LNCIS

Lecture Notes in Control and Information Sciences

398

Jan C. Willems
Shinji Hara
Yoshito Ohta
Hisaya Fujioka (Eds.)

Perspectives in Mathematical System Theory, Control, and Signal Processing

A Festschrift in Honor of Yutaka Yamamoto on the Occasion of his 60th Birthday

Springer

Lecture Notes in Control and Information Sciences 398

Jan C. Willems, Shinji Hara, Yoshito Ohta,
and Hisaya Fujioka (Eds.)

Perspectives in Mathematical System Theory, Control, and Signal Processing

A Festschrift in Honor of Yutaka Yamamoto on the Occasion of his 60th Birthday

Springer

Editors

Jan C. Willems

ESAT, K.U.Leuven
Kasteelpark Arenberg 10
3001 Leuven, Belgium
E-mail: Jan.Willems@esat.kuleuven.be

Shinji Hara

Department of Information
Physics and Computing
Graduate School of Information
Science and Technology
The University of Tokyo
7-3-1, Hongo, Bunkyo-ku
Tokyo 113-8656, Japan
E-mail: Shinji_Hara@ipc.i.u-tokyo.ac.jp

Yoshito Ohta

Control Systems Theory Group
Department of Applied Mathematics
and Physics
Graduate School of Informatics
Kyoto University
Yoshida-honmachi, Kyoto 606-8501, Japan
E-mail: Yoshito_ohta@i.kyoto-u.ac.jp

Hisaya Fujioka

Department of Applied Analysis and
Complex Dynamical Systems
Kyoto University
Kyoto 606-8501, Japan
E-mail: fujioka@i.kyoto-u.ac.jp

ISBN 978-3-540-93917-7 e-ISBN 978-3-540-93918-4

DOI 10.1007/978-3-540-93918-4

Lecture Notes in Control and Information Sciences ISSN 0170-8643

Library of Congress Control Number: 2009943558

Typeset & Cover Design: Scientific Publishing Services Pvt. Ltd., Chennai, India.

Printed in acid-free paper

5 4 3 2 1 0

springer.com

Dedicated to Yutaka Yamamoto
on the occation of his sixtieth birthday

Preface

This Festschrift, published on the occasion of the sixtieth birthday of Yutaka Yamamoto (‘YY’ as he is occasionally casually referred to), contains a collection of articles by friends, colleagues, and former Ph.D. students of YY. They are a tribute to his friendship and his scientific vision and oeuvre, which has been a source of inspiration to the authors.

Yutaka Yamamoto was born in Kyoto, Japan, on March 29, 1950. He studied applied mathematics and general engineering science at the Department of Applied Mathematics and Physics of Kyoto University, obtaining the B.S. and M.Sc. degrees in 1972 and 1974. His M.Sc. work was done under the supervision of Professor Yoshikazu Sawaragi. In 1974, he went to the Center for Mathematical System Theory of the University of Florida in Gainesville. He obtained the M.Sc. and Ph.D. degrees, both in Mathematics, in 1976 and 1978, under the direction of Professor Rudolf Kalman.

His stay at the Center for Mathematical System Theory and the influence of Rudy Kalman had a decisive influence on YY’s research outlook. He became deeply convinced of the importance of rigorous thinking, the relevance of clear problem formulation, and the value of questioning hypotheses. Kalman’s paper ‘On the general theory of control systems’ (1st IFAC Congress, Moscow, 1960) which YY read when he was a junior student in Kyoto opened his eyes to the relevance of mathematics in systems and control theory. Trained as a control engineer, his view of the field had been somewhat limited to classical control theory, and he became fascinated with the potential of mathematical concepts in engineering. This motivated him to study the theoretical aspects of systems and control, and it brought him eventually to the Center for Mathematical System Theory in Florida and to the style of doing research that was in vogue at the Center. He also learned to appreciate the importance that an international visitor program can have in a scientific environment. He later recreated many of these aspects in his own research group in Kyoto.

YY’s Ph.D. dissertation, entitled *Realization Theory of Infinite-Dimensional Linear Systems*, deals with the construction of state models for infinite-dimensional systems. In particular, the existence and uniqueness of canonical realizations for such systems is proven. This topic remained one of the central themes of his research throughout his career.

Through his Ph.D. work, YY realized that the construction of a canonical realization for infinite-dimensional systems is intrinsically difficult. While a canonical model can be realized through a closed subspace of the output function space, it is difficult to go beyond there. He discovered that part of the difficulty lies in the fact that the state is in general not determined by the output data over a finite-time interval. This led him to the concept of *pseudorational* impulse responses. This class of systems allows a fractional representation of the impulse response over the ring of distributions with compact support. The compactness of the support makes it possible to construct a state space from bounded-time data. This in turn gives rise to characterizations of spectra, reachability criteria, stability properties, etc.

After completing his Ph.D., Yutaka Yamamoto returned to Kyoto University, where he has had a position since. He rose through the ranks: assistant professor from 1978 to 1987, associate professor from 1987 to 1997, and professor since 1997. His present affiliation is Professor in the Department of Applied Analysis and Complex Dynamical Systems, the Graduate School of Informatics, Kyoto University.

The class of pseudorational impulse responses provides an excellent platform for studying a class of learning control systems, called repetitive control. The stability criteria for pseudorational class played a central role in YY's proof of the internal model principle for such systems.

Through the study of infinite-dimensional systems, YY became interested in sampled-data control systems. When he took interest in this area, the subject of sampled-data control had been quite messy. The setting of the problem involves two time sets: a continuous one and a discrete one. Due to this hybrid nature, formulas were complicated and not very transparent. There are often ripples between sampling points and intersample information may be lost. In a CDC paper in 1990, YY introduced a function space technique that allows one to model sampled-data systems with a single discrete time set, while describing the intersample behavior through a function space setting. This technique, now called *lifting*, has become standard in the study of sampled-data systems, and makes the whole theory completely transparent. It has become possible to optimize the intersample behavior with a digital controller.

While studying digital control systems, YY observed that the same problem is encountered in digital signal processing, where one also needs to reconstruct intersample signals from the sampled data. In order to do this, usually Shannon's sampling theorem is employed and the intersample signal is recovered with as low frequency components as possible. YY noticed that by assuming a model for the signal generator, the ideas of sampled-data control theory can greatly improve the high-frequency performance. This contribution led to three patents, one for digital/analog converters, one (jointly with Masaaki Nagahara) for a sample rate converter, and yet another for compression audio (jointly with Sanyo Electric Corporation). These ideas were implemented in the production of sound-processing Large Scale Integrated circuits by Sanyo. These LSIs expand the limited frequency range of compressed audio signals up to the range of compact discs and have been used in mobile phones, MP3 players, digital voice recorders, etc. The net production as of June 2009 reached 9 million chips.

YY's research laboratory at Kyoto University has become one of the leading centers in the field of Systems and Control in Japan. It hosts an active program of international visitors and seminars. These follow the style he met as a graduate student in the Center for Mathematical System Theory in Florida. Seminars are not passively listened to, but involve active audience participation and questioning, to the benefit of all. Well-known to visitors are YY's halting remarks *'just wait'* . This may refer to the fact that the visitor neglected to leave the visitor's uncivilized shoes in the corridor before entering YY's office, or to stop the visitor from erasing a sloppy formula from YY's whiteboard before the high tech mechanism mounted on the board could record a copy of the scribblings. But mostly the *'just wait'* means that a technical point was not perfectly clear. YY's visitors are always treated to a demonstration of the superior high-frequency performance of the 'YY filter' in the audio room of his laboratory. Even if the visitor's hearing is not up to the finesses of the high frequency drop-off, there is always Maria Callas or Wilhelm Furtwängler to compensate these inadequacies, not to mention an occasional glass of *Brunello di Montalcino* that makes the YY filter experience into an unforgettable one.

Yutaka Yamamoto has been deeply committed to improve the quality of the research efforts of Japan in the field of Systems and Control. He has written tutorial essays and given talks at Japanese control conferences, guiding young researchers in giving presentations and writing articles in English. YY has been very active in professional organizations in Japan and internationally. He has been chair of the MTNS steering committee, and the General Chair of the MTNS held in Kyoto in 2006. He has been vice president of the IEEE Control Systems Society from 2005 to 2008, and is an active participant in the board of SICE (the Society for Instrument and Control Engineers) and ISCIE (the Institute of Systems, Control and Information Engineers), two Japanese engineering societies. He was on the board of ISCIE from 2006, with a term as president, until 2009.

YY has published or edited 5 books, and wrote close to 200 journal or conference papers. His work has been honored by awards and best paper prizes. He received the prestigious George S. Axelby outstanding paper award in 1996 for the paper 'A Function Space Approach to Sampled-Data Control Systems and Tracking Problems', published in the *IEEE Transactions on Automatic Control*, volume AC-39, pages 703–712 (1994), and the best paper prize from SICE in 1987, 1990, and 1997. In 2007, he received the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology. In 2009, he also received the ISCIE Best Industrial Application Award (jointly with four other authors in Sanyo Electric Corporation) for his work on high frequency compensation for compressed digital audio using sampled-data control.

Yutaka Yamamoto married Mamiko in 1979 in Kyoto. They live in Kyoto, and have two children, Sho and Kaoru.

The editors are grateful to all authors for their efforts in writing their article, and for meeting the time schedule. We thank the editorial staff of Springer Verlag for accepting this collection of articles as a volume for the Lecture Notes in Control and Information Sciences, and for making this book available in time for the Symposium on *Systems, Control, and Signal Processing*.

This volume contains articles by invited speakers for the symposium to celebrate the sixtieth birthday of Yutaka Yamamoto in Kyoto University on March 29–31, 2010.

November 2009

The editors

Jan C. Willems (K. U. Leuven)
Shinji Hara (University of Tokyo)
Yoshito Ohta (Kyoto University)
Hisaya Fujioka (Kyoto University)

Selected Publications of Yutaka Yamamoto

1. Yamamoto, Y.: Some remarks on reachability of infinite-dimensional linear systems. Journal of Mathematical Analysis and Applications 74, 568–577 (1980)
2. Yamamoto, Y.: Realization theory of infinite-dimensional linear systems, Part I. Mathematical Systems Theory 15, 55–77 (1981)
3. Yamamoto, Y.: Module structure of constant linear systems and its applications to controllability. Journal of Mathematical Analysis and Applications 83, 411–437 (1981)
4. Yamamoto, Y.: Realization theory of infinite dimensional linear systems, Part II. Mathematical Systems Theory 15, 169–190 (1982)
5. Yamamoto, Y., Ueshima, S.: Canonical realizations of time delay systems. Systems and Control 27, 517–522 (1983)
6. Yamamoto, Y.: A note on linear input/output maps of bounded type. IEEE Transactions on Automatic Control 29, 733–734 (1984)
7. Yamamoto, Y., Ueshima, S., Morishita, T.: Input/output maps of bounded type and their approximate realizations. Systems and Control 28, 536–543 (1984)
8. Yamamoto, Y.: On infinite-dimensional pole-zero cancellations. International Journal of Control 41, 1621–1626 (1985)
9. Yamamoto, Y.: Realization of pseudo-rational input/output maps and its spectral properties. Memories of the Faculty of Engineering 47, 221–239 (1985)
10. Yamamoto, Y., Ueshima, S.: A new model for neutral delay-differential systems. International Journal of Control 43, 465–472 (1986)
11. Ueshima, S., Yamamoto, Y.: Approximate realization of time-delay systems. Transactions of SICE 22, 375–382 (1986)
12. Sugimoto, K., Yamamoto, Y.: On indices of linear systems. Transactions of SICE 22, 467–469 (1986)
13. Yamamoto, Y., Hara, S.: The internal model principle and stabilizability of repetitive control systems. Transactions of SICE 22, 830–834 (1986)
14. Fukushima, M., Yamamoto, Y.: A second-order algorithm for continuous-time nonlinear optimal control problems. IEEE Transactions on Automatic Control 31, 673–676 (1986)
15. Hara, S., Yamamoto, Y.: Stability of multivariable repetitive control systems — stability condition and characterization of stabilizing controllers. Transactions of SICE 22, 1256–1261 (1986)
16. Sugimoto, K., Yamamoto, Y.: New solution to the inverse regulator problem by the polynomial matrix method. International Journal of Control 45, 1627–1640 (1987)

17. Sugimoto, K., Yamamoto, Y.: LQ regulators for time delay systems: convergence of approximate solutions in the frequency domain. Transactions of SICE 23, 590–596 (1987)
18. Yamamoto, Y., Hara, S.: Relationships between internal and external stability for infinite-dimensional systems with applications to a servo problem. In: Proceedings of the 26th IEEE Conference on Decision and Control, pp. 1558–1563 (1987)
19. Hara, S., Yamamoto, Y., Omata, T., Nakano, M.: Repetitive control system — a new-type servo system. IEEE Transactions on Automatic Control 33, 659–668 (1988)
20. Sugimoto, K., Yamamoto, Y.: Solution to the Inverse regulator problem for discrete-time systems. International Journal of Control 48, 1285–1300 (1988)
21. Yamamoto, Y., Hara, S.: Relationships between internal and external stability for infinite-dimensional systems with applications to a servo problem. IEEE Transactions on Automatic Control 33, 1044–1052 (1988)
22. Yamamoto, Y.: Pseudo-rational input/output maps and their realizations: a fractional representation approach to infinite-dimensional systems. SIAM Journal on Control and Optimization 26, 1415–1430 (1988)
23. Fujinaka, T., Sugimoto, K., Yamamoto, Y., Katayama, T.: On pole-assignable region of discrete-time LQ regulators. Transactions of SICE 24, 1253–1259 (1988)
24. Yamamoto, Y.: On the convergence condition of learning control for linear systems. Transactions of SICE 24, 1331–1333 (1988)
25. Sugimoto, K., Yamamoto, Y., Fujinaka, T.: Pole-assignable region via successive application of LQ regulators. Transactions of SICE 25, 123–125 (1989)
26. Yamamoto, Y.: Reachability of a class of infinite-dimensional linear systems: an external approach with applications to general neutral systems. SIAM Journal on Control and Optimization 27, 217–234 (1989)
27. Sugimoto, K., Yamamoto, Y.: A computational method for doubly coprime factorization via polynomial matrices. Transactions of ISCIE 2, 241–246 (1989)
28. Sontag, E.D., Yamamoto, Y.: On the existence of an approximately coprime factorization for retarded systems. Systems & Control Letters 13, 53–58 (1989)
29. Sugimoto, K., Yamamoto, Y.: On successive pole assignment by linear quadratic optimal feedbacks. Linear Algebra and its Applications 122-124, 697–723 (1989)
30. Sugimoto, K., Yamamoto, Y.: On the generalized robustness of optimality of linear quadratic regulators. International Journal of Control 51, 521–533 (1990)
31. Sugimoto, K., Yamamoto, Y.: A polynomial matrix method for computing stable rational doubly coprime factorizations. Systems & Control Letters 14, 267–273 (1990)
32. Yamamoto, Y.: New approach to sampled-data control systems — a function space method. In: Proceedings of the 29th IEEE Conference on Decision and Control, pp. 1882–1887 (1990)
33. Kato, Y., Yamamoto, Y.: On learning of neutral networks with feedback connections. Transactions of ISCIE 4, 369–374 (1991)
34. Antoulas, A.C., Matsuo, T., Yamamoto, Y.: Linear deterministic realization theory. In: Mathematical System Theory — The Influence of Kalman, R.E., pp. 191–212. Springer, Heidelberg (1991)
35. Yamamoto, Y.: Equivalence of internal and external stability for a class of distributed systems. Mathematics of Control, Signals, and Systems 4, 391–409 (1991)
36. Yamamoto, Y., Hara, S.: Internal and external stability and robust stability condition for a class of infinite-dimensional systems. Automatica 28, 81–93 (1992)
37. Nakatani, T., Yamamoto, Y., Matsumoto, Y.: On composite neural networks. Transactions of ISCIE 5, 349–356 (1992)
38. Yamamoto, Y.: Learning control and related problems in infinite-dimensional systems. In: Trentelman, H.L., Willems, J.C. (eds.) Essays on Control: Perspectives in the Theory and its Applications, pp. 191–222. Birkhäuser, Basel (1993)

39. Yamamoto, Y.: On the state space and frequency domain characterization of H^∞-norm of sampled-data systems. Systems & Control Letters 21, 163–172 (1993)
40. Yamamoto, Y.: A function space approach to sampled-data control systems and tracking problems. IEEE Transactions on Automatic Control 39, 703–712 (1994)
41. Yamamoto, Y., Araki, M.: Frequency responses for sampled-data systems — their equivalence and relationships. Linear Algebra and its Applications 205-206, 1319–1339 (1994)
42. Hayakawa, Y., Hara, S., Yamamoto, Y.: H_∞ type problem for sampled-data control systems — a solution via minimum energy characterization. IEEE Transactions on Automatic Control 39, 2278–2284 (1994)
43. Zwart, H., Yamamoto, Y., Gotoh, Y.: Stability is realization-dependent: some examples. Systems & Control Letters 24, 25–31 (1995)
44. Hirata, K., Yamamoto, Y., Katayama, T., Tannenbaum, A.R.: The equivalence among the solutions of the H^∞ optimal sensitivity computation problem. Transactions of SICE 31, 1954–1961 (1995)
45. Yamamoto, Y., Khargonekar, P.P.: Frequency response of sampled-data systems. IEEE Transactions on Automatic Control 41, 166–176 (1996)
46. Nishida, H., Matsumoto, Y., Yamamoto, Y.: On low-dimensional models of neural networks. Transactions of SICE 32, 379–388 (1996)
47. Hirata, K., Yamamoto, Y., Tannenbaum, A.R., Katayama, T.: New solution to the two block H^∞ problem for infinite-dimensional stable plants. Transactions of SICE 32, 1416–1424 (1996)
48. Yamamoto, Y., Hirata, K., Tannenbaum, A.: Some remarks on Hamiltonians and the infinite-dimensional one block H^∞ problem. Systems & Control Letters 29, 111–117 (1996)
49. Yamamoto, Y.: A retrospective view on sampled-data control systems. CWI Quarterly 9, 261–276 (1996)
50. Khargonekar, P.P., Yamamoto, Y.: Delayed signal reconstruction using sampled-data control. In: Proceedings of the 35th IEEE Conference on Decision and Control, pp. 1259–1263 (1996)
51. Hirata, K., Yamamoto, Y., Tannenbaum, A.R., Katayama, T.: Skew Toeplitz solution to the H^∞ problem for infinite-dimensional unstable plants. Transactions of SICE 33, 1066–1071 (1997)
52. Yamamoto, Y., Madievski, A.G., Anderson, B.D.O.: Computation and convergence of frequency response via fast sampling for sampled-data control systems. In: Proceedings of the 36th IEEE Conference on Decision and Control, pp. 2157–2162 (1997)
53. Yamamoto, Y., Fujioka, H., Khargonekar, P.P.: Signal reconstruction via sampled-data control with multirate filter banks. In: Proceedings of the 36th IEEE Conference on Decision and Control, pp. 3395–3400 (1997)
54. Ishii, H., Yamamoto, Y.: Periodic compensation for sampled-data H^∞ repetitive control. In: Proceedings of the 37th IEEE Conference on Decision and Control, pp. 331–336 (1998)
55. Ishii, H., Yamamoto, Y.: Sampled-data H^∞ and H^2/H^∞ design of multirate D/A converter. Transactions of ISCIE 11, 586–593 (1998)
56. Yamamoto, Y., Khargonekar, P.P.: From sampled-data control to signal processing. In: Learning, Control and Hybrid Systems. LNCIS, vol. 241, pp. 108–126. Springer, Heidelberg (1998)
57. Yamamoto, Y.: Coprimeness of factorizations over a ring of distributions. In: Blondel, V.D., et al. (eds.) Open Problems in Mathematical Systems and Control Theory, pp. 281–284. Springer, Heidelberg (1998)

58. Yamamoto, Y., Hara, S.: Performance lower bound for a sampled-data signal reconstruction. In: Blondel, V.D., et al. (eds.) Open Problems in Mathematical Systems and Control Theory, pp. 277–280. Springer, Heidelberg (1998)
59. Yamamoto, Y., Madievski, A.G., Anderson, B.D.O.: Approximation of frequency response for sampled-data control systems. Automatica 35, 729–734 (1999)
60. Yamamoto, Y.: Digital control. Wiley Encylopedia of Electrical and Electronics Engineering 5, 445–457 (1999)
61. Hirata, K., Yamamoto, Y., Tannenbaum, A.: Computation of the singular values of Toeplitz operators and the gap metric. Systems & Control Letters 36, 327–338 (1999)
62. Wakasa, Y., Yamamoto, Y.: Control system design considering a tradeoff between evaluated uncertainty ranges and control performance. Asian Journal of Control 1, 49–57 (1999)
63. Ishii, H., Yamamoto, Y., Francis, B.A.: Sample-rate conversion via sampled-data H^∞ control. In: Proceedings of the 38th IEEE Conference on Decision and Control, pp. 3440–3445 (1999)
64. Yamamoto, Y., Nagahara, M., Fujioka, H.: Multirate signal reconstruction and filter design via sampled-data H^∞ control. In: Proceedings of the 14th International Symposium on Mathematical Theory of Networks and Systems (MTNS 2000), Perpignan, France (2000)
65. Hirata, K., Yamamoto, Y., Tannenbaum, A.: A Hamiltonian-based solution to the two-block H^∞ problem for general plants in H^∞ and rational weights. Systems & Control Letters 40, 83–96 (2000)
66. Fujioka, H., Wakasa, Y., Yamamoto, Y.: Optimization in control system design. Journal of the Operations Research Society of Japan 43, 48–70 (2000)
67. Nagahara, M., Yamamoto, Y.: A new design for sample-rate converters. In: Proceedings of the 39th IEEE Conference on Decision and Control, pp. 4296–4301 (2000)
68. Nagahara, M., Yamamoto, Y.: Sampled-data H^∞ design of interpolators. Transactions of ISCIE 14, 483–489 (2001)
69. Wakasa, Y., Yasufuku, D., Nagahara, M., Yamamoto, Y.: Sampled-data design of interpolators using the cutting-plane method. Transactions of SICE 38, 462–468 (2002)
70. Yamamoto, Y., Nagahara, M.: Digital filter design via sampled-data control theory. In: Hashimoto, K., Oishi, Y., Yamamoto, Y. (eds.) Control and Modeling of Complex Systems: Cybernetics in the 21st Century. Birkhäuser, Boston (2002)
71. Nagahara, M., Yamamoto, Y.: Sampled-data H^∞ design for digital communication systems. Transactions of ISCIE 16, 38–43 (2003)
72. Fujimoto, K., Shimazu, M., Yamamoto, Y.: Decision support for Internet users. Transactions of the Japanese Society for Artificial Intelligence 18, 36–44 (2003)
73. Yamamoto, Y., Anderson, B.D.O., Nagahara, M., Koyanagi, Y.: Optimizing FIR approximation for discrete-time IIR filters. IEEE Signal Processing Letters 10, 273–276 (2003)
74. Wakasa, Y., Hikita, M., Yamamoto, Y.: LMS adaptive filter design using semidefinite programming and its applications to active noise control. Transactions of ISCIE 16, 461–467 (2003)
75. Yasufuku, D., Wakasa, Y., Yamamoto, Y.: Adaptive digital filtering based on a continuous-time performance index. Transactions of SICE 39, 569–574 (2003)
76. Kashima, K., Yamamoto, Y., Nagahara, M.: Optimal wavelet expansion via sampled-data control theory. IEEE Signal Processing Letters 11, 79–82 (2004)
77. Yamaguchi, T., Nishikawa, A., Miyazaki, F., Shimada, J., Katoh, D., Kawakami, M., Ikeda, F., Yamamoto, Y.: Real-time Endoscopic Image Overlay System for Intra-Operative Localization of Tumors in Deformative Organs. Transactions of the Japanese Society for Medical and Biomedical Engineering 42(4), 318–327 (2004)

78. Ashida, S., Kakemizu, H., Nagahara, M., Yamamoto, Y.: Sampled-data audio signal compression with Huffman coding. In: Proceedings of the SICE Annual Conference 2004, pp. 972–976 (2004)
79. Fujimoto, K., Yamamoto, Y.: On accuracy of knowledge acquisition for decision making processes acquiring subjective information on the Internet. Transactions of the Japanese Society for Artificial Intelligence 19, 571–579 (2004)
80. Yamamoto, Y.: A new approach to signal processing via sampled-data control theory. Australian Journal of Electrical & Electronics Engineering 2, 141–148 (2005)
81. Kashima, K., Yamamoto, Y.: A new characterization of invariant subspaces of H^2 and applications to the optimal sensitivity problem. Systems & Control Letters 54, 539–545 (2005)
82. Nagahara, M., Wada, T., Yamamoto, Y.: Design of oversampling delta-sigma DA converters via H-infinity optimization. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Toulouse, France, vol. III, pp. 612–615 (2006)
83. Kashima, K., Özbay, H., Yamamoto, Y.: A Hamiltonian-based solution to the mixed sensitivity optimization problem for stable pseudorational plants. Systems & Control Letters 54, 1063–1068 (2005)
84. Fujiyama, K., Kano, H., Iwasaki, N., Kaibe, R., Yamamoto, Y.: High frequency compensation for compressed digital audio using sampled-data control. Transactions of ISCIE 20, 31–38 (2007)
85. Fuhrmann, P.A., Rapisarda, P., Yamamoto, Y.: On the state of behaviors. Linear Algebra and its Applications 424, 570–614 (2007)
86. Willems, J.C., Yamamoto, Y.: Behaviors defined by rational functions. Linear Algebra and its Applications 425, 226–241 (2007)
87. Kashima, K., Yamamoto, Y.: System theory for numerical analysis. Automatica 43, 1156–1164 (2007)
88. Kashima, K., Özbay, H., Yamamoto, Y.: Parameterization of suboptimal solutions of the Nehari problem for infinite-dimensional systems. IEEE Transactions on Automatic Control 53, 2369–2374 (2007)
89. Yamamoto, Y., Nagahara, M.: Signal processing via digital control theory. Journal of the Institute of Image Information and Television Engineers (2007)
90. Yamamoto, Y.: New Development of Digital Signal Processing via Sampled-data Control Theory. In: Modeling, Estimation and Control — Festschrift in honor of Giorgio Picci's 65th birthday. LNCIS, vol. 364, pp. 345–355. Springer, Heidelberg (2007)
91. Yamamoto, Y.: Pseudorational impulse responses — algebraic system theory for distributed parameter systems. SICE Journal of Control, Measurement, and System Integration (SICE JCMSI) 1, 51–57 (2008)
92. Kashima, K., Yamamoto, Y.: Finite rank criteria for H^∞ control of infinite-dimensional systems. IEEE Transactions on Automatic Control 53, 881–893 (2008)
93. Kashima, K., Yamamoto, Y.: On standard H^∞ control problems for systems with infinitely many unstable poles. Systems & Control Letters 57, 309–314 (2008)
94. Willems, J.C., Yamamoto, Y.: Behaviors described by rational symbols and the parametrization of the stabilizing controllers. In: Blondel, V., Boyd, S.P., Kimura, H. (eds.) Recent Advances in Learning and Control — Festschrift in honor of M. Vidyasagar's 60th birthday. LNCIS, vol. 371, pp. 263–277 (2008)
95. Nagahara, M., Yamamoto, Y., Khargonekar, P.P.: Stability of signal reconstruction filters via cardinal exponential splines. In: Proceedings of the 17th IFAC World Congress, Seoul, Korea, pp. 1414–1419 (2008)

Contents

Part I: Circuit Theory

Part II: Control

Part III: Graphs and Networks

Part IV: Mathematical System Theory

Part V: Modeling

Part VI: Signal Processing

Part VII: System Identification

Old and New Directions of Research in System Theory

Rudolf Kalman

Dedicated to Professor Yutaka Yamamoto on his 60th birthday by his doctorfather.

Abstract. The problem of classical electrical network synthesis (flourished between 1920–1970) is subjected to scientific critique. Conclusions: the first attacks on the problem were frustrated and eventually defeated by a naive over-reliance on engineering/physical intuition and shoving the mathematical issues whenever possible under the rug; now, by concentrating on essential mathematics, much of it known since the 19th century, research will be revived with spectacular prospects of scientific progress.

1 Newton

The first big result in System Theory was Newton's gravitational law (1686). Why so? Because it linked (perhaps forever) two things which were until then, and still are now, considered to be objectively separate: the natural real world of physics and the world of mathematics. Newton's theorem: *"The empirical description of motions in the solar system (known as Kepler's Laws) is abstractly the same as the physical model based on the inverse-square gravitational law"*. Newton had been studying this problem in the early 1680's, purely from the mathematical point of view using as "data" Kepler's Laws that had by then been known for over fifty years. No one knows if the great man had fully understood that this result was the decisive first step in creating system theory. Note, however, that Book III of the Principia is titled "System of the World". Newton certainly knew that he was penetrating into the secrets of the real (physical) world by the inspiration and active help of mathematics. Recall the full title of the Principia and the famous manifesto *"Hypotheses non fingo"* (1713). I discuss the issues in more detail in [9].

2 Foster and Cauer

It took more than 250 years until the next step was attempted, by an American engineer at Bell Telephone Laboratories, Ronald M. Foster (1896–1998). I came to

Rudolf Kalman
Swiss Federal Institute of Technology (ETH-Zürich), Switzerland

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 3–13.
springerlink.com

regard his 1924 paper, “A reactance theorem” [4] as a contribution as decisive as Newton’s, but in Foster’s case linking the man-made real world of engineering with mathematics.

Foster’s great enabler was Oliver Heaviside (1850–1925), as a genius not inferior to Kepler, who showed, with great insight but less than requisite precision, that the impedance of an electric *RCL* network (resistors, capacitors and inductors) can be described, with very high accuracy, as $Z(s) = p(s)/q(s)$, where p, q are polynomials with real coefficients and s an algebraic indeterminate (the famous Heaviside “operator”). If you give s the value $i2\pi f$, $Z(s)$ is a complex number, the physical measure of the impedance of the network at the frequency f. This is a fact, known for about 100 years. It was also, for a long time, regarded as the best way of thinking about impedance, avoiding the crazy idea of the Heaviside “operator”.

Heaviside’s $Z(s)$ — just like Kepler’s laws — is a precise external description of a system, in this case of an electric network. Foster’s theorem, in its modern form, says, “If $Z(s)$, viewed as a function of the complex variable s, has a certain property then Z can be realized in the physical sense, namely as an electric network built from two kinds of components selected from among the classical three: R (resistors), C (capacitors) and L (inductors), in such a way that the impedance of the resulting network is exactly the given $Z(s)$.” Properties of Z dictate which one and only one of the three choices *CL*, *RC*, or *RL* is the appropriate one for the realization. Foster’s technique of proof, expanding $Z(s)$ into partial fractions, leads to networks generated by series (or parallel) cascading of parallel (or series) pairs of the two components *CL*, *RC*, or *RL*. This results in two families of “Foster canonical forms” depending on whether one works with $Z(s)$ = “impedance” or its reciprocal $Y(s)$ = “admittance”. Many more details are given in [8, Chapters 3 and 4].

In the interest of historical correctness it should be noted that [4] treated only the *CL* (lossless) case; it was immediately seen, however, that the *RC* and *RL* cases followed trivially from it. Further consideration showed that *CL* was a borderline case of marginal interest. It need not be discussed here.

Foster’s proof was easy and pretty, so much so that a bright young student, Wilhelm Cauer (1900–1945), having just been awarded his diploma at Technische Hochschule Charlottenburg in Berlin, immediately noticed another easy way to realization: expanding $Z(s)$ as a continued fraction. The resulting networks were of the ladder type, again in two variants, the “Cauer canonical forms”. When Cauer finished there was a big surprise: although the Foster and Cauer networks bore no resemblance to one another they were equally effective: given a $Z(s)$ either they could both realize $Z(s)$ or neither could do the job.

Thus the field of “passive network synthesis” was born, a common offspring of two legitimate fathers. Again see [8, Chapters 3 and 4].

What, if anything, was wrong with all this?

Neither Foster nor Cauer gave an explicit realizability condition directly computable from the parameters (coefficients of p and q) of Z. This is clearly an unsatisfactory aspect of the theorems of Foster and Cauer and calls for an explanation.

In the 1920's, and until much later, electric *RCL* networks were always passive, meaning the components R, C, L had positive real values. This positivity condition on the components necessarily induces a certain condition on $Z(s)$. In 1924 such a positivity condition on Z was completely unknown, hence also unknown to Foster. But Foster noticed that the coefficients of the partial fraction expansion of $Z(s)$ could be identified with the component values of certain networks (later labeled as Foster "canonical" networks) — so component positivity had to be equivalent to Z positivity defined as the positivity of the coefficients created by the computation of the partial fraction expansion of Z. Cauer's idea resulted in another definition of Z positivity, namely the positivity of coefficients created by the computation of the partial fraction expansion of Z. The explanation for the surprise: the two definitions of Z positivity were proved to be equivalent via the idea of network realization (at that time called "synthesis"). Alas, these equivalent conditions were both algorithmic. Not algebraic, as Heaviside would have liked it.

Background Remark. Observe that partial fraction expansions (with real coefficients) exist whenever p or q have real roots while continued fraction expansions (with real coefficients) exist always. Hence the Foster-Cauer theorems, from the very beginning, could have been stated in a more general form: "Any $Z(s)$ has a quasi-realization (component values nonzero but not necessarily positive) by one of the two *RC* or *RL* network types provided a (Foster) partial fraction or a (Cauer) continued fraction expansion exists with *real* coefficients; when all such coefficients are positive then $Z(s)$ has a true (component values all positive) realization." Even more trivially, any expansion of Z with an "interpretation" as a network induces a quasi-realization; and if all coefficients of the expansion are positive, a (true) realization; each expansion of Z that corresponds to a network generates its own "Z positivity" condition. Let me warn the reader, however, that no expansion of Z beyond Foster-Cauer has found a permanent niche in the literature.

The proofs of Foster and Cauer gave few clues about the totality of possible realizations. Finding new realizations of Foster-Cauer impedances was a hot and prestigious research topic in theoretical electrical engineering departments, especially at MIT. Guillemin's book [8] contains a long discussion of various possibilities — but without yielding definitive results like an elegant characterization of Foster-Cauer impedances.

3 Brune

This state of affairs around 1925 was something of an international research scandal, or, better said, a rare opportunity. The obvious goal was to remove the restrictions implicit in Foster-Cauer realizations and find conditions on Z equivalent to realizability by a network composed of arbitrary interconnections of positive-valued R, C and L. The race was "won" by Otto Brune (1901–1982), a South African engineer, in the form of a spectacular MIT doctoral dissertation. It acknowledged only the influence of Cauer and was quickly published as [3].

The much-sought-after characterization of Z, as formulated by Brune, is the content of the following definition:

$Z(s)$ is (a) *positive real* (function) iff $\mathrm{Re}(s) > 0$ implies $\mathrm{Re}(Z(s)) > 0$.

And Brune's theorem is:

(a) The impedance $Z(s)$ of any electric network composed of passive components is positive real.
(b) If $Z(s)$ is positive real it is realizable by a network having as components passive (positive) R, C, L *as well as ideal transformers* T.

An "ideal transformer" T is physically the same as a pair of maximally closely coupled (positive) inductances. This is a 2-port network component while R, C, L are all 1-port components. Viewed as coupled inductors T is a thing which is physically possible but as a manufactured product it is very difficult to make it a good physical approximation of its mathematical description. Viewed abstractly T is a special kind of passive component such that there is no need to worry about it having a "value" that must be positive. Coupled resistors and coupled capacitors are likewise physically possible but have not generated interesting technology in the past.

Part (a) of Brune's theorem was proved by using physically-motivated arguments related to passivity (no energy is generated within the network). Part (b) was proved by exhibiting a realization algorithm.

The question arises again: what, if anything, is wrong with Brune's theorem and his realization algorithm?

Really, only one thing: An ideal transformer T is a cute theoretical invention but not a practical network component. And it is also mathematically awkward. Engineers ask: *where is the achievement?*

There is a lemma in linear algebra, arising from Quantum Mechanics in the late 1920's, that almost trivializes Brune's theorem: "Any pair of positive definite quadratic forms may be simultaneously diagonalized." It was well known by 1930, again because matrix theory had become energized by problems emerging from Quantum Mechanics, that any R, C, L network (allowing also arbitrary coupling between inductors as well as coupling between resistors and between capacitors) can be represented by a matrix triple $\{R, C, L\}$ — this amounts to packaging all R, C and L in the network into correspondingly labeled matrices. When all network components are passive then these matrices will be positive definite. The diagonalizing transformations needed in the lemma turn out to be system-compatible, in other words, they preserve impedance. Hence the lemma can be immediately applied to two-kinds-of-components networks, making contact with the Foster-Cauer developments. A triple of matrices, in general, cannot be transformed into pure diagonal form; applying the theorem to the triple $\{R, C, L\}$ it is possible to diagonalize, say, R and C but then L must be abandoned to its fate and will remain nondiagonal. Thus an arbitrary network cannot necessarily be reorganized preserving impedance but eliminating coupling between inductors. So a T occurring in a Brune realization cannot always be removed. Conclusion, based on the lemma: there is an essential difference between the RC, RL and the RCL cases.

4 Bott-Duffin

It was clearly understood that Brune's theorem makes no claim that T are unavoidable components for realizing any positive real Z, or unavoidable for realizing even some subclass of these Z's. There was no proof of the unavoidability of ideal transformers and what may have been lacking was a better method of realization. Nonetheless, 18 years after Brune the community of passive network synthesizers was shaken as if by an earthquake by another brilliant doctoral dissertation, this time in mathematics, by Raoul Bott (1923–2005). The new theorem required only half a journal page [2].

Bott's theorem: "If $Z(s)$ is positive real it can be realized by a network constructed from passive (positive) components R, C, L". The proof (a direct consequence of a technical lemma about "positive real") was constructive, it provided a realization algorithm with explicit formulas.

Ironically, the shock that Brune's theorem was to be superceded by something defying engineering intuition was instantly tempered by bitter disappointment: the realization created by Bott's theorem was hopelessly impractical, because the number of components required grew exponentially with $n = \deg(Z) = \max(\deg(p), \deg(q))$. See [8, Chapter 10] for a soon-after-the-event analysis. In 1953 as an undergraduate student at MIT I learned this material in Guillemin's graduate course while he was working on the book.

Passive network synthesis, theory and application, never came to grips with this strangely paradoxical situation. By 1970 the field was dead. Earlier, the 1950's saw the birth of "modern" system theory, which dealt with similar problems and generated many new ideas, results, techniques. So an autopsy of the corpse has become possible. What would we find? The preceding discussion points to two deeply buried issues:

(i) System passivity ("positive real") versus component passivity ("$R > 0$, $C > 0$, $L > 0$");
(ii) Minimality: how many components are needed and of what kind?

What is the system-theoretic meaning of: "Z(s) = positive real"? Brune's theorem claims that any R, C, L network, consisting only of passive (positive) components has this property. Indeed, this was Brune's main mathematical contribution. Unfortunately, his thesis does not contain a crucially important, and utterly trivial, clarifying remark, namely, that, while

$$\text{passive components} \implies \text{passive system}$$

(an intuitively obvious statement, acceptable in the network context to most physicists and engineers), the converse calls for a proof, and in fact the converse is obviously false in almost any system context. Very unfortunately, Brune's readers did not spot the big hidden issue: were the converse true, it would have justified "positive real" as a sharp characterization of a system built from passive components. Since the converse is not true, "positive real" is nothing more than a Z-testable condition for a system (as a whole) to be passive.

This misunderstanding had fateful consequences. It is recognizable on almost every page of [8], indeed Guillemin dedicated the book to *"Otto Brune who laid the mathematical foundations for modern realization theory"*. Wrong, but not Brune's fault.

"Modern" linear system theory studies systems as a whole, linear components linearly interconnected. Passivity as a linear system property has been routinely and thoroughly researched over the past 50 years, and it has been known for decades that "positive real" is an elegant and concise characterization of a passive linear system when it is viewed in its external (transfer function) description. Part (a) of Brune's theorem is a permanent contribution. But in trying to prove the converse, part (b), Brune had, no doubt unknowingly, violated Newton's command

"Hypotheses non fingo".

For the instruction of the common man here is Newton's inelegant Latin translated into colloquial English, "Be sure never to add anything extraneous to an already defined problem (even if that makes the problem easier to solve)."

In Brune's defense, it must be said that he did abide, perhaps also unknowingly, by Occam's Razor, "Do not pile up extraneous assumptions without dire need". He added just one component, T, and he needed that T very badly to prove his realization algorithm.

Occam's Razor is part of the erudition of all historians of science, some of whom may have forgotten, or never did know, that 400 years after Occam but 300 years before our time Newton had seen the issue with great clarity but only when he, Newton, was already 70 years old.

5 Foster Again; His Tactics and Catalog

Unlike his definition of "positive real", Brune's realization algorithm was never regarded as the last word, precisely because of its reliance on T. What could be done to avoid T? There is a brute-force way to check if T is unavoidable. One might proceed by enumerating all R, C, L networks and see whether or not they exhaust all positive real impedances. Here is what one would have to do:

(i) List all finite graphs (undirected, connected, no self loops, ...) with two distinguished vertices (the network terminals, or the 1-port, with respect to which impedance is to be defined). Each branch of such a graph is to correspond to a single (1-port) component R, C, L. This out rules out T (4 terminal points), transistors (3 terminals points), etc.

(ii) Then list all possible networks obtained by populating the branches of the graphs in (i) by R, C, L. Do this minimally, avoiding redundancies such as putting the same type of component on two parallel or two series branches. Exclude from the list networks identical under relabeling of vertices. Make sure there are not too many resistors, so use at most $\#R \leq \#C + \#L + 1$.

(iii) Continue by computing Z corresponding to each network of (ii). In doing so view the values of R, C, L as undetermined (but nonzero) real numbers. A branch with a component of zero impedance is a short circuit, the two end vertices of the branch must be identified, the branch becomes a self-loop and is eliminated. A branch with a component of zero admittance (infinite impedance) is a cut branch, it is deleted from the graph. Therefore zero (or infinite) component values are forbidden in this problem formulation, necessarily so to make the graph of the network well defined.. Each resulting Z will be a pair of polynomials with undetermined real coefficients except that coefficients may be determined as identically zero. Each such Z may correspond to several distinct networks. Record the results as a list of $\{$impedance, $\{$corresponding networks$\}\}$.

(iv) Using the (large) list of $\{$impedance, $\{$corresponding networks$\}\}$ generated in step (iii) invert the process: Given a Z, determine the component values in each of the networks corresponding to Z as explicit functions of the parameters (coefficients of p, q) of Z.

(v) Inequalities on component values to assure positivity are inequalities on coefficients of Z. See if these inequalities allow every positive real Z to be positively realized.

Assuming these steps can be carried out effectively, step (iv) is the solution of the R, C, L quasi-realization problem *without using* T; as before, “quasi-realization” means that component values, never zero, may be positive or negative. Because of this, the question still remains, “Does the set $\{$impedances, $\{$corresponding passive R, C, L networks$\}\}$ contain all positive real impedances?” Answering it is the task in step (v).

I have just described the research strategy explored by Foster in the 1940’s. It required calculations that must have spanned several years. The calculations were carried on up to and including the generic biquadratic case: $n = \deg(p) = \deg(q) = 2$, no identically zero coefficients (Z property); $\#R = \#C + \#L + 1$ (network property). The results, up to and including step (iv), were recorded in an unpublished M.S. (!!!) thesis [10]. They are of a truly amazing complexity, especially considering the fact that the (re)search was limited to first-order (linear) and second-order (quadratic) networks on which there exists, in mathematical physics and engineering, an ocean of literature without even a hint at the complexity that Foster had discovered.

To give an idea of the contents: over 130 graphs need to be considered (list (i)), there are exactly 108 distinct networks to be looked at (list (ii)), and roughly 1000 formulas, The generic biquadratic impedance alone has 59 distinct network realizations, of which 49 are generic networks (3 resistors) and 10 are subgeneric networks (2 resistors). There is absolutely no indication of how the computations were made — a pity. But the catalog is valuable experimental data, to be checked for accuracy and explained by theory.

Foster’s main objective, presumably, was to characterize the set $\{Z$ minimally realizable without $T\}$, step (v). In this, he failed — no results in [10] concerning step (v). Foster must have been frightened by the complexity revealed by his brute-force analysis of his question, his fears turning into despair when, just a few months after [10], Bott gave the answer to his question, by a totally different and very elegant approach.

Yet Foster need not have been discouraged, he was on the right track. He had separated the problem into two natural parts:

- realization without regard to positivity of components, and
- separate consideration of the passivity of Z and the positivity of the components of its realization.

The treatment of the first part is complete in [10], and there is no indication that the second part was attempted at all. Note also also that everything recorded in [10] revolves around minimal networks, yet “minimality”, as a system-theoretic notion, is nowhere used, defined or discussed. Bott’s theorem, with due hindsight, offers us a first glimpse into the still unexplored jungle of nonminimal realizations — it had nothing to do with Foster’s basic research strategy, Bott (wisely) dropped the topic immediately after his doctorate, and no one was able to do anything with it after him. Such is the psychology of scientific creativity — Foster never published anything significant about the subject after Bott. [10] remains buried in the library, outside the view of GOOOOGLE and its competitors. And yet ...

6 The Road Ahead

In Foster’s last serious publication known to me, there is [5, page 868] the following claim, without proof, presumably distilled from the results recorded in the catalog [10]:

“Let all undetermined coefficients of Z be positive (they are positive whenever all component values of a realization are positive) and consider the resultant Res(Z) of Z.

If Res(Z) > 0 then Z has a minimal quasi-realization of either the *RC* or *RL* type, and only of that type, depending on whether the invariant $B(Z) = a_0b_2 - a_2b_0$ is positive or negative.

If Res(Z) < 0 then Z has a minimal quasi-realization of the *RCL* type, and only of that type.”

The prima donna here is evidently Res(Z). I explain the terminology. The classical algebraic object known as the “resultant” (and there are also multiple resultants, subresultants, etc.) measures whether or not a pair of polynomials have a common factor, in which case the resultant is zero. It can be represented as the determinant of the classical Sylvester matrix — Res(Z) = det(Sylv(Z)) — and in many other ways; there is active and growing interest in such things in computational algebra, real algebraic geometry, theory of elimination, ... In [5] Foster calls the resultant “discriminant” (which is actually a special kind of resultant)—-this is a clue pointing to bad communication between algebraists and those, like Foster, outside algebra, as late as the 1960’s. Today that situation is greatly changed, there are special treatises dealing with resultants, ... See [7] and, more recently, [1].

It should be noted that this theorem (for $n > 2$ at present only a conjecture) is totally outside to the mathematical mainstream in Foster-Cauer-Brune environment of the 1920’s. Then Heaviside’s genius-class idea — that the “operator” s was an algebraic thing, not a complex number representing some kind of frequency — was viewed with conservative skepticism and the preferred way of looking at impedances was that $Z(s)$ was as a complex *function*. This has blocked any possibility of computing Res(Z). In fact, common factors of p, q were supposed to be "divided out" during some stage of the computation of Z from the network. This kind of nonsensical applied mathematics was eliminated only after the advent of "modern" system theory in the early 1960's.

That Foster's claim is true, up to and including the biquadratic case, seems to rest on checking each of the 108 cases in the Ladenheim catalog. It is clearly a theorem and a proof is badly needed.

The role of Res(Z) in the theorem is striking. Res(Z) is an algebraic object that disappears from sight when $Z(s)$ is viewed as a *function*. With the help of Res(Z) the possible minimal quasi-realizations of Z are classified into three disjoint types, *RC*, *RL*, *RCL*. The classification into the three types is exhaustive because $B(Z) = 0$ implies $\mathrm{Res}(Z) < 0$ and because Res(Z) = 0 means that deg(Z) is reduced by at least 1 so the networks corresponding to Z are nonminimal. Passivity plays no role at this point. But concrete, elegant, explicit conditions for true minimal realizations ($R > 0, C > 0, L > 0$) remain elusive. They seem to be an automatic consequence of condition $\mathrm{Res}(Z) > 0$ for two-kinds-of-components networks but in the *RCL* case the exact passivity conditions are opaque because they do not depend solely on the sign or value of Res(Z).

Example. Consider the Z-class given by $Z(s) = p(s)/q(s)$, $p(s) = a_1 s + a_0$, $q(s) = b_2 s^2 + b_1 s + b_0$ (a sub-biquadratic). There are a total of 7 minimal networks corresponding to this Z-class. The *RCL* networks are shown in Figures 1–3, corresponding to [10, cases 47–49]. The *RC* networks [10, cases 43–46] are shown in Figures 4–7. There are no *RL* networks in this Z-class because $B(Z) > 0$.

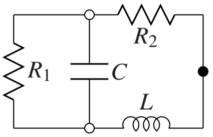


Fig. 1

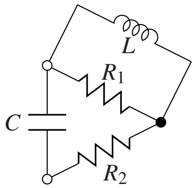


Fig. 2

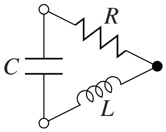


Fig. 3

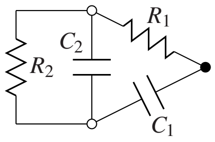


Fig. 4

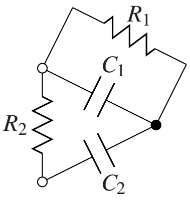


Fig. 5

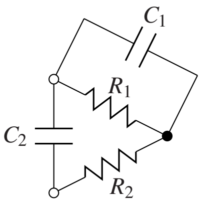


Fig. 6

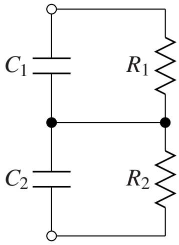


Fig. 7

Because of "minimality" the networks are restricted to the following component count: since deg(Z) = 2, #C + #L = 2; because Z has 5 nonzero coefficients the number of components is $5 - 1 = 4$, hence #R = 2. So the networks relevant to our

Z-class must have 4 branches. Except, Fig. 3 has only 3 branches, with C, L, and (only one) R. Why? As required by our chosen Z-class, the impedance of the network in Fig. 3 has five (not-identically-zero) coefficients, but these cannot be independent since the graph of the network of Fig. 3 has only 3 branches. Therefore there must exist an algebraic condition satisfied by the coefficients of Z of the network of Fig. 3. To find this condition we need to compute the coefficients of Z as a function of the network components R, C, L. For Fig. 3 this is easy, and in any case were given (correctly!) in [10] as

$$a_1 = L, \quad a_0 = R, \quad b_2 = CL, \quad b_1 = RC, \quad b_0 = 1.$$

(These expressions are obtained via the convention of specifying the component values as resistance, inductance (impedances) and capacitance (admittance); in our special case, the convention causes Z to be in a "normalized" form with a_0 equal to 1.) The condition we are looking for is easily seen to be

$$a_0 b_2 - a_1 b_1 = 0,$$

which is satisfied by any choice of nonzero R, C, L in Fig. 3.

The network in Fig. 3 is a realization of exactly that subclass of the our subbiquadratic Z for which the above condition holds. There are similar results for the generic biquadratic Z, as can be verified in [10].

The RC networks in Figures 4–7 were not entirely new in 1948. Fig. 7 is a Foster canonical form, Figures 5–6 are Cauer canonical forms, and Fig. 4 is a Foster-like form. Another interesting fact is that all these networks have industrial applications: they provide the required RIAA equalization in (long-playing) phonograph (records) preamplifiers. In a careful scholarly article [11] Stanley Lipschitz, a South African Engineer (like Brune), working at the University of Waterloo, researched the engineering literature pertaining to such amplifiers, including service manuals and circuit diagrams of various manufacturers, and concluded that those shown in Figures 4–7 were the "four most commonly used equalization networks" [11, page 480, Fig. 1 (a)–(d)]. Of course, Lipschitz's empirical approach didn't (couldn't) prove that there were no other such networks. What is seen here is a genuine experimental confirmation of discoveries arising from Foster's research strategy of the 1940's. Moreover, in a physiology research paper [6, page 525, Fig. 1] we find the same four networks. There are probably other examples that could be dug up.

7 A Reflection on the Preceding

Apparently at the time of [4] nobody noticed that it continued the Newtonian tradition: Newton disposed (up to the three-body problem) of the one-kind-of-component problem (movement of point masses under the gravitational force), then Foster did the same for the two-kinds-of-components problem, though he apparently did not see the problem clearly until 40 years later [5]. The three-kinds-of-components problem, in networks, is still largely open, but then there are also the (electron, proton, neutron) systems called atoms. Dare we dream of bringing DNA into the picture — it seems to be a four-kinds-of-components problem.

8 Conclusion

Much remains to be done. But the Promised Land has been sighted from several directions.

References

1. Apéry, F., Jouanolou, J.-P.: Élimination: le cas d'une variable, Hermann (2006)
2. Bott, R., Duffin, R.J.: Impedance synthesis without use of transformers. J. Applied Physics 20, 816 (1940)
3. Brune, O.: Synthesis of a finite two-terminal network whose driving-point impedance is a prescribed function of frequency. J. of Mathematical Physics 10, 191–236 (1931)
4. Foster, R.M.: A reactance theorem. Bell System Technical J. 3, 259–267 (1924)
5. Foster, R.M.: Academic and theoretical aspects of circuit theory. Proceedings IRE 50, 866–871 (1962)
6. Freygang, W.H., Trautwein, W.: The structural implications of the linear electrical properties of cardiac Purkinje strands. J. General Physiology 55, 524–547 (1970)
7. Gel'fand, I.M., Kapranov, M.M., Zelevinsky, A.V.: Discriminants, Resultants and Multidimensional Determinants. Birkhäuser (1994)
8. Guillemin, E.A.: Synthesis of Passive Neworks. Wiley (1957)
9. Kalman, R.: Discovery and invention: the Newtonian revolution in systems technology. J. Guidance, Control, and Dynamics 26, 833–837 (2003)
10. Ladenheim, E.L.: A synthesis of biquadratic impedances, M.S. thesis, supervised by R. M. Foster, Polytechnic Institute of Brooklyn (May 1948)
11. Lipschitz, S.P.: On RIAA equalization networks. J. Audio Engineering Society 27, 458–491 (1979)

Regular Positive-Real Functions and the Classification of Transformerless Series-Parallel Networks

Jason Zheng Jiang and Malcolm C. Smith

Abstract. This paper studies series-parallel electrical or mechanical networks using the recently introduced concept of regular positive-real functions. Previous work showed that series-parallel five-element networks with two reactive elements are always regular and that six such networks can realise all regular biquadratic immittances. In this paper we consider five- and six-element networks with three reactive elements. We describe a classification procedure to find an efficient subset of such networks which may realise any non-regular biquadratic that can be synthesised by this class of networks.

1 Introduction

The possibility to realise any positive-real function as the driving-point immittance of a network consisting of resistors, capacitors and inductors only was established by Bott and Duffin in [1]. The construction appears to be wasteful in terms of the number of elements used, however subsequent research has failed to solve the question of "minimal realisation". For about 20 years following the publication of [1] there was an attempt to classify simple networks by exhaustive enumeration. Nowadays this literature is hard to digest and verify, and there is no complete statement of the results that were obtained. In this paper we take a fresh look at the classification of series-parallel networks using the recently introduced concept of regular positive-real functions. The work is motivated by a new mechanical network element (the inerter) which has revived interest in passive network synthesis [16].

2 The Concept of Regularity and Its Properties

In this section we recall the concept of regularity and some of its properties given in [8, 9].

Jason Zheng Jiang · Malcolm C. Smith
Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, UK
e-mail: zj219@cam.ac.uk, mcs@eng.cam.ac.uk

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 15–25.
springerlink.com

Definition 1. A positive-real function $Z(s)$ is defined to be *regular* if the smallest value of $\operatorname{Re}(Z(j\omega))$ or $\operatorname{Re}\left(Z^{-1}(j\omega)\right)$ occurs at $\omega = 0$ or $\omega = \infty$.

Lemma 1. *Let $Z(s)$ be a regular positive-real function. Then $\alpha Z(s)$, $Z(\beta s)$, $Z\left(s^{-1}\right)$, $Z^{-1}(s)$ are all regular, where α, $\beta > 0$.*

Lemma 2. *Let $Z(s)$ be a regular positive-real function. Then $Z(s) + R$ and $Z^{-1}(s) + R^{-1}$ are both regular, where R is nonnegative.*

The next lemma follows from the fact that the impedance $Z(s)$ or admittance $Y(s)$ of any network that has all reactive elements of the same kind has $\operatorname{Re}(Z(j\omega))$ and $\operatorname{Re}(Y(j\omega))$ monotonic ([5, Chapter 2.2]).

Lemma 3. *Any network that has all reactive elements of the same kind can only realise regular immittances.*

Lemma 4. *Any network that has a path between the two external terminals 1 and $1'$ or a cut set ([15]) that places 1 and $1'$ in different connected parts consisting of one type of reactive element can only realise regular immittances.*

Proof. This follows since either the impedance or the admittance has a zero at $s = 0$ or ∞ (see [14, Theorem 2]).

We now focus attention on biquadratics

$$Z(s) = \frac{As^2 + Bs + C}{Ds^2 + Es + F}, \tag{1}$$

where $A, B, C, D, E, F \geq 0$. It is well known [4, 17, 2] that $Z(s)$ is positive real if and only if $\sigma = BE - \left(\sqrt{AF} - \sqrt{CD}\right)^2 \geq 0$.

The classification of networks is facilitated by the following transformations on the impedance $Z(s)$:

1. Multiplication by a constant α,
2. Frequency scaling: $s \longrightarrow \beta s$,
3. Frequency inversion: $s \longrightarrow s^{-1}$,
4. Impedance inversion: $Z \longrightarrow Z^{-1}$.

In network realisations, the first two transformations correspond to element scaling, the third to replacing inductors with capacitors of reciprocal values (and vice versa), and the fourth to taking the network dual. The third and fourth transformations together allow networks to be arranged into groups of four, which we call *network quartets* (see Fig. 1). Such families have appeared in [3] with the terminology “Untergruppe”. It should be noted that a network quartet may sometimes reduce to two or even one distinct network(s). It follows from Lemmas 1 and 2 that if a network can only realise regular immittances, then so will the other networks in the quartet and also the networks obtained by adding a resistor in series or in parallel with the original one.

$$\begin{tikzcd}
\mathrm{N}_a \arrow[rr, leftrightarrow, "\text{dual}"] \arrow[dd, leftrightarrow, "s \leftrightarrow s^{-1}"] & & \mathrm{N}_d \arrow[dd, leftrightarrow, "s \leftrightarrow s^{-1}"'] \\
 & & \\
\mathrm{N}_c \arrow[rr, leftrightarrow, "\text{dual}"] & & \mathrm{N}_b
\end{tikzcd}$$

Fig. 1 Transformations relating members of a network quartet.

