model $\hat{\mathbf{H}}^r \in \mathcal{RL}_\infty$ that well reproduces $\mathbf{H} \in \mathcal{L}_\infty$ (at least interpolates a large number of points). We assume that this implies a convergence in the \mathcal{L}_∞ sense, meaning that one is always able to find a rational function matching an irrational one defined over \mathcal{L}_∞ . Then, if $\hat{\mathbf{H}}^r \in \mathcal{RL}_\infty$ can in addition be projected onto $\hat{\mathbf{H}}_+ = P_{\mathcal{H}_\infty}(\hat{\mathbf{H}}^r) \in \mathcal{H}_\infty(\mathbb{C}_+)$ with negligible loss in the \mathcal{L}_∞ -norm, then the unstable part can be considered as irrelevant for the behaviour description of \mathbf{H} , which can thus be assumed stable. Otherwise, if the unstable part leads to important \mathcal{L}_∞ -norm mismatch, \mathbf{H} can be considered as unstable.

5 Conclusion

In this chapter, the interpolatory framework proposed by the Loewner setup, as introduced in the seminal paper [13], has been further used for the control design and the stability estimation. The singularity of the proposed approach is to show that the Loewner is not only a model approximation tool, but a complete dynamical-oriented tool. The main contributions of this work is twofold. First, to compare frequency-oriented data- and model-based control design approaches, showing that both lead to similar performances. Second, to suggest a method for approximating the stability of any \mathcal{L}_∞ functions (either rational or irrational). Both contributions are based on Loewner matrices. Through an academic example described by a linear PDE set, the Loewner framework has been used for different purpose, showing its impressive versatility and applicability to solve complex problems. To the authors perspective, this approach opens the fields for analysing and controlling irrational (infinite-dimensional) models in a relatively simple manner. Even if the approach does not stand as a completely closed solution, it can be viewed as an alternative for engineers and practitioners to deal with irrational models in a simple manner.

References

1. Antoulas, A., Lefteriu, S., Ionita, A.: Model reduction and approximation theory and algorithms. In: Benner, P., Cohen, A., Ohlberger, A., Willcox, K. (eds.) A tutorial introduction to the Loewner framework for model reduction. SIAM, Philadelphia. (2016)
2. Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems. Advanced Design and Control, SIAM, Philadelphia (2005)
3. Apkarian, P., Noll, D.: Nonsmooth \mathcal{H}_∞ synthesis. *IEEE Trans. Autom. Control* **51**(1), 71–86 (2006)
4. Beattie, C., Gugercin, S.: Realization-independent \mathcal{H}_2 -approximation. In: Proceedings of the 51st IEEE Conference on Decision and Control, pp. 4953–4958 (2012)
5. Cooman, A., Seyfert, F., Amari, S.: Estimating unstable poles in simulations of microwave circuits. In: IEEE/MTT-S International Microwave Symposium (2018)
6. Cooman, A., Seyfert, F., Olivi, M., Chevillard, S., Baratchart, L.: Model-free closed-loop stability analysis: a linear functional approach. *IEEE Trans. Microw. Theory Tech.* (2018)
7. Hoffman, K.: Banach Spaces of Analytic Functions. Prentice Hall (1962)
8. Kergus, P., Olivi, M., Poussot-Vassal, C., Demourant, F.: Data-driven reference model selection and application to L-DDC design (2019). [arXiv:1905.04003](https://arxiv.org/abs/1905.04003)
9. Kergus, P., Olivi, M., Poussot-Vassal, C., Demourant, F.: From reference model selection to controller validation: application to loewner data-driven control. *IEEE Control Syst. Lett.* **3**(4), 1008–1013 (2019)
10. Kergus, P., Poussot-Vassal, C., Demourant, F., Formentin, S.: Frequency-domain data-driven control design in the Loewner framework. In: Proceedings of the 20th IFAC World Congress, pp. 2095–2100. Toulouse, France (2017)
11. Kohler, M.: On the closest stable descriptor system in the respective spaces \mathcal{RH}_2 and \mathcal{RH}_∞ . *Linear Algebra Appl.* **443**, 34–49 (2014)
12. Magni, J.F.: Linear fractional representation toolbox for use with matlab. Technical report, Onera, Toulouse, France (2006). <http://w3.onera.fr/smac/>
13. Mayo, A.J., Antoulas, A.C.: A framework for the solution of the generalized realization problem. *Linear Algebra Appl.* **425**(2), 634–662 (2007)

14. Partington, J.: Linear operators and linear systems: an analytical approach to control theory, vol. 60. Cambridge University Press (2004)
15. Pontes, I.: Large-scale and infinite dimensional dynamical model approximation. Ph.D. thesis, Onera, ISAE, Toulouse University, Toulouse, France (2017)
16. Pontes Duff, I., Vuillemin, P., Poussot-Vassal, C., Briat, C., Seren, C.: Approximation of stability regions for large-scale time-delay systems using model reduction techniques. In: Proceedings of the 14th European Control Conference, pp. 356–361. Linz, Austria (2015)
17. Poussot-Vassal, C.: Large-scale dynamical model approximation and its applications. HDR, habilitation thesis, Onera, INP Toulouse, Toulouse, France (2019)

Applications

Oscillations in Biology



Jitendra K. Meena and Clifford C. Dacso

Abstract Experiments in molecular biology are generally performed in a static state; analytes are frozen in time and displayed as if there had been no changes over time. Yet, biology is dynamic, responding to internal regulatory demands, external stimuli and evolutionary pressures. There are fundamental rhythms in biological systems such as ovulation in humans and estrus in animals. At the cell level, day-night synchronization, the 24 h circadian cycle created by a feedback loop of gene circuits, affects vital functions. Disruption of the circadian cycle has been shown to have protean health consequences. In contrast, little is studied about the rhythms independent of 24 h clock. This article reviews the non-circadian rhythms in biological systems existing at the transcriptional level; inherent problems and biases in detecting these from biological data and the use of mathematical methods in overcoming these biases.

Keywords Genetic network · Transcription factor · Transcillation · BMAL1

Analysis of the cyclic nature of gene transcription has disclosed cycles other than circadian. Application of the eigenvalue decomposition of the matrix pencil to temporal transcription data, for example, demonstrated rhythms that are independent of the circadian cycle at 12 and 8 h. Experimental examination of the genes involved in this novel 12 h clock discloses them to be intimately involved in aspects of inflammation and energy. The eigenvalue pencil method adds to the analysis of gene transcription by disclosing non-circadian oscillations and suggesting those likely to be independent rather than harmonic of the dominant circadian rhythm. Additionally, plotting the eigenvalues on the unit circle discloses aspects of the fate of the transcribed RNA, whether the gene continues to oscillate, decays, or becomes disorganized and chaotic. It is thus unveiling a deeper mechanistic aspect of non-circadian rhythm to the functions of gene transcription.

J. K. Meena · C. C. Dacso (✉)

Departments of Biochemistry and Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX, USA

e-mail: cdacso@bcm.edu

© Springer Nature Switzerland AG 2022

C. Beattie et al. (eds.), *Realization and Model Reduction of Dynamical Systems*,
https://doi.org/10.1007/978-3-030-95157-3_20

1 Cycles in Biology

1.1 Introduction

In 1953, the arcane world of genetics intersected with the world at large with *un grand éclat* when Francis Crick and James D. Watson announced the discovery of the three-dimensional structure of DNA. It was such a momentous event that the *New York Times* recognized it with one column of type on page 17 (CLUE TO CHEMISTRY OF HEREDITY FOUND; American and Briton Report Solving Molecular Pattern of Vital Nucleic Acid TESTS BY X-RAY PLANNED Work Done in England, if It Is Confirmed, Should Make Biochemical History, 1953). Nonetheless, it did eventually shake the world and continues to as the mechanism by which DNA controls inheritance has become more evident by the day. The elucidation of the three-dimensional structure of DNA compelled understanding the mechanism by which the molecule self-replicated, thus disclosing the way in which hereditary information is passed on through the generations. That insight, when ramified through the ensuing years, has increased in complexity.

It is now clear, for example, that the substituent components of genetic information transfer are mutable and modifiable, even during the process of replication itself. Further, in animals, genes need to be turned on and off for their action to result in orderly biological function. The “Central Dogma” of genetics as described by Francis Crick in lecture notes from 1956 (Fig. 1) describes the one-way path by which DNA is decoded, recruits RNA, and results in the synthesis of proteins that build and operate the cell machinery. This function repeats at an astounding rate with the mammalian gene transcribed entirely within minutes. Gene transcription is initiated by the recruitment of transcription factors at the level of DNA followed by polymerization of ribonucleotides by the DNA-dependent RNA polymerase and termination by the specialized proteins that cleave off the nascent transcripts. Subsequently, the RNA molecule is spliced where non-coding areas called introns are removed, then exported from the nucleus, and translated by intricate machinery called ribosomes resulting in protein production. Throughout the life of a transcript, specialized molecular machineries regulate synthesis and turnover to maintain a dynamic equilibrium. Each transcript that is necessary for producing an individual protein is thus controlled by the genetic networks driven by the survival and the growing need of an individual cell. Such a regulation leads to rhythmicity of the transcript, and studying this rhythmic behavior allows us to unveil the underlying genetic network and evolutionary pattern of a biological system. With this in mind, it is curious that most analyses of cell functions look at the phenomena as if they were frozen in time despite the intricate dynamism of the process.

Evolution forms the core of biology. Indeed, Dobzhansky’s famous dictum is contained in the title of his seminal article, “Nothing In Biology Makes Sense Except In The Light of Evolution.” Reference [1] Evolution is not directional or teleologic; thus, it is not fruitful to ask, “what is the purpose of evolution?” Instead, the inherent

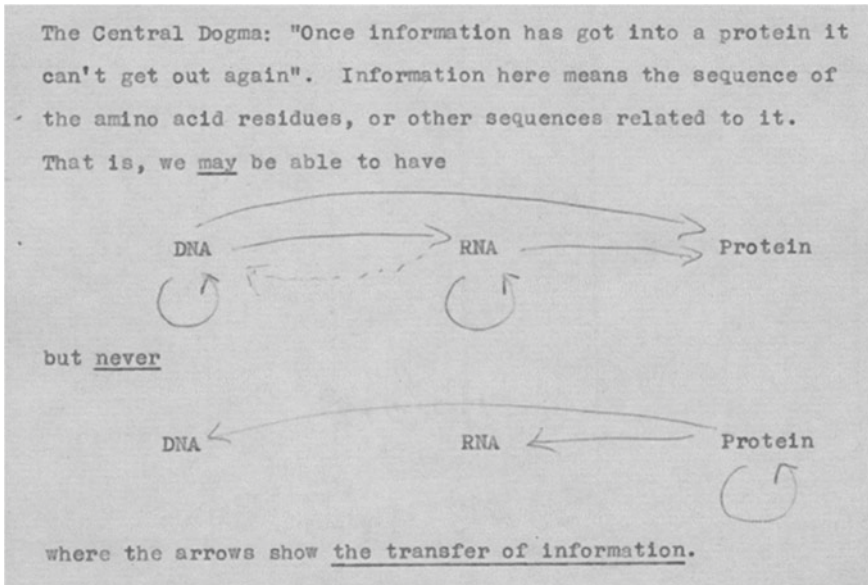


Fig. 1 The unidirectionality of transcription and translation with embedded cyclic activity. From Wellcome Library, Francis crick papers

stochasticity of biological properties allows for natural selection, the colloquial “survival of the fittest”. At a more macro level than the gene, periodicity remains ubiquitous in biology, generally characterized by a feedback loop. One such loop drives the rhythmic oscillations of electrical generation in the sinoatrial node of the mammalian heart, thus generating sufficient electricity to depolarize the myocardium and drive a regular heartbeat. Recent work has clarified that the generation and propagation of the electricity depends on the mutual entrainment of subcellular and membrane clocks [2].

This is not to say that there is no order in biological systems. From the constraints of physics to the appearance of lethal mutations, there are many spaces that are not permissible and there are “guard-rails” that exist to preserve the cell and thus the organism from the emergence of detrimental events. Additionally, there are many events in the biology of the cell that are cyclic, using the same process, and substrate many times.

1.2 *Krebs Cycle in Biological Systems for Energy Generation*

Cycles are parsimonious of resources as cyclic activity reconstitutes the substrate for use again the next time the cycle activates. An archetypical example of this is the Krebs, or tricarboxylic acid cycle (Fig. 2). In this cycle, energy is produced from

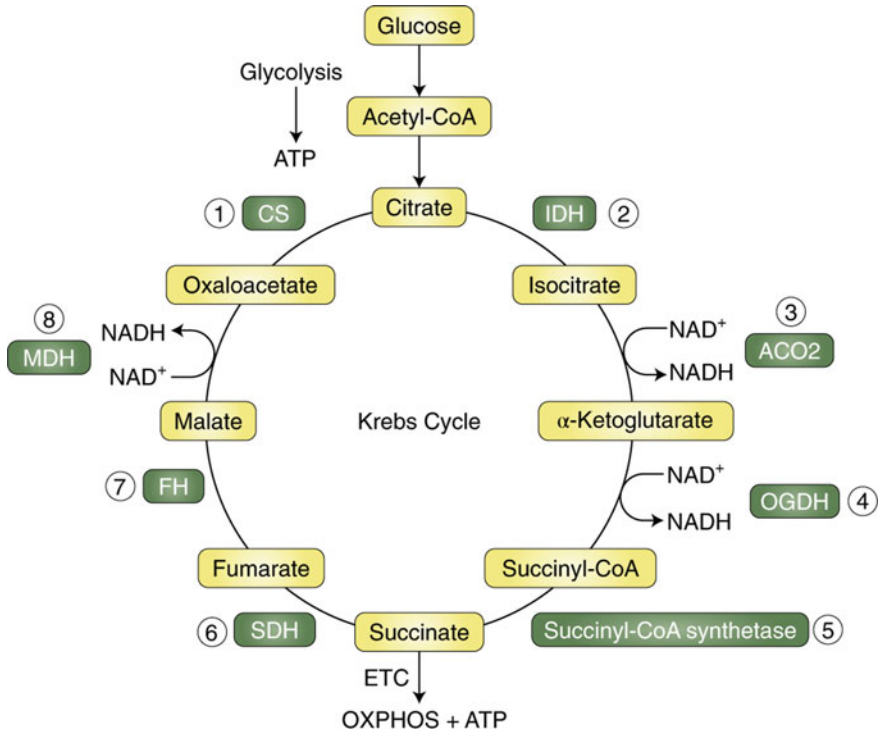


Fig. 2 The krebs cycle [3]

glucose in the mitochondrion of the cell animal. Glucose comes in at the top and is stored by a process called oxidative phosphorylation as a high energy phosphate bond of adenosine triphosphate, ATP. Through a beautiful series of chemical events, most of the energy is stripped out of the glucose molecule, but the constituents of the reaction are reconstituted to do the same thing over again. Moreover, this repeats as long as there are oxygen and glucose. It does not serve the majesty of biology to oversimplify this process as it is complex and full of variability. However, it is these features that render the mathematical modeling of biological processes difficult. One way to overcome the modeling challenge is to account for each of the substituents of the cycle and model the output of the cycle itself.

1.3 Essential Biological Rhythms

From the effects of the lunar cycle to extremely rapid oscillation in cell metabolism, cyclic activity forms the foundation of biological structure and function. Of great

| Biological Rhythm | Period |
|------------------------------------|----------------------------|
| Cell-level oscillations | |
| Central Pattern generators | 0.001-10 seconds |
| Cardiac pacemaker | 1 second |
| Aerobic glycolysis | 1-20 minutes |
| Insulin secretion in humans | 4-6 minutes |
| Murine somitogenesis | 2 hours |
| Trascription factors | 90- minutes-3 hours |
| Circatidal clock (marine animals) | 11-12 hours |
| Cell stress marker (animal) (XBP1) | 12 hours |
| Circadian cycle | 24 hours |
| Organism-level oscillators | |
| Estrus (cow) | 21 days |
| Ovarian cycle in humans | 28 days |
| Estrus (dogs) | 6 months |
| Seasonal affective disorder | 1 year |
| Periodic Cicadas | 17 years |

Fig. 3 Some common oscillators in biology. Adapted from [4]

interest, there are a large number of known oscillators in biology that are of enormous relevance to human health (Fig. 3).

1.4 Molecular Oscillators

A periodic function is critical to several fundamental biological processes. Best known, of course, is the circadian oscillator that has approximately 24-h rhythmicity. Its name translates from the Latin *circa*, around, and *dies*, day. It is instructive to spend a few words on the circadian oscillator as it has been well worked out in the past decade.

Because circadian rhythms are under the direct control of the gene, they appear in every cell in the body. An elegant and intricate feedback loop controls the circadian rhythm in cells under the control of an area of the brain called the suprachiasmatic nucleus, which synchronizes the circadian cycle to the ebb and flow of daylight. In humans, of course, the circadian rhythm turns things “on” in the daylight and “off” at night. Rodents, the species best studied for mechanisms of circadian biology, are nocturnal. Thus, the reverse is obtained with the same stimulus. Circadian disruption, the situation where there is dyssynchrony between the endogenous clocks and external cues, affects human physiology and contributes to many disorders (Fig. 4).

Recent work has elucidated a new and essential cycle to the cell metabolism machinery. Using tools developed by Ionita and Antoulas, investigators uncovered and characterized a 12 h rhythm in cells that is independent of the circadian clock [6, 7]. Subsequent work defines the role of a single gene, XBP1s, as a master regulator of the 12 h clock, much the same as *Per*, in *Drosophila* and *CLOCK:BMAL1* in

| Circadian-Dependent Function | Consequence of dyssynchrony |
|-------------------------------------|------------------------------------|
| Sleep/wake, feeding/fasting | Sleep duration, jet lag |
| Blood vessel tone | Atherosclerosis, hypertension |
| Immune function | Altered inflammatory response |
| Fat storage | Altered feeding behavior |
| Metabolic functions | Fatty liver disease |
| Hormone secretion | Hormone-specific dysfunction |

Fig. 4 Investigations in animals and humans confirm the consequences of circadian disruption [5]

eukaryotic cells, the former recognized by the Nobel Prize in Physiology or Medicine in 2017 [8–10].

Indeed, the periodicity of biology has been known since the time of the ancients. The ebb and flow of the seasons, coupled with the repeating patterns of the stars and other celestial bodies, were ample proof that periodic phenomena are a feature of the natural life. Observations of the periodicity of biology were also evident to the earliest observers. Mammals had specific versions of estrus defining fertility, reptiles molted and grew, and all species were born, lived, and died. However, these rhythms, regardless of their importance to the organism, are not periodic in the mathematical sense. The adaptation of the mathematical notion of periodicity to repetitive functions in biology discloses aspects of biology that would not have been perceived without the analysis (demonstrated in the following section).

2 Mathematical Tools for Biological Oscillation Discovery

2.1 Introduction

The previous section describes the nature of biological rhythms, the complexities associated with their regulation, and the difficulty in segregating the independent non-circadian rhythms when dominant circadian rhythms are present. This section reviews the merit of the tools developed by Antoulas and collaborators in disclosing such non-circadian rhythms in an unbiased manner [6].

Smets and Bartholomay develop the notion of periodicity in the mathematical sense as a “function y such that $y(t) = y(t + p)$ for all t in the domain of y and $p > 0$ [11]. Biological precision does not mimic this, yet the notion of periodicity is pervasive and a very useful construct to explain observed phenomena. Smets and Bartholomay proceed to construct a series of expressions to describe what they call “bioperiodicity”, a less strict example describing a phenomenon that oscillates between upper and lower limits in a repetitive fashion. The expressions thus derived,

however, are from a noiseless case and specifically eliminate stochasticity. Moreover, it is those two latter restrictions that limit the usefulness of their mathematical expressions in the understanding of biological phenomena.

Stochasticity is a feature of many biological phenomena. Variations in the cellular chemical milieu, temperature, physical proximity, as well as variations in molecular interaction, create variability in a biochemical process that, under strict laboratory conditions, could be non-probabilistic. Stochasticity, though, has been a confounder for understanding the directionality of cell organization and function. Cell functions must be coordinated temporally and spatially; the right components must come together at the right time for the cell to thrive and reproduce. Mammalian cells do this, in part, by use of checkpoints and breaks that assure that a process is at the appropriate stage of repair before another one is started. Assessment of oscillations in biological function is challenged by detecting the underlying rhythm in the face of the underlying confounding stochasticity.

Noise is inherent in the measurement of biological phenomena. The errors and mistranslations occurring in the process of protein synthesis play a significant role in the genesis of disease [12]. Also, surprisingly, such errors may be propitious and provide a mechanism for the organism to develop superiority in reproduction and the game of life in general [1]. Noise, however, can be very confusing when trying to tease out a biological phenomenon from the background.

Finally, as laboratory investigation becomes more complex with sophisticated instrumentation and complex biologics, lab-to-lab variation appears more often [13].

2.2 Matrix Pencil Method for Studying Oscillations in Transcription

The fundamental tools of biological rhythm detection in gene transcription assume the presence of a rhythm. However, this is not always so. Indeed, there are many genes whose transcription is DC, either on or off. Others can be induced to oscillate in the presence of a transcription factor such as the oncogene MYC (Meena JK, Wang J, Dacso CC, unpublished). Therefore, it is useful to examine gene transcription as a signal and apply a signal processing methodology to it. This strategy does not elide the fundamental problems of biological variation, stochasticity, and noise. Thus, it is useful to seek a mathematical strategy that would find a rhythm if one existed and help to distinguish it from harmonics of underlying rhythms. Strategies for detecting oscillatory rhythms have been extensively reviewed [14, 15]. The Fourier analysis methods (e.g., the DFT) decompose a signal in terms of oscillatory components exclusively. Therefore, it is hard to know whether the oscillations obtained by means of DFT are actually present in the biological system or not. Furthermore, DFT-related methods yield orthogonality of the various components, without regard to biological significance. The methods have benefits and liabilities but when discovery is the goal, the matrix pencil method serves well by both being agnostic to the presence

or absence of a rhythm [6, 7] and having the capacity to distinguish the potential of independence by demonstrating orthogonality. This allows us to observe stochastic behavior of a transcriptional rhythm due to perturbations caused by activation of a master transcription program.

The matrix pencil allows the reliable computation of the dominant oscillations in biological time-series data, circadian or non-circadian. Even more, the matrix pencil method might not reveal oscillatory behavior, depending on the data. The eigenvalue pencil method does not force orthogonality on the oscillatory components. Due to the orthogonality, an oscillation is not affected by further computations. In other words, discovered oscillations are independent in contrast to the other methods. Previous methods of oscillation computation suggested that if a non-circadian oscillation is found, it will be considered a harmonic of the circadian rhythm. The eigenvalue pencil paved the way to discover a 12 h autonomous rhythm in cells that is independent of circadian cycle [7, 16, 17].

Figure 5 is an example of how the Eigenvalue pencil revealed aspects of oscillation in gene function that are not apparent when the raw data are observed. The experiment involved the collection of mice liver samples every two hours for 48 h, isolating the RNA at each timepoint and hybridizing the RNA on a microarray to reveal the number of individual RNA species. The top panels are microarray data obtained from livers of entrained mice [18]. Model fitting smooths the curves and begins to disclose aspects of the genes, but it is not until the eigenvalue pencil is applied to data that the superimposition of oscillations is clear.

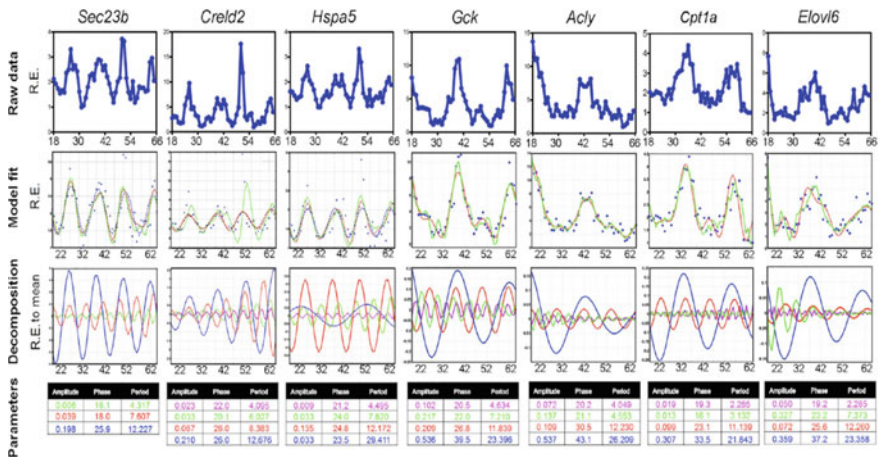


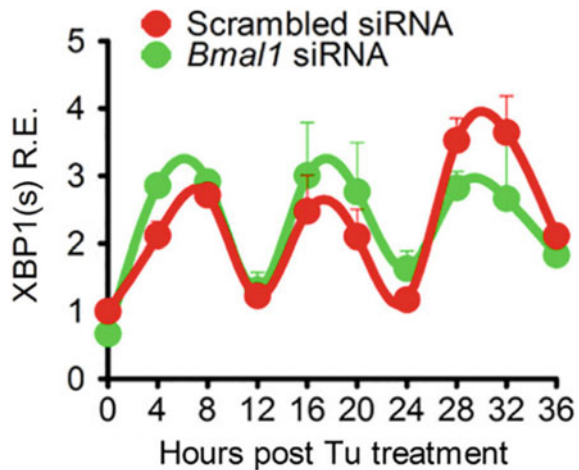
Fig. 5 Decomposition of gene oscillations by the eigenvalue/pencil method. Top row: raw microarray data; second row: fitting two different models (fifth-order approximation in dashed red and ninth-order approximation in solid green) to raw data (blue dots); third row: superimposed oscillations revealed by the ninth-order approximation; fourth row: amplitudes, phases, and periods of different oscillations with the color matching the different oscillations depicted in the third row [7]

This analysis allowed the detection of a suite of genes with 12 h cycles, but the question remained as to their independence from the dominant 24 h circadian clock. When examining the angles between the oscillations, several genes were discovered that were orthogonal to the 24 h cycle, thus suggesting independence [7, 16]. Further experiments that involved the deletion of a master regulator of circadian rhythm established the independence of a 12-h rhythm.

In the laboratory, it is possible to eliminate a gene (“knockout”) and thus see the effect on the whole animal from eliminating that regulatory function termed as “loss-of-function experiments”. To establish the independence of the newly discovered 12 h rhythm, cells were engineered that were deficient for BMAL1 gene (BMAL1 knockout), a core regulator of the circadian clock (24 h rhythm). When BMAL1 was knocked out, the 24 h circadian cycle was eliminated in the gene transcripts, but the 12 h rhythm remained. These experiments showed that perturbation of circadian rhythm regulator BMAL1 did not impact the periodicity of the 12 h oscillation regulator XBP1 (Fig. 6). This conclusively demonstrated a hitherto undescribed feature of transcription, a feature whose discovery depended on the use of the eigenvalue pencil in a field for which it was not originally designed.

Further use of eigenvalues is the depiction of a unit circle^{16, 17} whereby it becomes visually obvious that a point is oscillatory or not. Further, the construction of the unit circle describes an important feature of oscillating genes: what is their progress over time. The unit circle of eigenvalues offers a rapid and accurate method of distinguishing these features (Fig. 7).

Fig. 6 Data from Zhu et al., 2017. Gene expression data shows 12 h oscillation in the regulator gene XBP1. Control (red line) shows a robust oscillation in the expression of XBP1 transcript, which is not altered by knockout of circadian regulator Bmal1 (green line) [7]



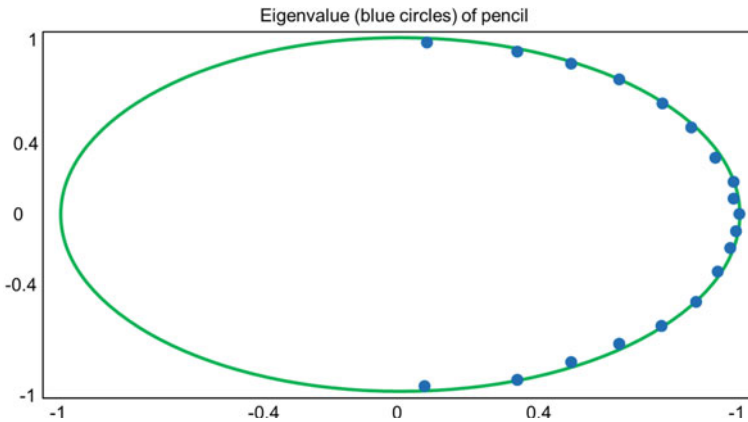


Fig. 7 Eigenvalues (blue dots) close to a unit circle (green line) where eigenvalue pencil method provides oscillatory poles (adapted from Antoulas et al. 2018) [16]

2.3 *Difficulty in Data Assessment and Analysis of Oscillation in Biological Data*

Other problems confound the detection of infradian and ultradian rhythms in biology. A major confounder is the synchronization of cellular clocks with the circadian clock, thus making it appear that the ultradian rhythms are harmonically dependent on the 24 h rhythm. Another important problem is the technical difficulty in performing experiments with sufficient granularity to overcome the Shannon-Nyquist limitations that determine the interval of points required to generate unambiguous curves [19]. A third problem is to interconnect the oscillations in different biological subsystems (e. g. links of transcriptional oscillation to the oscillation in the metabolic activity of a cell). Elucidating these problems will require a mathematical framework laid out by eigenvalue pencil and merging it with methodologies that analyze a dynamic network process such as graph signal processing, while also simultaneously testing the interpretation in a reasonable biological system.

3 Conclusion

In summary, we are just beginning to comprehend the complex biological systems through the generation of novel methodologies that involve innovative techniques and the application of mathematics. In this regard, the eigenvalue pencil method serves as a mathematical pipeline that allows the reliable computation of the dominant reduced-order models, thus informing a great deal about the oscillations present in a time-series type of data.¹⁶ Furthermore, applications of the eigenvalue pencil

method are broad. It is used to compute oscillation of non-circadian (12 h, XBP1-driven oscillation in transcriptome of mice liver), oscillation in the body temperature (Antoulas, Dacso, unpublished), non-circadian oscillations in transcriptome of cultured cells (Meena, Wang, Westbrook, & Dacso, in preparation), thus a significant advancement to both the mathematical and biological research.