3 Networks That Can Only Realise Regular Biquadratics

In this section we present a series of lemmas which will facilitate our classification of networks in the next section.

Lemma 5. ([6, 8]) *The network quartet in Fig.* 2 *(which contains only two distinct networks) can only realise regular immittances.*

Lemma 6. ([6]) *The network shown in Fig.* 3 *can only realise regular immittances.*

Lemma 7. ([7]) *The network shown in Fig.* 4 *can only realise regular biquadratic immittances.*

Lemma 8. ([7]) *The networks shown in Fig.* 5 *can only realise regular biquadratic immittances.*

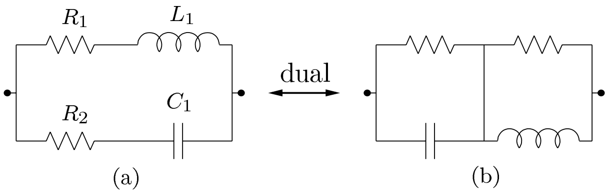


Fig. 2 A series-parallel network quartet (which reduces to two distinct networks) that can only realise regular immittances.

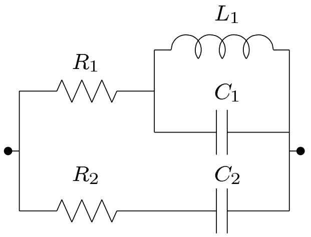


Fig. 3 A series-parallel five-element network with three reactive elements which can only realise regular immittances.

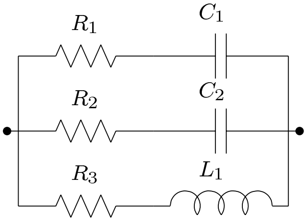


Fig. 4 A series-parallel six-element network with three reactive elements which can only realise regular biquadratics.

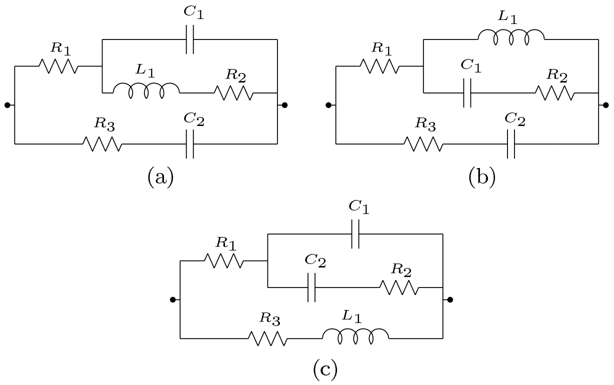


Fig. 5 Three series-parallel six-element networks with three reactive elements which can only realise regular biquadratics.

4 Series-Parallel Networks That Can Realise Non-regular Biquadratics

In this section we present a complete classification of low-complexity series-parallel networks using the concept of regular positive-real functions and the lemmas developed in Section 3.

Lemma 9. ([11]) *For arbitrary impedances* $Z_1(s)$*,* $Z_2(s)$ *and positive constants a, b, c, the networks of Figs.* 6 (a) *and* (b) *are equivalent under the transformations:* $a' = a(a+b)/b$*,* $b' = a+b$*,* $c' = c((a+b)/b)^2$ $[a = a'b'/(a'+b')$*,* $b = b'^2/(a'+b')$*,* $c = c'(b'/(a'+b'))^2]$.

Theorem 1. ([8, 6]) *A biquadratic immittance can be realised by series-parallel five-element networks with two reactive elements (and all element values nonnegative) if and only if it is regular.*

Theorem 2. *All series-parallel networks with three reactive and two resistive elements (and all element values nonnegative) can only realise regular immittances except for the network quartet of Fig.* 8.

Proof. If no distinction is made among the elements, there are 24 distinct two-terminal series-parallel structures with five elements [13, 12]. In Fig. 7, 12

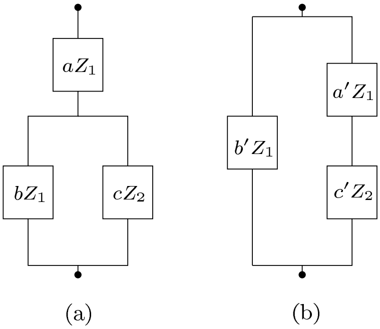


Fig. 6 Two equivalent networks related by the transformations in Lemma 9.

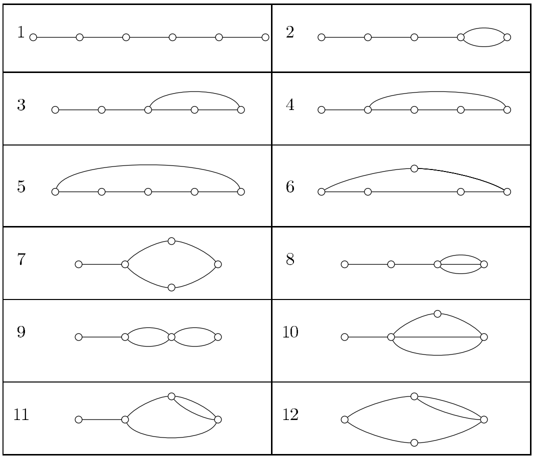


Fig. 7 One-Half of the Two-Terminal Five-Element Series-Parallel Structures.

five-element structures are shown which together with their duals make up the 24 structures. (These are the structures with 6 or 5 vertices and one half of those with 4 vertices.) Based on Lemma 1, the analysis may be performed on these structures only. Based on Lemma 3, we only need to investigate the networks containing both kinds of reactive elements. Therefore, it is sufficient to consider only the assignments of elements with two capacitors and one inductor, since immittances which are regular remain so after the $s \to s^{-1}$ transformation (Lemma 1).

It is straightforward to see using Lemmas 2, 4 and 5 that structures 1–5, 7–11 can only realise regular immittances. For example, in structure 7, only the case where the series element is a resistor needs to be considered by Lemma 4. Then, if the two capacitors are in the same or different parallel branches Lemma 4 applies and the network consists of a resistor in series with a regular immittance, which can only realise regular immittances by Lemma 2.

For structure 6, if there are two resistors or two reactive elements in the upper branch, Lemmas 2 and 4 show that the network can only realise regular immittances. When one resistor is in each branch the network can only realise regular immittances by Lemmas 4 and 5.

Structure 12 is regular if the lower branch contains two resistors or two capacitors, by Lemmas 2 and 4. If the lower branch contains an inductor and capacitor the series element in the upper branch must be a resistor (otherwise Lemma 4 applies) in which case Lemma 9 can be used on the upper branch to transform to a network with a resistor in parallel with a network which, by Lemma 4, can only realise regular immittances, and then Lemma 2 applies. If the lower branch contains a resistor and inductor the structure is regular by Lemma 5 or Lemma 4. It remains only to consider the case where the lower branch contains a resistor and capacitor. If the series element in the upper branch is a capacitor (resp. resistor) the network can only realise regular immittances by Lemma 4 (resp. Lemma 6). If the series element is an inductor the network takes the form of Fig. 8 (a). □

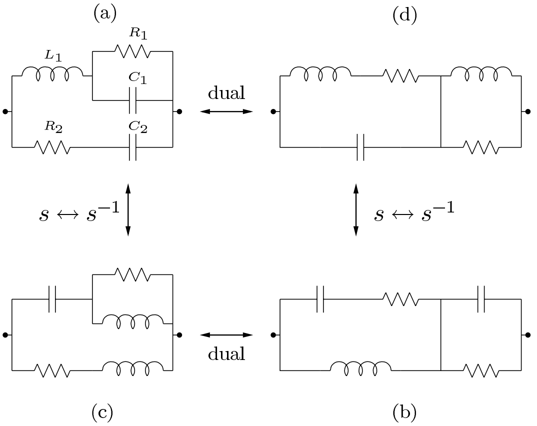


Fig. 8 The series-parallel three-reactive five-element network quartet that can realise non-regular biquadratics.

Theorem 3. *A non-regular biquadratic immittance can be realised by a series-parallel network with three reactive and three resistive elements if and only if it is realisable by some network in the four network quartets of Figs.* 10, 11, 12, 13.

Proof. If no distinction is made among the elements, there are 66 distinct two-terminal series-parallel structures with six elements [13]. These structures may be divided into two classes, any structure in one class having its dual in the other. Based on Lemma 1, the analysis may be performed upon only one class. In Fig. 9, all the series-parallel six-element structures in one class are presented according to a numbering of Vasiliu [19]. Based on Lemma 3, we only need to investigate the networks containing both kinds of reactive elements. Furthermore, it is sufficient to consider only the assignments of elements with two capacitors and one inductor by Lemma 1.

It is apparent that structures 1–6, 24 must reduce to a network with at most five elements. It can be checked that structures 7–23 will either reduce to a network with fewer than six elements or can only realise regular immittances by Lemmas 2 and 4.

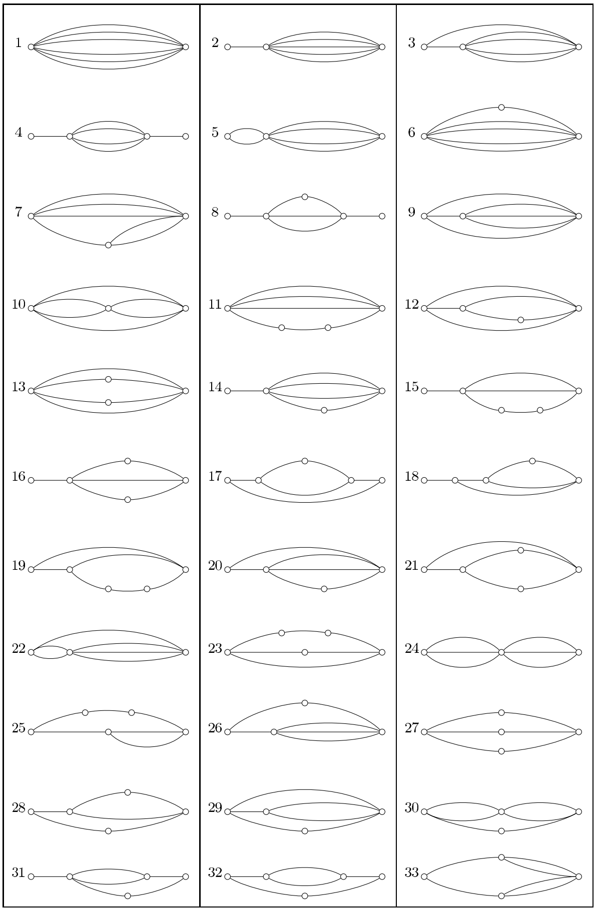


Fig. 9 One-Half of the Six-Element Series-Parallel Structures.

For structure 25, if it does not reduce to a five-element network the upper branch contains all three kinds of elements and the lower branch has the resistor in series. After applying the transformation of Fig. 6 to the lower branch, there is a resistor in parallel with a network which can only realise regular immittances by Lemma 4.

For structure 26, if it does not reduce to a five-element network the upper branch has a resistor and capacitor and the lower branch has a resistor in series. After applying the transformation of Fig. 6 to the lower branch, we obtain a network which has a resistor in parallel with the network of Fig. 3. The resulting network can only realise regular immittances by Lemmas 2 and 6.

For structure 27, if it does not reduce to a five-element network it takes the form of Fig. 4, which can only realise regular biquadratics by Lemma 7.

For structure 28 there must be one resistor in the lower branch, otherwise the network reduces to five-elements (using Lemma 9 in some cases). If there is a

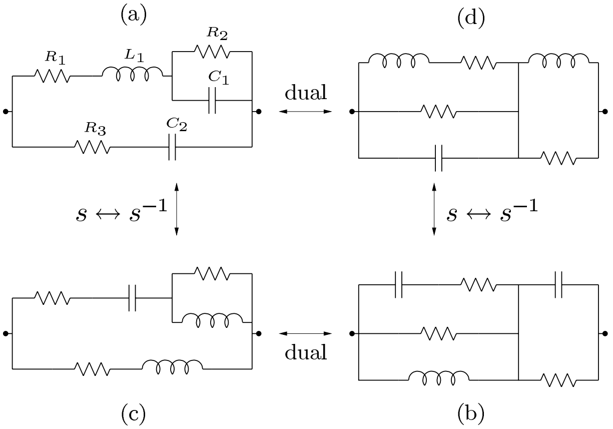


Fig. 10 One of the four network quartets of the series-parallel three-reactive six-element networks that can realise non-regular biquadratics.

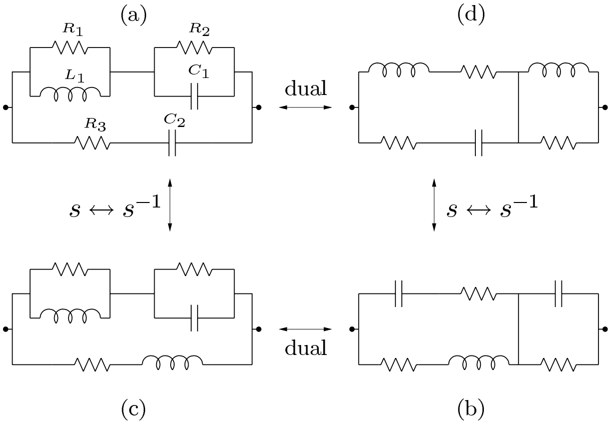


Fig. 11 One of the four network quartets of the series-parallel three-reactive six-element networks that can realise non-regular biquadratics.

resistor and inductor in the lower branch the network takes the form of Fig. 5(c) (or a form which can be transformed to it with two applications of Lemma 9) which can only realise regular biquadratic immittances by Lemma 8. It remains to consider the cases of a resistor and capacitor in the lower branch. If a capacitor is the series element (in the upper branch) the network can only realise regular immittances by Lemma 4. If the inductor is the series element (in the upper branch) it reduces to the network of Fig. 10 (a) using Lemma 9. If a resistor is the series element (in the upper branch) there are three cases: two of these (Figs. 5 (a) and (b)) are eliminated by Lemma 8. The third is eliminated using Lemma 9 followed by Lemmas 2 and 4.

For structure 29 the top branch must be a resistor by Lemma 4. The rest of the structure can only realise non-regular immittances if it takes the form of Fig. 8 (a)

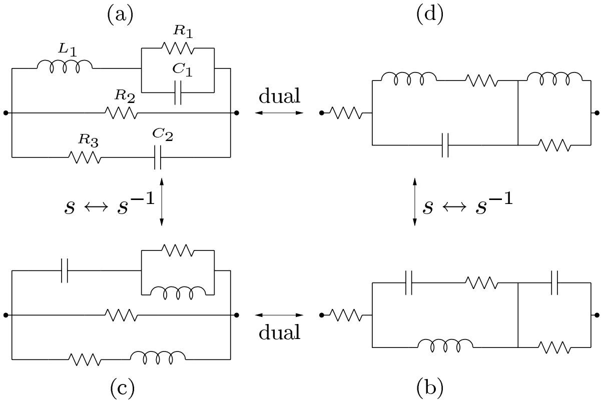


Fig. 12 One of the four network quartets of the series-parallel three-reactive six-element networks that can realise non-regular biquadratics.

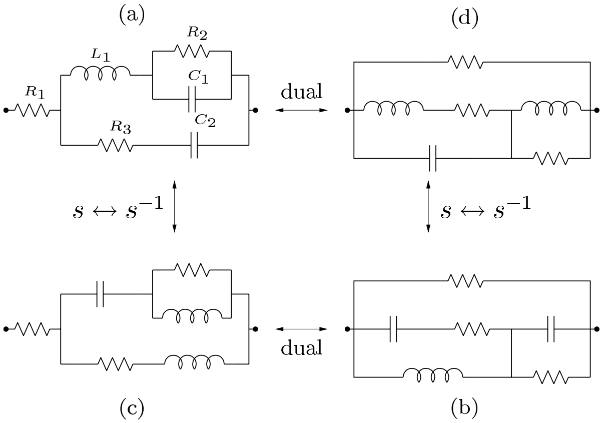


Fig. 13 One of the four network quartets of the series-parallel three-reactive six-element networks that can realise non-regular biquadratics.

(see the proof of Theorem 2). Hence, the only possibility to realise non-regular immittances is the network of Fig. 12 (a).

For structure 30 there must be one resistor in the lower branch otherwise the network reduces to at most five elements. If the other element in the lower branch is an inductor, the network can only realise regular immittances by Lemma 4 or it will reduce to a network with four elements. Hence the other element in the lower branch must be a capacitor and the only possibility to realise non-regular immittances is the network of Fig. 11 (a).

For structure 31 the series element must be a resistor by Lemma 4. Again the rest of the structure can only realise non-regular immittances if it takes the form of Fig. 8

(a) (see the proof of Theorem 2). Hence, the only possibility to realise non-regular immittances is the network of Fig. 13 (a).

For structure 32, if it does not reduce to a network with at most five elements, there must be one resistor in the lower branch. If the reactive element in the lower branch is an inductor the network can always be transformed to Fig. 5 (c) using Lemma 9, which means that the network can only realise regular biquadratics. If the reactive element in the lower branch is a capacitor there are only two cases which do not reduce to five-element networks. One of these can only realise regular immittances by Lemma 4, the other being the network of Fig. 10 (a).

For structure 33, we need consider only the case where there is a single resistor in the upper branch and two resistors in the lower branch, since if there are three resistors in one of the branches the network reduces to four elements. There must be a capacitor in the lower branch because otherwise the network can only realise regular immittances by Lemma 4 or the network reduces to five elements. Applying Lemma 9 to the lower branch we obtain a resistor in parallel with a network which can only realise non-regular immittances if it takes the form of Fig. 8 (a) (see the proof of Theorem 2). Hence, after transformations, the only possibility to realise non-regular immittances is the network of Fig. 12 (a). □

The networks contained in the quartet of Figs. 8, 10–13 have appeared in the work of Ladenheim [10] and Vasiliu [18] and [19]. The contribution of this paper has been to develop a procedure which proves that these five quartets are a complete set in the following sense: that they can realise all the non-regular biquadratics which can be synthesised by the two classes of networks considered. In [6], [7] the class of non-regular biquadratics realisable by these five quartets is described explicitly.

5 Conclusions

This paper has studied some simple classes of series-parallel networks using the recently introduced concept of regular positive-real functions. The concept has been shown to be useful to efficiently identify the most convenient and powerful networks to realise driving-point immittance functions.

References

1. Bott, R., Duffin, R.J.: Impedance synthesis without use of transformers. J. Appl. Phys. 20, 816 (1949)
2. Chen, M.Z.Q., Smith, M.C.: A note on tests for positive-real functions. IEEE Transactions on Automatic Control 54(2), 390–393 (2009)
3. Dittmer, G.: Zur realisierung von RLC-Brückenzweipolen mit zwei Reaktanzen und mehr als drei Widerständen (On the realisation of RLC two-terminal bridge networks with two reactive and more than three resistive elements). Nachrichtentechnische Zeitschrift 23, 225–230 (1970)
4. Foster, R.M., Ladenheim, E.L.: A class of biquadratic impedances. IEEE Trans. on Circuit Theory, 262–265 (1963)

5. Guillemin, E.A.: Synthesis of Passive Networks. John Wiley & Sons, Chichester (1957)
6. Jiang, J.Z., Smith, M.C.: Regular positive-real functions and five-element network realisations for electrical and mechanical networks (in preparation)
7. Jiang, J.Z., Smith, M.C.: Synthesis of positive-real functions with low-complexity series-parallel networks (in preparation)
8. Jiang, J.Z., Smith, M.C.: Regular positive-real functions and passive networks comprising two reactive elements. In: Proc. European Control Conference (ECC), August 2009, pp. 219–224 (2009)
9. Jiang, J.Z., Smith, M.C.: Synthesis of positive-real functions with low-complexity series-parallel networks. In: Proc. 48th IEEE Conference on Decision and Control (CDC), December 2009, pp. 7086–7091 (2009)
10. Ladenheim, E.L.: Three-reactive five-element biquadratic structures. IEEE Trans. Circuit Theory, 88–97 (1964)
11. Lin, P.M.: A theorem on equivalent one-port networks. IEEE Trans. Circuit Theory 12, 619–621 (1965)
12. Riordan, J.: Introduction to Combinatorial Analysis. John Wiley, New York (1958)
13. Riordan, J., Shannon, C.E.: The number of two-terminal series-parallel networks. J. Math. Phys. 21 (1942)
14. Seshu, S.: Minimal realizations of the biquadratic minimum function. IRE Transactions on Circuit Theory, 345–350 (December 1959)
15. Seshu, S., Reed, M.B.: Linear Graphs and Electrical Networks. Addison-Wesley, Reading (1961)
16. Smith, M.C.: Synthesis of mechanical networks: the inerter. IEEE Trans. Automatic Control 47(10), 1648–1662 (2002)
17. Van Valkenburg, M.E.: Introduction to Modern Network Synthesis. John Wiley & Sons, Chichester (1960)
18. Vasiliu, C.G.: Three-reactive five-element structures. IEEE Trans. on Circuit Theory 16, 99 (1969)
19. Vasiliu, C.G.: Series-parallel six-element synthesis of the biquadratic impedances. IEEE Trans. on Circuit Theory 17, 115–121 (1970)

Ports and Terminals

Jan C. Willems

Abstract. A terminal of an electrical circuit is a wire that allows the circuit to interact with its environment through a potential and a current. Interconnection is defined as variable sharing: two terminals share the same potential and current. A port of an electrical circuit is a set of terminals that satisfy port-KCL (Kirchhoff's current law). Power and energy that enter a circuit is defined for ports. Terminals are for interconnection, ports are for energy transfer. A port of a mechanical system is a set of terminals that satisfy port-KFL (Kirchhoff's force law).

1 Introduction

It is a pleasure to contribute an article to this Festschrift in honor of Yutaka Yamamoto on the occasion of his 60-th birthday. I had the privilege to develop a fruitful research collaboration with him over the last decade, leading to a number of articles [11]–[17] combining ideas from behavioral theory with system representations in terms of rational and pseudorational symbols. I am also grateful to him for hosting me on several pleasant extended visits to Kyoto University over this period.

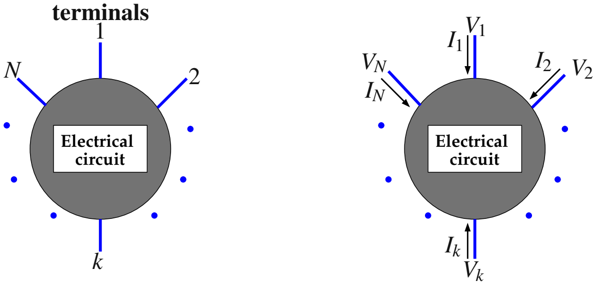
The aim of this article is to explain the distinction that should be made in physical systems between interconnection of systems on the one hand, and energy transfer between systems on the other hand. Interconnection happens via terminals, while energy transfer happens via ports. We consider systems that interact through terminals, as wires for electrical circuits, or pins for mechanical systems. We develop the ideas mainly in the context of electrical circuits, but, towards the end of the paper, we also study mechanical systems.

2 Behavioral Circuit Theory

We view a circuit as follows. An electrical circuit is a device, a black-box, with wires, called terminals, through which the circuit can interact with its environment. This interaction takes place through two real variables, *a potential and a current*, at each terminal. The current is counted positive when it flows into the circuit. For the basic concepts of circuit theory, see [2], [6], or [1]. The setting developed in [5] and [6] has the same flavor as our approach.

Jan C. Willems
ESAT/SCD (SISTA), K.U. Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
e-mail: `Jan.Willems@esat.kuleuven.be`
`http://www.esat.kuleuven.be/~jwillems`

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 27–36.
springerlink.com

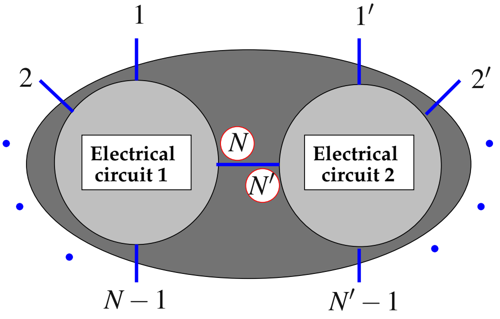


The *behavior* of N-terminal circuit is a subset $\mathscr{B} \subseteq \left(\mathbb{R}^{2N}\right)^{\mathbb{R}}$; $(V,I) \in \mathscr{B}$ means that the time-function $(V,I) = (V_1, V_2, \ldots, V_N, I_1, I_2, \ldots, I_N) : \mathbb{R} \to \mathbb{R}^N \times \mathbb{R}^N$ is compatible with the architecture and the element values of the circuit.

Circuit properties are conveniently defined in terms of the behavior.
A circuit obeys *Kirchhoff's voltage law* (KVL) if $(V_1, \ldots, V_N, I_1, \ldots, I_N) \in \mathscr{B}$ and $\alpha : \mathbb{R} \to \mathbb{R}$ imply $(V_1 + \alpha, \ldots, V_N + \alpha, I_1, \ldots, I_N) \in \mathscr{B}$.
A circuit obeys *Kirchhoff's current law* (KCL) if $(V_1, \ldots, V_N, I_1 \ldots, I_N) \in \mathscr{B}$ implies $I_1 + \cdots + I_N = 0$.

KVL means that the potentials are defined up to an arbitrary additive constant (that may change in time), while KCL means that the circuit stores no net charge.

3 Interconnection



We view interconnection as the connection of two terminals, as shown in the figure below. We start with two circuits, one with N and one with N' terminals. We assume that one terminal (terminal N) of the first circuit is connected to another terminal (terminal N') of the second circuit. The interconnection equations are

$$V_N = V_{N'} \quad \text{and} \quad I_N + I_{N'} = 0.$$

This yields a new circuit with $N + N' - 2$ terminals, with behavior $\mathscr{B}_1 \sqcap \mathscr{B}_2$ defined in terms of the behavior $\mathscr{B}_1$ of the first circuit and $\mathscr{B}_2$ of the second (we consider the connected terminals as internal to the interconnected circuit) as follows.

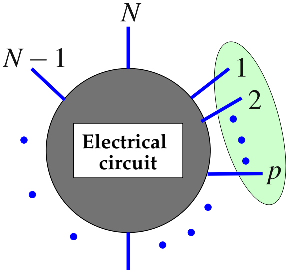
$$\begin{aligned}\mathscr{B}_1 \sqcap \mathscr{B}_2 := \{&(V_1,V_2,\ldots,V_{N-1},V_{1'},V_{2'},\ldots,V_{N'-1},I_1,I_2,\ldots,I_{N-1},I_{1'},I_{2'},\ldots,I_{N'-1}) \\ &\mid \exists\, V,I \text{ such that } \ (V_1,\ V_2,\ \ldots,\ V_{N\ -1},V,I_1,I_2,\ \ldots,I_{N\ -1},\ \ I) \in \mathscr{B}_1 \\ &\qquad\qquad\qquad\quad (V_{1'},V_{2'},\ldots,V_{N'-1},V,I_{1'},I_{2'},\ldots,I_{N'-1},-I) \in \mathscr{B}_2\}.\end{aligned}$$

The idea is that the connected terminals share the voltage and the current (up to a sign) of after interconnection. Note that the product of the shared variables has the dimension of power. The same idea of interconnection applies to the interconnection of two terminals of the same circuit, and to the connection of more terminals of two or more circuits way by connecting one pair of terminals at the time.

Interconnection preserves many circuit properties. In particular, if $\mathscr{B}_1$ and $\mathscr{B}_2$ obey KVL, or KCL, then so does $\mathscr{B}_1 \sqcap \mathscr{B}_2$.

4 Ports

In this section, we introduce a notion that is essential to the energy exchange of a circuit with its environment and between circuits. Consider a circuit with N terminals, and single out p terminals, which, for simplicity, we take to be the first p terminals.



Terminals $\{1,2,\ldots,p\}$ form a (electrical) *port* :⇔

$$(V_1,V_2,\ldots,V_p,V_{p+1},\ldots,V_N,I_1,I_2,\ldots,I_p,I_{p+1},\ldots,I_N) \in \mathscr{B} \\ \Rightarrow I_1+I_2+\cdots+I_p=0.$$

We call this relation *port KCL*. KCL implies that all the terminals combined form a port. It can be shown that for linear passive circuits satisfying KVL and KCL, port KCL is *equivalent* to *port KVL*, defined by

$$(V_1,\ldots,V_p,V_{p+1},\ldots,V_N,I_1,\ldots,I_p,I_{p+1},\ldots,I_N) \in \mathscr{B}, \text{ and } \alpha : \mathbb{R} \to \mathbb{R} \\ \Rightarrow \quad (V_1+\alpha,\ldots,V_p+\alpha,V_{p+1},\ldots,V_N,I_1,\ldots,I_p,I_{p+1},\ldots,I_N) \in \mathscr{B}.$$

If terminals $\{1,2,\ldots,p\}$ form a port, then we define the *power* that flows into the circuit at time t along these p terminals to be equal to

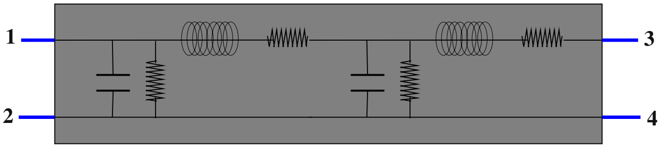
$$\mathit{power} \ = \ V_1(t)I_1(t)+V_2(t)I_2(t)+\cdots+V_p(t)I_p(t),$$

and the *energy* that flows into the circuit along these p terminals during the time-interval $[t_1,t_2]$ to be equal to

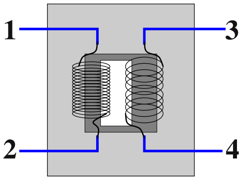
$$energy = \int_{t_1}^{t_2} (V_1(t)I_1(t) + V_2(t)I_2(t) + \cdots + V_p(t)I_p(t))\, dt.$$

Note that port KCL implies that the additive constant from KVL does not appear in the expressions of power and energy.

The above formulas for power and energy are not valid *unless these terminals form a port!* In particular, it is not possible to speak about the energy that flows into the circuit along a single wire — a conclusion that is physically quite obvious. Power and energy flow are not 'local' physical entities, but they involve action at a distance. Note that the terminals of a 2-terminal circuit that internally consists of the interconnection of circuits that all satisfy KVL and KCL form a 1-port, since KVL and KCL are preserved under interconnection. In particular, a 2-terminal circuit that is composed of resistors, capacitors, inductors, transformers, gyrators, memristors, etc. forms a 1-port. However, a pair of terminals of a circuit with more than two terminals rarely forms a 1-port. In particular, for the circuit shown below, the terminals $\{1,2,3,4\}$ form a port, but there is no reason why the terminal pairs $\{1,2\}$ and $\{3,4\}$ should form ports.



An example of an element that consists of more than one port is a transformer.



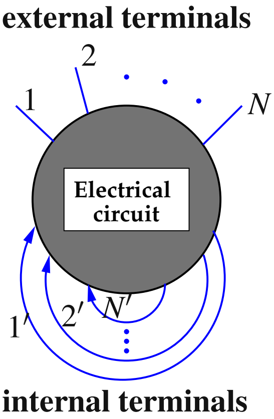
The behavioral equations of an ideal transformer are

$$V_1 - V_2 = n(V_3 - V_4),\ I_3 = -nI_1,\ I_1 + I_2 = 0,\ I_3 + I_4 = 0, \quad \text{with } n \text{ the } \textit{turns ratio}.$$

Clearly $\{1,2\}$ and $\{3,4\}$ form ports, and the energy that flows into the port $\{1,2\}$ is equal to the energy that flows out of the port $\{3,4\}$.

5 Internal Ports

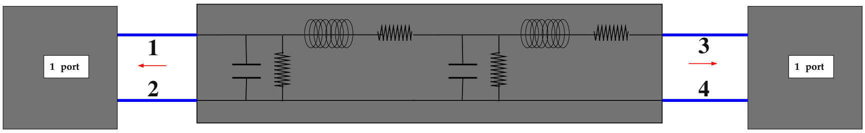
In order to study the energy flow inside a circuit, we introduce in this section circuits with both external and internal terminals. Consider a circuit with N external



terminals and also N' internal terminals. Assume that the internal terminals are directed.

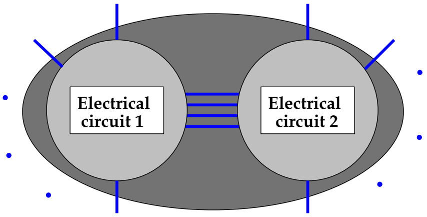
We can define the behavior of this circuit analogously as we did for circuits with only external terminals. A set of terminals, say $\{1', 2', \ldots, p'\}$, forms an *internal port* :⇔ for all elements of the behavior, $I_{1'} + I_{2'} + \cdots + I_{p'} = 0$. A circuit has in general *external ports*, consisting of only external terminals, *internal ports*, consisting of only internal terminals, and *mixed ports*, consisting of both external and internal terminals. The internal ports allow to consider the power and energy flow between parts of a circuit.

For example, it is possible this way to consider the energy transferred into the ports formed by terminals $\{1, 2\}$ and $\{3, 4\}$ of the circuit below, since these pairs form internal ports.



6 Terminals Are for Interconnection, Ports for Energy Transfer

As explained before, interconnection means that certain terminals share the same potential and current (up to a sign). This is distinctly different from stating that the power or the energy flows from one side of an interconnection to the other side. Power and energy involve ports, and this requires consideration of more than one terminal at the time. For example, the two circuits in the figure below share four terminals, but it is not possible to speak of the energy that flows from circuit 1 to circuit 2, unless the connected terminals form internal ports. Similarly, it is not possible to speak about the energy that flows from the environment into circuit 1, or from the environment into circuit 2, unless the external terminals of system 1 and of

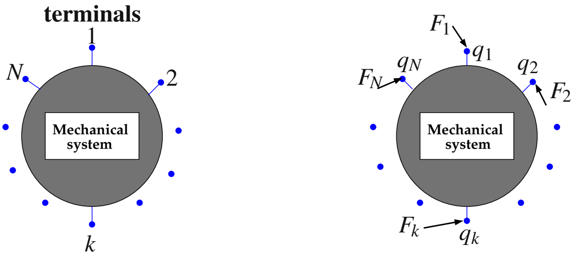


system 2 form ports. Of course, assuming KVL and KCL, the external terminals of the interconnected system always form a port.

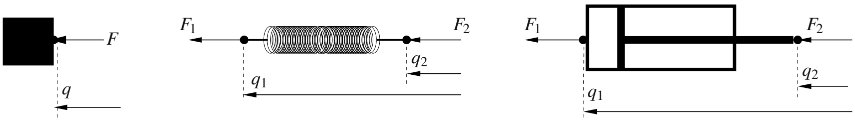
Setting up the behavioral equations of a circuit involves interconnection and variable sharing. Exchange of power and energy involves ports. Interconnections need not involve ports or power and energy transfer. These observations put into perspective power-based modeling methodologies of interconnected systems, as bond graphs [7, 3] and port-Hamiltonian systems [9, 4]. In [10] we propose a modeling methodology for interconnected systems based on *tearing, zooming, and linking*, which involves interconnection by sharing variables, but in which power considerations do not take a central place.

7 Mechanical Systems

We view a mechanical system as a device, a black box, with pins, called terminals, through which the system can interact with its environment. This interaction takes place through two vectors, *a position and a force*, for each terminal. Even though angles and torques play an important role in mechanical systems, we do not consider these here. The position and the force are elements of $\mathbb{R}$ for rectilinear motion, or of $\mathbb{R}^2$ for motions in the plane, or of $\mathbb{R}^3$ for spatial motion. We indicate the fact that we want to leave open which of these cases we consider by the notation $q_k : \mathbb{R} \to \mathbb{R}^\bullet$ and $F_k : \mathbb{R} \to \mathbb{R}^\bullet$.



The *behavior* of the mechanical system is a subset $\mathscr{B} \subseteq ((\mathbb{R}^\bullet)^{2N})^{\mathbb{R}}$; $(q,F) \in \mathscr{B}$ means that the position/force time-function $(q,F) = (q_1,q_2,\ldots,q_N,F_1,F_2,\ldots,F_N) : \mathbb{R} \to (\mathbb{R}^\bullet)^N \times (\mathbb{R}^\bullet)^N$ is compatible with the architecture and the element values of the mechanical system.



Basic building blocks for mechanical systems under rectilinear motion are masses, springs, and dampers. Their behavioral equations are

$$\begin{array}{llll} \text{mass:} & M\dfrac{d^2}{dt^2}q = F, & & \\ \text{spring:} & q_1 - q_2 = \rho(F_1), & F_1 + F_2 = 0, \\ \text{damper:} & \dfrac{d}{dt}q_1 - \dfrac{d}{dt}q_2 = d(F_1), & F_1 + F_2 = 0, \end{array}$$

with $\rho : \mathbb{R} \to \mathbb{R}$ the spring characteristic, and $d : \mathbb{R} \to \mathbb{R}$ the damper characteristic.

We now list some properties of mechanical systems that are conveniently defined in terms of the behavior.

A mechanical system is *invariant under uniform motions* if $(q_1, \ldots, q_N, F_1, \ldots, F_N) \in \mathscr{B}$ and $v : t \in \mathbb{R} \mapsto (a+bt) \in \mathbb{R}^\bullet$, $a, b \in \mathbb{R}^\bullet$, imply $(q_1 + v, \ldots, q_N + v, F_1, \ldots, F_N) \in \mathscr{B}$.

A mechanical system obeys *Kirchhoff's force law* (KFL) if $(q_1, q_2, \ldots, q_N, F_1, F_2, \ldots, F_N) \in \mathscr{B}$ implies $F_1 + F_2 + \cdots + F_N = 0$.

The spring and the damper obey KFL, but the mass does not. Invariance under uniform motions, a most basic premise of mechanics, is important in the sequel.

The interconnection of two mechanical systems is defined by interconnecting two terminals at the time, identifying the positions of the interconnected terminals, and putting the sum of the forces acting on the interconnected terminals equal to zero. The interconnecting equations are

$$q_N = q_{N'} \quad \text{and} \quad F_N + F_{N'} = 0.$$

Note that the product of the shared variables does not have the dimension of power.

This yields, with notation analogous to the one used for circuits,

$$\begin{aligned} \mathscr{B}_1 \sqcap \mathscr{B}_2 := \{ & (q_1, q_2, \ldots, q_{N-1}, q_{1'}, q_{2'}, \ldots, q_{N'-1}, F_1, F_2, \ldots, F_{N-1}, F_{1'}, F_{2'}, \ldots, F_{N'-1}) \\ & \mid \exists\, q, F \text{ such that } (q_1,\, q_2, \ldots, q_{N\,-1}, q, F_1,\, F_2,\, \ldots, F_{N\,-1},\, F) \in \mathscr{B}_1 \\ & \qquad (q_{1'}, q_{2'}, \ldots, q_{N'-1}, q, F_{1'}, F_{2'}, \ldots, F_{N'-1}, -F) \in \mathscr{B}_2 \}. \end{aligned}$$

This leads to interconnection of different terminals of the same mechanical system, and to interconnection of many pairs of terminals of two or more mechanical systems. Interconnection preserves invariance under uniform motion and KFL.

8 Mechanical Ports

We now introduce conditions that allows to study power and energy flow in mechanical systems. Consider a mechanical system, and single out p terminals, which, for simplicity, we take to be the first p terminals.

Terminals $\{1,2,\ldots,p\}$ form a (mechanical) *port* $:\Leftrightarrow$

$$(q_1,\ldots,q_p,q_{p+1},\ldots,q_N,F_1,\ldots,F_p,F_{p+1},\ldots,F_N)\in\mathscr{B},$$
$$\Rightarrow\quad F_1+F_2+\cdots+F_p=0.$$

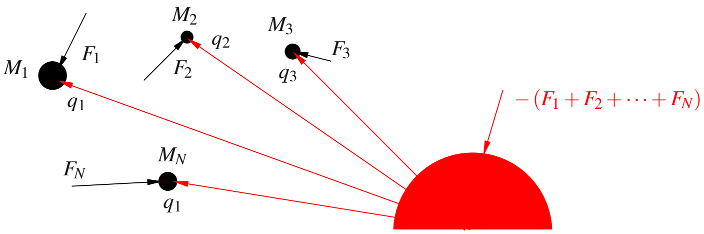
We call this relation *port KFL*. Note that KFL implies that all terminals combined form a port. Also, the external terminals of the interconnection of port devices form again a port. Note that including masses with external forces acting on them form a difficulty for KFL.

If terminals $\{1,2,\ldots,p\}$ form a port, then we define the *power* that flows into the mechanical system at time t along these p terminals and the *energy* that flows into the circuit along these p terminals on the time-interval $[t_1,t_2]$ to be equal to

$$\mathit{power}\ = F_1(t)^\top\frac{d}{dt}q_1(t)+F_2(t)^\top\frac{d}{dt}q_2(t)+\cdots+F_p(t)^\top\frac{d}{dt}q_p(t),$$

$$\mathit{energy}\ = \int_{t_1}^{t_2}\left(F_1(t)^\top\frac{d}{dt}q_1(t)+F_2(t)^\top\frac{d}{dt}q_2(t)+\cdots+F_p(t)^\top\frac{d}{dt}q_p(t)\right)dt.$$

The above formulas for power and energy are not valid *unless these terminals form a mechanical port!* Note that port KFL implies that power and energy are invariant under the additive constant that can be added to the velocities due to invariance under uniform motion. A mass, a spring and a damper obey invariance under uniform motion. A spring and a damper form a mechanical port, but a mass does not. The inerter [8] is a mass-like device that is a port. In order to be able to consider the energy that flows into a mechanical system, we should make sure that the total external force acting on the masses is zero. This can be obtained, albeit in a physically artificial way, by introducing a *'ground'*, an infinite mass that cannot be accelerated, on which the negative of the total force acts, and with respect to which positions are measured, as illustrated below.



We now compute the kinetic energy stored in N moving masses with masses $M_1,M_2,\ldots,M_N$, positions $q_1,q_2,\ldots,q_N\in\mathbb{R}^3$, and with forces $F_1,F_2,\ldots,F_N\in\mathbb{R}^3$ acting on them. By Newton's second law, $M_k\frac{d^2}{dt^2}q_k=F_k$. If we assume that KFL is satisfied, $F_1+F_2+\cdots+F_N=0$, then it is readily verified that

$$\frac{d}{dt}\left(\frac{1}{4}\sum_{i,j\in\{1,2,\ldots,N\}}\frac{M_i\,M_j}{M_1+M_2+\cdots+M_N}\left\|\frac{d}{dt}q_i-\frac{d}{dt}q_j\right\|^2\right)=\sum_{i\in\{1,2,..\ ,N\}}F_i^\top\frac{d}{dt}q_i.$$

Hence the kinetic energy equals

$$\mathscr{E}_{\text{kinetic}}=\frac{1}{4}\sum_{i,j\in\{1,2,\ldots,N\}}\frac{M_i\,M_j}{M_1+M_2+\cdots+M_N}\left\|\frac{d}{dt}q_i-\frac{d}{dt}q_j\right\|^2.$$

$\mathscr{E}_{\text{kinetic}}$ is invariant under uniform motions, as a physically meaningful quantity should be. The expression for $\mathscr{E}_{\text{kinetic}}$ can also be justified by computing the energy that can be stored in a spring or dissipated in a damper, mounted between the masses, while bringing all the masses to the same velocity. This expression is distinct from the classical expression of the kinetic energy,

$$\mathscr{E}_{\text{classical}}=\frac{1}{2}\sum_{i\in\{1,2,\ldots,N\}}M_i\left\|\frac{d}{dt}q_i\right\|^2.$$

In fact, without requiring KFL, there holds

$$\frac{d}{dt}\left(\frac{1}{2}\sum_{i\in\{1,2,\ldots,N\}}M_i\left\|\frac{d}{dt}q_i\right\|^2\right)=\sum_{i\in\{1,2,\ldots,N\}}F_i^\top\frac{d}{dt}q_i.$$

The classical expression $\mathscr{E}_{\text{classical}}$ for the kinetic energy can be made compatible with the expression for $\mathscr{E}_{\text{kinetic}}$ by assuming the presence of an infinite mass at rest on which the force $-(F_1+F_2+\cdots+F_N)$ acts without accelerating it, and applying the formula for $\mathscr{E}_{\text{kinetic}}$.

Acknowledgements. The SISTA research program of the K.U. Leuven is supported by the Research Council KUL: GOA AMBioRICS, CoE EF/05/006 Optimization in Engineering (OPTEC), IOF-SCORES4CHEM; by the Flemish Government: FWO: projects G.0452.04 (new quantum algorithms), G.0499.04 (Statistics), G.0211.05 (Nonlinear), G.0226.06 (cooperative systems and optimization), G.0321.06 (Tensors), G.0302.07 (SVM/Kernel), G.0320.08 (convex MPC), G.0558.08 (Robust MHE), G.0557.08 (Glycemia2), G.0588.09 (Brain-machine) research communities (ICCoS, ANMMM, MLDM); G.0377.09 (Mechatronics MPC) and by IWT: McKnow-E, Eureka-Flite+, SBO LeCoPro, SBO Climaqs; by the Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011) ; by the EU: ERNSI; FP7-HD-MPC (INFSO-ICT-223854); and by several contract research projects.

References

1. Anderson, B.D.O., Vongpanitlerd, B.: Network Analysis and Synthesis. A Modern Systems Approach. Prentice Hall, Englewood Cliffs (1972)
2. Belevitch, V.: Classical Network Theory. Holden-Day (1968)

3. Gawthrop, P.J., Bevan, G.P.: Bond-graph modeling. Control Systems Magazine 27, 24–45 (2007)
4. Jeltsema, D., Scherpen, J.M.A.: Multidomain modeling of nonlinear networks and systems. Control Systems Magazine 29, 28–59 (2009)
5. McMillan, B.: Introduction to formal realizability theory. The Bell System Technical Journal 31, 217–299, 541–600 (1952)
6. Newcomb, R.W.: Linear Multiport Synthesis. McGraw-Hill, New York (1966)
7. Paynter, H.M.: Analysis and Design of Engineering Systems. MIT Press, Cambridge (1961)
8. Smith, M.C.: Synthesis of mechanical networks: the interter. IEEE Transactions on Automatic Control 47, 1648–1662 (2002)
9. van der Schaft, A.J.: Interconnection of port-Hamiltonian systems and composition of Dirac structures. Automatica 43, 212–225 (2007)
10. Willems, J.C.: The behavioral approach to open and interconnected systems. Control Systems Magazine 27, 46–99 (2007)
11. Willems, J.C., Yamamoto, Y.: Behaviors defined by rational functions. Linear Algebra and Its Applications 425, 226–241 (2007)
12. Willems, J.C., Yamamoto, Y.: Behaviors described by rational functions and the parametrization of the stabilizing controllers. In: Blondel, V., Boyd, S., Kimura, H. (eds.) Recent Advances in Learning and Control. Springer Lecture Notes in Control and Information Sciences, vol. 371, pp. 263–278 (2008)
13. Willems, J.C., Yamamoto, Y.: Linear differential behaviors described by rational symbols. In: Proc. 17th IFAC World Congress, Seoul, pp. 1266–1272 (2008)
14. Willems, J.C., Yamamoto, Y.: Parametrization of the set of regular and superregular stabilizing controllers. In: Proc. 46th IEEE Conference on Decision and Control, New Orleans, pp. 458–463 (2007)
15. Willems, J.C., Yamamoto, Y.: Behaviors defined by rational functions. In: Proc. 45th IEEE Conference on Decision and Control, San Diego, pp. 550–552 (2006)
16. Yamamoto, Y., Willems, J.C.: Path integrals and Bézoutians for pseudorational transfer functions. To appear in Proc. 48th IEEE Conference on Decision and Control, Shanghai (2009)
17. Yamamoto, Y., Willems, J.C.: Behavioral controllability and coprimeness for a class of infinite-dimensional systems. In: Proc. 47th IEEE Conference on Decision and Control, pp. 1513–1518 (2008)

On the Sample Complexity of Probabilistic Analysis and Design Methods

Teodoro Alamo, Roberto Tempo, and Amalia Luque

Abstract. In this paper, we study the sample complexity of probabilistic methods for control of uncertain systems. In particular, we show the role of the binomial distribution for some problems involving analysis and design of robust controllers with finite families. We also address the particular case in which the design problem can be formulated as an uncertain convex optimization problem. The results of the paper provide simple explicit sample bounds to guarantee that the obtained solutions meet some pre-specified probabilistic specifications.

This paper is dedicated to Yutaka Yamamoto on the occasion of his 60th birthday.

Keywords: randomized algorithms, probabilistic robustness, uncertain systems, sample complexity.

1 A Randomized Approach to Analysis and Design of Control Systems

In recent years, research on probabilistic analysis and design methods for systems and control has significantly progressed. Specific areas where we have seen convincing developments include uncertain and hybrid systems [11], [13]. A key technical ingredient of this approach is the use of the theory of rare events and large deviation inequalities which suitably bound the tail of the probability distribution. These inequalities are crucial in the area of Statistical Learning Theory [12, 13]. The use of this theory for feedback design of uncertain systems has been initiated

Teodoro Alamo and Amalia Luque
Departamento de Ingeniería de Sistemas y Automática, Universidad de Sevilla,
Escuela Superior de Ingenieros, Camino de los Descubrimientos s/n. 41092 Spain
e-mail: {alamo,amalia}@cartuja.us.es

Roberto Tempo
IEIIT-CNR, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy
e-mail: roberto.tempo@polito.it

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 39–50.
springerlink.com

in [13]. Recently, significant improvements regarding the sample complexity have been provided in [1]. For the special case of convex optimization problems, the scenario approach has been introduced in [3] for probabilistic controller design.

In this section we first introduce some preliminary notation and definitions as well as two randomized strategies. In Section 2 we provide bounds for the binomial distribution which are used in Section 3 to analyze the probabilistic properties of different schemes involving randomization. The paper draws to a close in Section 4.

We assume that a probability measure $\mathrm{Pr}_{\mathscr{W}}$ over the sample space $\mathscr{W}$ is given. Given $\mathscr{W}$, a collection of N independent identically distributed (i.i.d.) samples $\mathrm{w} = \{w^{(1)},\dots,w^{(N)}\}$ drawn from $\mathscr{W}$ is said to belong to the Cartesian product $\mathscr{W}^N = \mathscr{W} \times \cdots \times \mathscr{W}$ (N times). Moreover, if the collection w of N i.i.d. samples $\{w^{(1)},\dots,w^{(N)}\}$ is generated from $\mathscr{W}$ according to the probability measure $\mathrm{Pr}_{\mathscr{W}}$, then the *multisample* w is drawn according to the probability measure $\mathrm{Pr}_{\mathscr{W}^N}$. The scalars $\eta \in (0,1)$ and $\delta \in (0,1)$ denote probabilistic parameters. Furthermore, $\ln(\cdot)$ is the natural logarithm and e is the Euler number. For $x \in \mathrm{I\!R}$, $x > 0$, $\lfloor x \rfloor$ denotes the largest integer smaller than or equal to x.

Typically, for a robustness problem, the design parameters, along with different auxiliary variables, are parameterized by means of a decision variable vector θ, which is denoted as "design parameter", and is restricted to a set Θ. On the other hand, the uncertainty w is bounded in the set $\mathscr{W}$. That is, each element $w \in \mathscr{W}$ represents one of the admissible uncertainty realizations. We also consider a binary measurable function $g : \Theta \times \mathscr{W} \to \{0,1\}$ and a real measurable function $f : \Theta \times \mathscr{W} \to \mathrm{I\!R}$ which serve to formulate the specific design problem under attention. In a control context, the binary function $g : \Theta \times \mathscr{W} \to \{0,1\}$, is defined as

$$g(\theta,w) := \begin{cases} 0 & \text{if } \theta \text{ meets control specifications for } w \\ 1 & \text{otherwise.} \end{cases}$$

Given $\theta \in \Theta$, there might be a subset of the elements of $\mathscr{W}$ for which the constraint $g(\theta,w) = 0$ is not satisfied. This concept is rigorously formalized by means of the notion of "probability of violation" which is now introduced.

Definition 1 (probability of violation). Consider a probability measure $\mathrm{Pr}_{\mathscr{W}}$ over $\mathscr{W}$ and let $\theta \in \Theta$ be given. The probability of violation of θ for the function $g : \Theta \times \mathscr{W} \to \{0,1\}$ is defined as

$$E(\theta) := \mathrm{Pr}_{\mathscr{W}}\,\{\, w \in \mathscr{W} \;:\; g(\theta,w) = 1 \,\}.$$

Given $\theta \in \Theta$, it is generally difficult to obtain the exact value of the probability of violation $E(\theta)$ since this requires the solution of a multiple integral. However, we can approximate its value using the concept of empirical mean. For given $\theta \in \Theta$, the empirical mean of $g(\theta,w)$ with respect to the multisample $\mathrm{w} = \{w^{(1)},\dots,w^{(N)}\}$ is defined as

$$\hat{E}(\theta,\mathrm{w}) := \frac{1}{N}\sum_{i=1}^{N} g(\theta,w^{(i)}).$$

Clearly, the empirical mean $\hat{E}(\theta,\mathrm{w})$ is a random variable. Since $g(\cdot,\cdot)$ is a binary function, $\hat{E}(\theta,\mathrm{w})$ is always within the closed interval $[0,1]$.

The utility of randomized algorithms stems from the fact they can circumvent the complexity of nonconvex design problems of the type

$$\min_{\theta\in\Theta} J(\theta) \ \text{ subject to } g(\theta,w)=0, \text{ for all } w\in\mathscr{W} \tag{1}$$

where $J:\Theta\to(-\infty,\infty)$ is a measurable function which normally represents the controller performance. In this setting, one can draw N i.i.d. samples $\{w^{(1)},\ldots,w^{(N)}\}$ from $\mathscr{W}$ according to probability $\mathrm{Pr}_{\mathscr{W}}$ and solve the sampled optimization problem

$$\min_{\theta\in\Theta} J(\theta) \ \text{ subject to } g(\theta,w^{(\ell)})=0,\ \ell=1,\ldots,N. \tag{2}$$

Since obtaining a global solution to the previous problem is a difficult task in the general case, we analyze in this paper the probabilistic properties of any suboptimal feasible solution. If one allows at most m violations of the N constraints, the following sampled problem can be used to obtain a probabilistic relaxation to the original problem (1)

$$\min_{\theta\in\Theta} J(\theta) \ \text{ subject to } \sum_{\ell=1}^{N} g(\theta,w^{(\ell)})\leq m. \tag{3}$$

The idea of allowing some violations of the constraints is not new and can be found, for example, in the context of identification [2]. The randomized strategies corresponding to problems (2) and (3) have been recently studied in [1], see also [11, 13]. In order to analyze the probabilistic properties of any feasible solution to problem (3), we introduce the following definition.

Definition 2 (probability of failure). Given N, $\eta\in(0,1)$, the integer m where $0\leq m\leq N$ and $g:\Theta\times\mathscr{W}\to\{0,1\}$, the probability of failure, denoted by $p(N,\eta,m)$ is defined as

$$\begin{aligned} p(N,\eta,m) := \mathrm{Pr}_{\mathscr{W}^N}\{\mathrm{w}\in\mathscr{W}^N \ :\ &\text{There exists } \theta\in\Theta \\ &\text{such that } \hat{E}(\theta,\mathrm{w})\leq\frac{m}{N} \text{ and } E(\theta)>\eta\}. \end{aligned}$$

We remark that the probability of failure is slightly different from the probability of one-sided constrained failure introduced in [1]. Therefore, if the probability of failure $p(N,\eta,m)$ is no greater than δ then every feasible solution $\theta\in\Theta$ to problem (3) satisfies $E(\theta)\leq\eta$ with probability no smaller than $1-\delta$. From a practical point of view, the objective is to obtain explicit expressions yielding a minimum number of samples N such that the inequality $p(N,\eta,m)\leq\delta$ holds.

We notice that the probability of failure can be easily bounded by the binomial distribution if Θ consists of a unique element. That is, if $\Theta=\{\hat{\theta}\}$ is a singleton, then

$$
\begin{aligned}
p(N,\eta,m) &= \mathrm{Pr}_{\mathscr{W}^N}\left\{\mathrm{w}\in\mathscr{W}^N \,:\, \hat{E}(\hat{\theta},\mathrm{w}) \le \frac{m}{N} \text{ and } E(\hat{\theta}) > \eta\right\} \\
&= \mathrm{Pr}_{\mathscr{W}^N}\left\{\mathrm{w}\in\mathscr{W}^N \,:\, \sum_{\ell=1}^{N} g(\hat{\theta},w^{(\ell)}) \le m \text{ and } E(\hat{\theta}) > \eta\right\} \\
&\le \mathrm{Pr}_{\mathscr{W}^N}\left\{\mathrm{w}\in\mathscr{W}^N \,:\, \sum_{\ell=1}^{N} g(\hat{\theta},w^{(\ell)}) \le m \text{ and } E(\hat{\theta}) = \eta\right\} \\
&= \sum_{i=0}^{m} \binom{N}{i} \eta^i (1-\eta)^{N-i}. \qquad (4)
\end{aligned}
$$

On the other hand, if Θ consists of an infinite number of elements, a deeper analysis involving Statistical Learning Theory is needed [11], [13]. In Subsection 3.3 of this paper, we address this problem under the assumption that Θ consists of a finite number of elements.

In Subsection 3.4 we study the probabilistic properties of the optimal solution of problem (2) under the assumption that $g(\theta,w)=0$ is equivalent to $f(\theta,w)\le 0$, where $f:\Theta\times\mathscr{W}\to\mathbb{R}$ is a convex function with respect to θ in Θ. In this case the result is not expressed in terms of probability of failure because it applies only to the optimal solution of problem (2), and not to every feasible solution.

2 Explicit Sample Size Bounds for the Binomial Distribution

Given a positive integer N and a nonnegative integer m, $m\le N$, and $\eta\in(0,1)$, the binomial distribution is given by

$$B(N,\eta,m) := \sum_{i=0}^{m} \binom{N}{i} \eta^i (1-\eta)^{N-i}.$$

The problem we address in this section is the explicit computation of the *sample complexity*, i.e. a function $\tilde{N}(\eta,m,\delta)$ such that the inequality $B(N,\eta,m)\le\delta$ holds for any $N\ge\tilde{N}(\eta,m,\delta)$, where $\delta\in(0,1)$. As it will be illustrated in the following section, the inequality $B(N,\eta,m)\le\delta$ plays a fundamental role in probabilistic analysis and design methods. Although some explicit expressions are available, e.g. the multiplicative and additive forms of Chernoff bound [5], the results obtained in this paper are tuned on the specific inequalities stemming from the control problems described in Section 3. Because of space limitations reasons, the proofs of the statements of this section are not included. The proofs are given in a technical report, which is available upon request.

The following technical lemma provides an upper bound for the binomial distribution $B(N,\eta,m)$.

Lemma 1. *Suppose that $\eta>0$ and that the nonnegative integers m and the positive integer N satisfy $m\le N$. Then,*

$$B(N,\eta,m) = \sum_{i=0}^{m} \binom{N}{i} \eta^i (1-\eta)^{N-i} \leq a^m \left(\frac{\eta}{a} + 1 - \eta\right)^N, \ \forall a \geq 1.$$

We notice that each particular choice of $a \geq 1$ provides an upper bound for $B(N,\eta,m)$. When using Lemma 1 to obtain a given sample complexity result, the chosen value for a plays a significant role.

Lemma 2. *Given $\delta \in (0,1)$ and the nonnegative integer m, suppose that the integer N and the scalars $\eta \in (0,1)$ and $a > 1$ satisfy the inequality*

$$N \geq \frac{1}{\eta}\left(\frac{a}{a-1}\right)\left(\ln\frac{1}{\delta} + m\ln a\right). \tag{5}$$

Then, $m \leq N$ and

$$B(N,\eta,m) = \sum_{i=0}^{m} \binom{N}{i} \eta^i (1-\eta)^{N-i} \leq \delta.$$

Obviously, the best sample size bound is obtained taking the infimum with respect to $a > 1$. However, a suboptimal value easily follows setting a equal to the Euler constant, which yields the sample size bound

$$N \geq \frac{1}{\eta}\left(\frac{\mathrm{e}}{\mathrm{e}-1}\right)\left(\ln\frac{1}{\delta} + m\right).$$

Since $\frac{\mathrm{e}}{\mathrm{e}-1} < 1.59$, we obtain $N \geq \frac{1.59}{\eta}\left(\ln\frac{1}{\delta} + m\right)$. If $m > 0$ then the choice $a = 1 + \frac{\ln\frac{1}{\delta}}{m} + \sqrt{2\frac{\ln\frac{1}{\delta}}{m}}$ provides a less conservative bound (which is very close to the optimal one based on extensive numerical experiments) at the price of a more involved expression.

Corollary 1. *Given $\delta \in (0,1)$ and the nonnegative integer m, suppose that the integer N and the scalar $\eta \in (0,1)$ satisfy the inequality*

$$N \geq \frac{1}{\eta}\left(m + \ln\frac{1}{\delta} + \sqrt{2m\ln\frac{1}{\delta}}\right). \tag{6}$$

Then,

$$B(N,\eta,m) = \sum_{i=0}^{m} \binom{N}{i} \eta^i (1-\eta)^{N-i} \leq \delta. \tag{7}$$

This corollary improves upon the explicit expression obtained when using the multiplicative form of the Chernoff bound [11], which turns out to be

$$N \geq \frac{1}{\eta}\left(m + \ln\frac{1}{\delta} + \sqrt{\left(\ln\frac{1}{\delta}\right)^2 + 2m\ln\frac{1}{\delta}}\right).$$

3 Sample Complexity for Probabilistic Analysis and Design

We now illustrate some control problems in the context of randomized algorithms where one encounters inequalities of the form

$$B(N,\eta,m) = \sum_{i=0}^{m} \binom{N}{i} \eta^i (1-\eta)^{N-i} \leq \delta.$$

In particular we show how the results of the previous section can be used to obtain explicit sample size bounds guaranteeing that the probabilistic solutions resulting from different randomized approaches meet some pre-specified probabilistic properties.

3.1 Worst Case Performance Analysis

We recall here a result presented in [10] for the probabilistic worst case performance analysis.

Theorem 1. *Suppose that given function* $f : \Theta \times \mathscr{W} \to \mathrm{IR}$*, and* $\hat{\theta} \in \Theta$*, the multisample* $\mathrm{w} = \{w^{(1)},\ldots,w^{(N)}\}$ *is drawn from* $\mathscr{W}^N$ *according to probability* $\mathrm{Pr}_{\mathscr{W}^N}$*. Suppose also that*

$$\gamma = \max_{\ell=1,\ldots,N} f(\hat{\theta}, w^{(\ell)}).$$

If

$$N \geq \frac{\ln \frac{1}{\delta}}{\ln \frac{1}{1-\eta}},$$

then $\mathrm{Pr}_{\mathscr{W}}\{w \in \mathscr{W} \,:\, f(\hat{\theta},w) > \gamma\} \leq \eta$ *with probability no smaller than* $1-\delta$*.*

The proof of this statement, that can be found in [10], relies on the fact that $\mathrm{Pr}_{\mathscr{W}}\{w \in \mathscr{W} \,:\, f(\hat{\theta},w) > \gamma\} \leq \eta$ with probability no smaller than $1-(1-\eta)^N$. Therefore, it suffices to take N such that $B(N,\eta,0) = (1-\eta)^N \leq \delta$.

3.2 Analysis of the Probability of Violation

In the following theorem we provide a sample complexity result that characterizes how the empirical mean converges in probability to the true probability of violation.

Theorem 2. *Given* $\hat{\theta} \in \Theta$*,* ρ*,* η *with* $0 \leq \rho < \eta < 1$ *and* $\delta \in (0,1)$*, if*

$$N \geq \frac{\ln \frac{1}{\delta}}{(\sqrt{\eta} - \sqrt{\rho})^2}$$

then $\mathrm{Pr}_{\mathscr{W}^N}\{\mathrm{w} \in \mathscr{W}^N \,:\, \hat{E}(\hat{\theta},\mathrm{w}) \leq \rho$ *and* $E(\hat{\theta}) > \eta\} \leq \delta$*.*

Proof. We notice that

$$\mathrm{Pr}_{\mathscr{W}^N}\{\mathrm{w} \in \mathscr{W}^N \,:\, \hat{E}(\hat{\theta},\mathrm{w}) \leq \rho \text{ and } E(\hat{\theta}) > \eta\} =$$

$$\mathrm{Pr}_{\mathscr{W}^N}\left\{\mathrm{w} \in \mathscr{W}^N \,:\, \hat{E}(\hat{\theta},\mathrm{w}) \leq \frac{\lfloor \rho N \rfloor}{N} \text{ and } E(\hat{\theta}) > \eta\right\} \leq B(N,\eta,\lfloor \rho N \rfloor).$$

Therefore it suffices to show that the proposed sample size bound guarantees $B(N,\eta,\lfloor \rho N \rfloor) \leq \delta$. Using Corollary 1 and taking into account that $\rho N \geq \lfloor \rho N \rfloor$ we obtain that this is in fact the case if

$$\begin{aligned} N &\geq \frac{1}{\eta}\left(\ln\frac{1}{\delta} + \rho N + \sqrt{2\rho N \ln\frac{1}{\delta}}\right) \\ &= \frac{1}{\eta}\left(\sqrt{\ln\frac{1}{\delta}} + \sqrt{\rho N}\right)^2 - \frac{2-\sqrt{2}}{\eta}\left(\sqrt{\rho N \ln\frac{1}{\delta}}\right). \end{aligned}$$

This inequality is satisfied if

$$N \geq \frac{1}{\eta}\left(\sqrt{\ln\frac{1}{\delta}} + \sqrt{\rho N}\right)^2.$$

Equivalently, $\left(\sqrt{\eta} - \sqrt{\rho}\right)\sqrt{N} \geq \sqrt{\ln\frac{1}{\delta}}$ which yields $N \geq \frac{\ln\frac{1}{\delta}}{(\sqrt{\eta}-\sqrt{\rho})^2}$. □

For small values of $\gamma = \frac{\rho}{\eta}$, the obtained sample size using Lemma 2 is

$$\frac{\ln\frac{1}{\delta}}{\eta(1-\sqrt{\gamma})^2} \approx \frac{\ln\frac{1}{\delta}}{\eta}.$$

This bound is significantly better (for small values of η and γ) than that corresponding to the additive form of the Chernoff bound [5], which for this case has a sample complexity given by

$$\frac{\ln\frac{1}{\delta}}{2(\eta-\rho)^2} = \frac{\ln\frac{1}{\delta}}{2\eta^2(1-\gamma)^2} \approx \frac{\ln\frac{1}{\delta}}{2\eta^2}.$$

On the other hand, the multiplicative form of the Chernoff bound [11] provides the sample size bound

$$\frac{2\eta\ln\frac{1}{\delta}}{(\eta-\rho)^2} = \frac{2\ln\frac{1}{\delta}}{\eta(1-\gamma)^2}$$

which is worse than that provided by Theorem 2 for small values of $\gamma = \frac{\rho}{\eta}$. Finally, we remark that the bound presented in Theorem 2 can be also obtained by means of a result stated in [9], which is the so-called Okamoto bound.

3.3 Finite Families for Design

We consider here the nonconvex sampled problem (3) for the case when Θ consists of a set of finite cardinality n_C. As a motivation consider the case when, after an appropriate normalization procedure, the design parameter set is rewritten as $\hat{\Theta} = \{ \theta \in \mathbb{R}^{n_\theta} \ : \ \|\theta\|_\infty \leq 1 \}$. Suppose also that a gridding approach is adopted. For each component θ_j, $j = 1, \ldots, n_\theta$ of the design parameters $\theta \in \mathbb{R}^{n_\theta}$, only n_{C_j} equally spaced values are considered. That is, θ_j is constrained into the set $\Upsilon_j = \{ -1 + \frac{2(t-1)}{(n_{C_j}-1)} \ : \ t = 1, \ldots, n_{C_j} \}$. With this gridding, the following finite cardinality set $\Theta = \{ [\theta_1, \ldots, \theta_{n_\theta}]^\top \ : \ \theta_j \in \Upsilon_j, \ j = 1, \ldots, n_\theta \}$ is obtained. We notice that the cardinality of the set is $n_C = \prod_{j=1}^{n_\theta} n_{C_j}$. Another situation in which the finite cardinality assumption holds is when a finite number of random samples in the space of design parameter are drawn according to a given probability, see e.g. [6, 7, 14].

The following property states the relation between the binomial distribution and the probability of failure under this finite cardinality assumption.

Lemma 3. *Suppose that the cardinality of Θ is no larger than n_C. Then,*

$$p(N, \eta, m) \leq n_C \sum_{i=0}^{m} \binom{N}{i} \eta^i (1-\eta)^{N-i} = n_C B(N, \eta, m).$$

Proof. Denote $\tilde{n}_C \leq n_C$ the cardinality of Θ. Therefore, Θ can be rewritten as $\Theta = \{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(\tilde{n}_C)}\}$. Then,

$$\begin{aligned} p(N, \eta, m) &= \Pr_{\mathscr{W}^N}\{\mathrm{w} \in \mathscr{W}^N \ : \ \text{There exists } \theta \in \Theta \\ &\qquad \text{such that } \hat{E}(\theta, \mathrm{w}) \leq \frac{m}{N} \text{ and } E(\theta) > \eta\} \\ &\leq \sum_{j=1}^{\tilde{n}_C} \Pr_{\mathscr{W}^N}\{\mathrm{w} \in \mathscr{W}^N \ : \ \hat{E}(\theta^{(j)}, \mathrm{w}) \leq \frac{m}{N} \text{ and } E(\theta^{(j)}) > \eta\} \\ &\leq \tilde{n}_C \sum_{i=0}^{m} \binom{N}{i} \eta^i (1-\eta)^{N-i} \leq n_C \sum_{i=0}^{m} \binom{N}{i} \eta^i (1-\eta)^{N-i}. \qquad \square \end{aligned}$$

Consider now the optimization problem (3). It follows from Lemma 3 that in order to guarantee that every feasible solution $\hat{\theta} \in \Theta$ satisfies $E(\hat{\theta}) \leq \eta$ with probability no smaller than $1 - \delta$, it suffices to take N such that $n_C B(N, \eta, m) \leq \delta$, where n_C is an upper bound on the cardinality of Θ. As it will be shown next, the required sample complexity in this case grows with the natural logarithm of n_C. This means that we can consider finite families with high cardinality and still obtain reasonable sample complexity bounds.

Theorem 3. *Suppose that the cardinality of Θ is no larger than n_C. Given the nonnegative integer m, $\eta \in (0,1)$ and $\delta \in (0,1)$, if*

$$N \geq \inf_{a>1} \frac{1}{\eta} \left(\frac{a}{a-1} \right) \left(\ln \frac{n_C}{\delta} + m \ln a \right)$$

then $p(N,\eta,m) \leq \delta$. *Moreover, if*

$$N \geq \frac{1}{\eta}\left(m + \ln\frac{n_C}{\delta} + \sqrt{2m\ln\frac{n_C}{\delta}}\right)$$

then $p(N,\eta,m) \leq \delta$.

Proof. From Theorem 3 we have that $p(N,\eta,m) \leq \delta$ provided that $B(N,\eta,m) \leq \frac{\delta}{n_C}$. The two claims of the property now follow directly from Lemma 2 and Corollary 1 respectively. □

We remark that taking a equal to the Euler constant, the following sample size bound

$$N \geq \frac{1}{\eta}\left(\frac{\mathrm{e}}{\mathrm{e}-1}\right)\left(\ln\frac{n_C}{\delta} + m\right)$$

is obtained from Theorem 3. If $m > 0$ then a suboptimal value for a is given by

$$a = 1 + \frac{\ln\frac{n_C}{\delta}}{m} + \sqrt{2\frac{\ln\frac{n_C}{\delta}}{m}}.$$

Suppose now that a multisample w is drawn from $\mathscr{W}^N$ according to probability $\Pr_{\mathscr{W}^N}$ and that $0 \leq \rho < \eta < 1$. Then, every feasible solution θ of the optimization problem

$$\min_{\theta\in\Theta} J(\theta) \text{ subject to } \hat{E}(\theta,\mathrm{w}) \leq \rho \tag{8}$$

satisfies $E(\theta) \leq \eta$, with probability no smaller than $1-\delta$, provided that

$$N \geq \frac{\ln\frac{n_C}{\delta}}{(\sqrt{\eta}-\sqrt{\rho})^2}$$

and that the cardinality of Θ is not larger than n_C. The proof of this statement follows the same lines as the proof of Theorem 3 and it is not included here because of space limitations. This result improves upon a similar one presented in [8], in which the sample size N grows as $\frac{\ln\frac{n_C}{\delta}}{2(\eta-\rho)^2}$.

3.4 Optimal Robust Optimization for Design

In this subsection, we study the so-called scenario approach for robust control introduced in [3], see also [4] for recent results in this area. Suppose that in order to address the general semi-infinite optimization problem (1), one resorts to randomization. That is, N i.i.d. samples $\{w^{(1)},\ldots,w^{(N)}\}$ from $\mathscr{W}$ according to probability $\Pr_{\mathscr{W}}$ are drawn and one solves the following problem

$$\min_{\theta\in\Theta} J(\theta) \;\text{ subject to } g(\theta, w^{(\ell)}) = 0,\ \ell = 1,\ldots,N. \tag{9}$$

We consider here the particular case in which $J(\theta) = c^\top \theta$, the constraint $g(\theta,w) = 0$ is convex in Θ for all $w \in W$, the solution of (9) is unique[1]. These assumptions are now stated precisely.

Assumption 1 (convexity). *Let $\Theta \subset \mathbb{R}^{n_\theta}$ be a convex and closed set. We assume that*

$$J(\theta) := c^\top \theta \quad \text{and} \quad g(\theta,w) := \begin{cases} 0 \text{ if } f(\theta,w) \le 0, \\ 1 \text{ otherwise} \end{cases}$$

where $f : \Theta \times \mathscr{W} \to [-\infty,\infty]$ is convex in Θ for any fixed value of $w \in \mathscr{W}$.

Assumption 2 (feasibility and uniqueness). *The optimization problem (9), for all possible multisample extractions $\{w^{(1)}, \ldots, w^{(N)}\}$, is always feasible and attains a unique optimal solution. Moreover, its feasibility domain has a nonempty interior.*

We state here a result proved in [4] that relates the binomial distribution to the probabilistic properties of the optimal solution obtained from (9).

Lemma 4. *Let Assumptions* 1 *and* 2 *hold. Suppose that N, $\eta \in (0,1)$ and $\delta \in (0,1)$ satisfy the following inequality*

$$\sum_{i=0}^{n_\theta - 1} \binom{N}{i} \eta^i (1-\eta)^{N-i} \le \delta. \tag{10}$$

Then, with probability no smaller than $1-\delta$, the optimal solution $\hat{\theta}_N$ to the optimization problem (9) *satisfies the inequality $E(\hat{\theta}_N) \le \eta$.*

We now state an explicit sample size bound to guarantee that the probability of violation is smaller than η with probability at least $1-\delta$.

Theorem 4. *Let Assumptions* 1 *and* 2 *hold. Given $\eta \in (0,1)$ and $\delta \in (0,1)$, if*

$$N \ge \inf_{a>1} \left(\frac{a}{\eta(a-1)}\right)\left(\ln\frac{1}{\delta} + (n_\theta - 1)\ln a\right) \tag{11}$$

or

$$N \ge \frac{1}{\eta}\left(\ln\left(\frac{1}{\delta}\right) + n_\theta - 1 + \sqrt{2(n_\theta - 1)\ln\frac{1}{\delta}}\right) \tag{12}$$

then, with probability no smaller than $1-\delta$, the optimal solution $\hat{\theta}_N$ to the optimization problem (9) *satisfies the inequality $E(\hat{\theta}_N) \le \eta$.*

Proof. From Lemma 4 it follows that it suffices to take N such that $B(N,\eta,n_\theta - 1) \le \delta$. Both inequalities (11) and (12) guarantee that $B(N,\eta,n_\theta - 1) \le \delta$ (see Lemma 2 and Corollary 1 respectively). This completes the proof. □

[1] We remark that this uniqueness assumption can be relaxed in most cases, as shown in Appendix A of [3].

We remark that a sample size bound which depends linearly on $\frac{1}{\eta}$ is obtained taking a equal to the Euler constant

$$N \geq \frac{1}{\eta}\left(\frac{\mathrm{e}}{\mathrm{e}-1}\right)\left(\ln\frac{1}{\delta}+n_\theta-1\right).$$

This bound always improves upon other recent bounds given in the literature, see e.g. [1]. If $n_\theta > 1$ then a suboptimal value for a is given by

$$a = 1+\frac{\ln\frac{1}{\delta}}{n_\theta-1}+\sqrt{2\frac{\ln\frac{1}{\delta}}{n_\theta-1}}.$$

4 Conclusion

In this paper we have derived sample complexity results for various analysis and design problems related to uncertain systems. In particular we provided new results which guarantee that a binomial distribution expression is smaller than a prespecified value. These results are subsequently exploited for the analysis of worst case performance and constraint violation. With regard to design problems we considered the case of finite cardinality of controller families and the special case when the design problem can be recast as a robust convex optimization problem.

Acknowledgement. This work was supported by the MCYT-Spain and the European Commision which funded this work under projects DPI2007-66718-C04-01 and FP7-223866 respectively.

References

1. Alamo, T., Tempo, R., Camacho, E.F.: Randomized strategies for probabilistic solutions of uncertain feasibility and optimization problems. IEEE Transactions on Automatic Control 54(11), 2545–2559 (2009)
2. Bai, E., Cho, H., Tempo, R., Ye, Y.: Optimization with few violated constraints for linear bounded error parameter estimation. IEEE Transactions on Automatic Control 47(7), 1067–1077 (2002)
3. Calafiore, G., Campi, M.C.: The scenario approach to robust control design. IEEE Transactions on Automatic Control 51(5), 742–753 (2006)
4. Campi, M.C., Garatti, S.: The exact feasibility of randomized solutions of robust convex programs. SIAM Journal on Optimization 19, 1211–1230 (2008)
5. Chernoff, H.: A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Annals of Mathematical Statistics 23, 493–507 (1952)
6. Fujisaki, Y., Kozawa, Y.: Probabilistic robust controller design: probable near minimax value and randomized algorithms. In: Calafiore, G., Dabbene, F. (eds.) Probabilistic and Randomized Methods for Design under Uncertainty. Springer, London (2006)
7. Koltchinskii, V., Abdallah, C.T., Ariola, M., Dorato, P., Panchenko, D.: Improved sample complexity estimates for statistical learning control of uncertain systems. IEEE Transactions on Automatic Control 45(12), 2383–2388 (2000)

8. Luedtke, J., Ahmed, S.: A sample approximation approach for optimization with probabilistic constraints. SIAM Journal on Optimization 19(2), 674–699 (2008)
9. Okamoto, M.: Some inequalities relating to the partial sum of binomial probabilities. Annals of the Institute of Statistical Mathematics 10(1), 29–35 (1959)
10. Tempo, R., Bai, E.-W., Dabbene, F.: Probabilistic robustness analysis: explicit bounds for the minimum number of samples. Systems & Control Letters 30, 237–242 (1997)
11. Tempo, R., Calafiore, G., Dabbene, F.: Randomized Algorithms for Analysis and Control of Uncertain Systems. Communications and Control Engineering Series. Springer, London (2005)
12. Vapnik, V.N.: Statistical Learning Theory. John Wiley & Sons, New York (1998)
13. Vidyasagar, M.: A Theory of Learning and Generalization: With Applications to Neural Networks and Control Systems. Springer, London (1997)
14. Vidyasagar, M.: Randomized algorithms for robust controller synthesis using statistical learning theory. Automatica 37, 1515–1528 (2001)

A Constant Factor Approximation Algorithm for Event-Based Sampling

Randy Cogill, Sanjay Lall, and João P. Hespanha

Abstract. We consider a control system in which sensor data is transmitted from the plant to a receiver over a communication channel, and the receiver uses the data to estimate the state of the plant. Using a feedback policy to choose when to transmit data, the goal is to schedule transmissions to balance a trade-off between communication rate and estimation error. Computing an optimal policy for this problem is generally computationally intensive. Here we provide a simple algorithm for computing a suboptimal policy for scheduling state transmissions which incurs a cost within a factor of six of the optimal achievable cost.

1 Introduction

We consider a control system in which sensor data is transmitted from the plant to a receiver over a communication channel, and the receiver uses the data to estimate the state of the plant. Sending data more frequently leads to increased use of limited communication resources, but also allows the average estimation error to be reduced. Conversely, of course we may reduce the use of the channel if we are willing to allow larger estimation errors.

Randy Cogill
Department of Systems and Information Engineering
University of Virginia, Charlottesville, VA 22904, U.S.A.
e-mail: rcogill@virginia.edu

Sanjay Lall
Department of Aeronautics and Astronautics
Stanford University, Stanford, CA 94305, U.S.A.
e-mail: lall@stanford.edu

João P. Hespanha
Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106, U.S.A.
e-mail: hespanha@ece.ucsb.edu

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 51–60.
springerlink.com

We consider feedback policies for choosing when to transmit data. That is, instead of simply choosing a transmission rate, at the plant measurements are used to decide whether to transmit data to the controller. This type of measurement is called *Lebesgue* or *event-based* sampling in [2]. Several other authors have considered both control and filtering problems using such sampling schemes, in particular [2, 7, 8, 9, 18, 26, 28].

The plant is modeled by a discrete-time linear system, and at each time step the channel allows exact transmission of the state. The cost function of interest in this problem is a weighted sum of the estimation error and the transmission rate. The optimal controller for a given weight then lies on the Pareto-optimal trade-off curve, and choosing the weight allows one to select the trade-off between rate and error.

For this cost function, the problem of finding the optimal policy was considered in [27], where the authors show that the problem of computing an optimal scheduling policy can be addressed in the framework of Markov decision processes, and consequently the value iteration algorithm can be used to compute an optimal policy. Although this provides an algorithm for computing an optimal policy, the computation required to compute such a policy quickly becomes prohibitive as the system's state dimension increases.

Since the optimal policy is very difficult to compute, we consider *approximately optimal* policies. Specifically, the main result of this paper is to give a simple algorithm for computing a policy, and show that this policy is guaranteed to achieve a cost within a factor of six of the optimal achievable cost. This result is Theorem 1 below.

Approximation algorithms have been widely used for addressing computationally intractable problems. While some NP-hard problems may be approximated to arbitrary accuracy, others may not be approximated within any constant factor. It is therefore extremely promising that the particular problem of rate-error trade-off considered in this paper is approximable within a constant factor of six. It is not currently known whether policies achieving better approximation ratios may be efficiently obtained.

Finally, due to space constraints, all proofs have been omitted from this paper. Proofs of the theorems in this paper can be found in [6].

2 Problem Formulation

Here we will present the problem that will be considered throughout this paper. In the following subsection, it will be shown how this problem is a generalization of the problem of networked estimation.

We have dynamics

$$e_{t+1} = (1 - a_t)Ae_t + w_t \qquad e_0 = 0, \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$, and for each $t \in \mathbb{N}$ the state is $e_t \in \mathbb{R}^n$ and the action is $a_t \in \{0,1\}$. Here $w_0, w_1, \dots$ is a sequence of independent identically distributed Gaussian random vectors, with $w_t \sim \mathscr{N}(0,\Sigma)$, where $\Sigma \succ 0$. Define the function $r : \mathbb{R}^n \times \{0,1\} \to \mathbb{R}$ to be the cost at time t, given by

$$r(e_t, a_t) = (1-a_t)e_t^T Q e_t + \lambda a_t \tag{2}$$

where $Q \succ 0$ and $\lambda > 0$. We would like to choose a state-feedback control policy $\mu : \mathbb{R}^n \to \{0,1\}$ to make the average cost incurred by the policy μ small. Here the average cost J is defined as

$$J(\mu) = \limsup_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} \mathbf{E}\big(r(e_t, \mu(e_t))\big) \tag{3}$$

See [1] for background on this choice of cost. Here, each a_t is determined according to the static state-feedback policy $a_t = \mu(e_t)$, and then the sequence $e_0, e_1, \dots$ is a Markov process. Therefore, the problem of choosing a policy which minimizes the cost J is can be addressed using the theory of Markov decision processes. The cost J given by equation (3) is called the *average per-period cost*, and we focus specifically on the problem of choosing a policy to minimize this. For convenience, define the space of policies

$$\mathscr{P} = \{ f : \mathbb{R}^n \to \{0,1\} \mid f \text{ is measurable} \}$$

Then the above problem can be stated as follows.

Problem 1 (Rate-Error Trade-off). Given A, $\Sigma \succ 0$, $Q \succ 0$, $\lambda > 0$, and $\gamma > 0$, find a state feedback policy $\mu \in \mathscr{P}$ such that

$$J(\mu) \le \gamma$$

Minimizing the cost J balances a trade-off between the average size of e_t, as measured by the quadratic form defined by Q, and the frequency with which e_t is reset to the level of the noise by setting $a_t = 1$. The problem of computing an optimal policy was considered in [27], and a numerical procedure for finding such a policy was given. However, the computation required to compute an optimal policy increases rapidly with the state dimension. In the following section we present an easily computable and easily implementable policy for this problem which incurs a cost within a provable bound of the optimal achievable cost. Specifically, we focus our attention on the set of problem instances where Q and A are such that $A^T Q A - Q \preceq 0$ and $Q \succ 0$. In particular, this implies that $\rho(A) \le 1$ and the system is therefore at least marginally stable. We show that in this case there is a simple policy which always achieves a cost within a factor of six of the optimal cost. It is worth noting that, in general, both the policy which always transmits and the policy which never transmits may achieve cost arbitrarily far from optimal.

2.1 Application to Networked Estimation

Suppose we have the dynamical system

$$\begin{aligned} x_{t+1} &= Ax_t + w_t \quad x_0 = 0 \\ y_t &= a_t x_t \end{aligned}$$

where for each $t \in \mathbb{N}$ the state $x_t \in \mathbb{R}^n$ and $a_t \in \{0,1\}$. As above, $w_0, w_1, \ldots$ is a sequence of independent identically distributed zero mean Gaussian random vectors with covariance $\Sigma \succ 0$. We have a per-period cost of

$$c(x_t, a_t, b_t) = (1 - a_t)(x_t - b_t)^T Q(x_t - b_t) + \lambda a_t \tag{4}$$

and we would like to choose two controllers. The first is the function $\mu : \mathbb{R}^n \to \{0,1\}$, and the second is the sequence of functions ϕ_t indexed by t where $\phi_t : \{0,1\}^t \times \mathbb{R}^{nt} \to \mathbb{R}^n$. These are connected according to

$$\begin{aligned} a_t &= \mu(x_t) \\ b_t &= \phi_t(a_0, \ldots, a_{t-1}, y_0, \ldots, y_{t-1}) \end{aligned}$$

Again, we are interested in the cost

$$J(\mu, \phi_0, \phi_1, \ldots) = \limsup_{N \to \infty} \frac{1}{N} \sum_{t=0}^{N-1} \mathbf{E}\big(r(x_t, a_t, b_t)\big)$$

The interpretation is shown in Figure 1, where the linear dynamics $x_{t+1} = Ax_t + w_t$ is denoted by G. The dashed lines indicate a communication channel. At each time t the transmitter μ chooses whether to transmit the signal x_t to the receiver ϕ. Each transmission costs λ. The receiver would like to estimate the state x_t of G, and choose b_t to minimize the error $x_t - b_t$ as measured by the quadratic form Q. The cost r is used to compute the trade-off, parametrized by λ, of estimation error against frequency of transmissions.

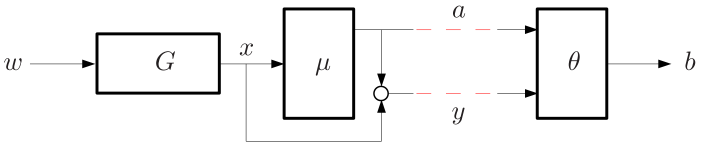


Fig. 1 Networked Estimation.

The estimator ϕ considered in Xu and Hespanha [27] is as follows. Let $b_t = \phi_t(a_0, \ldots, a_{t-1}, y_0, \ldots, y_{t-1})$, and define ϕ by the realization

$$b_{t+1} = (1 - a_t)Ab_t + a_t Ax_t \quad b_0 = 0$$

If the random variables $a_0, a_1, \ldots$ are independent of $x_0, x_1, \ldots$ then this is the time-varying Kalman filter, and b_t is the minimum mean square error estimate of x_t given measurements $y_0, \ldots, y_{t-1}$.

We now have the dynamics

$$\begin{bmatrix} x_{t+1} \\ b_{t+1} \end{bmatrix} = \begin{bmatrix} A & 0 \\ a_t A & (1-a_t)A \end{bmatrix} \begin{bmatrix} x_t \\ b_t \end{bmatrix} + \begin{bmatrix} I \\ 0 \end{bmatrix} w_t$$

We change coordinates to

$$\begin{bmatrix} e_t \\ f_t \end{bmatrix} = \begin{bmatrix} I & -I \\ 0 & I \end{bmatrix} \begin{bmatrix} x_t \\ b_t \end{bmatrix}$$

to give

$$\begin{bmatrix} e_{t+1} \\ f_{t+1} \end{bmatrix} = \begin{bmatrix} (1-a_t)A & 0 \\ a_t A & A \end{bmatrix} \begin{bmatrix} e_t \\ f_t \end{bmatrix} + \begin{bmatrix} I \\ 0 \end{bmatrix} w_t$$

In these coordinates, the cost c specified in equation (4) is exactly equal to the cost (2), and e evolves according to the dynamics (1). With this choice of ϕ therefore the optimal choice of μ is found by solving the RATE-ERROR TRADE-OFF problem.

3 Main Results

In this section we present the main result of this paper, which is that for a slightly restricted version of the RATE-ERROR TRADEOFF problem, there is a simple policy which achieves cost within a constant factor of optimal. Define for convenience

$$J_{\text{opt}} = \inf_{\mu \in \mathscr{P}} \left(\liminf_{N \to \infty} \frac{1}{N} \sum_{t=0}^{N-1} \mathbf{E}\big(r(e_t, \mu(e_t))\big) \right)$$

The policy that we consider is a simple quadratic threshold policy. The main result of this paper is as follows.

Theorem 1. *Suppose $A \in \mathbb{R}^{n \times n}$, $Q \succ 0$, $\Sigma \succ 0$, and $A^T QA - Q \preceq 0$. Then there exists a unique matrix $M \in \mathbb{S}^n$ satisfying*

$$\frac{1}{1 + \operatorname{trace}(\Sigma M)} A^T MA - M + \frac{Q}{\lambda} = 0 \tag{5}$$

Furthermore, define the policy μ by

$$\mu(e) = \begin{cases} 0 & \text{if } e^T Me \le 1 \\ 1 & \text{otherwise} \end{cases} \tag{6}$$

For this policy, the cost satisfies

$$J(\mu) \le 6J_{opt} \tag{7}$$

Proof. The result follows immediately from Theorems 2 and 3 which are proved below.

Note that implementation of the policy μ requires an algorithm for computing the unique solution M of equation (5). It is easily shown that this equation can be solved by performing a bisection search and solving a sequence of Lyapunov equations.

4 Bounds for the Communication Cost

4.1 Upper Bounds

We are now ready to the upper bound on $J(\mu)$ is obtained, where μ is the policy in (6). The following lemma provides the upper bound and also shows that one may use semidefinite programming, combined with a line search, to find policies that minimize this upper bound.

Lemma 1. *Suppose $M \succeq 0$ and $H \succeq 0$ are symmetric positive semidefinite matrices, and $\alpha \in \mathbb{R}$. If*

$$\begin{aligned} A^T HA - H + Q - \alpha M &\preceq 0 \\ (\lambda - \alpha)M - H &\preceq 0 \\ \alpha - \lambda &\leq 0 \\ \alpha &\geq 0 \end{aligned} \tag{8}$$

Then the policy

$$\mu(e) = \begin{cases} 0 & \textit{if } e^T Me \leq 1 \\ 1 & \textit{otherwise} \end{cases}$$

achieves a cost which satisfies

$$J(\mu) \leq \operatorname{trace}(\Sigma H) + \alpha$$

We now make use of this result to provide an explicit upper bound.

Theorem 2. *Suppose $A \in \mathbb{R}^{n \times n}$, $Q \succ 0$, $\Sigma \succ 0$ and $A^T QA - Q \preceq 0$. Let M be the unique solution to*

$$\frac{1}{1 + \operatorname{trace}(\Sigma M)} A^T MA - M + Q/\lambda = 0$$

Then the policy

$$\mu(e) = \begin{cases} 0 & \textit{if } e^T Me \leq 1 \\ 1 & \textit{otherwise} \end{cases}$$

achieves

$$J(\mu) \leq \frac{2\lambda \operatorname{trace}(\Sigma M)}{1 + \operatorname{trace}(\Sigma M)}$$

4.2 Lower Bounds

For the class of instances of RATE-ERROR TRADEOFF with A and Q satisfying $A^TQA - Q \preceq 0$, we can show that the policy μ of equation (6) achieves a cost within a constant factor of optimal. To complete the presentation of the main result of this paper, we now determine a lower bound on J_{opt} which guarantees that for this class of instances,

$$J(\mu) \le 6J_{\text{opt}}$$

This result can be established using the lemmas below, the proofs of which can be found in [6].

Lemma 2. *Suppose $Y \succeq 0$ and $q \in \mathbb{R}^n$, and $w \sim \mathscr{N}(0,\Sigma)$ is a Gaussian random vector. Let f be the random variable*

$$f = (q+w)^T Y (q+w)$$

Then

$$\mathbf{E}f = q^T Y q + \text{trace}(\Sigma Y) \tag{9}$$

$$\begin{aligned}\mathbf{E}(f^2) = {} & (q^TYq)^2 + 4q^TY\Sigma Yq + \big(\text{trace}(\Sigma Y)\big)^2 \\ & + 2\,\text{trace}(\Sigma Y\Sigma Y) + 2q^TYq\,\text{trace}(\Sigma Y)\end{aligned} \tag{10}$$

and further

$$\mathbf{E}(f^2) \le (q^TYq)^2 + 6q^TYq\,\text{trace}(\Sigma Y) + 3\big(\text{trace}(\Sigma Y)\big)^2$$

Lemma 3. *Suppose there exists a positive semidefinite matrix $C \succeq 0$ and $s \in \mathbb{R}$ such that*

$$\begin{aligned}\big(s - 6\,\text{trace}(C\Sigma)\big)A^TCA - sC + Q &\succeq 0 \\ s^2 &\le 4\lambda \\ A^TCA - C &\preceq 0\end{aligned} \tag{11}$$

Then for all policies $\mu \in \mathscr{P}$

$$J(\mu) \ge s\,\text{trace}(C\Sigma) - 3\big(\text{trace}(C\Sigma)\big)^2$$

Lemma 4. *Suppose there exists $M \succeq 0$ such that*

$$\begin{aligned}\frac{1}{1+\text{trace}(\Sigma M)}A^TMA - M + Q/\lambda &= 0 \\ A^TMA - M &\preceq 0\end{aligned}$$

Then for all policies $\mu \in \mathscr{P}$ we have

$$J(\mu) \ge \frac{\lambda\,\text{trace}(\Sigma M)}{3\big(1+\text{trace}(\Sigma M)\big)}$$

Lemma 5. *Suppose $Q \succ 0$ and $A^T QA - Q \preceq 0$, and $\alpha \in \mathbb{R}$ satisfies $0 \leq \alpha < 1$. Then there exists a unique $M \in \mathbb{S}^n$ such that*

$$\alpha A^T MA - M + Q = 0 \tag{12}$$

and the matrix M is positive definite and satisfies

$$A^T MA - M \preceq 0$$

Finally, the lemmas above can be combined to obtain the following theorem.

Theorem 3. *Suppose $A \in \mathbb{R}^{n\times n}$, $Q \succ 0$, $\Sigma \succ 0$ and $A^T QA - Q \preceq 0$. Let M be the unique solution to*

$$\frac{1}{1+\mathrm{trace}(\Sigma M)} A^T MA - M + Q/\lambda = 0$$

Then for all policies $\mu \in \mathscr{P}$ we have

$$J(\mu) \geq \frac{\lambda\, \mathrm{trace}(\Sigma M)}{3\big(1+\mathrm{trace}(\Sigma M)\big)}$$

5 Conclusions

In this paper we considered a simple, yet fundamental estimation problem involving balancing the trade-off between communication rate and estimation error in networked linear systems. This paper extended work of [27], where it was shown that this problem can be posed as a Markov decision process. Here we show that there is a simple, easily computable suboptimal policy for scheduling state transmissions which incurs a cost within a factor of six of the optimal achievable cost.

References

1. Arapostathis, A., Borkar, V.S., Fernandez-Gaucherand, E., Ghosh, M.K., Marcus, S.I.: Discrete-time controlled markov processes with average cost criterion: a survey. Optimization 31(2), 282–344 (1993)
2. Åström, K.J., Bernhardsson, B.M.: Comparison of Riemann and Lebesgue sampling for first order stochastic systems. In: Proceedings of the 41st IEEE Conference on Decision and Control, pp. 2011–2016 (2002)
3. Cervin, A., Åström, K.J.: On limit cycles in event-based control systems. In: Proceedings of the 46th IEEE Conference on Decision and Control, pp. 3190–3195 (2007)
4. Cervin, A., Johannesson, E.: Sporadic control of scalar systems with delay, jitter and measurement noise. In: Proceedings of the 17th IFAC World Congress (2008)
5. Cogill, R., Lall, S.: Suboptimality bounds in stochastic control: a queueing example. In: Proceedings of the 2006 American Control Conf., pp. 1642–1647 (2006)

6. Cogill, R., Lall, S., Hespanha, J.P.: A constant factor approximation algorithm for optimal estimation subject to communication costs. In: Proceedings of the 2007 American Control Conference, pp. 305–311 (2007)
7. Hespanha, J.P., Ortega, A., Vasudevan, L.: Towards the control of linear systems with minimum bit-rate. In: Proc. 15th Int. Symp. on the Mathematical Theory of Networks and Systems (August 2002)
8. Hristu-Varsakelis, D., Kumar, P.R.: Interrupt-based feedback control over a shared communication medium. In: Proceedings of the 41st IEEE Conference on Decision and Control, pp. 3223–3228 (2002)
9. Imer, O.C., Başar, T.: Optimal estimation with limited measurements. In: Proceedings of the 44th IEEE Conference on Decision and Control, pp. 1029–1034 (2005)
10. Imer, O.C., Başar, T.: Optimal estimation with scheduled measurements. Appl. Comput. Math. 4(2), 92–101 (2005)
11. Imer, O.C., Başar, T.: Optimal control with limited controls. In: Proceedings of the 2006 American Control Conf., pp. 298–303 (2006)
12. Imer, O.C., Başar, T.: To measure or to control: optimal control with scheduled measurements and controls. In: Proceedings of the 2006 American Control Conf., pp. 1003–1008 (2006)
13. Mazo Jr., M., Tabuada, P.: On event-triggered and self-triggered control over sensor/actuator networks. In: Proceedings of the 47th IEEE Conference on Decision and Control, pp. 435–440 (2008)
14. Kofman, E., Braslavsky, J.: Level crossing sampling in feedback stabilization under data-rate constraints. In: Proceedings of the 45th IEEE Conference on Decision and Control, pp. 4423–4428 (2006)
15. Baras, J.S., Rabi, M., Moustakides, G.V.: Adaptive sampling for linear state estimation. Submitted to SIAM Journal on Control and Optimization (2008)
16. Miskowicz, M.: Bandwidth requirements for event driven observations of continuous time variable. In: Proceedings of the 7th IFAC Workshop on Discrete-Event Systems (WODES 2004), pp. 475–480 (2004)
17. Miskowicz, M.: Efficiency of level-crossing sampling for bandlimited gaussian random processes. In: Proc. 2006 IEEE International Workshop on Factory Communication Systems, pp. 137–142 (2006)
18. Rabi, M., Baras, J.S.: Sampling of diffusion processes for real-time estimation. In: Proceedings of the 43rd IEEE Conference on Decision and Control, pp. 4163–4168 (2004)
19. Rabi, M., Baras, J.S.: Level-triggered control of a scalar linear system. In: Proceedings of the 2007 Mediterranean Conference on Control and Automation (2007)
20. Rabi, M., Baras, J.S., Moustakides, G.: Multiple sampling for estimation on a finite horizon. In: Proceedings of the 45th IEEE Conference on Decision and Control, pp. 1351–1357 (2006)
21. Rabi, M., Johansson, K.H.: Event-triggered strategies for industrial control over wireless networks. In: Proceedings of the 4th Annual Wireless Interned Conference (2008)
22. Tabuada, P.: Event-triggered real-time scheduling of stabilizing control tasks. IEEE Trans. Automatic Control 52(9), 1680–1685 (2007)
23. Triantafyllopoulos, K.: On the central moments of the multidimensional Gaussian distribution. The Mathematical Scientist 28, 125–128 (2003)
24. Wang, X., Lemmon, M.: Self-triggered feedback control systems with finite-gain stability. To appear in IEEE Trans. Automat. Control (2009)

25. Wang, X., Lemmon, M.D.: Event-triggered broadcasting across distributed networked control systems. In: Proceedings of the 2008 American Control Conf., pp. 3139–3144 (2008)
26. Xu, Y., Hespanha, J.P.: Communication logics for networked control systems. In: Proceedings of the 2004 American Control Conf., pp. 572–577 (2004)
27. Xu, Y., Hespanha, J.P.: Optimal communication logics in networked control systems. In: Proceedings of the 43rd IEEE Conference on Decision and Control, pp. 3527–3532 (2004)
28. Yook, J.K., Tilbury, D.M., Soparkar, N.R.: Trading computation for bandwidth: reducing communication in distributed control systems using state estimators. IEEE Trans. Control System Technol. 4(10), 503–518 (2002)

A Unified Approach to Decentralized Cooperative Control for Large-Scale Networked Dynamical Systems

Shinji Hara

Abstract. A class of large-scale networked dynamical systems with decentralized information structures such as multi-agent dynamical systems can be represented by an LTI system with a generalized frequency variable. This paper summarizes recent results on the fundamental analysis, namely stability conditions, $\mathscr{H}_2$-norm computations, and $\mathscr{H}_\infty$-norm conditions, for LTI systems with generalized frequency variables. We also show an analytic condition for the existence of protein level's periodic oscillations in cyclic gene regulatory networks as an application.

1 Introduction

There are many large scale systems in both nature and artificial systems, which consist of a bunch of subsystems interacted each other, and our target dynamical systems in modern engineering have become more and more complex and subject to multitude of system dimensions. To cope with these challenges, many studies of different approaches in a variety of areas have been made in the last decade. One of the bulk flows in these studies is the decentralized cooperative control of the multi-agent dynamical systems (See e.g., [12] and references therein.). There have been many researches in the form of proposing a specific approach within an individual problem formulation, but very few results are available so far to provide a unifying theoretical framework.

This situation motivates us to establish a unified approach for the analysis and synthesis of multi-agent dynamical systems in which agents with dynamics exchange information each other and autonomously cooperate. To this end our research group recently proposed a linear time-invariant (LTI) system with a generalized frequency variable as one of the unifying expressions for multi-agent dynamical systems [3, 4]. Specifically, the transfer function $\mathscr{G}(s)$ representing the overall

Shinji Hara
Department of Information Physics and Computing, Graduate School of Information Science and Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
e-mail: Shinji_Hara@ipc.i.u-tokyo.ac.jp

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 61–72.
springerlink.com

dynamics of a multi-agent system is described by simply replacing s by a rational function $\phi(s)$ in a transfer function $G(s)$, i.e., $\mathscr{G}(s) := G(\phi(s))$. We call $\phi(s)$ the generalized frequency variable, because s in a continuous-time transfer function represents the frequency variable. The system description has a potential to provide a theoretical foundation for analyzing and designing homogeneous large-scale networked dynamical systems in a variety of areas. For example, the framework of the generalized frequency variable can be applied to the analysis and synthesis of central pattern generators (CPGs) [8] and gene-protein regulatory networks [1, 14, 7] as well as consensus and formation problems as surveyed in [12].

The very fundamental properties including controllability/observability have been discussed in [3, 4], and Reference [16] provided two systematic ways of stability check, namely an algebraic condition and LMI condition, which are different from graphical tests in [13, 12]. A Hurwitz type stability criterion for characteristic polynomials with complex coefficients in [2] was used for the derivation of the former condition, and it can be reduced to a set of LMIs by applying a generalized Lyapunov theorem in [6]. Recently, $\mathscr{H}_2$ and $\mathscr{H}_\infty$ norm computations for fairly general class of multi-agent dynamical systems have been investigated in [5]. The $\mathscr{H}_2$ norm computations are useful for evaluating a variety of control performances including rapidness of consensus, and the $\mathscr{H}_\infty$ norm relates conditions for robust stability and robust performances, which bring us a systematic treatment of heterogeneous multi-agent dynamical systems.

This paper summarizes recent results on the fundamental analysis for LTI systems with generalized frequency variables. After introducing the system description in Section 2, stability conditions, $\mathscr{H}_2$-norm computations, and $\mathscr{H}_\infty$-norm conditions are shown in Sections 3, 4, 5, respectively. Section 6 is devoted to an application to cyclic gene regulatory networks, where an analytic condition for the existence of protein level's periodic oscillations.

We use the following notation. The sets of real, positive real, complex and natural numbers, are denoted by $\mathbb{R}$, $\mathbb{R}_+$. $\mathbb{C}$, and $\mathbb{N}$, respectively. The complex conjugate of $z \in \mathbb{C}$ is denoted by $\bar{z}$. For a matrix A, its transpose and complex conjugate transpose are denoted by A^T and A^*, respectively. For a square matrix A, the set of eigenvalues is denoted by $\sigma(A)$. The symbols $\mathbf{S}_n$ and $\mathbf{S}_n^+$ stand for the sets of $n \times n$ real symmetric matrices and its positive definite subsets. For matrices A and B, $A \otimes B$ means their Kronecker product. The open left-half complex plane and the closed right-half complex plane are denoted by $\mathbb{C}_-$ and $\mathbb{C}_+$, respectively.

2 Linear Time-Invariant System with Generalized Frequency Variable

We here define a class of linear time-invariant (LTI) systems with generalized frequency variables and provide their dynamical system representations in the frequency and time domains. Specifically, consider the LTI system described by the transfer function

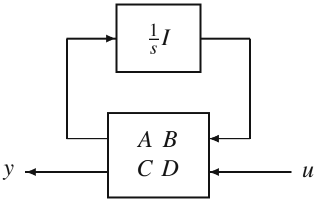


Fig. 1 LFT representation of $G(s)$

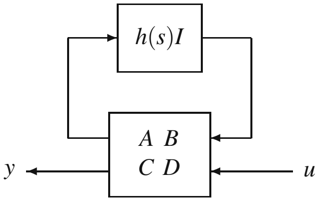


Fig. 2 LFT representation of $\mathscr{G}(s)$

$$\mathscr{G}(s) = C\left(\frac{1}{h(s)}I_n - A\right)^{-1} B + D = \mathscr{F}_u\left(\begin{bmatrix} A & B \\ C & D \end{bmatrix}, h(s)I_n\right), \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, $D \in \mathbb{R}^{p\times m}$, $h(s)$ is a single-input single-output, νth-order, strictly proper transfer function, and $\mathscr{F}_u$ denotes the upper linear fractional transformation. The system $\mathscr{G}(s)$ can be viewed as an interconnection of n identical agents, each of which has the internal dynamics $h(s)$ rather than an integrater in the standard system depicted in Fig. 1,. As shown in Fig. 2, the interconnection structure is specified by A, and the input-output structure for the whole system is specified by B, C, and D. Defining the transfer function

$$G(s) = C(sI_n - A)^{-1}B + D, \tag{2}$$

the system can be described as

$$\mathscr{G}(s) = G(\phi(s)), \quad \phi(s) := 1/h(s). \tag{3}$$

Note that the variable s in (2) characterizes frequency properties of the transfer function $G(s)$ and that $\mathscr{G}(s)$ is generated by simply replacing s by $\phi(s)$ in G. Hence, we say that the system (1) is described by the transfer function G with the *generalized frequency variable* $\phi(s)$ [3, 4].

Let $h(s)$ have a minimal realization $h(s) \sim (A_h, b_h, c_h, 0)$, where $A_h \in \mathbb{R}^{\nu\times\nu}$, $b_h \in \mathbb{R}^{\nu}$, $c_h \in \mathbb{R}^{1\times\nu}$. It can be readily shown that a realization of $\mathscr{G}(s)$ is given by $\mathscr{G}(s) \sim (\mathscr{A}, \mathscr{B}, \mathscr{C}, \mathscr{D})$, where

$$\begin{aligned} &\mathscr{A} = I_n \otimes A_h + A \otimes (b_h c_h) \in \mathbb{R}^{n\nu\times n\nu}, && \mathscr{B} = B \otimes b_h \in \mathbb{R}^{n\nu\times m}, \\ &\mathscr{C} = C \otimes c_h \in \mathbb{R}^{p\times n\nu}, && \mathscr{D} = D \in \mathbb{R}^{p\times m}. \end{aligned} \tag{4}$$

It should be also noticed that $(\mathscr{A}, \mathscr{B}, \mathscr{C}, \mathscr{D})$ is a minimal realization if (A, B, C, D) and $(A_h, b_h, c_h, 0)$ are both minimal realizations [3, 4].

3 Stability Conditions

Thanks to the preserving property on minimality of the realizations, we have the equivalence between the BIBO stability and the internal stability. In other words, The LTI system with generalized frequency variable $\mathscr{G}(s)$ given by (3) is BIBO stable (all the poles of $\mathscr{G}(s)$ are in $\mathbb{C}_-$), if and only if

$$\mathscr{H}_A(s) := \left(\frac{1}{h(s)}I - A\right)^{-1} = (\phi(s)I - A)^{-1}, \tag{5}$$

is exponentially stable or $\mathscr{H}_A(s)$ is proper and analytic in the closed right half complex plane. In other words, we can check the stability of an LTI system with generalized frequency variable $\phi(s) = 1/h(s)$ from the pair $(A, h(s))$, and we have the following theorem [16].

Theorem 1. *Let a matrix $A \in \mathbb{R}^{n\times n}$, and a strictly proper rational function $h(s) = n(s)/d(s)$ be given and define $\mathscr{H}_A(s)$ by* (5) *and $p(\lambda, s)$ by*

$$p(\lambda, s) := d(s) - \lambda n(s). \tag{6}$$

Suppose that $n(s)$ and $d(s)$ are coprime. The following five statements are equivalent.

(i) *$\mathscr{H}_A(s)$ is stable.*
(ii) *For all $\lambda \in \sigma(A)$, all the eigenvalues of $A_h + \lambda b_h c_h$ belong to the open left-half complex plane.*
(iii) *$\sigma(A) \subset \Lambda := \{\, \lambda \in \mathbb{C} \mid p(\lambda, s)$ is Hurwitz $\}$.*
(iv) *$\sigma(A) \subset \bigcap_{k=1}^{\nu} \Sigma_k$,*
where $\Sigma_k := \{\, \lambda \in \mathbb{C} \mid l_{\ell_k}(\lambda)^* \Phi_k l_{\ell_k}(\lambda) > 0 \,\}$.
(v) *For each $k = 1, 2, \ldots, \nu$, there exists $X_k \in \mathbf{S}_n^+$ such that*

$$L_{\ell_k}(A)^T (\Phi_k \otimes X_k) L_{\ell_k}(A) > 0. \tag{7}$$

Here,

$$l_\ell(\lambda) := \begin{bmatrix} 1 \\ \lambda \\ \vdots \\ \lambda^{\ell-1} \end{bmatrix}, \quad L_\ell(A) := \begin{bmatrix} I \\ A \\ \vdots \\ A^{\ell-1} \end{bmatrix},$$

and the positive integer $\ell_k \in \mathbb{N}$ and $\Phi_k \in \mathbf{S}_{\ell_k}$ for $k = 1, 2, \ldots, \nu$ are specified by applying a Hurwitz-type stability test for polynomials with complex coefficients in [6] *to the corresponding closed-loop characteristic polynomial $p(\lambda, s)$.*

It is clear that Condition (ii) is equivalent to that of (i), and the former will be used for the the $\mathscr{H}_2$-norm computation in the next section. The equivalence among (i), (iii), (iv), and (v) was proved in [16], Condition (iv), or a Hurwitz type condition, gives us an algebraic condition so that all the eigenvalues of A should belong to guarantee the stability of the total system. Condition (v), or Lyapunov type condition, provides an LMI feasibility problem, where we need no prior computation of the set of all eigenvalues of A.

The following is an algorithm based on Condition (v) for checking stability of the LTI system with generalized frequency variable. It includes a simple example of $h(s) = (cs+d)/(s^2+as+b)$, where we assume that $n(s) = cs+d$ and $d(s) = s^2+as+b$ are coprime to illustrate the proposed procedure.

Algorithm: Let $h(s)$ and A be given.

1. Write $h(s) = n(s)/d(s)$ with $n(s)$, $d(s)$ coprime. Define the polynomial $p(\lambda,s)$ by (6).

$$\begin{aligned} p(\lambda,s) &= s^2+as+b-\lambda(cs+d) \\ &= s^2+(a-cx-jcy)s+b-dx-jdy, \end{aligned}$$

   where x and y are the real and imaginary part of λ, respectively.
2. Obtain a necessary and sufficient condition for $p(\lambda,s)$ to be a Hurwitz polynomial in s by applying a Hurwitz type stability test for polynomials with complex coefficients in [2]. Let $\Delta_k(\lambda,\bar{\lambda}) > 0$ $(k = 1,2,\dots,\nu)$ be the resulting condition.

$$\begin{aligned} &\Delta_1(\lambda,\bar{\lambda}) = a-cx = a-c\lambda-c\bar{\lambda} > 0, \\ &\Delta_2(\lambda,\bar{\lambda}) = \begin{vmatrix} a-cx & 0 & dy \\ 1 & b-dx & cy \\ 0 & -cy & a-cx \end{vmatrix} \\ &= -\frac{1}{2}c^2d(\lambda^2\bar{\lambda}+\lambda\bar{\lambda}^2)+\frac{1}{2}(acd+3bc^2-d^2)\lambda\bar{\lambda} \\ &+\frac{1}{4}\left(adc+bc^2+d^2\right)(\lambda^2+\bar{\lambda}^2)-\frac{1}{2}(2abc+ad^2)(\lambda+\bar{\lambda})+a^2b > 0. \end{aligned}$$

3. Obtain the maximum degree ℓ_k-1 of λ in $\Delta_k(\lambda,\bar{\lambda})$ and the coefficient matrix Φ_k of $\Delta_k(\lambda,\bar{\lambda})$ for $k = 1,2,\dots,\nu$, where the (i,j) entry of Φ_k is the coefficient of $\lambda^{i-1}\bar{\lambda}^{j-1}$ in $\Delta_k(\lambda,\bar{\lambda})$.

$$\ell_1 = 2, \quad \ell_2 = 3, \quad \Phi_1 = \begin{bmatrix} 2a & -c \\ -c & 0 \end{bmatrix},$$

$$\Phi_2 = \frac{1}{4}\begin{bmatrix} 4a^2b & -2a^2d-4abc & acd+bc^2+d^2 \\ -2a^2d-4abc & 6acd+2bc^2-2d^2 & -2c^2d \\ acd+bc^2+d^2 & -2c^2d & 0 \end{bmatrix}.$$

4. For $k = 1,2,\dots,\nu$, form the LMIs

$$L_{\ell_k}(A)^T(\Phi_k \otimes X_k)L_{\ell_k}(A) > 0, \quad X_k \in \mathbf{S}_n^+. \tag{8}$$

5. Check whether there exists X_k satisfying (8) for each $k = 1,2,\dots,\nu$. If feasible/infeasible, then the system is stable/unstable.