It is fashionable in cell biology to append “omics” to a feature, thus describing its independence and utility. Genomics (sequencing of genomes), metabolomics (profiling of cellular metabolome), proteomics (profiling of cellular proteins), and so on now grace the scientific literature. These are also avenues where understanding the oscillation dynamics is a must for us to have a system-wide understanding. The addition of time to biological observations coupled with the eigenvalue pencil method augments the usefulness in “omics” such as the possibility of “temporal transcriptomics” or in the case of oscillatory functions in transcription the portmanteau word, “transcillation” (Jitendra Meena, Ph.D., personal communication). Jargon aside, temporal analysis is increasingly recognized as a critical feature of understanding the cell, organs, and organisms in general. To this end, new tools in mathematics, dynamic signal processing, and biological systems are required. A platform that allows the integration of the mathematical solutions to biological questions will be the key to solve the problems of stochasticity in functions as well as noise from biological variability. Application of the eigenvalue pencil to unpacking a biological function disclosed a new area of investigation that has the promise of significant explanatory power. Similar cross-disciplinary collaborations are likely to be fruitful and should be pursued by interested investigators.

References

1. Dobzhansky, T.: Nothing in biology makes sense except in the light of evolution. *Am. Biol. Teach.* **35**, 125–129 (1973)
2. Lakatta, E.G., Vinogradova, T.M., Maltsev, V.: The missing link in the mystery of normal automaticity of cardiac pacemaker cells. *Ann. N.Y. Acad. Sci.* **1123**, 41–57 (2008). <https://doi.org/10.1196/annals.1420.006>
3. Ryan, D.G., et al.: Coupling Krebs cycle metabolites to signalling in immunity and cancer. *Nat. Metab.* **1**, 16–33 (2019). <https://doi.org/10.1038/s42255-018-0014-7>
4. Goldbeter, A.: Dissipative structures and biological rhythms. *Chaos* **27**, (2017). <https://doi.org/10.1063/1.4990783>
5. Maury, E.: Off the clock: from circadian disruption to metabolic disease. *Int. J. Mol. Sci.* **20** (2019). <https://doi.org/10.3390/ijms20071597>
6. Ionita, A.C., Antoulas, A.C.: Matrix pencils in time and frequency domain system identification. In: Qiu, L., Chen, J., Iwasaki, T., Fujioka, H. (eds.) *Developments in Control Theory Towards Global Control IET Control Engineering*, pp. 79–87. The Institution of Engineering and Technology (2012).
7. Zhu, B., et al.: A cell-autonomous mammalian 12 h clock coordinates metabolic and stress rhythms. *Cell Metab* **25**, 1305–1319 e1309 (2017). <https://doi.org/10.1016/j.cmet.2017.05.004>
8. Pan, Y., et al.: 12 h clock regulation of genetic information flow by XBP1s. *PLoS Biol.* **18**(1), (2020). <https://doi.org/10.1371/journal.pbio.3000580>
9. Bargiello, T.A., Jackson, F.R., Young, M.W.: Restoration of circadian behavioural rhythms by gene transfer in *Drosophila*. *Nature* **312**, 752–754 (1984). <https://doi.org/10.1038/312752a0>

10. Ewer, J., Rosbash, M., Hall, J.C.: An inducible promoter fused to the period gene in *Drosophila* conditionally rescues adult per-mutant arrhythmicity. *Nature* **333**, 82–84 (1988). <https://doi.org/10.1038/333082a0>
11. Smets, P., Bartholomay, A.F.: Repetitive pattern and biological periodicity—a mathematical interpretation. *Math. Biosci.* **10**, 333–351 (1971)
12. Moghal, A., Mohler, K., Ibba, M.: Mistranslation of the genetic code. *FEBS Lett* **588**, 4305–4310 (2014). <https://doi.org/10.1016/j.febslet.2014.08.035>
13. Yamada, K.M., Hall, A.: Reproducibility and cell biology. *J. Cell Biol.* **209**, 191–193 (2015). <https://doi.org/10.1083/jcb.201503036>
14. Zielinski, T., Moore, A.M., Troup, E., Halliday, K.J., Millar, A.J.: Strengths and limitations of period estimation methods for circadian data. *PLoS ONE* **9**,(2014). <https://doi.org/10.1371/journal.pone.0096462>
15. Hughes, M.E., et al.: Guidelines for genome-scale analysis of biological rhythms. *J. Biol. Rhythms* **32**, 380–393 (2017). <https://doi.org/10.1177/0748730417728663>
16. Antoulas, A.C., et al.: A novel mathematical method for disclosing oscillations in gene transcription: a comparative study. *PLoS ONE* **13**(9),(2018). <https://doi.org/10.1371/journal.pone.0198503>
17. Zhu, B., et al.: Unveiling “Musica Universalis” of the cell: a brief history of biological 12 h rhythms. *J Endocr Soc.* **2**(7), 727–752 (2018). <https://doi.org/10.1210/js.2018-00113>
18. Hughes, M.E., et al.: Harmonics of circadian gene transcription in mammals. *PLoS Genet* **5**,(2009). <https://doi.org/10.1371/journal.pgen.1000442>
19. Shannon, C.E.: Communication in the presence of noise. *Proc. IRE* **37**, 10–21 (1949). <https://doi.org/10.1109/JRPROC.1949.232969>

Model-Order Reduction for Coupled Flow and Linear Thermal-Poroelasticity with Applications to Unconventional Reservoirs



Horacio Florez, Eduardo Gildin, and Patrick Morkos

Abstract This work focuses on the development of model reduction workflows for coupled flow and geomechanics arising in Ultra-Low Permeability (ULP) reservoir simulation. ULP challenges conventional simulators because they require multi-physics couplings, e.g., flow, thermal, and geomechanics couplings, which poses a severe burden regarding computational efforts. We tackle this problem by implementing a workflow for a two-step Proper Orthogonal Decomposition/Discrete Empirical Interpolation Method (POD-DEIM) model reduction approach for flow and geomechanics. More specifically, we perform the standard offline training stage on displacements as primary variables to create a basis for each primary variable using POD. During the online phase, we project the residual and Jacobian that arise from both poroelasticity and rate-independent poroelasticity into the given basis to reduce one-way coupled flow and geomechanics computations. We approximate the tensors, for the energy equation, to minimize the serial-time. We consider the role of the heterogeneity and material models such as Von Mises and investigate the benefits of hyper-reduction via DEIM on the nonlinear functions. Our results, which focus on linear and nonlinear thermo-poroelasticity, show that our Model-Order-Reduction (MOR) algorithm provides substantial single and double digits speedups, up to 50X if we combine with multi-threading assembling or DEIM and perform MOR on both physics.

Keywords Model reduction · Geomechanics · Porous media flow · POD-DEIM · Reservoir simulation

H. Florez
Deepcast.ai, Houston, TX, USA

E. Gildin (✉) · P. Morkos
Petroleum Engineering Department, Texas A&M University, College Station, TX, USA
e-mail: egildin@tamu.edu

P. Morkos
e-mail: patrickmorkos@tamu.edu

1 Introduction

Unconventional or ULP reservoirs concentrate the bulk of hydrocarbon production activity in the US. Rapid assessment of the subsurface integrity and environmental impacts of hydraulic fracturing (fracking) campaigns used in these reservoirs are paramount for sustainable production in unconventional reservoir developments. Geomechanical effects are often encountered in subsurface flow problems, and they can affect reservoir deformation, well integrity and flow response. The geomechanical behavior of hydraulic-fracturing is of importance to environmental effects as it can be used to model induced microseismicity and caprock integrity. Coupled flow and geomechanics simulations can model such systems, but they are computationally expensive. Coupled flow and geomechanics simulations are the basis for such developments. However, these simulation models are highly elaborate and require solving nonlinear algebraic systems of millions of equations and unknowns. Such simulations allow assessing the induced stress changes that hydrocarbon production or the injection of fluids in a reservoir (RS) induce in the surrounding rock mass. These studies often involve compaction and subsidence that could produce harmful and costly effects such as in wells casing, cap-rock stability, faults reactivation, and environmental issues as well.

Computational demands are one of the main drivers in selecting the proper model resolution and coupling mechanisms in such reservoirs. The multiscale nature of the geological formation requires solving discretized models of millions of gridcells for flow only—usually solved by finite volume techniques. In addition, coupled flow with another set of higher number of unknowns for the geomechanics (usually discretized by finite element techniques) may extend the domain beyond the flow-reservoir model. This is especially daunting in uncertainty quantification (UQ) and production optimization frameworks where multiple runs of the coupled problem are required to explore the full spectrum of possible scenarios. Therefore, it becomes necessary to obtain some form of reduced complexity models, e.g., using model-order reduction (MOR), to mitigate the high computational cost of such simulations, and eventually allow for real-time decision-making during the production cycle of such reservoirs.

MOR has been successfully applied for reservoir simulation (flow only) to alleviate the high computational cost [1–5] but, its application to coupled multiphysics, as in the problem herein, has been minimal. MOR, in this setting, has been developed based on the so-called *snapshot-based* MOR, whereby the reduction is attained by the POD method [6]. Simulation of single- and two-phase flows in porous media, as in the case of the work presented here, is a highly nonlinear process due to the nature of the fluid-rock interaction. In order to handle such non-linearities, POD-DEIM frameworks [7, 8] and the Trajectory-Piecewise-Linear (POD-TPWL) have been implemented [2]. For the problem herein, we rely on a global-in-time POD implementation whose projection matrix remains constant in time, since ULP exhibits a monotone behavior that does not require a local-in-time POD basis as in [9–11]. This is to say that reservoir pressure behavior is smooth and its dynamics can be

expressed by snapshots taken directly for the entire time span as opposed to taking localized-in-time snapshots to capture varying dynamics. However, it is well known that POD heavily depends on the set of collected snapshots obtained in the offline step of the algorithm, and judicious selection of such snapshots is an important step in our framework.

There exists a vast literature of model reduction for structural mechanical problems [12, 13], and in particular, for the structure persevering MOR for second-order systems whereby, the mass, stiffness and damping matrices are preserved as in the original form [14]. It has been, however, not the path taken in the area of geomechanics. The reader can find a helpful summary of recent MOR developments in geomechanics in [10, 11]. Particular applications of POD-based MOR techniques for mechanics are considered in [15–18]. To the best of our knowledge, [10, 11, 19] were about the first applications of MOR for coupled flow and geomechanics using POD-like approaches. In [20] a MOR method for coupled flow and geomechanics is developed within the POD-TPWL framework. These references include bridges between POD and Newton/Krylov solvers, homogenization, and domain decomposition. In order to tackle the non-linearities, we use the DEIM approach [7, 8]. DEIM relies on projecting into an oblique subspace that is built based on snapshots of the primary variables. However, the original papers consider the Galerkin projection; we instead utilize the Petrov-Galerkin projection that is more robust [9, 10, 21]. It is important to notice that DEIM approximates the tensors by using a subspace based on snapshots of a representative nonlinear term, not the primary variables. For rate-independent plasticity, we rely on the incremental finite element formulation whose numerical algorithm usually defines an implicit function for integration of the rate constitutive equations of the model. Fully-linearized finite element models have been used for path-dependent materials [22]. Here, the authors relied on linearization of the incremental stress updating procedure, rather than the rate stress-strain tangential relation, to achieve quadratic rates of convergence in the gradual finite element process. Detailed derivations can be found in [23, 24].

In this work, we continue our efforts [11] where a POD-based projection MOR scheme for coupled single-phase flow and thermal-poroelasticity attained substantial speedups, up to 50X. Here, we consider two-phase flow and add the DEIM based hyper-reduction for the thermal and mechanic parts. We see that we can still attain tremendous speedups. We presented an early version of this research in a conference paper [25]. The present work revises this last one and introduces a two-step DEIM process that, from our numerical experiments, improves the speedup compared with earlier findings [26]. We organize the remainder of the paper as follows. Section 2 presents the mathematical models of the governing partial differential equations. In particular, Sect. 2.1 revises the discretization of the flow equations and briefly comments on the MOR model for flow. In Sect. 3 we propose the MOR method and also shortly discuss POD-DEIM properties. Then, we present a numerical example of our MOR algorithms in Sect. 4. Finally, in the last sections we state concluding remarks, future work, and acknowledgments respectively.

2 Mathematical Model for Thermo-Poroelasticity

We discuss the governing equations for linear homogeneous isotropic thermo-poroelasticity and their finite element (FE) formulation. For the sake of brevity, we provide most relevant information and refer to [24, 27–29] for detailed and more complete description. We consider a bounded domain $\Omega \subset \mathbb{R}^3$ and its boundary is $\Gamma = \partial\Omega$. The time interval of interest $]0, \mathfrak{S}[$. We describe flow and transport models in porous media by a set of partial differential equations describing the conservation of mass as a function of pressure, saturation, and temperature. For simplicity, we neglect the inertial and temperature effects and focus on the two-phase oil-water system. Thus the mass balance equation for each phase is:

$$\nabla \cdot (\rho_\gamma \mathbf{v}_\gamma) + \frac{\partial (\rho_\gamma \phi S_\gamma)}{\partial t} - \rho_\gamma q_\gamma = 0 \quad \text{in } \Omega \times]0, \mathfrak{S}[, \gamma = \{w, o\} \quad (1)$$

where \mathfrak{S} is the simulation time, ρ_γ is the fluid phase density, \mathbf{v}_γ is the fluid phase superficial velocity, ϕ is the porosity of the rock, S_γ denotes the fluid phase saturation, q_γ is the volumetric source/sink term and the subscript $\gamma \in \{o, w\}$ indicates the oil and water phases, respectively. We apply Darcy's law:

$$\mathbf{v}_\gamma = -\frac{k_{r\gamma}}{\mu_\gamma} \mathbf{K} \cdot (\nabla p_\gamma - \rho_\gamma g \nabla h) \quad \text{in } \Omega \times]0, \mathfrak{S}[, \quad (2)$$

where \mathbf{K} denotes the permeability tensor, μ_γ is the fluid phase viscosity, $k_{r\gamma}$ denotes the relative permeability of each phase (which is a function of water saturation), p_γ is the phase pressure, g is the constant of gravity acceleration and h denotes depth. Combining Eqs. (1) and (2) yields

$$\nabla \cdot \left[-\frac{\rho_\gamma k_{r\gamma}}{\mu_\gamma} \mathbf{K} \cdot (\nabla p_\gamma - \rho_\gamma g \nabla h) \right] + \frac{\partial (\rho_\gamma \phi S_\gamma)}{\partial t} - \rho_\gamma q_\gamma = 0. \quad (3)$$

The general oil-water model is completed by enforcing the saturation constraint $S_o + S_w = 1$ and by specifying a capillary pressure relationship $p_c(S_w) = p_o - p_w$. As we can see, Eq. (3) is nonlinear. We take $p = p_o$ and S_w to be the primary unknown variables, from which p_w and S_o can be easily computed. We can solve Eq. (3) numerically using a several time-discretization schemes, but here we implement the fully-implicit (i.e., implicit pressure and saturation) finite volume procedure for the two-phase flow solution. One should note that discretization of the geomechanics equations is based on the finite element method. The reader can find more details in [30, 31], where other time discretization methods, such as implicit pressure-explicit saturation (IMPES) is described. Finally, the porosity ϕ is considered a function of the following variables:

$$\phi = \phi^0 + \alpha \cdot (\nabla \cdot \mathbf{u} - \varepsilon_v^0) + \frac{1}{M} (p - p^0), \quad (4)$$

where the additional parameters are, accordingly, α which is the Biot's constant, \mathbf{u} represents the displacement vector, while ε_v^0 is the initial volumetric strain. Herein M is the Biot's modulus, see [32], while ϕ^0 and p^0 denote a reference or initial state. The common boundary conditions (BCS) for Eq. (1) imply no-flow. We should also consider an initial or reference pressure and saturation distribution in the whole domain. We begin from the equilibrium equation for a quasi-steady process for the mechanics part:

$$\begin{aligned} -\nabla \cdot \sigma &= \mathbf{f}_u \text{ in } \Omega \\ \mathbf{u} &= 0 \text{ on } \Gamma_D^u \\ \mathbf{t} &= \sigma \cdot \hat{\mathbf{n}} \text{ on } \Gamma_N^u \end{aligned} \tag{5}$$

where σ is the stress tensor, \mathbf{f}_u corresponds to the vector of body forces, such as gravity, for instance. Herein $\hat{\mathbf{n}}$ is the outer unitary normal vector as usual. We can decompose the BCS Γ , in Dirichlet type, i.e., Γ_D^u , and Neumann type BCS, i.e., Γ_N^u , and the external tractions are known or prescribed. Hooke's law combined with Biot's poroelastic theory defines σ by the following expression:

$$\sigma = \mathbf{C} : \varepsilon - [\alpha (p - p^0) + 3K\beta(T - T^0)] \delta ; \mathbf{C} = \lambda \delta \otimes \delta + 2G\mathbf{I}, \tag{6}$$

where ε is the strain vector, $T = T(x, t)$ is the temperature, T^0 is a reference temperature, \mathbf{C} is the elastic moduli, β corresponds to the coefficient of thermal dilatation while K is the bulk modulus. The Kronecker delta is denoted by δ while λ , and G , are the Lamé constants, and \mathbf{I} represents the fourth-order identity tensor.

Let \mathcal{T}_h be a non-degenerate, quasi-uniform conforming partition of Ω composed of convex elements. We can derive a finite element formulation for Eq. (5) upon multiplying it by a virtual displacement and integrating it by parts. We omit details herein for the sake of conciseness but refer the reader to [24, 28, 33] for a detailed treatment. We take a finite-dimensional subspace of the continuous Sobolev spaces [24, 28, 34] as FE space, thus:

$$\mathcal{C}_k(\mathcal{T}_h) = \{v \in L^2(\Omega) : \forall e \in \mathcal{T}_h, v|_e \in \mathbb{P}_k(e)\}, \tag{7}$$

where $\mathbb{P}_k(e)$ represents the space of polynomials of total degree less than or equal to k , $\mathcal{C}_k(\mathcal{T}_h)$ are continuous along the given element's edges. We introduce the next notation to improve readability of the derivations to follow, $(\cdot)_+ \equiv (\cdot) + 1$, which operates on superscripts, for instance, $n_+ \equiv n + 1$, where n and $n + 1$ indicate consecutive time levels. We also define the linear operator, $\Delta(\cdot)^i_j \equiv (\cdot)^i - (\cdot)^j$, that defines the discrepancy between different instances of a vector, i.e., $\Delta \mathbf{x}_i^{n_+} \equiv (\mathbf{x}^{n_+} - \mathbf{x}^i)$; for consecutive snapshots the operator reduces to $\Delta \mathbf{x}^{n_+} \equiv (\mathbf{x}^{n_+} - \mathbf{x}^n)$ [5].

We obtain the loose coupling approach between flow and mechanics in different ways. More information can be retrieved from [35] and references therein. Equation (8) shows one possible choice, where one solves the displacements first by taking the pressures from the previous time step. Next, one updates the pressures by using the newest displacements:

$$\mathbf{K}_u \cdot \mathbf{u}^{n+} = \mathbf{f}_u + \mathbf{Q}_u (\Delta \mathbf{p}_0^n) \quad (8)$$

where \mathbf{p} is the discretized pressures, \mathbf{K}_u is the stiffness matrix. We provide the expressions for \mathbf{K}_u and \mathbf{Q}_u in [24, 28, 33]. One can define an iterative coupling scheme in different ways, but they all derive from the loose coupling approach by incorporating an internal iteration to update lagged quantities. For further details, please refer to [4, 29, 36]. Also notice that for thermal stresses, one can derive an equivalent pressure drop, after Eq. (6), that renders Eq. (8) unchanged as shown by [33]. The next subsection present the discretization of the nonlinear flow equations to solve for the pressure and saturation. A similar approach can be taken for the energy equations and temperature fields, but for the sake of space, we leave them out. The reader is referred to [11, 24].

2.1 Discretization of the Flow Equations

The flow equation as described in Eq. 3 is discretized using a finite volume approach. The reader can refer to [37] for more detailed exposition of the space and time discretization methods used in reservoir simulation. Most commercial reservoir simulators rely on the following system of discrete equations which for the wetting phase reads [30]:

$$\frac{\Delta(\phi \rho_\gamma \mathbf{S}_\gamma)^{n+}}{\Delta t^n} + \text{div}(\rho \mathbf{v})_\gamma^{n+} = \mathbf{q}_\gamma^{n+} \quad (9)$$

where \mathbf{q}_γ^{n+} has been converted to appropriate volumetric units as compared to Eq. 1 and Darcy's law implies:

$$\mathbf{v}_\gamma^{n+} = \frac{\mathbf{K} k_{r\gamma}}{\mu_\gamma^{n+}} [\text{grad}(\mathbf{p}_\gamma^{n+}) - g \rho_\gamma^{n+} \text{grad}(z)] \quad (10)$$

where $\mathbf{S}_\gamma, \mathbf{p}_\gamma \in \mathbb{R}^{n_b}$ denote the vector with, one saturation, and one pressure value per cell, respectively, while ϕ is a diagonal matrix with the cellwise porosity values, and n_b is the number of cells in the mesh and z is the discretized depth. We remark that both porosity and density are functions of the pressure. We thus can rewrite equation (9) as:

$$\mathcal{R}_\gamma \equiv \frac{\Delta \eta_\gamma^{n+}}{\Delta t^n} + \text{div}(\rho \mathbf{v})_\gamma^{n+} - \mathbf{q}_\gamma^{n+} = \mathbf{0}, \quad (11)$$

where $\eta \equiv (\phi \rho_\gamma \mathbf{S}_\gamma)$ is an algorithmic variable. The Jacobian would be:

$$\mathcal{J}_\gamma = \frac{\partial \mathcal{R}_\gamma}{\partial \mathbf{x}} = \frac{1}{\Delta t^n} \frac{\partial \Delta \eta_\gamma^{n+}}{\partial \mathbf{x}} - \frac{\partial (\text{div}(\rho \mathbf{v})_\gamma^{n+} - \mathbf{q}_\gamma^{n+})}{\partial \mathbf{x}}, \quad (12)$$

where the state vector $\mathbf{x} \equiv (\mathbf{p}, \mathbf{S}_w)^T \in \mathbb{R}^N$ and $N \equiv 2 \cdot n_b$. Thus for the oil-water system (9), the global-in-space tensors become:

$$\mathcal{R}^{(\mathbf{x})} \equiv \left\{ \begin{array}{c} \mathcal{R}_o \\ \mathcal{R}_w \end{array} \right\}; \mathcal{J}^{(\mathbf{x})} \equiv \left\{ \begin{array}{c} \mathcal{J}_o \\ \mathcal{J}_w \end{array} \right\} \quad (13)$$

where $\mathcal{R}^{(\mathbf{x})} \in \mathbb{R}^N$ and $\mathcal{J}^{(\mathbf{x})} \in \mathbb{R}^{N \times N}$. Section 3.1 (Algorithm 2) covers in detail how we propose to reduce the computational cost to solve the nonlinear systems that arise from the discretization in space and time of coupled flow and geomechanics equations.

3 Proposed MOR Method

We introduce here a regularized Petrov-Galerkin Newton-Raphson (PGNR) algorithm. We start with the global MOR method and then comment about obtaining an oblique subspace via POD. Since we consider staggered one- or two-way coupled systems here, i.e., we do not solve all variables simultaneously; we present a general treatment that segregates the primary variables, namely, pressure, water-saturation, temperature, and displacements. We create a separate POD basis for every variable and refer to a generic variable; namely, $\omega \in \{\mathbf{x}, \mathbf{u}_h, \mathbf{T}_h\}$, where \mathbf{x} is cell-centered while $(\cdot)_h$ indicates the FE approximation of those quantities.

3.1 Global MOR Algorithm

After discretizing the governing equations of the problem of interest, we end up with a parameterized nonlinear dynamic computational model described by the large-scale system of algebraic equations denoted by $\mathcal{R}^{(\omega)}$:

$$\mathcal{R}^{(\omega)}(\omega^{(n_+)}; \omega^{(n)}; \dots; \omega^{(0)}; \underline{\chi}_\omega) = \underline{0}, \quad (14)$$

here $\omega^{(\ell)}$, $\ell = 0, \dots, (n_+)$ are successive solutions (or snapshots) at timesteps ℓ and only $\omega^{(n_+)}$ is unknown, $\omega^{(\ell)} \in \mathbb{R}^{N^{(\omega)}}$ are the solutions to equation (14). Herein $N^{(\omega)}$ is the number of unknown degrees-of-freedom (DOF) for every variable, i.e., the nonlinear system rank, $\underline{\chi}_\omega \in \mathbb{R}^{d_\omega}$ is the vector of input parameters such as material properties, initial condition and BCS, etc. For the problems of interest here, the function $\mathcal{R}^{(\omega)}: \mathbb{R}^{N^{(\omega)}} \times \mathbb{R}^{d_\omega} \rightarrow \mathbb{R}^{N^{(\omega)}}$ is nonlinear with respect to its first argument.

The Jacobian $\mathcal{J}^{(\omega)}$ of the nonlinear system $\mathcal{R}^{(\omega)}$ is defined by a matrix of $N^{(\omega)} \times N^{(\omega)}$ partial derivatives

$$\mathcal{J}_{ij}^{(\omega)} \equiv \frac{\partial \mathcal{R}_i^{(\omega)}}{\partial \omega_j^{(n_+)}}; i, j = 1, \dots, N^{(\omega)}. \tag{15}$$

Model reduction aims at projecting equations (14) and (15) to a smaller subspace spanned by the truncated solution snapshots of Eq. (8). Many implementations have been investigated for reservoir simulation, i.e., multi-phase flow as in Eq. (9) and in recent years they have been extended to multiphase flow by [1–4]. In the particular case of solving the nonlinear system of equations, as shown above, we employ the Newton-Raphson approach, and we know that it is a local method. Thus, its convergence relies on the choice of the initial point. Providing an appropriate initial guess for a given problem may be cumbersome, so we develop a modified strategy to the Newton iteration by adding a globalization strategy to improve robustness. The line-search procedure below seeks convergence regardless of the initial point. Towards this end, we introduced a suitable merit function in [38], which guarantees that Newton’s direction guides to a solution:

$$M(\tilde{\omega}) = \|\mathcal{R}^{(\omega)}(\Xi(\tilde{\omega}))\|^2, \Xi(\tilde{\omega}) = \omega_o + \Phi_\omega \tilde{\omega}, \tilde{\omega} \in \mathbb{R}^{\tilde{n}^{(\omega)}}. \tag{16}$$

Here $\Phi_\omega \in \mathbb{R}^{N^{(\omega)} \times \tilde{n}^{(\omega)}}$ is the subspace matrix and $\tilde{n}^{(\omega)}$ is the reduced space dimension. The PGNR Algorithm 1 proposed in [9, 38] yields: find $\omega \in \mathbb{R}^{N^{(\omega)}}$, $\mathcal{R}^{(\omega)}(\omega) = \mathbf{0}$. We rediscovered the Levenberg-Marquardt regularization (ι in Algorithm 1) and proposed such as algorithm in [9, 38], also a readable summary can be found at [11]. For the sake of brevity, we only note that the term $\iota \mathbf{I}$ and $(1 + \varrho \|\mathcal{R}^{(\omega)}\|)$ with $\varrho \in (0, 1)$ are regularization parameters. We enumerate the steps below.

Algorithm 1: Petrov-Galerkin steps

- 1 Solve for Newton increment: $(\Theta_\omega^T \Theta_\omega + \iota \mathbf{I}) \cdot \Delta \tilde{\omega} = -\Theta_\omega^T \frac{\mathcal{R}^{(\omega)}}{1 + \varrho \|\mathcal{R}^{(\omega)}\|}$;
 - 2 Decompression step: $\Delta \omega = \Phi_\omega \Delta \tilde{\omega}$;
 - 3 Sufficient decrease (globalization): Find $\varsigma \in (0, 1]$;
 $\|\mathcal{R}^{(\omega)}(\omega + \varsigma \Delta \omega)\| < g(\varsigma) \|\mathcal{R}^{(\omega)}(\omega)\|, \quad g(\varsigma) < 1$;
 - 4 Update: $\omega = \omega + \varsigma \Delta \omega$;
-

It should be noted that this is a global algorithm that retains its q -quadratic rate of convergence moving on the affine subspace $\omega_o + \Phi_\omega \tilde{\omega}$, $\tilde{n}^{(\omega)} \ll N^{(\omega)}$ where $\Theta_\omega = \mathcal{J}^{(\omega)} \cdot \Phi_\omega$. We present the algorithm’s pseudo-code below. Notice that for linear elasticity the residual and Jacobian reduce to (see Eq. (8)):

$$\begin{aligned} \mathcal{R}^{(u_h)} &= \mathbf{K}_u \cdot \mathbf{u}^{n_+} - \mathbf{f}_u + \mathbf{Q}(\mathbf{p}^{(n)} - \mathbf{p}^{(0)}) \\ \mathcal{J}^{(u_h)} &= \mathbf{K}_u. \end{aligned} \tag{17}$$

Algorithm 2: The pseudo-code depicts the PGNR algorithm**Data:** $\omega = \omega^0$, ϵ^{TOL} , δ^{TOL} and K^{MAX} , $bMOR$ and Φ_ω , $bRegld_1$, $bLSearch$ **Result:** Solution to $\mathcal{R}^{(\omega)}(\omega) = \mathbf{0}$ or error message

```

1  $\epsilon = 1.0$ ,  $\varsigma = 1.0$ ,  $\delta = 1.0$ ,  $k = 0$ ,  $bRegld = (bRegld1 \parallel bRegld2)$ ;
2 while  $((\epsilon > \epsilon^{TOL}) \parallel (\delta > \delta^{TOL})) \&\& (k < K^{MAX})$  do
3   if  $(k == 0)$  then
4      $[\mathcal{R}^{(\omega)}, \mathcal{J}^{(\omega)}] = ResidualJacobian(\omega)$ ;
5   else
6      $\mathcal{J}^{(\omega)} = Jacobian(\omega)$ ;
7   if  $bMOR$  then
8      $\Theta_\omega = \mathcal{J}^{(\omega)} \cdot \Phi_\omega$ ,  $\Delta\omega = (\Theta_\omega^T \cdot \Theta_\omega)^{-1} \cdot (\Theta_\omega^T \cdot \mathcal{R}^{(\omega)})$ ;
9      $\Delta\omega = \Phi_\omega \cdot \Delta\omega$ ;
10  else
11    if  $bRegld$  then
12       $\iota = \rho \|\mathcal{R}^{(\omega)}\|$ ;
13      if  $bRegld1$  then
14         $\mathcal{J}^{(\omega)} = \mathcal{J}^{(\omega)} + \iota \mathbf{I}$ ;
15      if  $bRegld2$  then
16         $\mathcal{R}^{(\omega)} = \frac{\mathcal{R}^{(\omega)}}{1 + \iota}$ ;
17     $\Delta\omega = (\mathcal{J}^{(\omega)})^{-1} \cdot \mathcal{R}^{(\omega)}$ ;
18  if  $bLSearch$  then
19     $[\varsigma, \mathcal{R}^{(\omega)}] = LineSearch(\omega, \Delta\omega, \mathcal{R}^{(\omega)})$ ;
20   $\omega = \omega - \varsigma \Delta\omega$ ;
21  if  $bMOR$  then
22     $\delta = \|\Theta_\omega^T \cdot \mathcal{R}^{(\omega)}\|$ ;
23  else
24     $\delta = \|\mathcal{R}^{(\omega)}\|$ ;
25   $\epsilon = \varsigma \cdot \|\Delta\omega\|$ ,  $k = k + 1$ ;
26 if  $(k < K^{MAX})$  then
27   return  $\omega$ ;
28 else
29   print "Could not converge...";

```

In Algorithm 2 for simplicity, we take $\iota = \rho \|\mathcal{R}^{(\omega)}(\omega^{(k)})\|$. We employ as stopping criteria for the algorithm the norm of the residual, $\delta_k = \|\mathcal{R}^{(\omega)}(\omega^{(k)})\|$, the maximum number of iterations, and the difference between two successive iterations, $\epsilon_k = \|\omega^{(k)} - \omega^{(k-1)}\|$. The functions *ResidualJacobian()* and *Jacobian()* evaluate, for the specific problem, the residual and the Jacobian of the nonlinear system. We refer the reader to [9] for a full explanation of the above algorithm. Notice that DEIM impacts the functions *ResidualJacobian()* and *Jacobian()* where we must obtain those tensors only on the rows that the DEIM indices dictate. The subspace matrix Φ_ω is dense; thus the product $\mathcal{J}^{(\omega)} \cdot \Phi_\omega$ should exploit the Jacobian's sparsity to achieve performance as well as the product $\Theta_\omega^T \cdot \Theta_\omega$ since Θ_ω is also sparse.