The proposed algorithm can systematically check the stability condition for the LTI system with the generalized frequency variable $\phi(s) = 1/h(s)$ and the interconnection matrix A. Recall that stability of $\mathscr{G}(s)$ can also be determined through existing methods, e.g., Lyapunov inequality $\mathscr{A}\mathscr{X}+\mathscr{X}\mathscr{A}^T < 0$. However, the size of the matrices in this LMI is $n\nu \times n\nu$. On the other hand, the main stability theorem is given in terms of ν LMIs in which the sizes of the matrices are $n \times n$. Thus, the computational burden associated with our stability condition can be much less than that of the standard Lyapunov method if the system dimension is high.

4 $\mathscr{H}_2$-Norm Computations

This section is devoted to the $\mathscr{H}_2$ norm computations for $\mathscr{G}(s)$, where we assume that $D = O$ to assure the boundedness of the norm and that A is diagonalizable for the notational simplicity.

The following theorem can be derived for the $\mathscr{H}_2$ norm computation [5].

Theorem 2. *For a given stable $\mathscr{G}(s)$ with $D = O$ in* (1). *We assume that A is diagonalizable and that A is represented by $A = T\Lambda T^{-1}$ with a non-singular matrix T and $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. Then, we have*

$$\|\mathscr{G}\|_2^2 = \mathrm{tr}\Big[\big(\Theta \otimes (c_h^* c_h)\big)P\Big] = \mathrm{tr}\Big[\big(\Pi \otimes (b_h b_h^*)\big)Q\Big], \tag{9}$$

where $\Pi := T^{-1}BB^*(T^{-1})^*$, $\Theta := T^*C^*CT$, and the (i, j) block of P, Q for $i = 1, 2, \ldots, n$, $j = i, i+1, \ldots, n$ denoted by $P_{ij}, Q_{ij} \in \mathbb{C}^{\nu \times \nu}$, $(i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, n)$ are the unique solutions of the following Sylvester equations:

$$\hat{A}_i P_{ij} + P_{ij} \hat{A}_j^* = -\pi_{ij} b_h b_h^*, \tag{10}$$

$$\hat{A}_i^* Q_{ij} + Q_{ij} \hat{A}_j = -\theta_{ij} c_h^* c_h, \tag{11}$$

where $\hat{A}_i := A_h + \lambda_i b_h c_h$, and π_{ij}, θ_{ij} are the (i, j) elements of Π and Θ, respectively.

If we restrict the class of systems, we have the more compact result.

Corollary 1. *For a given stable $\mathscr{G}(s)$ with $D = O$ in* (1). *We assume that A is a normal matrix which is represented by $A = T\Lambda T^{-1}$ with a unitary matrix T and $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ and that $B = I$. Then, we have*

$$\|\mathscr{G}\|_2^2 = \mathrm{tr}\Big[\big((T^* C^* CT) \otimes (c_h^* c_h)\big)P\Big], \tag{12}$$

where P is the block diagonal matrix defined by $P = \mathrm{diag}(P_1, P_2, \ldots, P_n), P_i \in \mathbb{C}^{\nu \times \nu}$ with P_i $(i = 1, 2, \ldots, n)$ being the unique solutions of the Lyapunov equation

$$\hat{A}_i P_i + P_i \hat{A}_i^* = -b_h b_h^* \tag{13}$$

where $\hat{A}_i := A_h + \lambda_i b_h c_h$.

The proof is straightforward from Theorem 2 using the facts T is unitary and $\Pi = I$. The corollary exploits the normal structure of the interconnection matrix A to reduce the computational complexity. We only need to solve n independent Lyapunov equations with size $\nu \times \nu$ for the $\mathscr{H}_2$ norm computation.

Another corollary which corresponds to the results in [11, 10] is given if we further assume the $C = I_n$ in the corollary.

5 $\mathscr{H}_\infty$-Norm Conditions

The following theorem provides two exact conditions for the $\mathscr{H}_\infty$ norm of a special class of $\mathscr{G}(s)$ [5].

Theorem 3. *For a given positive number $\gamma > 0$ and an LTI system with generalized frequency variable $\mathscr{G}(s)$ represented by* (1)*, we assume that A is a normal matrix, i.e., $AA^* = A^*A$, $B = I_n$, $C = I_n$, and $D = O_n$. Then the following statements are equivalent.*

(i) $\|\mathscr{G}\|_\infty < \gamma$

(ii) *For all $\lambda \in \sigma(A)$,*

$$\left\| \frac{h}{1-\lambda h} \right\|_\infty < \gamma. \tag{14}$$

(iii) *For all $\lambda \in \sigma(A)$ and $\phi \in \Phi := \{1/h(j\omega) : \omega \in \mathbb{R}\}$,*

$$\left| \frac{1}{\phi - \lambda} \right| < \gamma. \tag{15}$$

Thus, the H_∞ norm calculation of the system (1) can be decomposed into that of n subsystems with a complex coefficient in this special case. Condition (iii) gives a graphical test, and there are three ways to compute the H_∞ norm based on Condition (ii), where we use the γ iteration to get the precise value. The three methods are based on the bounded real lemma (LMI), the eigenvalue condition of the associated Hamiltonian matrix, and the corresponding polynomial KYP lemma. See [5] for the details.

There are two remarks on Theorem 3.

- Since the H_∞ norm constraint corresponds to the allowable uncertainty bound in the context of robust control, the conditions for $\|\mathscr{G}\|_\infty < \gamma$ in Theorem 3 relates a robust stability condition. As shown in [5] the corresponding class of perturbations to the nominal system $h(s)I$ is the feedback type, and the class of perturbed system without inter connection matrix A is given by

$$\mathbf{H}_\gamma := \{\ (I + h(s)\Delta(s))^{-1}h(s) \mid \Delta(s) \in \mathbf{\Delta}_\gamma\ \}.$$

$$\mathbf{\Delta}_\gamma := \left\{ \Delta(s) \,\middle|\, \text{proper \& stable, } \|\Delta\|_\infty \le \frac{1}{\gamma} \right\},$$

  In other words, the result for the H_∞ norm computation can be applied to the allowable uncertainty bound of feedback type to assure the feedback stability. Hence, this is a powerful tool for examination of the stability of heterogeneous multi-agent dynamical systems. It should be noticed that the result relates an investigation in [13] for the identical diagonal perturbation case.

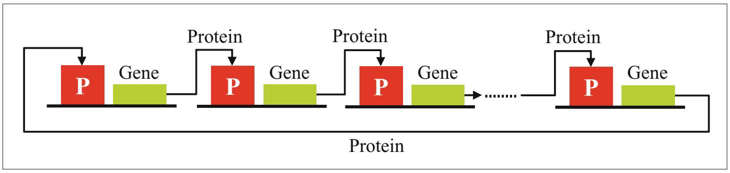


Fig. 3 Gene regulatory network with cyclic structure [7]

- The same approach can be applied to the wider class of systems if A is a normal matrix. One of the interesting classes is a feedback loop system in the H_∞ loop shaping design, since it relates the normalized coprime factor perturbations and it includes both the sensitivity and complementarity sensitivity functions. The target transfer function is given by

$$\mathscr{L}(s) := \begin{bmatrix} A \\ I \end{bmatrix} (I - h(s)A)^{-1} \begin{bmatrix} h(s)I \ I \end{bmatrix},$$

and the exact H_∞ norm condition is derived in [5].

6 Application to Gene Regulatory Networks

This section is devoted to an application of the stability results in Section 3 to large-scale cyclic gene regulatory networks, where we show an analytic condition for the existence of protein level's periodic oscillations.

The well-known central dogma of molecular biology is that protein is synthesized following the two steps called transcription and translation: genes on a DNA are first transcribed into mRNAs, and then a mRNA is translated into one or multiple copies of corresponding proteins. Further, some proteins, called transcription factors, are known to activate or repress the transcription of other genes. Then, such chemical interactions between transcription factors and genes can be described by gene regulatory networks.

We here consider the gene regulatory network where a transcription factor of one gene activates or represses the transcription of another gene in a cyclic way as depicted in Fig. 3 [15]. Note that this cyclic feedback structure is one of the substantial chemical pathways in living organisms, and is also observed in metabolic pathways, tissue growth regulations, cellular signaling pathways and neuron models (see [9] and references therein). Then, the dynamics of the above cyclic gene regulatory networks is modeled as, for $i = 1, 2, \cdots, N$,

$$\begin{aligned} \dot{r}_i(t) &= -a_i r_i(t) + \beta_i f_i(p_{i-1}(t)), \\ \dot{p}_i(t) &= c_i r_i(t) - b_i p_i(t), \end{aligned} \tag{16}$$

where $r_i \in \mathbb{R}_+$ and $p_i \in \mathbb{R}_+$ denote the concentrations of the i-th mRNA and its corresponding protein synthesized in the i-th gene, respectively. Let $p_0(t) := p_N(t)$ and $r_0(t) := r_N(t)$ for the sake of notational simplification. Positive constants a_i, b_i, c_i

and β_i represent the followings: a_i and b_i denote the degradation rates for the i-th mRNA and protein, respectively; c_i and β_i denote the translation and transcription rates, respectively. A monotonic function $f_i(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ represents either activation or repression of the transcription. As one pair of the candidates for $f_i(\cdot)$, the following Hill functions are often introduced in biochemical characterization:

$$f_i(p_{i-1}) = \frac{1}{1+p_{i-1}^{\nu}}, \quad f_i(p_{i-1}) = \frac{p_{i-1}^{\nu}}{1+p_{i-1}^{\nu}} \tag{17}$$

We here assume that the gene regulatory network has odd number of repressive interactions between genes, or it holds that

$$\delta := \left(\frac{df_1}{dp_N}\right) \cdot \left(\frac{df_2}{dp_1}\right) \cdots \left(\frac{df_N}{dp_{N-1}}\right) < 0. \tag{18}$$

We can show that the protein concentrations of the cyclic gene regulatory network either (i) converge to an equilibrium state, or (ii) oscillate periodically [7]. Thus, if such a unique equilibrium state is locally unstable, there exists a set of initial values so that protein concentrations do not converge to the equilibrium level and eventually enter into a non-constant periodic orbit.

In order to analyze the local stability of (16), we now consider its linearized model around the equilibrium state:

$$\begin{bmatrix} \dot{r}_i \\ \dot{p}_i \end{bmatrix} = \begin{bmatrix} -a_i & 0 \\ c_i & -b_i \end{bmatrix} \begin{bmatrix} r_i \\ p_i \end{bmatrix} + \begin{bmatrix} \beta_i \\ 0 \end{bmatrix} u_i, \quad u_i := \zeta_i p_{i-1}, \tag{19}$$

where $\zeta_i := f_i'(p_{i-1}^*)$. Note that condition (18) is equivalent to $\prod_{i=1}^{N} \zeta_i < 0$.

If we assume that the network consists of nearly the same bases of genes, it can be assumed that $a_1 = \cdots = a_N (=: a)$ and $b_1 = \cdots = b_N (=: b)$. Then, the network system can be represented by an LTI system with generalized frequency variable represented by

$$\mathscr{G}(s) := (\phi(s)I - A)^{-1}, \quad \phi(s) := 1/h(s), \tag{20}$$

where

$$h(s) := \frac{1}{(T_a s+1)(T_b s+1)}; \quad T_a := 1/a, \quad T_b := 1/b \tag{21}$$

and

$$A := \begin{bmatrix} 0 & 0 & 0 & \cdots & \zeta_1 R_1^2 \\ \zeta_2 R_2^2 & 0 & 0 & \cdots & 0 \\ 0 & \zeta_3 R_3^2 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \zeta_N R_N^2 & 0 \end{bmatrix}. \tag{22}$$

Applying Theorem 1 to the system leads to

$$\Omega_+ := \left\{ -y + jx \in \mathbb{C} \;\middle|\; y < \frac{1}{4}Q^2x^2 - 1 \right\} \tag{23}$$

as the instability region Ω_+ for the eigenvalues matrix A defined by (22), where

$$Q := \frac{\sqrt{T_aT_b}}{(T_a + T_b)/2} \left(= \frac{\sqrt{ab}}{(a+b)/2} \right). \tag{24}$$

Note that Q is the ratio between the arithmetic and geometric means of mRNA and protein time constants, T_a and T_b.

We can see from the special structure of A that the eigenvalues of A are simply computed as, for $k = 1, 2, \cdots, N$,

$$\lambda_k = Le^{j(2k-1)\pi/N}, \quad L := \left| \prod_{\ell=1}^{N} R_\ell^2 \zeta_\ell \right|^{\frac{1}{N}}. \tag{25}$$

This together with the instability region Ω_+ yields the following analytic criterion for the existence of periodic oscillations [7].

Theorem 4. *Consider the cyclic gene regulatory network with gene dynamics* (16) *where $a_i = a$, $b_i = b$, $(i = 1, 2, \cdots, N)$, and its linearized system $\mathscr{H}(s)$ in* (20). *Assume that the condition in* (18) *holds. Then, there exist the periodic oscillations of protein concentrations p_i $(i = 1, 2, \cdots, N)$, if the following condition holds:*

$$L := \left| \prod_{\ell=1}^{N} R_\ell^2 \zeta_\ell \right|^{\frac{1}{N}} > \frac{2\left(-\cos(\frac{\pi}{N}) + \sqrt{\cos^2(\frac{\pi}{N}) + Q^2 \sin^2(\frac{\pi}{N})} \right)}{Q^2 \sin^2(\frac{\pi}{N})} \tag{26}$$

We briefly remark on the relation between our results and the conventional ones. The analysis by Samad *et al.* [14] was performed based on the direct computation of the eigenvalues of the Jacobian matrix, and then analytic criteria for $N = 2, 3$ cases were presented. We can easily see that their results coincide with the ones in Theorem 2 where $N = 2, 3$.

7 Conclusion

In this paper, we have considered LTI systems with generalized frequency variables $\phi(s)$, described as $C(\phi(s)I - A)^{-1}B + D$ as one of the unifying expressions for homogeneous multi-agent dynamical systems. Such systems arise from interconnections of multiple identical subsystems, where $h(s) := 1/\phi(s)$ is the common subsystem dynamics, and A is the connectivity matrix characterizing the information exchange among subsystems.

This paper summarized recent results on the fundamental analysis, namely stability conditions, $\mathscr{H}_2$-norm computations, and $\mathscr{H}_\infty$-norm conditions, for LTI systems with generalized frequency variables. We have also shown an analytic condition for the existence of protein level's periodic oscillations in cyclic gene regulatory networks as an application. Although the class of such systems can be directly applied only to homogeneous multi-agent dynamical systems, it has a potential to be applied to heterogeneous multi-agent dynamical systems if we consider the robust stability conditions.

Acknowledgements. This work was supported in part by Japan Science and Technology Agency and by Grant-in-Aid for Scientific Research (A) of the Ministry of Education, Culture, Sports, Science and Technology, Japan, No. 21246067 and No. 21656106. The author would like to thank Professors T .Iwasaki and T.-H. Kim, Mr. H. Tanaka, and Mr. Y. Hori for doing the joint work on this research topic.

References

1. Banks, H., Mahaffy, J.: Stability of cyclic gene models for systems involving repression. J. of Theoretical Biology 74, 323–334 (1978)
2. Frank, E.: On the zeros of polynomials with complex coefficients. Bulletin of the American Mathematical Society 52(2), 144–157 (1946)
3. Hara, S., Hayakawa, T., Sugata, H.: Stability analysis of linear systems with generalized frequency variables and its applications to formation control. In: Proc. of the 46th IEEE Conference on Decision and Control, pp. 1459–1466 (2007)
4. Hara, S., Hayakawa, T., Sugata, H.: LTI systems with generalized frequency variables: a unified framework for homogeneous multi-agent dynamical systems. SICE J. of Control, Measurement, and System Integration 2(5), 299–306 (2009)
5. Hara, S., Iwasaki, T., Tanaka, H.: H_2 and H_∞ Norm Computations for LTI Systems with Generalized Frequency Variables, The University of Tokyo, Mathematical Engineering Technical Reports, METR 2009-49 (2009), `http://www.keisu.t.u-tokyo.ac.jp/research/techrep/`
6. Hinrichsen, D., Pritchard, A.: Mathematical Systems Theory I: Modelling, State Space Analysis, Stability And Robustness. Texts in Applied Mathematics, vol. 48. Springer, Heidelberg (2005)
7. Hori, Y., Kim, T.-H., Hara, S.: Graphical and Analytic Criteria for the Existence of Protein Level Oscillations in Cyclic Gene Regulatory Networks. To appear in Proc. of the 48th IEEE Conference on Decision and Control (2009)
8. Iwasaki, T.: Multivariable harmonic balance for central pattern generators. Automatica 44, 3061–3069 (2008)
9. Jovanović, M.R., Arcak, M., Sontag, E.D.: A passivity-based approach to stability of spatially distributed systems with a cyclic interconnection structure. IEEE Trans. on Automatic Control 53, 75–86 (2008)
10. Li, Z., Duan, Z., Huang, L.: $\mathscr{H}_\infty$ control of networked multi-agent systems. J. of Systems Science and Complexity 22(1), 35–48 (2009)
11. Massioni, P., Verhaegen, M.: Distributed control for identical dynamically coupled systems: a decomposition approach. IEEE Trans. on Automatic Control 54(1), 124–135 (2009)

12. Olfati-Saber, R., Fax, J.A., Murray, R.M.: Consensus and cooperation in networked multi-agent systems. Proc. of the IEEE 95(1), 215–233 (2007)
13. Polyak, B.T., Tsypkin, Y.Z.: Stability and robust stability of uniform system. Automation and Remote Control 57(11), 1606–1617 (1996)
14. Samad, H.E., Vecchio, D.D., Khammash, M.: Repressilators and promotilators: loop dynamics in synthetic gene networks. In: Proc. of the 2005 American Control Conference, pp. 4405–4410 (2005)
15. Smith, H.L.: Oscillations and multiple steady states in a cyclic gene model with repression. J. Math. Biol. 25, 169–190 (1987)
16. Tanaka, H., Hara, S., Iwasaki, T.: LMI stability condition for linear systems with generalized frequency variables. In: Proc. of the 7th Asian Control Conference, pp. 136–141 (2009)

Control and Stabilization of Linear Equation Solvers

Uwe Helmke and Jens Jordan

Abstract. This work connects with and partially extends the pioneering work by the group of Y. Yamamoto on systems theory techniques for numerical algorithms ([22, 18]). We consider iterative linear equation solvers such as Richardson iteration, GMRES(m) and more generally, Krylov subspace methods from a control theoretic viewpoint. The motivation for this research lies in the need to improve convergence properties by suitable feedback design strategies as well as extending the applicability of linear equation solvers to wider classes of, possibly non-normal, matrices. We derive necessary as well as sufficient conditions for controllability of polynomially shifted linear equation solvers and consider optimal control feedback strategies via Riccati equations.

1 Introduction

One of the main tasks in numerical linear algebra is to solve large systems of linear equations $Ax = b$. The literature provides a large number of iterative solution methods and software packages. Nevertheless, it is fair to say, that these methods work only under severe restrictions on the matrix A. To this date there is no iterative equation solver known, that works for a generic set of matrices. Krylov subspace methods, such as conjugate gradient iteration or GMRES(m) work well for positive definite or normal matrices. However, for non-normal matrices the dynamics of Krylov methods can become quite complex and is far from being fully understood. In fact, for indefinite symmetric matrices, GMRES(m) exhibit continua of non-trivial equilibrium points that may prevent the algorithm to converge to the desired solution. The situation becomes worse for matrices far from normality, forcing the algorithm to loose fast local convergence or create regimes of chaotic behavior in the phase space; see Embree [6].

Uwe Helmke and Jens Jordan
University of Würzburg, Germany
e-mail: {helmke,jordan}@mathematik.uni-wuerzburg.de

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 73–82.
springerlink.com

The variables of such methods – such as shifts or relaxation parameters – can be regarded as control parameters. Thus we obtain control systems, which can be studied with the various tools from control theory. One of the first approaches in this direction can be traced back to the early work of Bellman [2]. In section 14.9 *Computing as a control process* Bellman writes:

> We have already referred to the fact that computing can be considered to be a control process in which we want to blend complexity, time, and accuracy in some appropriate way. It should, in addition, be treated in many cases as an adaptive control process in which results of the previous calculation are used to guide the subsequent calculations, producing not only a choice of parameters but a choice of algorithms

During the past ten years a number of feedback control and stabilization tasks for numerical algorithms have been considered. This includes the work by Gustafsson, Lundh and Söderlind [8] and Gustafsson [9] on step-size control in ODE-solvers; investigations on controllability of eigenvalue methods and linear equation solvers by Fuhrmann and Helmke [11], Helmke and Wirth [15], Helmke, Jordan and Lanzon [14], Jordan [16]; the work by Wakasa and Yamamoto [22] and Kashima and Yamamoto [18] on robust stability of linear equation solvers, and the textbook by Bhaya and Kaszkurewicz [3].

Throughout this paper, $\mathbb{F}$ denotes either the field of real numbers $\mathbb{R}$ or the field of complex numbers $\mathbb{C}$, respectively. We consider three different types of iteration schemes for solving a linear equation $Ax = b$: the Richardson method, restarted polynomial iteration and linear control schemes. *Richardson's method* refers to the bilinear control system on $\mathbb{F}^n$

$$x_{t+1} = x_t + u_t(b - Ax_t), \quad x_0 \in \mathbb{F}^n.$$

Clearly, a fixed point of this iteration is a solution of the linear equation $Ax = b$. The literature proposed different shift strategies for certain families of matrices; see, e.g., Opfer and Schober [19], Smorlaski and Saylor [20], Golub and Overton [10], Calvetti and Reichel [4]. In particular, the constant shift strategy $u_t = u$ yields the trivial splitting method, which converges if and only if $Spec(I - uA)$ lies in the unit disc. Another interesting shift strategy is given by the feedback law $u_t = r_t^* A r_t / \|Ar_t\|^2$ with $r_t = b - Ax_t$. This approach yields GMRES(1) i.e.,

$$x_{t+1} = \arg \min_{x \in x_t + \operatorname{Span}(b - Ax_t)} \|b - Ax\|.$$

It is known, that GMRES(1) converges if $A + A^*$ is positive definite, but not for arbitrary matrices.

A generalization of Richardson's method are *restarted polynomial iterations of order m*

$$x_{t+1} = (I - p_t(A)A)x_t + p_t(A)b, \quad x_0 \in \mathbb{F}^n. \tag{1}$$

Here the controls p_t are polynomials of degree at most m. Polynomial restarted iteration can be regarded as restarted Krylov methods. See Sorensen [21] for an overview on Krylov methods and polynomial restarting. Note that this setting includes GMRES(m); see Eiermann, Ernst and Schneider [5], Joubert [17] and Embree [6] for convergence results. A somewhat counter-intuitive phenomenon appears, i.e. increasing the number m of inputs in such feedback schemes does not necessarily improve the convergence properties. In particular, Embree [6] constructed simple examples where GMRES(1) converges while GMRES(2) stagnates.

To improve controllability properties we introduce *linear control schemes* as an alternative to the bilinear Richardson method. Explicitly, we consider

$$x_{t+1} = (I - A)x_t + Bu_t + b, \quad x_0 \in \mathbb{R}^n \tag{2}$$

that has $A^{-1}b$ as an fixed point for the zero control $u_t = 0$. Here, the choice of B can be used to improve the convergence behavior. Generically, convergent shift strategies $u_t = Kx_t$ can be constructed using standard linear quadratic controller design. This results in a globally convergent iterative algorithm, called LQRES, for solving linear systems; see [13].

This works builds on the pioneering research by Yutaka Yamamoto and his group, who proposed robust control ideas for solving systems of linear equations. It is a pleasure to acknowledge the inspiration we gained from reading their papers. Happy birthday, Yutaka!

2 Controllability of Restarted Polynomial Iterations

We begin with an analysis of the controllability properties of Richardson's method

$$x_{t+1} = x_t + u_t(b - Ax_t), \quad x_0 \in \mathbb{F}^n \tag{3}$$

with controls $u_t \in \mathbb{F}$. Clearly, for any input sequence $(u_t)_{t\in\mathbb{N}}$ the state trajectory $(x_t)_{t\in\mathbb{N}}$ converges to a solution of $Ax = b$ if and only if the sequence of residuals $r_t := b - Ax_t$ converges to zero. Thus, the dynamics of (3) is equivalent to that of the residuals

$$r_{t+1} = (I - u_t A)r_t. \tag{4}$$

Given any initial condition $r_0 \in \mathbb{F}^n$, we consider the reachable set $\mathscr{R}(r_0)$ of (4), i.e. the set of states of which can be reached from r_0 with a finite number of control steps. We say that the Richardson method is controllable on $M \subset \mathbb{F}^n$ if $M \subset \mathscr{R}(r_0)$ holds for all $r_0 \in M$.

Let $\mathscr{W}$ be the set of proper A-invariant subspaces and $N_A := \mathbb{F}^n \setminus \bigcup_{W\in\mathscr{W}} W$. Clearly, for any A-invariant subspace W, $r_0 \in W$ implies $\mathscr{R}(r_0) \subset W$. In the following we assume that A is cyclic, i.e. there exists $x \in \mathbb{F}^n$ such that $x, Ax, \ldots, A^{n-1}x$ is a basis of $\mathbb{F}^n$. Clearly, the set of cyclic matrices is dense in the set of all matrices and for any cyclic matrix A, the subset N_A is open and dense in $\mathbb{F}^n$. In the complex case $\mathbb{F} = \mathbb{C}$, easy algebraic arguments similar to the analysis of inverse power iterations in [11], [16] show, that $N_A \subset \mathscr{R}(r_0)$ for any $r_0 \in N_A$.

Theorem 1. *Let $\mathbb{F} = \mathbb{C}$ and $A \in \mathbb{C}^{n \times n}$ be cyclic. Then, Richardson's method is controllable on N_A.*

The real case $\mathbb{F} = \mathbb{R}$ turns out to be much more complicated. In the following we show that Richardson's method is controllable on N_A if A is real diagonalizable. Our arguments use some results of Jordan [16], where reachable sets of iteration schemes are analyzed in a general framework using the semigroup structure of reachable sets. Consider $S_A := \{\prod_{t=1}^{T}(I - u_t A) \,|\, T \in \mathbb{N}, u_t \in \mathbb{R}\} \subset \mathbb{R}^{n \times n}$. Clearly, S_A is a semigroup with respect to matrix multiplication. Moreover,

$$\mathscr{R}(r_0) = S_A r_0 := \{S r_0 \,|\, S \in S_A\}.$$

In [16] it is shown, that the invertible elements of S_A form a group, provided the eigenvalues of A are real. This yields the following controllability result.

Theorem 2. a) *There exists an open set of invertible cyclic real $n \times n$-matrices A for which such Richardson's method is controllable on N_A. In particular this is the case if A has n different real eigenvalues.*

b) *For $n > 2$, Richardson's method is not controllable on N_A, if A has a purely imaginary eigenvalue.*

Proof. The semigroup $S_A \cap GL_n(\mathbb{R})$ is a subgroup of the centralizer group $Z(A) := \{B \in GL_n(\mathbb{R}) \,|\, BA = AB\}$. Let G_A be the smallest subgroup of $Z(A)$ such that $S_A \cap GL_n(\mathbb{R}) \subset G_A$. We show that (i) $G_A = Z_A$ and thus $S_A \cap GL_n(\mathbb{R}) = G_A$ provided all eigenvalues of A are real and (ii) Z_A acts transitively on N_A. Then, $S_A \cap GL_n(\mathbb{R}) = G_A$ follows, since $N_A = (S_A \cap GL_n(\mathbb{R}))r_0$ is for any $r_0 \in N_A$ a subset of $\mathscr{R}(r_0)$.

Let m_A be the minimal polynomial of A. Since A is cyclic, $Z_A = \{p(A) \,|\, p \in \mathbb{R}[x] \text{ coprime to } m_A\}$ (see [7]). Any invertible $p(A)$ can be written in the form $p(A) = \prod_{t=1}^{N}(A - u_t I)(A - v_t I)^{-1}$ with $N \in \mathbb{N}$ and $u_t, v_t \in \mathbb{R} \setminus$ Spec A (see [16]). Thus, (i) holds, since $\mathrm{T}Z_{A^{-1}} \subset G_A \subset Z_A$ and $Z_{A^{-1}} = Z_A$.

Recall that N_A consists of all points $x \in \mathbb{R}^n$, which are not in a proper A-invariant subset. Therefore, $x, Ax, \ldots, A^{n-1}x$ is a basis of $\mathbb{R}^n$ if $x \in N_A$. In other words, any $w \in N_A$ can be written in the form $w = p(A)x$ with some polynomial $p \in \mathbb{R}[x]$ of degree at most $n-1$. Moreover, $p(A)$ is invertible, since it maps the basis $x, Ax, \ldots, A^{n-1}x$ on the basis $w, Aw, \ldots, A^{n-1}w$. Thus, Z_A acts transitively ob N_A.

Now we show statement b). Without loss of generality we assume

$$A = \begin{pmatrix} A_1 & A_2 \\ 0 & A_3 \end{pmatrix} \text{ with } A_1 = \begin{pmatrix} 0 & \operatorname{Im}\lambda \\ -\operatorname{Im}\lambda & 0 \end{pmatrix}.$$

For $N \subset \mathbb{R}^n$ we denote the topological closure of N with $\overline{N}$. We show that $0 \notin \overline{S_A r_0}$ for any $r_0 \in N_A$, and thus $\mathscr{R}(r_0) \setminus N_A \neq \emptyset$. Since G_A is abelian and acts transitively on N_A, any $r \in N_A$ can be written as $r_0 = \tilde{S}_1^{-1}\tilde{S}_2 r$ with $\tilde{S}_1, \tilde{S}_2 \in S_A \cap GL_n(\mathbb{R})$. Therefore, it is enough to show $0 \notin \overline{S_A r_0}$ for one $r_0 \in N_A$, since $S_n r \to 0$ with $r \in N_A$ and $S_n \in S_A, n \in \mathbb{N}$ implies $S_n r_0 \to 0$ and therefore $0 \in \overline{S_A r_0}$. Let $r_0 = (1, 1, 1 \ldots, 1)^\top$. Assuming that $\{0\} \cap \overline{S_A r_0} \neq \emptyset$. Then there exists a sequence

$$s_n := \begin{pmatrix} B_n & B'_n \\ 0 & B''_n \end{pmatrix} \in S_A,$$

with $B_n \in \mathbb{R}^{2\times 2}$ such that $B_n(1,1)^\top \to 0$ for $n \to \infty$. Since

$$B_n \subset P(A_1) := \left\{ \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \middle| \, a^2 + b^2 \neq 0 \right\}$$

and $\det B_n = \det(\prod_{t=1}^N (I - u_t A_1)) \geq 1$ we obtain

$$\|B_n(1,1)^\top\|_2 = \sqrt{(a+b)^2 + (a-b)^2} = \sqrt{2\det(B_n)} \geq \sqrt{2}.$$

Thus $\{0\} \cap \overline{S_A r_0} = \emptyset$.

The existence of a shift strategy $u = (u_t)_{t\in\mathbb{N}}$ such that (4) converges to a solution of $Ax = b$ implies $0 \in \overline{\mathscr{R}(r_0)}$. Thus, as an immediate consequence of Theorem 2 we obtain necessary condition for the existence of shift strategies such that Richardson methods converges to a solution of $Ax = b$.

Corollary 1. *If $A \in \mathbb{R}^{n\times n}$ has a purely imaginary eigenvalue and $x_0 \in \mathbb{R}^n$ is any generic initial condition, then $Ax = b$ can not be solved by any Richardson method.*

Finally, we consider *restarted polynomial iteration of degree m*

$$x_{t+1} = x_t - p_t(A)(b - Ax_t), \tag{5}$$

where $p_t \in \mathbb{R}[x]$ with $\deg p_t < m$. Such methods are also called *restarted Krylov methods*, since

$$x_{n+1} \in x_t + \mathscr{K}_m(A, r_t)$$

where $\mathscr{K}_m(A, r_t) := \operatorname{span}(r_t, Ar_t \ldots, A^{m-1} r_t)$ denotes the *Krylov space* with respect to A and $r_t := b - Ax_t$. . Similar to Richardson's method, the dynamics of the iteration can be equivalently described by the dynamics of the residual sequence $(r_t)_{t\in\mathbb{N}}$. We obtain

$$r_{t+1} = (I - Ap_t(A))r_t. \tag{6}$$

An polynomial interpolation argument shows, that the corresponding system semigroup coincides with the centralizer group, provided $m \geq 2$. This implies the following controllability result.

Theorem 3. *Let A be invertible and cyclic. The restarted polynomial iteration scheme of degree $m \geq 2$ is controllable on N_A.*

3 Linear Control Schemes

Another approach to design iterative methods for solving linear equations $Ax = b$ is based on *linear controller design techniques*. Here an additional tuning parameter

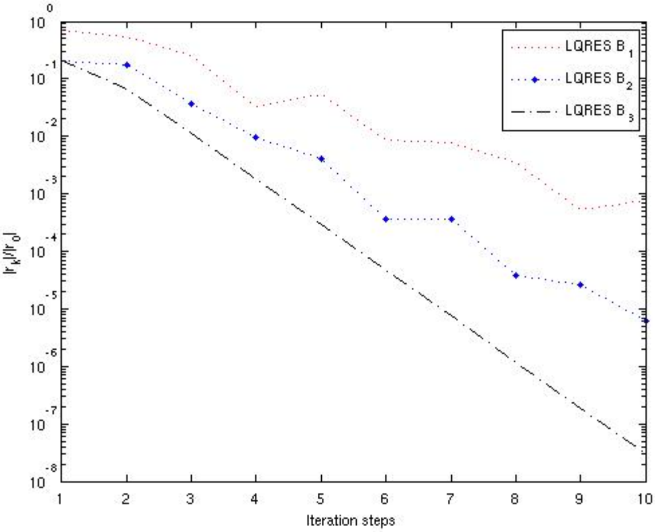


Fig. 1 Comparison of LQRES with different parameters B. We apply LQERS on the example problem of Embree (see [6]). The example is known to produce extreme behavior for restarted GMRES algorithms. In particular GMRES(2) fails to converge while GMRES(1) converges. We compare the relative residuals after n outer iteration steps. The algorithm converges for different parameters $B1 = (3, 1, 1)^\top$, $B2 = [B1, (1, 1, 1)^\top]$ and $B3 = [B1, (-1, -2, -3)^\top]$. However, the speed of convergence can be tuned by the choice of B.

arises, the choice of a suitable input matrix B. Thus, given $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$, we want to find a matrix $B \in \mathbb{R}^{n\times m}$ and a shift sequence $u_t \in \mathbb{R}^m, t \in \mathbb{N}$, satisfying $\lim_{t\to\infty} u_t = 0$ such that

$$x_{t+1} = (I - A)x_t + Bu_t + b \tag{7}$$

converges to $A^{-1}b$. Without loss of generality we assume that b lies in the image space of B. Otherwise we replace B by $\tilde{B} := [b, B] \in \mathbb{R}^{n\times(m+1)}$. Assuming that A is invertible, we have

$$x = A^{-1}b = \sum_{j=0}^{n-1} \alpha_j (I - A)^j b$$

for some $\alpha_j \in \mathbb{R}$, $j = 0, \ldots, n-1$. Thus $A^{-1}b$ is in the reachable subspace defined by $(I - A, B)$. The dynamics of the residuals $r_t := b - Ax_t$ is given by the linear system

$$r_{t+1} = (I - A)r_t - ABu_t,\ r_0 \in \mathbb{R}^n.$$

Based on standard ideas from Riccati-based linear quadratic controller design, we construct an explicit shift sequence such that (7) converges globally to the solution of $Ax = b$. Thus, under the assumption that $(I - A, -AB)$ is discrete-time stabilizable, r_t converges to zero for the feedback law $u_t = -Kr_t$ with

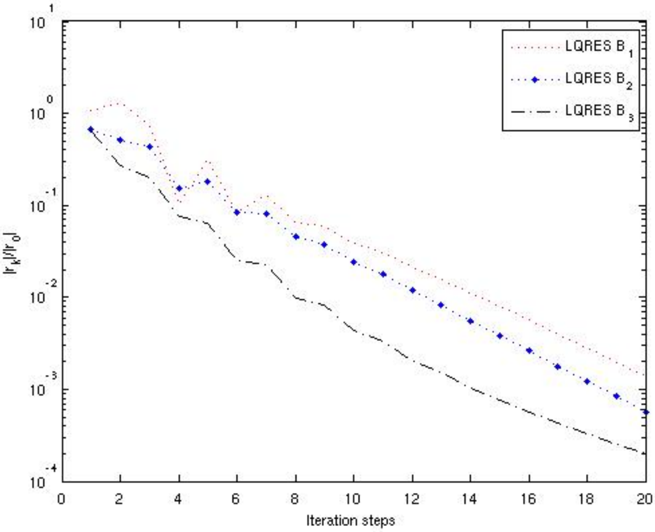


Fig. 2 LQRES applied on a Hilbert matrix of dimension 5 and $b = (1,0,0,0,0)^\top$. The elements of the Hilbert matrices are given by $a_{i,j} = \frac{1}{i+j-1}$. It is known that this matrix is poorly conditioned. We compare the relative residuals after n outer iteration steps. We observe that the speed of convergence increases when the number of columns of B gets larger.

$$K = -(I_m + B^\top A^\top PAB)^{-1} B^\top A^\top P(I-A). \tag{8}$$

Here P is the unique solution of the algebraic Riccati equation

$$P = I_n + \tilde{A}^\top P\tilde{A} + (\tilde{B}^\top P\tilde{A})^\top (I_m + \tilde{B}^\top P\tilde{B})^{-1} \tilde{B}^\top P\tilde{A} \tag{9}$$

with $\tilde{A} = I - A$ and $\tilde{B} = -AB$. This yields the following LQRES-algorithm [13, 14].

Algorithm (LQRES)

(i) *Choose B such that $(I-A, -AB)$ is stabilizable*
(ii) *Calculate the unique positive definite solution of the Riccati equation (9).*
(iii) *Calculate K as in Equation (8).*
(iv) *Iterate the closed loop system*

$$x_{t+1} = (I-A)x_t + BK(b - Ax_t) + b. \tag{10}$$

Classical LQG theory implies $\sum_{t=0}^{\infty}(\|r_t\|^2 + \|u_t\|^2) = r_0^\top P r_0$. Therefore the sequence of residuals $(r_t)_{t\in\mathbb{N}}$ converges to a zero, thus proving

Theorem 4. *If $(I-A, -AB)$ is stabilizable then LQRES converges to a solution of $Ax = b$.*

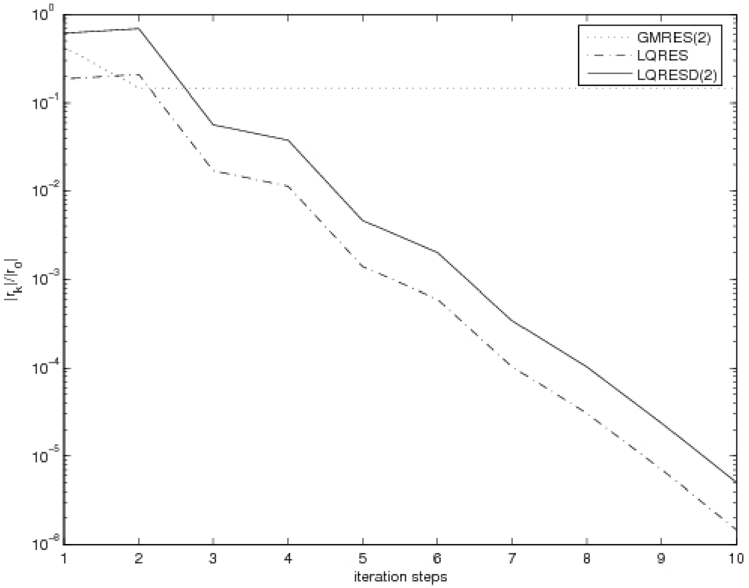


Fig. 3 Comparisons of GMRES(2), LQRES and LQRESD(2) on the example of Embree (See Figure 1). We compare the relative residuals after *n* outer iteration steps. As stated in [6], GMRES(2) diverges. LQRES and the cheaper LQRES(2) converge.

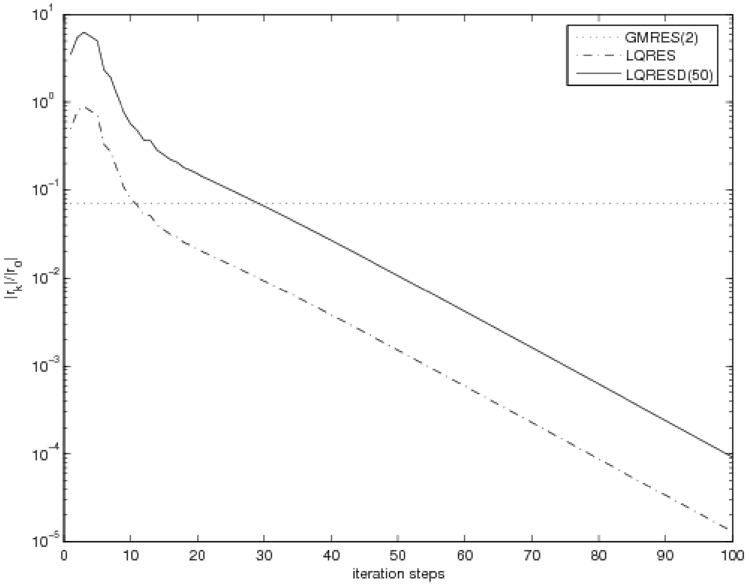


Fig. 4 Comparisons of GMRES(2), LQRES and LQRESD(50). We compare the relative residuals after *n* outer iteration steps. Here *A* is a randomly generated problem of size 50 × 50. Here *B* is a randomly generated 50 × 3-matrix.

We now address some of the difficulties and advantages of the above method. Certainly, a solution to step (i) does not exist for arbitrary choices of A. However, for generic choices of A, step (i) is always solvable. Moreover, the freedom in choosing B can be exploited to improve convergence speed (see Figure 1 and Figure 2). In any case, the existence of a matrix B with $(I-A,\, -AB)$ stabilizable is far less restrictive than assuming cyclicity of A. If A is invertible (which we always assume) then we can choose $m=n$ and any invertible B will do the job. However, this would increase the computational complexity in inverting the $m\times m$ matrix in (8),(9). Thus, a good design of B has to meet a trade-off between the size of B and the ease to satisfy the stabilizability constraint.

If the eigenvalues λ of A satisfy $|1-\lambda|<1$, then one can always choose $B=0$. In this case, LQRES coincides with the Richardson's method with constant shifts $u\equiv 1$. In particular, LQRES may converge in cases, where Richardson's iteration fails for all possible shift strategies. An example is

$$A=\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad b=\begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Corollary 1 shows that $Ax=b$ is not solvable for any Richardson method. In contrast, $(I-A,-AB)$ is stabilizable with $B:=b$ and thus LQRES converges.

If m is relatively small, then step (iii) does not cause serious numerical problems. However, the computationally expensive part lies in the preconditioning process of solving the algebraic Riccati equation (9). In fact, any known method for solving the algebraic Riccati equation is more expensive then solving the linear equation $Ax=b$. Nevertheless, we believe that variations of LQRES, e.g. by using suboptimal techniques for solving equation (9), may yield attractive numerical alternatives to established linear equation solvers. One possible approach in this direction might be model predictive control, where the solution to the algebraic Riccati equation P is replaced by the solution P_M after M steps of the dynamic Riccati difference equation. Preliminary numerical experiments with this method LQRESD(M) have shown useful convergence and stability properties (see Figure 3 and Figure 4). Another approach might be based on open loop optimal control strategies, e.g. via relaxed dynamic programming. Thus, the sky seems open for further investigations and improvements of iterative linear equation solvers.

Acknowledgements. Dedicated to the 60th Birthday of Yutaka Yamamoto.

References

1. Beattie, A., Embree, M., Sorensen, D.C.: Convergence of polynomial restart Krylov methods for eigenvalue computation. Rice University, Report 03-08 (2003)
2. Bellman, R.E.: Introduction to the Mathematical Theory of Control Processes. Nonlinear processes, vol. II. Academic Press, New York (1971)
3. Bhaya, A., Kaszkurewicz, E.: Control Perspectives on Numerical Algorithms and Matrix Problems. In: Advances in Design and Control, vol. 10. SIAM, Philadelphia (2006)

4. Calvetti, D., Reichel, L.: An adaptive Richardson iteration method for indefinite linear systems. Numerical Algorithms 12, 125–149 (1996)
5. Eiermann, M., Ernst, O.G., Schneider, O.: Analysis of acceleration strategies for restarted minimal residual methods. J. Comp. Appl. Math. 123, 261–292 (2000)
6. Embree, M.: The tortoise and the hare restart GMRES. SIAM Review 45, 259–266 (2003)
7. Fuhrmann, P.A.: A Polynomial Approach to Linear Algebra. Springer, New York (1996)
8. Gustafsson, K., Lundh, M., Söderlind, G.S.: A PI stepsize control for the numerical solution of ordinary differential equations. BIT 28, 270–287 (1988)
9. Gustafsson, K.: Control theoretic techniques for stepsize selection in explicit Runge-Kutta methods. ACM Trans. Math. Softw. 17(4), 533–554 (1991)
10. Golub, G.H., Overton, M.L.: The convergence of inexact Chebyshev and Richardson iterative methods for solving linear systems. Numer. Math. 53, 571–593 (1988)
11. Helmke, U., Fuhrmann, P.A.: Controllability of matrix eigenvalue algorithms: the inverse power method. Systems & Control Letters 41, 57–66 (2000)
12. Helmke, U., Jordan, J.: Numerics versus control. In: Rosenthal, J., Gilliam, D. (eds.) Mathematical Systems Theory in Biology, Communications, Computations and Finance. IMA Volumes in Mathematics and its Applications, vol. 134, pp. 223–236. Springer, New York (2002)
13. Helmke, U., Jordan, J.: Optimal control of iterative solution methods for linear systems of equations. Proc. Appl. Math. Mech. 5(1), 163–164 (2005)
14. Helmke, U., Jordan, J., Lanzon, A.: A control theory approach to linear equation solvers. In: Proceedings of 17th International Symposium on Mathematical Theory of Network and Systems (MTNS), Kyoto, Japan (2006)
15. Helmke, U., Wirth, F.: On controllability of the real shifted inverse power iteration. Systems & Control Letters 43, 9–23 (2001)
16. Jordan, J.: Reachable sets of numerical iteration schemes. Ph.D. thesis, University of Würzburg (2008)
17. Joubert, W.: On the convergence behavior of the restarted GMRES algorithm for solving nonsymmetric linear systems. Numerical Linear Algebra with Applications 1(5), 427–447 (1994)
18. Kashima, K., Yamamoto, Y.: System theory for numerical analysis. Automatica 43(7), 1156–1164 (2007)
19. Opfer, G., Schober, G.: Richardson's iteration for nonsymmetric matrices. Linear Algebra and its Appl. 58, 343–361 (1984)
20. Smolarski, D.C., Saylor, P.E.: An optimum iterative method for solving any linear systems with a square matrix. BIT 28, 163–178 (1988)
21. Sorensen, D.C.: Numerical methods for large eigenvalue problems. Acta Numerica 11, 519–584 (2002)
22. Wakasa, Y., Yamamoto, Y.: An iterative method for solving linear equations by using robust methods. In: Proceedings of SICE first Annual Conference on Control Systems, pp. 451–454 (2001)

Nonlinear Output Regulation: A Unified Design Philosophy

Alberto Isidori

Abstract. In this paper, we extend to the case of higher relative degree and uncertain high-frequency gain a recently proposed method for the solution of problems of output regulation for nonlinear systems. The method is applicable to systems with possibly unstable zero dynamics, in which case the problem at issue is addressed by rendering an appropriate auxiliary system stable, with a constrain on the gain. The dynamics of this auxiliary system only depend on the zero dynamics of the controlled plant, while the constraint on the gain only depends on some parameters of the exogenous input which have to be tracked.

Keywords: nonlinear output regulation, nonminimum-phase systems, robust control.

This paper is dedicated to Yutaka Yamamoto on the occasion of his 60-th birthday.

1 Introduction

In this Chapter, we present a unified approach to the solution of problems of *output regulation* for minimum-phase and non-minimum-phase nonlinear systems. We consider a system having relative degree $r > 1$ between *control input* $u \in \mathbb{R}$ and *regulated output* $e \in \mathbb{R}$ described in normal form as

Alberto Isidori
Dipartimento di Informatica e Sistemistica, Università di Roma "La Sapienza",
Via Ariosto 25, 00185 Rome, Italy
C.A.SY. — Dipartimento di Elettronica, Informatica e Sistemistica,
University of Bologna, 40136 Bologna, Italy
e-mail: albisidori@dis.uniroma1.it

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 83–94.
springerlink.com

$$\begin{aligned} \dot{w} &= s(w) \\ \dot{z} &= f(w,z,\xi,\zeta) \\ \dot{\xi} &= A\xi + B\zeta \\ \dot{\zeta} &= q(w,z,\xi,\zeta) + b(w,z,\xi,\zeta)u \\ e &= C\xi \end{aligned} \tag{1}$$

in which $z \in \mathbb{R}^m$, $\xi \in \mathbb{R}^{r-1}$, and

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdot & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 \\ 0 \\ \cdot \\ 0 \\ 1 \end{pmatrix}, \qquad C = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Here, $w \in \mathbb{R}^s$ is a vector of inputs which cannot be controlled and include *exogenous* commands, disturbances and model uncertainties. This exogenous input is assumed to be a (undefined) member of the family of all solutions of a fixed ordinary differential equation of the form

$$\dot{w} = s(w) \tag{2}$$

obtained when the initial condition $w(0)$ is allowed to vary on a prescribed set W. This system is usually referred to as the *exosystem*. The initial states of (1) and (2) are assumed to range over fixed compact sets X and W, the latter being invariant under the dynamics of (2). Motivated by well-known standard design procedures (see e.g. [5]), we assume throughout that the *measured output* y coincides with the entire *partial state* $(\xi_1,\dots,\xi_{r-1},\zeta)$ i.e.

$$y = \mathrm{col}(\xi_1,\dots,\xi_{r-1},\zeta).$$

The states w and z, on the contrary, *are not* available for measurement and this is what makes the problem challenging.

The functions characterizing the model (1) are assumed to be smooth functions of their arguments. In addition, we assume the existence of a pair of real numbers (b_0,b_1) such that

$$0 < b_0 \le b(w,z,\xi,\zeta) \le b_1. \tag{3}$$

The problem of output regulation is to design a controller

$$\begin{aligned} \dot{\xi} &= \varphi(\xi,y) \\ u &= \gamma(\xi,y) \end{aligned}$$

with initial state in a compact set Ξ, yielding a closed-loop system in which

- the positive orbit of $W \times X \times \Xi$ is bounded,
- $\lim_{t\to\infty} e(t) = 0$, uniformly in the initial condition (on $W \times X \times \Xi$).

We observe that, in the general setup presented above, the vector w of exogenous inputs may well include (constant) uncertain parameters, which are hence assumed to range on a given compact set. Thus, if a controller solves the problem at issue, the goal of asymptotic regulation is achieved robustly with respect to (constant) parameter uncertainties.

The theory of output regulation of nonlinear systems, which uses a combination of geometry and nonlinear dynamical systems theory, was initiated by pioneering works of [9, 8]. Since these early contributions, the theory has experienced a tremendous growth, culminating in the recent development of design methods able to handle the case of parametric uncertainties affecting the autonomous (linear) system which generates the exogenous signals (such as in [15]), the case of nonlinear exogenous systems (such as in [2]), or a combination thereof (as in [13]). A thorough presentation of several recent advances in this area can also be found in the recent books [11, 7, 14]. Most of these contributions, though, only address the case of systems having a stable zero dynamics. In the recent paper [3], a method applicable to a class of systems with possibly unstable zero dynamics was proposed, which is based on a re-interpretation of a method discussed earlier in [12]. Systems considered in [3] were systems having relative degree 1 and unitary high-frequency gain. In this paper we extend the method to the case of higher relative degree and uncertain high-frequency gain. If the zero dynamics of the systems are stable, the method essentially reduces to the design method of [13]. On the other hand, if the zero dynamics of the system are unstable, the method consists in rendering an appropriate auxiliary system stable, while respecting a constrain on a suitably defined gain. The dynamics of the auxiliary system only depend on the zero dynamics of the controlled plant, while the constraint on the gain only depends on some parameters of the exogenous input which have to be tracked.

2 The Plant and the Basic Assumptions

The point of departure for the solution of the problem is, as usual, the assumption of the existence of a (smooth) manifold which can be rendered invariant by feedback and on which the regulated output vanishes (see [9]). In the case of system (1), this amounts to the assumption of the existence of a smooth map $\pi : W \to \mathbb{R}^m$ satisfying

$$\frac{\partial \pi}{\partial w} s(w) = f(w, \pi(w), 0, 0) \qquad \forall w \in W\,. \tag{4}$$

This being the case, it is readily seen that the set

$$\mathscr{S}^* = \{(w, z, \xi, \zeta) : w \in W, z = \pi(w), \xi = 0, \zeta = 0\}$$

is rendered invariant by the control

$$u^*(w) = -\frac{q(w, \pi(w), 0, 0)}{b(w, \pi(w), 0, 0)} \tag{5}$$

and, indeed, the regulated variable $e = \xi_1$ vanishes on this set. The input $u^*(w)$ is the input which forces e to remain identically zero.

The second step in the solution of the problem usually consists in making assumptions that make it possible to build an "internal" model for the control $u^*(w)$. In a series of recent papers, it was shown how these assumptions could be progressively weakened, moving from the so-called assumption of "immersion into a linear observable system" (as in [8]) to "immersion into a nonlinear uniformly observable system" (as in [2]) to the recent results of [13], in which it was shown that no assumption is in fact needed for the construction of an internal model if only continuous (thus possibly not locally Lipschitz) controllers are acceptable. Motivated by these latest developments we assume, in what follows, the existence of $d \in \mathbb{N}$, a map $F : \mathbb{R}^d \to \mathbb{R}^d$, a $d \times 1$ column vector G_0, a map $\gamma : \mathbb{R}^d \to \mathbb{R}$ and a map $\tau : W \to \mathbb{R}^d$ satisfying

$$\begin{aligned} \frac{\partial \tau}{\partial w} s(w) &= F(\tau(w)) + G_0 \gamma(\tau(w)) && \forall w \in W \\ u^*(w) &= \gamma(\tau(w)) && \forall w \in W\,. \end{aligned} \tag{6}$$

Coherently with the assumptions on (1), $F(\cdot)$, $\gamma(\cdot)$ and $\tau(\cdot)$ are assumed to be smooth maps.

Remark 1. Under the (mild) assumption that

$$L_s^d u^*(w) = \phi(u^*(w), L_s u^*(w), \ldots, L_s^{d-1} u^*(w))\,, \tag{7}$$

for some $d \in \mathbb{N}$ and some smooth $\phi(x_1, \ldots, x_d)$ and all $w \in W$, conditions (6) can be fulfilled by taking

$$F(x) = \begin{pmatrix} x_2 \\ \cdots \\ x_d \\ \phi(x_1, \ldots, x_d) \end{pmatrix} - G_0 x_1 \qquad \gamma(x) = x_1\,, \tag{8}$$

in which case

$$\tau(w) = \mathrm{col}(u^*(w), \ldots, L_s^{d-1} u^*(w))\,.$$

This includes the (classical) case of linear internal models. Recent advances in the theory of nonlinear observers (see e.g. [13]) show that, if d is large enough, and $F(x) = F_0 x$ with F_0 Hurwitz and (F_0, G_0) controllable, a C^1 map $\tau(\cdot)$ and a C^0 map $\gamma(\cdot)$ which do fulfill (6) always exist. ◁

3 The Control

We consider, in what follows, a dynamic controller, with internal state (φ, η), "driven" by the measured variables (ξ, ζ). The control in question is modelled by equations of the form

$$\begin{aligned} u &= \gamma(\eta) + \beta \dot{N}(\varphi) + v \\ \dot{\varphi} &= L(\varphi) - Mv \\ \dot{\eta} &= F(\eta) + G_0[\gamma(\eta) + v] \\ v &= -k[\zeta - H\xi - N(\varphi)] \end{aligned} \tag{9}$$

in which $F(\cdot), G_0, \gamma(\cdot)$ satisfy (6) for some $\tau(\cdot)$, while $L(\cdot), N(\cdot)$, M, H are smooth maps and, respectively, vectors of appropriate dimensions, and β, k are real numbers. It is assumed (without loss of generality) that

$$\frac{\partial N}{\partial \varphi} M = 0 \tag{10}$$

in which case

$$\dot{N}(\varphi) = \frac{\partial N}{\partial \varphi} L(\varphi) \,.$$

Changing ζ into

$$\theta = \zeta - H\xi - N(\varphi)$$

yields a closed-loop system of the form

$$\begin{aligned} \dot{w} &= s(w) \\ \dot{z} &= f(w, z, \xi, H\xi + N(\varphi) + \theta) \\ \dot{\xi} &= A\xi + B[H\xi + N(\varphi) + \theta] \\ \dot{\varphi} &= L(\varphi) - Mv \\ \dot{\eta} &= F(\eta) + G_0[\gamma(\eta) + v] \\ \dot{\theta} &= Q(w, z, \xi, H\xi + N(\varphi) + \theta) + b(w, z, \xi, H\xi + N(\varphi) + \theta)[\gamma(\eta) + v] \\ &\qquad + \Delta(w, z, \xi, H\xi + N(\varphi) + \theta)\dot{N}(\varphi) \,, \end{aligned}$$

in which we have set

$$\begin{aligned} Q(w, z, \xi, \zeta) &= q(w, z, \xi, \zeta) - H(A\xi + B\zeta) \\ \Delta(w, z, \xi, \zeta) &= b(w, z, \xi, \zeta)\beta - 1 \,, \end{aligned}$$

with control

$$v = -k\theta \,.$$

Remark 2. Note that, in case the coefficient $b(w, z, \xi, \zeta)$ only depends on the measured variables (ξ, ζ), one can choose $\beta = 1/b$, obtaining in this way $\Delta(w, z, \xi, \zeta) = 0$. ◁

The system thus obtained can be regarded as a system with input v and output θ, having relative degree 1, in which v to is chosen as $v = -k\theta$, that is as a negative output feedback. To facilitate the analysis, we bring this system in normal form.