Taking advantage of the sparsity would prove critical to reduce the overhead on such operations. The system of equations in line 8 is mostly full and tiny, its rank is $\tilde{n}^{(\omega)}$, which strongly suggest using a direct frontal solver for tackling it. Finally, notice that DEIM also affects *LineSearch* since we require to evaluate the residual there. It is also essential to check convergence upon projecting in the reduced space while running MOR as indicated in line 22. Indeed, part of the speedup appears because the reduced model performs fewer Newton-steps per timestep compared to the FOM.

3.2 Global-in-Time POD Algorithm (GPOD)

We can now propose our Global-in-Time (GPOD), POD-based algorithm for the flow Eqs. (9). We build an oblique subspace by performing POD analysis of snapshots of the combined primary variable \mathbf{x}^n ; thus the Petrov-Galerkin Newton-Raphson scheme implies [5, 9]:

$$\begin{cases} (\Theta_x^T \cdot \Theta_x) \cdot \Delta \tilde{\mathbf{x}} = \Theta_x^T \cdot \mathcal{R}^{(\mathbf{x})} \\ \mathbf{x}^{n+} = \mathbf{x}^n - \Phi_x \Delta \tilde{\mathbf{x}} \end{cases} \quad (18)$$

where $\Theta_x = \mathcal{J}^{(\mathbf{x})} \cdot \Phi_x$ and Φ_x is the subspace matrix (see Sect. 3.1). From here on, the notation $(\tilde{\cdot})$ refers to variables in the reduced space. We now present the Global-in-Time DEIM (GDEIM) approximation for the energy equation.

3.3 Two-Step Global-in-Time DEIM (GDEIM) Process

In the original DEIM paper [7], we notice that one can capture the snapshots of both primary variables and relevant nonlinear terms during the very same Full-Order Model (FOM) run. We refer to this approach as the one-step DEIM process. However, there are other possibilities. Indeed, it is possible to improve the affinity with the projection space if we, instead, capture the snapshots of nonlinear terms in a second step, after creating the oblique subspace. In other words, we capture the snapshots of the nonlinear terms at the decompressed states, so, we are filtering away high-frequency features that make such DEIM process expensive from the computational standpoint, i.e., deterioration of condition number the DEIM basis [26]. The alternative process that we refer as two-step GDEIM approach implies a first step in which we run the FOM to save snapshots of the primary variables. We then build an oblique subspace upon performing POD analysis on the variable ω . Once that we obtain, Φ_ω , we can then execute GPOD and store the snapshots of the nonlinear terms, which are more affine to Φ_ω . The whole idea entails improving the condition number of the resulting approximated DEIM Petrov-Galerkin system by removing unwelcome high-frequency features that tend to appear in the alternative one-step process. We

present the algorithm below. Preliminary numerical results suggest that the two-step process certainly improves the performance as aimed, but certain numerical deterioration while comparing to GPOD may remain.

Algorithm 3 depicts a modified DEIM approach (GDEIM). We changed our implementation software paradigm, in particular, memory allocation, to implement the above procedure. The function “RunFOM_wDEIM_Step1” runs the FOM and allocates memory to only storing snapshots of ω . Upon completion, we dump the snapshot matrix to disk in step 2 and perform POD analysis on line 3. We then run a variant of GPOD in line 4, in which we also allocate memory to store snapshots of the nonlinear terms, i.e., “RunGPOD_wDEIM_Step2”. The last two functions are specialized versions for each task. We save information about the nonlinear terms to disk in line 5 and perform POD analysis and build DEIM interpolants and selection matrices in step 6. Notice that the object “nlterm” handles the nonlinear terms. The implementation was mostly straightforward but required developing the specialized functions to allocate memory properly. The algorithm below focuses on building the POD and DEIM basis, running the test case FOM and GDEIM remain the same.

Algorithm 3: Two-step GDEIM process

```

/* 1st step: run FOM to capture  $\omega^{(n)}$  */
1 RunFOM_wDEIM_Step1(bWriteSol);
2 SaveSnpshts();
/* POD analysis on the primary variables... */
3 BuildPODBasis(dEnergy);
/* 2nd step: GPOD-only to capture nonlinear terms at
   decompressed states */
4 RunGPOD_wDEIM_Step2(bWriteSol);
5 nlterm.SaveSnpshts();
/* Build DEIM interpolants and selection matrixes */
6 nlterm.BuildDEIMInterp([dEnergy dEnergy]);

```

4 Numerical Example

We implemented these deterministic models through the one-way coupling between the Matlab Reservoir Simulator Toolbox (MRST) [30] and the Integrated Parallel Finite Element Analysis (IPFA), whose main features we described in [24, 39, 40]. MRST solves the flow equations via a fully-implicit approach based on cell-centered finite differences while IPFA employs continuous Galerkin finite elements to compute the induced poroelastic displacements and thus stresses. We hook up both codes by memory through a Matlab MEX interface which allows calling native C code from dynamic link libraries. IPFA employs standard continuous Lagrange polynomials as shape functions for the space discretization. All examples herein were run on a MacBook Pro laptop equipped with an Intel(R) Quad-Core(TM) i7-4870HQ CPU @ 2.5GHz and 16 GB of RAM. The authors chose this laptop for the sake of convenience, in particular, the availability of debugging tools free of

charge. Aside, one can achieve some level of parallelism due to the multi-core technology. All numerical simulations reported below utilized the Incomplete LU (ILU) as preconditioner and conjugate gradient as the iterative solver. We visualized in a postprocessor called “LogProc” which is a proprietary application, see [41].

Given the matrix $\Lambda^{(\omega)} \in \mathbb{R}^{N^{(\omega)} \times N^s}$, we define the root-mean-square norm (rms) as:

$$\|\Lambda\|_{rms}^{(\omega)} = \sqrt{\frac{1}{\tilde{n}^{(\omega)} \cdot N^s} \sum_{i=1}^{N^{(\omega)}} \sum_{j=1}^{N^s} \left(\Lambda_{ij}^{(\omega)}\right)^2}; \Lambda_{ij}^{(\omega)} = |\omega_{i(FOM)}^j - \omega_{i(ROM)}^j|, \quad (19)$$

and we employ this definition to compute the error between FOM and ROM for transient problems, where i refers to spatial discretization while j implies temporal stations. We also define the following running times for profiling purposes: $(\delta t)^*$, where $*$ = {ovl, asb, slv}, where the prefixes stand for overall, assembling, and solving runtimes. Here, $(\tilde{\delta t})^*$ and $(\delta t)^*$ refer to the computational time of the reduced and full-order systems, respectively. The ratio of assembling to overall running time per Newton iteration is:

$$r_{asb} = \frac{(\delta t)^{asb}}{(\delta t)^{asb} + (\delta t)^{slv}}. \quad (20)$$

We also express the speedup \mathcal{S} as the ratio between the FOM and ROM overall running times:

$$\mathcal{S} = \frac{(\delta t)^{ovl}}{(\tilde{\delta t})^{ovl}}. \quad (21)$$

We present an example involving the simulation of an ULP reservoir, including two-phase flow and geomechanics commonly encountered in West Texas. In this case, we reduce both physics to obtain substantial speedups and benefit from a two-way coupled flow and mechanics approach. Here, we start by updating the porosity after Eq. (4) and employ the fixed-stress approach in which we expect in every coupling step to iterate in this fashion to update lagged quantities and achieve convergence.

4.1 Example: Two-Phase Flow in an Unconventional RS

We revisit the ULP synthetic model that we presented in [10, 11] and whose geometry is shown in Figs. 1 and 2. The model consists of two fractured horizontal wells, separated by a distance equal to $2 \cdot s_w$. We apply a uniform pressure p_{bh} along the transverse fractures, which are divided by a distance of s_f . We hold the bottom-hole pressure constant throughout production. The ULP model is homogeneous, isotropic, poroelastic, and is bounded by layers with similar mechanical properties. Flow only

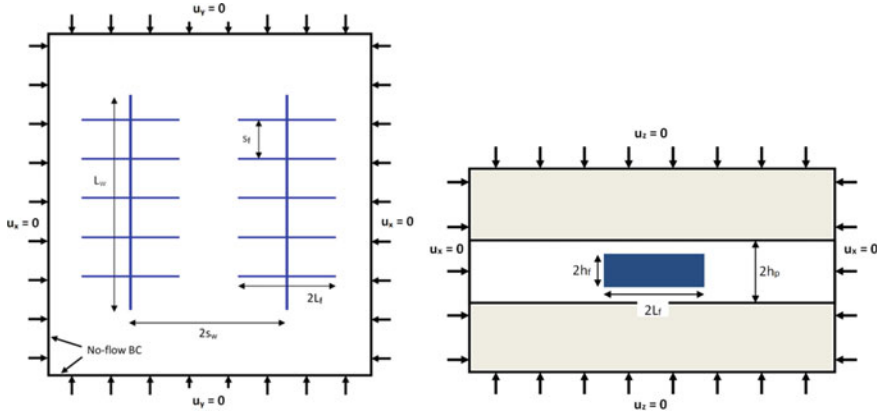


Fig. 1 Geometry and BCS in the horizontal and vertical planes, according to [42]

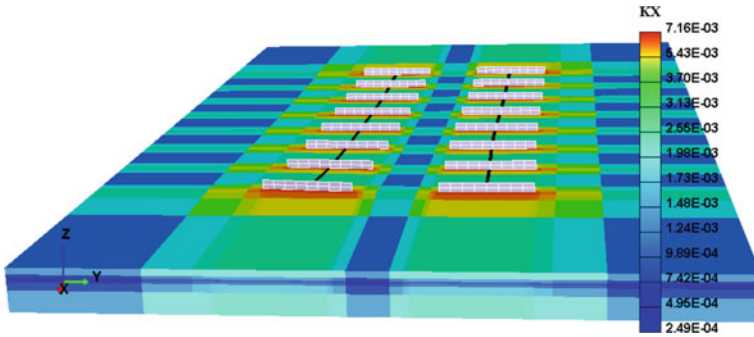


Fig. 2 Sample curvilinear mesh and permeability field

occurs within the reservoir and does not leak into the surrounding strata. We assume no-flow and far-field boundary conditions for each physics respectively. We studied perfectly transverse fractures in the offset well to simplify the analysis. We also know that the stresses in the infill region are impacted only by poroelastic effects as concluded by [42]. Table 1 compiles all relevant parameters for this two-phase flow simulation that mimics a representative case of liquid-rich shale development. We also developed a preprocessor which generates the given graded mesh. We vary the number of fractures to assess the performance of our ROM.

Figure 2 depicts a cut-away (upper half) representation that highlights the well trajectories in bold black and the fractures that are planar and perpendicular to the well path. The preprocessor generates a graded mesh and also populates properties by using the same blending rationale. The resulting permeability field is slightly heterogeneous to represent the stimulated reservoir volume best. As an illustration, Fig. 3 pictures the mesh and permeability field for a representative case that includes 16 fractures. We should emphasize that the data taken in Table 1 and the examples

Table 1 Unconventional RS model's parameters

| Parameter | Value [Unit] |
|------------------------------|----------------------------|
| Pay-zone half-height h_p | 400 ft |
| Fracture half-height h_f | 50 ft |
| Fracture half-length L_f | 200 ft |
| Stage spacing s_f | $\approx 428, 200, 158$ ft |
| Well length L_w | ≈ 3000 ft |
| Well spacing s_w | ≈ 1000 ft |
| Number of fractures | 8, 16, 20 |
| Matrix permeability k_m | 0.3 μ d |
| Fracture permeability k_f | 10 μ d |
| SRV's permeability k_{SRV} | 1.5 μ d |
| Young's modulus E | 2000 Ksi |
| Biot modulus M | 850 Ksi |
| Reservoir pressure p_R | 10000 Psi |
| Bottomhole pressure p_{bh} | 7000 Psi |
| Condensate viscosity μ | 0.25 cp |
| Porosity ϕ | 0.05 |
| Poisson's ratio ν | 0.2 |
| Biot coefficient α | 0.7 |

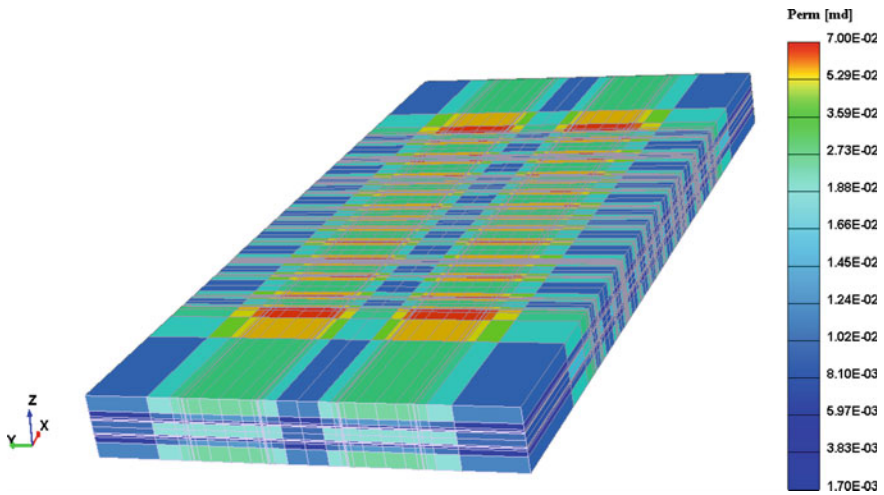


Fig. 3 Permeability field for the 16 fractures case

described here are realistic representation of hydraulic fracturing models used for unconventional reservoir simulation. We have not, however, compared our results with commercial-of-the-shelf simulators.

Regarding the oil-water system we assume that the fluid viscosity ratio is $\mu_w/\mu_o = 0.1$, and the relative permeability curves are [5]:

$$k_{ro}(S_w) = k_{ro}^0 \left(\frac{1 - S_w - S_{or}}{1 - S_{wr} - S_{or}} \right)^a, \quad (22)$$

$$k_{rw}(S_w) = k_{rw}^0 \left(\frac{S_w - S_{wr}}{1 - S_{wr} - S_{or}} \right)^b, \quad (23)$$

where k_{ro}^0 and k_{rw}^0 are the endpoint relative permeabilities, S_{or} and S_{wr} are the irreducible oil and water saturations, respectively. Herein, we set $S_{or} = S_{wr} = 0.2$, $k_{ro}^0 = k_{rw}^0 = 1$ and $a = b = 2$. The system is assumed to be slightly compressible, with compressibility $c_o = 1.0 \times 10^{-5} \text{Psia}^{-1}$, $c_w = 5.0 \times 10^{-7} \text{Psia}^{-1}$. The viscosity of the water phase is $\mu_w = 1$ cp, and the viscosity of the oil phase depends on pressure as $\mu_o = \mu_o^{ref} \cdot \exp[-c_{\mu_o} \cdot (p - p^{ref})]$, where $\mu_o^{ref} = 3$ cp, $c_{\mu_o} = 2.0 \times 10^{-6}$ and $p^{ref} = 14.7$ Psia. The initial water saturation is 0.2, and the initial pressure is in hydrostatic equilibrium with the top layer at 10000 Psia.

For the sake of simplicity, we consider the same domain for flow and mechanics. We ran a series of models with 8, 16 (see Fig. 3) and 24 fractures for four years with a fully implicit scheme with a timestep size of 10 days. We utilize multi-threading assembling and post-processing for the mechanics. We employ four threads that in average provide a 4X speedup over the mentioned serial tasks. Notice that we assemble the matrix \mathbf{K}_u only once at the initialization. Postprocessing, the stresses per iteration, can take up to 3 secs that add to the serial time which is around 15% for the overall computational time for the case of 8 fractures.

Let us provide a few comments on the numerical simulations associated with the ULP model. Figure 4 depicts pressure field snapshots after every year (4 years total) of production from left-to-right and top-to-bottom, respectively. We observe a reasonable symmetry and monotonic depletion scenario where the formation was primarily stimulated in the vicinity of the fractures. Naturally, in the event of having a higher density of fractures, we could increase the depletion and production rate. Figure 5 depicts the oil and water well's rates for this example. We observe the typical hyperbolic curves with a sudden depletion after 200 days where they become relatively flat. The barrels produce are larger than expected because the fracture thicknesses for this example are not realistic. However, it is still a meaningful example.

We coupled the mechanics with MRST for this model with 16 fractures, and run the one-way scheme (8) every six months. We observe a reasonably symmetric and monotone depletion scenario where we stimulated the formation only in the vicinity of the fractures due to the very small permeability values. Figure 6 depicts snapshots of the mean-stress σ_m that behaves proportionally to the pressure drop in the RS. It

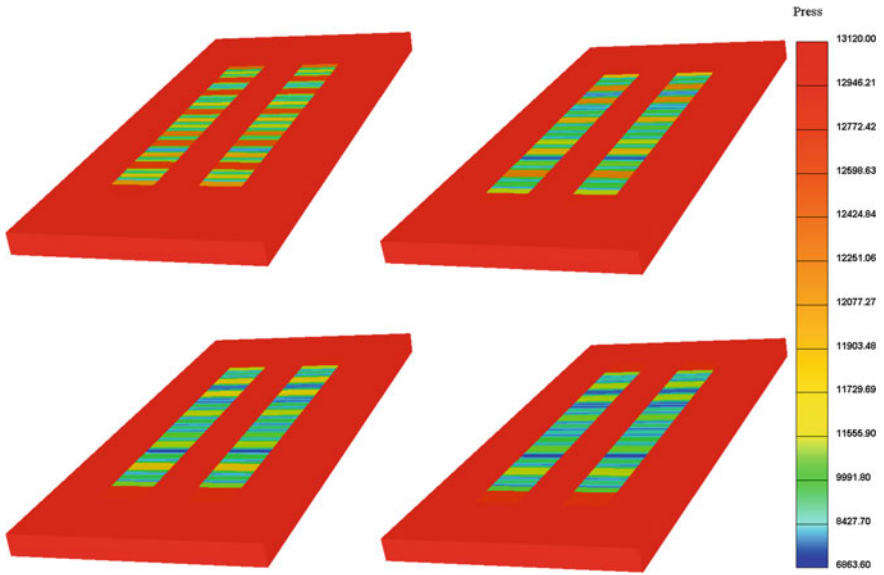


Fig. 4 Pressure snapshots for the case with 16 fractures

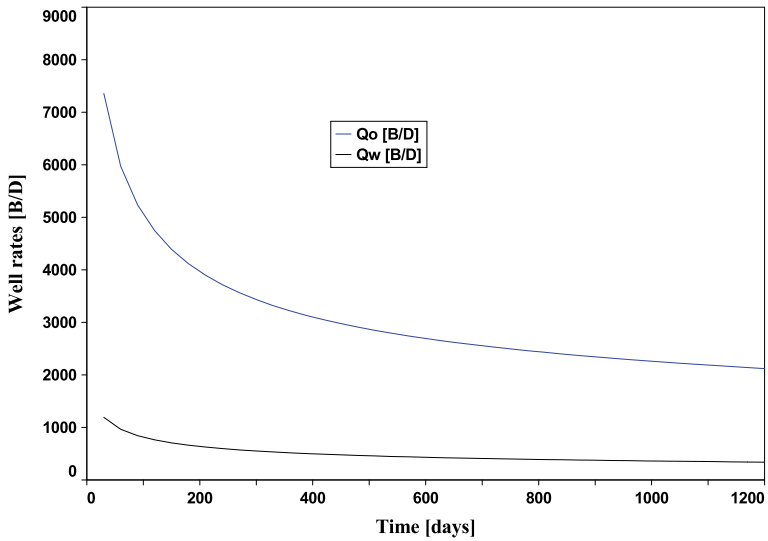


Fig. 5 Well's production curves for oil and water

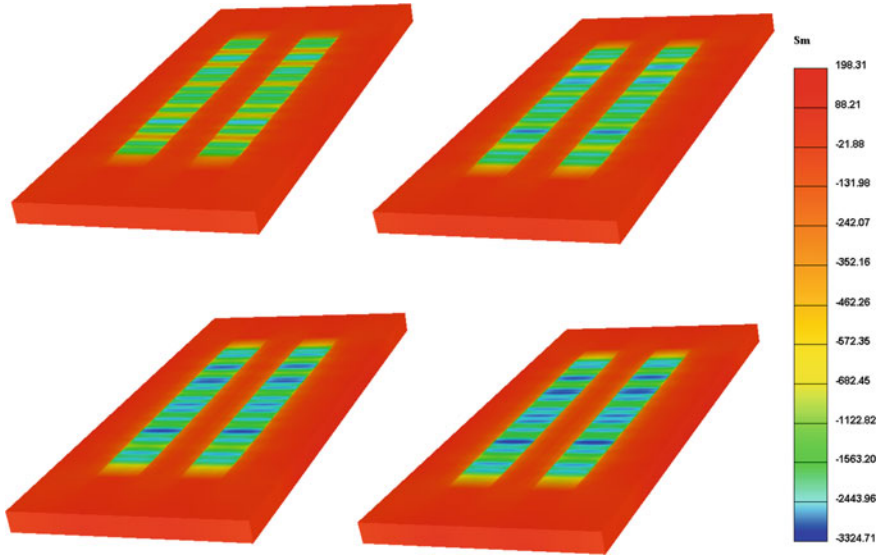


Fig. 6 Mean-stress snapshots for the case with 16 fractures

Table 2 Unconventional RS ROM’s performance: flow only

| # | $N^{(x)}$ | $\tilde{n}^{(x)}$ | $\tau_{\text{POD}}^{(x)}$ [%] | $(\delta t)^{\text{ovl}}$, FOM | $(\delta t)^{\text{ovl}}$, ROM | $\ \underline{\Delta}\ _{rms}^{(p)}$ | $\ \underline{\Delta}\ _{rms}^{(S_w)}$ | \mathcal{S} |
|----|-----------|-------------------|-------------------------------|------------------------------------|------------------------------------|--------------------------------------|--|---------------|
| 8 | 49152 | 8 | 93.3 | 19.2 | 0.70 | 0.0032 | 0.0001 | 27.5 |
| 16 | 98304 | 8 | 93.3 | 49.4 | 1.36 | 0.0035 | 0.0001 | 36.2 |
| 24 | 147456 | 8 | 93.3 | 124.9 | 2.45 | 0.0040 | 0.0001 | 50.8 |

seems that a third of the latter gets transferred as induced poroelastic stresses but only in the vicinity of the fractures. There are no significant stress shadows produced in the infill regions (Table 2).

For the MOR results ahead we consider 120 snapshots for flow, i.e., every ten days, while for mechanics, we couple every two months, i.e., $N^s = 20$ snapshots. Table (2) shows preliminary numerical results for the flow-only reduced model in which we attain substantial speedups as indicated. The results shown here are representative of our testing (online computations) different for training (offline). We ran a sensitivity study by varying the number of hydraulic fractures (#), and thus increasing the number of degrees of freedoms (DOF) for the full order model, namely $N^{(x)}$. Here, $\tilde{n}^{(x)}$ refers to the size of the reduced-order model. Indeed, the size of these problems dictates that the average solving ratio per Newton step is about 97%, which means that most of the time per iteration goes to solving the system of equations. Now, the Petrov-Galerkin projection only requires about 8 significant modes to solve an equivalent system, which provides tremendous speedups. We truncate the POD and

Table 3 Unconventional RS ROM's performance: coupled problem

| # | $N^{(u_h)}$ | $\tilde{n}^{(u_h)}$ | $\tau_{\text{POD}}^{(u_h)}$ [%] | $(\delta t)^{\text{OVI}}$, FOM | $(\delta t)^{\text{OVI}}$, ROM | $\ \underline{\Delta}\ _{rms}^{(u_h)}$ | \mathcal{S} |
|----|-------------|---------------------|---------------------------------|------------------------------------|------------------------------------|--|---------------|
| 8 | 78962 | 14 | 30.0 | 20.7 | 0.69 | 5.18e-8 | 30.0 |
| 16 | 157554 | 14 | 30.0 | 51.8 | 1.52 | 3.99e-8 | 38.2 |
| 24 | 236146 | 14 | 30.0 | 156.2 | 3.05 | 1.04e-7 | 51.2 |

DEIM modes based on a simple energy criterion as in [10, 11] that relies on the decay of their respective singular values. We did not try different compression ratios in the numerical experiments. Errors on the primary variables remain smaller than 0.001%, and the speedup improves with the problem size.

Table 3 compares the performance of the different reduced models for coupled flow and mechanics. In addition, we only varied the number of fractures to control the problem size and thus explore the performance of our MOR approach regarding the number of DOF. Notice that the Singular Value Decomposition (SVD) requires a significant amount of RAM and depends on these two numbers, N , and N^s . CR for mechanics is above 99.96% with three-quarters of significant snapshots for most cases. We observe from the runtime data that MOR plus multi-threaded provide a substantial speedup for all problems, and the latter improves with increasing problem size. Most of the speedup arises for savings in solving the sparse systems for both physics but mostly for two-phase flow. Errors are small, in particular, if we compare them with the magnitudes of the pressure and DISP. All of these ROM could adequately reproduce the FOM behavior, and they provide a substantial speedup, i.e., double digits. The speedup for the mechanics is slightly disappointing, e.g., about five times, if we compared with the case in which we couple with single-phase flow. There may be another issue as the performance of the communication between Matlab and IPFA as well as the fact that we are taking the same mesh for flow and mechanics, while we supposed to extend it beyond the pay-zone.

5 Conclusions

We presented herein a MOR algorithm that provides substantial single and double digits speedups, up to 50X if we combine it with multi-threading processing or DEIM and perform MOR on both physics, for one-way coupled problems involving thermo-poroelasticity. We highlight the remarkable MOR compression ratio above 99.9% for elasticity. The approach is particularly useful to speed up solving the sparse system for the inner iteration in convolution like problems which produces significant time savings compared to the serial FOM. The latter is also true for problems that exhibit long serial times, for instance, while assembling the Jacobian and Residual for both physics and post-processing to compute stresses, as long as

the serial time per iteration is shorter than solving the sparse system of equations. It is our observation that POD-DEIM certainly speeds up assembling the tensors, but there is another factor that emerges from the code profiling information. The time spent to solve the Petrov-Galerkin system tends to increase. Indeed, it seems that the DEIM approximation tends to deteriorate the condition number of the approximate system and thus we should expect longer runtimes while solving which decreases the speedup that otherwise we should attain with projection-only. Future work will provide a seamless framework to introduce a hyper-reduction scheme that does not increase the solving time considerably. We will also add more physics by coupling the two-phase flow code with rate-independent plasticity which we expect will behave similarly to the coupled nonlinear energy equation.

Acknowledgements The first author acknowledges “GeoFeMOR, LLC” for allowing access to IPFA and LogProc and permitting to publish these results. E.G acknowledges partial support through the Energi Simulation Research Chair at Texas A&M. We also acknowledge the thorough reading and the many improvements and suggestions made by the reviewers of this manuscript.