Since $b(w,z,\xi,\zeta)$ is bounded as in (3), by the method of the characteristics one can obtain a globally defined change of coordinates

$$X \;:\; \eta \;\mapsto\; x = X(w,z,\xi,\varphi,\eta,\theta)$$

in which X satisfies

$$\frac{\partial X}{\partial \eta} G_0 + \frac{\partial X}{\partial \theta} b(w,z,\xi,H\xi + N(\varphi) + \theta) = 0\,.$$

At $\theta = 0$, the map X is the identity map, namely $X(w,z,\xi,\varphi,\eta,0) = \eta$ which in turn implies

$$\left[\frac{\partial X}{\partial \eta}\right]_{\theta=0} = I\,.$$

Actually, it is not difficult to find a closed form for X, which turns out to be

$$X(w,z,\xi,\varphi,\eta,\theta) = \eta - G_0 \int_0^\theta \frac{1}{b(w,z,\xi,H\xi + N(\varphi) + t)}\,dt\,.$$

From this, using our earlier assumption (10), it is readily seen that

$$\frac{\partial X}{\partial \varphi} M = 0\,.$$

Likewise, by the method of the characteristics one can obtain a globally defined change of coordinates

$$K \;:\; \varphi \;\mapsto\; \chi = K(w,z,\xi,\varphi,\theta)$$

in which K satisfies

$$\frac{\partial K}{\partial \varphi} M - \frac{\partial K}{\partial \theta} b(w,z,\xi,H\xi + N(\varphi) + \theta) = 0\,.$$

At $\theta = 0$, the map K is the identity map, namely $K(w,z,\xi,\varphi,0) = \varphi$ which in turn implies

$$\left[\frac{\partial K}{\partial \varphi}\right]_{\theta=0} = I\,.$$

The inverses of K and X define a pair of maps

$$\begin{aligned} \varphi &= \hat{K}(w,z,\xi,\chi,\theta) \\ \eta &= \hat{X}(w,z,\xi,\chi,x,\theta) \end{aligned}$$

which, at $\theta = 0$, are identities in χ and – respectively – in x, that is

$$\hat{K}(w,z,\xi,\chi,0) = \chi\,, \qquad \hat{X}(w,z,\xi,\chi,x,0) = x\,.$$

Changing coordinates in this way yields a system of the form

$$\begin{aligned}
\dot w &= s(w)\\
\dot z &= f(w,z,\xi,H\xi+N(\hat K)+\theta)\\
\dot \xi &= A\xi+B[H\xi+N(\hat K)+\theta]\\
\dot \chi &= \frac{\partial K}{\partial \varphi}\Big[L(\hat K)+M\Big(\frac{Q}{b}(w,z,\xi,\theta+N(\hat K)+H\xi)+\frac{\Delta}{b}\dot N(\hat K)+\gamma(\hat X)\Big)\Big]+R_\chi \qquad (11)\\
\dot x &= F(\hat X)-G_0\Big(\frac{Q}{b}(w,z,\xi,\theta+N(\hat K)+H\xi)+\frac{\Delta}{b}\dot N(\hat K)\Big)+R_x\\
\dot \theta &= Q(w,z,\xi,H\xi+N(\hat K)+\theta)+b(w,z,\xi,H\xi+N(\hat K)+\theta)[\gamma(\hat X)+v]\\
&\qquad +\Delta(w,z,\xi,H\xi+N(\hat K)+\theta)\dot N(\hat K),
\end{aligned}$$

in which, for readability, we have omitted the indication of the arguments of $\hat K$, $\hat X$ and Δ/b, and we have set

$$\begin{aligned}
R_\chi &= \frac{\partial K}{\partial w}s(w)+\frac{\partial K}{\partial z}f(w,z,\xi,H\xi+N(\hat K)+\theta)+\frac{\partial K}{\partial \xi}(A\xi+B[H\xi+N(\varphi)+\theta])\\
R_x &= \frac{\partial X}{\partial w}s(w)+\frac{\partial X}{\partial z}f(w,z,\xi,H\xi+N(\hat K)+\theta)+\frac{\partial X}{\partial \xi}(A\xi+B[H\xi+N(\varphi)+\theta])\\
&\quad +\frac{\partial X}{\partial \varphi}L(\hat K)\,.
\end{aligned}$$

Note that at $\theta=0$ both these terms vanish, because at $\theta=0$ the map K is simply an identity in φ and the map X is simply an identity in η.

The system obtained in this way can be seen as feedback interconnection of a system with input θ and state (w,z,ξ,χ,x) and of a system with input (w,z,ξ,χ,x) and state θ. As a matter of fact, setting

$$\mathbf{p}=\mathrm{col}\{w,z,\xi,\chi,x\}$$

the system cam be viewed as a system of the form

$$\begin{aligned}
\dot{\mathbf{p}} &= \mathbf{F}(\mathbf{p})+\mathbf{G}(\mathbf{p},\theta)\theta\\
\dot\theta &= \mathbf{H}(\mathbf{p})+\mathbf{H}(\mathbf{p},\theta)\theta+\mathbf{b}(\mathbf{p},\theta)v
\end{aligned} \qquad (12)$$

with control to be chosen as

$$v=-k\theta\,. \qquad (13)$$

The advantage of seeing system (11) in this form is that we can appeal to the following result (see e.g. [13]).

Theorem 1. *Consider a system of the form* (12) *with v as in* (13)*. The functions* $\mathbf{F}(\mathbf{p})$ *and* $\mathbf{H}(\mathbf{p})$ *are locally Lipschitz and the functions* $\mathbf{G}(\mathbf{p},\theta)$ *and* $\mathbf{H}(\mathbf{p},\theta)$ *are continuous. Let* $\mathbf{P}$ *be an arbitrary fixed compact set. Suppose that* $\mathbf{b}(\mathbf{p},\theta) > 0$ *for all* $(\mathbf{p},\theta)$*. Suppose there exists a set* $\mathscr{A}$ *which is locally exponentially stable for*

$$\dot{\mathbf{p}} = \mathbf{F}(\mathbf{p}),$$

with a domain of attraction that contains the set $\mathbf{P}$*. Suppose also that*

$$\mathbf{H}(\mathbf{p}) = 0, \qquad \forall \mathbf{p} \in \mathscr{A}.$$

Then, for any choice of a compact set Θ*, there is a number* k^* *such that, for all* $k > k^*$*, the set* $\mathscr{A} \times \{0\}$ *is locally exponentially stable, with a domain of attraction that contains* $\mathbf{P} \times \Theta$*.*

If the assumption of this Theorem are fulfilled and, in addition, the regulated variable $e = \xi_1$ vanishes $\mathscr{A}$, we conclude that the proposed controller is able to solve the problem of output regulation.

4 The Structure of the Core Subsystem

All of the above suggests the use of the degrees of freedom in the choice of the parameters of the controller in order to fulfill the hypotheses of Theorem 1. At $\theta = 0$ we have

$$\hat{K} = \chi, \quad \hat{X} = x, \quad \frac{\partial K}{\partial \varphi} = 0, \quad R_\chi = 0, \quad R_x = 0,$$

and hence, in the (w,z,ξ,χ) coordinates, system $\dot{\mathbf{p}} = \mathbf{F}(\mathbf{p})$ reads as

$$\begin{aligned}
\dot{w} &= s(w)\\
\dot{z} &= f(w,z,\xi,H\xi + N(\chi))\\
\dot{\xi} &= A\xi + B(H\xi + N(\chi))\\
\dot{\chi} &= L(\chi) + M\Big(\frac{Q}{b}(w,z,\xi,H\xi + N(\chi)) + \frac{\Delta}{b}\dot{N}(\chi) + \gamma(x)\Big)\\
\dot{x} &= F(x) - G_0\Big(\frac{Q}{b}(w,z,\xi,H\xi + N(\chi)) + \frac{\Delta}{b}\dot{N}(\chi)\Big).
\end{aligned} \tag{14}$$

while the map $\mathbf{H}(\mathbf{p})$ reads as

$$\begin{aligned}
&Q(w,z,\xi,H\xi + N(\chi)) + b(w,z,\xi,H\xi + N(\chi))\gamma(x)\\
&\qquad + \Delta(w,z,\xi,H\xi + N(\chi))\dot{N}(\chi).
\end{aligned} \tag{15}$$

Theorem 1 above identifies an auxiliary problem which, if solved, makes it possible to use the controller (9) for the solution of the problem of output regulation for the original plant: shape the internal model $\{F(x), G_0, \gamma(x)\}$ and find, if possible, a triplet $\{L(\chi), M, N(\chi)\}$ in such a way that system (14) possesses a compact invariant set $\mathscr{A}$ which is locally exponentially stable and attracts all admissible initial conditions, and that both ξ_1 and the map (15) vanish on this set.

Recall now that, by assumption, there exists $\pi(w)$ and $\tau(w)$ satisfying (4) and (6). Hence, it is readily seen that if $L(0) = 0$ and $N(0) = 0$, the set

$$\mathscr{A} = \{(w, z, \xi, \chi, x) : w \in W,\ z = \pi(w),\ \xi = 0,\ \chi = 0,\ x = \tau(w)\}$$

is a compact invariant set of (14). Moreover, by construction, the map (15) vanishes on this set. Trivially, also ξ_1 vanishes on this set. Thus, it is concluded that if the set $\mathscr{A}$ can be made local exponentially stable, with a domain of attraction that contains the compact set of all admissible initial conditions, the proposed controller, for large k, solves the problem of output regulation.

System (14) is not terribly difficult to handle. As a matter of fact, it can be regarded as interconnection of three much simpler subsystems. To see this, set

$$\begin{aligned} z_{\mathrm{a}} &= z - \pi(w) \\ \tilde{x} &= x - \tau(w) \end{aligned}$$

and define

$$f_{\mathrm{a}}(w, z_{\mathrm{a}}, \xi, \zeta) = f(w, z_{\mathrm{a}} + \pi(w), \xi, \zeta) - f(w, \pi(w), 0, 0)$$

$$h_{\mathrm{a}}(w, z_{\mathrm{a}}, \xi, \zeta) = \frac{Q}{b}(w, z_{\mathrm{a}} + \pi(w), \xi, \zeta) - \frac{Q}{b}(w, \pi(w), 0, 0)$$

and

$$\Delta_{\mathrm{a}}(w, z_{\mathrm{a}}, \xi, \zeta) = \frac{\Delta}{b}(w, z_{\mathrm{a}} + \pi(w), \xi, \zeta) = \beta - \frac{1}{b(w, z_{\mathrm{a}} + \pi(w), \xi, \zeta)}\,.$$

In the new coordinates thus introduced, the invariant manifold $\mathscr{A}$ is simply the set

$$\mathscr{A} = \{(w, z_{\mathrm{a}}, \xi, \chi, \tilde{x}) : w \in W,\ (z_{\mathrm{a}}, \xi, \chi, \tilde{x}) = (0, 0, 0, 0)\}\,.$$

Bearing in mind (4), (6) and (5), it is readily seen that

$$\dot{z}_{\mathrm{a}} = f_{\mathrm{a}}(w, z_{\mathrm{a}}, \xi, H\xi + N(\chi))$$

and

$$\frac{Q}{b}(w, z, \xi, H\xi + N(\chi)) = h_{\mathrm{a}}(w, z_{\mathrm{a}}, \xi, H\xi + N(\chi)) - \gamma(\tau(w))\,.$$

In view of this, using again (6), the core subsystem (14) can be seen as a system with input $\bar{u}$ and output $\bar{y}$ defined as

$$
\begin{aligned}
\dot{w} &= s(w)\\
\dot{z}_{\mathrm{a}} &= f_{\mathrm{a}}(w, z_{\mathrm{a}}, \xi, H\xi + N(\chi))\\
\dot{\xi} &= A\xi + B(H\xi + N(\chi))\\
\dot{\chi} &= L(\chi) + M[h_{\mathrm{a}}(w, z_{\mathrm{a}}, \xi, H\xi + N(\chi)) + \Delta_{\mathrm{a}}(w, z_{\mathrm{a}}, \xi, H\xi + N(\chi))\dot{N}(\chi) + \bar{u}]\\
\dot{\tilde{x}} &= F(\tilde{x} + \tau(w)) - F(\tau(w)) - G_0[h_{\mathrm{a}}(w, z_{\mathrm{a}}, \xi, H\xi + N(\chi))\\
&\qquad + \Delta_{\mathrm{a}}(w, z_{\mathrm{a}}, \xi, H\xi + N(\chi))\dot{N}(\chi)]\\
\bar{y} &= \gamma(\tilde{x} + \tau(w)) - \gamma(\tau(w))
\end{aligned}
\tag{16}
$$

subject to unitary output feedback $\bar{u} = \bar{y}$.

System (16), in turn, can be seen as the cascade of an "inner loop" consisting of a subsystem, which we call the "auxiliary plant", modelled by equations of the form

$$
\begin{aligned}
\dot{w} &= s(w)\\
\dot{z}_{\mathrm{a}} &= f_{\mathrm{a}}(w, z_{\mathrm{a}}, \xi, H\xi + u_{\mathrm{a}})\\
\dot{\xi} &= (A + BH)\xi + Bu_{\mathrm{a}}\\
y_{\mathrm{a}} &= h_{\mathrm{a}}(w, z_{\mathrm{a}}, \xi, H\xi + u_{\mathrm{a}}) + \Delta_{\mathrm{a}}(w, z_{\mathrm{a}}, \xi, H\xi + u_{\mathrm{a}})v_{\mathrm{a}},
\end{aligned}
\tag{17}
$$

controlled by

$$
\begin{aligned}
\dot{\chi} &= L(\chi) + M[y_{\mathrm{a}} + \bar{u}]\\
u_{\mathrm{a}} &= N(\chi)\\
v_{\mathrm{a}} &= \dot{N}(\chi),
\end{aligned}
\tag{18}
$$

cascaded with a system, which we call a "weighting filter", modelled by equations of the form

$$
\begin{aligned}
\dot{\tilde{x}} &= F(\tilde{x} + \tau(w)) - F(\tau(w)) - G_0 y_{\mathrm{a}}\\
\bar{y} &= \gamma(\tilde{x} + \tau(w)) - \gamma(\tau(w)).
\end{aligned}
\tag{19}
$$

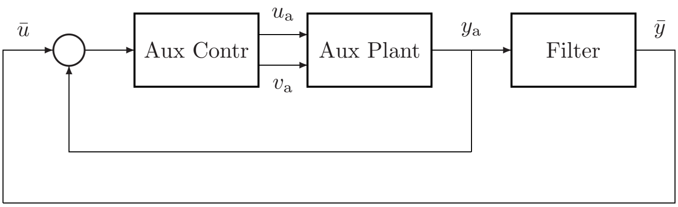


Fig. 1 The feedback structure of system (14)

All of this is depicted in Fig. 1. Having interpreted system (14) as the system resulting from a unitary output feedback on system (16), the idea is now to use the degrees of freedom in the design to make the latter a stable system and to force its gain to be a simple contraction.

5 The Asymptotic Properties of the Core Subsystem

System (16) is the cascade of two subsystems: the “inner loop”, consisting of (17) and (18), and the “filter” (19). An obvious prerequisite for stability is the stability of both subsystems of the cascade. Stability of the filter (19) is not an issue. As a matter of fact, appealing to the results of [6] and proceeding as in [2], it is not difficult to prove the existence of a filter (with $F(\cdot)$ and $\gamma(\cdot)$ as in Remark 1) which is globally input-to-state stable, actually with a linear gain function.

As far as the inner loop is concerned, the simplest situation in which the design paradigm outlined above can be successfully implemented is the case in which the controlled plant is globally asympotically and locally exponentially *minimum phase*, i.e. satisfies the following assumptions (see e.g. [1]):

- the function $f(w,z,\xi,\zeta)$ in (1) does not depend on $\xi_2,\xi_3,\ldots,\xi_{r-1},\zeta$
- there exists a smooth positive definite and proper function $V(z_\mathrm{a})$, with quadratic bounds for small $|z_\mathrm{a}|$, satisfying

$$\frac{\partial V}{\partial z_\mathrm{a}} f_\mathrm{a}(w,z_\mathrm{a},0,0) \leq -\alpha(|z_\mathrm{a}|)$$

  some class $\mathscr{K}_\infty$ function $\alpha(\cdot)$ which is quadratic for small values of the argument.

In this case, in fact, setting $M=0, N=0$, and letting $\dot{\chi}=L(\chi)$ to be any arbitrary globally stable system, it is always possible, by known methods, to find a vector H that makes the inner loop stable (in a semiglobal sense) with an arbitrarily small linear gain function.

If, on the contrary, the plant *is not* minimum-phase, a more sophisticated design is necessary, seeking $L(\cdot),M,N(\cdot)$ and H in such a way as to obtain — whenever possible — a stable inner loop, with a gain function which, composed with the gain function of the filter (19), would respect the small gain condition required for the stability of (14). A number of relevant cases in which this is possible have been recently presented in the literature (see [3] and [4]). They include the complete solution of the problem in the case of a (linear) controlled plant having an arbitrary number of zeros at the origin (while all other zeros have negative real part) and a discussion of the case in which the controlled plant has a zero with positive real part. In the latter case, it has been shown that the method is applicable if the frequencies which characterize the harmonic components of the exogenous input exceed a minimal value determined by the gain needed to make the inner loop stable. This is a nice example showing that, in a non-minimum phase system, a tradeoff exists between stability and performance. In fact, the minimal gain needed to stabilize the unstable zero dynamics of the original plant determines a *lower limit* on the frequencies of the exogenous inputs for which the desired *tracking properties* can be achieved.

References

1. Byrnes, C.I., Isidori, A.: Asymptotic stabilization of minimum phase nonlinear systems. IEEE Trans. Automat. Control 36, 1122–1137 (1991)
2. Byrnes, C.I., Isidori, A.: Nonlinear internal models for output regulation. IEEE Trans. Automat. Control 49, 2244–2247 (2004)
3. Delli Priscoli, F., Marconi, L., Isidori, A.: A dissipativity-based approach to output regulation of non-minimum-phase systems. Systems & Control Lett. 58, 584–591 (2009)
4. Delli Priscoli, F., Marconi, L., Isidori, A.: A method for robust regulation of non-mimimum-phase linear systems. In: Proc. IFAC ROCOND (2009)
5. Esfandiari, F., Khalil, H.: Output feedback stabilization of fully linearizable systems. Int. J. Control 56, 1007–1037 (1992)
6. Gauthier, J.P., Kupka, I.: Deterministic Observation Theory and Applications. Cambridge Univ. Press, Cambridge (2001)
7. Huang, J.: Nonlinear Output Regulation: Theory and Applications. SIAM, Philadephia (2004)
8. Huang, J., Lin, C.F.: On a robust nonlinear multivariable servomechanism problem. IEEE Trans. Automat. Control 39, 1510–1513 (1994)
9. Isidori, A., Byrnes, C.I.: Output regulation of nonlinear systems. IEEE Trans. Automat. Control 25, 131–140 (1990)
10. Isidori, A.: Nonlinear Control Systems, vol. II. Springer, London (1999)
11. Isidori, A., Marconi, L., Serrani, A.: Robust Autonomous Guidance: An Internal-Model Approach. Springer, London (2003)
12. Marconi, L., Isidori, A., Serrani, A.: Non-resonance conditions for uniform observability in the problem of nonlinear output regulation. Systems & Control Lett. 53, 281–298 (2004)
13. Marconi, L., Praly, L., Isidori, A.: Output stabilization via nonlinear Luenberger observers. SIAM J. Control and Optimization 45, 2277–2298 (2006)
14. Pavlov, A., van de Wouw, N., Nijmeijer, H.: Uniform Output Regulation of Nonlinear Systems: A Convergent Dynamics Approach. Birkhäuser, Boston (2006)
15. Serrani, A., Isidori, A., Marconi, L.: Semiglobal nonlinear output regulation with adaptive internal model. IEEE Trans. Autom. Control 46, 1178–1194 (2001)

An Estimated General Cross Validation Function for Periodic Control Theoretic Smoothing Splines

Maja Karasalo, Xiaoming Hu, and Clyde F. Martin

Abstract. In this paper, a method is developed for estimating the optimal smoothing parameter ε for periodic control theoretic smoothing splines. The procedure is based on general cross validation (GCV) and requires no a priori information about the underlying curve or level of noise in the measurements. The optimal ε is the minimizer of an estimated GCV cost function, which is derived based on a discretization of the L_2 smoothing problem for periodic control theoretic smoothing splines.

1 Introduction

In this paper, we consider the problem of estimating representations of objects or contours using a type of continuous closed curves, the periodic control theoretic smoothing splines. The splines are retrieved from noisy measurements of an unknown, underlying contour. Intended applications include mapping, identification and path planning for autonomous agents. The focus of this paper is the issue of the level of smoothing, determined by the magnitude of the so called *smoothing parameter* ε.

It is well known that an interpolating spline generated from noisy measurements yields a poor estimate of the underlying curve, as the spline will pass through every measurement point. An interpolating spline may be regarded as a smoothing spline with $\varepsilon = \infty$, while a periodic smoothing spline with $\varepsilon \to 0$ approaches a circle. The theory of regular smoothing splines, and the particular issue of choice of smoothing parameter, is treated in [13, 14, 15]. Control theoretic smoothing splines have been

Maja Karasalo and Xiaoming Hu
Department of Mathematics, Royal Institute of Technology, Stockholm, Sweden
e-mail: {karasalo,hu}@math.kth.se

Clyde F. Martin
Department of Mathematics and Statistics, Texas Tech University, Lubbock, Texas, USA
e-mail: clyde.f.martin@ttu.edu

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 95–104.
springerlink.com

studied in [1]–[7] and it has been shown in [5] that such splines, where the curve is found through minimizing a cost function, act as filters and are better suited for noisy measurements. A thorough treatment of control theoretic smoothing splines is provided in the book [8].

The particular type of periodic control theoretic smoothing spline explored in this paper has been previously presented in [9, 10, 11]. These publications cover error convergence properties for a recursive formulation of the smoothing spline problem. Experimental results indicate that the convergence is fairly robust with respect to the choice of smoothing parameter ε, but a formal method of finding the optimal value of ε has so far been lacking in our work. In this paper, a method is developed for determining the appropriate level of smoothing, assuming the shape of the underlying contour, as well as the level of noise in the measurements, is unknown. We propose an estimate of the general cross validation (GCV) function, based on the estimated influence matrix for the smoothing spline problem. A general expression for the influence matrix based on Bernoulli polynomials is derived in [12]. However, this expression is computationally heavy. In [16, 17, 18], the trace of the influence matrix is estimated and an estimate of the GCV function itself is obtained by Taylor expansion. [19, 20] and others use singular value decomposition to estimate the GCV function. In this paper, on the other hand, we derive an estimate of the influence matrix and GCV function, based directly upon a discretization of the underlying spline problem. The method is straightforward and easy to implement, and the accuracy can be chosen arbitrarily by adjusting the number of discretization points.

The paper is organized as follows. The contour estimation problem is formally stated in Section 2. In Section 3, we derive a discretization of the optimization problem. The GCV approach to optimal smoothing is reviewed in Section 4 and the specific estimated GCV function is introduced. This constitutes the main contribution of the paper. Finally, simulation results are presented in Section 5 and conclusions are drawn in Section 6.

2 Preliminaries

Consider the problem of reconstructing continuous, closed curves in $\mathbb{R}^2$ from noisy and sparse measurement data. This problem arises for instance in mapping applications for mobile robots. A formal problem statement follows.

Given a data set $D = \{(t_i, z_i) : i = 1, ..., N\}$, *where* $t_i \in [0, 2\pi]$ *is the* polar coordinate *angle and* z_i *is the radius in polar coordinates. If* $z_i = r(t_i) + \xi_i$, $\xi_i \in N(0, \sigma^2)$, *where* $r(t_i)$ *are samples from a closed smooth curve, how to find the function* $r(t)$ *that best represents the underlying curve, with respect to smoothness and closeness to measurement data?*

The solution is found by solving the following polar second derivative L_2 smoothing problem:

Problem 1

$$\min_{u(t)\in L_2[0,T]} \quad J(u,r)=\int_0^{2\pi} u(t)^2dt+\frac{\varepsilon^2}{N}\sum_{i=1}^{N}(r(t_i)-z_i)^2 \tag{1}$$

$$\begin{aligned}\text{s. t. } r''(t) &= u(t)\\ r(0) &= r(2\pi)\\ r'(0) &= r'(2\pi).\end{aligned} \tag{2}$$

The solution of Problem 1 is the optimal compromise between smoothness of the output curve, due to the integral term in $J(u,r)$, and faithfulness to the data set, due to the summation term. The magnitude of the smoothing parameter $\varepsilon > 0$ determines how much credibility is given to measurement data. A large value brings the spline close to the data points, while a small value yields a smoother spline and thus more filtering. The main topic of this paper is how to determine the best choice of ε, based on the discretization of Problem 1 reviewed in the next section.

3 Discretization

Using a proper choice of approximation formulas, Problem 1 is reduced to an unconstrained quadratic programming problem (*QP*), suitable for numeric implementation. With this particular choice of discretization the periodic boundary condition is embedded in the *QP*, facilitating the analysis of convexity and solvability for the problem.

Let the vectors $\hat{r}=\{\hat{r}_m\}$ and $\hat{u}=\{\hat{u}_m\}$ be the discretizations of the spline $r(t)$ and control $u(t)$, and let $\hat{t}=\{\hat{t}_m\}$ be the corresponding discretization of t. Here $m=1,...,M$ and the sampling rate h is defined so that $(M+1)h=2\pi$. We emphasize that $(\hat{t}_m,\hat{r}_m)$ are equidistant samples from the spline $r(t)$ while (t_i,z_i), $i=1,\dots,N$ are noisy measurement data from the true curve. Let $z=\{z_i\}$ denote the vector of radius measurements. Note that when the continuous function $r(t)$ is expressed in discretized form as a vector pair $(\hat{t},\hat{r})$, the periodicity constraint translates to $(\hat{t}_1,\hat{r}_1)=(\hat{t}_{M+1},\hat{r}_{M+1})$, where $M+1$ indicates the point after the last point of the vector. $\hat{u}_m$ can be approximately expressed as functions of $\hat{r}$, using numerical differentiation:

$$\hat{u}_m=\hat{r}''_m=(1/h^2)(\hat{r}_{m-1}-2\hat{r}_m+\hat{r}_{m+1}). \tag{3}$$

Construct the matrix Φ:

$$\begin{aligned}\Phi_{m,m-1} &= \Phi_{m,m+1}=1\\ \Phi_{1,M} &= \Phi_{M,1}=1\\ \Phi_{m,m} &= -2\\ \Phi_{j,l} &= 0 \quad \text{otherwise.}\end{aligned} \tag{4}$$

Then $\hat{u}=\frac{1}{h^2}\Phi\hat{r}$. Note that the periodicity is implicitly expressed in Φ. The discretization of the integral is

$$\int_0^{2\pi} u(t)^2 dt \approx \frac{1}{h^3} \hat{r}^T \Phi^T \Phi \hat{r}. \tag{5}$$

For the summation term, construct the $M \times N$ matrix F:

$$F_{m,i} = \begin{cases} 1 & \text{if } \hat{t}_m = t_i, \text{ for some } m \in [1,M],\ i \in [1,N] \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

F expands a vector of N measurements to an M-vector with the measurements inserted at the corresponding sampling times and zeros elsewhere. We get

$$\frac{\varepsilon^2}{N}\sum_{i=1}^{N}(r(t_i) - z_i)^2 \approx \frac{\varepsilon^2}{N}\left(\hat{r}^T F F^T \hat{r} + z^T z - 2(Fz)^T \hat{r}\right). \tag{7}$$

Remark 1. We assume that the sampling times are unique so that there is one data point for each value of i. In reality, t_i may not coincide with a grid point t_m, but with $M >> N$, choosing the grid point closest to t_i gives a negligible error.

Finally, the discretization of Problem 1 is

Problem 2

$$\min_{\hat{r}} J(\hat{r}) = \tfrac{1}{2}\hat{r}^T H \hat{r} + c^T \hat{r} \tag{8}$$

$$\begin{aligned} \text{where } H &= \tfrac{1}{h^3}\Phi^T \Phi + \tfrac{\varepsilon^2}{N} F F^T \\ c &= -\tfrac{\varepsilon^2}{N} F z, \end{aligned}$$

Proposition 1. *Define* $\mathbf{F} \triangleq (\varepsilon^2/N) F F^T$. *Assume that* $\mathbf{F}_{mm} > 0$ *for at least one value of* $m \in [1,M]$ *(this is equivalent to having a non-empty data set). Then the QP* (8) *has a unique solution.*

Proof. If H is positive definite, the *QP* (8) is strictly convex and has the unique solution

$$\hat{r} = -H^{-1} c. \tag{9}$$

The matrix $\frac{1}{h^3}\Phi^T \Phi$ is by construction a symmetric positive semidefinite matrix, but not positive definite since $\text{rank}(\Phi) = M - 1$. The single zero eigenvalue of Φ has the eigenvector $v_0 = [1, 1, \ldots, 1]^T$. $\mathbf{F}$ is diagonal with nonnegative elements and has rank N. Now, any $x \in \mathbb{R}^M$ can be decomposed as

$$x = v + \alpha v_0, \quad \alpha \in \mathbb{R}, \quad v \perp v_0, \tag{10}$$

so that

$$x^T \Phi^T \Phi x = v^T \Phi^T \Phi v \geq 0, \tag{11}$$

with equality only for $v = 0$. Therefore,

$$x^T H x = \frac{1}{h^3} v^T \Phi^T \Phi v + x^T \mathbf{F} x. \tag{12}$$

If x is nonzero, at least one of v and α is nonzero. If $v \neq 0$, $\frac{1}{h^3} v^T \Phi^T \Phi v > 0$, and consequently $x^T H x > 0$. If $v = 0$, $\alpha \neq 0$ we get

$$x^T H x = \alpha^2 v_0^T \mathbf{F} v_0 > 0. \tag{13}$$

□

In the next section, the general cross validation method is reviewed and a specific GCV function for the *QP* (8) is proposed, that estimates the optimal value of the smoothing parameter ε.

4 Generalized Cross-Validation

The general cross validation method for smoothing splines was first developed by Wahba *et al.* in for instance [12], [13]. Here, a review is provided for general cross validation as proposed by [12], where the smoothing parameter is estimated for a problem similar to Problem 1. Subsequently we derive an estimate of the general cross validation function for Problem 1.

Background

[12], [13] study a smoothing spline problem with solution $g_{N,\lambda}(t)$, defined by

$$g_{N,\lambda}(t) \triangleq \underset{g:g^{(n)}(t)\in L^2}{\arg\min}\ \lambda \int_0^1 (g^{(n)}(t))^2 dt + \frac{1}{N}\sum_{i=1}^{N}(g(t_i) - z_i)^2, \tag{14}$$

where $g^{(n)}$ is the n-th derivative of $g(t)$ and (t_i, z_i) are noisy samples of an underlying function $g_{true}(t)$. $g_{N,\lambda}(t)$ is a linear function of the data z_i. In particular, this means that

$$\left[g_{N,\lambda}(t_1), \ldots, g_{N,\lambda}(t_N)\right]^T = S(\lambda)\left[z_1, \ldots, z_N\right]^T \tag{15}$$

for a unique matrix $S(\lambda)$, denoted the *influence matrix*. $S_{i,j}(\lambda)$ is a measure of how much the measurement z_j influences the spline $g_{N,\lambda}$ at t_i.

Ideally, the smoothing parameter λ should be chosen to minimize the average square error, defined by

$$R(\lambda) \triangleq \frac{1}{N}\sum_{i=1}^{N}(g_{N,\lambda}(t_i) - g_{true}(t_i))^2. \tag{16}$$

However, minimizing $R(\lambda)$ requires knowledge of $g_{true}(t)$, at least at the measurement points. In [12] an estimate λ_{GCV} of the minimizer of (16) is found by minimizing the *Generalized Cross Validation* (GCV) function $V_{\mathrm{GCV}}(\lambda)$. Denote by $g_{N,\lambda}^{[k]}(t)$ the smoothing spline obtained when removing the k-th data point prior to minimizing the cost function (14). Then $V_{\mathrm{GCV}}(\lambda)$ is defined by

$$V_{\mathrm{GCV}}(\lambda) \triangleq \frac{1}{N}\sum_{k=1}^{N} \omega_k(\lambda)(g_{N,\lambda}^{[k]}(t_k) - z_k)^2 = \frac{\frac{1}{N}\|(I - S(\lambda))z\|^2}{\left(\frac{1}{N}\,\mathrm{trace}(I - S(\lambda))\right)^2}, \tag{17}$$

$$\text{for } \omega_k(\lambda) \triangleq \left(\frac{1 - s_{kk}(\lambda)}{\frac{1}{N}\,\mathrm{trace}(I - S(\lambda))}\right)^2, \tag{18}$$

where term k in the sum is a measure of how well the spline $g_{N,\lambda}^{[k]}(t)$ predicts the data point z_k, and $\omega_k(\lambda)$ is a weight that compensates for unequal spacing of the data. It is shown in [12] that $\lambda_{\mathrm{GCV}} = \arg\min_{\lambda}\ V_{\mathrm{GCV}}(\lambda)$ has the following appealing property:

$$\lim_{N\to\infty} \frac{E(R(\lambda_{\mathrm{GCV}}))}{\min_{\lambda} E(R(\lambda))} = 1, \tag{19}$$

where $E(\cdot)$ is the expectation value of $(\cdot)$. In other words, the expected mean square error using λ_{GCV} tends to the minimum possible expected mean square error as $N \longrightarrow \infty$.

In the next section, an estimate of the influence matrix for Problem 1 is derived.

GCV for Periodic Control Theoretic Smoothing Splines

In this section, we derive a GCV cost function for periodic control theoretic smoothing splines, based on an estimate $\hat{S}(\varepsilon)$ of the influence matrix for Problem 1. This constitutes the main result of the paper.

The estimate $\hat{S}(\varepsilon)$ is computed from the discretization reviewed in Section 3 and takes into consideration the constraints (2).

This paper follows the convention used in [9, 10, 11]. To clarify, the relation between the smoothing parameters in (1) and (14) is $\varepsilon^2 = 1/\lambda$. $\hat{S}(\varepsilon)$ should satisfy

$$\left[r_{N,\varepsilon}(t_1), \ldots, r_{N,\varepsilon}(t_N)\right]^T = \hat{S}(\varepsilon)\left[z_1, \ldots, z_N\right]^T, \tag{20}$$

where $r_{N,\varepsilon}(t)$ is the optimal solution to Problem 1, given the data set (t_i, z_i) and the smoothing parameter ε. Recall that the vectors $\hat{t}, \hat{r} \in \mathbb{R}^M$ constitute the discretization of the spline $r(t)$. Let $\hat{r}_{t_i}$ denote the element of $\hat{r}$ corresponding to $r(t_i)$. The matrix F defined by (6) "picks out" those elements of $\hat{r}$:

$$\begin{bmatrix} \hat{r}_{N,\varepsilon,t_1} \\ \vdots \\ \hat{r}_{N,\varepsilon,t_N} \end{bmatrix} = F^T \begin{bmatrix} \hat{r}_{N,\varepsilon,\hat{t}_1} \\ \vdots \\ \hat{r}_{N,\varepsilon,\hat{t}_M} \end{bmatrix} = F^T \hat{r}. \tag{21}$$

Equation (9) yields

$$\hat{r}_{N,\varepsilon} = -H^{-1}c = \frac{\varepsilon^2}{N}H^{-1}Fz = \frac{\varepsilon^2}{N}\left(\frac{1}{h^3}\Phi^T\Phi + \frac{\varepsilon^2}{N}FF^T\right)^{-1} Fz. \tag{22}$$

Therefore, we define

$$\hat{S}(\varepsilon) \triangleq \frac{\varepsilon^2}{N} F^T \left(\frac{1}{h^3} \Phi^T \Phi + \frac{\varepsilon^2}{N} F F^T \right)^{-1} F \tag{23}$$

$$\hat{V}_{\mathrm{GCV}}(\varepsilon) \triangleq \frac{\frac{1}{N}\|(I - \hat{S}(\varepsilon))z\|^2}{\left(\frac{1}{N}\operatorname{trace}(I - \hat{S}(\varepsilon))\right)^2}. \tag{24}$$

[18] states that $S(\lambda)$ is symmetric and positive semidefinite. It is straightforward to show that the estimate (23) retains these properties.

5 Simulations

In this section, simulation results are provided to demonstrate properties of the proposed GCV method. First, an example is shown to illustrate the usefulness and performance of the GCV method. Then we investigate whether the asymptotic result (19) holds for the estimate $\varepsilon_{\mathrm{GCV}}$. Throughout this section, the following notation is used:

$$\begin{aligned} \varepsilon_{\mathrm{GCV}} &= \arg\min_{\varepsilon} \quad \hat{V}_{\mathrm{GCV}}(\varepsilon) \\ \varepsilon_{\min} &= \arg\min_{\varepsilon} \quad E(R(\varepsilon)), \quad \text{with } R(\varepsilon) \triangleq \tfrac{1}{N}\textstyle\sum_{i=1}^{N} (r_{N,\varepsilon}(t_i) - r_{true}(t_i))^2 \\ r_{N,\varepsilon} &= \text{the spline computed using } (N,\varepsilon) \\ r_{true} &= \text{the underlying contour.} \end{aligned} \tag{25}$$

In simulation, the feasible region for ε was restricted to the interval $\Delta\varepsilon = [1, 1000]$. Deviations of $\varepsilon_{\mathrm{GCV}}$ were computed as $\|\varepsilon_{\mathrm{GCV}} - \varepsilon_{\min}\|/\Delta\varepsilon$.

Importance of Choice of Smoothing

Here, an example is provided to demonstrate advantages of optimal smoothing. With an added noise $\sigma = 0.1 mean(r_{true})$ and using $N = 100$, $M = 800$, splines are generated with $\varepsilon = \varepsilon_{\min}$, $\varepsilon = \varepsilon_{\mathrm{GCV}}$, $\varepsilon = 10\varepsilon_{\min}$ and $\varepsilon = 0.1\varepsilon_{\min}$ to compare results for different values of ε. Splines are shown in Figure 1, while the resulting cost functions $R(\varepsilon)$ and $\hat{V}_{\mathrm{GCV}}(\varepsilon)$ are shown in Figure 2. The advantages of optimal smoothing are clear from the figures. Due to space limitations, only one example contour is included. More examples are available at `http://www.math.kth.se/~karasalo/GCV.pdf`. In total, simulations were run for 25 test cases, with σ, N and M as above and the mean deviation of $\varepsilon_{\mathrm{GCV}}$ was less than 7%.

Asymptotic Properties

In this section we investigate whether the asymptotic result (19) holds for the estimate $\varepsilon_{\mathrm{GCV}}$, i.e. if

$$\lim_{N\to\infty} \frac{E(R(\varepsilon_{\mathrm{GCV}}))}{\min_{\varepsilon} E(R(\varepsilon))} = 1. \tag{26}$$

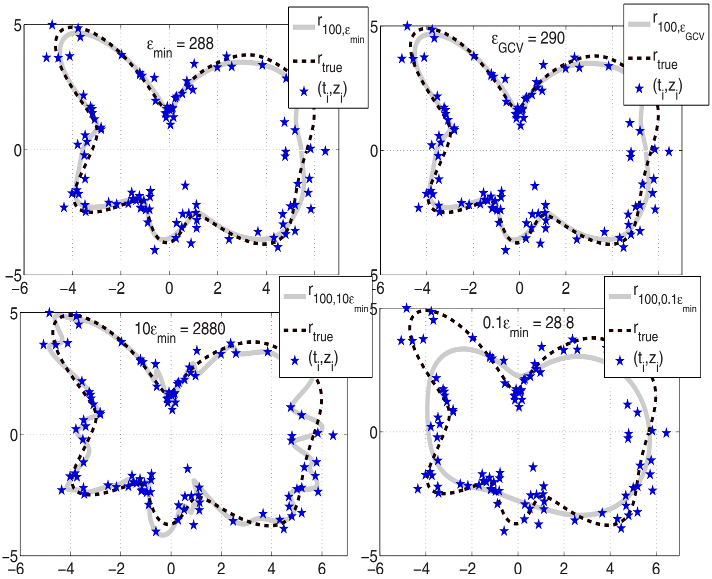


Fig. 1 Importance of Choice of Smoothing: Splines generated with different values of ε. Top left: $\varepsilon = \varepsilon_{\min}$. Top right: $\varepsilon = \varepsilon_{\text{GCV}}$. Bottom left: $\varepsilon = 10\varepsilon_{\min}$. Bottom right: $\varepsilon = 0.1\varepsilon_{\min}$.

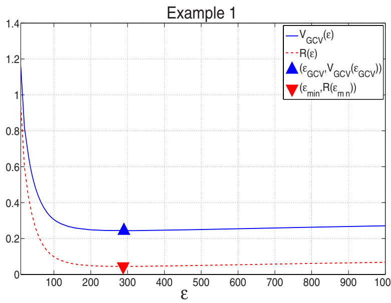


Fig. 2 Importance of Choice of Smoothing: Cost functions.

This asymptotic optimality may be regarded as the most important property of the smoothing parameter. We have performed simulations for $M = 1000$ and $N = \{1, 2, \ldots, 1000\}$ for 25 arbitrary contours and a noise level of $\sigma = 0.1mean(r_{true})$. Results are provided in Figure 3. We show mean values of ε_{GCV}, $\varepsilon_{\min}$, $R(\varepsilon_{\text{GCV}})$ and $R(\varepsilon_{\min})$ for the 25 contours. ε_{GCV} was generally a fair estimate of $\varepsilon_{\min}$, with a mean deviation of about 11%. $R(\varepsilon_{\text{GCV}})$ stays close to $R(\varepsilon_{\min})$ and shows a clear decrease as N increases. Finally, resulting error quotients are shown for increasing N. $R(\varepsilon_{\text{GCV}})/\min_{\varepsilon} R(\varepsilon)$ is decreasing toward 1, as expected.

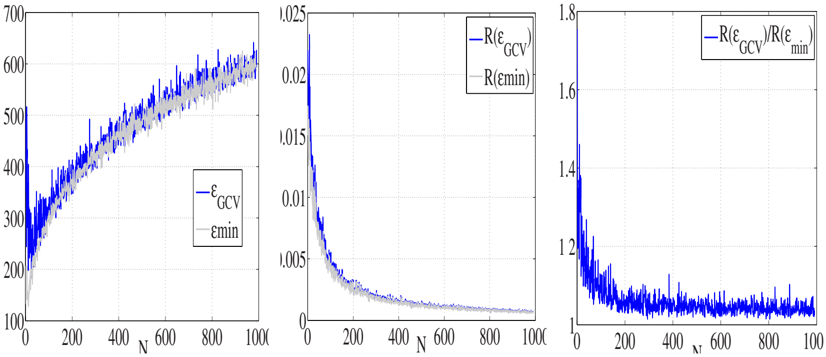


Fig. 3 **Asymptotic Properties:** Mean values of over 10 test contours as $N \to M$. Left: Mean values of ε. Middle: Mean values $R(\varepsilon_{\rm GCV})$ and $R(\varepsilon_{\rm min})$. Right: $R(\varepsilon_{\rm GCV})/\min_\varepsilon R(\varepsilon)$. Mean values over 10 test cases for $N \to M$.

6 Conclusions

In this paper, a general cross validation function was derived based on a discretization of a periodic control theoretic smoothing spline problem. An estimate of the optimal smoothing parameter ε was found by minimizing a GCV cost function $\hat{V}_{\rm GCV}(\varepsilon)$, without a priori information about the underlying closed curve or the quality of data. Theoretical and simulation results regarding properties of $\hat{V}_{\rm GCV}(\varepsilon)$ and the corresponding influence matrix $\hat{S}(\varepsilon)$ were provided and an example was shown to illustrate the usefulness and performance of the method.

References

1. Martin, C.F., Smith, J.: Approximation, interpolation and sampling. Differential Geometry: The Interface between Pure and Applied Mathematics, Contemp. Math. 68, 227–252 (1987)
2. Martin, C.F., Sun, S., Egerstedt, M.: Control theoretic smoothing splines. IEEE Transactions on Automatic Control 45, 2271–2279 (2000)
3. Egerstedt, M., Martin, C.F., Sun, S.: Optimal control, statistics and path planning. Math. Comput. Modelling 33, 237–253 (2001)
4. Egerstedt, M., Martin, C.F.: Statistical estimates for generalized splines. ESAIM: Control, Optimization and Calculus of Variations 9, 553–562 (2003)
5. Zhou, Y., Dayawansa, W., Martin, C.F.: Control theoretic smoothing splines are approximate linear filters. Comm. in Information and Systems 4, 253–272 (2004)
6. Kano, H., Fujioka, H., Egerstedt, M., Martin, C.F.: Optimal smoothing spline curves and contour synthesis. In: Proc. of the 16th IFAC World Congress (2005)
7. Kano, H., Egerstedt, M., Fujioka, H., Takahashi, S., Martin, C.F.: Periodic Smoothing Splines. Automatica 44, 185–192 (2008)
8. Egerstedt, M., Martin, C.F.: Control Theoretic Splines: Optimal Control, Statistics and Path Planning. Princeton Series in Applied Mathematics (to appear, 2010)

9. Karasalo, M., Hu, X., Martin, C.F.: Localization and mapping using recursive smoothing splines. In: Proc. of the 2007 European Control Conference, ECC (2007)
10. Karasalo, M., Hu, X., Martin, C.F.: Contour reconstruction and matching using recursive smoothing splines. In: Modeling, Estimation and Control, pp. 193–206. Springer, Heidelberg (2007)
11. Karasalo, M., Piccolo, G., Kragic, D., Hu, X.: Contour reconstruction using recursive smoothing splines — algorithms and experimental validation. Robotics and Autonomous Systems 57, 617–628 (2009)
12. Craven, P., Wahba, G.: Smoothing noisy data with spline functions. Numer. Math. 31, 377–403 (1979)
13. Wahba, G.: Spline Models for Observational Data. CBMS-NSF Series. SIAM, Philadelphia (1990)
14. Eubank, R.L.: Nonparametric Regression and Spline Smoothing, 2nd edn. CRC Press, Boca Raton (1999)
15. Reinsch, C.H.: Smoothing by spline functions. Numer. Math. 10, 177–183 (1967)
16. Hutchinson, M.F., de Hoog, F.R.: Smoothing noisy data with spline functions. Numer. Math. 47, 99–106 (1985)
17. Hutchinson, M.F., de Hoog, F.R.: An efficient method for calculating smoothing splines using orthogonal transformations. Numer. Math. 50, 311–319 (1987)
18. Hutchinson, M.F.: A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. Communications in Statistics — Simulation and Computation 19(2), 433–450 (1990)
19. Eldén, L.: A note on the computation of the generalized cross-validation function for ill-conditioned least squares problems. BIT Numerical Mathematics 24(4) (1984)
20. Shahraray, B., Anderson, D.J.: Optimal estimation of contour properties by cross-validated regularization. IEEE Transactions on Pattern Analysis and Machine Intelligence 11(6) (1989)

Stable $\mathscr{H}^\infty$ Controller Design for Systems with Time Delays

Hitay Özbay

Abstract. One of the difficult problems of robust control theory is to find strongly stabilizing controllers (i.e. stable controllers leading to stable feedback system) which satisfy a certain $\mathscr{H}^\infty$ performance objective. In this work we discuss stable $\mathscr{H}^\infty$ controller design methods for various classes of systems with time delays. We consider sensitivity minimization problem in this setting for SISO plants. We also discuss a suboptimal design method for stable $\mathscr{H}^\infty$ controllers for MIMO plants.

This paper is dedicated to Yutaka Yamamoto on the occasion of his 60th birthday.

1 Introduction

In this paper we will give an overview of recent results on design for various types of systems with time delays. The problem of finding a *stable* stabilizing controllers has been studied since 1970s, see [4, 8, 12, 18, 19] for finite dimensional systems and [1, 5, 6, 10, 16] for systems. This list is by no means complete; the reader can find various approaches and results from the references of the papers listed here.

In particular, [6] considers a class of SISO time delay systems with possibly infinitely many poles in $\mathbb{C}_+$. Under the condition that the number of zeros in $\mathbb{C}_+$ is finite, stable stabilizing controllers achieving a desired sensitivity level can be found using Nevanlinna-Pick interpolation.

Another approach for finding stable $\mathscr{H}^\infty$ controllers is to use the parameterization of all controllers achieving a desired $\mathscr{H}^\infty$ performance level, then look for a feasible free parameter which stabilizes the controller. In the context of time delay systems, this method has been studied in [5] where the suboptimal controller structure of [3, 17] is used.

By extending a result of [21], it is possible to obtain a large subset of all stable stabilizing controllers for a class of systems with time delays, [10]. Then, in this subset, we can search for controllers satisfying a desired $\mathscr{H}^\infty$ performance level.

Hitay Özbay
Dept. of Electrical and Electronics Eng., Bilkent University, Ankara, TR-06800, Turkey,
Currently on sabbatical leave at INRIA, Paris-Rocquencourt, France
e-mail: `hitay@bilkent.edu.tr`

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 105–113.
springerlink.com

Definitions of various stable controller design problems are given in Section 2. In Section 3 we discuss the Nevanlinna-Pick interpolation approach from [6] for stable $\mathscr{H}^\infty$ controller design for SISO time delay systems. The result of [10] is illustrated with an example in Section 4. Concluding remarks are made in Section 5.

2 Problem Definition and Preliminary Remarks

Consider the feedback system shown in Figure 1, where C is the controller and P is the plant. We say that the system is stable if $S := (1+PC)^{-1}$, PS and CS are in $\mathscr{H}^\infty$; in this case we say that C stabilizes P and write $C \in \mathscr{C}(P)$, where $\mathscr{C}(P)$ represents the set of all controllers stabilizing P. All stable stabilizing controller are denoted by $\mathscr{C}_\infty(P) := \mathscr{C}(P) \cap \mathscr{H}^\infty$.

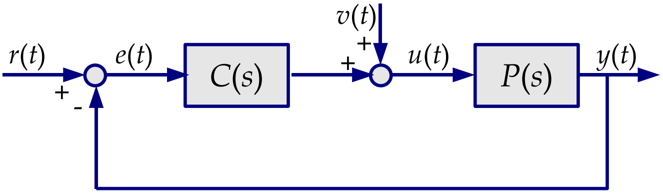


Fig. 1 Feedback System

We can define the following problems.

SS0 Given P find a controller C in $\mathscr{C}_\infty(P)$.
SS1 Given P, W_1 and $\rho > 0$, find a controller $C \in \mathscr{C}_\infty(P)$ such that $\|W_1 S\|_\infty \leq \rho$.
SS2 Given P, W_1, W_2 and $\rho > 0$, find a controller $C \in \mathscr{C}_\infty(P)$ such that

$$\left\| \begin{bmatrix} W_1 S \\ W_2(1-S) \end{bmatrix} \right\|_\infty \leq \rho.$$

SS0PD Given P find (if possible) a controller $C \in \mathscr{C}(P)$ such that

$$C(s) = K_p + K_d \frac{s}{\tau_d \; s+1}$$

for some $K_p, K_d \in \mathbb{R}$ and $\tau_d > 0$.

In this paper we will discuss SS0 and SS1 for various classes of time delay systems. The problem SS2 is a difficult one; it can be solved by trying to find a feasible free parameter in the parameterization of all suboptimal controllers, see [5]. Due to page limitations, we will also leave SS0PD aside, but it can be solved by finding a characterization of the set of all stabilizing (K_p, K_d) pairs for each fixed $\tau_d > 0$, see e.g. [13] and its references. An alternative approach for SS0PD would be to use the results of [7, 11], where a simple but conservative design method is proposed for proportional plus derivative (PD) controller synthesis for systems with time delays.

For finite dimensional systems, it is well known that the problem SS0 is solvable if and only if P satisfies the PIP (the number of poles between every pair of blocking zeros on the extended real axis is even), [19]. This result remains valid for a large class of time delay systems, see e.g. [1].

Let us consider a plant in the form

$$P(s) = N(s)/D(s) \tag{1}$$

where $N, D \in \mathscr{H}^\infty$ are strongly coprime, [14]. Assume that N has finitely many zeros, $z_1, \dots, z_\ell$ (assume they are distinct for simplicity) in the extended right half plane, $\mathbb{R}_{+e} = \mathbb{R}_+ \cup \{\infty\}$. A controller $C \in \mathscr{H}^\infty$ is in $\mathscr{C}(P)$ if and only if $U, U^{-1} \in \mathscr{H}^\infty$, where $U = D + NC$. Note that when $C \in \mathscr{H}^\infty$ we have $U(z_i) = D(z_i)$. The problem of finding a feasible U is solvable if and only if the set $\{D(z_1), \dots, D(z_\ell)\}$ is sign invariant, which is equivalent to PIP.

3 Nevanlinna-Pick Interpolation for Stable $\mathscr{H}^\infty$ Controller Design

Consider the plant (1) defined in the previous section with ensuing assumptions. Besides zeros on the positive real axis, plant may have other zeros in $\mathbb{C}_+$, let us enumerate them as $z_{\ell+1}, \dots, z_n$, and assume that they are distinct. Let $D(z_i) > 0$ for all $i = 1, \dots, \ell$ (i.e., PIP is satisfied). In order to find a controller $C \in \mathscr{C}_\infty(P)$ we can construct a unimodular U (i.e. $U, U^{-1} \in \mathscr{H}^\infty$) such that

$$U \; : \; \mathbb{C}_+ \to \mathbb{W}_\gamma \quad \text{with} \quad U(z_i) = D(z_i) \quad i = 1, \dots, n \tag{2}$$

where the range $\mathbb{W}_\gamma$ is defined as

$$\mathbb{W}_\gamma := \{re^{j\theta} \in \mathbb{C} \; : \quad \varepsilon < r < \gamma \, , \quad -\pi < \theta < \pi\} \tag{3}$$

for some sufficiently small number $\varepsilon > 0$ and a finite number $\gamma > \varepsilon$. Note that $U(s)$ should not take negative values for $s \in \mathbb{R}_{+e}$ (otherwise U^{-1} does not exists because in that case $U(s)$ takes both positive and negative values for $s \in \mathbb{R}_+$ meaning that it has a zero in $\mathbb{R}_+$), so negative real axis is excluded from $\mathbb{W}_\gamma$. Clearly γ should be large enough so that $D(z_i) \in \mathbb{W}_\gamma$ for all $i = 1, \dots, n$. Also note that with the above definition we guarantee the upper bounds $\|U\|_\infty < \gamma$ and $\|U^{-1}\|_\infty < \varepsilon^{-1}$. Once a feasible U is found, the controller is given by

$$C(s) = \frac{U(s) - D(s)}{N(s)}$$

which is stable by interpolation conditions, and we have $S = DU^{-1}$ and $PS = NU^{-1}$.

For technical reasons, assume for the moment that the plant does not have a zero at $+\infty$, i.e. all z_i's are finite. Since $\mathbb{W}_\gamma$ is a simply connected domain there is a conformal map

$$\phi_\gamma \; : \; \mathbb{W}_\gamma \rightarrow \mathbb{D}.$$

Let φ be a conformal map from $\mathbb{C}_+$ to $\mathbb{D}$. Define

$$\alpha_i = \varphi(z_i) \in \mathbb{D}, \qquad \beta_i = \phi_\gamma(U(z_i)) \in \mathbb{D}, \quad i = 1, \ldots, n.$$

Then, finding a bounded analytic U satisfying (2) is equivalent to finding a bounded analytic function

$$\vartheta \; : \; \mathbb{D} \rightarrow \mathbb{D} \qquad \text{such that} \quad \vartheta(\alpha_i) = \beta_i, \quad i = 1, \ldots, n.$$

This is the Nevanlinna-Pick problem and it is solvable if and only if a Pick matrix is positive definite, [3, 20]. The associated Pick matrix is constructed from α_i's and β_i's, which depend on the original problem data z_i's, $D(z_i)$'s and γ. If this problem is feasible, then U can be found from ϑ as

$$U(s) = \phi_\gamma^{-1}(\vartheta(\varphi(s))).$$

Thus SS0 can be solved from the above procedure. Note that when the plant has a zero at $+\infty$, then under the φ this point is mapped to a point on the unit circle. So, we need to construct ϑ from $\overline{\mathbb{D}}$ to $\mathbb{D}$. This case requires a slight extension of the classical Nevanlinna-Pick interpolation; for a solution see Section 2.11.3 of [3].

Although γ puts a bound on $\|U^{-1}\|_\infty$, in order to find a controller for SS1 we need to have a bound for $\|W_1 S\|_\infty = \|W_1 D U^{-1}\|_\infty$. For this purpose, let us first consider an inner-outer factorization of $D = D_i D_o$ and assume D_o is invertible in $\mathscr{H}^\infty$. If the plant does not have a pole on the Im-axis then this assumption holds, and D_o^{-1} can be seen as part of N. So, we can take $D = D_i$ and under this assumption $\|W_1 S\|_\infty = \|W_1 U^{-1}\|_\infty$. Let $W_1^{-1} \in \mathscr{H}^\infty$ and define

$$F(s) := \frac{1}{\rho} W_1(s) U^{-1}(s).$$

Under the above assumptions, the problem SS1 is solvable if and only if there exists an F such that $F, F^{-1} \in \mathscr{H}^\infty$ with

$$F \; : \; \mathbb{C}_+ \rightarrow \mathbb{W}_1 \quad \text{and} \;\; F(z_i) = \frac{W_1(z_i)}{\rho \, D(z_i)} \quad i = 1, \ldots, n.$$

By using the conformal maps as defined above, this problem can be transformed to a Nevanlinna-Pick problem. Once a feasible F is found a controller solving SS1 is given by

$$C = \frac{\rho^{-1} W_1 F^{-1} - D}{N},$$

which is stable by interpolation conditions and it leads to $S = \rho D W_1^{-1} F$ satisfying the $\mathscr{H}^\infty$ performance condition:

$$\|W_1 S\|_\infty = \|\rho F\|_\infty \leq \rho.$$

In [6] the function F is considered to be in the form $F(s) = e^{-G(s)}$. Since $F^{-1}(s) = e^{G(s)}$ and $\|F^{-1}\|_\infty < \varepsilon^{-1}$, we are looking for a bounded analytic G such that associated interpolation conditions hold and

$$G : \mathbb{C}_+ \to \mathbb{C}_+^{\sigma_o} := \{s \in \mathbb{C}_+ : 0 < \mathrm{Re}(s) < \sigma_o = \ln(\varepsilon^{-1})\},$$

where $\varepsilon > 0$ is as in (3). Again, by a series of conformal maps construction of a feasible G can be reduced to a Nevanlinna-Pick problem, see [6] for details.