References

1. Ghommem, M., Gildin, E., Ghasemi, M.: Complexity reduction of multiphase flows in heterogeneous porous media. *SPE J.* (2015)
2. He, J., Durlofsky, L.J.: Reduced-order modeling for compositional simulation by use of trajectory piecewise linearization. *SPE J.* **19**(05), 858–872 (2014)
3. Tan, X., Gildin, E., Trehan, S., Yang, Y., Hoda, N. et al.: Trajectory-Based DEIM TDEIM model reduction applied to reservoir simulation. In: *SPE Reservoir Simulation Conference*, Society of Petroleum Engineers, (2017)
4. Yoon, H., Kim, J., et al.: Rigorous modeling of coupled flow and geomechanics in largely deformable anisotropic geological systems. In: *50th US Rock Mechanics/Geomechanics Symposium*. American Rock Mechanics Association (2016)
5. Tan, X., Gildin, E., Florez, H., Trehan, S., Yang, Y., Hoda, N.: Trajectory-based DEIM (TDEIM) model reduction applied to reservoir simulation. *Comput. Geosci.* **23**(1), 35–53 (2019)
6. Sirovich, L.: Turbulence and the dynamics of coherent structures, I–III. *Q. Appl. Math.* **45**(3), 561–590 (1987)
7. Chaturantabut, S., Sorensen, D.C.: Nonlinear model reduction via discrete empirical interpolation. *SIAM J. Sci. Comput.* **32**(5), 2737–2764 (2010)
8. Chaturantabut, S., Sorensen, D.C.: Application of POD and DEIM on dimension reduction of non-linear miscible viscous fingering in porous media. *Math. Comput. Modell. Dyn. Syst.* **17**(4), 337–353 (2011)
9. Florez, H., Argáez, M.: A model-order reduction method based on wavelets and POD to solve nonlinear transient and steady-state continuation problems. *Appl. Math. Model.* **53**, 12–31 (2018)
10. Florez, H., Gildin, E.: Model-order reduction applied to coupled flow and geomechanics. In: *Proceedings of the ECMOR XVI—16th European Conference on the Mathematics of Oil Recovery*, Barcelona, Spain, EAGE (2018)
11. Florez, H., Gildin, E.: Global/local model order reduction in coupled flow and linear thermoelasticity. *Comput. Geosci.* (2019)
12. Gildin, E., Antoulas, A.C., Sorensen, D., Bishop, R.H.: Model and controller reduction applied to structural control using passivity theory. *Struct. Control Health Monit.* **16**(3), 319–334 (2009)

13. Qu, Z.-Q.: *Model Order Reduction Techniques with Applications in Finite Element Analysis*. Springer (2014)
14. Chahlaoui, Y., Gallivan, K.A., Vandendorpe, A., Van Dooren, P.: Model reduction of second-order systems. In: Benner, P., Sorensen, D.C., Mehrmann, V. (eds.), *Dimension Reduction of Large-Scale Systems*, pp. 149–172, Springer, Berlin, Heidelberg (2005)
15. Kerfriden, P., Gosselet, P., Adhikari, S., Bordas, S.: Bridging proper orthogonal decomposition methods and augmented Newton-Krylov algorithms: an adaptive model order reduction for highly nonlinear mechanical problems. *Comput. Methods Appl. Mech. Eng.* **200**, 850–866 (2011)
16. Hernández, J., Oliver, J., Huespe, A.E., Caicedo, M., Cante, J.: High-performance model reduction techniques in computational multiscale homogenization. *Comput. Methods Appl. Mech. Eng.* **276**, 149–189 (2014)
17. Corigliano, A., Dossi, M., Mariani, S.: Model order reduction and domain decomposition strategies for the solution of the dynamic elastic-plastic structural problem. *Comput. Methods Appl. Mech. Eng.* **290**, 127–155 (2015)
18. Kerfriden, P., Passieux, J.-C., Bordas, S.P.-A.: Local/global model order reduction strategy for the simulation of quasi-brittle fracture. *Int. J. Numer. Methods Eng.* **89**(2), 154–179 (2012)
19. Florez, H.: Applications of model-order reduction to thermo-poroelasticity. In: *51st US Rock Mechanics/Geomechanics Symposium*. American Rock Mechanics Association (2017)
20. Jin, Z.L., Garipov, T., Volkov, O., Durlofsky, L.: Reduced-order modeling of coupled flow-geomechanics problems. In: *Society of Petroleum Engineers, SPE* (2019)
21. Carlberg, K., Bou-Mosleh, C., Farhat, C.: Efficient non-linear model reduction via a least-squares Petrov-Galerkin projection and compressive tensor approximations. *Int. J. Numer. Meth. Eng.* **86**, 155–181 (2011)
22. de Souza Neto, E., Peric, D., Owen, D.: *Computational Methods for Plasticity: Theory and Applications*. Wiley, UK (2008)
23. Simo, J., Taylor, R.: Consistent tangent operators for rate independent elastoplasticity. *Comput. Methods Appl. Mech. Eng.* **48**, 101–118 (1985)
24. Florez, H.: *Domain Decomposition Methods for Geomechanics*. Ph.D. thesis, The University of Texas at Austin (2012)
25. Florez, H., Gildin, E.: Model-order reduction of coupled flow and geomechanics in Ultra-Low Permeability (ULP) reservoirs, No. 193911 in *SPE Reservoir Simulation Conference*, (Galveston, Texas) (2019)
26. Florez, H., Gildin, E.: Model-order reduction for two-phase flow: projection and hyper-reduction. Technical report, Texas A&M University (2019)
27. Lewis, R., Schrefler, B.: *The Finite Element Method in the Static and Dynamic Deformation and Consolidation of Porous Media*, 2nd edn. Wiley, New York (1998)
28. Florez, H., Wheeler, M.: A mortar method based on NURBS for curved interfaces. *Comput. Methods Appl. Mech. Eng.* **310**, 535–566 (2016)
29. Dean, R., Gai, X., Stone, C., Minkoff, S.: A comparison of techniques for coupling porous flow and geomechanics, No. 79709 in *SPE Reservoir Simulation Symposium*, (Houston), SPE (2003)
30. Lie, K.-A., Krogstad, S., Ligaarden, I.S., Natvig, J.R., Nilsen, H.M., Skaflestad, B.: Open-source matlab implementation of consistent discretisations on complex grids. *Comput. Geosc.* **16**, 297–322 (2012)
31. Aziz, K., Settari, A.: *Petroleum Reservoir Simulation*. Elsevier Applied Science Publishers (1986)
32. Coussy, O.: *Poromechanics*. Wiley, New York (2004)
33. Florez, H.: Linear Thermo-Poroelasticity and geomechanics. In: Pacurar, R. (ed.), *Finite Element Method—Simulation, Numerical Analysis and Solution Techniques*, InTech Open 2018, , Chap. 10, pp. 223–242 (2018). ISBN 978-953-51-3849-5
34. Phillips, P.: *Finite Element Methods in Linear Poroelasticity: Theoretical and Computational Results*. Ph.D. thesis, The University of Texas at Austin (2005)

35. Kim, J.: A new numerically stable sequential algorithm for coupled finite-strain elastoplastic geomechanics and flow. *Comput. Methods Appl. Mech. Eng.* **335**, 538–562 (2018)
36. Kim, J., Tchelepi, H.A., Juanes, R., et al.: Rigorous coupling of geomechanics and multiphase flow with strong capillarity. *SPE J.* **18**(06), 1–123 (2013)
37. Lie, K.-A.: *An Introduction to Reservoir Simulation Using MATLAB/GNU Octave: User Guide for the MATLAB Reservoir Simulation Toolbox (MRST)*. Cambridge University Press (2019)
38. Florez, H., Argáez, M.: In: Reyhanoglu, M. (ed.), *A Reduced Order Gauss-Newton Method for Nonlinear Problems Based on Compressed Sensing for PDE Applications*, Chap. 1, pp. 1–20. in *Nonlinear Systems—Volume 1*, InTech Open (2018). ISBN 978-953-51-6134-9
39. Florez, H., Wheeler, M., Rodriguez, A., Monteagudo, J.: *Domain Decomposition Methods Applied to Coupled Flow-Geomechanics Reservoir Simulation*, No. 141596 in *SPE Reservoir Simulation Symposium*, (The Woodlands, Texas), (2011)
40. Florez, H., Wheeler, M., Rodriguez, A.: A mortar method based on NURBS for curved interfaces. In: *Proceedings of the 13th European Conference on the Mathematics of Oil Recovery (ECMOR XIII)*, Biarritz, France (2012)
41. Florez, H., Borges, B.: *LogProc's Technical Manual*. Version 1.2, GeoFeMOR, LLC. Adelphi, MD (2018)
42. Roussel, N., Florez, H., Rodriguez, A.A.: Hydraulic Fracture Propagation from Infill Horizontal Wells, in *SPE Annual Technical Conference and Exhibition held in New Orleans*. Society of Petroleum Engineers, Louisiana (2013)

Challenges in Model Reduction for Real-Time Simulation of Traction Chain Systems



Roxana Ionuțiu

Abstract Real-time simulation for closed-loop controller testing represents one of the most challenging playing fields in industrial mathematics. Compared to offline simulation, where more sophisticated but time-consuming numerical methods can accurately simulate complex non-linear dynamical systems, in real-time simulation only a limited execution time is available—per time-step—to solve these systems. This calls for dynamical system modeling and simulation approaches which must simultaneously satisfy stringent accuracy and performance requirements. In this paper we present solutions and open problems in modeling, simulating and controlling dynamical systems in real-time, focusing on electric transportation applications. In particular, the applicability of model-order reduction is analyzed for speeding-up the real-time simulation of traction chain systems.

Keywords Real-time simulation · Hardware-in-the-loop · Control · Traction systems · Model reduction

1 Introduction

Digital simulation is today the de facto technology which allows systems of growing complexity to be tested in virtual environments, with high accuracy, at increasingly faster speeds, and with low risk of damaging equipment. This is especially valuable in industrial control development. There, in order to accurately test a controller running at its nominal operation speed, the controller has to be connected in closed-loop to a model of the plant which runs in *real-time*, i.e. at the same rate as the actual physical system. This setup can be achieved nowadays using *hardware-in-the-loop* (HIL) testing on a digital *real-time simulator* (RTS) [1].

An area which extensively uses the benefits of an RTS is the control of traction chain systems for electric vehicles (e.g., railways, trolleybuses, electric buses, etc.). In

R. Ionuțiu (✉)

Principal Engineer Real-Time Simulation, Traction, ABB Switzerland A.G., Traction, Austrasse, Turgi 5300, Switzerland
e-mail: roxana.ionutiu@ch.abb.com

© Springer Nature Switzerland AG 2022

C. Beattie et al. (eds.), *Realization and Model Reduction of Dynamical Systems*,
https://doi.org/10.1007/978-3-030-95157-3_22

409

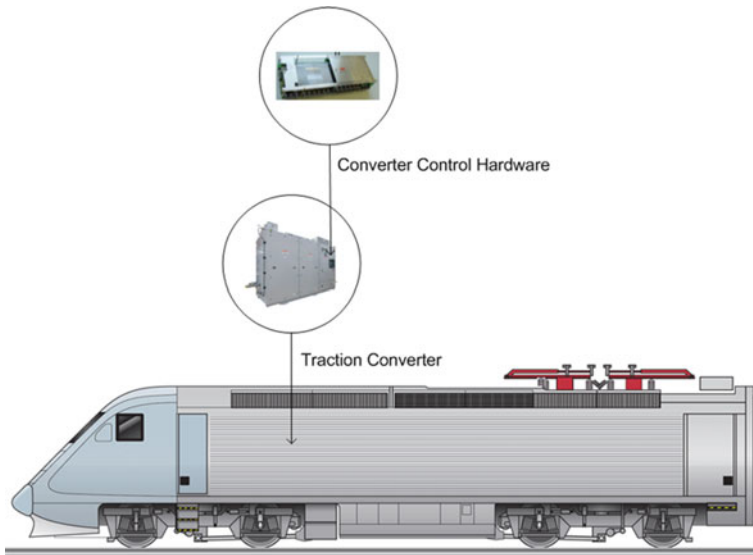


Fig. 1 Multiple unit train with traction converter and controller

a locomotive for instance, the *controller* operates on the power electronic components of the traction converter inside the train, as shown in Fig. 1. The traction converter, together with associated components such as transformers or on-board energy storage systems, are responsible for converting the vehicle power supply (AC, DC, diesel engines) into three phase variable-frequency power suitable for driving the motors, and into auxiliary power needed for ventilation, lighting, battery charging, three-phase on board grid, etc. The controller is the converter's "brain" which, based on its software and hardware, ensures that the energy conversion process delivers peak power at wheel while minimizing energy losses.

In the Traction Product Group of ABB Switzerland in Turgi, RTS is the key technology for testing the complexity of traction converter applications. The RTS provides a testing environment where the controller operates on a mathematical representation of the traction converter and the vehicle. On the RTS, the so called *physical model* reproduces, with high fidelity, the electrical and mechanical domains involved in the energy conversion process (Fig. 2). Thus, using the aforementioned HIL setup, the converter control software, running on the actual hardware (i.e. DPSs, FPGAs), can operate on the physical model as if it was controlling the real system (Fig. 3).

Achieving real-time capable physical models is non-trivial especially in power converter applications, because the plant models have fast dynamics and frequent switching behavior. The ABB Traction RTS technology has developed methods suitable for modeling and simulating such systems using constrained simulation time-steps. This paper will show how, in addition, model order reduction could be integrated into the existing RTS modeling and simulation framework to further

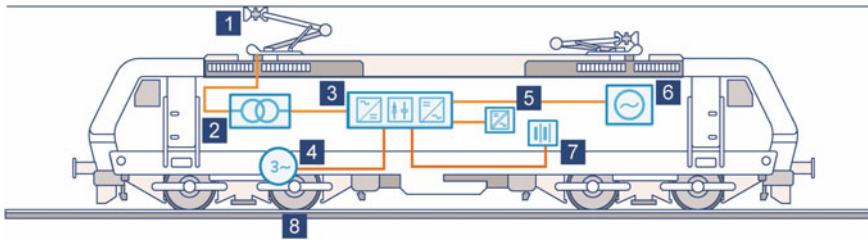
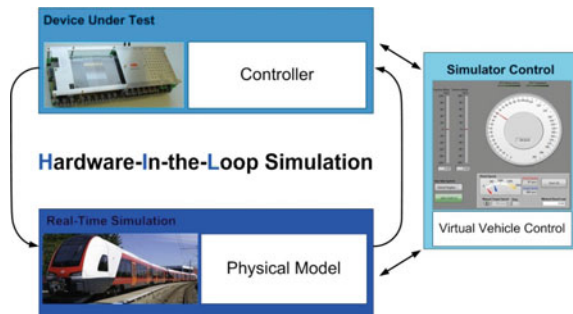


Fig. 2 RTS modeling domains for traction chain systems: 1. Power supply, overhead line, pantograph and main circuit breaker, 2. Transformer, 3. Traction converter (line converter, DC-link with tuned filter, motor converter), 4. Motor, gear, axle 5. Auxiliary converter, 6. Diesel engine generator, 7. Energy storage, 8. Wheel-rail contact

Fig. 3 Hardware-In-the-Loop (HIL) simulation



improve real-time performance. In particular we show how, by exploiting the time-discretization inherent to real-time simulation, reduced-order models can be easily obtained and re-inserted in a full physical model, thereby improving computational performance. The paper will focus on the reduction of linear parts of the traction chain system, while the reduction of complete traction chain systems including switching and non-linear components remains for further research.

The structure of this paper is as follows: Sect. 2 explains the fundamentals of real-time simulation; Sect. 3 details on how physical models are solved numerically in discrete-time; the application of model reduction in the real-time simulation context is shown in Sect. 4; Sect. 5 shows real-time experiments using the proposed reduction framework, based on examples from traction applications. Section 6 concludes.

2 Fundamentals of Real-Time Simulation

Industrial RTS simulators can differ in scale, the hardware used, or specifics of plant modeling, but are nevertheless challenged by the same fundamental issues: ensuring fast and accurate physical models while satisfying critical timing constraints.

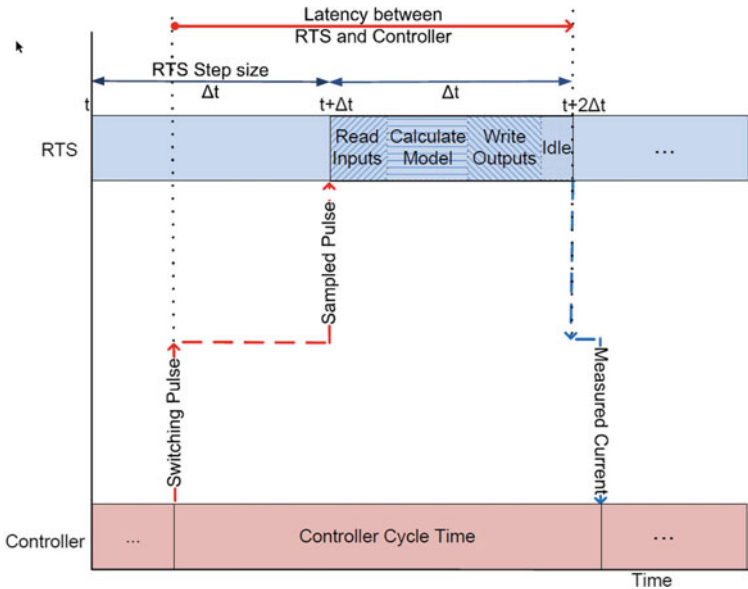


Fig. 4 A maximum delay of twice the RTS step-size occurs between the time the controller sends switching pulses to the RTS, until the corresponding current measurement is output to the controller. The step-size of the RTS must be as small as possible to ensure minimal latency between RTS and the controller. Latency is usually compensated for by the control software

The crux of the RTS is achieving the synchronization between real time and simulated time. In other words, the simulated inputs, states, and outputs reached on the RTS after a certain simulation time, should be the same as when this time had passed on the real system. The rate at which the real-time clock achieves this synchronization is the step-size of the RTS. Its length is fixed, but must be carefully chosen as to ensure the optimal timing and the desired accuracy. Figure 4 illustrates this. On one hand, the RTS step-size must be as small as possible in order to guarantee simulation accuracy, and to minimize the latency of the closed-loop system. On the other hand, the RTS step-size must be long enough to permit the necessary operations (reading inputs, updating states, writing outputs, and providing an idle time margin) to be completed in one simulation step Δt without causing overruns.¹

In power electronic converters the semiconductor devices switch with frequencies extending into the kHz range, making real-time simulation especially difficult compared to other slower automation applications such as robotics. The high frequency switching imposes controller cycle times in the order of $100 - 300 \mu s$, which in turn impose yet smaller cycle times for the real-time simulation of the physical model (between $1 - 50 \mu s$). Having hardware with exceptional computing performance is necessary but not sufficient to ensure that the operations are fast and that the system

¹ An overrun occurs when the RTS calculation time per time-step exceeds the defined fixed simulation step-size Δt , which in turn results in incorrect RTS-Controller closed-loop behavior.

runs in real time. The physical models and the numerical routines to solve them must also be optimized for maximum efficiency.

A distinctive feature of traction converters is the presence of switches (circuit breakers, IGBTs or diodes) changing state at unknown points in time. This challenges the mathematical modeling task: different switch configurations generate different circuit topologies, and hence different sets of differential equations which must be recognized and solved for in real-time. While modeling and simulating switched non-linear systems in real-time is a challenging topic in itself, this paper will only focus on a linear part of the traction chain system for model order reduction. The models are solved numerically in discrete time on the RTS, as explained next.

3 Discrete-Time Simulation

In industrial control applications, state space rather than differential algebraic equations are used to describe dynamical systems. Therefore this paper will focus on the state-space modeling approach. Hence, by expressing the Kirchhoff's current and voltage laws (KCL and KVL) at each node and for each loop in an electrical circuit, a set of ordinary differential equations (ODEs) results, which can be expressed in *state-space* form as follows:

$$\Sigma \begin{cases} \dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \\ \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \end{cases}, \quad (1)$$

where the states \mathbf{x} are the currents through inductors and voltages across capacitors, \mathbf{u} are the inputs of the system and \mathbf{y} are the outputs (the measured quantities of interest, for instance certain voltages or currents). For a system with n states, p inputs and m outputs the corresponding matrix dimensions are: $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times p}$, $\mathbf{C} \in \mathbb{R}^{m \times n}$, $\mathbf{D} \in \mathbb{R}^{m \times p}$. The system's transfer function is:

$$\mathbf{H}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D} \in \mathbb{C}^{m \times p}. \quad (2)$$

The continuous equations (1) can be solved numerically in discrete time. For this purpose, a discretized version $\widehat{\Sigma}$ of (1) is needed, where the states (and outputs) at a new time-step can be computed from previous states and the present (and/or past) inputs, for instance:

$$\widehat{\Sigma} \begin{cases} \mathbf{x}_{t+\Delta t} = \widehat{\mathbf{A}}\mathbf{x}_t + \widehat{\mathbf{B}}\mathbf{u}_{t+\Delta t} \\ \mathbf{y}_{t+\Delta t} = \widehat{\mathbf{C}}\mathbf{x}_{t+\Delta t} + \widehat{\mathbf{D}}\mathbf{u}_{t+\Delta t} \end{cases}, \quad (3)$$

where the discretized matrices $\widehat{\mathbf{A}}$, $\widehat{\mathbf{B}}$, $\widehat{\mathbf{C}}$, $\widehat{\mathbf{D}}$ are derived from the continuous \mathbf{A} , \mathbf{B} , \mathbf{C} , \mathbf{D} depending on the chosen discretization method. Discretization methods are intimately related to the numerical integration employed for finding the solution $\mathbf{x}(t)$ of ODEs as in (1). Reference [2] provides a comprehensive overview of numerical methods for solving dynamical systems on digital computers. The numerical

integration/discretization approaches which are suitable for real-time simulation are described next.

3.1 Numerical Integration

There are several ways to discretize the continuous state-space system (1), each resulting in a different approximation of its exact solution. A popular choice for fixed-step integration (as is required in real-time simulation) is the Backward Euler (BE) method [2]. With a step-size $t + \Delta t$, the BE discretized matrices of system (3) are:

$$\widehat{\mathbf{A}} := (\mathbf{I} - \Delta t \mathbf{A})^{-1} \quad (4)$$

$$\widehat{\mathbf{B}} := (\mathbf{I} - \Delta t \mathbf{A})^{-1} \Delta t \mathbf{B} \quad (5)$$

$$\widehat{\mathbf{C}} := \mathbf{C}, \quad \widehat{\mathbf{D}} := \mathbf{D}. \quad (6)$$

It is important to note that the necessary discretization steps (4)–(5) produce matrices $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ which no longer retain any of the structure or sparsity which could be inherent to the continuous time matrices \mathbf{A} and \mathbf{B} . Furthermore, the discretization step is typically done once and offline on a host computer, rather than online on the real-time processor. Hence, the discretization step itself does not affect the online real-time simulation performance (or the so called *turnaround time*). The repeating computations performed online which do influence the turnaround time are the state and output updates (3), which are based on the non-sparse and non-structure preserving discretized matrices (4)–(5) (whether original or reduced).

These observations enable one to perform any model reduction on the original, continuous time system (1), without employing special structure or sparsity preserving reduction methods. The reduced model will also be discretized before being reinserted in the real-time simulation environment, as will be explained in Sect. 4.

3.1.1 Higher Order Methods

For improved integration accuracy higher order fixed-step numerical integration methods are often needed. These can be derived from an approximation to the analytical solution [3] of the continuous differential equation (1):

$$\mathbf{x}(t) = e^{\mathbf{A}t} \mathbf{x}(0) + \int_0^t e^{\mathbf{A}(t-\tau)} \mathbf{B} \mathbf{u}(\tau) d\tau. \quad (7)$$

It can be shown² that the analytical solution (7) of the state \mathbf{x} evaluated at a new time-step $t + \Delta t$ can be expressed as a function of the state \mathbf{x}_t at the previous time-step t as follows:

$$\mathbf{x}_{t+\Delta t} = e^{\mathbf{A}\Delta t}\mathbf{x}_t + \int_t^{t+\Delta t} e^{\mathbf{A}(t+\Delta t-\tau)}\mathbf{B}\mathbf{u}(\tau)d\tau. \quad (8)$$

As mentioned in [4], two approximations are necessary to implement (8) numerically:

1. the approximation of the matrix exponential $e^{\mathbf{A}\Delta t}$
2. the approximation of the input \mathbf{u} during integration.

For instance, the Backward Euler discretized system matrices (4–5) can be shown to be special cases of (8) with the following approximations for the matrix exponential and input:

$$e^{\mathbf{A}\Delta t} \approx (\mathbf{I} - \Delta t\mathbf{A})^{-1}, \quad \mathbf{u}(\tau)|_t^{t+\Delta t} \approx \mathbf{u}_{t+\Delta t}. \quad (9)$$

By approximating the matrix exponential $e^{\mathbf{A}\Delta t}$ for instance with more terms of a Taylor expansion, higher-order methods are derived and used in real-time simulation.

4 Model Order Reduction in Switched-Linear Systems

Model order reduction (MOR) methods aim to approximate the behavior of a large dynamical system in an efficient manner, so that the resulting approximation error is small. Other requirements are: the preservation of important system properties, of its physical interpretation, and an efficient implementation. In short, starting from an n -dimensional continuous-time state space system Σ , a *reduced*, continuous-time k -dimensional system Σ_r is sought:

$$\Sigma_r \begin{cases} \dot{\mathbf{x}}_r(t) = \mathbf{A}_r\mathbf{x}_r(t) + \mathbf{B}_r\mathbf{u}(t) \\ \mathbf{y}_r(t) = \mathbf{C}_r\mathbf{x}_r(t) + \mathbf{D}\mathbf{u}(t) \end{cases}, \quad (10)$$

where $k \ll n$, so that the output approximation error $\|\mathbf{y}(t) - \mathbf{y}_r(t)\|$ (in appropriate norm) is small. Through reduction, the number of inputs and outputs is the same as in the reduced model, however the internal variables and the system matrices are reduced in dimension: $\mathbf{x}_r \in \mathbb{R}^k$, $\mathbf{A}_r \in \mathbb{R}^{k \times k}$, $\mathbf{B}_r \in \mathbb{R}^{k \times p}$, $\mathbf{C}_r \in \mathbb{R}^{m \times k}$. The inputs of the reduced system are the same as for the original. The transfer function of Σ_r is:

$$\mathbf{H}_r(s) = \mathbf{C}_r(s\mathbf{I}_r - \mathbf{A}_r)^{-1}\mathbf{B}_r + \mathbf{D} \in \mathbb{C}^{m \times p}. \quad (11)$$

² Proof is omitted as it lies outside the scope of the paper.

The unifying approach for obtaining a reduced model from an original system is via a Petrov-Galerkin³ projection [5]:

$$\Sigma_r(\mathbf{A}_r, \mathbf{B}_r, \mathbf{C}_r, \mathbf{D}_r) = (\mathbf{W}^* \mathbf{A} \mathbf{V}, \mathbf{W}^* \mathbf{B}, \mathbf{C} \mathbf{V}, \tilde{\mathbf{D}}), \quad (12)$$

where $\mathbf{V}, \mathbf{W} \in \mathbb{R}^{n \times k}$ are matrices whose $k \ll n$ columns form bases for the relevant subspaces pertaining to the reduction method chosen. In this projection framework it is common to set $\mathbf{D}_r = \mathbf{D}$, but other scenarios are possible, as described in [5].

The governing principle behind all reduction methods is that, after a suitable decomposition is found, the non-dominant⁴ internal variables are eliminated from the system. Reduction methods differ in the way the decomposition is performed. This in turn dictates how the projection matrices \mathbf{V} and \mathbf{W} are constructed. Roughly speaking, reduction methods are classified into (a) spectral-, and (b) Krylov-based. Among the volumes which give a comprehensive coverage of the various methods are [3, 6]. A comparison of spectral and Krylov-based methods is available in [7]. Without attempting to give an exhaustive review for the applicability of the various reduction methods in the context of real-time simulation, in this paper we focus on a selection of spectral and Krylov-based methods, as shown next.

4.1 Dominant Poles and Moment Matching at DC

From the spectral methods, the EVD (eigenvalue decomposition) approach called *modal approximation* [8] constructs a reduced model by interpolating the *dominant poles* of the original system, based on the eigenvalue decomposition of \mathbf{A} .

Krylov-based reduction methods achieve system approximation by matching moments (i.e. terms from the Taylor expansion) of the transfer function (2) of the original system. One of the most popular reduction methods for electrical circuits is PRIMA [9], which matches moments of the original system at frequency $s_0 = 0$, namely at DC.

Here, we construct a continuous-time reduced order model Σ_r (12) by computing projection matrices \mathbf{W}, \mathbf{V} which interpolate a selection of the original system's dominant poles, together with an additional moment matching at DC.

4.2 Reinserting Reduced-Order Models in Simulations

Reinserting reduced-order models in simulation environments is a widely studied problem. For multi-input multi-output systems special reduction methods exist which preserve the structure of the original input/output matrices \mathbf{B} and \mathbf{C} , so that a reduced

³ When $\mathbf{W} = \mathbf{V}$ the projection is of Galerkin type.

⁴ Non-dominant here refers to internal variables which contribute the least to the system's response.

model can be re-synthesized via R, L, C elements and reinserted as a netlist in the simulation tool [10]. For a selection of such methods we refer to SPRIM [11], PACT [12, 13], SparseRC [14], TurboMOR [15], PartMOR [16] and others.

In this paper we show that no synthesis of reduced order models, and hence no structure preserving reduction methods, are needed to reinsert reduced models in real-time simulation environments. Since all models used in real-time simulation are time-discretized, not even the original discretized $\widehat{\Sigma}$ model (3) preserves any of the input/output structure of the original continuous time Σ model (1). Similarly, any continuous-time reduced model Σ_r (12) is discretized as explained in Sect. 3 to obtain a reduced, discretized system $\widehat{\Sigma}_r$, before being re-inserted in a real-time compatible simulation environment such as Matlab-Simulink [17]. Therefore, in real-time simulation, reduced order models can be easily reused for co-simulation with other “non-reduced” parts of a plant model. In Sect. 5 we show the results obtained by reinserting a reduced line-side model in the actual real-time simulation without any structure preservation.

5 Experiments

We demonstrate the effect of model order reduction in real-time simulation based on an industrial example. The model under test is shown in Fig. 5. We compare the behavior of the original and reduced model both in frequency domain, as well as in time domain. We analyze the reduction accuracy, as well as the performance gained through the reduction procedure. Finally, we draw conclusions on the results obtained and on open problems which ensue from applying model-reduction to real-time models.

An important problem in the control of traction converters is the supervision of line voltages, currents, and the DC-link voltage inside the converter. A circuit example used in the simulation of a DC voltage supply, line components, and DC-link is shown in Fig. 5. In the complete real-time simulation test, this model is coupled to remaining models comprising the complete traction chain system (e.g. to parts 3 – 8 from the modeling domains previously shown in Fig. 2). This particular part of the traction chain is chosen as an MOR example since it is linear, whereas the other components include either switching devices (IGBTs, diodes), or non-linear behavior such as motor saturation. Finding an appropriate MOR framework for the entire switched and non-linear traction chain system remains nevertheless a relevant problem for future research.

The rectangular marking from Fig. 5 shows the circuit part which is reduced. The original system has 11 states, 2 inputs and 4 outputs. The partitioning from Fig. 5 ensures that both the original $\widehat{\Sigma}$ and reduced $\widehat{\Sigma}_r$ state space systems are easily coupled to the remaining DC-link capacitor in real-time simulation. This is done via a voltage-current input-output relationship as shown in Fig. 5: i.e. the output current $y2_iLLF$ becomes an input current $iLLF$ to the DC-link capacitor, and the DC-link output voltage UD becomes the input voltage $u2_UD$. During the RTS simulation

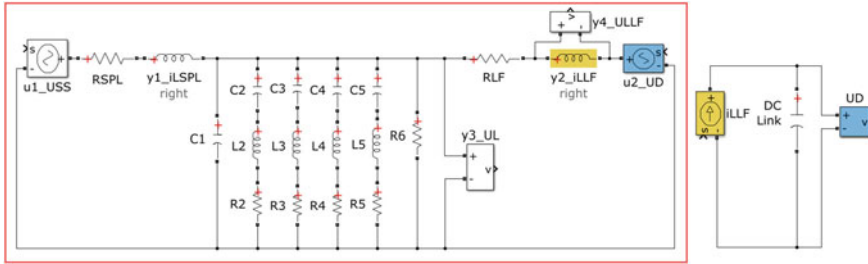


Fig. 5 Electrical circuit for modeling a DC supply, catenary line, line filters and DC-link. The system to be reduced (in rectangle) is coupled to the main DC-link via a voltage-current relationship. The inputs of the state space system are: $u1_USS$, the supply voltage, and $u2_UD$, the DC-link voltage as measured across the main DC-link capacitor. The outputs are: $y1_ISPL$ (the supply current), $y2_iLF$ (the line filter current), $y3_UL$ (the line voltage) and $y4_ULLF$ (the voltage across the line filter inductor)

the measured line voltage $y3_UL$ and DC-link voltage UD are recorded, while a sequence is carried out which closes supply and line specific circuit breakers (not shown).

5.1 Dominant Pole Matching

We reduce the original state space system by interpolating 7 of its poles: the 3 most dominant imaginary pole pairs and 1 real pole, as shown in Fig. 6. The computation of the 3 most dominant imaginary pole pairs was done using the MIMO *samdp* algorithm of [18]. The dominance criterion used is that of poles which are well observable and controllable in the transfer function, as is described in detail in [8, Chapter 4.2]. We match the additional real pole as well, since we aim to preserve the DC characteristics of the system.

The frequency response of the original and resulting reduced system is shown in Fig. 7 (for simplicity only the maximum singular value of the frequency response is shown). One can observe a perfect match at the dominant peaks in the frequency response and an overall match across the entire frequency range.

In model order reduction it is common to judge the quality of a reduced order model based on its frequency response characteristics, without simulating the original and reduced models in the time domain. For real-time simulation, it is crucial that the time-domain behavior is also preserved. This is because in the HIL setup, the controller monitors relevant voltage and current measurements continuously over time, as it also does on the real system.

In Figs. 8 and 9 we show the line and DC-link voltages of the circuit, as recorded on the real-time simulator.

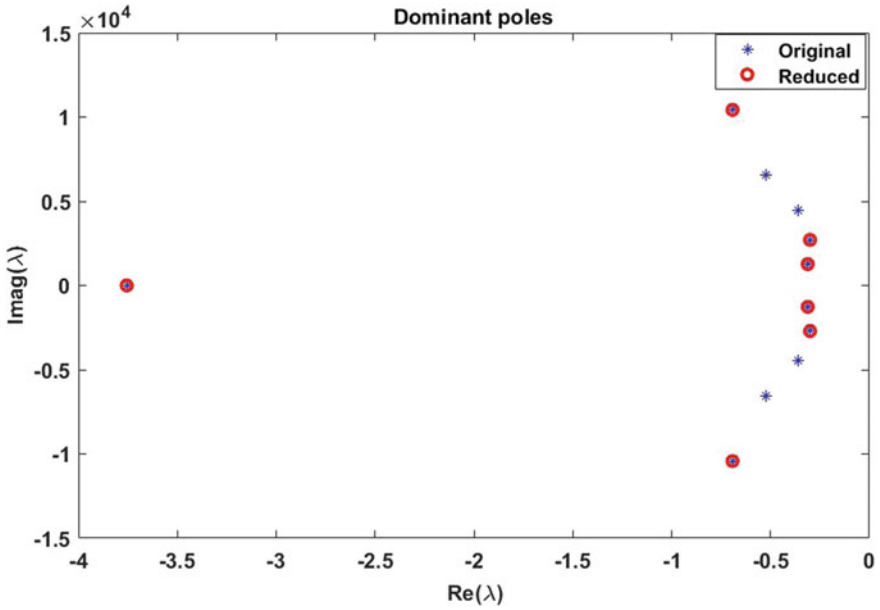


Fig. 6 Dominant poles of original and reduced system. Reduction is by dominant pole matching

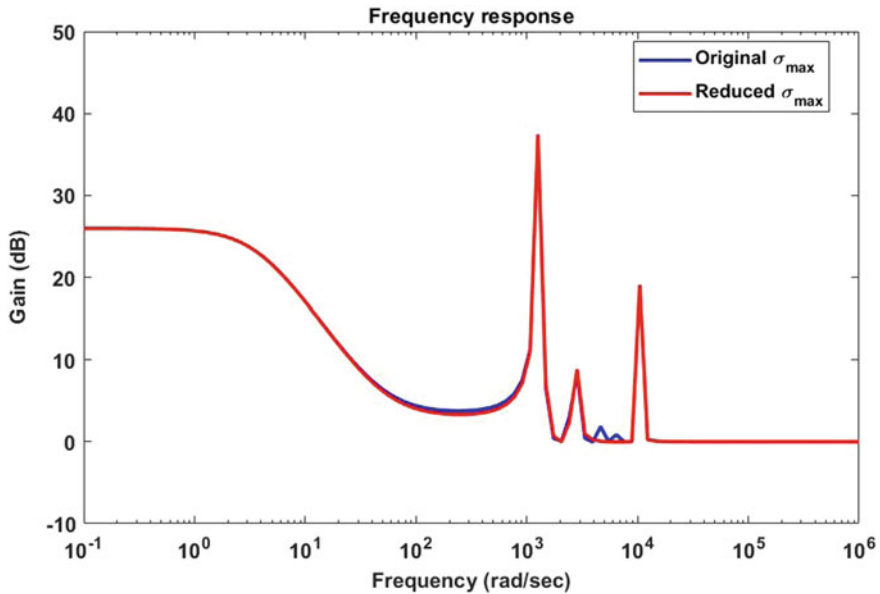


Fig. 7 Frequency response of original and reduced system. Reduction is by dominant pole matching

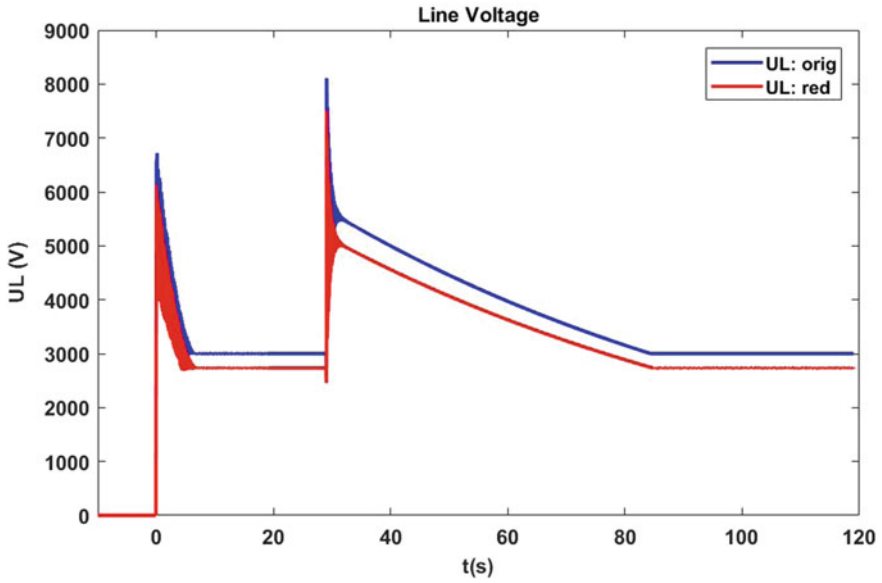


Fig. 8 Line voltage. Reduction is by dominant pole matching

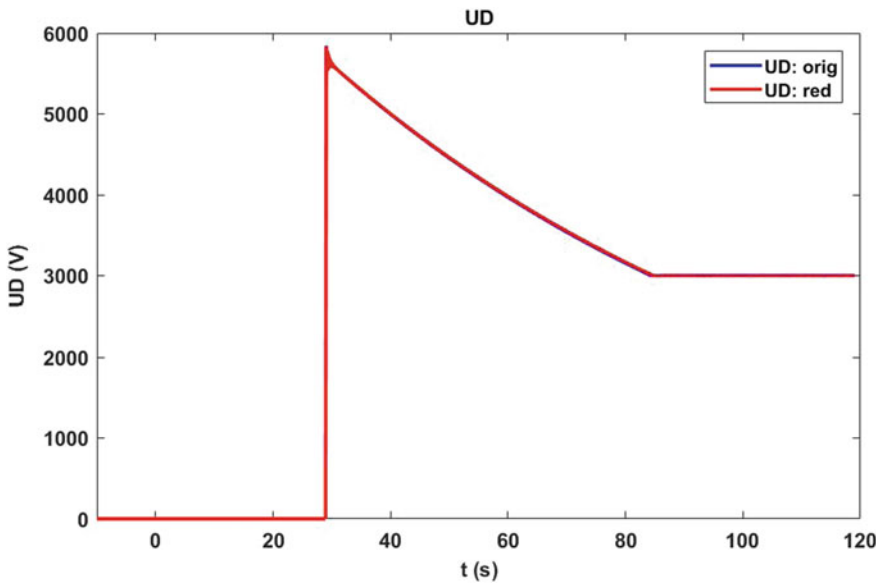


Fig. 9 DC-link voltage. Reduction is by dominant pole matching

One can see from Fig. 8 that the reduced order model has a lower DC voltage (below 3000V) than the original model. This is undesirable, since the controller is programmed to see a DC value of 3000V in the line voltage, and hence will throw an undervoltage fault if it will operate on the reduced model. We note that in this particular operation mode the controller only monitors the line and DC-link voltage, without applying a closed loop control, hence it is important that the plant model behavior is correct also in open loop.

In Fig. 9 on the other hand, we see that the capacitor is charged to the correct value of 3000V, both in the original and the reduced model.

5.2 Dominant Pole and DC Moment Matching

The DC mismatch of UL in Fig. 8 suggests that one could improve the reduced model by matching an additional moment at $s = 0$. In this section we show the results obtained based on this approach.

Figure 10 shows that, by matching one moment of the original transfer function at $s = 0$ (along with 3 imaginary dominant pole pairs), additional poles are generated in the reduced model which are not present in the original. The frequency response of this reduced model is shown in Fig. 11, displaying again a very good match against the original frequency response.

As before, we re-insert the discretized reduced model in the RTS and measure UL and UD in time domain. The results are shown in Figs. 12 and 13. While this time the desired DC value of 3000V is preserved due to the moment matching at $s = 0$, the transient response of the reduced model no longer matches that of the original due to the additional poles which were generated in Fig. 10.

These results calls for new reduction methods which preserve the original system's DC behavior and dominant poles, while not introducing other undesired poles through reduction. For instance, the applicability of balanced truncation for linear switched systems [19] presents a relevant next step for future work.

5.3 Reduction in Execution Time

As explained in Sect. 2, in real-time simulation the turnaround time (TaT) must be as small as possible, and always shorter than the simulation step-size. A step-size in the order of $50\mu s$ is typical for processor-based simulations of a locomotive and converter as shown in Fig. 2.

In Fig. 14 we measure the TaT of the full RTS simulation of the train, where its line side is simulated both using the original and the reduced model. For the experiments in this paper the RTS is using a dSPACE 1006 real-time system [20]. The TaT of the reduced model is approximately $2\mu s$ smaller than of the original. Recall from Fig. 4 that the TaT includes both model calculation as well as IO processing time. From a

TaT of $25\mu s$, approximately $10\mu s$ are needed for the actual model calculation time, while approximately $15\mu s$ are needed for input/output transfer between the simulator and the controller. Since the number of IOs is the same in the reduced and original models, the reduction in model execution time of $2\mu s$ (from $10\mu s$) represents a 20% improvement in computational performance.

6 Conclusion

In this paper we proposed and analyzed a framework for improving the real-time simulation performance of traction chain systems using model order reduction methods. Our results have shown that:

- Models can be reduced in continuous time and reinserted for co-simulation in the real-time simulation environment naturally, without requiring structure preserving or re-synthesis methods.
- Off-the shelf model-reduction methods are applicable only to a limited extent when simulating electrical circuits which are controlled via external devices in real-time. We saw that preserving both the dominant poles of a model and the DC behavior are necessary to guarantee simulation accuracy; nevertheless, finding an appropriate combination of these two requirements remains an open problem.

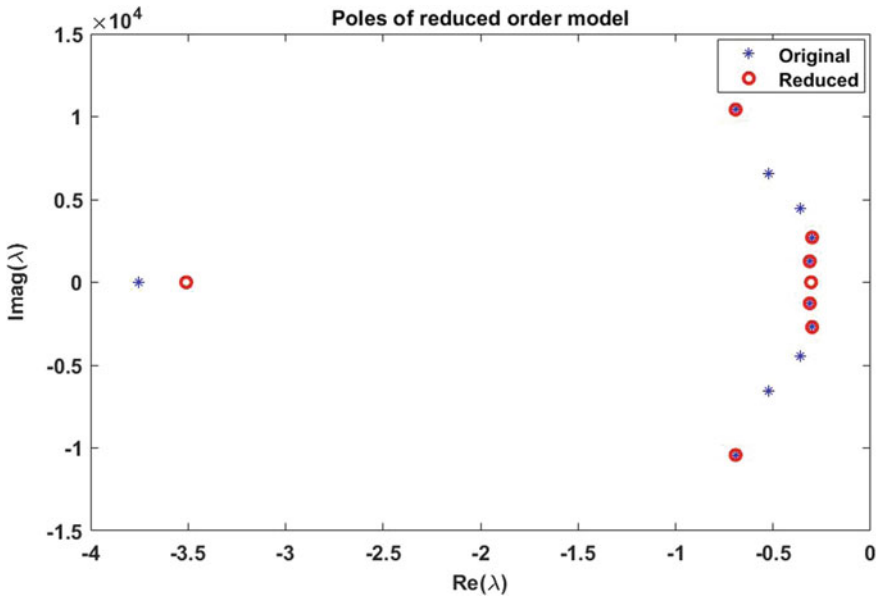


Fig. 10 Poles of original and reduced system. Three dominant pole pairs are matched, and additional poles are generated from one moment matched at DC

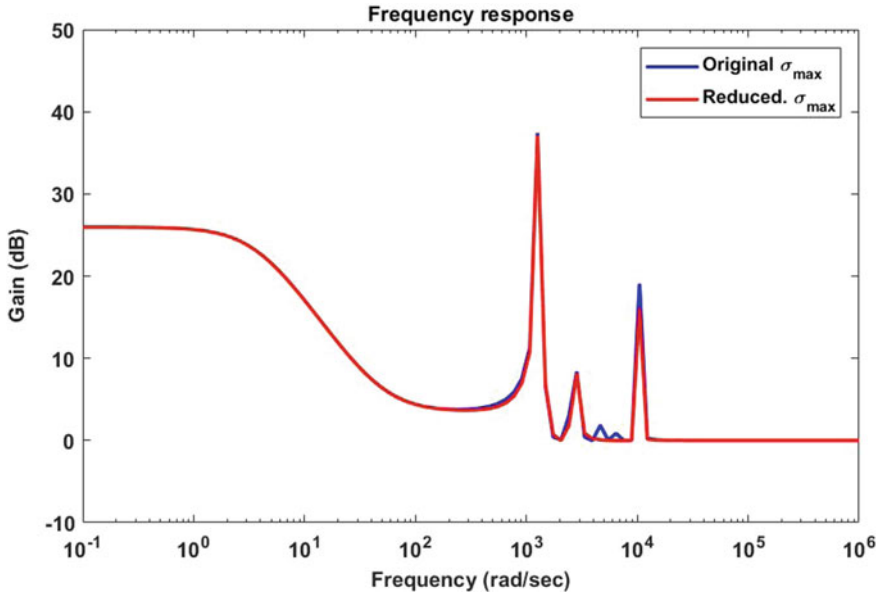


Fig. 11 Frequency response of original and reduced system. Reduction is by dominant pole and DC moment matching

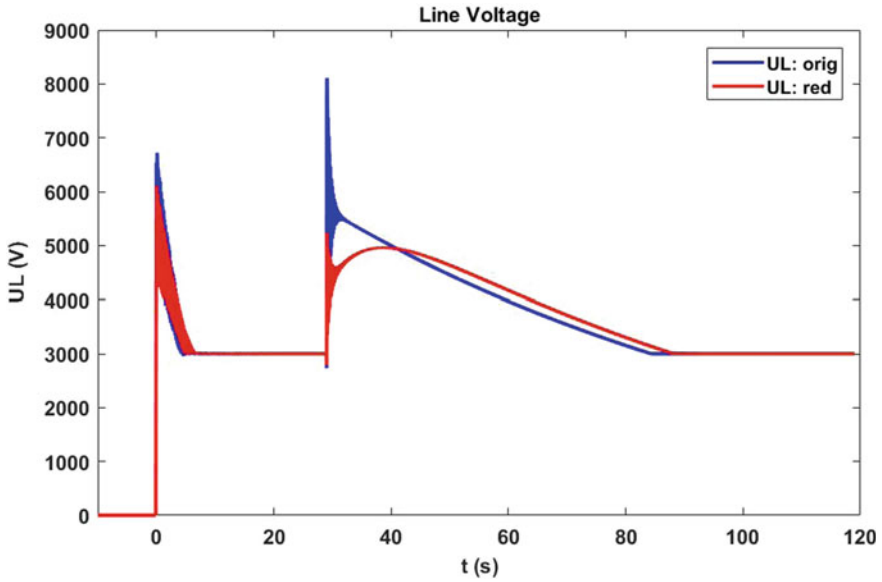


Fig. 12 Line voltage. Reduction is by dominant pole and DC moment matching

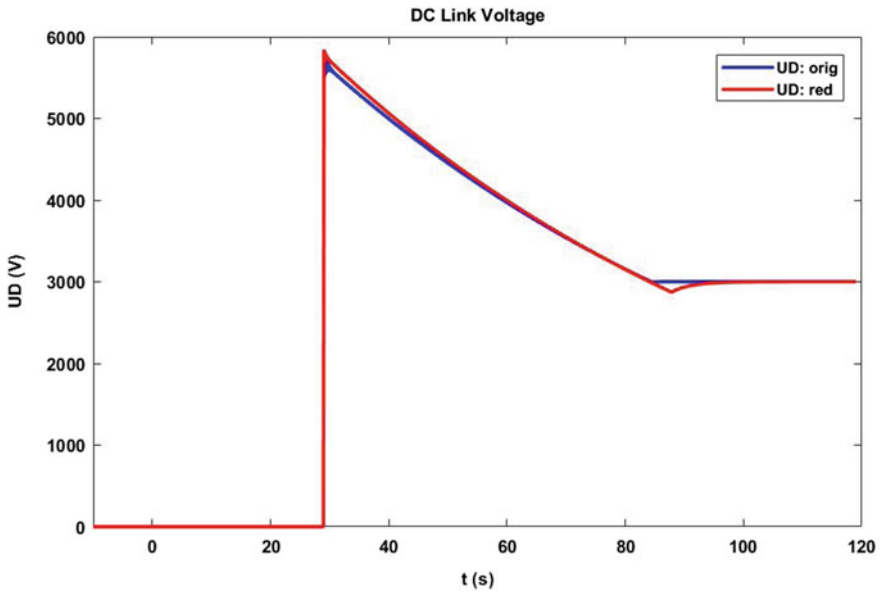


Fig. 13 DC-link voltage. Reduction is by dominant pole and DC moment matching

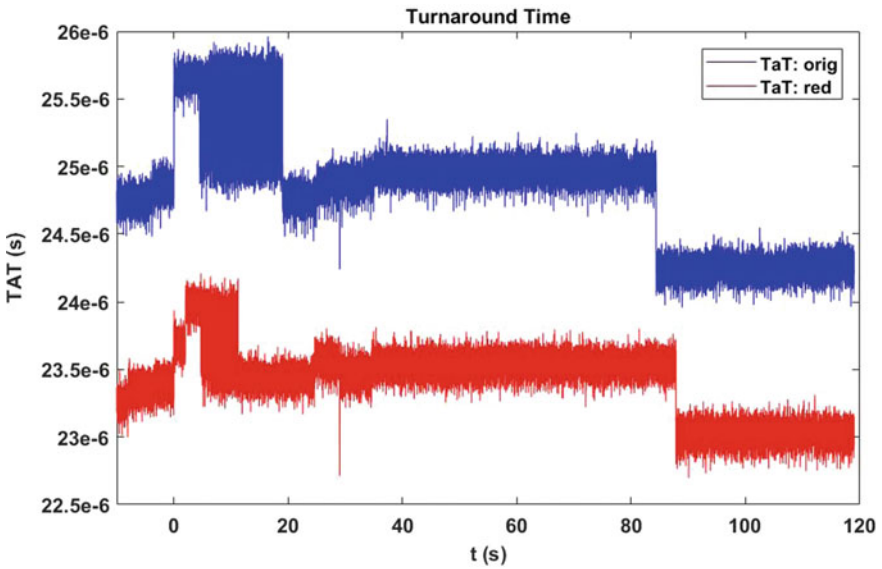


Fig. 14 Turnaround time improvement through model reduction

- Model reduction can significantly improve the real-time simulation performance.

While model order reduction today is widely used in offline simulation, its application in industrial real-time simulation is less explored. As demands for wider real-time test coverage increase, MOR could facilitate this requirement by improving the simulation performance. Finding reduced order models which are sufficiently accurate both in frequency and time domain remains a challenging problem for real-time simulation and control.

References

1. Terwiesch, P., Keller, T., Scheiben, E.: Rail vehicle control system integration testing using digital hardware-in-the-loop simulation. *IEEE Trans. Control Syst. Technol.* **7**, 352–362 (1999)
2. Cellier, F.E., Kofman, E.: *Continuous System Simulation*. Springer
3. Antoulas, A.C.: *Approximation of Large-Scale Dynamical Systems*. SIAM, Philadelphia, PA (2005)
4. Dufour, C., Mahseredjian, J., Bélanger, J., Naredo, J.L.: An advanced real-time electromagnetic simulator for power systems with a simultaneous state-space nodal solver. In: *Transmission and Distribution Conference and Exposition: Latin America (T D-LA)*, 2010 IEEE/PES, pp. 349–358 (2010)
5. Antoulas, A.C., Beattie, C.A., Gugercin, S.: Interpolatory mode reduction for large-scale dynamical systems. In: Mohammadpour, J., Grigoriadis, K.M. (eds.), *Efficient Modeling and Control of Large-Scale Systems*. Springer (2010)
6. Benner, P., Mehrmann, V., Sorensen, D. (eds.): *Dimension Reduction of Large-Scale Systems*, vol. 45. *Lecture Notes in Computational Science and Engineering*. Springer (2005)
7. Ionutiu, R., Lefteriu, S., Antoulas, A.C.: Comparison of model reduction methods with applications to circuit simulation. In: Ciuprina, G., Ioan, D. (eds.), *Scientific Computing in Electrical Engineering SCEE 2006*, vol. 11, pp. 3–24. *Mathematics in Industry*. Springer (2007)
8. Rommes, J.: *Methods for eigenvalue problems with applications in model order reduction*. Ph.D. thesis, Utrecht University (2007)
9. Odabasioglu, A., Celik, M., Pileggi, L.T.: PRIMA: passive reduced-order interconnect macromodeling algorithm. *IEEE Trans. CAD Integr. Circuits Syst.* **17**(8), 645–654 (1998)
10. Ionutiu, R.: *Model order reduction for multi-terminal systems with applications to circuit simulation*. Ph.D. thesis, Jacobs University Bremen and Technische Universiteit Eindhoven (2011)
11. Freund, R.W.: SPRIM: structure-preserving reduced-order interconnect macromodeling. In: *IEEE/ACM International Conference on Computer Aided Design*, 2004. ICCAD-2004, pp. 80–87 (2004)
12. Kerns, K.J., Yang, A.T.: Stable and efficient reduction of large, multiport RC networks by pole analysis via congruence transformations. *IEEE Trans. CAD Integr. Circuits Syst.* **16**(7), 734–744 (1997)
13. Kerns, K.J., Yang, A.T.: Preservation of passivity during RLC network reduction via split congruence transformations. *IEEE Trans. CAD Integr. Circuits Syst.* **17**(7), 582–591 (1998)
14. Ionutiu, R., Rommes, J., Schilders, W.H.A.: SparseRC: sparsity preserving model reduction for RC circuits with many terminals. *IEEE Trans. CAD Integr. Circuits Syst.* **30**(12), 1828–1841 (2011)
15. Oyaro, D., Triverio, P.: TurboMOR-RC: an efficient model order reduction technique for RC networks with many ports. *IEEE Trans CAD Integr. Circuits Syst.* **35**(10), 1695–1706 (2016)
16. Miettinen, P., Honkala, M., Roos, J., Valtonen, M.: PartMOR: partitioning-based realizable model-order reduction method for RLC circuits. *IEEE Trans. CAD Integr. Circuits Syst.* **30**(3), 374–387 (2011)

17. Simulink: MATLAB version 9.2.0.556344 (R2017a). The MathWorks Inc., Natick, MA (2017)
18. Rommes, J., Martins, N.: Efficient computation of multivariable transfer function dominant poles using subspace acceleration. *IEEE Trans. Power Syst.* **21**(4), 1471–1483 (2006)
19. Gosea, I.V., Petreczky, M., Antoulas, A.C., Fiter, C.: Balanced truncation for linear switched systems. *Adv. Comput. Math.* **44**(6), 1845–1886 (2018)
20. www.dSPACE.com

Sparse Representation for Sampled-Data H^∞ Filters



Masaaki Nagahara and Yutaka Yamamoto

Abstract We consider the problem of discretization of analog filters and propose a novel method based on sampled-data H^∞ control theory with sparse representation. For optimal discretization, we adopt minimization of the H^∞ norm of the error system between a (delayed) target analog filter and a digital system consisting of an ideal sampler, the zero-order hold, and an FIR (finite impulse response) filter. Also, for digital implementation, we propose a sparse representation of the FIR filter to reduce the number of nonzero coefficients with the ℓ^1 norm regularization. We show that this multi-objective optimization is reducible to a convex optimization problem, which can be solved efficiently by numerical computation. We then extend the design method to multi-rate filters, and show a design example. We also give an application to the feedback filter design of delta-sigma modulators.

Keywords Sparse representation · Compressed sensing · Sampled-data control · Filter design · Delta-sigma modulation · Convex optimization

Preface

It is a pleasure for us to contribute an article to this Festschrift in honor of Athanasios Antoulas (hereafter Thanos) on the occasion of his 70th birthday.

Both Thanos and the second author (hereafter “I”) studied for Ph.D. under the late professor R. E. Kalman in the late 1970s. While Thanos did his Ph.D. at ETH Zürich, I did my Ph.D. at the University of Florida in Gainesville.

After completing our Ph.D. studies, Thanos got a job in the U.S., and I returned to Japan. So we were geographically quite apart. But we have had various interactions through conferences and through occasional mutual visits to our places. Thanos liked Japan, and he visited Kyoto several times. It was a pleasure to introduce the traditional Japanese culture to him, for example, Shugakuin imperial villa, gardens in various temples, Daimonji farewell fire, etc. They are part of the highest representation of

M. Nagahara (✉)

The University of Kitakyushu, Hibikino 1-1, Wakamatsu, Kitakyushu, Fukuoka 808-0135, Japan
e-mail: nagahara@ieee.org

Y. Yamamoto

Kyoto University, Yoshida Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
e-mail: yy@i.kyoto-u.ac.jp

© Springer Nature Switzerland AG 2022

C. Beattie et al. (eds.), *Realization and Model Reduction of Dynamical Systems*,
https://doi.org/10.1007/978-3-030-95157-3_23

the Japanese culture, and he appreciated them very much. On the other hand, he also enjoyed a Japanese “izakaya” restaurant (a Japanese style pub that serves a variety of foods), and we both had a good time together.

Although we did not have much intersection in our research interests, we had some in-depth discussions during these visits, on modeling of linear systems—large-scale systems on his side, and infinite-dimensional systems on my side.

Throughout the years, he has shown great leadership in the modeling of large-scale systems. This volume is a tribute to his long-term contributions, and we feel it appropriate to dedicate the present paper on sparse representation in sampled-data filters.

1 Introduction

Sparse representation is a kind of model reduction, by which the number of parameters that represent the behavior of a system is minimized [6]. The number of active parameters is measured by using the ℓ^0 (pseudo-) norm, which is defined by the number of nonzero elements in a vector. The sparse representation is done by minimizing the ℓ^0 norm of the parameter vector under constraints (e.g. linear equations). This is however a combinatorial optimization and actually NP hard [20]. To avoid this computational difficulty, the ℓ^1 norm is used instead of the ℓ^0 norm. The ℓ^1 norm minimization under linear constraints becomes linear programming (or convex optimization), and the solution is numerically tractable. A surprising fact is that the ℓ^1 solution is identical to the original ℓ^0 solution under some conditions (see e.g. [2, 4, 29]). This idea has recently been applied to optimal control using the L^0 norm, called *maximum hands-off control* in [13, 14].