Now we want to give an example from [6] for the class of plants which can be handled in the above framework. Consider

$$P(s) = \frac{(s+1)+4e^{-3s}}{(s+1)+2(s-1)e^{-2s}} = \frac{1e^{-0s} + \left(\frac{4}{s+1}\right)e^{-3s}}{1e^{-0s} + 2\left(\frac{s-1}{s+1}\right)e^{-2s}} =: \frac{R(s)}{T(s)}$$

where $R(s)$ has four zeros in $\mathbb{C}_+$: $z_{1,2} \approx 0.31 \pm j0.85$ and $z_{3,4} \approx 0.1 \pm j2.7$, so define

$$N_i(s) = \prod_{i=1}^{4} \frac{s - z_i}{s + z_i}.$$

Note that relative degree of the plant is zero hence $+\infty$ is not a zero of P, so we do not have to deal with interpolation conditions at the boundary. Also, the plant has infinitely many poles in $\mathbb{C}_+$; in this situation we define

$$\bar{T}(s) := e^{-2s}T(-s)\left(\frac{s-1}{s+1}\right) = 2 + \left(\frac{s-1}{s+1}\right)e^{-2s}$$

and check that $\bar{T}(s)$ is stable and it does not have zeros in $\mathbb{C}_+$. Thus the plant admits the following coprime factorization

$$P(s) = \frac{N_i(s)N_o(s)}{D_i(s)} \quad \text{with} \quad D_i(s) = \frac{T(s)}{\bar{T}(s)}, \quad N_o(s) = \frac{R(s)}{N_i(s)}\,\frac{1}{\bar{T}(s)}.$$

If we choose $\sigma_o = \ln(\varepsilon^{-1}) = 3$, i.e. $\varepsilon = e^{-3} \approx 0.05$, and $W_1(s) = (1+0.1s)/(s+1)$, then we can find a solution for SS1 with $\rho = 1.0815$, and the resulting F is given as

$$F(s) = \exp\left(-\frac{\sigma_o}{2} - j\frac{\sigma_o}{\pi}\ln\left(\frac{1+\widetilde{G}(s)}{1-\widetilde{G}(s)}\right)\right) \quad \text{where}$$
$$\widetilde{G}(s) \approx j\frac{-0.99(s-3.473)(s+1)(s^2-0.03s+7.56)}{(s+3.415)(s+1.007)(s^2+0.034s+7.57)}.$$

As $\varepsilon \to 0$ we see that the smallest ρ for which SS1 is solvable decreases to 1.0726.

At this point we should mention that the zeros $z_{3,4}$ have not been taken into account in [6], so the numerical example given there is not correct (it is correct only for a plant with two zeros $z_{1,2}$ in $\mathbb{C}_+$ with same interpolation conditions). It is interesting that $z_{1,2}$ are the dominant zeros in the sense that when interpolation conditions due to $z_{3,4}$ are ignored the smallest ρ for which SS1 is solvable can be computed to be 1.0704 as $\varepsilon \to 0$.

4 Suboptimal Stable $\mathscr{H}^\infty$ Controllers

In this section we first consider SS0 for MIMO plants in the form $P = D^{-1}N$, where all entries of $N(s)$ and $D(s)$ are in $\mathscr{H}^\infty$. A controller C is in $\mathscr{C}_\infty(P)$ if all entries of C are in $\mathscr{H}^\infty$, and $U = D + NC$ is unimodular, i.e. U and U^{-1} have all its entries in $\mathscr{H}^\infty$. In this setting N, D, C, U are appropriate size matrices whose entries are in $\mathscr{H}^\infty$. For notational convenience, without specifying the matrix size we write $D, N, C, U \in \mathscr{H}^\infty$.

The system given below illustrates one possible class of plants which can be studied in this framework:

$$P(s) = \frac{(s-4)e^{-3hs}}{(s+1-2e^{-0.4s})} \begin{bmatrix} \frac{1}{s+2} & \frac{-1}{s+4} & \frac{1}{s+3} \\ 0 & 0 & \frac{e^{-hs}}{s+1+e^{-s}} \end{bmatrix}, \quad h > 0 \tag{4}$$

which can be factored as $P(s) = D(s)^{-1}N_i(s)N_o(s)N_1(s)$ where N_i is inner, N_o is finite dimensional outer and N_1 is right invertible infinite dimensional outer matrix:

$$N_i(s) = \frac{s-4}{s+4}\, e^{-3hs} \begin{bmatrix} 1 & 0 \\ 0 & e^{-hs} \end{bmatrix}, \qquad N_o(s) = \frac{1}{s+1} I,$$

$$N_1(s) = \frac{s-p}{s+1-2e^{-0.4s}} \begin{bmatrix} \frac{s+4}{s+2} & -1 & \frac{s+4}{s+3} \\ 0 & 0 & \frac{s+4}{s+1+e^{-s}} \end{bmatrix}$$

and $D(s) = \dfrac{s-p}{s+1} I$ with $p > 0$ being the only root of $s + 1 - 2e^{-0.4s} = 0$ in $\mathbb{C}_+$ (note that $p \approx 0.5838$). For this plant, a controller $C \in \mathscr{H}^\infty$ is in $\mathscr{C}_\infty(P)$ if and only if

$$U = D + N_i N_o N_1 C$$

is unimodular. Note that N_1 admits a right inverse

$$N_1^\dagger(s) = \frac{s+1-2e^{-0.4s}}{s-p} \begin{bmatrix} 2\frac{s+2}{s+4} & 0 \\ 1 & \frac{s+1+e^{-s}}{s+3} \\ 0 & \frac{s+1+e^{-s}}{s+4} \end{bmatrix} \in \mathscr{H}^\infty.$$

If we define $C = N_1^\dagger C_1$ where $C_1 \in \mathscr{H}^\infty$ is free, then this controller is in $\mathscr{C}_\infty(P)$ if $U = D + N_i N_o C_1$ is unimodular.

Let $R := (D - I)$, then $C \in \mathscr{C}_\infty(P)$ if $C_1 \in \mathscr{H}^\infty$ satisfies

$$\|R + N_i N_o C_1\|_\infty < 1. \tag{5}$$

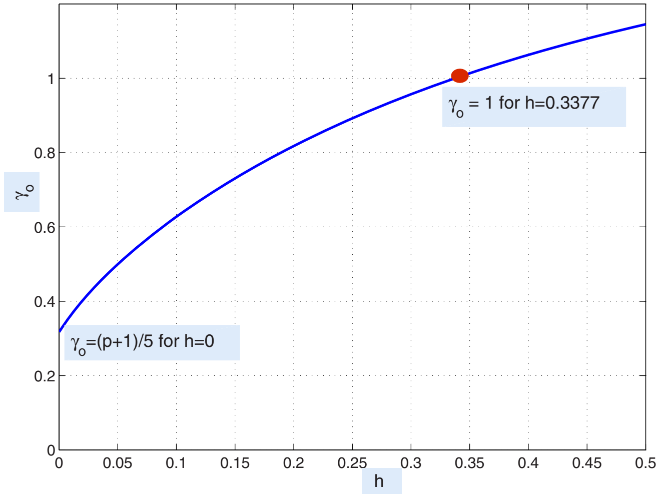


Fig. 2 γ_o versus h

The problem of finding a suitable C_1 is an $\mathscr{H}^\infty$ control problem and can be solved using one of many alternative techniques from the literature, see e.g. [9]. For the numerical example given above, the problem (5) has a solution if and only if

$$\gamma_o := \inf_{Q\in\mathscr{H}^\infty} \left\| \frac{p+1}{s+1} - \frac{(s-4)}{(s+4)(s+1)}\, e^{-4hs} Q \right\|_\infty < 1. \tag{6}$$

Using the results of [3, 9] we can compute $\gamma_o < (p+1)$ from the smallest root ω_o of

$$\tan^{-1}\omega_o + 2\tan^{-1}\frac{\omega_o}{4} + 4h\omega_o = \pi, \qquad \text{where} \quad \omega_o = \sqrt{\frac{(p+1)^2}{\gamma_o^2} - 1}.$$

Figure 2 shows γ_o as a function of h. It implies that for the given plant we can find a controller $C \in \mathscr{C}_\infty(P)$ using this method if and only if $h < 0.3377$.

Let us now study SS1 for the SISO version of the plants considered in this section, $P = N/D$. A controller $C = Q \in \mathscr{H}^\infty$ solves SS1 if $U = D + NQ$ is unimodular and $\|\rho^{-1}W_1 D U^{-1}\|_\infty \leq 1$, equivalently

$$|\rho^{-1}W_1(j\omega)D(j\omega)| \leq |D(j\omega) + N(j\omega)Q(j\omega)|, \qquad \omega \in \mathbb{R}.$$

Using $R := D - 1$ we see that a sufficient condition for the above is

$$|\rho^{-1}W_1(j\omega)D(j\omega)|^2 + |R(j\omega) + N(j\omega)Q(j\omega)|^2 \leq 1/2 \qquad \omega \in \mathbb{R}.$$

Assume that $\rho > \sqrt{2}\|W_1 D\|_\infty$, then we can find $V_\rho \in \mathscr{H}^\infty$ such that $V_\rho^{-1} \in \mathscr{H}^\infty$ and

$$|V_\rho(j\omega)|^2 = \frac{1}{2} - |\rho^{-1} W_1(j\omega) D(j\omega)|^2 \qquad \omega \in \mathbb{R}.$$

With this spectral factorization, SS1 is solvable if

$$\gamma_1 := \inf_{Q_1 \in \mathscr{H}^\infty} \|V_\rho^{-1} R + N Q_1\|_\infty < 1. \tag{7}$$

If (7) holds, then $C = V_\rho Q_1$ is an admissible solution of SS1 for all $Q_1 \in \mathscr{H}^\infty$ satisfying $\|V_\rho^{-1} R + N Q_1\|_\infty < 1$.

Let us now consider this problem for the plant $P = N/D$

$$D(s) = \frac{s-p}{s+1}, \qquad N(s) = \frac{s-4}{(s+4)(s+1)} e^{-4hs},$$

with $p = 0.5838$ and $h > 0$. Take $\rho = 2$ and $W_1(s) = \frac{s+1}{10s+1}$, and check that $\rho > \sqrt{2}\|W_1 D\|_\infty = \sqrt{2}p$. Below table shows the values of γ_1 for varying h. We see that the largest h for which we can find a solution to SS1 using this method is 0.1354.

h	0	0.01	0.05	0.10	0.13	0.1354	0.14	0.15	0.2
γ_1	0.45	0.52	0.71	0.89	0.98	0.9991	1.013	1.041	1.165

It is interesting to compare the results of this table with Figure 2. For each fixed h we have $\gamma_1 > \gamma_o$. This is expected since SS1 is more stringent than SS0. In fact, due to added conservatism in our approach to SS1, for each fixed h we have that $\gamma_1 \to \sqrt{2}\gamma_o$ as $\rho \to \infty$.

5 Conclusions

Stable $\mathscr{H}^\infty$ controller design problems are discussed and two alternative methods are illustrated for two different classes of plants with time delays. Here we considered the sensitivity minimization problem only. Generalization of the proposed methods to mixed sensitivity minimization is a non-trivial problem which remains unsolved.

References

1. Abedor, J.L., Poolla, K.: On the strong stabilization of delay systems. In: Proc. of the 28th IEEE Conf. on Decision and Control, Tampa FL, December 1989, pp. 2317–2318 (1989)
2. Cheng, P., Cao, Y.-Y., Sun, Y.: On Strong $\gamma_k - \gamma_{cl}$ H_∞ stabilization and simultaneous $\gamma_k - \gamma_{cl}$ H_∞ control. In: Proc. of the 46th IEEE Conf. on Decision and Control, New Orleans, LA, December 2007, pp. 5417–5422 (2007)
3. Foias, C., Özbay, H., Tannenbaum, A.: Robust Control of Infinite Dimensional Systems: Frequency Domain Methods. Lecture Notes in Control and Information Sciences, vol. 209. Springer, London (1996)

4. Gumussoy, G., Özbay, H.: Remarks on strong stabilization and stable H^∞ controller design. IEEE Transactions on Automatic Control 50, 2083–2087 (2005)
5. Gumussoy, G., Özbay, H.: Stable H^∞ controller design for time delay systems. International Journal of Control 81, 546–556 (2008)
6. Gumussoy, S., Özbay, H.: Sensitivity minimization by strongly stabilizing controllers for a class of unstable time-delay systems. IEEE Transactions on Automatic Control 54, 590–595 (2009)
7. Gündeş, A.N., Özbay, H., Özgüler, A.B.: PID controller synthesis for a class of unstable MIMO plants with I/O delays. Automatica 43, 135–142 (2007)
8. Halevi, Y.: Stable LQG controllers. IEEE Transactions on Automatic Control 39, 2104–2106 (1994)
9. Kashima, K., Yamamoto, Y., Özbay, H.: Parameterization of suboptimal solutions of the Nehari problem for infinite-dimensional systems. IEEE Transactions on Automatic Control 52(12), 2369–2374 (2007)
10. Özbay, H.: On strongly stabilizing controller synthesis for time delay systems. In: Proc. of the 17th IFAC World Congress, Seoul, Korea, July 2008, pp. 6342–6346 (2008)
11. Özbay, H., Gündeş, A.N.: Resilient PI and PD controller designs for a class of unstable plants with I/O delays. Applied and Computational Mathematics 6, 18–26 (2007)
12. Özbay, H., Gündeş, A.N.: Strongly stabilizing controller synthesis for a class of MIMO plants. In: Proc. of the 17th IFAC World Congress, Seoul, Korea, July 2008, pp. 359–363 (2008)
13. Saadaoui, K., Elmadssia, S., Benrejeb, M.: Stabilizing first-order controllers for n-th order all pole plants with time delay. In: Proc. of the 16th Mediterranean Conf. on Control and Autom., Ajaccio, France, June 2008, pp. 812–817 (2008)
14. Smith, M.C.: On stabilization and existence of coprime factorizations. IEEE Transactions on Automatic Control 34, 1005–1007 (1989)
15. Smith, M.C., Sondergeld, K.P.: On the order of stable compensators. Automatica 22, 127–129 (1986)
16. Suyama, K.: Strong stabilization of systems with time-delays. In: Proc. IEEE Industrial Electronics Society Conference, pp. 1758–1763 (1991)
17. Toker, O., Özbay, H.: H^∞ optimal and suboptimal controllers for infinite dimensional SISO plants. IEEE Transactions on Automatic Control 40, 751–755 (1995)
18. Vidyasagar, M.: Control System Synthesis: A Factorization Approach. MIT Press, Cambridge (1985)
19. Youla, D.C., Bongiorno, J.J., Lu, C.N.: Single-loop feedback stabilization of linear multivariable dynamical plants. Automatica 10, 159–173 (1974)
20. Zeren, M., Özbay, H.: Comments on Solutions to the combined sensitivity and complementary sensitivity problem in control systems. IEEE Transactions on Automatic Control 43(5), 724 (1998)
21. Zeren, M., Özbay, H.: On the strong stabilization and stable $\mathscr{H}^\infty$-controller design problems for MIMO systems. Automatica 36, 1675–1684 (2000)

Dynamic Quantization for Control

Toshiharu Sugie, Shun-ichi Azuma, and Yuki Minami

Abstract. This paper overviews a series of the authors' recent contributions to dynamic quantizer design for control. The problem considered here is to find a dynamic quantizer such that the resulting quantized system is an optimal approximation of an ideal unquantized system.We show here a fundamental solution to this problem and briefly review several results toward real applications.

1 Introduction

As a bridge between the infinite and finite worlds, the quantization has been a key issue in science and engineering fields. For example, it can be seen in

• *signal processing*	analog-to-digital conversion,
• *information theory*	source coding,
• *statistics*	cluster analysis,
• *operations research*	facility location.

By the quantization, one can convert infinitely large number of noise-corrupted data to compact data.

Also in the systems and control field, the quantization has attracted much attention in last decade, due to the increasing need for hybrid control and networked control (e.g., see [1, 2, 3, 4]). From various points of view, many results have been obtained so far; for example, the quantizer coarseness for stabilization has been characterized in e.g., [5, 6, 7, 8, 9], and good quantizers (or switching controllers) for control have been developed in e.g., [1, 2, 10, 11].

Toshiharu Sugie and Shun-ichi Azuma
Graduate School of Informatics, Kyoto University; Uji, Kyoto 611-0011, Japan
e-mail: {sugie, sazuma}@i.kyoto-u.ac.jp

Yuki Minami
Department of Control Engineering, Maizuru National College of Technology;
Maizuru, Kyoto 625-8511, Japan
e-mail: minami@maizuru-ct.ac.jp

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 115–124.
springerlink.com

On the other hand, the authors have been interested in "dynamic quantizers", which map continuous-valued signals into discrete-valued ones depending on the past history of both signals. Compared with static quantizers, dynamic quantizers have much better performance, which has motivated us to pursue their potential for control. So far, for a class of dynamic quantizers, called the $\Delta\Sigma$ modulators in the signal processing community [12], we have obtained several key results which clarify the optimal quantization structure and the performance limitation in control systems. In this paper, we briefly review a series of the authors' results in [13, 14, 15, 16, 17].

The problem addressed here is as follows: when a plant and a controller are given for the quantized feedback system in Fig. 1 (a), find a dynamic quantizer such that the system in (a) optimally approximates the usual feedback system in Fig. 1 (b), in terms of the input-output relation. If the problem is solved with small approximation error, one can directly apply controllers designed for the usual system in (b) to the quantized system in (a), even if the plant input is restricted to belonging to a *fixed* discrete set. This gives a big advantage to construct quantized control systems subject to discrete-valued signal constraints.

In the following sections, we first derive an expression of the performance of dynamic quantizers. Based on this, an optimal dynamic quantizer in a closed form is presented. Finally, the authors' recent studies toward real applications are introduced.

Notation: Let $\mathbf{R}$, $\mathbf{R}_+$, and $\mathbf{N}$ be the real number field, the set of positive real numbers, and the set of natural numbers (positive integers), respectively. We denote by 0 the zero matrix of appropriate dimensions. For the matrix $M := \{M_{ij}\}$, let $\mathrm{abs}(M)$ denote the matrix composed of the absolute value of each element, i.e., $\mathrm{abs}(M) = \{|M_{ij}|\}$, and let M^+ be the pseudo-inverse. For the vector sequences $X := (x_1, x_2, \ldots)$ and $Y := (y_1, y_2, \ldots)$, we use $X - Y$ to express the vector sequence $(x_1 - y_1, x_2 - y_2, \ldots)$. For the vector x, the matrix M, and the vector sequence X, the symbols $\|x\|$, $\|M\|$, and $\|X\|$ express their ∞-norms (i.e., $\|X\| := \sup_{i\in\mathbf{N}} \|x_i\|$).

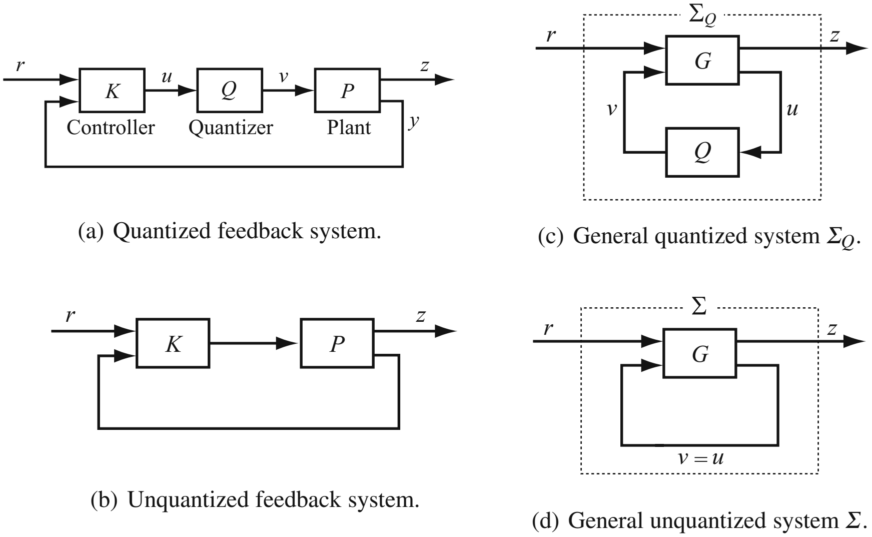
2 Dynamic Quantizer Design Problem

Consider the feedback system Σ_Q in Fig. 1 (c), which is a generalized version of the quantized feedback system in (a).

The system G is given by

$$G : \begin{cases} x(k+1) &= Ax(k) + B_1 r(k) + B_2 v(k), \\ \quad z(k) &= C_1 x(k) + D_1 r(k), \\ \quad u(k) &= C_2 x(k) + D_2 r(k) \end{cases} \tag{1}$$

where $x \in \mathbf{R}^n$ is the state, $r \in \mathbf{R}^p$ and $v \in \mathbf{R}^m$ are the inputs, $z \in \mathbf{R}^l$ and $u \in \mathbf{R}^m$ are the outputs, $k \in \{0\} \cup \mathbf{N}$ is the discrete time, and $A \in \mathbf{R}^{n\times n}$, $B_1 \in \mathbf{R}^{n\times p}$, $B_2 \in \mathbf{R}^{n\times m}$, $C_1 \in \mathbf{R}^{l\times n}$, $C_2 \in \mathbf{R}^{m\times n}$, $D_1 \in \mathbf{R}^{l\times p}$, $D_2 \in \mathbf{R}^{m\times p}$ are constant matrices. The initial state is given as $x(0) = x_0$ for $x_0 \in \mathbf{R}^n$.



(a) Quantized feedback system.

(b) Unquantized feedback system.

(c) General quantized system Σ_Q.

(d) General unquantized system Σ.

Fig. 1 Quantized and unquantized (usual) feedback systems.

On the other hand, Q is the dynamic quantizer in the following form:

$$Q : \begin{cases} \xi(k+1) = \mathcal{A}\xi(k) + \mathcal{B}_1 u(k) + \mathcal{B}_2 v(k), \\ \quad v(k) \quad = q(\mathcal{C}\xi(k) + u(k)) \end{cases} \tag{2}$$

where $\xi \in \mathbf{R}^{\mathcal{N}}$ is the state of dimension $\mathcal{N}$, $u \in \mathbf{R}^m$ is the input, $v \in \mathbf{V}^m := \{0, \pm d, \pm 2d, \ldots\}^m$ is the output, and $\mathcal{A} \in \mathbf{R}^{\mathcal{N}\times\mathcal{N}}$, $\mathcal{B}_1, \mathcal{B}_2 \in \mathbf{R}^{\mathcal{N}\times m}$, $\mathcal{C} \in \mathbf{R}^{m\times\mathcal{N}}$ are constant matrices. The function $q : \mathbf{R}^m \to \mathbf{V}^m$ is the nearest-neighbor static quantizer, where $\mathbf{V}$ is the discrete set on which each output takes its value and $d \in \mathbf{R}_+$ is the quantization level. The initial state is given as $\xi(0) = 0$ for guaranteeing that Q is drift-free, i.e., $v(k) = 0$ for $u(k) = 0$ $(k = 0, 1, \ldots)$. This quantizer determines its output depending upon its current input and past input sequence.

Next, we prepare some symbols. For the system Σ_Q, let $Z_Q(x_0, R)$ denote the controlled output sequence $(z(1), z(2), \ldots, z(\infty))$ for the initial state x_0 and the reference input $R := (r_0, r_1, \ldots) \in \ell_\infty^p$ (i.e., $x(0) = x_0$ and $r(k) = r_k$), and let $z_Q(k, x_0, R)$ be the output at time k. In addition, we consider the feedback system Σ in Fig. 1 (d), corresponding to a generalized version of (b), for which the symbols $Z(x_0, R)$ and $z(k, x_0, R)$ are similarly defined. Then as a performance index of Q, we define the *maximum output difference*:

$$E(Q) := \sup_{(x_0, R) \in \mathbf{R}^n \times \ell_\infty^p} \| Z_Q(x_0, R) - Z(x_0, R) \|. \tag{3}$$

This is the difference between the system Σ_Q in Fig. 1 (c) and the ideal system Σ in (d), in terms of the input-output relation. Then our problem is formulated as follows.

Problem 1. For the system Σ_Q, suppose that the quantization level $d \in \mathbf{R}_+$ is given and assume that Σ is stable (the matrix $A + B_2C_2$ is Schur).
(i) Determine the value of $E(Q)$ for given Q.
(ii) Find a Q (i.e., parameters $(\mathcal{N}, \mathcal{A}, \mathcal{B}_1, \mathcal{B}_2, \mathcal{C})$) minimizing $E(Q)$ under the stability condition:

$$\sup_{k \in \mathbf{N}} \|x(k)\| < \infty, \quad \sup_{k \in \mathbf{N}} \|\xi(k)\| < \infty \quad (\forall (x_0, R) \in \mathbf{R}^n \times \ell_\infty^p). \tag{4}$$

□

Problems (i) and (ii) respectively correspond to the analysis and design problems. If $E(Q)$ is small, the input-output relation of the ideal system Σ is almost preserved in Σ_Q. This provides us a practical method of control systems design with discrete-valued signal constraints. For example, consider the feedback system in Fig. 1 (a), and suppose that P has to be actuated by discrete-valued signal. Then the performance would be good with

- any controller K achieving desirable performance in the ideal system in Fig. 1 (b),
- any dynamic quantizer Q such that $E(Q)$ is small.

Note that Problem 1 is nonconvex. In fact, since Z_Q involves a discontinuous function q, the function $\|Z_Q(x_0, R) - Z(x_0, R)\|$ is nonconvex with respect to (x_0, R). Furthermore, the problem (ii) is a minimax problem between (x_0, R) and $(\mathcal{N}, \mathcal{A}, \mathcal{B}_1, \mathcal{B}_2, \mathcal{C})$.

3 Analytical Solutions

Under the following assumptions, we can obtain an analytical solution to Problem 1.

(A1) rank $D_2 = m$ (D_2 is full row rank).
(A2) For $\tilde{A} := A + B_2C_2$, there exists a $k \in \{0\} \cup \mathbf{N}$ such that $C_1\tilde{A}^0B_2 = C_1\tilde{A}^1B_2 = \cdots = C_1\tilde{A}^{k-1}B_2 = 0$ (if $k \geq 1$) and rank $C_1\tilde{A}^kB_2 = l$.
(A3) The system Σ' in Fig. 2 has no unstable system zero.

Note in (A2) that $C_1\tilde{A}^kB_2$ ($k = 0, 1, \ldots$) express the impulse response matrices of the system Σ' in Fig. 2 (from s to z).

Even if these assumptions do not hold, a practical solution to Problem 1 is obtained, which will be explained in Remark 1 and Section 4.

3.1 Performance Expression

A solution to Problem 1 (i) is given as follows.

Theorem 1 ([14, 15]). *For the system Σ_Q, assume* (A1) *(*(A2) *and* (A3) *are not necessary). If*

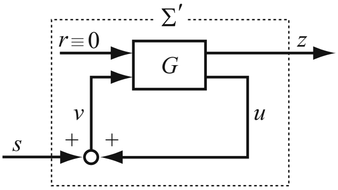


Fig. 2 A subsystem of quantized system Σ_Q.

$$[C_1 \;\; 0]\begin{bmatrix} \tilde{A} & B_2\mathcal{C} \\ 0 & \mathcal{A}+\mathcal{B}_2\mathcal{C} \end{bmatrix}^k \begin{bmatrix} 0 \\ \mathcal{B}_1+\mathcal{B}_2 \end{bmatrix} = 0 \quad (\forall k \in \{0\} \cup \mathbf{N}), \tag{5}$$

then

$$E(Q) = \left\| \sum_{k=0}^{\infty} \operatorname{abs}\left([C_1 \;\; 0]\begin{bmatrix} \tilde{A} & B_2\mathcal{C} \\ 0 & \mathcal{A}+\mathcal{B}_2\mathcal{C} \end{bmatrix}^k \begin{bmatrix} B_2 \\ \mathcal{B}_2 \end{bmatrix} \right) \right\| \frac{d}{2}; \tag{6}$$

otherwise

$$E(Q) = \infty. \tag{7}$$

□

Theorem 1 gives an exact expression of $E(Q)$, which enables us to compute the value of $E(Q)$ for given dynamic quantizer Q.

The intuitive meaning of this result is as follows. Let us introduce the new variable $w \in [-d/2, d/2]^m$:

$$w(k) := q(\mathcal{C}\xi(k)+u(k)) - (\mathcal{C}\xi(k)+u(k)), \tag{8}$$

which expresses the quantization error of the static quantizer q in (2). This allows us to represent Q as

$$Q : \begin{cases} \xi(k+1) = (\mathcal{A}+\mathcal{B}_2\mathcal{C})\xi(k) + (\mathcal{B}_1+\mathcal{B}_2)u(k) + \mathcal{B}_2 w(k), \\ \quad v(k) \quad = \mathcal{C}\xi(k) + u(k) + w(k) \end{cases} \tag{9}$$

and to formally regard Q as a linear system with the external inputs u and w. With this expression, the error system for Σ_Q and Σ is illustrated as Fig. 3, where H is a subsystem (which is linear) of (9). Then (5) means that the impulse response matrices from r to $z_Q - z$ are zero. Thus if (5) does not hold, $\|z_Q - z\|$ can be arbitrarily large by some large r, which gives (7). On the other hand, the right hand side of (6) is composed of

- the impulse response matrices from w to $z_Q - z$,
- the upper bound of the static quantization error w, i.e., $d/2$.

So it follows that the right hand side represents the influence of the static quantization error on the output difference $z_Q - z$.

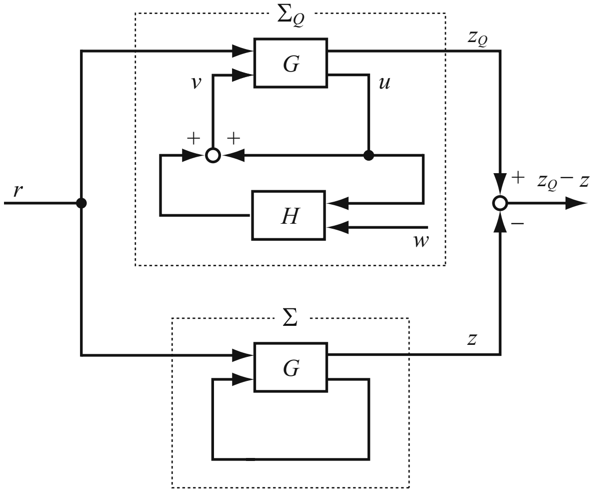


Fig. 3 Error system between quantized system Σ_Q and unquantized system Σ.

Remark 1. Even if (A1) is not satisfied, the weak version of Theorem 1, in which the right-hand side of (6) becomes an upper bound of $E(Q)$, holds. Therefore, although it is rather conservative, the value of $E(Q)$ can be estimated. □

3.2 Optimal Dynamic Quantizers

We next show a solution to Problem 1 (ii).

Theorem 2 ([13, 15]). *For the system Σ_Q, assume* (A1)–(A3). *Then a solution to Problem* 1 (ii) *is given by*

$$Q^* : \begin{cases} \xi(k+1) = \tilde{A}\xi(k) - B_2u(k) + B_2v(k), \\ v(k) \quad = q(-(C_1\tilde{A}^\tau B_2)^+ C_1\tilde{A}^{\tau+1}\, \xi(k) + u(k)) \end{cases} \tag{10}$$

and the minimum value of $E(Q)$ is given by

$$E(Q^*) = \|C_1\tilde{A}^\tau B_2\|\frac{d}{2} \tag{11}$$

where τ is the value of k satisfying the condition in (A2). □

Theorem 2 provides an optimal quantizer, where Assumptions (A1) and (A2) relate to the minimality of $E(Q^*)$ and (A3) does to the stability of Σ_Q.

This result explains an optimal quantization structure as follows. Suppose that Q^* is applied to the error system in Fig. 3. Then the impulse response matrices from r to $z_Q - z$ are $(0, 0, \ldots)$, and those from w to $z_Q - z$ are given by

$$(0,\ldots,\ 0,\ \underset{\substack{\uparrow \\ \tau\text{-th}}}{C_1\tilde{A}^{\tau}B_2},\ 0,\ 0,\ldots) \tag{12}$$

which, actually, corresponds to the minimum. So Q^* plays a role to satisfy (5) and to reduce the signal transfer from w to $z_Q - z$ as small as possible.

Example 1. Consider the system Σ_Q for the feedback system in Fig. 1 (a). Here, P and K are the discrete-time plant and controller obtained from the continuous-time ones

$$P_c : \begin{cases} \dot{x}_P(t) = \begin{bmatrix} 0.1 & 3 \\ -0.8 & 2 \end{bmatrix} x_P(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} v(t), \\ z(t) \ \ = [1\ \ 2]x_P(t), \quad y(t) = [1\ \ 0]x_P(t), \end{cases}$$

$$K_c : \begin{cases} \dot{x}_K(t) = \begin{bmatrix} -10 & 3 \\ -14.7 & -6.1 \end{bmatrix} x_K(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} r(t) + \begin{bmatrix} 10.1 \\ 11.5 \end{bmatrix} y(t), \\ u(t) \ \ = -[2.4\ \ 8.1]x_K(t) + r(t) \end{cases}$$

and the zero-order hold with the sampling period $h := 0.1$. For Q^*, the quantization level is given by $d := 2$.

Fig. 4 shows the simulation result on the time responses of the system Σ_Q, where $x_0 := [0.5\ -0.5\ \ 0\ \ 0]^\top$ ($x := [x_P^\top\ \ x_K^\top]^\top$) and $r(k) \equiv 0$. In addition, the output response of Σ in Fig. 1 (b) is also shown by the thin line in the third figure, where x_0 and r are set to the same values. Though v is a coarse discrete-valued signal in Σ_Q, we see that the output behavior of Σ_Q is almost the same as that of Σ. □

4 Advanced Topics

To apply to real systems, the above basic theory has to be generalized. In this section, we introduce recent results toward this direction.

Numerical Optimization Based Design of Dynamic Quantizers

Since the above discussion holds under several assumptions, the results (especially, Theorem 2) can be applied to a limited class of systems. As an alternative, a design method based on numerical optimization has been developed in [14]. There, by exploiting special structure of Problem 1, the problem is reduced into a linear programming problem, which allows us to efficiently optimize the dynamic quantizer without any strong assumption. A MATLAB implementation entitled "ODQ toolbox" is available at the web site [18].

Decentralized Dynamic Quantizers

As shown in Fig. 5 (a), it is often necessary to have a decentralized structure in the quantizer. In [16], the above results have been extended to the decentralized

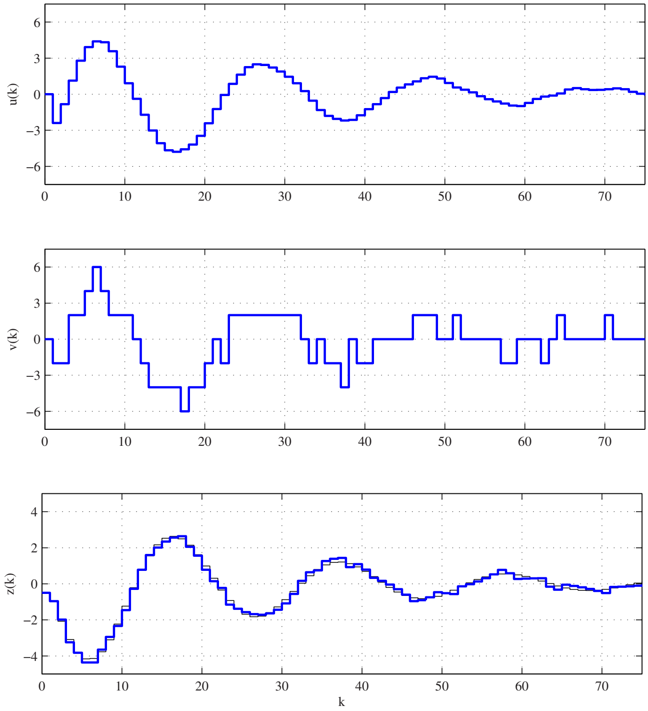
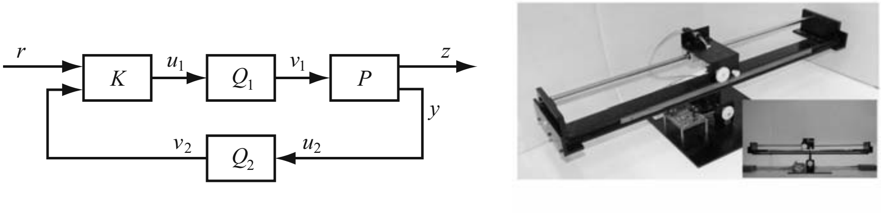


Fig. 4 Responses of optimally quantized system Σ_Q (thick lines) and output response of unquantized system Σ (thin line).

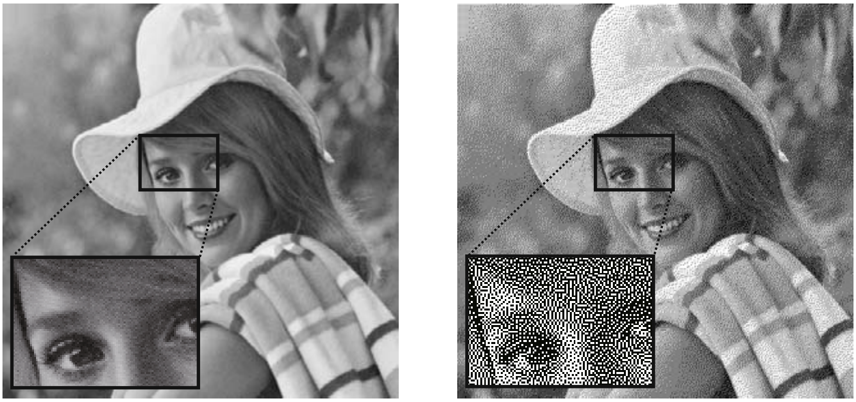


(a) Quantized feedback system with decentralized quantizer.

(b) Seesaw-cart system.

Fig. 5 Decentralized dynamic quantizers [16].

case. Furthermore, an experimental evaluation has been performed with the seesaw-cart system in Fig. 5 (b). There, it is successfully achieved to stabilize the unstable system under the severe condition that the plant input takes one of three values and the controller input does one of seven values.



(a) Original image (8-bit colors).

(b) Halftone image (2-bit colors).

Fig. 6 Binary halftoning by a 2D optimal dynamic quantizer [17].

n-Dimensional Dynamic Quantizers

In [17], the authors have extended the optimal quantizers to an n-dimensional (n-D) version. This can be used not only for control of n-D systems but also for image processing. Fig. 6 shows an example of applying the result to the halftone image processing, which is to transform a grayscale image to a binary image keeping the quality to the eye. This shows the potential of our framework to other fields out of control.

5 Conclusion

The authors' recent results on the control-oriented dynamic quantizers have been reviewed. We hope that this will be utilized to real control applications.

Acknowledgment. The authors would like to thank Mr. Ryosuke Morita, Kyoto University, for supplying data.

References

1. Bemporad, A., Morari, M.: Control of systems integrating logic, dynamics, and constraints. Automatica 35(3), 407–427 (1999)
2. Liberzon, D.: Switching in Systems and Control. Birkhäuser, Boston (2003)
3. Nair, G.N., Fagnani, F., Zampieri, S., Evans, R.J.: Feedback control under data rate constraints: an overview. Proc. of the IEEE 95(1), 108–137 (2007)
4. Goodwin, G.C., Silva, E.I., Quevedo, D.E.: A brief introduction to the analysis and design of networked control systems. In: Proc. of Chinese Control and Decision Conference, pp. 1–13 (2008)

5. Wong, W.S., Brockett, R.W.: Systems with finite communication bandwidth constraints II: Stabilization with limited information feedback. IEEE Trans. on Automatic Control 44(5), 1049–1053 (1999)
6. Brockett, R.W., Liberzon, D.: Quantizer feedback stabilization of linear systems. IEEE Trans. on Automatic Control 45(7), 1279–1289 (2000)
7. Elia, N., Mitter, S.K.: Stabilization of linear systems with limited information. IEEE Trans. on Automatic Control 46(9), 1384–1400 (2001)
8. Fu, M., Xie, L.: The sector bound approach to quantized feedback control. IEEE Trans. on Automatic Control 50(11), 1698–1711 (2005)
9. Liberzon, D., Nesic, D.: Input-to-state stabilization of linear systems with quantized state measurements. IEEE Trans. on Automatic Control 52(5), 767–781 (2007)
10. Quevedo, D.E., Goodwin, G.C., De Dona, J.A.: Finite constraint set receding horizon quadratic control. Int. J. of Robust and Nonlinear Control 14(4), 355–377 (2004)
11. Canudas-de-Wit, C., Rubio, F.R., Fornés, J., Gómez-Estern, F.: Differential coding in networked controlled linear systems. In: Proc. of 2006 American Control Conf., pp. 4177–4182 (2006)
12. Bourdopoulos, G.I.: Delta-Sigma Modulators: Modeling, Design and Applications. Imperial College Press, London (2003)
13. Azuma, S., Sugie, T.: Optimal dynamic quantizers for discrete-valued input control. Automatica 44(2), 396–406 (2008)
14. Azuma, S., Sugie, T.: Synthesis of optimal dynamic quantizers for discrete-valued input control. IEEE Trans. on Automatic Control 53(9), 2064–2075 (2008)
15. Minami, Y., Azuma, S., Sugie, T.: Optimal dynamic quantizers for discrete-valued input feedback control. In: Proc. 46th IEEE Conference on Decision and Control, pp. 2259–2264 (2007)
16. Minami, Y., Azuma, S., Sugie, T.: Optimal decentralized dynamic quantizers for discrete-valued input control: a closed form solution and experimental evaluation. In: Proc. of 2009 American Control Conf., pp. 4367–4372 (2009)
17. Minami, Y., Azuma, S., Sugie, T.: Optimal feedback quantizers for *n*-dimensional systems with discrete-valued input. To appear in Nonlinear Analysis: Hybrid Systems (2009)
18. Morita, R., Azuma, S., Minami, Y., Sugie, T.: ODQ Toolbox, `http://www.robot.kuass.kyoto-u.ac.jp/~morita/odqtoolbox_1.0.0b.zip`

Convergence of Periodic Gossiping Algorithms

Brian D.O. Anderson*, Changbin Yu**, and A. Stephen Morse***

Abstract. In deterministic gossiping, pairs of nodes in a network holding in general different values of a variable share information with each other and set the new value of the variable at each node to the average of the previous values. This occurs by cycling, sometimes periodically, through a designated sequence of nodes. There is an associated undirected graph, whose vertices are defined by the nodes and whose edges are defined by the node pairs which gossip over the cycle. Provided this graph is connected, deterministic gossiping asymptotically determines the average value of the initial values of the variables across all the nodes. The main result of the paper is to show that for the case when the graph is a tree, all periodic gossiping sequences including all edges of the tree just once actually have the same rate of convergence. The relation between convergence rate and topology of the tree is also considered.

Keywords: gossiping algorithms, multi-agent systems, graphs.

Brian D.O. Anderson
The Australian National University and NICTA Ltd., Canberra ACT 2600 Australia
e-mail: Brian.Anderson@anu.edu.au

Changbin(Brad) Yu
The Australian National University, Canberra ACT 2600 Australia
e-mail: Brad.Yu@anu.edu.au

A. Stephen Morse
Yale University, New Haven, CT 06520, USA
e-mail: morse@yale.edu

* His work is supported by NICTA and the Australian Research Council's Discovery Project DP-0877562.

** His work is supported by the Australian Research Council through an Australian Postdoctoral Fellowship under DP-0877562.

*** His work is supported by grants from the US Air Force Office of Scientific Research, the US Army Research Office, the National Science Foundation, and by a gift from the Xerox Corporation.

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 127–138.
springerlink.com

1 Introduction

We postulate a set of $n+1$ agents each holding a value of a scalar variable, and exchanging information according to a certain protocol with the aim of all arriving at common knowledge of the average of the $n+1$ initial values. Such a problem is typically termed a *consensus problem*. Consensus problems are treated in many references, for example [1, 2, 3, 6, 8, 9, 11, 12, 13, 14, 15, 16, 17].

A key aspect of any consensus problem is that agents normally exchange information with a limited subset of the other agents. This aspect is most easily modelled by a graph, whose vertices correspond to the agents, and whose edges correspond to the agent pairs between which information can be exchanged. The graph is directed if exchange is one way, and undirected if two-way; undirected graphs are much more commonly considered. In this form of consensus, and in a discrete-time framework, the value at time $(k+1)$ at agent i is set equal to the average of the values at time k at agent i and all its neighbors. Among the variants on the basic problem, some of the preceding references have also considered problems involving consensus with a leader, time-delays, time-varying connection graphs, random graphs, and the introduction of shift register storage at individual agents. Virtually every result requires an assumption of *connectivity* in the underlying graph, which is intuitively reasonable. A key issue in any consensus problem is defining the rate of convergence, or equivalently, the time constant governing convergence (which is usually exponential).

A special type of consensus algorithm is exemplified by a *gossiping* algorithm. In a gossiping algorithm, at any instant of time, at most one pair of agents can interact, i.e. exchange and average their values. Gossiping references include [4, 5, 7, 10] An underlying graphical structure is assumed for a gossiping algorithm, and the graph must be connected. One can conceive of synchronous or asynchronous selection of edges, and random or deterministic selection of edges. In the latter case, selection of edges on a periodic basis provides an attractive analytical framework.

In this work, we consider gossiping algorithms where the underlying graph is a tree, the simplest connected graph of course, and there is a deterministic periodic protocol causing each edge to be activated once in the underlying period. Each individual gossip can be described by a doubly stochastic matrix (call it a gossiping matrix), and the composition of n successive gossips (the number of edges in a tree on a graph of $n+1$ vertices) corresponds to a product of such gossiping matrices, which is again a doubly stochastic matrix. The eigenvalue with second largest magnitude of this matrix effectively defines the speed of convergence of the algorithm (the largest eigenvalue of course is 1). Obviously, one might expect different speeds for different orderings in the product of the same n gossiping matrices. In fact, this is not the case. It will be shown that the convergence rates are the same for all possible periodic gossip sequences of a given graph with a tree structure.

In the next sections, we shall first consider first a tree that is simply a path graph, and then consider a tree which has just one node of degree greater than 2, before considering more general trees. While we show that the ordering of edges within a tree is immaterial in determining the spectrum of the composition of the individual gossip matrices over the cycle, the shape of the tree is relevant. Trees with the same number of edges do not all have the same associated convergence rate.

2 Gossiping in a Path Graph Tree

Consider a tree with $n+1$ vertices, forming a single path. See Figure 1. There are edges between vertices $i, i+1$ for $i = 1, \dots, n$. Call S_i the matrix which is the direct sum of the $(i-1) \times (i-1)$ identity matrix, the 2×2 matrix with $1/2$ for each entry, and the $(n-i-1) \times (n-i-1)$ identity matrix. Then S_i is the stochastic matrix modelling gossiping between nodes $i, i+1$. Let us call such a gossiping matrix a *primitive* gossiping matrix. For future reference, we observe, noting the proof is easy, that

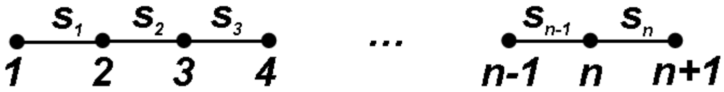


Fig. 1 A path graph tree; S_i is the gossiping matrix when nodes $i, i+1$ gossip

Lemma 1. *Let the S_i be as defined above. Then the matrices S_i, S_j commute if and only if $|i-j| \geq 2$, i.e. if and only if S_i and S_j correspond to nonadjacent edges.*

Below, we state a theorem, omitting the proof because of space limitations, which draws on this lemma and the facts that:

1. For any two square matrices A, B of the same size the eigenvalues of AB and of BA are the same.
2. (Consequence of the above). The spectrum of any product $A_1 A_2 \dots A_n$ is unchanged by cyclic permutation of the A_i

For a given tree, let us call a gossiping sequence *complete* when the edges of the tree are ordered, and gossips are executed in the corresponding order; thus each edge is used once and only once to define a complete sequence. Call the associated product of primitive gossip matrices a *composite* gossiping matrix. Call an infinite gossiping sequence *periodic* when the same complete sequence is repeatedly used. The main result now for a path tree is:

Theorem 1. *Let the S_i be as defined above. Let $\pi : \{1, 2, \dots, n\} \to \{1, 2, \dots, n\}$ be an arbitrary permutation. Then the eigenvalues of $S_1 S_2 \dots S_n$ are the same as the eigenvalues of $S_{\pi(1)} S_{\pi(2)} \cdots S_{\pi(n)}$*

Proofs of all theorems are omitted due to length limitations. In some cases, outlines will be provided. The tools used in proving Theorem 1 are the lemma above, and the two well-known facts following it. This means that if the particular primitive gossiping matrices S_i are replaced by more general matrices, T_i say, with the property that T_i, T_j commute if $|i-j| \geq 2$, then the theorem will hold with the S_i replaced by T_i. Note that the convergence rate associated with a periodic gossip sequence is given by the second largest magnitude of an eigenvalue of the composite gossiping matrix, and so the theorem in effect is asserting that the convergence rate to the average of the variables at each node is independent of the ordering of the complete gossip sequence.

There is a second important invariant of the ordering that we will make use of below. The proof of the theorem is not so obvious, but actually rests on very similar observations to those used in proving Theorem 1.

Theorem 2. *Let the S_i be as defined above. Let $\pi : \{1,2,\ldots,n\} \to \{1,2,\ldots,n\}$ be an arbitrary permutation and let $T = S_{\pi(1)}S_{\pi(2)}\ldots S_{\pi(n)}$ be the associated product. Partition T as*

$$T = \begin{bmatrix} A & b \\ c & d \end{bmatrix} \tag{1}$$

where d is scalar. Define the transfer function associated with node $n+1$ as

$$w(z) = d + c(zI - A)^{-1}b \tag{2}$$

Then $w(z)$ is independent of the ordering of the primitive gossiping matrices in T.

Theorem 1 actually almost follows from Theorem 2, in view of the easily derived formula:

$$|zI - T| = [z - w(z)]|zI - A| \tag{3}$$

In the event that $\{A,b,c,d\}$ is a *minimal* realization of $w(z)$, Theorem 1 is an immediate consequence of Theorem 2.

3 Trees with One Node of Degree Greater Than 2

In this section, we start to generalize the trees for which we can make statements about many gossip matrix product orderings having the same spectra (and indeed

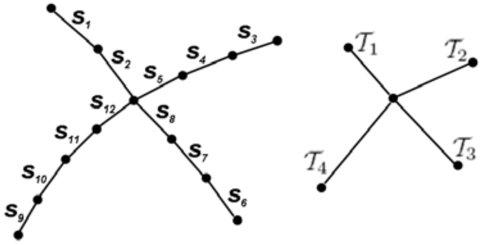


Fig. 2 First generalization of path graph tree; S_i denote gossiping matrices and T_i products of gossiping matrices.

the same transfer function associated with a certain vertex). The main result deals with trees in which there is just one node with degree greater than 2, see Figure 2, and establishes that again, there is a single spectrum independent of the product ordering. We term such trees *star trees*.

It is helpful to fix some notation. Suppose a star tree has $(n+1)$ nodes with one node of degree $p > 2$. Number this as the $(n+1)$-th. Index the paths between this node and the leaf nodes as $1, 2, \ldots, p$ with the length of the paths given by $n_1, n_2, \ldots, n_p$, so that $n = \sum_i n_i$. Number the nodes by working from the leaf end of the first path to the last node before the node of degree p, then by working from the outer end of the second path to the last node before the node of degree p, and so on. Order links in the same manner. The associated primitive gossiping matrices are $S_1, S_2, \ldots, S_{n_1}$ for path 1, then $S_{n_1+1}, \ldots, S_{n_1+n_2}$ for path 2, and so on, through to S_n. For each $i = 1, 2, \ldots, p$, define a composite gossiping matrix by

$$T_i = S_{n_{i-1}+1} S_{n_{i-1}+2} \ldots S_{n_i} \tag{4}$$

where also $n_0 = 0$

Using this notation, we first record a lemma in which the number of possible spectra is greatly reduced. The proof runs similarly to the proof of the result for path graphs, but again is omitted.

Lemma 2. *Consider a star tree with $(n+1)$ nodes and p paths from leaf nodes to a vertex of degree p, with numbering as described above. Then for an arbitrary complete gossip, the spectrum of the product of the gossip matrices will be the same as the spectrum of $T_1 T_2 \ldots T_p$ or some permutation thereof. The permutation in question can be obtained by deleting all S_i from the product corresponding to the complete gossip, other than those with $i = n_1, i = n_1 + n_2, \ldots, i = n$ and then replacing $S_{n_1+n_2+\ldots+n_j}$ by T_j.*

If for example, $p = 3$, there are apparently two possible spectra which might result, corresponding to the spectra of $T_1 T_2 T_3$ and $T_1 T_3 T_2$. Note that all other four orderings in a product of the three T_i are cyclic permutations on one of these two, and thus have the same spectrum. For the tree of Figure 2, there are apparently six spectra, viz. those of $T_1 T_2 T_3 T_4$, $T_1 T_2 T_4 T_3$, $T_1 T_3 T_2 T_4$, $T_1 T_3 T_4 T_2$, $T_1 T_4 T_2 T_3$, $T_1 T_4 T_3 T_2$. Nevertheless, we prove in the theorem below, perhaps surprisingly, that all such spectra are in fact the same.

Theorem 3. *Adopt the same hypotheses as in Lemma 2. Then for an arbitrary complete gossip, the spectrum of the product of the gossip matrices is independent of the ordering of the individual gossips. A formula for the characteristic polynomial whose roots determine the spectrum is obtainable as follows. Let the matrix T_i be partitioned as*

$$T_i = \begin{bmatrix} I_{n_1} & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & I_{n_2} & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & A_i & 0 & \dots & b_i \\ 0 & 0 & \dots & 0 & I_{n_{i+1}} & \dots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & c_i & 0 & \dots & d_i \end{bmatrix} \tag{5}$$

Here, b_i, c_i, d_i are respectively a column n_i-vector, a row n_i-vector and a scalar, while A_i is an $n_i \times n_i$ matrix. Define $w_i(z) = d_i + c_i(zI - A)^{-1}b_i$. Then the eigenvalues of the complete gossiping matrix are the zeros of $1 - z^{-1}\prod_{i=1}^{p} w_i(z)$.

Outline of Proof: From the preceding lemma, we know that the spectrum will be the same as the spectrum of $T_1T_2 \dots T_p$ or the same product but after permutation of the T_i. To prove the claim of the theorem, it is enough to show that the spectrum of $T_1T_2 \dots T_p$ is given by the zeros of $1 - z^{-1}\prod_{i=1}^{p} w_i(z)$. Then the desired result will follow since this last expression is independent of the ordering of the T_i. We demonstrate the formula for the case $p = 3$; the method of proof is obviously generalizable. One can compute

$$T_1T_2T_3 = \begin{bmatrix} A_1 & b_1c_2 & b_1d_2c_3 & b_1d_2d_3 \\ 0 & A_2 & b_2c_3 & b_2d_3 \\ 0 & 0 & A_3 & b_3 \\ c_1 & d_1c_2 & d_1d_2c_3 & d_1d_2d_3 \end{bmatrix} \tag{6}$$

One can check that this composite gossiping matrix is also the state update matrix for the closed-loop system depicted in the Fig. 3. The spectrum of the associated state update matrix corresponds to the zeros of the characteristic polynomial of the closed-loop system depicted in the figure. Consider the elementary redrawing of the figure, shown in Fig. 4, for which the characteristic polynomial formula follows easily.

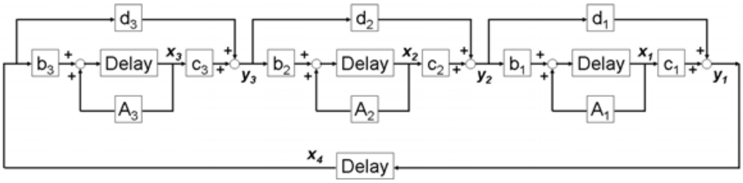


Fig. 3 The corresponding closed-loop system.

Remark 1. The particular form of the S_i is not relevant in the above result. Each T_i could be the gossiping matrix of a rather large tree, rather than a path graph.

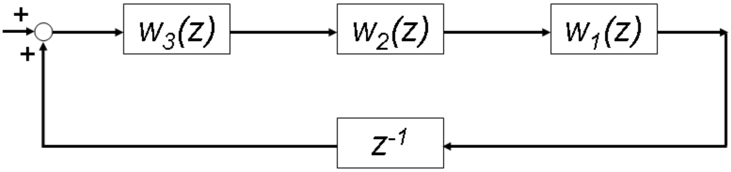


Fig. 4 Elementary redrawing of Fig. 3.

Remark 2. Partition the composite matrix $T_1T_2T_3$ in a similar manner to each of the T_i, thereby defining an $n\times n$ matrix A, column and row n-vectors b and c and a scalar d. Then it is straightforward to check for the transfer function associated with node $n+1$ that there holds

$$w(z) := d + c(zI - A)^{-1}b = \prod_{i=1}^{3} w_i(z) \tag{7}$$

This second conclusion is a general one of course, and we sum it up as follows:

Theorem 4. *Adopt the same hypothesis as in Theorem* 3. *Let T be the matrix of the complete gossip, and write it as*

$$T = \begin{bmatrix} A & b \\ c & d \end{bmatrix} \tag{8}$$

Associate with node $n+1$ the transfer function $w(z) = d + c(zI - A)^{-1}b$. Then $w(z) = \prod_{i=1}^{p} w_i(z)$.

We are now ready to consider general trees.

4 General Trees

In effect, we will treat general trees using an induction process. One of the inductive steps is as follows.