In this chapter, we adapt the idea of sparse representation to a problem of FIR (finite impulse response) approximation (or discretization) of *analog* filters. In digital signal processing, FIR filters are preferred to IIR (infinite impulse response) filters because of the following merits:

Merits of FIR filters

- **Stability:** FIR filters are always stable (all poles are at $z = 0$).
- **Implementability:** they can be easily implemented in digital devices.
- **Robustness:** they are more robust against coefficient quantization than IIR filters that may cause limit cycles.

On the other hand, a lot of design methods are known for IIR digital filters, e.g., Butterworth, Chebyshev and Elliptic filters [21]. To enjoy the merits of FIR filters mentioned above, we need to approximate a given IIR *digital* filter by an FIR filter. For this purpose, optimization-based approximation methods have been proposed in [10, 30]. These methods can be used if we are given a target IIR digital filter.

In practical applications, we often need to implement an *analog* filter on a digital device as an FIR filter. An RLC analog filter (an electrical circuit consisting of resistors, inductors, and capacitors) is one example and a PID controller is another. To obtain an FIR digital filter that well approximates a given analog filter, we usually take the following two conventional steps:

FIR approximation by truncation

1. Discretize an analog filter to an IIR digital filter via the step-invariant transformation or the bilinear (or Tustin) transformation [3].
2. Truncate the impulse response of the IIR digital filter to obtain an FIR one, or use more sophisticated approximation methods as in [10, 30].

However, it is more preferable if an FIR digital filter is obtained *directly* from the original analog filter. For this purpose, a direct design method has been proposed in [17], which is based on the theory of sampled-data H^∞ control [16]. In this chapter, we extend these design methods to *sparse* FIR digital filter design. Key ideas are summarized below.

Key ideas of the proposed method

- Measure the approximation error by the sampled-data H^∞ norm.
- Use the KYP (Kalman-Yakubovich-Popov) lemma that reduces the H^∞ optimization problem to optimization described in an LMI (linear matrix inequality).
- Adopt the ℓ^1 norm regularization to obtain sparse FIR coefficients.

We also extend the result to multi-rate systems that consists of an upsampler and a fast hold. Design examples of filter discretization and delta-sigma modulation are shown to illustrate the effectiveness of our methods.

The remainder of this article is organized as follows. In Sect. 2, we give the mathematical formulation of our design problem. Section 3 is the main section where we give a procedure for the solution. Then we extend the method to multi-rate systems in Sect. 4. We show a design example of analog filter discretization in Sect. 5, and an application to delta-sigma modulator design in Sect. 6. In Sect. 7, we will make a conclusion.

2 Problem Formulation

The problem we consider here is to discretize an analog filter for digital implementation. Let $K_c(s)$ be the transfer function of an analog filter. We suppose that the

transfer function $K_c(s)$ is stable, real-rational,¹ and proper. To discretize this filter, let us consider the digital system shown in Fig. 1. In this block diagram, $K(z)$ is an FIR digital filter of length M , described by

$$K(z) = \sum_{k=0}^{M-1} a_k z^{-k}, \tag{1}$$

Note that $K(z)$ is always stable since it has poles only at $z = 0$. The system \mathcal{S}_h before $K(z)$ is the ideal sampler with fixed sampling period $h > 0$ that converts a continuous-time signal $u(t)$ into a discrete-time signal $\{v[0], v[1], v[2], \dots\}$ as

$$v[n] = (\mathcal{S}_h u)[n] = u(nh), \quad n = 0, 1, 2, \dots$$

The system \mathcal{H}_h after $K(z)$ is the zero-order hold that produces a continuous-time signal $\hat{y}(t)$ from a discrete-time signal $\{\psi[0], \psi[1], \psi[2], \dots\}$ as

$$\hat{y}(t) = \sum_{n=0}^{\infty} \psi[n] \phi(t - nh), \quad t \in [0, \infty),$$

where $\phi(t)$ is a box function defined by

$$\phi(t) = \begin{cases} 1, & \text{if } t \in [0, h), \\ 0, & \text{otherwise.} \end{cases}$$

Our problem is to obtain the FIR digital filter coefficients a_0, a_1, \dots, a_{M-1} in (1) so that the sampled-data system

$$\mathcal{K} := \mathcal{H}_h K(z) \mathcal{S}_h = \mathcal{H}_h \left(\sum_{k=0}^{M-1} a_k z^{-k} \right) \mathcal{S}_h$$

well approximates the input/output behavior of the analog filter $K_c(s)$.

Let $\alpha(t)$ denote the impulse response (or the inverse Laplace transform) of $K_c(s)$. The digital filter $K(z)$ (or the filter coefficients a_0, a_1, \dots, a_{M-1}) is designed to

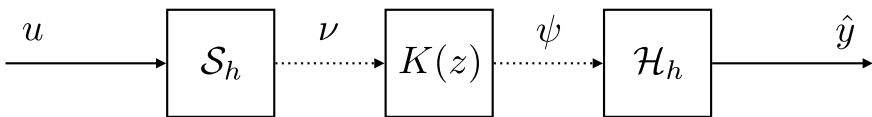


Fig. 1 Digital system \mathcal{K} consisting of ideal sampler \mathcal{S}_h , digital filter $K(z)$, and zero-order hold \mathcal{H}_h with sampling period h

¹ A real-rational transfer function is a rational function of s with real coefficients.

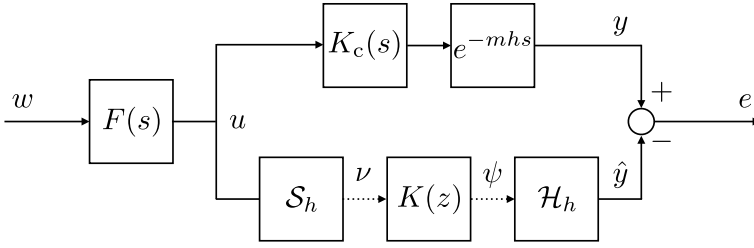


Fig. 2 Error system $\mathcal{E}(K)$

produce a continuous-time signal \hat{y} after the zero-order hold \mathcal{H}_h that approximates the *delayed* output

$$y(t - l) = (\alpha * u)(t - l) = \int_0^{t-l} \alpha(\tau)u(t - l - \tau)d\tau$$

of $K_c(s)$ for an analog input u . This reconstruction delay plays an important role in the approximation performance. That is, if we have a large enough delay $l > 0$, then the approximation error may become small. See the discussion in [19, 32] for details. For simplicity of numerical computation, we assume that l is an integer multiple of h , that is, $l = mh$, where m is a nonnegative integer.

Figure 2 shows the block diagram of the error system

$$\mathcal{E}(K) := (e^{-mhs} K_c - \mathcal{H}_h K \mathcal{S}_h)F, \tag{2}$$

where $F(s)$ is a stable, real-rational, and strictly proper transfer function, which defines the analog characteristic of the input signals in the frequency domain.

The strict properness of $F(s)$ is essential in the reconstruction problem to avoid a trivial solution $K(z) = 0$. If $F(s)$ is bi-proper (i.e. the degree of the numerator is equal to that of the denominator), then the ideal sampler becomes an unbounded operator from L^2 to ℓ^2 [3, Section 9.3], and hence if $K(z) \neq 0$ then the error system also becomes unbounded. Therefore, $K(z)$ should be zero if $F(s)$ is bi-proper. The use of $F(s)$ means that we consider the input signals that belong to the following signal subspace of $L^2[0, \infty)$:

$$\Omega_F := \{Fw : w \in L^2[0, \infty)\}.$$

Then, the problem of H^∞ -optimal discretization proposed in [17] is described as follows.

Problem 1 (H^∞ FIR Discretization) *Given a target filter $K_c(s)$, analog characteristic $F(s)$, sampling period h , and delay step m , find the filter coefficients a_0, a_1, \dots, a_{M-1} of $K(z)$ given by (1) that minimizes*

$$\|\mathcal{E}(K)\|_\infty = \|(e^{-mhs} K_c - \mathcal{H}_h K \mathcal{S}_h) F\|_\infty = \sup_{\substack{w \in L^2 \\ w \neq 0}} \frac{\|(e^{-mhs} K_c - \mathcal{H}_h K \mathcal{S}_h) F w\|_{L^2}}{\|w\|_{L^2}}.$$

Clearly, if the FIR filter length M is larger, the H^∞ optimal filter may give a better performance. However, for implementation, the filter length cannot be arbitrarily large and the number of coefficients should be limited. To solve this implementation problem, we propose to adapt the idea of *sparse representation*, also known as *compressed sensing* [13, 29], to our discretization problem. Sparse representation attempts to reduce the number of nonzero elements of a vector by adopting the ℓ^0 norm as a regularization term. Then, borrowing the technique of convex relaxation by using ℓ^1 norm instead of the non-convex ℓ^0 norm, we formulate the sparse FIR filter design as follows.

Problem 2 (Sparse FIR Discretization) *Given a target filter $K_c(s)$, analog characteristic $F(s)$, sampling period h , and delay step m , find the filter coefficients a_0, a_1, \dots, a_{M-1} of $K(z)$ given by (1) that minimizes*

$$J(a) = \|\mathcal{E}(K)\|_\infty + \lambda \|a\|_1, \tag{3}$$

where λ is a fixed positive integer and a is the coefficient vector defined by

$$a = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_{M-1} \end{bmatrix} \in \mathbb{R}^M. \tag{4}$$

Note that if we take $\lambda = 0$, then Problem 2 is equivalent to Problem 1.

Remark 1 We have assumed that $K_c(s)$ is stable. If $K_c(s)$ has no poles on the imaginary axis, the proposed method can be used for unstable $K_c(s)$ as follows: factorize it as $K_c(s) = K_s(s)K_{as}(s)$, where $K_s(s)$ is stable and $K_{as}(s)$ is anti-stable, and then discretize $K_s(s)$ and $K_{as}(-s)$ by the proposed method. Such an unstable filter is often used in signal processing as a *non-causal* filter [21].

3 Sparse FIR Filter Design via Sampled-Data H^∞ Optimization

In this section, we give convex optimization formulation to numerically compute the sparse FIR filter coefficients of Problem 2. The key technique is *fast sample/hold* approximation [16, 17].

First, we define the lifting operator for discrete-time signals. For a discrete-time signal $x = \{x[0], x[1], x[2], \dots\}$, the discrete-time lifting (also known as blocking) is defined by

$$\mathbf{L}_N x := \left\{ \left[\begin{array}{c} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{array} \right], \left[\begin{array}{c} x[N] \\ x[N+1] \\ \vdots \\ x[2N-1] \end{array} \right], \dots \right\}.$$

Note that the lifting operator \mathbf{L}_N can be written as

$$\mathbf{L}_N = (\downarrow N) \begin{bmatrix} 1 \\ z \\ \vdots \\ z^{N-1} \end{bmatrix},$$

where $\downarrow N$ is the downsampling operator [28] defined by

$$\downarrow N : \{x[k]\}_{k=0}^\infty \mapsto \{x[0], x[N], x[2N], \dots\}.$$

For the lifting operator \mathbf{L}_N , its inverse \mathbf{L}_N^{-1} is defined as

$$\mathbf{L}_N^{-1} := [1 \ z^{-1} \ \dots \ z^{-N+1}] (\uparrow N),$$

where $\uparrow N$ is the upsampling operator [28] defined by

$$\uparrow N : \{x[k]\}_{k=0}^\infty \mapsto \{x[0], \underbrace{0, \dots, 0}_{N-1}, x[1], 0, \dots\}.$$

With the lifting operator, we define the discrete-time lifting of a discrete-time system by

$$\mathbf{lift} \left(\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right], N \right) := \mathbf{L}_N \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right] \mathbf{L}_N^{-1} = \left[\begin{array}{c|ccc} A^N & A^{N-1}B & A^{N-2}B & \dots B \\ \hline C & D & 0 & \dots 0 \\ CA & CB & D & \ddots \vdots \\ \vdots & \vdots & \vdots & \ddots 0 \\ CA^{N-1} & CA^{N-2}B & CA^{N-3}B & \dots D \end{array} \right],$$

where we use the packed notation

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right] := C(zI - A)^{-1}B + D.$$

By the discrete-time lifting, we obtain the fast sample/hold approximation E_N of the sampled-data error system $\mathcal{E}(K)$ in (2) as

$$E_N(z) = (z^{-mN} K_N(z) - H_N K(z) S_N) F_N(z), \quad (5)$$

where

$$\begin{aligned}
 K_N(z) &= \mathbf{lift}(S_{h/N} K_c \mathcal{H}_{h/N}, N), & F_N(z) &= \mathbf{lift}(S_{h/N} F \mathcal{H}_{h/N}, N), \\
 H_N &= \underbrace{[1, \dots, 1]}_N^\top, & S_N &= \underbrace{[1, 0, \dots, 0]}_{N-1}.
 \end{aligned} \tag{6}$$

We note that $K_N(z)$ and $F_N(z)$ are discrete-time, finite-dimensional, linear, and time-invariant systems as shown in [16]. Let $\{A_K, B_K, C_K, D_K\}$ be a minimal realization of the FIR filter $K(z)$ in (1), that is,

$$\begin{aligned}
 A_K &:= \begin{bmatrix} 0 & 1 & 0 \\ \vdots & \ddots & \ddots \\ 0 & \dots & 0 & 1 \\ 0 & \dots & 0 & 0 \end{bmatrix}, & B_K &:= \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \\
 C_K &:= [a_{M-1} \ a_{M-2} \ \dots \ a_1], & D_K &:= a_0.
 \end{aligned} \tag{7}$$

From (5), we obtain

$$E_N(z) = \left[\begin{array}{ccc|c} A_1 & 0 & 0 & B_1 \\ 0 & A_2 & 0 & B_2 \\ 0 & B_K C_2 & A_K & 0 \\ \hline C_1 - H_N D_K C_2 - H_N C_K & & & 0 \end{array} \right] =: \left[\begin{array}{c|c} A & B \\ \hline C(a) & D(a) \end{array} \right], \tag{8}$$

where $\{A_1, B_1, C_1\}$ and $\{A_2, B_2, C_2\}$ are state-space matrices of strictly proper transfer functions $z^{-mN} K_N(z) F_N(z)$ and $S_N F_N(z)$, respectively, and a is the coefficient vector defined in (4). The important property of the state-space representation in (8) is that the FIR coefficient vector a is independent of A and B . From this property with the KYP lemma [17], Problem 2 is finally approximated to the following convex optimization:

Sparse optimization

Give a parameter $\lambda > 0$. Find $a \in \mathbb{R}^M$ and $\gamma > 0$ that minimizes

$$J(a, \gamma) := \gamma + \lambda \|a\|_{\ell^1} \tag{9}$$

subject to

$$\begin{bmatrix} A^\top X A - X & A^\top X B & C(a)^\top \\ B^\top X A & B^\top X B - \gamma I & D(a)^\top \\ C(a) & D(a) & -\gamma I \end{bmatrix} < 0, \tag{10}$$

$X > 0.$

This problem can be efficiently solved via numerical optimization softwares such as SDPT3 [26], SeDuMi [25], or cvx [7, 8] on MATLAB.

Note that the solution to the above optimization is an approximated solution to the original problem (Problem 2). If N is sufficiently large, the approximation error becomes tolerable and the error converges to zero as $N \rightarrow \infty$ (see [31] for details). Note also that if the approximation number N is very large, then the optimization will take a lot of computational time. However, we can empirically say that in many cases, N can be small, say $N \leq 10$ (see also the design example in Sect. 5).

In summary, we have derived a computationally efficient method with convex optimization for sparse FIR approximation of analog filters based on the fast sample/hold approximation in (5) and the KYP lemma.

4 Extension to Multi-rate Filters

The result in the previous section can be extended to multi-rate filters [16, 17]. Here a multi-rate filter utilizes the upsampling operator $\uparrow L$ with an integer $L \geq 2$ to upsample the output of a slow sampler. The multi-rate signal processing system with upsampling factor L is defined by

$$\mathcal{K}_L := \mathcal{H}_{h/L} K(z) (\uparrow L) \mathcal{S}_h = \mathcal{H}_{h/L} \left(\sum_{k=0}^{M-1} a_k z^{-k} \right) (\uparrow L) \mathcal{S}_h.$$

The block diagram of this system is shown in Fig. 3.

The design objective is to find a sparse FIR filter $K(z)$ given by (1) that reduces the sampled-data H^∞ norm of the multi-rate error system $\mathcal{E}_L(K)$ defined by

$$\mathcal{E}_L(K) := (e^{-mhs} K_c - \mathcal{H}_{h/L} K (\uparrow L) \mathcal{S}_h) F. \tag{11}$$

Figure 4 shows the block diagram of the error system $\mathcal{E}_L(K)$. Now, our problem is described as follows:

Problem 3 (Sparse Multi-rate FIR Discretization) *Given target filter $K_c(s)$, analog characteristic $F(s)$, sampling period h , delay step m , and upsampling ratio L , find the filter coefficients a_0, a_1, \dots, a_{M-1} of $K(z)$ given by (1) that minimizes*

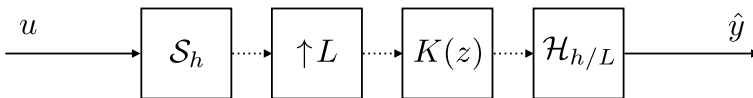


Fig. 3 Multi-rate system \mathcal{K}_L consisting of ideal sampler \mathcal{S}_h , upsampler $\uparrow L$, digital filter $K(z)$, and fast hold $\mathcal{H}_{h/L}$

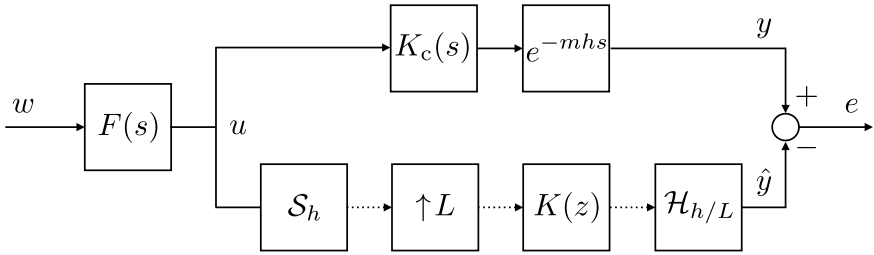


Fig. 4 Multi-rate error system $\mathcal{E}_L(K)$

$$J_L(a) = \|\mathcal{E}_L(K)\|_\infty + \lambda \|a\|_1. \tag{12}$$

We then use the fast sample/hold approximation in Sect. 3 for this system. Let us assume $N = Lp$ for some positive integer p . Then, we obtain the fast sample/hold approximation of \mathcal{E}_L as

$$E_{L,N}(z) = (K_N(z)z^{-mN} - \tilde{H}_N \tilde{K}(z)S_N)F_N(z), \tag{13}$$

where $K_N(z)$, $F_N(z)$ and S_N are given in (6), and

$$\begin{aligned} \tilde{H}_N &= \mathbf{blkdiag}\{\underbrace{\mathbf{1}_p, \mathbf{1}_p, \dots, \mathbf{1}_p}_L\}, \quad \mathbf{1}_p = \underbrace{[1, \dots, 1]}_p^\top, \\ \tilde{K}(z) &= \mathbf{lift}(K(z), L)[\underbrace{1, 0, \dots, 0}_{L-1}]^\top =: \left[\begin{array}{c|c} \tilde{A}_K & \tilde{B}_K \\ \hline \tilde{C}_K & \tilde{D}_K \end{array} \right], \\ \tilde{A}_K &:= A_K^L, \quad \tilde{B}_K := A_K^{L-1}B_K, \\ \tilde{C}_K &:= \begin{bmatrix} C_K \\ C_K A_K \\ \vdots \\ C_K A_K^{L-1} \end{bmatrix}, \quad \tilde{D}_K := \begin{bmatrix} D_K \\ C_K B_K \\ \vdots \\ C_K A_K^{L-2} B_K \end{bmatrix}. \end{aligned} \tag{14}$$

Note that matrices A_K , B_K , C_K , and D_K in (14) are defined in (7). From (13), we obtain

$$E_{L,N}(z) = \left[\begin{array}{ccc|c} A_1 & 0 & 0 & B_1 \\ 0 & A_2 & 0 & B_2 \\ 0 & \tilde{B}_K C_2 & \tilde{A}_K & 0 \\ \hline C_1 - \tilde{H}_N \tilde{D}_K C_2 & -\tilde{H}_N \tilde{C}_K & 0 & 0 \end{array} \right] =: \left[\begin{array}{c|c} \tilde{A} & \tilde{B} \\ \hline \tilde{C}(a) & \tilde{D}(a) \end{array} \right], \tag{15}$$

where $\{A_1, B_1, C_1\}$ and $\{A_2, B_2, C_2\}$ are state-space matrices of strictly proper transfer functions $z^{-mN} K_N(z)F_N(z)$ and $S_N F_N(z)$, respectively, and a is the FIR coefficient vector given by (4) which is to be designed.

Finally, our problem (Problem 3) is also reduced to finding $a \in \mathbb{R}^M$ that minimizes the cost function $J_L(a)$ given in (12), which can be described as a convex optimization by using the KYP lemma.

Sparse optimization (multi-rate systems)

Given a parameter $\lambda > 0$, find $a \in \mathbb{R}^M$ and $\gamma > 0$ that minimize

$$\tilde{J}(a, \gamma) := \gamma + \lambda \|a\|_{\ell^1} \quad (16)$$

subject to

$$\begin{bmatrix} \tilde{A}^\top X \tilde{A} - X & \tilde{A}^\top X \tilde{B} & \tilde{C}(a)^\top \\ \tilde{B}^\top X \tilde{A} & \tilde{B}^\top X \tilde{B} - \gamma I & \tilde{D}(a)^\top \\ \tilde{C}(a) & \tilde{D}(a) & -\gamma I \end{bmatrix} < 0, \\ X > 0.$$

5 Example: Analog Filter Discretization

In this section, we show an example of FIR filter design to illustrate the effectiveness of the proposed method. We assume that the target analog filter $K_c(s)$ is given by

$$K_c(s) = \frac{0.02567 s^2 + 0.2636}{s^3 + 0.594 s^2 + 0.9388 s + 0.2636}. \quad (17)$$

This is a third-order elliptic filter with 3 dB passband peak-to-peak ripple, 50 dB stopband attenuation, and 1 (rad/sec) cut-off frequency. The transfer function in (17) is obtained by MATLAB command `ellip(3, 3, 50, 1, 's')`. Figure 5 shows the frequency response of $K_c(s)$.

We set the sampling period $h = 1$ (sec), the upsampling ratio $L = 2$ (i.e., we consider a multi-rate system), and delay step $m = 2$. The analog characteristic $F(s)$ of the input signals is chosen as

$$F(s) = \frac{1}{s + 1}.$$

The fast sample/hold approximation factor N is chosen as $N = 6$. The FIR filter length M is fixed to be 64. We set $\lambda = 0.01$ for the sparse optimization in (16). We use `cvx` [7, 8] on MATLAB with the SDPT3 solver [26] to solve (16).

Under these parameters, we design the H^∞ -optimal FIR filter based on [17], and the proposed sparse FIR filter discussed in Sect. 4. Figure 6 shows the impulse response (or the FIR coefficients) of the H^∞ -optimal FIR filter. We see that the

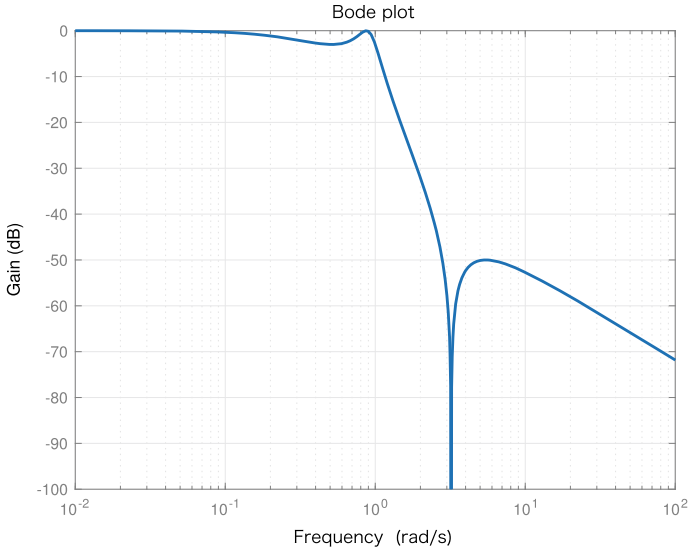


Fig. 5 Target analog filter $K_c(s)$

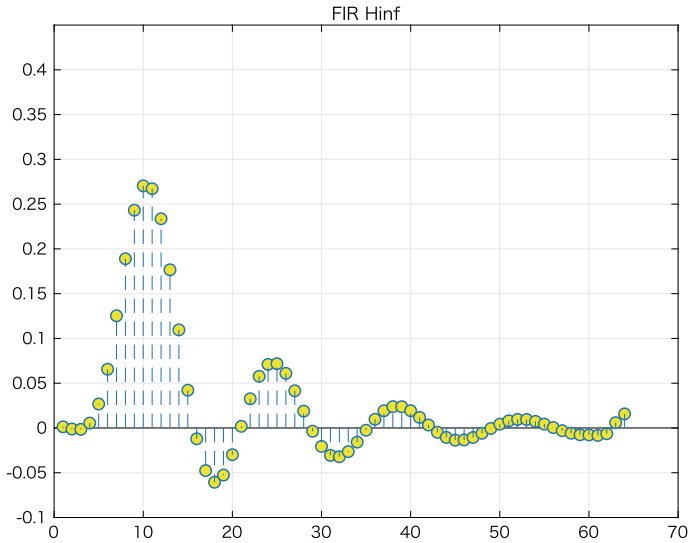


Fig. 6 Impulse response of the H^∞ -optimal FIR filter

response shows an oscillatory characteristic. On the other hand, Fig. 7 shows the impulse response of the proposed sparse FIR filter. Many of the coefficients are zero or very small (i.e. sparse), and only 17 coefficients have significant values, which is a preferable characteristic of an FIR filter for digital implementation.

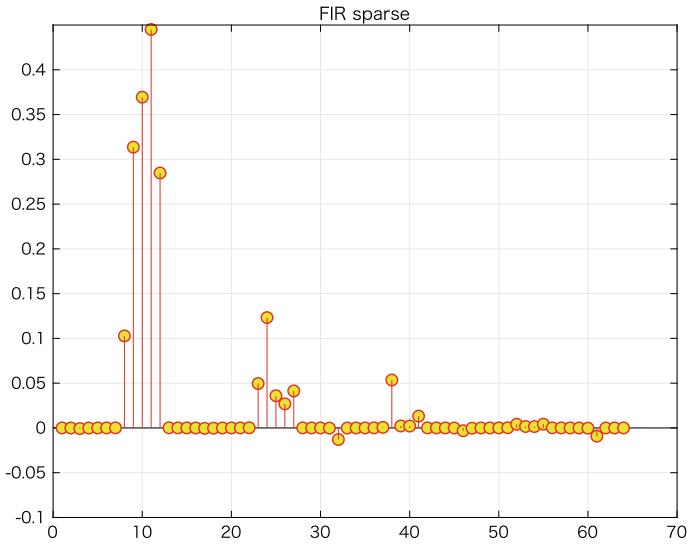


Fig. 7 Impulse response of the sparse filter

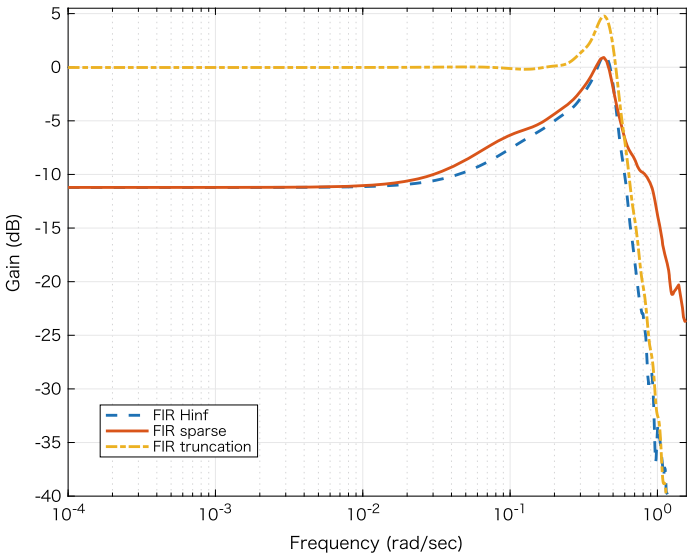


Fig. 8 Gain frequency response of sampled-data error system \mathcal{E}_L : H^∞ -optimal FIR filter (dashed line), sparse FIR filter (solid line), and truncation of H^∞ -optimal IIR filter (dash-dotted line)

To see the difference of performance between the H^∞ -optimal FIR filter and the truncated FIR filter, we show the gain of the frequency response of the sampled-data error system \mathcal{E}_L defined in (11) in Fig. 8. In this figure, we also show the response with the truncated H^∞ -optimal IIR filter proposed in [16] with the same length $M = 64$ as the proposed sparse filter. The truncated FIR filter shows a large approximation error and results in a larger H^∞ norm of the error system. Compared with that, the proposed sparse FIR filter and the H^∞ -optimal FIR filter show similar responses, although the proposed filter is much sparser than the H^∞ -optimal FIR filter. This is a merit of the use of sparse optimization.

6 Application to Delta-Sigma Modulator Design

Here, we show an application of the sparse FIR filter design to delta-sigma ($\Delta\Sigma$) modulators [24]. A delta-sigma modulator is a dynamic quantizer with a static uniform quantizer and a feedback digital filter. The feedback filter is implemented in a digital device as an FIR filter, which is chosen to minimize the quantization error. In [15], the feedback filter is designed to minimize the frequency response gain of the noise transfer function in a pre-specified low frequency range with an H^∞ norm constraint for stability. The optimization is described as an LMI optimization using the generalized KYP lemma [9].