Lemma 3. *Let $\mathscr{T}_0$ be an arbitrary tree of $(n+1)$ vertices, and let v_0 be an arbitrary vertex. Augment the tree by adding an edge from v_0 to a new vertex v, and call this new tree $\mathscr{T}$. Adopt a node ordering such that v_0, v are the second last and last node, i.e. the $(n+1)$-th and $(n+2)$-th node of $\mathscr{T}$. Suppose that a complete gossiping matrix for $\mathscr{T}_0$ is given by*

$$T_0 = \begin{bmatrix} A_0 & b_0 \\ c_0 & d_0 \end{bmatrix} \tag{9}$$

where d_0 is a scalar, and suppose that the following two properties hold:

1. *The spectrum of T_0 is independent of the ordering used to generate it.*
2. *The transfer function $w_0(z) = d_0 + c_0(zI - A_0)^{-1}b_0$ associated with node v_0, numbered $n+1$, is independent of the ordering used to generate T_0*

Then the corresponding properties hold for any complete gossiping matrix T for $\mathscr{T}$, i.e. if

$$T = \begin{bmatrix} A & b \\ c & d \end{bmatrix} \tag{10}$$

where d is a scalar, then the spectrum of T and the transfer function $w(z) = d + c(zI - A)^{-1}b$ associated with node v, numbered $n+2$, are independent of the ordering used to form T.

This lemma can be proved by appeal to Theorem 3 to establish the spectral result. The transfer function result can be obtained by adjusting the argument of Theorem 4, which calculates the transfer function of v_0, in order to calculate the transfer function associated with v.

Finally, we can state our main result.

Theorem 5. *Let $\mathscr{T}$ be a tree, and let T be a composite gossip matrix associated with a complete gossip sequence. Then the spectrum of T is independent of the order of the individual gossips. Further, the transfer function associated with any node is also independent of the order of the gossips.*

Proof. Any tree may be regarded as being built by a sequence of operations which successively add branches to an existing tree. In the light of the last Lemma 3 and the earlier theorems, an inductive proof exists to establish the spectral result. To establish the transfer function result, choose a node in the tree. Associate with each of the branches incident on that node a subtree. The associated transfer function is by a trivial variant on Theorem 4 a product of the individual transfer functions for the individual trees, and the inductive argument again applies.

5 Examples

In this section, we provide a series of simulations verifying the analytical results. Moreover, extensive simulations for 12-edge trees with one node of degree greater than 2 partially unveil some relationships between the topology and the convergence rate (as measured by the second largest eigenvalue). We attempt to summarize the observed patterns here.

Simulations for path tree graphs are easily created and the results are immediate. For more general trees, we used a pseudo-code for generating a *sufficiently* random tree and random gossip order. Simulation results are consistent with analysis, that is, the spectrum is independent of the order and only determined by the topology of the tree.

```
Specify the number of nodes $V$ in the tree
Create a root node (1), add node (2) and edge (1,2)
Set (2) to be the forking node

WHILE (there is unconnected node(s))
DO {
    – generate a random branch number $K$
    – create $K$ new nodes (or until there are $V$ nodes)
    – edges between every new node to forking node
    – select a new forking node among the $K$ new nodes
}

Create the gossip matrix for every edge
Pseudo random permutation of all edges and order them accordingly
Obtain the product of gossip matrices in that order
```

With a view to linking the convergence rate of a complete gossip to the topology of a tree, we conducted simulations for 12-edge trees with one node of degree greater than 2; for convenience, we call such trees the "star-trees".

Let us denote the tree with $[n_1, n_2, n_3, \ldots]$ where the n_i denote the length (i.e. number of edges) of the branches, and the n_i are sorted in descending order.

The results are summarized in Table 1. Note that the path tree-[12], which is the same as [11,1] or any $[n, 12-n]$, is included for comparison only.

From the table, we observe two patterns (with small exceptions denoted with $*$):

Pattern 1. For 12-edge star-trees with at least one branch length exceeding 4, there exists a partial order of these trees corresponding to the convergence rate of the gossiping sequence

Pattern 2. For 12-edge star-trees with at least 8 branches and maximum branch length 3, the more branches it has, the slower its corresponding gossiping sequence converges. For trees with same number of branches, no specific rules are derived for their order.

It can be noted that there are a number of entries which have no specific pattern, and more investigation is required to understand these. Moreover, it is important to verify the validity of the above patterns for star-trees of other sizes. We would expect that the convergence rate will be affected by at least the depth (length of the longest branch) and the breadth (number of branches) of the tree. We are conducting more simulations to explore these points further.

Table 1 12-edge star-trees and their second largest eigenvalue

Tree	$|\lambda_2|$	remark
[12]	0.9427	path tree
[11,1]	0.9427	path tree
[10,2]	0.9427	path tree
[10,1,1]	0.9413	Pattern 1 follows
[9,2,1]	0.9393	
[9,1,1,1]	0.9380	
[8,3,1]	0.9368	
[8,2,2]	0.9344	
[8,2,1,1]	0.9333	
[7,3,2]	0.9285	
[6,3,2,1]	0.9145	
[5,4,3]	0.9045	
[5,4,2,1]	0.9045	
[5,2,1,...,1]	0.8849	
[5,1,...,1]	0.8830 *	Exception to Pattern 2
[4,4,4]	0.8848 *	no patterns observed here
[4,4,3,1]	0.8737	onwards due to exceptions
[4,4,2,2]	0.8780 *	
[4,1,...,1]	0.8393	
[3,3,3,3]	0.8256	
[3,3,2,2,2]	0.7500	
[3,2,2,2,2,1]	0. 7500	
[3,2,1,...,1]	0.7500	
[3,1,...,1]	0.7532 *	
[2,2,2,2,2,2]	0.7667 *	
[2,2,2,2,2,1,1]	0.7510 *	
[2,2,2,2,1,1,1,1]	0.7366	Pattern 2 follows
[2,2,2,1,...,1]	0.7409	
[3,2,1,...,1]	0.7500	
[3,1,...,1]	0.7532	
[2,2,1,...,1]	0.7673	
[2,1,...,1]	0.8000	
[1,1,...,1]	0.8315	

6 Conclusions

The main result of this paper is that for a fixed tree, a composite gossip matrix obtained by multiplying together primitive gossip matrices corresponding to each edge of the tree has a spectrum that is independent of the ordering of these matrices. This means that it is straightforward to determine the convergence rate of a periodic gossip. The rate depends on the ‘shape’ of the tree. Our examples of 12-edge star-trees show the complexity of making a connection of the two. More research is required to properly understand what aspects of a tree are associated with the fastest convergence rates.

The reader may also wonder what will happen in the case of graphs with cycles; for example, could it be that the main result of this paper actually applies to more general graphs? The answer is no. We have verified for a graph which is a pure cycle of 6 nodes and 6 edges that the convergence rate for a periodic gossip does indeed depend on the ordering of the edges in a complete gossip sequence.

References

1. Angeli, D., Bliman, P.-A.: Extension of a result by moreau on stability of leaderless multi-agent systems. In: Proc. of the 44th IEEE Conference on Decision and Control, pp. 759–764 (2005)
2. Bertsekas, D.P., Tsitsiklis, J.N.: Parallel and Distributed Computation: Numerical Methods. Prentice Hall, Englewood Cliffs (1989)
3. Blondel, V.D., Hendrickx, J.M., Olshevsky, A., Tsitsiklis, J.N.: Convergence in multiagent coordination, consensus and flocking. In: Proc. of 44th IEEE Conference on Decision and Control, pp. 2996–3000 (2005)
4. Boyd, S., Ghosh, A., Prabhakar, B., Shah, D.: Gossip algorithms: design, analysis and applications. In: Proc. of the 24th Annual Joint Conference of the IEEE Computer and Communications Socsieties (INFOCOM), pp. 1653–1664 (2005)
5. Boyd, S., Ghosh, A., Prabhakar, B., Shah, D.: Randomized gossip algorithms. IEEE Transactions on Information Theory 52, 2508–2530 (2006)
6. Cao, M., Spielmam, D.A., Morse, A.S.: A lower bound on convergence of a distributed network consensus algorithm. In: Proc. of the 44th IEEE Conference on Decision and Control, pp. 2356–2361 (2005)
7. Cao, M., Spielman, D.A., Yeh, E.M.: Accelerated gossip algorithms for distributed computation. In: Proc. of the 44th Annual Allerton Conference on Communication, Control, and Computation, pp. 952–959 (2006)
8. Jadbabaie, A., Lin, J., Morse, A.S.: Coordination of groups of mobile autonomous agents using nearest neighbor rules. IEEE Transactions on Automatic Control 48, 988–1001 (2003)
9. Li, S., Wang, H.: Multi-agent coordination using nearest neighbor rules: revisiting the Vicsek model (2004), arXiv:cs/0407021v2, `http://arxiv.org/abs/cs.MA/0407021`
10. Liu, J., Anderson, B.D.O., Cao, M., Morse, A.S.: Analysis of accelerated gossip algorithms. To appear in Proc. of the 48th IEEE Conference on Decision and Control (2009)
11. Morse, A.S., Cao, M., Anderson, B.D.O.: Reaching a consensus in a dynamically changing environment — a graphical approach. SIAM Journal on Control and Optimization 47, 575–600 (2008)
12. Moreau, L.: Stability of multi-agent systems with time-dependent communication links. IEEE Transactions on Automatic Control 50, 169–182 (2005)
13. Olfati-Saber, R., Murray, R.M.: Consensus problems in networks of agents with switching topology and time-delays. IEEE Transactions on Automatic Control 49, 1520–1533 (2004)
14. Ren, W., Beard, R.W.: Distributed Consensus in Multi-vehicle Cooperative Control: Theory and Applications. Springer, London (2008)

15. Ren, W., Beard, R.W., Atkins, E.M.: A survey of consensus problems in multi-agent coordination. In: Proc. of 2005 American Control Conference, pp. 1859–1864 (2005)
16. Tsitsiklis, J.N., Bertsekas, D.P., Athans, M.: Distributed asynchronous deterministic and stochastic gradient optimization algorithms. IEEE Transactions on Automatic Control 31, 803–812 (1986)
17. Vicsek, T., Czirok, A., Ben-Jacob, E., Cohen, I., Schochet, O.: Novel type of phase transitions in a system of self-driven particles. Physical Review Letters 75, 1226–1229 (1995)

Distributed PageRank Computation with Link Failures

Hideaki Ishii and Roberto Tempo

Abstract. The Google search engine employs the so-called PageRank algorithm to rank the search results by quantifying the importance of each web page. In this paper, we continue our recent work on distributed randomized computation of PageRank, where the pages locally determine their values by communicating with linked pages. In particular, we propose a distributed randomized algorithm with limited information, where only part of the linked pages is required to be contacted. This is useful to enhance flexibility and robustness in computation and communication.

This paper is dedicated to Yutaka Yamamoto on the occasion of his 60th birthday.

1 Introduction

The performance of search engines heavily relies on the capability of listing search results so that users can quickly have access to the desired information. One effective and objective way to quantify the importance or popularity of the web pages is by simply examining the link structure of the web. The so-called PageRank algorithm at Google follows such an idea and ranks pages higher when they have links from more important pages (see, e.g., [3, 4, 19]).

To execute the PageRank algorithm, however, the size of the web poses serious difficulties. Google is said to have over 8 billion web page indices and moreover computes the PageRank in a centralized fashion. In view of the rapid growth of the web, it is critical to develop more efficient computational methods. In this regard, a line of current research is towards distributed computation of the PageRank. In [25], block structures in the web are exploited to apply Markov chain methods

Hideaki Ishii
Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, 4259 Nagatsuta-cho, Midori-ku, Yokohama 226-8502, Japan
e-mail: ishii@dis.titech.ac.jp

Roberto Tempo
IEIIT-CNR, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy
e-mail: roberto.tempo@polito.it

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 139–150.
springerlink.com

while the work of [1] utilizes techniques from Monte Carlo simulation. In [6, 18], the application of numerical analysis methods known as asynchronous iterations [2] is discussed. Other works include [17], where adaptive methods allocate computational resources depending on the rate of convergence.

In our recent work [13], we developed a distributed randomized approach for the PageRank computation; for recent advances on probabilistic methods in systems and control, see [22]. The approach is distributed in that each page computes its own PageRank value locally by communicating with the pages that are connected by direct links. That is, each page exchanges its value with the pages to which it links and those linked to it. Randomization is with respect to the time that each page decides to initiate the communication. The time is randomly chosen and is independent among the pages. Hence, there is no need of a fixed order among the pages or a leader agent that specifies the pages to start updates. It is also stressed that relatively small communication and computation are required for the agents. On the other hand, in [14], we considered a centralized scheme for computing the bounds on the PageRank values when the web data contains uncertainties.

In this paper, we further explore the approach of [13] to enhance flexibility and robustness under limited information; an earlier version of this work has appeared in [15]. Specifically, we are interested in situations where each page initiating an update contacts only part of its linked pages. We continue to work in the probabilistic setting, and such pages are determined in a random manner. The links not used for communication at the time of updates will be referred to as the *failing* links. This feature would be useful, for example, when the computation/communication load among the pages must be reduced, but the rate of updates should be kept at the same level. In this respect, this scheme is more flexible than that in [13] because in addition to the rate of updates for each page, the rate for link selection may be specified. This algorithm can be also applied when communication is unreliable due to link failures and/or packet losses. In such a case, it may not be possible to contact all linked pages at the same time. A simple way to model packet losses is to consider them as an outcome of Bernoulli random processes, which has been widely adopted in the fields of networked control and consensus; see, e.g., [7, 8, 11, 12, 20]. This channel model can be incorporated into the proposed scheme.

As discussed in [13], it is important to note that the proposed distributed randomized approach has been motivated by the recent development in the multi-agent problems. In particular, our approach has strong ties with the stochastic versions of the consensus problems (e.g., [9, 21, 23, 24]). From the viewpoint of consensus, it is natural to treat the web as a network of agents capable of local computation as well as communication with neighbors. It is further emphasized that there are similarities at the technical level. In the algorithm for PageRank computation, stochastic matrices play a crucial role, but in a slightly different form than consensus problems.

This paper is organized as follows: We first provide an overview of the PageRank problem in Section 2. This is followed by Section 3, where we summarize the distributed approach of [13]. In Section 4, we present a distributed algorithm which allows for link failures and prove its convergence. We illustrate the results through a numerical example in Section 5. Finally, in Section 6, concluding remarks are given.

Notation: For vectors and matrices, inequalities are used to denote entry-wise inequalities: For $X,Y \in \mathbb{R}^{n\times m}$, $X \le Y$ implies $x_{ij} \le y_{ij}$ for $i = 1,\dots,n$ and $j = 1,\dots,m$; in particular, we say that the matrix X is nonnegative if $X \ge 0$ and positive if $X > 0$. A probability vector is a nonnegative vector $v \in \mathbb{R}^n$ such that $\sum_{i=1}^n v_i = 1$. By a stochastic matrix, we refer to a column-stochastic matrix, i.e., a nonnegative matrix $X \in \mathbb{R}^{n\times n}$ with the property that $\sum_{i=1}^n x_{ij} = 1$ for $j = 1,\dots,n$. Let $\mathbf{1} \in \mathbb{R}^n$ be the vector with all entries equal to 1 as $\mathbf{1} := [1 \cdots 1]^T$. Similarly, $S \in \mathbb{R}^{n\times n}$ is the matrix with all entries being 1. The norm $\|\cdot\|$ for vectors is the Euclidean norm.

2 The PageRank Problem

The PageRank problem is now briefly described based on, e.g., [3, 4, 19]. Consider a network of n web pages indexed from 1 to n. The network is represented by the directed graph $\mathscr{G} = (\mathscr{V},\mathscr{E})$. Here, $\mathscr{V} := \{1,2,\dots,n\}$ is the set of vertices corresponding to the web page indices while $\mathscr{E} \subset \mathscr{V}\times\mathscr{V}$ is the set of edges representing the links among the pages. The vertex i is connected to the vertex j by an edge, i.e., $(i,j) \in \mathscr{E}$, if page i has an outgoing link to page j, or in other words, page j has an incoming link from page i.

The objective of the PageRank algorithm is to assign some measure of importance to each web page. The PageRank value, or simply the value, of page $i \in \mathscr{V}$ is a real number denoted by $x_i^* \in [0,1]$. The values are ordered: $x_i^* > x_j^*$ implies that page i is more important than page j.

The pages are ranked according to the rule that a page having links from important pages is also important. This is done in such a way that the value of one page equals the sum of the contributions from all pages that have links to it. Specifically, we define the value of page i by

$$x_i^* = \sum_{j\in\mathscr{L}_i} \frac{x_j^*}{n_j},$$

where $\mathscr{L}_i := \{j : (j,i) \in \mathscr{E}\}$, i.e., this is the set of page indices that are linked to page i, and n_j is the number of outgoing links of page j. It is customary to normalize the total of all values as $\sum_{i=1}^n x_i^* = 1$.

Let the values be in the vector form as $x^* \in [0,1]^n$. Then, the PageRank problem can be restated as

$$x^* = Ax^*, \;\; x^* \in [0,1]^n, \;\; \sum_{i=1}^n x_i^* = 1, \tag{1}$$

where the link matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ is given by

$$a_{ij} := \begin{cases} \frac{1}{n_j} & \text{if } j \in \mathscr{L}_i, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The value vector x^* is a nonnegative unit eigenvector corresponding to the eigenvalue 1 of A. In general, for this eigenvector to exist and then to be unique, it is sufficient that the web as a graph is strongly connected[1]. However, the web is known not to be strongly connected. Thus, the problem is slightly modified as follows.

First, note that in the real web, the so-called dangling nodes, which are pages having no links to others, are abundant. To simplify the discussion, we redefine the graph by bringing in artificial links. As a result, the link matrix A becomes a stochastic matrix, having at least one eigenvalue equal to 1.

Next, to guarantee the uniqueness of this eigenvalue, a modified version of the values has been introduced in [3] as follows: Let $m \in (0,1)$, and let the modified link matrix $M \in \mathbb{R}^{n\times n}$ be defined by

$$M := (1-m)A + \frac{m}{n}S. \tag{3}$$

This matrix is a positive stochastic matrix. By the Perron theorem [10], the eigenvalue is simple and is the unique one with maximum modulus, and the corresponding eigenvector is positive. Hence, we redefine the value vector x^* by using M as

$$x^* = Mx^*, \;\; x^* \in [0,1]^n, \;\; \sum_{i=1}^{n} x_i^* = 1. \tag{4}$$

We note that in the original paper [3], a typical value for m is chosen as $m = 0.15$. This value is employed in the rest of the paper.

Because of the large dimension of the link matrix M, the computation of the value vector x^* relies on the power method. That is, x^* is computed through the recursion

$$x(k+1) = Mx(k) = (1-m)Ax(k) + \frac{m}{n}\mathbf{1}, \tag{5}$$

where $x(k) \in \mathbb{R}^n$ and the initial condition $x(0) \in \mathbb{R}^n$ is a probability vector. Notice that the second equality can be established because A is stochastic and thus $x(k)$ is a probability vector, resulting in $Sx(k) \equiv \mathbf{1}$.

The following lemma shows that, using this method, we can asymptotically find the value vector (e.g., [10]).

Lemma 1. For any initial condition $x(0)$, in the update scheme (5) using the modified link matrix M, it holds that $x(k) \to x^*$ as $k \to \infty$.

We now comment on the convergence rate of this scheme. Denote by $\lambda_1(M)$ and $\lambda_2(M)$ the largest and the second largest eigenvalues of M in magnitude. Then, for the power method applied to M, the asymptotic rate of convergence depends on the ratio $|\lambda_2(M)/\lambda_1(M)|$. Since M is a positive stochastic matrix, we have $\lambda_1(M) = 1$ and $|\lambda_2(M)| < 1$. Furthermore, it is shown in [19] that the structure of the link matrix M leads us to the bound $|\lambda_2(M)| \leq 1-m$. For the value $m = 0.15$, the asymptotic rate of convergence is bounded by 0.85.

[1] A directed graph is said to be strongly connected if for any two vertices $i, j \in \mathscr{V}$, there is a sequence of edges which connects i to j.

3 Distributed Randomized Approach

In this section, we summarize the distributed randomized algorithm for computing the PageRank values from [13].

Consider the web with n pages described in Section 2. The basic protocol employed in this scheme is as follows: At each time k, the page i initiates its PageRank value update (i) by sending its value to the pages to which it is linked and (ii) by requesting the pages that link to it for their values. All pages involved here update their values based on the newly available information.

These updates can take place in a fully distributed and randomized manner. The decision to make an update is a random variable. In particular, this is determined under a given probability $\alpha \in (0,1]$ at each time k, and hence, the decision can be made locally at each page. The probability α is however a global parameter, and all pages in the web use the same value.

Formally, the proposed distributed update scheme is described as follows. Let $\eta_i(k) \in \{0,1\}$, $i = 1,\dots,n$, be i.i.d. Bernoulli processes given by

$$\eta_i(k) = \begin{cases} 1 & \text{if page } i \text{ initiates an update at time } k, \\ 0 & \text{otherwise} \end{cases}$$

for $k \in \mathbb{Z}_+$, where their probability distributions are specified by

$$\alpha = \mathrm{Prob}\big\{\eta_i(k) = 1\big\}. \tag{6}$$

The process $\eta_i(k)$ is generated at the corresponding page i, and when its value is 1, then the page will follow the protocol outlined above so that an update is initiated. Let $\eta(k) := [\eta_1(k) \ \cdots \ \eta_n(k)]$ be the notation in a vector form.

Now, consider the distributed update scheme given by

$$x(k+1) = (1-\hat{m})A_{\eta(k)}x(k) + \frac{\hat{m}}{n}\mathbf{1}, \tag{7}$$

where $x(k) \in \mathbb{R}^n$ is the state whose initial condition satisfies $x(0) \geq 0$ and $\sum_{i=1}^{n} x_i(0) = 1$; $\hat{m} \in (0,1)$ is the parameter used instead of m in the centralized case, and let

$$\hat{m} = \frac{[1-(1-\alpha)^2]m}{1-m(1-\alpha)^2}. \tag{8}$$

The distributed link matrices A_q for $q \in \{0,1\}^n$ are given as follows:

$$\big(A_q\big)_{ij} := \begin{cases} a_{ij} & \text{if } q_i = 1 \text{ or } q_j = 1, \\ 1 - \sum_{h:\ q_h=1} a_{hj} & \text{if } q_i = 0 \text{ and } i = j, \\ 0 & \text{if } q_i = q_j = 0 \text{ and } i \neq j, \end{cases} \qquad i,j \in \mathscr{V}. \tag{9}$$

These matrices have the following properties: (i) If $q_i = 1$, then the ith column and the ith row are the same as those in the original link matrix A. (ii) If $q_i = 0$, then the

ith diagonal entry is chosen so that the entries of the ith column add up to 1. (iii) All other entries are 0. Hence, these matrices are constructed to be stochastic.

In this scheme, each page i also computes the time average of its own state x_i. Let $y(k)$ be the average of the past and current states $x(0),\ldots,x(k)$ as

$$y(k) := \frac{1}{k+1}\sum_{\ell=0}^{k} x(\ell),\ \ k \in \mathbb{Z}_+. \tag{10}$$

We say that, for the distributed update scheme, the PageRank value x^* is obtained through the time average y if, for each initial condition $x(0)$, $y(k)$ converges to x^* in the mean-square sense as follows:

$$E\left[\left\|y(k) - x^*\right\|^2\right] \to 0, \quad k \to \infty. \tag{11}$$

This type of convergence is known as ergodicity for stochastic processes.

For completeness, we restate the main result of [13].

Theorem 1. Consider the distributed update scheme (7). For any update probability $\alpha \in (0,1]$, the PageRank value x^* is obtained through the time average y as in (11).

We comment on this distributed update scheme. As can be seen in (7), the scheme can be implemented decentrally. Clearly, each page communicates only with pages sharing direct links. Such links correspond to the nonzero entries of the link matrix A. The parameter α determines the probability of updates to occur and thus the communication load among the pages. At page i, the amount of computation is fairly small since the state $x_i(k)$ and its time average $y_i(k)$ are scalars.

This distributed update scheme can also be viewed as a generalization of the original centralized scheme (5) in Section 2. By using the update probability of $\alpha = 1$, all pages initiate their updates at all times. In this case, we have $\eta(k) \equiv [1\ \cdots\ 1]$ and thus, the distributed link matrix $A_{\eta(k)}$ is equal to the original A. Furthermore, the parameter $\hat{m}$ coincides with m.

4 A Distributed Scheme with Link Failures

In this section, we extend the distributed approach to handle situations where only part of the links are used for communication when a page initiates an update. That is, we examine how an update can be carried out when not all values from linked pages are available; we say that link failures occur in this case. We continue to work in the probabilistic setting and assume that such links are randomly selected. This scheme would be useful when the communication load among the pages must be reduced or when some pages cannot be reached because of link failures and/or packet losses.

The set of failing links where no communication takes place at time k is denoted by $\Delta(k)$. This is a subset of the edges that link to or from the pages initiating the updates; we denote such a set by $\mathscr{E}_{\eta(k)}$, which is formally defined by

$$\mathscr{E}_q := \big\{(i,j) \in \mathscr{E} :\ q_i = 1 \text{ or } q_j = 1\big\}, \quad q \in \{0,1\}^n. \tag{12}$$

For the set $\varDelta(k)$ at time k, we assume that if $(i,j) \in \varDelta(k)$ and $(j,i) \in \mathscr{E}_{\eta(k)}$, then $(j,i) \in \varDelta(k)$ for $(i,j) \in \mathscr{E}_{\eta(k)}$. This represents symmetry in the link failures; if a link from one page to another is failing at time k, then the link in the other direction must be failing as well. The set $\varDelta(k)$ is a random process specified by the link failure probability $\delta \in [0,1)$ under the probability distribution

$$\delta = \mathrm{Prob}\big\{(i,j) \in \varDelta(k) \mid \eta(k) = q\big\}, \quad \forall (i,j) \in \mathscr{E}_q,\ q \in \{0,1\}^n,\ k \in \mathbb{Z}_+. \tag{13}$$

This shows that the links through which information of other pages is not transmitted are probabilistically selected under a fixed probability. Such failure models are employed in the context of networked control and consensus [7, 8, 11, 12, 20].

To take account of failing links, consider the distributed update scheme given by

$$x(k+1) = (1-\hat{m})A_{\eta(k),\varDelta(k)}x(k) + \frac{\hat{m}}{n}\mathbf{1}, \tag{14}$$

where $x(k) \in \mathbb{R}^n$, the initial condition $x(0) \geq 0$ satisfies $\sum_{i=1}^n x_i(0) = 1$, and $\hat{m} \in (0,1)$ is the parameter used instead of m in the centralized case. The matrices $A_{q,\mathscr{D}}$ for $q \in \{0,1\}^n$ and $\mathscr{D} \subset \mathscr{E}_q$ are the distributed link matrices with link failures.

The objective here is to design this distributed update scheme by finding the appropriate link matrices $A_{q,\mathscr{D}}$ and the parameter $\hat{m}$ so that the PageRank values are computed through the time average y of the state x. We follow an approach similar to that in Section 3 and, in particular, construct the link matrices so that they possess the stochastic property.

Distributed link matrices and their average: The first step in the design is to introduce the distributed link matrices and analyze their properties.

Let the distributed link matrix with link failures be given as follows:

$$\big(A_{q,\mathscr{D}}\big)_{ij} := \begin{cases} 0 & \text{if } (j,i) \in \mathscr{D}, \\ (A_q)_{ij} & \text{if } (j,i) \notin \mathscr{D} \text{ and } i \neq j, \\ 1 - \sum_{\substack{h \in \mathscr{V},\ h \neq j \\ (j,h) \notin \mathscr{D}}} (A_q)_{hj} & \text{if } i = j \end{cases} \tag{15}$$

for $q \in \{0,1\}^n$, $\mathscr{D} \subset \mathscr{E}_q$, and $i,j \in \mathscr{V}$. Note that by definition, $(i,i) \notin \mathscr{E}_q$, $\forall i,q$.

By the definition of link failures, if the link $(j,i) \in \mathscr{E}$ is failing, then the (i,j) entry of the link matrix must be equal to zero. The link matrices defined above take account of such zero entries, but are still designed to be stochastic. This property is critical in showing the convergence of the scheme. In practice, this structure implies that if page j initiates an update and sends its value to page h over a link that is potentially failing, it must know whether page h received the value (and used it for its own update) or not. This can be observed in the (j,j) entry in (15) since it consists of the (h,j) entry of A_q.

We now analyze the average dynamics of the distributed update scheme determined by the link matrices just introduced. We define the average link matrix by

$$\overline{A} := E\left[A_{\eta(k),\Delta(k)}\right], \tag{16}$$

where $E[\cdot]$ is the expectation with respect to the processes $\eta(k)$ and $\Delta(k)$. This matrix $\overline{A}$ is nonnegative and stochastic because all $A_{\eta(k),\Delta(k)}$ share this property.

The following proposition shows that the average link matrix $\overline{A}$ has a clear relation to the original link matrix A.

Proposition 1. (i) The average link matrix $\overline{A}$ given in (16) can be expressed as

$$\overline{A} = \left[1-\delta-(1-\delta)(1-\alpha)^2\right]A + \left[\delta+(1-\delta)(1-\alpha)^2\right]I. \tag{17}$$

(ii) There exists a vector $z_0 \in \mathbb{R}+^n$ which is an eigenvector corresponding to the eigenvalue 1 for both matrices A and $\overline{A}$.

Mean-square convergence of the distributed scheme: In order to show the convergence property of the distributed update scheme, we now introduce the modified version of the link matrices. First, we rewrite the update scheme of (14) in its equivalent form as

$$x(k+1) = M_{\eta(k),\Delta(k)}x(k), \tag{18}$$

where the matrices $M_{q,\mathscr{D}}$ for $q \in \{0,1\}^n$ and $\mathscr{D} \subset \mathscr{E}_q$ are given by

$$M_{q,\mathscr{D}} := (1-\hat{m})A_{q,\mathscr{D}} + \frac{\hat{m}}{n}S. \tag{19}$$

These matrices are called the modified distributed link matrices. This equivalent form of (18) can be obtained because the link matrices A_q are stochastic matrices; thus, the state $x(k)$ remains a probability vector for all k, implying $Sx(k) \equiv \mathbf{1}$.

Also, let the average matrix of $M_{\eta(k),\Delta(k)}$ be

$$\overline{M} := E[M_{\eta(k),\Delta(k)}]. \tag{20}$$

Here, the distributed link matrices are positive stochastic matrices, which means that the average matrix $\overline{M}$ enjoys the same property.

The next step in designing the update scheme is to determine the parameter $\hat{m}$. The specific aim here is to show that the average of the modified distributed link matrices and the link matrix M from (3) share an eigenvector corresponding to the eigenvalue 1. Since such an eigenvector is unique for M, it is necessarily equal to the value vector x^*.

Similarly to the case in Section 3, the parameter $\hat{m}$ is chosen differently from m in the centralized scheme. Let $\hat{m}$ be given by

$$\hat{m} = \frac{\left[1-\delta-(1-\delta)(1-\alpha)^2\right]m}{1-m\left[\delta+(1-\delta)(1-\alpha)^2\right]}. \tag{21}$$

The next lemma states an important property of the link matrices for this $\hat{m}$.

Lemma 2. The parameter $\hat{m}$ in (21) and the average link matrices $\overline{M}$ in (20) have the following properties:

(i) $\hat{m} \in (0,1)$ and $\hat{m} \leq m$.
(ii) $\overline{M} = \frac{\hat{m}}{m} M + \left(1 - \frac{\hat{m}}{m}\right) I$.
(iii) For the average matrix $\overline{M}$, the eigenvalue 1 is simple and is the unique eigenvalue of maximum modulus. The value vector x^* is the corresponding eigenvector.

We can show by (iii) in the lemma that, in an average sense, the distributed update scheme obtains the correct values, i.e., $E[x(k)] = \overline{M}^k x(0) \to x^*$ as $k \to \infty$.

We are now ready to state the main result of this paper.

Theorem 2. Consider the distributed scheme with link failures in (14). For any update probability $\alpha \in (0,1]$ and link failure probability $\delta \in [0,1)$, the PageRank value x^* is obtained through the time average y in (10) as $E\left[\left\|y(k) - x^*\right\|^2\right] \to 0$, $k \to \infty$.

The proof follows along similar lines as that in [13]. Specifically, one way to establish the convergence is by the general Markov chain results of, e.g., [5]. Another approach is to employ the proof developed in the paper [13] by adapting it to the current update scheme. This proof is found to be useful to study the rate of convergence and to include an update termination feature. Under this feature, each page is allowed to stop its update when an approximate value is obtained; this is important because computation as well as communication loads can be reduced.

We have remarks on the asymptotic rate of convergence for the average state $E[x(k)]$. Similarly to the discussion in Section 2, the convergence rate is dominated by the second largest eigenvalue $\lambda_2(\overline{M})$ in magnitude. By $|\lambda_2(M)| \leq 1 - m$ as we have seen in Section 2 and (ii) in Lemma 2, this eigenvalue can be bounded as

$$|\lambda_2(\overline{M})| = \frac{\hat{m}}{m} |\lambda_2(M)| + 1 - \frac{\hat{m}}{m} \leq \frac{1 - m}{1 - m\big[\delta + (1 - \delta)(1 - \alpha)^2\big]}.$$

It is clear that this bound is a decreasing function of α and an increasing function of δ. That is, higher probability α in updates and/or smaller δ results in faster average convergence. Faster convergence is, nevertheless, realized by additional computation and communication, which are affected by both α and δ.

5 Numerical Example

In this section, we present a numerical example to verify the efficacy of the results. We generated a web with 1,000 pages ($n = 1{,}000$), where the links among the pages were randomly determined. The first ten pages are designed to have high PageRank values and are linked from over 90% of the pages. For other pages, the numbers of links are between 2 and 333. The parameter m was taken as $m = 0.15$.

Simulations were carried out using three algorithms: The first one is the original distributed scheme in Section 3, which is run without any link failures and is

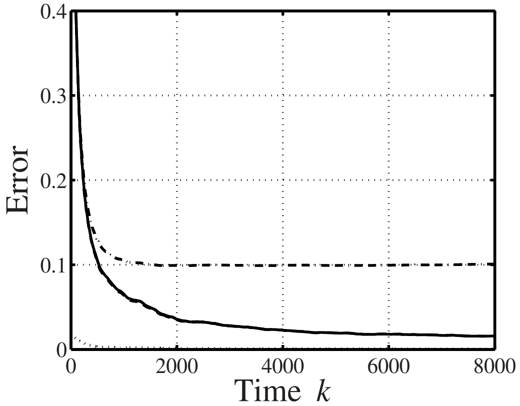


Fig. 1 The error $e(k)$ in the PageRank: $||e(k)||_1$ for the original scheme without link failures (dashed line), the original scheme (dash-dot line), and for the proposed scheme (solid line); $||e(k)||_\infty$ for the proposed scheme (dotted line).

for reference. The second one is the same original scheme, but failing links are present, that is, when some values from linked pages are not available, they are considered to be zero. The last one is the proposed scheme with link failures in Section 4. For all three cases, the probability of update for the pages was taken as $\alpha = 0.01$.

We executed the algorithms with the link failure probability $\delta = 0.02$. Sample paths of the state x were computed from time 0 to 8,000. The initial state $x(0)$ was taken the same for all algorithms and was randomly chosen as a probability vector. We computed the error $e(k) := y(k) - x^*$ in the PageRank value estimate. In Fig. 1, we show the ℓ_1 norm of $e(k)$. The original scheme without link failures (dashed line) and the proposed scheme (solid line) have comparable performance, and the difference is not visible in the plot. In contrast, for the original scheme with failing links, the error stops decreasing and stays at a relatively high level of 0.1. This is interesting because the probability δ of link failures is quite small, but has a significant effect. One reason is that the original link matrices are in effect no longer stochastic. As a result, the final value $y(k)$ at $k = 8,000$ for this scheme is not a probability vector and, in fact, we obtained $\sum_i y_i(k) = 0.900$. We also plotted in Fig. 1 the ℓ_∞ norm of the error $e(k)$ for the proposed case. This corresponds to the maximum individual error. We observe that it rapidly decreases.

6 Conclusion

In this paper, we studied extensions of the distributed randomized approach for the PageRank computation proposed in [13]. We considered the effect of link failures under which not all the links are used for communication in the update. This scheme is, in particular, useful to model failures in the network as well as to reduce the communication/computation load for the pages. In future research, we will address issues related to aggregation of webpages for PageRank computation [16].

References

1. Avrachenkov, K., Litvak, N., Nemirovsky, D., Osipova, N.: Monte Carlo methods in PageRank computation: when one iteration is sufficient. SIAM J. Numer. Anal. 45, 890–904 (2007)
2. Bertsekas, D.P., Tsitsiklis, J.N.: Parallel and Distributed Computation: Numerical Methods. Prentice-Hall, Englewood Cliffs (1989)
3. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks & ISDN Systems 30, 107–117 (1998)
4. Bryan, K., Leise, T.: The $25,000,000,000 eigenvector: the linear algebra behind Google. SIAM Rev. 48, 569–581 (2006)
5. Cogburn, R.: On products of random stochastic matrices. Contemp. Math. 50, 199–213 (1986)
6. de Jager, D.V., Bradley, J.T.: Asynchronous iterative solution for state-based performance metrics. In: Proc. ACM SIGMETRICS, pp. 373–374 (2007)
7. Elia, N.: Remote stabilization over fading channels. Systems & Control Letters 54, 238–249 (2005)
8. Fagnani, F., Zampieri, S.: Average consensus with packet drop communication. SIAM J. Control and Optim. 48, 102–133 (2009)
9. Hatano, Y., Mesbahi, M.: Agreement over random networks. IEEE Trans. Autom. Control 50, 1867–1872 (2005)
10. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge Univ. Press, Cambridge (1985)
11. Imer, O.Ç., Yüksel, S., Başar, T.: Optimal control of LTI systems over unreliable communication links. Automatica 42, 1429–1439 (2006)
12. Ishii, H., Hara, S.: A subband coding approach to control under limited data rates and message losses. Automatica 44, 1141–1148 (2008)
13. Ishii, H., Tempo, R.: A distributed randomized approach for the PageRank computation, Parts 1 and 2. In: Proc. 47th IEEE Conf. on Decision and Control, pp. 3523–3534 (2008). Also, to appear in Trans. Autom. Control (2010)
14. Ishii, H., Tempo, R.: Computing the PageRank variation for fragile web data. SICE J. Control, Measurement, and System Integration 2, 1–9 (2009)
15. Ishii, H., Tempo, R.: Distributed PageRank computation with link failures. In: Proc. 2009 American Control Conference, pp. 1976–1981 (2009)
16. Ishii, H., Tempo, R., Bai, E.W., Dabbene, F.: Distributed randomized PageRank computation based on web aggregation. In: Proc. 48th IEEE Conf. on Decision and Control, pp. 3026–3031 (2009)
17. Kamvar, S., Haveliwala, T., Golub, G.: Adaptive methods for the computation of PageRank. Linear Algebra and its Appl. 386, 51–65 (2004)
18. Kollias, G., Gallopoulos, E., Szyld, D.B.: Asynchronous iterative computations with Web information retrieval structures: the PageRank case. In: Joubert, G.R., et al. (eds.) Parallel Computing: Current and Future Issues of High-End Computing. NIC Series, vol. 33, pp. 309–316. John von Neumann-Institut for Computing, Julich (2006)
19. Langville, A.N., Meyer, C.D.: Google's PageRank and Beyond: The Science of Search Engine Rankings. Princeton Univ. Press, Princeton (2006)
20. Sinopoli, B., Schenato, L., Franceschetti, M., Poolla, K., Jordan, M.I., Sastry, S.S.: Kalman filtering with intermittent observations. IEEE Trans. Autom. Control 49, 1453–1464 (2004)
21. Tahbaz-Salehi, A., Jadbabaie, A.: A necessary and sufficient condition for consensus over random networks. IEEE Trans. Autom. Control 53, 791–795 (2008)

22. Tempo, R., Calafiore, G., Dabbene, F.: Randomized Algorithms for Analysis and Control of Uncertain Systems. Springer, London (2005)
23. Tempo, R., Ishii, H.: Monte Carlo and Las Vegas randomized algorithms for systems and control: an introduction. European J. Control 13, 189–203 (2007)
24. Wu, C.W.: Synchronization and convergence of linear dynamics in random directed networks. IEEE Trans. Autom. Control 51, 1207–1210 (2006)
25. Zhu, Y., Ye, S., Li, X.: Distributed PageRank computation based on iterative aggregation-disaggregation methods. In: Proc. 14th ACM Conf. on Information and Knowledge Management, pp. 578–585 (2005)

Predicting Synchrony in a Simple Neuronal Network

Sachin S. Talathi and Pramod P. Khargonekar

1 Introduction

Human brain is comprised of hierarchically structured networks of neurons with feedforward and feedback connections across the hierarchy [17]. It has a remarkable ability to process sensory information and generate motor actions at millisecond time scales [20]. In recent years, new theories have emerged that view the brain as an active and adaptive system in which there is a close connection between cognition and action [5]. Instead of viewing cognition as building universal, context independent models of the external world [4, 14], cognition is considered to play an important role in the generation of correct action responses in a context dependent adaptive manner [19]. Accordingly, the modern perspective is to relate cognitive functions with coherent behavior of large number of neural populations [3, 19, 35]. This modern view has been particularly relevant for understanding the binding problem which deals with the question of how does the brain integrate sensory information on object properties (color, shape, ...) to identify the object as a coherent whole [27]. Since many objects in the world are multi-sensory, a coherent representation of the object requires integration of responses across different sensory modalities including in various combinations the haptic, visual, olfactory and auditory properties. It is believed that neural synchrony at millisecond level precision is crucial in implementing such an integration across different cortical regions. Although there has been a growing interest in understanding the role for neural synchrony in cognitive processes involved in normal brain functioning, a number of recent works have also examined the relevance of neural synchronization in various neurological diseases such as epilepsy [21], schizophrenia [24], autism [26] and Alzheimers disease [30].

Sachin S. Talathi
Department of Biomedical Engineering, University of Florida
e-mail: stalathi@bme.ufl.edu

Pramod P. Khargonekar
Department of Electrical and Computer Engineering, University of Florida
e-mail: ppk@ufl.edu

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 151–162.
springerlink.com

Neural synchrony generally refers to the fact that large numbers of heterogeneous interconnected neurons fire in a precisely coordinated manner to generate very distinct oscillations in different frequency bands. A number of theoritcal studies have employed computational models of neurons and their interactions to in order to understand the mechanisms underlying the generation of such synchronized oscillations [1, 9, 12, 34]. While for large neuronal networks, detailed computational simulations is the main approach [2, 36, 37, 39], it is possible to conduct analytical investigations for small networks with the hope that these analytical tools will shed light on synchronized behavior in large neuronal networks [16, 29, 31, 38].

In recent years, there has been much work in the control theory community on consensus in networks of dynamical systems. In particular, the survey paper by Paley et al. [23] contains a large list of references on oscillator models and collective action. Indeed, neural synchrony is a form of consensus and collective behavior. The possibility of controlling pathological brain synchrony in neurological diseases such as Parkinson's disease and epilepsy through electrical stimulation, has also spurred interest in the control community to device novel control algorithms, that are based on intrinsic neural populations dynamics [6]. As such, we expect that in the future there will be new research directions at the interface of research efforts by computational neuroscientists and control theorists.

In this paper, our goal is to demonstrate how weak coupling theory and spike time response curves can be used to analyzed patterns of synchrony in a small network of interacting neurons. We present our analysis of phase locked synchronous states emerging in a simple unidirectionally coupled interneuron network (UCIN) comprising of two heterogeneously firing neuron models coupled through a biologically realistic inhibitory synapse. The paper is divided into following sections. Section 2 provides the mathematical background on the neuron models used in the present study and the weak coupling theory of interacting pulse coupled oscillators. In section 3, we analyze patterns of synchrony in the UCIN using weak coupling theory and nonlinear interaction map derived using spike time response curves. Section 4 presents some concluding remarks.

We are very pleased to dedicate this paper to Professor Yutaka Yamamoto on his 60th birthday. One of us (PPK) has had the good fortune and privilege of being friends with him for nearly 30 years, and also collaborated with him in research on sampled-data control and filtering problems. The present paper is loosely connected to signal processing research in that the brain is one of the most remarkable signal processing and understanding machines in existence.

2 Background

In this section, we will briefly describe some background material from computational neuroscience which is relevant for analysis of synchrony in neuronal networks.

2.1 Mathematical Model of Neuronal Dynamics

We model neuronal dynamics following the universally accepted Hodgkin-Huxley formalism [13] (conductance based neuron models [28]). The basic neuron model satisfies the following current-balance equation for the flow of current through the neuronal membrane.

$$C_M \frac{dv(t)}{dt} + I_{na}(t, v(t)) + I_k(t, v(t)) + I_L(v(t)) + I_{dc} = 0 \tag{1}$$

where, t is typically measured in ms, $v(t)$ is the neuronal membrane potential in mV, and $C_M \frac{dv(t)}{dt}$ is the capacitive component of the membrane current, with C_M being the membrane capacitance in μF/cm^2. The current through voltage gated sodium channel is $I_{na}(t, v(t)) = g_{na} m^3(t) h(t)(v(t) - E_{na})$ and the current through the voltage gated potassium channel is $I_k(t, v(t)) = g_k n^4(t)(v(t) - E_k)$. The leak current resulting from passive flow of all other ions through the membrane is modeled through $I_l = g_l(v(t) - E_l)$. Here g_C and E_C (C≡na,k,l), represent the maximal conductance in mS/cm^2 and the reversal potential for ion channels in mV respectively. The intrinsic firing frequency of each neuron is dependent on the constant current I_{dc} input to each neuron. The variables $X(t) \equiv \{m(t), h(t), n(t)\}$ which represent the fraction of open ion channels, satisfy the following first order kinetic equation

$$\frac{dX(t)}{dt} = \phi(\alpha_X(v(t))(1 - X(t)) - \beta_X(v(t))X(t)) \tag{2}$$

The model parameters are set to those obtained by Wang and Buzsaki (WB) [37] to simulate the dynamics of a fast spiking cortical interneuron. Specifically the functional form for $\alpha_X(V)$ and $\beta_X(V)$ (X≡m,n,h) are provided in table 1 and the model parameters are (g_{na},g_k,g_l)=(35,9,0.1) mS/cm^2; (E_{na},E_k,E_l)=(50,−90,−65) mV and $\phi = 5$.

Table 1 Functions α_X and β_X for the WB neuron model

X	$\alpha_X(V)$	$\beta_X(V)$
m	$\frac{0.1(V+35)}{1-e^{-0.1(V+35)}}$	$4e^{-(V+60)/18}$
h	$0.07e^{-(V+58)/20}$	$\frac{1}{1+e^{-0.1(V+28)}}$
n	$\frac{0.01(V+34)}{1-e^{-0.1(V+34)}}$	$0.125e^{-(V+44)/88}$

2.2 Phase for Limit Cycle Oscillators

Following the well known approach due to Winfree [40], Guckenheimer [11] and Ermentrout [8, 9], we will simplify the analysis of neuronal dynamical networks by using transformation to phase variables. In general, the neuron model described in equations 1 and 2 can be written as

$$\frac{dx}{dt} = f(x,\alpha) \tag{3}$$

We will assume that this system has a normally hyperbolic attracting limit cycle $x_0(t)$ with period T_0 which is a function of parameter α such that $x_0(t+T_0) = x_0(t)$. Equation 3 can then be simplified by defining a scalar phase variable $\phi(x_0) \in [0$ $1)$ such that the phase evolution has a simple form $d\phi/dt = 1/T_0$. Thus, with each point on the limit cycle, there is a unique associated phase.

Now consider a point x_* in the basin of attraction of the limit cycle $x_0(t)$. It is then clear that there is a unique phase $\phi_* \in [0\ 1)$ such that the trajectories of dynamical system defined in 3 starting with initial conditions x_* and $x_0(\phi_* T_0)$ converge asymptotically. We define phase of the point x_* to be ϕ_*. The set of points x_* in the basin of attraction with a given phase ϕ_* define an *isochrone* [40]. With the notion of phase defined in the vicinity of the limit cycle through isochrons, the nonlinear system (3) then induces a differential equation for phase in the basin of attraction:

$$\frac{d\phi}{dt} = g(x(t)) \tag{4}$$

It is important to observe that $g(x) = 1/T_0$ if $x \equiv x_0$.

2.3 Weakly Coupled Oscillators

In order to analyze interactions among neurons and the effect of external stimulus, let us now introduce a small periodic force $\varepsilon p(x,t) = \varepsilon p(x,t+P)$ with period P (ε measure the strength of the forcing term) which is in general different from T_0:

$$\dot{x} = f(x,\ \alpha) + \varepsilon p(x,t) \tag{5}$$

Using the notion of isochrons defined above, the phase dynamics for equation 5 in the neighborhood of the unperturbed system $x_0(t)$ can now be written as

$$\frac{d\phi(x)}{dt} = \omega_0 + \varepsilon \bigtriangledown_{\mathbf{x}} \phi . p(x,t) \tag{6}$$

For weak coupling $\varepsilon << 1$, the deviation of x from the limit cycle x_0 is negligible, and in the first order approximation we can evaluate the rhs of eq 6 on the limit cycle:

$$\frac{d\phi(x)}{dt} \approx \omega_0 + \varepsilon \bigtriangledown_{x} \phi . p(x_0,t) \tag{7}$$

On the limit cycle, there is one-one correspondence between the state variable x and the phase ϕ. We therefore have a closed equation for phase:

$$\frac{d\phi}{dt} = \omega_0 + \varepsilon H(\phi,t) \tag{8}$$

where $H(\phi,t) = Z(\phi).p(x_0(\phi),t)$ is unit period function of ϕ and P period function of t referred to as the "averaged" interaction function [8]. The function $Z(\phi) := \bigtriangledown_x \phi$ is purely a function of the oscillator limit cycle and captures the effect of perturbation on the phases. It is commonly referred to as the infinitesimal phase response curve (iPRC) or the linear response function [15]. It can be shown that $Z(\phi)$ is the adjoint eigenfunction for the linearization of the differential equation given in equation 3, about the stable limit cycle $x_0(t)$ [8, 9], which naturally turns out to be a linear periodic system. Recently, a computationally efficient algorithm using properties of this linear periodic system has been proposed in [10].

A special case of the above setup arises when the periodic perturbation εp is the output of another neuron. In this case, $H(\phi,t) = H(\phi,\phi^{'}) = Z(\phi).p(\phi,\phi^{'})$, where $\phi^{'}$ represents the phase variable for the driver neuron. In the case of weak coupling, to the extent that the change in phase ϕ, $d\phi/dt << \omega_0$ over one cycle of unperturbed oscillator, the effective perturbation can be approximated by averaged perturbation over one cycle of the unperturbed oscillator [7],

$$H(\phi,\phi^{'}) = \int_0^1 d\theta Z(\phi+\theta).p(\phi+\theta,\phi^{'}+\theta) \tag{9}$$

In case the perturbation is an independent function of the driver and the driven oscillators, i.e., $p(\phi,\phi^{'}) \equiv p(\phi^{'}).q(\phi)$, equation 9 can be written as a correlation integral

$$H(\phi^{'}-\phi) = \int_0^1 d\theta Z(\theta-(\phi^{'}-\phi)).p(\theta).q(\theta-(\phi^{'}-\phi)) \tag{10}$$

and the phase dynamics of the perturbed oscillator is given by

$$\frac{d\phi}{dt} = \omega_0 + \varepsilon H(\phi^{'}-\phi) \tag{11}$$

3 Analysis of Phase Locked States in a Coupled Neuronal Network

As mentioned in the Introduction, analysis of synchrony in networks of interacting heterogeneous neurons is important for understanding information processing in the brain. Here we present a relatively simple example of two synaptically coupled WB neurons (see Figure 1a inset). We analyze phase locked states for this special network.

3.1 Description of the Network

As shown in Figure 1, WB neuron labelled A fires at intrinsic frequency $\omega_A(I_{dc}^A)$ and receives periodic synaptic perturbation from a WB neuron B, which fires with intrinsic frequency $\omega_B(I_{dc}^B)$ with $I_{dc}^B \neq I_{dc}^A$. The synaptic coupling is modeled as: $I_s = gs(v_B(t),t)(E_R - v_A(t))$, where g is the strength of synaptic coupling,

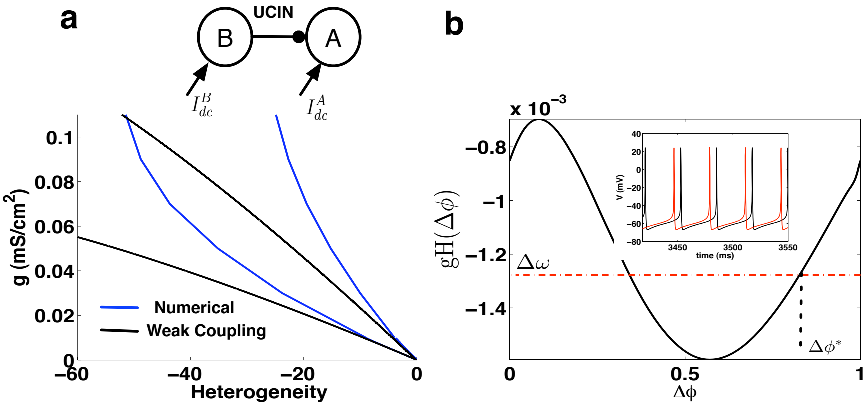


Fig. 1 (a) Arnold tongue for the phase locked states of the unidirectionally coupled interneuron network (UCIN). The region bounded by blue curves represent the Arnold tongue for the UCIN generated through numerical simulation of equations 1,2 and 16 for the two coupled neurons and the synapse. The region bounded by curve in black represents the steady state solution to equation 18 resulting from weak coupling approximation. Inset shows the schematic diagram of the UCIN. (b) Description of procedure to determine steady state fixed point solution to equation 14, for the specific case of heterogeneity H=-4 % resulting from the choice of $I_{dc}^A = 0.5\ \mu$A/cm^2 producing $\omega_A \approx 32.2$ Hz and $I_{dc}^A = 0.48\ \mu$A/cm^2 producing $\omega_B = 30.9$ Hz. The synaptic coupling strength is set at $g = 0.0052$ mS/cm^2. The fixed point $\Delta\phi^*$ corresponds to the stable state solution of equation 14 satisfying conditions in equations 15 and 16 respectively. Inset shows the time series of membrane potential (neuron A in black and neuron B in red) for the particular case considered, when the weak coupling approximation is able to predict the existence of stable phase locked state $\Delta\phi^*$.

v_X $\{$X$\equiv$A,B$\}$ is the membrane potential, $E_R = -75$ mV is the reversal potential of the synapse and $s(v,t)$ represents the fraction of neurotransmitters bound to the membrane of the post-synaptic cell (neuron A) resulting from the release of these neurotransmitters by neuron B at any given time. It satisfies the following ordinary differential equation

$$\frac{ds(t)}{dt} = \frac{s_0(v_B(t)) - s(t)}{\tau(s_1 - s_0(v_B(t)))} \tag{12}$$

where $s_0(v) = 0.5(1 + \tanh(100(v - 0.1)))$ and the parameters τ and s_1 are set such that the synaptic rise time $\tau_R = \tau(s_1 - 1) = 0.1$ ms and the synaptic decay time $\tau_D = \tau s_1 = 8$ ms.

3.2 Weak Coupling Approach

If ϕ_A and ϕ_B represent the phase variables of A and B respectively, then we have in the weak coupling limit

$$\begin{aligned}\frac{d\phi_A}{dt} &= \omega_A + gH(\Delta\phi)\\ \frac{d\phi_B}{dt} &= \omega_B \end{aligned} \tag{13}$$

where $H(\Delta\phi) = \int_0^1 Z(\theta - \Delta\phi)s(\theta).(E_R - v(\theta - \Delta\phi))d\theta$ and $\Delta\phi = \phi_B - \phi_A$. The ordinary differential equation for the phase difference $\Delta\phi$ is:

$$\frac{d(\Delta\phi)}{dt} = \Delta\omega - gH(\Delta\phi) \tag{14}$$

where $\Delta\omega = \omega_B - \omega_A$ is the difference in the intrinsic firing rates of the two coupled neurons. *Stable fixed point solution $\Delta\phi^*$ of equation 14 corresponds to the phase locked state of synchronous oscillations between the two coupled neurons in the UCIN.* The fixed point solution satisfies

$$H(\Delta\phi^*) = \Delta\omega/g \tag{15}$$

and the local stability of $\Delta\phi^*$ is guaranteed provided

$$\frac{dH(\Delta\phi)}{d\Delta\phi}|_{\Delta\phi^*} > 0 \tag{16}$$

Two key parameters of the UCIN that influence phase locking behavior are the heterogeneity $\Delta\omega$ and the synaptic coupling strength g. We will now use weak coupling theory to estimate the set of $\{\Delta\omega, g\}$ which corresponds to phase locked synchronous states of UCIN and compare these to the set $\{\Delta\omega, g\}$ which result in phase locked solutions of UCIN using full nonlinear model as described through equations 1, 2 and 12.

Weak coupling theory estimate of $\{\Delta\omega, g\}$ can be obtained by solving equations 15 and 16. Detailed explanation of the this computation is provided in Figure 1b. The resulting domain of $\{\Delta\omega, g\}$ is referred to as the Arnold Tongue [25], which is depicted as the region bounded by black curves in Figure 1a. Arnold tongue for full UCIN is obtained by fixing the firing frequency of neuron A to $\omega_A \approx 32Hz$ and varying the intrinsic firing frequency of neuron B ω_B by changing the dc drive I_{dc}^B on to the neuron B, thereby varying the degree of heterogeneity $H = 100\frac{\omega_B - \omega_A}{\omega_A}$ in the intrinsic firing rates of the two coupled neurons. The phase locked states correspond to the value of synaptic strength g that result in $< \omega_A > / \omega_B = 1$, where $< \omega_A >$ is the frequency of neuron A when the UCIN settles into steady state. The Arnold tongue so obtained is the region bounded by curves in blue in Figure 1a. We see from Figure 1a that the weak coupling theory based estimate of the Arnold tongue matches that generated through numerical simulations only in the vicinity of $\{0,0\}$. However, there is a significant mismatch for higher values of synaptic strength and heterogeneity in the network.