Let us assume the feedback filter is an FIR filter of order $M - 1$ as given in (1). Then the noise transfer function is given by

$$H(z) = 1 + K(z) = 1 + \sum_{k=0}^{M-1} a_k z^{-k}. \quad (18)$$

This is the transfer function from the quantization noise, which is assumed to be an additive noise, to the quantizer output. Although the quantization is a non-linear operation, the linearized system $H(z)$ is very important for the design of delta-sigma modulators [24].

We here consider the low-pass modulator, where the noise transfer function (or the sensitivity function of the feedback system) is a high-pass filter that stops a low frequency band. In other words, the signal transfer function (or the complementary sensitivity function), which is the transfer function from the input to the output, is a low-pass system. The gain of the noise transfer function in the low frequency band should be sufficiently attenuated to obtain a good performance for low-pass signals. However, due to the waterbed effect [5], we cannot arbitrarily attenuate the gain in the low frequency band without increasing the gain at high frequencies. To obtain an optimal solution, we propose to describe the design problem as an H^∞ optimal control problem. Let $\Omega \in (0, \pi)$ be the bandwidth of the low-pass modulator. Then the optimization is described as a min-max optimization:

$$\underset{H: \text{stable}}{\text{minimize}} \quad \max_{\omega \in [0, \Omega]} |H(e^{j\omega})| \quad \text{subject to} \quad \max_{\omega \in [0, \pi]} |H(e^{j\omega})| < \beta, \quad (19)$$

where $\beta > 0$ is an upper bound of the H^∞ norm of the noise transfer function. If we consider an FIR filter for $H(z)$ as in (18), we obtain the optimization for the filter coefficients as

$$\begin{aligned} & \underset{a_0, \dots, a_{M-1} \in \mathbb{R}}{\text{minimize}} && \max_{\omega \in [0, \Omega]} \left| 1 + \sum_{k=0}^{M-1} a_k e^{-j\omega k} \right| \\ & \text{subject to} && \max_{\omega \in [0, \pi]} \left| 1 + \sum_{k=0}^{M-1} a_k e^{-j\omega k} \right| < \beta. \end{aligned} \quad (20)$$

This is equivalently reduced to an LMI optimization by the generalized KYP lemma [15].

Then, borrowing the idea of sparse representation mentioned in the previous sections, we seek a sparse coefficient vector for the feedback filter to add an ℓ^1 regularization term to the cost function in (20). By using the generalized KYP lemma, we obtain the following optimization problem.

Sparse feedback filter design in delta-sigma modulator

Given a parameter $\lambda > 0$, find $a \in \mathbb{R}^M$ and $\gamma > 0$ that minimize

$$J(a, \gamma) := \gamma + \lambda \|a\|_{\ell^1}$$

subject to

$$\begin{aligned} & \begin{bmatrix} A^\top Y A + Z A + A^\top Z - Y - 2Z \cos \Omega & A^\top Y B + Z B & a \\ & B^\top Y A + B^\top Z & B^\top Y B - \gamma^2 & 1 \\ & a^\top & 1 & -1 \end{bmatrix} < 0, \\ & \begin{bmatrix} A^\top X A - X & A^\top X B & a \\ & B^\top X A & B^\top X B - \beta^2 & 1 \\ & a^\top & 1 & -1 \end{bmatrix} < 0, \\ & X > 0, \\ & Z > 0. \end{aligned}$$

This problem is also a convex optimization problem and can be efficiently solved via numerical optimization softwares.

We show a design example with $\Omega = \pi/32$ and $\beta = 1.5$. Figure 9 shows the obtained sparse filter coefficients with $\lambda = 10$. Also, we show the full coefficients obtained by the optimization (20) with $\lambda = 0$. We see from the result that the sparse filter has just 4 non-zero coefficients. We then compute the frequency response of the noise transfer function $H(z)$ in (18). For comparison, we also compute the 4-

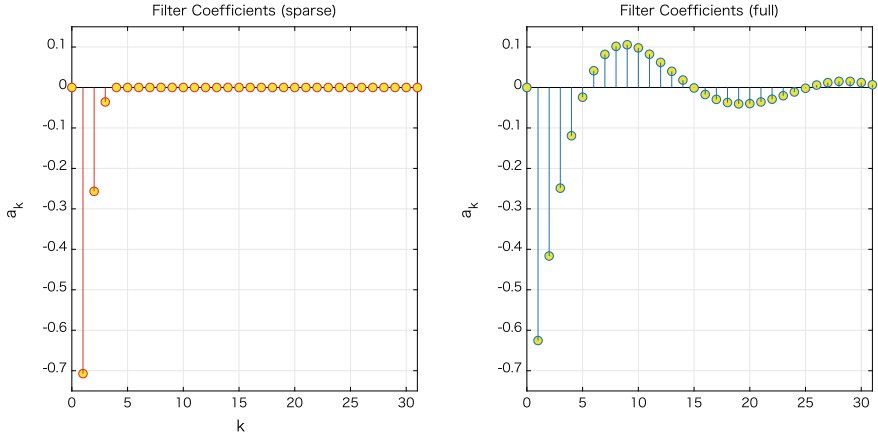


Fig. 9 Filter coefficients: sparse filter (left) and full filter (right)

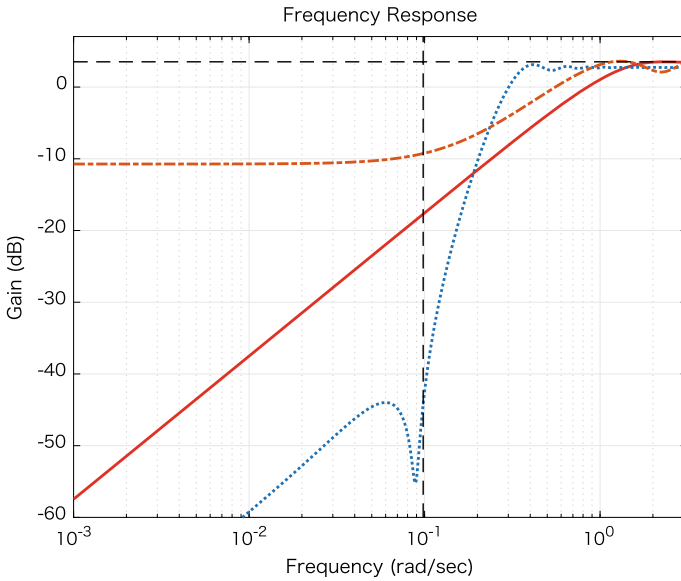


Fig. 10 Frequency response of the noise transfer function: sparse filter (solid), truncated filter (dash-dotted), and ideal response (dotted)

tap truncated coefficients from the full coefficients. Figure 10 shows the frequency responses of the noise transfer functions by the sparse filter, the truncated filter, and the ideal full filter. The truncated filter shows a poor response in the low frequency; it does not have enough attenuation. On the other hand, the sparse filter shows a better response. For more details, see [18].

7 Conclusion

In this chapter, we have proposed a design method of a sparse FIR filter that optimally approximates a given analog filter with minimization of a sampled-data H^∞ performance index and enhancement of the sparsity of the FIR coefficients simultaneously. The design problem is described as a convex optimization problem, which can be efficiently solved by numerical optimization softwares. We have also extended the proposed method to multi-rate systems. A design example has shown to illustrate the effectiveness of the proposed method. We also show an application to the design of feedback filters in delta-sigma modulators. Future work includes multiplier-free implementation of H^∞ -optimal FIR digital filters as discussed in [11, 23]. Also, it is an important but open problem to give a practical sufficient conditions for guaranteeing the ℓ^1 solution of Problem 2 to be also ℓ^0 optimal. Here a practical condition means that it is easier to check in view of computational complexity than directly solving the original ℓ^0 problem that is known to be NP hard [20, 29].

References

1. Anderson, B.D.O.: A system theory criterion for positive real matrices. *SIAM J. Control Optim.* **5**(2), 171–182 (1967)
2. Candes, E.J., Romberg, J.K., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pure Appl. Math.* **59**, 1207–1223 (2006)
3. Chen, T., Francis, B.A.: *Optimal Sampled-data Control Systems*. Springer (1995)
4. Donoho, D.L.: Compressed sensing. *IEEE Trans. Inf. Theory* **52**(4), 1289–1306 (2006)
5. Doyle, J., Francis, B.A., Tannenbaum, A.: *Feedback Control Theory*. Macmillan Publishing (1990)
6. Elad, M.: *Sparse and Redundant Representations*. Springer (2010)
7. Grant, M., Boyd, S.: Graph implementations for nonsmooth convex programs. *Recent Advances in Learning and Control*, pp. 95–110. Springer-Verlag Limited (2008)
8. Grant, M., Boyd, S.: CVX: Matlab software for disciplined convex programming, version 1.21 (2011). <http://cvxr.com/cvx>
9. Iwasaki, T., Hara, S.: Generalized KYP lemma: unified frequency domain inequalities with design applications. *IEEE Trans. Automat. Contr.* **50**(1), 41–59 (2005)
10. Kootsookos, P.J., Bitmead, R.B., Green, M.: The Nehari shuffle: FIR(q) filter design with guaranteed error bounds. *IEEE Trans. Signal Process.* **40**(8), 1876–1883 (1992)
11. Macleod, M.D., Dempster, A.G.: Multiplierless FIR filter design algorithms. *Signal Process. Lett. IEEE* **12**(3), 186–189 (2005)
12. Nagahara, M.: Min-max design of FIR digital filters by semidefinite programming. *Applications of Digital Signal Processing*, pp. 193–210. InTech (2011)
13. Nagahara, M.: *Sparsity Methods for Systems and Control*. Now Publishers (2020)
14. Nagahara, M., Quevedo, D.E., Nešić, D.: Maximum hands-off control: A paradigm of control effort minimization. *IEEE Trans. Automat. Contr.* **61**(3), 735–747 (2016)
15. Nagahara, M., Yamamoto, Y.: Frequency domain min-max optimization of noise-shaping delta-sigma modulators. *IEEE Trans. Signal Process.* **60**(6), 2828–2839 (2012)
16. Nagahara, M., Yamamoto, Y.: Optimal discretization of analog filters via sampled-data H^∞ control theory. In: *Proceedings of the 2013 IEEE Multi-Conference on Systems and Control (MSC 2013)*, pp. 527–532 (2013)

17. Nagahara, M., Yamamoto, Y.: FIR digital filter design by sampled-data H^∞ discretization. In: Proceedings of the 19th IFAC World Congress, pp. 3110–3115 (2014)
18. Nagahara, M., Yamamoto, Y.: Sparse representation of feedback filters in delta-sigma modulator. In: Proceedings of the 21st IFAC World Congress (2020). (to appear)
19. Nagahara, M., Ogura, M., Yamamoto, Y.: H^∞ design of periodically nonuniform interpolation and decimation for non-band-limited signals. *SICE J. Control Meas. Syst. Integr.* **4**(5), 341–348 (2011)
20. Natarajan, B.K.: Sparse approximate solutions to linear systems. *SIAM J. Comput.* **24**(2), 227–234 (1995)
21. Oppenheim, A.V., Schaffer, R.W.: *Discrete-Time Signal Processing*, 3rd edn. Prentice Hall (2009)
22. Rantzer, A.: On the Kalman-Yakubovich-Popov lemma. *Syst. Control Lett.* **28**(1), 7–10 (1996)
23. Samuelli, H.: An improved search algorithm for the design of multiplierless FIR filters with powers-of-two coefficients. *IEEE Trans. Cir. Syst.* **36**(7), 1044–1047 (1989)
24. Schreier, R., Temes, G.C.: *Understanding Delta-Sigma Data Converters*. Wiley Interscience (2005)
25. Sturm, J.F.: Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optim. Methods Softw.* **11**(12), 625–653 (1999). <http://sedumi.ie.lehigh.edu/>
26. Toh, K.C., Todd, M.J., Tütüncü, R.H.: SDPT3—a Matlab software package for semidefinite programming, version 1.3. *Optim. Methods Softw.* **11**(1), 545–581 (1999)
27. Tuqan, J., Vaidyanathan, P.P.: The role of the discrete-time Kalman-Yakubovitch-Popov lemma in designing statistically optimum FIR orthonormal filter banks. In: Proceedings of the 1998 IEEE International Symposium on Circuits and Systems, 1998. ISCAS '98, vol. 5, pp. 122–125 (1998)
28. Vaidyanathan, P.P.: *Multirate Systems and Filter Banks*. Prentice Hall (1993)
29. Vidyasagar, M.: *An Introduction to Compressed Sensing*. SIAM (2020)
30. Yamamoto, Y., Anderson, B.D.O., Nagahara, M., Koyanagi, Y.: Optimizing FIR approximation for discrete-time IIR filters. *IEEE Signal Process. Lett.* **10**(9), 273–276 (2003)
31. Yamamoto, Y., Madievski, A.G., Anderson, B.D.O.: Approximation of frequency response for sampled-data control systems. *Automatica* **35**(4), 729–734 (1999)
32. Yamamoto, Y., Nagahara, M., Khargonekar, P.P.: Signal reconstruction via H^∞ sampled-data control theory—Beyond the Shannon paradigm. *IEEE Trans. Signal Process.* **60**(2), 613–625 (2012)

Analysis of a Reduced Model of Epithelial–Mesenchymal Fate Determination in Cancer Metastasis as a Singularly-Perturbed Monotone System



M. Ali Al-Radhawi and Eduardo D. Sontag

Abstract Tumor metastasis is one of the main factors responsible for the high fatality rate of cancer. Metastasis can occur after malignant cells transition from the epithelial phenotype to the mesenchymal phenotype. This transformation allows cells to migrate via the circulatory system and subsequently settle in distant organs after undergoing the reverse transition from the mesenchymal to the epithelial phenotypes. The core gene regulatory network controlling these transitions consists of a system made up of coupled SNAIL/miRNA-34 and ZEB1/miRNA-200 subsystems. In this work, we formulate a mathematical model of the core regulatory motif and analyze its long-term behavior. We start by developing a detailed reaction network with 24 state variables. Assuming fast promoter and mRNA kinetics, we then show how to reduce our model to a monotone four-dimensional system. For the reduced system, monotone dynamical systems theory can be used to prove generic convergence to the set of equilibria for all bounded trajectories. The theory does not apply to the full model, which is not monotone, but we briefly discuss results for singularly-perturbed monotone systems that provide a tool to extend convergence results from reduced to full systems, under appropriate time separation assumptions.

Keywords Monotone systems · Generic convergence · Cancer metastasis · Epithelial-mesenchymal transition · Model reduction · Singular perturbations

M. Ali Al-Radhawi
Department of Electrical and Computer Engineering, Northeastern University,
360 Huntington Avenue, Boston, MA 02115, USA
e-mail: malirdwi@northeastern.edu

E. D. Sontag (✉)
Departments of Bioengineering and Electrical and Computer Engineering,
Departments of Mathematics and Chemical Engineering (affiliate), Northeastern University,
360 Huntington Avenue, Boston, MA 02115, USA
e-mail: e.sontag@northeastern.edu

Laboratory of Systems Pharmacology, Program in Therapeutic Science,
Harvard Medical School, Boston, MA 02115, USA

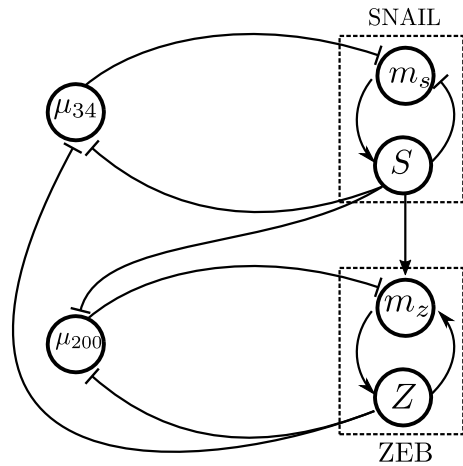
1 Introduction

Realistic dynamical models of physical systems are often very complex and high-dimensional, and this is especially so in molecular cell biology. Effective analysis often requires simplifications through model reduction techniques. A traditional approach to model reduction in biology is to take advantage of time-scale separation. Indeed, most interesting processes in biology are made up of subsystems which operate at different time scales, thus allowing fast subprocesses to be “averaged out” at the observational time scale. The rigorous mathematical analysis of time scale separation, and in particular the mathematical field of *singular perturbations*, owe much to the pioneering work of Tikhonov [1]. Singular perturbation theory now plays a key role in both science and engineering [2].

There are many examples of the ubiquitousness of time-scale separation in molecular biology. An early example is the study of enzymatic reactions, where the derivation of *Michaelis-Menten* kinetics in 1913 [3] is still widely used [4]. Another example is provided by Gene Regulation Networks (GRNs) which naturally have multiple levels of time scales: external stimuli change Transcription Factor (TF) activities in milliseconds, promoter kinetics equilibrate in seconds, transcription and translation take minutes, and protein kinetics are in the order of tens of minutes to hours [5].

In this paper we study a GRN that determines cell-fate in the metastasis of cancerous tumors. This network regulates a transition between two cell types: *epithelial cells*, which line the external and internal surfaces of many organs, and *mesenchymal stem cells*, which are multipotent connective tissue cells that can differentiate into other type of cells such as muscles, bone, etc. Bidirectional transitions between these two cell types can happen, and they are referred to as “Epithelial to Mesenchymal” transitions (EMT) and “Mesenchymal to Epithelial” transitions (MET). During the EMT process, a cell loses adhesion to neighbouring cells, and becomes more invasive and migratory. It is worth noting that both EMTs and METs are part of normal developmental processes such as embryogenesis and tissue healing. Nevertheless, they are of one of main mechanisms of tumor metastasis. After undergoing EMTs, cancerous cells travel through the blood as Circulating Tumor Cells (CTCs). These CTCs settle in other organs by undergoing METs, and they subsequently multiply, thus giving rise to metastatic tumors [6, 7]. Because of their key role in cancer, the identification of the GRNs enabling EMTs/METs has been a focus of a large research effort. Transcription factors (TFs) such as SNAIL, SLUG, TWIST, and ZEB1 have been studied in great detail [8, 9]. Cellular signals such as p54, Notch, EGF, Wnt, HIF-1 α , and others can induce EMTs/METs [6]. These signals act on a core four-component network that involves the upstream SNAIL/miR-34 circuit and the downstream ZEB1/miR-200 circuit [10, 11]. The conceptual organization for such a circuit is depicted in Fig. 1. Each pair reproduces the standard toggle switch architecture, i.e., mutual inhibition. However, this design differs by having a mixed inhibition mechanism: the TFs ZEB1 and SNAIL inhibit the miRNAs (μ_{34} , μ_{200}) at the transcriptional level, while μ_{34} , μ_{200} inhibit SNAIL and ZEB1, respectively, at the translational level. Therefore, such circuits are known as *chimeric* circuits [10].

Fig. 1 The core EMT/MET network [10], which involves proteins, mRNAs, microRNAs (miRNAs) and the underlying genes which are not explicitly depicted. The mRNAs of SNAIL and ZEB1 are denoted by m_s , m_z , respectively. An arrow of the form “ \rightarrow ” denotes activation, while “ \dashv ” denotes inhibition. The detailed reaction network model is presented in Sect. 4



One of the contributions of this work is to translate the conceptual diagram in Fig. 1 into a precise mathematical model. Our model, while not completely novel, clarifies and provides explicit details of several features of models found in the literature. More importantly, we analyze the long-term behavior of a reduced model obtained under a natural and biologically realistic time scale separation assumption. We prove a theorem guaranteeing “almost-sure” convergence of trajectories to steady states. This paper also serves to motivate a powerful but relatively unknown theorem on singularly perturbed monotone systems. A central and well-known result for monotone systems is Hirsch’s Generic Convergence Theorem [12–16], which guarantees that almost every bounded solution of a strongly monotone system converges to the set of steady states. This theory has been widely applied to biochemical systems [17–19]. However, many biological models are not monotone. For example, monotonicity with respect to orthant cones rules out negative feedback loops, which are key components of homeostasis and adaptive systems. Indeed, the core EMT/MET network that is the focus of this paper has a negative feedback loop between S and m_s (see Fig. 1). Nevertheless, we can take advantage of the fact that transcription happens at a faster time-scale than translation. Moreover, despite the fact that miRNAs and mRNAs belong to the same class of biochemical molecules, comparison of their half-lives reveals that mRNAs have much shorter half-lives than miRNAs and proteins [20, 21]. Hence, the negative loop in the core EMT/MET network is a “fast” loop. Intuitively, negative loops that act at a comparatively fast time scale should not affect the main characteristics of monotone behavior. This has been rigorously studied in [22] (see the Ph.D. thesis [23] for more details), using tools from geometric singular perturbation theory.

The paper is organized as follows. In Sect. 2 we model the core EMT/MET network in detail, and derive a reduced model. Section 3 reviews several basic definitions and theorems about monotone systems, and in Sect. 4 we analyze the reduced model theoretically and we review results that can be used to extend convergence from reduced to full systems.

2 Modeling of the EMT/MET System

A mathematical model for the EMT/MET system has been presented in [10, 24] based on several assumptions that include detailed balance. Here, we drop such assumptions, and develop a more “mechanistic” model via the framework of Chemical Reaction Networks (CRN). We review the notation briefly [25, 26]. A CRN is a set of species $\mathcal{S} = \{Z_1, \dots, Z_n\}$ and a set of reactions $\mathcal{R} = \{R_1, \dots, R_\nu\}$. A reaction R_j can be written as: $\sum_{i=1}^n \alpha_{ij} Z_i \rightarrow \sum_{i=1}^n \beta_{ij} Z_i$.

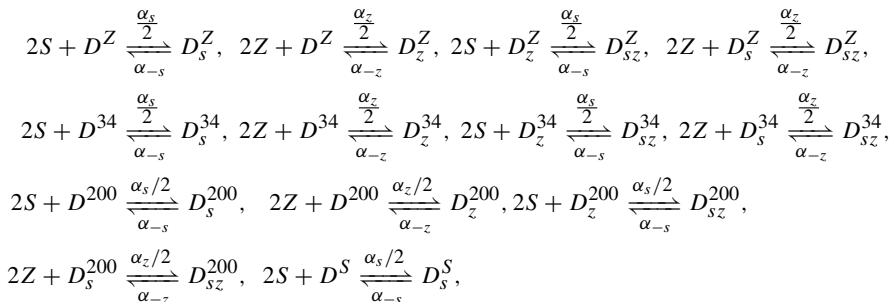
A stoichiometry matrix $\Gamma \in \mathbb{R}^{n \times \nu}$ is defined elementwise as $[\Gamma]_{ij} = \beta_{ij} - \alpha_{ij}$. The reactions are associated with a rate function $R : \mathbb{R}_{\geq 0}^n \rightarrow \mathbb{R}_{\geq 0}^\nu$. We assume that R takes the form of Mass-Action kinetics: $R_j(z) = \prod_{i=1}^n k_j z_i^{\alpha_{ij}}$, where k_j is the kinetic constant. Let $z(t) \in \mathbb{R}_{\geq 0}^n$ be the vector of species concentrations at time t . The associated ODE can be written as: $\dot{z} = \Gamma R(z)$.

We model GRNs as CRNs via the central dogma of molecular biology (see [27] for a detailed framework). Each gene is associated with promoter states, an mRNA state, and a protein state. As mentioned in the introduction, we can assume that the promoter kinetics and the mRNA kinetics are fast. In this section, we will derive the dynamics of the fast and slow systems. We will show that the model can be reduced from *24 dimensions* to *four dimensions*. The slow states are $S, Z, \mu_{200}, \mu_{34}$. We will start with promoter kinetics.

2.1 Promoter Dynamics

For each gene j , we denote the promoter states by species of the form D_i^j . The superscript denotes the gene, while the subscript denotes the occupancy of the binding sites. Let $D^{34}, D^{200}, D^S, D^Z$ be the unbound promoters for $\mu_{34}, \mu_{200}, \text{SNAIL},$ and ZEB1 , respectively. For instance, D_s^{34} denotes S binding to the promoter of μ_{34} , while D_{sz}^{34} denotes both S, Z binding to the promoter of μ_{34} .

Hence, the CRN describing promoter dynamics for the EMT/MET network (Fig. 1) can be written as:



The concentration of a species Y will be denoted, if no other notation is used, by $[Y]$. The reaction structure implies that we have the conservation law $\sum_i [D_i^j](t) = E_j$ for each gene, where $E_j, j \in \{S, Z, 34, 200\}$ is the total concentration available in the medium. Hence, E_j stays constant during the course of the reaction. Note that each TF is assumed to bind to its promoter as a dimer. The same analysis can be repeated if the TF binds as an n -mer for some integer $n \geq 1$.

Since binding/unbinding kinetics are usually fast [5], we will approximate all the promoter states with their quasi steady state (QSS) approximations. We are interested in deriving the expression for the *active* promoter states. Since both S, Z repress μ_{34}, μ_{200} , then the active states are the unbound promoter states D^{34}, D^{200} , respectively. The promoter D^S is the active state for SNAIL since it is a self-repressing gene. Finally, D_{sz}^Z is the active state for ZEB1 because it is activated by both SNAIL and self-binding.

We will consider ZEB1 as an example. Let $d_z(t) := [[D^Z], [D_s^Z], [D_z^Z], [D_{sz}^Z]]^T(t)$. Using the reactions above, we can write the ODE for the promoters of Z to get:

$$\dot{d}_z = P_z(s, z)d_z := \begin{bmatrix} -\alpha_s s^2 - \alpha_z z^2 & \alpha_{-s} & \alpha_{-z} & 0 \\ \alpha_s s^2 & -\alpha_{-s} - z^2 \alpha_z & 0 & \alpha_{-z} \\ \alpha_z z^2 & 0 & -\alpha_{-z} - s^2 \alpha_s & \alpha_{-s} \\ 0 & z^2 \alpha_z & s^2 \alpha_s & -\alpha_{-s} - \alpha_{-z} \end{bmatrix} d_z. \tag{1}$$

Note that s, z are slow variables. Setting the derivatives to zero and substituting the conservation law we get:

$$[D_{sz}^Z]_{qss} = \frac{E_Z s^2 z^2}{(s^2 + A_s)(z^2 + A_z)},$$

where $A_s := \alpha_{-s}/\alpha_s, A_z := \alpha_{-z}/\alpha_z$. Similarly, we can define the matrices $P_s(s, z), P_{34}(s, z), P_{200}(s, z)$. Hence, we get:

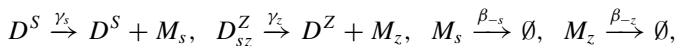
$$[D^{200}]_{qss} = \frac{E_{200} A_s A_z}{(s^2 + A_s)(z^2 + A_z)}, [D^S]_{qss} = \frac{E_S A_s}{(s^2 + A_s)}, [D^{34}]_{qss} = \frac{E_{34} A_s A_z}{(s^2 + A_s)(z^2 + A_z)}.$$

2.2 RNA Dynamics

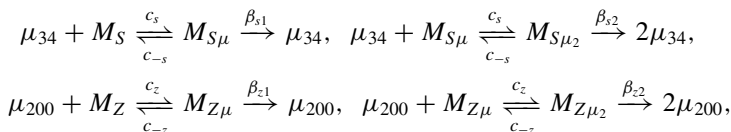
Translational inhibition by miRNAs is achieved by an RNA-induced Silencing Complex (RISC). RISC binds to the target mRNA and degrades it via the Argonaute protein [28]. Here, we assume a simple model in which miRNA binds to the target mRNA to inhibit translation and activate degradation. In general, multiple miRNAs can bind

to a single mRNA. Here we assume that each mRNA can have up to two binding sites. For each additional miRNA binding, translation is inhibited and degradation is accelerated.

The CRN for the transcription of SNAIL and ZEB1 can be written as follows:



where M_s, M_z are the species denoting the mRNAs of SNAIL and ZEB1, respectively. The CRN for the miRNA-mRNA reaction (with two binding sites) can be written as follows:



Since miRNAs actively degrade mRNAs, we assume that:

$$\beta_s < \beta_{s1} < \beta_{s2}, \quad \beta_z < \beta_{z1} < \beta_{z2}. \quad (2)$$

Remembering that μ_{200} , and z are the slow variables, we can write the ODE for ZEB1 translational dynamics as follows (with $m_z(t) := [[M_z], [M_{z\mu}], [M_{z\mu_2}]]^T(t)$):

$$\begin{aligned} \dot{m}_z &= Q_z(\mu_{200})m_z + b_z D_{s_z}^Z \\ &:= \begin{bmatrix} -c_z\mu_{200} - \beta_z & c_{-z} & 0 \\ c_z\mu_{200} & -\beta_{z1} - c_z\mu_{200} - c_{-z} & c_{-z} \\ 0 & c_z\mu_{200} & -c_{-z} - \beta_{z2} \end{bmatrix} m_z + \begin{bmatrix} \gamma_z \\ 0 \\ 0 \end{bmatrix} D_{s_z}^Z \end{aligned} \quad (3)$$

Since (3) is a linear system in μ_{200} , the quasi-steady state can be solved easily by matrix inversion. Substituting the QSS for $[D^Z]$ we get:

$$m_{z,qss} = \frac{\gamma_z E_Z s^2 z^2}{(s^2 + A_s)(z^2 + A_z)} \begin{bmatrix} \beta_{z2}c_z\mu_{200} + e_{z1} \\ e_{z2}\mu_{200} \\ c_z^2\mu_{200}^2 \end{bmatrix} \frac{1}{\beta_{z2}c_z^2\mu_{200}^2 + e_{z3}c_z\mu_{200} + e_{z1}\beta_z}, \quad (4)$$

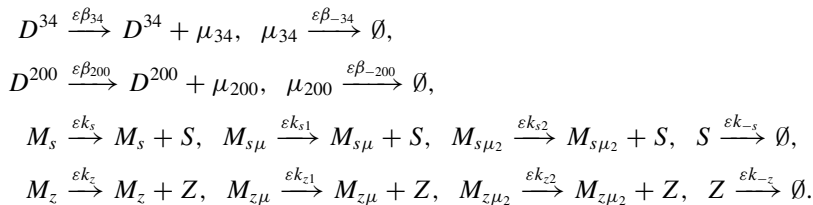
where $e_{z1} := \beta_{z1}\beta_{z2} + (\beta_{z1} + \beta_{z2})c_{-z} + c_{-z}^2$, $e_{z2} := c_z(\beta_{z2} + c_{-z})$, $e_{z3} := \beta_z\beta_{z2} + \beta_{z1}\beta_{z2} + \beta_{z1}c_{-z}$. The dynamics of m_s can be derived similarly and we get:

$$m_{s,qss} = \frac{\gamma_s E_S A_s}{(s^2 + A_s)} \begin{bmatrix} \beta_{s2}c_s\mu_{34} + e_{s1} \\ e_{s2}\mu_{34} \\ c_s^2\mu_{34}^2 \end{bmatrix} \frac{1}{\beta_{s2}c_s^2\mu_{34}^2 + e_{s3}c_s\mu_{34} + e_{s1}\beta_s}, \quad (5)$$

where $e_{s1} := \beta_{s1}\beta_{s2} + (\beta_{s1} + \beta_{s2})c_{-s} + c_{-s}^2$, $e_{s2} := c_s(\beta_{s2} + c_{-s})$, $e_{s3} := \beta_s\beta_{s2} + \beta_{s1}\beta_{s2} + \beta_{s1}c_{-s}$.

2.3 Slow Dynamics

The complete reaction model requires modeling the production of the proteins and the miRNAs. The CRN can be written as follows:



The kinetic constants are multiplied by ε to emphasize that the corresponding reactions are slow. Note that since miRNA inhibits translation, the mRNA-miRNA complexes have lower translation rates than raw mRNAs. Hence, we have:

$$k_s > k_{s1} > k_{s2}, \quad k_z > k_{z1} > k_{z2}. \quad (6)$$

The model above assumes that Z is produced only when both Z , and S are bound to ZEB1's promoter. But this is not realistic, since constitutive transcription is always present at low levels. Hence, we model this by a reaction $\emptyset \xrightarrow{\varepsilon\delta_z} Z$ with δ_z small. This will help us also show generic convergence for the reduced system in Sect. 4.

After substituting the QSSs for the fast variables in the slow system, all the terms corresponding to promoter and mRNA reactions vanish. Hence, the corresponding ODE can be written as follows:

$$\begin{aligned}
 \dot{\mu}_{34} &= \beta_{34}[D^{34}] - \beta_{-34}\mu_{34} = \frac{E_{34}A_zA_s\beta_{34}}{(s^2 + A_s)(z^2 + A_z)} - \beta_{-34}\mu_{34} \\
 \dot{s} &= [k_s k_{s1} k_{s2}]m_s - k_{-s}s = [k_s k_{s1} k_{s2}]m_{s,qss} - k_{-s}s \\
 \dot{\mu}_{200} &= \beta_{200}[D^{200}] - \beta_{-200}\mu_{200} = \frac{E_{200}\beta_{200}A_zA_s}{(s^2 + A_s)(z^2 + A_z)} - \beta_{-200}\mu_{200} \\
 \dot{z} &= \delta_z + [k_z k_{z1} k_{z2}]m_z - k_{-z}z = \delta_z + [k_z k_{z1} k_{z2}]m_{z,qss} - k_{-z}z.
 \end{aligned} \quad (7)$$

where $m_{z,qss}$, $m_{s,qss}$ are given by (4), (5), respectively.

3 Monotone Systems and Singular Perturbations

3.1 Monotone Systems

In this section, we review basic definitions and results regarding monotone systems. We base our discussion on [13, 16, 23].

A nonempty, closed set $C \subset \mathbb{R}^N$ is said to be a *cone* if $C + C \subset C$, $\alpha C \subset C$ for all $\alpha > 0$, and $C \cap (-C) = \{0\}$. For each cone C , a partial order on \mathbb{R}^N can be associated. For any $x, y \in \mathbb{R}^N$, we define:

$$\begin{aligned} x \geq y &\Leftrightarrow x - y \in C \\ x > y &\Leftrightarrow x - y \in C, x \neq y. \end{aligned}$$

When C° is not empty, we can define $x \gg y \Leftrightarrow x - y \in C^\circ$.

In this paper, we only consider cones which are *orthants* of \mathbb{R}^N . In order to identify the various orthants, let $\sigma \in \{\pm 1\}^N$ and let

$$C_\sigma = \{z \in \mathbb{R}^n \mid \sigma_i z_i \geq 0, 1 \leq i \leq n\}$$

be the corresponding orthant cone. Let \preceq_σ be the corresponding partial order.

A set $W \subseteq \mathbb{R}^N$ is said to be p -convex, if W contains the line joining x and y whenever $x \preceq y$, $x, y \in W$. Hence, equipped with a partial ordering on a p -convex and *open* set W we study the ordinary differential equation:

$$\frac{dz}{dt} = F(z), \tag{8}$$

where $F : W \rightarrow \mathbb{R}^N$ is a C^1 vector field. We assume that the system is forward invariant with respect to W . In our application in this paper we will use $W = \mathbb{R}_+^N$, the *open* positive orthant.

We are interested in a special class of equations which preserve the partial order along all trajectories.

Definition 1 The flow ϕ_t of (8) is said to have positive derivatives on a set $W \subseteq \mathbb{R}^N$, if $[\frac{\partial}{\partial z} \phi_t(z)]x \in C^\circ$ for all $x \in C \setminus \{0\}$, $z \in W$, and $t > 0$.

Definition 2 The system (8) is called *monotone* (resp. strongly monotone) on a set $W \subseteq \mathbb{R}^N$ if for all $t > 0$ and all $z_1, z_2 \in W$,

$$z_1 \geq z_2 \Rightarrow \phi_t(z_1) \geq \phi_t(z_2) \text{ (resp. } \phi_t(z_1) \gg \phi_t(z_2) \text{ when } z_1 \neq z_2).$$

Establishing that a flow has positive derivatives can be performed by verifying the irreducibility of the Jacobian as the following theorem states:

Proposition 1 Assume that (8) is monotone with respect to an orthant cone C . If $\frac{\partial F}{\partial z}(z)$ is irreducible for all $z \in W$ then the flow ϕ_t of (8) has positive derivatives on the set $W \subseteq \mathbb{R}^N$.

The following result can be interpreted as saying that having positive derivatives is an “infinitesimal” version of strong monotonicity.

Proposition 2 Let $W \subset \mathbb{R}^N$ be p -convex and open. If the flow ϕ_t has positive derivatives in W , then the associated ODE is strongly monotone on W .

Hence, we have the following corollary:

Corollary 1 *Let (8) and W be given as above. Assume that (8) is monotone with respect to an orthant cone C . If $\frac{\partial F}{\partial z}(z)$ is irreducible for all $z \in W$, then the system (8) is strongly monotone with respect to C on W .*

This is a rephrasing of the well-known Hirsch Generic Convergence Theorem:

Theorem 1 *Assume that (8) is strongly monotone on a p -convex open set $W \subseteq \mathbb{R}^N$. Let $W^c \subseteq W$ be defined as the subset of points whose forward orbit has compact closure in W . If the set of equilibria is totally disconnected, then the forward trajectory starting from almost every point in W^c converges to an equilibrium.*

By Corollary 1, it is sufficient to check monotonicity and irreducibility to establish generic convergence.

3.2 Graphical Characterization of Monotonicity

Monotonicity with respect to orthant cones can be characterized via *Kamke’s conditions*. We review the relevant material from [16, 18]. Let $\sigma \in \{\pm 1\}^N$ and let C_σ be the corresponding orthant as defined above. Let $\Sigma = \text{diag}(\sigma)$, i.e., a diagonal matrix with σ as the diagonal. Then, the following holds:

Theorem 2 (Kamke’s conditions) *Let (8) be given and let ϕ_t be the associate flow. Let $J = \frac{\partial F}{\partial z}$ be the corresponding Jacobian. Let $W \subseteq \mathbb{R}^N$ be a p -convex open set. Assume that there exists $\sigma \in \{\pm 1\}^N$ such that $\Sigma J \Sigma$ is Metzler on W (i.e., all non-diagonal entries are non-negative). Then the corresponding flow is monotone on W with respect to the partial order \preceq_σ .*

The last proposition gives a useful characterization for monotonicity with respect to orthant cones; however, it requires checking 2^N possible sign combinations. Alternatively, a simple graphical criteria can be stated for a graph derived from the Jacobian. Informally, it states that the system is orthant monotone if every loop has a net positive sign. We state it more formally next.

We say that a Jacobian J is *sign-stable* on a set W if for each $i \neq j$, $\text{sgn}(J_{ij})$ is constant on W . Define a signed directed graph G with vertices $\{1, \dots, n\}$. There is an edge connecting vertices i to j if the partial derivative J_{ij} does not vanish identically. The sign of the edge is equal to sign of J_{ij} . A *loop* is any sequence of edges (without regard to direction) that does not traverse a vertex twice and it starts and ends with the same vertex. The *sign* of a loop is the product of the signs of the constituent edges. It is worth noting that diagonal entries, i.e., “self-loops”, have no effect on the validity of the results.

Theorem 3 (Positive Loop Property) *Let (8) be given. Let $J = \frac{\partial F}{\partial z}$ be the corresponding Jacobian. Assume J is sign-stable. Let $W \subseteq \mathbb{R}^N$ be a p -convex open*

set. Let G be the graph defined above. If every loop has a positive sign, then there exists $\sigma \in \{\pm 1\}^N$ such that $\Sigma J \Sigma$ is Metzler on W (i.e., all non-diagonal entries are non-negative).

3.3 Singular Perturbation Model Reduction

The process of mathematical modeling of physical processes involves usually reduction of the dynamics of fast variables. For instance, self-loops, i.e., terms corresponding to the diagonal entries of the Jacobian J , are often approximations of fast dynamics. Hence, it is intuitive to expect that the theorems in the previous section hold for *sufficiently fast* negative self-loops.

Hence, we consider a system in the singularly perturbed form:

$$\begin{aligned} \frac{dx}{dt} &= f_0(x, y, \varepsilon) \\ \varepsilon \frac{dy}{dt} &= g_0(x, y, \varepsilon), \end{aligned} \tag{9}$$

where $f_0 : \mathbb{R}_+^n \times \mathbb{R}_+^m \times [0, \bar{\varepsilon}] \rightarrow \mathbb{R}_+^n$, $g_0 : \mathbb{R}_+^n \times \mathbb{R}_+^m \times [0, \bar{\varepsilon}] \rightarrow \mathbb{R}_+^m$ are smooth bounded functions and $\bar{\varepsilon} > 0$ is fixed. Furthermore, we assume that the equilibrium set is totally disconnected for all $\varepsilon < \bar{\varepsilon}$. These assumptions are automatically satisfied in our case since Mass-Action kinetics give rise to polynomial systems.

For $0 < \varepsilon \ll 1$, the dynamics of x are much slower than y . If $\varepsilon \neq 0$, we can change the time scale to $\tau = t/\varepsilon$, and study the equivalent form:

$$\begin{aligned} \frac{dx}{d\tau} &= \varepsilon f_0(x, y, \varepsilon) \\ \frac{dy}{d\tau} &= g_0(x, y, \varepsilon). \end{aligned} \tag{10}$$

Model reduction via singular perturbations requires solving the fast system at a quasi-steady state, hence we assume that there exists a smooth bounded function

$$m_0 : \mathbb{R}_m^+ \rightarrow \mathbb{R}_n^+$$

such that $g_0(x, m_0(x), 0) = 0$ for all $x \in \mathbb{R}_n^+$. From the previous section it can be seen that m_0 exists and is unique as we have derived all the QSS expressions uniquely.

For simplicity, let $z = y - m_0(x)$. Hence, the fast system (10) can be written as:

$$\begin{aligned} \frac{dx}{d\tau} &= \varepsilon f_1(x, z, \varepsilon) \\ \frac{dz}{d\tau} &= g_1(x, z, \varepsilon), \end{aligned} \tag{11}$$

where

$$\begin{aligned}
 f_1(x, z, \varepsilon) &= f_0(x, z + m_0(x), \varepsilon), \\
 g_1(x, z, \varepsilon) &= g_0(x, z + m_0(x), \varepsilon) - \varepsilon \left[\frac{\partial}{\partial x} m_0(x) \right] f_1(x, z, \varepsilon).
 \end{aligned}$$

When $\varepsilon = 0$, the system (11) becomes

$$\frac{dz}{d\tau} = g_1(x, z, 0), \quad x(\tau) \equiv x_0 \in W_x. \tag{12}$$

Verifying the stability of the fast system is a standard and an intuitive requirement for singular perturbation methods as in the original Tikhonov theorem [1]. Hence we need to check that:

1. the equilibrium $z = 0$ of (12) is globally asymptotically stable on $\{z \mid z + m_0(x_0) \in \mathbb{R}_+^m\}$ for all $x_0 \in \mathbb{R}_n^+$.
2. all eigenvalues of the Jacobian $\frac{\partial}{\partial y} g_0(x, m_0(x), 0)$ have negative real parts for every $x \in \mathbb{R}_n^+$.

We study this in the next section.

4 Analysis of the Core EMT/MET Network

In our analysis in the Sect. 2 we have considered fast and slow reactions separately. We will use now the notation from Sect. 3. Let $x(t) = [\mu_{34}, s, \mu_{200}, z]^T(t)$ be the state of the slow system, and let $y(t) = [d_s^T, d_z^T, d_{34}^T, d_{200}^T, m_s^T, m_z^T](t)$ be the state of the fast system. Hence, the full system can be written in the form (9). We denote the reduced system (7) as $\dot{x} = G(x)$.

4.1 Stability of the Fast Dynamics

In this subsection, we want to study global asymptotic stability of the fast system, and to show that its Jacobian is Hurwitz.

For a fixed slow variable x , the fast dynamics of ZEB1, SNAIL, miRNA-34 , miRNA-200 are decoupled from each other as can be seen from the CRNs introduced before. Furthermore, all the systems are linear. Let us analyze ZEB1 as an example. From (1),(3) we get

$$\begin{bmatrix} \dot{d}_z \\ \dot{m}_z \end{bmatrix} = \begin{bmatrix} P_z(s, z) & 0 \\ B_z & Q_z(s, z) \end{bmatrix} \begin{bmatrix} d_z \\ m_z \end{bmatrix},$$

where $B_z := [O_{3 \times 3}, b_z]$ and $O_{3 \times 3} \in \mathbb{R}^{3 \times 3}$ is a zero matrix. The dynamics of d_z are decoupled from m_z . Since P_z is an irreducible Metzler matrix with principal

eigenvalue of 0, Perron-Frobenius Theorem [29] implies that all other eigenvalues are strictly negative. In fact, they can be computed as $\{-\alpha_s s^2 - \alpha_{-s}, -\alpha_z z^2 - \alpha_{-z}, -\alpha_s s^2 - \alpha_{-s} - \alpha_z z^2 - \alpha_{-z}\}$. Since the total number of promoters is conserved, we can reduce the dimension of the ODE by one. It follows that the reduced promoter dynamics are globally asymptotically stable. The same argument applies to the promoters of Z , μ_{34}, μ_{200} .

We turn to the dynamics of m_z . Due to the block triangular structure we only need to study the eigenvalues of $Q_z(s, z)$. This matrix is also Metzler and the principal eigenvalue can be shown to be negative. Hence, it is Hurwitz. This can be proven alternatively with a linear Lyapunov function $V(m_z) = \mathbf{1}^T m_z$. The time-derivative is $\dot{V}(m_z) = -[\beta_z \beta_{z1} \beta_{z2}]m_z < 0$ for all $m_z \neq 0$. The same argument applies to the translational dynamic of SNAIL. Hence, the fast system is globally asymptotically stable and its Jacobian is Hurwitz.

4.2 Monotonicity of the Reduced System

We study the reduced system (7). We will utilize Kamke’s conditions as in Theorem 2. Hence, we compute the Jacobian J for the slow system (7):

$$J = \begin{bmatrix} -\beta_{-34} - \frac{2E_{34}A_s A_z \beta_{34}s}{(s^2 + A_s)^2(z^2 + A_z)} & 0 & -\frac{2E_{34}A_s A_z \beta_{34}z}{(s^2 + A_s)^2(z^2 + A_z)} \\ J_{21} & J_{22} & 0 \\ 0 & -\frac{2E_{200}A_s A_z \beta_{200}s}{(s^2 + A_s)^2(z^2 + A_z)} - \beta_{-200} & -\frac{2E_{200}A_s A_z \beta_{200}z}{(s^2 + A_s)^2(z^2 + A_z)} \\ 0 & J_{42} & J_{43} & J_{44} \end{bmatrix},$$

where

$$J_{21} = -\frac{h_1(\mu_{34})E_s A_s \gamma_s}{(s^2 + A_s)(\beta_{s2}c_s^2 \mu_{34}^2 + e_{s3}c_s \mu_{34} + e_{s1}\beta_s)^2},$$

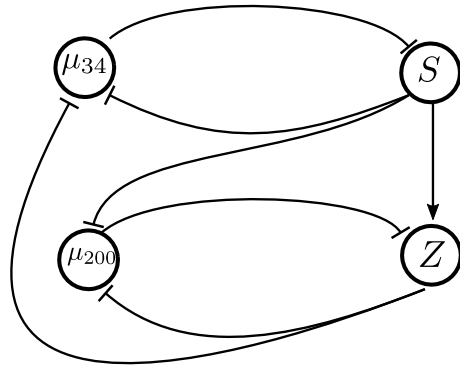
where $h(\mu_{34}) := c_s(\beta_{s2} + c_{-s})^2(\beta_{s1} + c_{-s})(\beta_{s1}k_s - \beta_s k_{s1}) + 2c_s^2 \mu_{34}(k_s \beta_{s2} - \beta_s k_{s2})(c_{-s} + \beta_{s1})(c_{-s} + \beta_{s2}) + c_s^3 \mu_{34}^2(\beta_{s2}(\beta_{s2}k_s - \beta_s k_{s2}) + (\beta_{s2} + c_{-s})(\beta_{s2}k_{s1} - \beta_{s1}k_{s2}))$. Note that $h(\mu_{200}) > 0$ for all parameters if (2), (6) are satisfied. Hence $J_{21} < 0$. Similarly, we can also calculate J_{43} to show that it is negative whenever (2), (6) are satisfied.

Finally, using (4) we compute J_{42} as:

$$J_{42} = \frac{2E_Z \gamma_z A_s s z^2}{(s^2 + A_s)^2(z^2 + A_z)} \frac{k_z(\beta_{z2}c_z \mu_{200} + e_{z1}) + k_{z1}(e_{z2}\mu_{200}) + k_{z2}c_z^2 \mu_{200}^2}{\beta_{z2}c_z^2 \mu_{200}^2 + e_{z3}c_z \mu_{200} + e_{z1}\beta_z} > 0.$$

Remark 1 Conditions (2), (6) are stronger than we need. In fact, it can be seen from the expression of h above that it is sufficient to have the protein production ratio of the raw mRNA greater than the corresponding one for the miRNA-RNA complex.

Fig. 2 Graph of the reduced circuit. In the terminology of Theorem 3, “→” has a + sign, and “-” has a - sign



The signs of J_{22}, J_{44} do not play a role since they correspond to self-loops. Therefore, we have the following sign pattern for the Jacobian under the conditions (2),(6):

$$\begin{bmatrix} * & - & 0 & - \\ - & * & 0 & 0 \\ 0 & - & * & - \\ 0 & + & - & * \end{bmatrix},$$

which can be represented via the graph depicted in Fig. 2. It can be easily seen that every loop is positive. Hence, by Theorem 3 there exists $\sigma \in \{\pm 1\}^N$ such that $\Sigma J \Sigma$ is Metzler. Therefore, by Theorem 2 we get that the reduced system (7) is monotone on $W = \mathbb{R}_+^4$. Monotonicity holds with respect to the orthant cone specified by $\sigma = [-1, 1, -1, 1]^T$ (see Sect. 3.2).

It can be verified that the Jacobian is irreducible on the open positive orthant \mathbb{R}_+^n . Hence, by Corollary 1 the reduced system is strongly monotone on \mathbb{R}_+^n .

Fix an initial condition $x_0 \in \mathbb{R}_+^n$, and let $x(t) := \varphi(t; x_0)$ be the corresponding solution of the slow system. In order to infer generic convergence to the set of equilibria, we need the strong order preserving property to hold on the ω -limit set of the solution. Hence, we need to show that $\omega(\varphi(t; x_0)) \cap \partial \mathbb{R}_+^n = \emptyset$. In other words, we need to show that the slow system is *persistent*. This can be shown as follows. Recall that the slow system (7) is given as $\dot{x} = G(x)$. It can be verified from (7) that $\forall i, G_i(z) > 0$ whenever $z_i = 0$. Hence, there exists $\eta > 0$ such that $\dot{x}_i(t) > 0$ whenever $x_i(t) \in [0, \eta)$. This implies that the boundary has no equilibria and is repelling (i.e., the vector field is pointing away from the boundary). We proceed by seeking contradiction. W.l.o.g, assume that $\exists x^* \in \omega(\varphi(t; x_0)) \cap \partial \mathbb{R}_+^n$ with $x_i^* = 0$ (the i -coordinate). Hence, there exists a sequence $\{t_k\}_{k=1}^\infty$ such that $x_i(t_k) > 0$ and $x_i(t_k) \rightarrow x_i^* = 0$. Therefore, for each right neighborhood \mathcal{N} of 0 there exists t^* for which $x_i(t^*) \in \mathcal{N}$ and $\dot{x}_i(t^*) < 0$, which is a contradiction. Hence, the reduced system is persistent.

Therefore, we state the following theorem.

Theorem 4 Consider the reduced system (7). Let $W^c \subset \mathbb{R}_+^4$ be the subset of points whose forward orbit has a compact closure in \mathbb{R}_+^4 . Then, the forward trajectory starting from almost every point in W^c converges to an equilibrium.

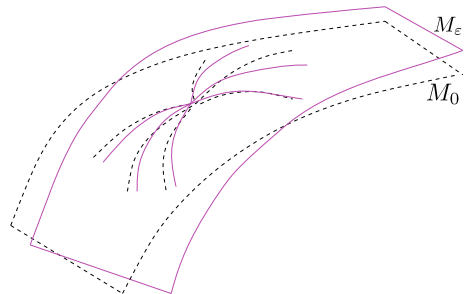
4.3 Remarks on Generic Convergence for Singularly Perturbed Monotone Systems

An interesting general question for singularly perturbed systems is as follows. Assuming that the slow system is strongly monotone, *does the full system obey generic convergence properties?* In other words, suppose that the flow ψ_t^0 of the slow system (set $\varepsilon = 0$ in (9)):

$$\frac{dx}{dt} = f_0(x, m_0(x), 0) \tag{13}$$

has strong monotonicity properties that guarantee almost-global convergence, meaning convergence to equilibria for all initial states except for those states in a set of measure zero (or, in a topological formulation, a nowhere dense set), but that the complete system is not monotone (so that no such theorem can be applied to it). Still, one may expect that the almost-convergence result can be lifted to the full system, at least for small $\varepsilon > 0$. An obstruction to this argument is that there is no *a priori* reason for the exceptional set to have a pre-image which has zero measure (or is nowhere dense). Nonetheless, a positive result along these lines was developed in [22] via the use of geometric invariant manifold theory to examine the fibration structure and utilize an asymptotic phase property [30–32]. Figure 3 illustrates the idea. The ODE restricted to the invariant manifold M_ε can be seen as a regular perturbation of the slow ($\varepsilon=0$) ODE. In [13], it has been noted that a C^1 regular perturbation of a flow with positive derivatives inherits the generic convergence properties. So, solutions in the manifold will generally be well-behaved, and asymptotic phase means that trajectories close to M_ε can be approximated by solutions in M_ε , and hence they also approach the set of equilibria if trajectories on M_ε do. We refer the reader to [22] for

Fig. 3 Illustration of the manifolds M_0 and M_ε . The figure shows two key properties of M_ε . First, M_ε is close to M_0 . Second, the trajectories on M_ε converge to steady-states if those on M_0 do



a technical discussion. In principle, this approach could be applied to systems such as ours, to conclude almost global convergence, under mild technical conditions on domains of validity for equation. We omit the details of the application here.

5 Conclusions

We have conducted a model reduction analysis for the EMT/MET system. Bounded trajectories of the reduced EMT/MET system generically converge to steady states, assuming sufficiently fast promoter and mRNA kinetics. Therefore, such a model cannot admit oscillations nor chaotic behavior. There are many further directions for research, which are being pursued by the authors, including the generalization to the case of miRNA binding to more than two sites on the mRNA molecule.

Acknowledgements We thank the reviewers of this manuscript for the meticulous reading and the constructive remarks. This research was supported by NSF grants 1716623 and 1849588.

References

1. Tikhonov, A.N.: Systems of differential equations containing small parameters in the derivatives. *Matematicheskii sbornik* **73**(3), 575–586 (1952)
2. Kokotovic, P., Khalil, H.K., O’Reilly, J.: *Singular perturbation methods in control: analysis and design*. SIAM (1999)
3. Michaelis, L., Menten, M.L.: Die kinetik der invertinwirkung. *Biochem Z* **49**, 333–369 (1913)
4. Gunawardena, J.: Time-scale separation–Michaelis and Menten’s old idea, still bearing fruit. *FEBS J.* **281**(2), 473–488 (2014)
5. Alon, U.: *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman and Hall/CRC (2006)
6. Thiery, J.P., Acloque, H., Huang, R.Y.J., Nieto, M.A.: Epithelial–mesenchymal transitions in development and disease. *Cell* **139**(5), 871–890 (2009)
7. Lambert, A.W., Pattabiraman, D.R., Weinberg, R.A.: Emerging biological principles of metastasis. *Cell* **168**(4), 670–691 (2017)
8. De Craene, B., Berx, G.: Regulatory networks defining EMT during cancer initiation and progression. *Nat. Rev. Cancer* **13**(2), 97 (2013)
9. Lamouille, J., Xu, S., Derynck, R.: Molecular mechanisms of epithelial–mesenchymal transition. *Nat. Rev. Mol. Cell Biol.* **15**(3), 178 (2014)
10. Lu, M., Jolly, M.K., Levine, H., Onuchic, J.N., Ben-Jacob, E.: MicroRNA-based regulation of epithelial–hybrid–mesenchymal fate determination. *Proc. Nat. Acad. Sci.* **110**(45), 18144–18149 (2013)
11. Kolch, W., Halasz, M., Granovskaya, M., Kholodenko, B.N.: The dynamic control of signal transduction networks in cancer cells. *Nat. Rev. Cancer* **15**(9), 515 (2015)
12. Hirsch, M.: Differential equations and convergence almost everywhere in strongly monotone flows. *Contemp. Math.* **17**, 267–285 (1983)
13. Hirsch, M.: Systems of differential equations that are competitive or cooperative II: convergence almost everywhere. *SIAM J. Math. Anal.* **16**, 423–439 (1985)
14. Hirsch, M.: Stability and convergence in strongly monotone dynamical systems. *J. Reine Angew. Math.* **383**, 1–53 (1988)

15. Hirsch, M., Smith, H.L.: Monotone dynamical systems. Handbook of Differential Equations, Ordinary Differential Equations, vol. 2. Elsevier, Amsterdam (2005)
16. Smith, H.L.: Monotone dynamical systems: an introduction to the theory of competitive and cooperative systems. Mathematical Surveys and Monographs, vol. 41. AMS, Providence, RI (1995)
17. Angeli, D., Ferrell, J.E., Sontag, E.D.: Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems. Proc. Nat. Acad. Sci. **101**(7), 1822–1827 (2004)
18. Sontag, E.D.: Monotone and near-monotone biochemical networks. Syst. Synth. Biol. **1**(2), 59–87 (2007)
19. Angeli, D., De Leenheer, P., Sontag, E.D.: Graph-theoretic characterizations of monotonicity of chemical networks in reaction coordinates. J. Math. Biol. **61**(4), 581–616 (2010)
20. Sharova, L.V., Sharov, A.A., Nedorezov, T., Piao, Y., Shaik, N., Ko, M.S.H.: Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. DNA Res. **16**(1), 45–58 (2008)
21. Guo, Y., Liu, J., Elfenbein, S.J., Ma, Y., Zhong, M., Qiu, C., Ding, Y., Lu, J.: Characterization of the mammalian miRNA turnover landscape. Nucl. Acids Res. **43**(4), 2326–2341 (2015)
22. Wang, L., Sontag, E.D.: Singularly perturbed monotone systems and an application to double phosphorylation cycles. J. Nonlinear Sci. (2008). <https://doi.org/10.1007/s00332-008-9021-2>
23. Wang, L., Sontag, E.D.: Singularly perturbed monotone systems and an application to double phosphorylation cycles. J. Nonlinear Sci. **18**(5), 527–550 (2008)
24. Lu, M., Jolly, M.K., Gomoto, R., Huang, B., Onuchic, J., Ben-Jacob, E.: Tristability in cancer-associated microRNA-TF chimera toggle switch. J. Phys. Chem. B **117**(42), 13164–13174 (2013)
25. Érdi, P., Tóth, J.: Mathematical Models of Chemical Reactions: Theory and Applications of Deterministic and Stochastic Models. Manchester University Press (1989)
26. Ali Al-Radhawi, M., Angeli, D., Sontag, E.D.: A computational framework for a Lyapunov-enabled analysis of biochemical reaction networks. PLoS Comput. Biol. **16**(2), e1007681 (2020)
27. Ali Al-Radhawi, M., Del Vecchio, D., Sontag, E.D.: Multi-modality in gene regulatory networks with slow promoter kinetics. PLoS Comput. Biol. **15**(2), e1006784 (2019)
28. Ender, C., Meister, G.: Argonaute proteins at a glance. J. Cell Sci. **123**(11), 1819–1823 (2010)
29. Berman, A., Plemmons, R.J.: Nonnegative Matrices in the Mathematical Sciences. SIAM (1994)
30. Fenichel, N.: Geometric singular perturbation theory for ordinary differential equations. J. Diff. Equ. **31**, 53–98 (1979)
31. Jones, C.K.R.T.: Geometric singular perturbation theory. In: Dynamical Systems (Montecatini Terme). Lecture Notes in Mathematics, vol. 1609. Springer, Berlin (1994)
32. Nipp, K.: Smooth attractive invariant manifolds of singularly perturbed ODEs. Seminar für Angewandte Mathematik. Eidgenössische Technische Hochschule, In Research Report (1992)