3.3 Analysis of the Strong Coupling Case

In this section, we will introduce the concept of spike time response curves (STRC's), and demonstrate its utility in the analysis of phase locked states in the UCIN in the regions where weak coupling theory fails. In order to motivate the concept of STRC, consider a spontaneously firing neuron with period T_0. At time t following a voltage peak in the firing cycle of a neuron a perturbation, e.g., a depolarizing current pulse is applied. It shifts the time of the next voltage peak as in Figure 2. Let T_j (j=1,2,...) represent the times of j^{th}voltage peak after the perturbation. The quantities $\Phi_{j,\alpha} = \frac{T_0 - T_j}{T_0}$, which measure the shift in the phase of neuron in response to a perturbing stimulus are called the STRC's. The parameter α corresponds to the dependence of STRC on the characteristics of the perturbation input. If the perturbation impulse occurs through a chemical synapse, a case of particular importance to the analysis of the UCIN we consider here, α represents the set of synaptic parameters such as the rise time of the synapse τ_R, the decay time of the synapse τ_D, the synaptic reversal potential E_R and the synaptic coupling strength g. The STRC's can be computed numerically by solving the nonlinear dynamical equations for a given neuron receiving the perturbation and measuring the length of subsequent firing cycles [1]. For the network under consideration, we specifically computed STRC's for an intrinsically firing WB neuron with frequency $\omega \approx 32$ Hz receiving perturbation through an inhibitory synapse with reversal potential $E_R = -75$ mV, $g = 0.15$ mS/cm^2 $\tau_D = 8$ ms, $\tau_R = 0.1$ ms. In Figure 3a, we show the STRC's $\Phi_j(\delta t)$ (j=1,2) as a function of the time δt at which it receives the perturbation.

We will now use these STRC's obtained numerically for different levels of synaptic coupling strengths g to derive a nonlinear map for the evolution of phase difference $\Delta\phi$ between the two coupled neurons in the UCIN. In Figure 3b, we show a schematic diagram of spike times of the two neurons when they are phase locked. Let t_X^n $\{$X$\equiv$A,B$\}$ be the time of n^{th} spike generated from neurons A and B respectively. Define δ^n to be the time in the n^{th} firing cycle of neuron A when it receives synaptic perturbation from neuron B. If P_A^n represents the length of n^{th} firing cycle of neuron A, then from Figure 3b, we have $P_A^n = t_A^{n+1} - t_A^n = F_A^n + R_A^n$, where $F_A^n = \delta^n + T_A\Phi_2(\delta^{n-1})$ is the entrained firing interval defining the time elapsed between the firing of neuron A at time $t_A^n$and the firing of pre-synaptic neuron B at time t_B^n [22]. In writing this equation, we assume that the oscillator returns back on to the limit cycle in between periodic perturbations. Analysis of phase locked state

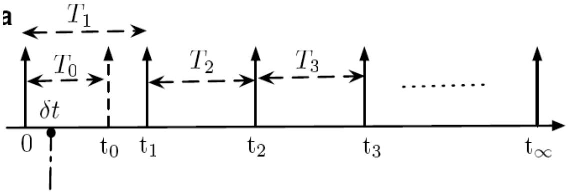


Fig. 2 Schematic diagram representing the effects of external perturbation on the subsequent firing periods of a neuron firing with intrinsic period T_0

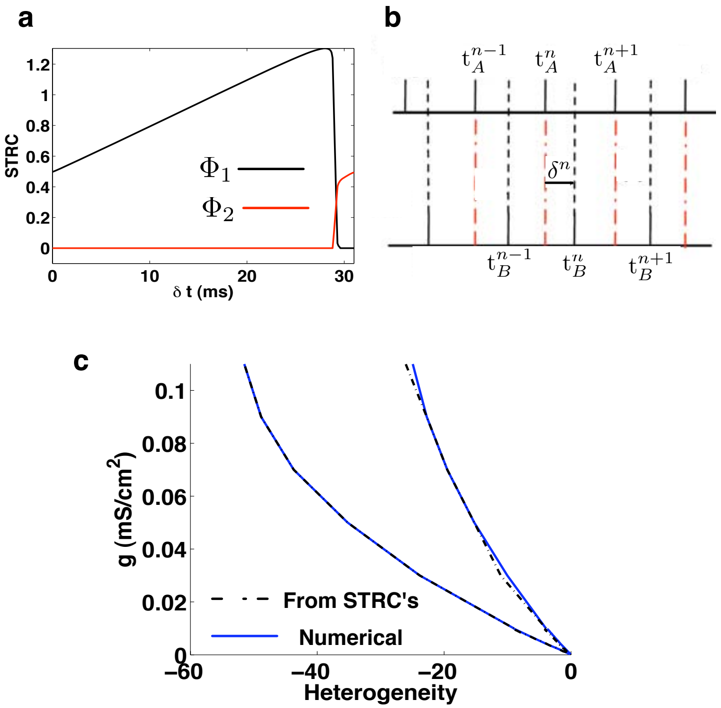


Fig. 3 (a) The STRC's Φ_1 (black line) and Φ_2 (red line) for a WB neuron receiving perturbation through an inhibitory synapse with parameters $E_R = -75$mV, $\tau_D = 8$ ms, and $\tau_R = 0.1$ ms. All higher order STRC's Φ_j (j>2) are zero. (b) Schematic diagram representing spike timing for neurons A and B in the unidirectionally coupled interneuron network (UCIN) when they are phase locked. (c) Arnold tongue for the phase locked states of the UCIN represented by region bounded by blue lines obtained through numerical simulation of equations 1, 2 and 12. The curves in black are obtained as a steady state solution to the nonlinear map in equation 18.

between oscillators when this condition is not met has been recently performed by [32, 33] and is beyond the scope of this article.The entrained recovery interval defining the time interval between the firing of the pre-synaptic neuron at time t_B^n and the next firing of neuron A at time t_A^{n+1} is then given as $R_A^n = T_A(1+\Phi_1(\delta^n)) - \delta^n$, which follows from the definition of Φ_1. From Figure 3b, we have

$$T_B = t_B^{n+1} - t_B^n = R_A^n + F_A^{n+1} \tag{17}$$

resulting in the following nonlinear map for the evolution of δ^n

$$\delta^{n+1} = \delta^n + T_B - T_A(1+\Phi_1(\delta^n) + \Phi_2(\delta^n)) \tag{18}$$

The stable fixed point solution of the nonlinear map given through equation 18 represent the phase locked state of the UCIN. The fixed point δ^* of equation 18 satisfies

$1+\Phi_\infty(\delta^*) = T_B/T_A$, where $\Phi_\infty(x) = \Phi_1(x)+\Phi_2(x)$. The local stability of δ^* requires $0 < \frac{d\Phi_\infty(x)}{dx}|_{x=\delta^*} < 2$. In Figure 3c, the curves shown in black enclose the region of stable fixed point solution to equation 18. We see that the nonlinear map derived from STRC's is successfully able to predict the Arnold tongue corresponding to the phase locked solution for the UCIN even in the strong coupling limit.

4 Discussion

Our primary aim is to develop a research program at the intersection of control, signal processing and computational neuroscience. Here we focused on analysis of synchrony in a simple network of two heterogeneous neurons interacting through a strong inhibitory synapse, the UCIN. Weak coupling theory is general and has proved effective in the analysis of large homogeneous neuronal networks interacting through weak coupling. [18]. It has limitations for the analysis of realistic biological networks [2]. Our analysis of the UCIN specifically demonstrates the limited applicability of weak coupling theory in predicting synchronous phase locked states of the network. We show that nonlinear maps derived from STRC's can better predict synchronous phase locked states generated by the network in more biologically realistic conditions of moderate to high heterogeneity and strong synaptic interactions.

Understanding the dynamics of large heterogeneous neuronal networks is a major area of research, see e.g., [2]. Analysis of such large networks is primarily done using numerical simulations. Mathematical analysis of small networks (UCIN, for example) has been shown to be fruitful in providing insights into the dynamics of large neuronal networks. In particular, White et al. [38], used a simple two cell network to shed light on two distinct mechanisms which synchrony in large networks can be lost. Similarly, Skinner et al. [29], have shown that coherent states observed in two cell networks in the presence of heterogeneity are preserved in large neuronal networks suggesting that a strategy of analysis on small network dynamics might be a useful way to understand the contribution of biophysical parameters in the generation of synchronous states in large biological networks. Finally, it is hoped that interdisciplinary efforts will lead to new mathematical analysis techniques that will apply to large heterogeneous neuronal networks.

References

1. Acker, C.D., Kopell, N., White, J.A.: Synchronization of strongly coupled excitatory neurons: relating network behavior to biophysics. J. Comput. Neurosci. 15(1), 71–90 (2003)
2. Bartos, M., Vida, I., Jonas, P.: Synaptic mechanisms of synchronized gamma oscillations in inhibitory interneuron networks. Nat. Rev. Neurosci. 8(1), 45–56 (2007)
3. Beer, R.D.: Dynamical approaches to cognitive science. Trends Cogn. Sci. 4(3), 91–99 (2000)
4. Biederman, I.: Recognition-by-components: a theory of human image understanding. Psychol. Rev. 94(2), 115–147 (1987)

5. Clark, A.: An embodied cognitive science? Trends Cogn. Sci. 3(9), 345–351 (1999)
6. Danzl, P., Moehlis, J.: Spike timing control of oscillatory neuron models using impulsive and quasi-impulsive charge balanced inputs. In: Proc. 2008 American Control Conf., pp. 171–176 (2008)
7. Ermentrout, G.B., Kopell, N.: Frequency plateaus in a chain of weakly coupled oscillators. SIAM J. Math. Anal. 15, 215–237 (1984)
8. Ermentrout, G.B., Kopell, N.: Multiple pulse interactions and averaging in systems of coupled neural oscillators. J. Math. Biol. 29, 195–217 (1991)
9. Ermentrout, G.B.: Type I membranes, phase resetting curves, and synchrony. Neural Comput. 8(5), 979–1001 (1996)
10. Govaerts, W., Sautois, B.: Computation of the phase response curve: a direct numerical approach. Neural Comput. 18(4), 817–847 (2006)
11. Guckenheimer, J., Holmes, P.: Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields. Springer, Heidelberg (2002)
12. Hansel, D., Mato, G., Meunier, C.: Synchrony in excitatory neural networks. Neural Comput. 7(2), 307–337 (1995)
13. Hodgkin, A., Huxley, A.: A quantitative description of membrane current and its application to conduction and excitation in nerve. J. Physiol. 117, 500–544 (1952)
14. Hubel, D.H., Wiesel, T.N.: Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. J. Neurophysiol. 28, 229–289 (1965)
15. Izhikevich, E.M., Ermentrout, G.B.: Phase model. Scholarpedia 3, 1487 (2008)
16. Jeong, H., Gutkin, B.: Synchrony of neural oscillations controlled by gabaergic reversal potentials. Neural Comput. 19, 706–729 (2007)
17. Kandel, E.R., Schwartz, J.H., Jessel, T.M.: Principles of Neural Science. McGraw-Hill, New York (2000)
18. Kuramoto, Y.: Collective behavior of coupled oscillators. In: Handbook of Brain Theory and Neural Networks. MIT Press, Cambridge (1995)
19. Markman, A.B., Dietrich, E.: Extending the classical view of representation. Trends Cogn. Sci. 4(12), 470–475 (2000)
20. Molholm, S., Ritter, W., Murray, M.M., Javitt, D.C., Schroeder, C.E., Foxe, J.J.: Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study. Brain Res. Cogn. Brain Res. 14(1), 115–128 (2002)
21. Neidermeyer, E.: Epileptic Seizure Disorders. In: Electroencephalography: Basic Principles, Clinical Applications and Related Fields. Lippincott Williams & Wilkins (2005)
22. Oprisan, S.A., Canavier, C.C.: Stability analysis of ring of pulse coupled oscillators: the effect of phase resetting in the second cycle is important at synchrony and for long pulses. Differential Equations and Dynamical Systems 9, 243–258 (2001)
23. Paley, D.A., Leonard, N.E.: Oscillator models and collective motion. IEEE Control Systems Mag. 27, 89–105 (2007)
24. Phillips, W.A., Silverstein, S.M.: Convergence of biological and psychological perspectives on cognitive coordination in schizophrenia. Behav. Brain Sci. 26(1), 65–82, discussion 82–137 (February 2003)
25. Pikovsky, A., Rosenblum, M., Kurths, J.: Synchronization a Universal Concept in Nonlinear Sciences. Cambridge University Press, UK (2001)
26. Polleux, F., Lauder, J.M.: Toward a developmental neurobiology of autism. Ment. Retard Dev. Disabil. Res. Rev. 10(4), 303–317 (2004)
27. Roskies, A.L.: The binding problem. Neuron 24(1), 7–9 (1999)
28. Skinner, F.K.: Conductance based models. Scholarpedia 1(11), 1408 (2006)

29. Skinner, F.K., Chung, J.Y.J., Ncube, I., Murray, P.A., Campbell, S.A.: Using heterogeneity to predict inhibitory network model characteristics. J. Neurophysiol. 93(4), 1898–1907 (2005)
30. Stam, C.J., van der Made, Y., Pijnenburg, Y.A.L., Scheltens, P.: EEG synchronization in mild cognitive impairment and Alzheimer's disease. Acta Neurol. Scand. 108(2), 90–96 (2003)
31. Talathi, S.S., Hwang, D.-U., Ditto, W.L.: Spike timing dependent plasticity promotes synchrony of inhibitory networks in the presence of heterogeneity. J. Comput. Neurosci. 25(2), 262–281 (2008)
32. Talathi, S.S., Hwang, D.-U., Miliotis, A., Carney, P.R., Ditto, W.L.: Predicting synchrony in heterogeneous pulse coupled oscillators. Phys. Rev. E 80, 021908 (2009)
33. Talathi, S.S., Hwang, D.U., Carney, P.R., Ditto, W.L.: Synchrony with shunting inhibition in a feed forward inhibitory network. J. Comp. Neurosci. (in press, 2010)
34. Van Vreeswijk, C., Abbott, L.F., Ermentrout, G.B.: When inhibition not excitation synchronizes neural firing. J. Comput. Neurosci. 1(4), 313–321 (1994)
35. Varela, F., Lachaux, J.P., Rodriguez, E., Martinerie, J.: The brainweb: phase synchronization and large-scale integration. Nat. Rev. Neurosci. 2(4), 229–239 (2001)
36. Vida, I., Bartos, M., Jonas, P.: Shunting inhibition improves robustness of gamma oscillations in hippocampal interneuron networks by homogenizing firing rates. Neuron 49(1), 107–117 (2006)
37. Wang, X.J., Buzsáki, G.: Gamma oscillation by synaptic inhibition in a hippocampal interneuronal network model. J. Neurosci. 16(20), 6402–6413 (1996)
38. White, A., Chow, C., Ritt, J., Trevino, C., Kopell, N.: Synchronization and oscillatory dynamics in heterogeneous mutually inhibitory neurons. J. Comput. Neurosci. 5, 5–16 (1998)
39. Whittington, M.A., Traub, R.D., Jefferys, J.G.: Synchronized oscillations in interneuron networks driven by metabotropic glutamate receptor activation. Nature 373(6515), 612–615 (1995)
40. Winfree, A.T.: The Geometry of Biological Time, 2nd edn. Springer, Heidelberg (2001)

On the Stability and Instability of Padé Approximants[⋆,⋆⋆]

Christopher I. Byrnes and Anders Lindquist

Abstract. Over the past three decades there has been interest in using Padé approximants K with $n = \deg(K) < \deg(G) = N$ as "reduced-order models" for the transfer function G of a linear system. The attractive feature of this approach is that by matching the moments of G we can reproduce the steady-state behavior of G by the steady-state behavior of K, for certain classes of inputs. Indeed, we illustrate this by finding a first-order model matching a fixed set of moments for G, the causal inverse of a heat equation. A key feature of this example is that the heat equation is a minimum phase system, so that its inverse system has a stable transfer function G and that K can also be chosen to be stable. On the other hand, elementary examples show that both stability and instability can occur in reduced order models of a stable system obtained by matching moments using Padé approximants and, in the absence of stability, it does not make much sense to talk about steady-state responses nor does it make sense to match moments. In this paper, we review Padé approximants, and their intimate relationship to continued fractions and Riccati equations, in a historical context that underscores why Padé approximation, as useful as it is,

Christopher I. Byrnes
Department of Mathematics, Division of Optimization and Systems Theory, Royal Institute of Technology, 100 44 Stockholm, Sweden, and the Center for Research in Scientific Computation, North Carolina State University, Rayleigh, North Carolina 27695, USA
e-mail: chrisbyrnes@wustl.edu

Anders Lindquist
Department of Mathematics, Division of Optimization and Systems Theory, Royal Institute of Technology, 100 44 Stockholm, Sweden
e-mail: alq@math.kth.se

[⋆] This paper is dedicated to Yutaka Yamamoto, who has contributed to our understanding of a wide class of research problems and applications, including the algebraic systems theory of lumped and distributed parameter systems.

[⋆⋆] This research was supported in part by grants from AFOSR, Swedish Research Council, Swedish Foundation for Strategic Research, ACCESS Linnaeus Center and the Göran Gustafsson Foundation.

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 165–175.
springerlink.com

is not an approximation in any sense that reflects stability. Our main results on stability and instability states that if $N \geq 2$ and $\ell, r \geq 0$ with $0 < \ell + r = n < N$ there is a non-empty open set $U_{\ell,r}$ of stable transfer functions G, having infinite Lebesque measure, such that each degree n proper rational function K matching the moments of G has ℓ poles lying in $\mathbb{C}^-$ and r poles lying in $\mathbb{C}^+$. The proof is constructive.

1 Introduction

The power moments

$$E(X^k) = \int_{-\infty}^{+\infty} x^k p(x) dx \tag{1}$$

of a random variable X defined on $\mathbb{R}$ have played a prominent role in probability ever since their use by Chebychev in his proof of the Central Limit Theorem. Their importance is largely due to their interpretation in terms of the Taylor coefficients

$$\phi_X^{(k)}(0) = i^k E(X^k)$$

of the characteristic function

$$\phi_X(\xi) = \int_{-\infty}^{\infty} \mathrm{e}^{i\xi x} p(x) dx = \hat{p}(\xi),$$

which is the Fourier transform of the probability density function.

Similarly, if $G(s) = C(sI - A)^{-1}B$ is the transfer function of a strictly proper linear systems (A, B, C), then the moments of G may be defined [14, pp. 112–113] as

$$\eta_k = (-1)^k \frac{d^k G}{ds^k}(0). \tag{2}$$

If $\sigma(A) \subset \mathbb{C}^-$, the moments of the system coincide with the the power moments

$$\eta_k = (-1)^k \frac{d^k G}{ds^k}(0) = \int_0^{\infty} t^k g(t) dt$$

of the impulse response $g(t) = C\mathrm{e}^{At}B$, for $k \geq 0$. For example, η_0 is the DC gain, $-CA^{-1}B$, of the system. In this case, since whenever $\lim_{t \to \infty} f(t)$ exists and $s\hat{f}(s)$ has no poles in the closed right plane we have

$$\lim_{t \to \infty} f(t) = \lim_{s \to 0} s\hat{f}(s),$$

for any other *stable* linear system whose transfer function $K(s)$ satisfies

$$\frac{d^k K}{ds^k}(0) = (-1)^k \eta_k, \;\; 0 \leq k \leq d \tag{3}$$

the difference between the responses to a fixed polynomial input $u(t) = a_0 + \cdots + a_d t^d$ will decay to zero as $t \to \infty$. In particular, any lower order *stable* interpolant K will have the same step response as G. Of course, similar remarks about steady-state behavior apply to the more general moment matching problem for the data

$$\eta_k(s_0) = \int_0^\infty t^k g(t) \mathrm{e}^{-s_0 t} dt = (-1)^k \frac{d^k G}{ds^k}(s_0)$$

whenever $s_0 = i\omega_0$ and G and K are stable, as the next example shows.

Example 1. Consider the controlled heat equation system [8]:

$$z_t(x,t) = z_{xx}(x,t) \tag{4}$$

$$z(0,t) = 0, \tag{5}$$

$$z_x(1,t) = u \tag{6}$$

$$z(x,0) = \varphi(x). \tag{7}$$

$$y(t) = z(1,t), \tag{8}$$

with transfer function

$$H(s) = \frac{\sinh(\sqrt{s})}{\sqrt{s}\cosh(\sqrt{s})}.$$

We wish to design a stable controller $K(s)$ so that the cascade interconnection $H(s)K(s)$ provides steady state tracking of the desired output $y_R(t)$ when driven by the input $y_R(t)$. In fact, since the heat equation has a stable, causal inverse system

$$z_t(x,t) = z_{xx}(x,t) \tag{9}$$

$$z(0,t) = 0, \tag{10}$$

$$z_x(1,t) = y_r \tag{11}$$

$$z(x,0) = \psi(x). \tag{12}$$

$$u_r(t) = z(1,t), \tag{13}$$

with transfer function $G(s) = H^{-1}(s)$, one can indeed use G as a feedforward controller. On the other hand, if the reference trajectory is given, for example, by $y_R(t) = A\sin(2t)$ then a *finite dimensional* cascade controller can be obtained by using any rational stable function satisfying the interpolation conditions

$$K(2i) = G(2i) = 1.0856 + 0.6504i, \tag{14}$$

$$K(-2i) = G(-2i) = 1.0856 - 0.6504i, \tag{15}$$

rounding to four decimals. Indeed, driving

$$K(s) = 1.4108\,\frac{s - .1525}{s+1} \tag{16}$$

with $y_R(t)$ produces the steady-state control law, $u_R(t) = 1.2655\sin(2t + 0.5397)$.

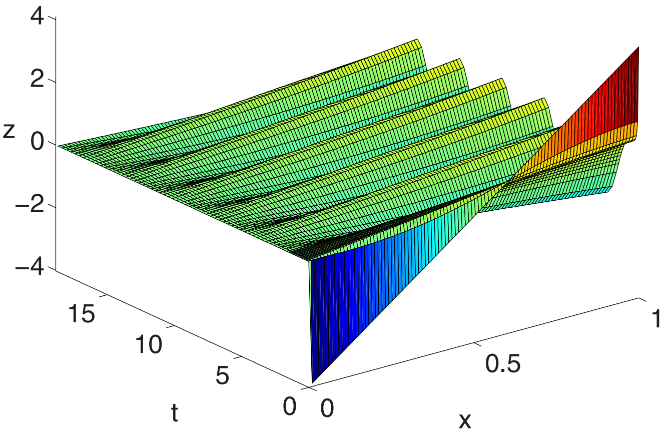


Fig. 1 Plot of solution surface for the cascade connection HK driven by y_r.

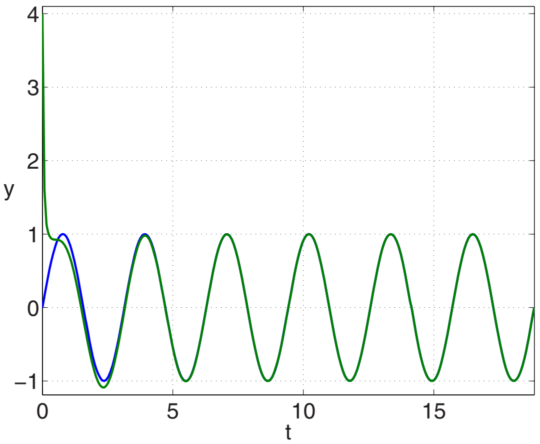


Fig. 2 Plot of $y(t)$ compared with $y_r(t)$

In the simulations depicted above, we have taken initial condition $\varphi(x) = -4(1-2x)$. The steady state behavior of the state trajectory is illustrated in Figure 1. The steady state behavior of the output trajectory is illustrated in Figure 2.

In contrast to our first example, however, even interpolation data generated by a stable second order system need not have a stable first order interpolant.

Example 2. Consider the critically damped harmonic oscillator with transfer function

$$G(s) = \frac{1}{s^2+2s+1} \tag{17}$$

and the induced one-parameter family of interpolation problems

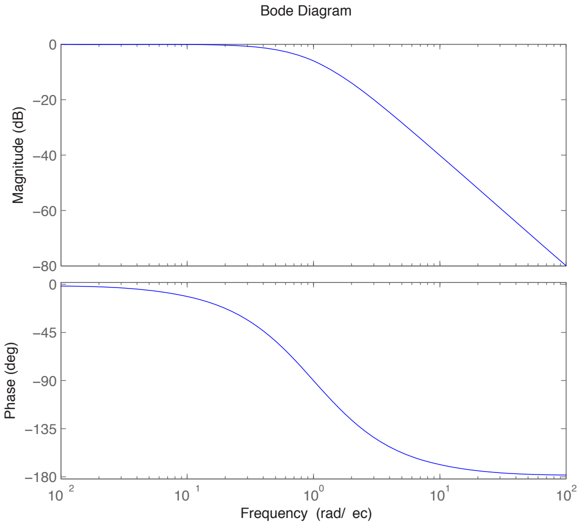


Fig. 3 Bode plots for $G(s) = \dfrac{1}{s^2+2s+1}$

$$K_\omega(i\omega) = G(i\omega),\ K_\omega(\infty) = G(\infty) = 0, \tag{18}$$

where for any fixed $\omega \in \mathbb{R}$ we seek a first order, stable interpolant K_ω.

First note that $-\pi/2 < \angle G(i\omega) < 0$ for any stable, strictly proper G with a positive high-frequency gain, while $\pi/2 < \angle G(i\omega) < \pi$ for any stable, strictly proper G with a negative high-frequency gain. On the other hand, $-\pi < \angle G(i\omega) < -\pi/2$ for $\omega > 1$, as is illustrated in Figure 3. In particular, the interpolation problem (18) has no stable, first order solution when $\omega > 1$.

Our final example illustrates the existence of stable rational interpolants for an open set of interpolation data.

Example 3. Consider the stable, minimum phase system with transfer function

$$G_\varepsilon(s) = \frac{s+1+\varepsilon}{s^2+2s+1} \tag{19}$$

and the one-parameter family of interpolation problems

$$K_\varepsilon(i) = G_\varepsilon(i),\ K_\varepsilon(\infty) = G_\varepsilon(\infty) = 0, \tag{20}$$

where for any fixed $\varepsilon \in \mathbb{R}$ we seek a first order, stable interpolant K_ε. Of course, for $\varepsilon = 0$, we can take $K_0(s) = \dfrac{1}{s+1}$. More generally, a stable first-order interpolant exists whenever $-1 < \varepsilon < 1$. Indeed, in this case we have $-\pi/2 < \angle G_\varepsilon(i\omega) < 0$ from which it is easy to construct a stable first order interpolant K_ε.

As Example 1 illustrates, there is potential use for such approximants K with $\deg(K) < \deg(G)$ as "reduced-order models" for G (see, e.g., [1]) when the class of inputs is restricted to sinusoids of a given frequency, *provided the interpolant K is stable*. On the other hand, Examples 2 and 3 show that both stability and instability can occur in reduced order models of a stable system obtained by matching moments. In this paper we shall develop some qualitative results about the stability and instability of strictly proper rational functions which match a sequence of moments of a rational transfer function at $s = 0$. We expect that similar results hold for moments computed along the imaginary axis. Roughly speaking, any transfer function K, stable or not, matching $\eta_k(0)$, for $k = 0, \dots, \tilde{n} < d$ is a *Padé approximation* to G. In Section 2, we review Padé approximants in more rigorous detail in a historical context that underscores why Padé approximation, as useful as it is, is not an approximation in any sense that reflects stability. In Section 3, we state our main results on stability and instability.

2 Padé Approximants, Continued Fractions and Riccati Equations

Over the past three decades there has been interest in using Padé approximants K with $\deg(K) < \deg(G)$ as "reduced-order models" for G (see, e.g., [1]). Rigorously, a Padé form of type (m,n) for G is a pair of polynomials (P,Q) with $\deg(P) \leq m$, $\deg(Q) \leq n$ such that

$$Q(s)G(s) - P(s) = O(s^{n+m+1}) \tag{21}$$

as $s \longrightarrow 0$. If $n = 0$, then (up to constant) P is the Taylor polynomial T_m of degree m. If $n,m \geq 1$ then $K(s) = P(s)/Q(s)$ is the ratio of two polynomials so that one might expect to obtain better approximations to G than T_m and, in many senses, this is true, explaining in part the ubiquity of Padé approximants. We shall be interested in the case $m \leq n$ and note that whenever

$$G(s) - K(s) = O(s^{n+m+1}) \tag{22}$$

as $s \longrightarrow 0$, then (21) holds. As Example 4 shows, the converse, however, is not true in general.

Padé approximants have found a remarkably wide array of applications in mathematics, engineering and science [18]. In particular, Padé's advisor, Hermite [13], used Padé approximants in 1873 to prove that e is transcendental. Euler [9] had already proved that e is irrational in 1739, by developing a continued fraction expansion for $e^{1/z}$ and evaluating at $z = 1$ to obtain

$$\alpha = \alpha_0 + \cfrac{1}{\alpha_1 + \cfrac{1}{\alpha_2 + \cfrac{1}{\alpha_3 + \cdots}}} \tag{23}$$

where $(\alpha_0, \alpha_1, \alpha_2, \cdots) = (2,1,2,1,1,4,1,1,6,\cdots)$. Since a number is rational if and only if its continued fraction expansion is finite, Euler concludes that e is irrational, but his proof that the continued fraction does not terminate is a remarkable method for summing a continued fraction by solving a Riccati equation. In 1775, Euler [10] returned to this observation in a paper (see also [3]) in which he shows that any continued fraction of the form

$$f(z) = \cfrac{1}{\pi_1(z) + \cfrac{1}{\pi_2(z) + \cfrac{1}{\pi_3(z) + \cdots}}} \tag{24}$$

can be summed by solving a Riccati differential equation and that the solution of any Riccati equation can be expressed as a continued fraction of the form (24). As one of several examples, he gives the continued fraction

$$\frac{e^{2/z}+1}{e^{2/z}-1} = z + \cfrac{1}{3z + \cfrac{1}{5z + \cfrac{1}{7z + \cdots}}} \tag{25}$$

for the hyperbolic function $\coth(1/z)$ which, when evaluated at $z = 2$, gives another proof that e is irrational.

Recall that a best rational approximant to a real number r is a rational number p/q such that $|r - p/q|$ is smaller than any other rational approximation with a smaller denominator. Among the remarkable properties of continued fraction expansions of a real number r is that the rational numbers obtained from the partial sums p_n/q_n obtained from $(\alpha_0, \alpha_1, \ldots, \alpha_n, 0, \ldots)$ turn out to be the sequence of best rational approximants to r and any best rational approximant to r arises in this way. For example, the continued fraction expansion of π yields the sequence $3/1, 22/7, 333/106, \ldots$ of best rational approximants. In general, one can show [12, p. 151] the stronger result that for any $p/q \neq p_n/q_n$

$$0 < q \leq q_n \implies |qr - p| > |q_n r - p_n| \tag{26}$$

Similarly, the partial sums obtained from a continued fraction expansion (24) for a function $f(z)$ form a sequence of Padé approximants (22).

Example 4. Padé approximants can be formed at any point in the extended complex plane, including $s = \infty$ as is treated in [18]. For example, given the Laurent expansion

$$G(s) = \gamma_0 + \gamma_1/s + \gamma_2/s^2 + \ldots\ , \tag{27}$$

consider the problem of finding partial realizations for the sequence of Markov parameters $(\gamma_1, \gamma_2, \gamma_2 \ldots) = (0,1,0,1,0,0,\ldots)$ generated by the fourth order linear system with transfer function $G(s) = (s^2+1)/s^4$ having a continued fraction expansion

$$G(s) = \frac{s^2+1}{s^4} = \cfrac{1}{s^2 - 1 + \cfrac{1}{s^2+1}} \tag{28}$$

Indeed while (21) has a solution of type $(1,1)$ the rational form of this expression in (22) does not, reflecting the fact that there is no partial realization for degree 1. On the other hand, $G_2(s) = 1/(s^2-1)$ is a second order partial realization obtained by truncating the continued fraction expansion. For a generic $G(s)$, the polynomials $\pi_i(s)$ will be linear functions [11, 17].

Remark 1. By analogy with the use of continued fractions in number theory, one might conclude that Padé approximants can be thought of as the "best" rational approximants to $f(z)$. However, while (26) is similar to (21) and $|r - p/q|$ is similar to (22), *best* in the sense of real and rational numbers is measured by absolute values of differences of real numbers while *best* for Padé approximants is measured by degrees of differences of polynomials and rational functions, which in general will not detemine the location of poles or zeros.

3 Main Results

The set of proper rational functions

$$\mathrm{Rat}^*(N) = \left\{ G : G(s) = \frac{p(s)}{q(s)},\ \deg(p) = \deg(q) = N,\ (p,q) = 1 \right\} \tag{29}$$

can be parameterized as an open, dense subset of $\mathbb{R}^{2N+1}$ using the coefficients of the polynomials

$$p(s) = p_N s^N + \cdots + p_1 s + p_0, \quad q(s) = s^N + q_{N-1} s^{N-1} + \cdots + q_0$$

We call $G \in \mathrm{Rat}^*(N)$ *stable* if all of its poles lie in the open left half plane $\mathbb{C}^-$ and *completely unstable* if all of it poles lie in the open right half plane $\mathbb{C}^+$. We are also interested in the number ℓ of poles of a rational function K lying in $\mathbb{C}^-$ and the number r of poles of K lying in $\mathbb{C}^+$. Thus, $\ell + r = n = \deg(K)$.

Theorem 1. *Suppose $N \geq 2$ and $\ell, r \geq 0$ with $0 < \ell + r = n < N$. For each pair ℓ, r there is a non-empty open cone $U_{\ell,r} \subset \mathrm{Rat}^*(N)$ of stable transfer functions G such that each degree n proper rational function K satisfying* (3) *with $d = 2n$ has ℓ poles lying in $\mathbb{C}^-$ and r poles lying in $\mathbb{C}^+$.*

In particular, for each n there does not exist a stable reduced order model of degree n for an open set of stable G having infinite Lebesgue measure.

Corollary 1. *Suppose $N \geq 2$. For each n satisfying $1 \leq n < N$ there is a non-empty open cone $U_n \subset \mathrm{Rat}^*(N)$ of stable transfer functions G such that each rational K satisfying* (3) *with $d = 2n$ is completely unstable.*

On the other hand, we have the following parallel positive result.

Corollary 2. *Suppose $N \geq 2$. For each n satisfying $1 \leq n < N$ there is a non-empty open cone $V_n \subset \mathrm{Rat}^*(N)$ of stable transfer functions G such that each rational K satisfying* (3) *with $d = 2n$ is stable.*

Proof. Each of the subsets

$$W_1^N = \{G \in \mathrm{Rat}^*(N) : G(0) \neq 0,\ G(\infty) \neq 0\},\ W_2^N = \{G \in \mathrm{Rat}^*(N) : q_0 \neq 0\}$$

is open and dense in $\mathrm{Rat}^*(N)$ and so is their intersection $W^N = W_1^N \cap W_2^N$. The function

$$T : W^N \to W^N \text{ defined by } T(G)(s) = G(1/s)$$

is a homeomorphism since it is continuous and its own inverse. Moreover, the map $s \to 1/s$ leaves both $\mathbb{C}^-$ and $\mathbb{C}^+$ invariant. Therefore, it suffices to prove Theorem 1 on W^N replacing (3) with the partial realization problem

$$\frac{d^kK}{ds^k}(\infty) = (-1)^k\gamma_k,\ 0 \leq k \leq 2n, \tag{30}$$

where $\gamma_0, \gamma_1, \gamma_2, \dots$ are the Markov parameters given by (27). Since solutions to the partial realization theorem are unchanged under multiplication by a non-zero constant, it is clear that the open sets described in Theorem 1 are cones and that it therefore suffices to prove that they are non-empty. Since we are interested only in the number of poles in open half-planes and stability, we can also suppress γ_0 so that we may assume that G is strictly proper. In this case, we are interested in the open dense set $U_N = W^N \cap V^N$ where

$$V^N = \{G : \det(\gamma_{i+j-1})_{i,j=1,\dots,N} \neq 0\}$$

which is known to be open and dense [5]. For any $G \in V^N$, any degree n rational function K satisfying (3) with $d = 2n$ is unique and can be constructed using the following algorithm.

Following [11], we associate a parameter sequence $\rho = (\rho_1, \dots, \rho_{2N})$ to each $G \in V^N$, where $\rho \in \mathscr{V}^N = \{\rho : \rho_i \neq 0,\ i = 1, \dots, 2N\}$. In [6, Lemma 1], it is shown that the map $\phi : V^N \to \mathscr{V}^N$ defined by $\phi(G) = \rho$ is a homeomorphism. From ρ one can [11] construct $K(s) = P_n(s)/Q_n(s)$ from the three-term recursions:

$$P_n(s) = (s - \rho_{2n})P_{n-1}(s) - \rho_{2n-1}P_{n-2}(s);\ P_0 = 0,\ P_{-1} = -1 \tag{31}$$

$$Q_n(s) = (s - \rho_{2n})Q_{n-1}(s) - \rho_{2n-1}Q_{n-2}(s);\ Q_0 = 1,\ Q_{-1} = 0. \tag{32}$$

Finally, in [6, Lemma 3], an open dense subset $\mathscr{U}_N \subset U_N$ is constructed so that the map $\rho \to (Q_N, Q_{N-1}, \rho_1)$ is a global, continuous change of coordinates.

Matters being so, we are now prepared to conclude the proof of Theorem 1 by induction on N. For $N = 2$, we have $n = 1$ and so either $\ell = 1,\ r = 0$ or $\ell = 0,\ r = 1$. In the first case, we construct the open subset

$$U_{1,0} = \{(Q_2, Q_1, \rho_1) \in \mathscr{U}_2 : Q_2 \text{ is stable, } \; Q_1 \text{ is stable, } \rho_1 \neq 0\}.$$

In the second case, we construct the open subset

$$U_{0,1} = \{(Q_2, Q_1, \rho_1) \in \mathscr{U}_2 : Q_2 \text{ is stable, } \; Q_1 \text{ is unstable, } \rho_1 \neq 0\}.$$

The latter construction is a special case of [6, Theorem 1], which was done in the case $n = r = N-1$.

We now assume Theorem 1 is true for $N-1$. In particular, for every $1 \leq \ell + r = n \leq N-2$ there exists an open subset $U_{\ell,r} \subset \mathscr{U}_{N-1}$ of stable rational functions P_{N-1}/Q_{N-1} so that the degree n partial realization has ℓ poles in $\mathbb{C}^-$ and r poles in $\mathbb{C}^+$. In the parameter sequence coordinates, we need to supplement the open set of corresponding $(\rho_1, \dots, \rho_{2N-2})$ by adding two more coordinates $\widetilde{\rho}_{2N-1}, \widetilde{\rho}_{2N}$ is such a way that Q_N is stable and the corresponding subset $U_{\ell,r} \subset \mathscr{U}_N$ of points $(\rho_1, \dots, \rho_{2N-2}, \widetilde{\rho}_{2N-1}, \widetilde{\rho}_{2N})$ is open. We first choose $\widetilde{\rho}_{2N} < 0$, so that the first term $d(s) = (s - \widetilde{\rho}_{2N})Q_{N-1}(s)$ appearing in the expression (32) for Q_N is a Hurwitz polynomial. We next write $n(s) = Q_{n-2}(s)$ and $k = -\widetilde{\rho}_{2N-1}$ so that (32) is the closed loop denominator $d(s) + kn(s)$ for the feedback system consisting of the stable open-loop system $g(s) = n(s)/d(s)$ with the feedback law $u = -ky$. In particular, for $\widetilde{\rho}_{2N-1}$ sufficiently small, the closed-loop system is stable and Q_N is a Hurwitz polynomial. Therefore, we have proved Theorem 1 for $n \leq N-2$.

Finally, suppose $n = N-1$. For any decomposition $\ell + r = n$, in the (Q_N, Q_{N-1}, ρ_1) coordinates on $\mathscr{U}_N$ we shall choose Q_N to be a Hurwitz polynomial and Q_{N-1} to have ℓ poles in $\mathbb{C}^-$ and r poles in $\mathbb{C}^+$. The corresponding subset $U_{\ell,r} \subset \mathscr{U}_N$ is again clearly open.

Remark 2. The importance of continued fractions in the deterministic partial realization problem was recognized in [15] and developed more comprehensively in [11], using the results in [16, 17]. These results were used in [6] to study the stability and instability properties of partial realizations, early results which are now generalized by Theorem 1. The inductive proof of Theorem 1 is constructive in each step and is phrased in terms of basic facts about root-loci. This is intimately related to the stability and instability proofs given in [6] using the Nyquist stability criterion. The geometry of the deterministic partial realization problem and its smooth parameterizations were studied in [5] using differential topology. The stochastic realization problem, which has proven much harder to analyze, was most recently studied using methods from algebraic geometry and differential topology in [7] in which it is shown, among other things, that there is no generic value for the degree of a minimal partial stochastic realization of a given covariance sequence $(\gamma_0, \dots, \gamma_n)$, in contrast to the deterministic partial realization problem.

References

1. Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems. SIAM, Philadelphia (2005)
2. Baker Jr., G.: Essentials of Padé Approximants. Academic Press, London (1975)

3. Bernlohr, C.R.: Continued fractions and the Riccati equation: an annotated translation of Euler, M.A. thesis, Ohio State University (1985); An annotated translation of [10]. The summation of a continued fraction whose indices constitute an arithmetic progression while all numerators are ones; Also a solution of the Riccati equation is taught through such fractions
4. Brezinski, C.: Padé-Type Approximation and General Orthogonal Polynomials. Birkhäuser Verlag, Basel (1980)
5. Brockett, R.W.: The geometry of partial realization problems. In: Proc. 17th IEEE Conf. on Decision and Control, pp. 1048–1052 (1978)
6. Byrnes, C.I., Lindquist, A.: Stability and instability of partial realizations. Systems & Control Letters 2, 99–105 (1982)
7. Byrnes, C.I., Lindquist, A.: On the partial stochastic realization problem. IEEE Trans. Automat. Control 42, 1049–1070 (1997)
8. Curtain, R., Zwart, H.: An Introduction to Infinite-Dimensional Linear System Theory. Springer, Berlin (1995)
9. Euler, L.: De Fractionibus Continuis Dissertatio. In: Proc. National Academy of St. Petersburg (1744) (in Latin); reprinted in Opera Omnia, Series I, vol. 14. English translation by Wyman, M.F., Wyman, B.F.: Math. Systems Theory, vol. 18, pp. 295–328 (1985)
10. Euler, L.: Summatio fractionis continuae cuius indices progressionem arithmeticam constituunt dum numeratores omnes sunt unitares ubi simul resolutio aequationes Riccatianae per huiusmodi fractiones docetur. Opulusca Analytica 2 (1785); reprinted in Opera Omnia, series I, vol. 23, Dulac, H. (ed.) pp. 174–194. Eneström no. 595
11. Gragg, W.B., Lindquist, A.: On the partial realization problem. Linear Algebra and its Appl. 50, 277–319 (1983)
12. Hardy, G.H., Wright, E.M.: An Introduction to the Theory of Numbers, 3rd edn. Oxford University Press, Oxford (1954)
13. Hermite, C.: Sur la fonction exponentielle. C. R. Acad. Sci., Paris 77, 18–24, 74–79, 226–233, 285–293 (1873); reprinted in Œuvres III, pp. 150–181
14. Kailath, T.: Linear Systems. Prentice-Hall, N.J. (1980)
15. Kalman, R.E.: On partial realizations, transfer functions and canonical forms. Acta Polytech. Scand. MA31, 9–39 (1979)
16. Lanczos, C.: An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. J. Res. Nat. Bur. Standards 45, 255–282 (1950)
17. Magnus, A.: Certain continued fractions associated with the Padé table. Math. Z. 78, 361–374 (1962)
18. Nikishin, E.M.: The Padé Approximants. In: Proc. of the Int. Congress of Mathematicians, Helsinki, pp. 623–630 (1978)

On the Use of Functional Models in Model Reduction

Paul A. Fuhrmann[⋆] and Uwe Helmke[⋆⋆]

Abstract. It has been known for some time that the Sylvester equation plays a significant role in interpolation, feedback control, observer theory and model reduction problems. In this paper, in place of state space techniques, we use polynomial models to replace the standard Sylvester equation by a polynomial version. The polynomial Sylvester equation is closely related to a Bezout equation. We use this functional setting to unify various model reduction techniques.

Keywords: model reduction, interpolation, Hankel norm approximation, balanced truncation, tensor products.

1 Introduction

It has been known for some time that the *matrix Sylvester equation*

$$AX - BX = C \tag{1}$$

plays a significant role in interpolation, feedback control, observer theory and model reduction problems. For an excellent introduction to the model reduction area, the area of focus of this paper, we refer to Antoulas [1] where one can find many instances how the Sylvester equation appears in model reduction tasks. In this paper, in place of state space techniques, we use polynomial models to replace the standard Sylvester equation by a polynomial version. Thus, instead of (1) we consider the *polynomial Sylvester equation*

Paul A. Fuhrmann
Department of Mathematics, Ben-Gurion University of the Negev, Beer Sheva, Israel

Uwe Helmke
Universität Würzburg, Institut für Mathematik, Würzburg, Germany

[⋆] Partially supported by the ISF under Grant No. 1282/05.

[⋆⋆] Partially supported by the DFG under Grant HE 1858/12-1.

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 177–187.
springerlink.com

$$D_2(z)X_1(z) - X_2(z)D_1(z) = R(z), \tag{2}$$

in two unknowns X_1, X_2, being rectangular polynomial matrices. Here the nonsingular polynomial matrices D_1, D_2 are suitable functional models for the matrices A, B and $R(z)$ is related to C. A possible choice would be $D_1(z) := zI - A, D_2(z) = zI - B$.

What is the advantage of such a functional model approach to the Sylvester equation? A first answer is in the reduced computational complexity of solving associated polynomial Sylvester equations. For example, if A, B are companion matrices of monic polynomials, then (2) becomes a scalar Bezout equation, for which fast solution methods are available. This should be compared with the standard approach to solving (1), that reformulates the equation via column vec operations and Kronecker matrix product as a standard linear equation

$$(A \otimes I - I \otimes \tilde{B})\mathrm{vec(X)} = \mathrm{vec(C)}. \tag{3}$$

Despite the increased size of the Kronecker form matrices, quite a lot of the special structure in the matrices A, B, C, X gets lost by this process. Thus one is asking for less invasive approaches that preserve the parametric information on the data. Polynomial models provide the right language for doing so.

A second answer is linked to the pivotal role functional models play in unifying frequency domain data and state space representations. This is exactly the idea underlying polynomial models; see e.g. Fuhrmann [3, 8]. In fact, the Sylvester equation provides the link between various methods from model reduction, making its appearance in both projection type and Krylov type methods; see Gallivan, Vanderdorpe and Van Dooren [13]; Genin and Vandendorpe [14] as well as in interpolation results; see Gallivan, Vandendorpe and Van Dooren [12], Fanizza, Karlsson, Lindquist and Nagamune [2], Sorensen [21]. The drawback in most current approaches to the model reduction problem using state space construction is that both the geometric information, and even more so, the interpolation information has to be encoded in an indirect way. This can be overcome by using functional models which exhibit in a clearer way the functional properties of a transfer function. For example, the connection between orthogonal projections in the Hardy space context and the Lyapunov equation is most clearly described in Fuhrmann and Gombani [9]. In the case of scalar transfer functions, both Hankel norm approximations as well as balanced approximation can be approached via the use of a polynomial Sylvester like equation which turns out to be a derivative of the polynomial Bezout equation. The use of the polyomial Sylvester equation allows us to interpret directly the results in terms of interpolation. These methods were worked out in Fuhrmann [6, 7] and extended in Fuhrmann and Ober [11].

Throughout this paper we will deal only with the case of scalar transfer functions, i.e. with SISO systems. Due to severe space constraints, we have not been able to give anything close to a detailed description of how the polynomial Sylvester equation enters into model reduction theory via interpolation methods. Instead, we focus on describing just a few instances, such as partial realizations, Hankel norm approximation and projection methods. Other topics, such as balanced truncation, rational approximation and connections to Bezoutians and finite section Hankels are left out.

2 Polynomial Models and the Sylvester Equation

2.1 Polynomial Models and Tensor Products

We briefly introduce the language of polynomial models and explain some recent ideas concerning tensor products. Polynomial models were introduced in Fuhrmann [3], which is still a good source for the full details. For simplicity, we restrict ourselves completely to the scalar case of single input-single output systems and associated polynomial models of scalar polynomials.

The canonical decomposition of Laurent series over a field $\mathbb{F}$ into its polynomial and strictly proper parts, respectively, leads immediately to the direct sum representation $\mathbb{F}((z^{-1})) = \mathbb{F}[z] \oplus z^{-1}\mathbb{F}[[z^{-1}]]$, with corresponding projections π_+, π_- on $\mathbb{F}[z], z^{-1}\mathbb{F}[[z^{-1}]]$ respectively given by

$$\pi_+ \sum_{-\infty}^{N} f_i z^i = \sum_{i=0}^{N} f_i z^i, \quad \pi_- \sum_{-\infty}^{N} f_i z^i = \sum_{i=-\infty}^{-1} f_i z^i.$$

For any $f \in \mathbb{F}((z^{-1}))$, we refer to $\mathrm{Res}\,(f) = f_{-1}$ as the **residue** of f.

Given a monic polynomial $q(z) \in \mathbb{F}[z]$, we define a projection $\pi_q : \mathbb{F}[z] \to \mathbb{F}[z]$ by

$$\pi_q f = q\pi_-(q^{-1}f) \quad \text{for } f \in \mathbb{F}[z]. \tag{4}$$

Clearly, $\mathrm{Ker}\,\pi_q = q\mathbb{F}[z]$. We define the **polynomial model** X_q by

$$X_q = \mathrm{Im}\,\pi_q. \tag{5}$$

It follows from (4) that $f \in X_q$ if and only if $q^{-1}f$ is strictly proper. Moreover, X_q is a finite-dimensional $\mathbb{F}$ vector space of dimension $\dim X_q = \deg q(z)$. Introducing in X_q an $\mathbb{F}[z]$-module structure by

$$p \cdot f = \pi_q(pf), \quad p \in \mathbb{F}[z],\ f \in X_q, \tag{6}$$

the map $\pi_q : \mathbb{F}[z] \longrightarrow X_q$ becomes a surjective $\mathbb{F}[z]$-homomorphism. Hence, using $\mathrm{Ker}\,\pi_q = q\mathbb{F}[z]$, we have the $\mathbb{F}[z]$-isomorphism with the quotient module

$$X_q \simeq \mathbb{F}[z]/q(z)\mathbb{F}[z]. \tag{7}$$

A special case of (6) is the **shift operator** $S_q : X_q \longrightarrow X_q$ by

$$(S_q f)(z) = zf(z) - q(z)\xi_f, \qquad f \in X_q, \tag{8}$$

where $\xi_f = (q^{-1}f)_{-1}$ is the residue of $q^{-1}f$. Thus S_q defines the module action on X_q, i.e. $p \cdot f = p(S_q)f$.

On the space $\mathbb{F}((z^{-1}))$ of truncated Laurent series, we introduce a bilinear form, given, for $f(z)=\sum_{j=-\infty}^{n_f} f_j z^j$ and $g(z)=\sum_{j=-\infty}^{n_g} g_j z^j$, by

$$[f,g]=\sum_{j=-\infty}^{\infty} f_j g_{-j-1}. \tag{9}$$

Clearly, the sum in (9) is well defined, as only a finite number of summands are nonzero. Given a subspace $M\subset\mathbb{F}((z^{-1}))$, we let $M^{\perp}=\{f\in F(z)|[m,f]=0,\forall m\in M\}$. It is easy to check that $\mathbb{F}[z]^{\perp}=\mathbb{F}[z]$. We can identify the dual of the space of $\mathbb{F}[z]$ with $z^{-1}\mathbb{F}[[z^{-1}]]$. By defining a new pairing

$$\langle f,g\rangle=[d^{-1}f,g]=[f,d^{-1}g] \tag{10}$$

for all $f,g\in X_d$. Under the pairing introduced in (10), we can identify the dual space of X_d with X_d. Moreover, we have $S_d^*=S_d$. For the full details of duality in the context of polynomial models, see Fuhrmann [4, 8].

Given two monic polynomials $q(z),\overline{q}(z)\in\mathbb{F}[z]$ of degrees $\deg q=n,\deg\overline{q}=\overline{n}$, respectively, with associated polynomial models $X_q,X_{\overline{q}}$, one can construct tensor product spaces in four different ways. For us the following construction of the Kronecker product model will be sufficient. We refer to Fuhrmann and Helmke [10] for the full details. We define the $\mathbb{F}$-Kronecker product of two scalar polynomials $q,\overline{q}\in\mathbb{F}[z]$ as the map $q\otimes_{\mathbb{F}}\overline{q}:\mathbb{F}[z,w]\longrightarrow\mathbb{F}[z,w]$ given by $(q\otimes_{\mathbb{F}}\overline{q})x(z,w)=q(z)x(z,w)\overline{q}(w)$. This map induces a projection $\pi_{q\otimes_{\mathbb{F}}\overline{q}}$ in $\mathbb{F}[z,w]$ defined by

$$\pi_{q\otimes_{\mathbb{F}}\overline{q}}\,x(z,w)=(q\otimes_{\mathbb{F}}\overline{q})(\pi_-^z\otimes\pi_-^w)(q\otimes_{\mathbb{F}}\overline{q})^{-1}x(z,w). \tag{11}$$

We obtain the **$\mathbb{F}$-Kronecker product model** as $X_{q(z)\overline{q}(w)}:=\operatorname{Im}\pi_{q\otimes_{\mathbb{F}}\overline{q}}$. By inspection, $X_{q(z)\overline{q}(w)}$ coincides with the $\mathbb{F}$-vector space of all polynomials in two variables $x(z,w)\in\mathbb{F}[z,w]$ for which $q(z)^{-1}x(z,w)\overline{q}(w)^{-1}$ is strictly proper in both variables. Thus the elements $x\in X_{q(z)\overline{q}(w)}$ are exactly the polynomials $x(z,w)$ that have degrees $<n$ in z and $<\overline{n}$ in w, respectively. $X_{q(z)\overline{q}(w)}$ is thus a finite-dimensional vector space of dimension $\dim X_{q(z)\overline{q}(w)}=n\overline{n}$. Moreover, $X_{q(z)\overline{q}(w)}$ has a natural $\mathbb{F}[z,w]$-module structures, given by

$$p(z,w)\cdot x(z,w)\;=\;\pi_{q\otimes_{\mathbb{F}}\overline{q}}(p(z,w)x(z,w)),\quad x(z,w)\in X_{q(z)\overline{q}(w)} \tag{12}$$

where $p(z,w)\in\mathbb{F}[z,w]$. Therefore, the Kronecker product model has two independent, commuting shift operators, one by multiplication with z and the other by multiplication with w.

2.2 The Polynomial Sylvester Equation

Given matrices $A\in\mathbb{F}^{p\times p}$ and $B\in\mathbb{F}^{m\times m}$, the classical Sylvester operator is the $\mathbb{F}$-linear map $S_{A,B}:\mathbb{F}^{p\times m}\longrightarrow\mathbb{F}^{p\times m}$ defined by $S_{A,B}X=AX-XB$. Of course, the Sylvester equation then is, for $C\in\mathbb{F}^{p\times m}$, the linear matrix equation

$$AX - XB = C. \tag{13}$$

How do we solve such linear matrix equations? One straightforward way would be to vectorize the equation. Thus, using column vec-operations we identify the matrix space $\mathbb{F}^{p\times m}$ with $\mathbb{F}^{pm}$ and (13) becomes equivalent with the standard linear equation in mp variables $(A \otimes I - I \otimes \tilde{B})\mathrm{vec(X)} = \mathrm{vec(C)}$. Obviously, such a process has several drawbacks. One is the large complexity of solving such a equations, which grows in the order of $(mp)^3$; thus the standard Gaussian elimination process cannot be applied for large scale problems. Second, the structure of the matrices and solutions may get lost in the Kronecker-vec description and cannot be easily recovered. An example of this type is passivity preserving model reduction for passive systems. In this case, the special structure is totally obliterated by the vec-operations. For the purpose of model reduction, we find it therefore more interesting, as well as useful, to consider the Sylvester map when A,B have functional representations in terms of suitable polynomial model spaces associated with nonsingular polynomial matrices. This can be done in full generality, without any assumption on A,B. However, in the scalar case that we consider here, this restriction amounts to require that A,B are cyclic matrices with characteristic polynomials $q,\overline{q}$, respectively. In either case, it is convenient to use tensor products of polynomial models.

Given arbitrary monic polynomials $q \in \mathbb{F}[z], \overline{q} \in \mathbb{F}[w]$, the **polynomial Sylvester operator** is the linear map $\mathscr{S} : X_{q(z)\overline{q}(w)} \to X_{q(z)\overline{q}(w)}$ defined by

$$(\mathscr{S}x)(z,w) = \pi_{q\otimes\overline{q}}((z-w)x(z,w)). \tag{14}$$

Using (8), it can be shown that the Sylvester map can be written as follows

$$(\mathscr{S}x)(z,w) = [zx(z,w) - x(z,w)w] - [q(z)\overline{p}(w) - p(z)\overline{q}(w)]$$

for suitable polynomials $\overline{p}(z), p(z)$ for which $p(z)/q(z) = \overline{p}(z)/\overline{q}(z)$ is strictly proper. Note that every nonzero scalar two variable polynomials $x(z,w), t(z,w) \in X_{q(z)\overline{q}(w)}$ are such that $q(z)^{-1}t(z,w)\overline{q}(w)^{-1}$ and $q(z)^{-1}x(z,w)\overline{q}(w)^{-1}$ are strictly proper in the variables z,w. Therefore, the Sylvester equation

$$\mathscr{S}x(z,w) = \pi_{q(z)\overline{q}(w)}(z-w)x(z,w) = c(z,w) \tag{15}$$

is solvable in $X_{q(z)\overline{q}(w)}$ if and only if there exists polynomials $p(z)$ and $\overline{p}(z)$ with $p/q, \overline{p}/\overline{q}$ strictly proper for which

$$q(z)\overline{p}(z) - p(z)\overline{q}(z) + c(z,z) = 0. \tag{16}$$

We will refer to (16) as the **polynomial Sylvester equation** (**PSE**). In that case, the solution is given by

$$x(z,w) = \frac{q(z)\overline{p}(w) - p(z)\overline{q}(w) + c(z,w)}{z-w}. \tag{17}$$

Similarly, a two-variable polynomial matrix $x(z,w) \in X_{q(z)\overline{q}(w)}$ solves the **homogeneous polynomial Sylvester equation** (**HPSE**), if and only if there exist polynomials $p(z)$ and $\overline{p}(z)$ with $p/q, \overline{p}/\overline{q}$ strictly proper and

$$q(z)\overline{p}(z) - p(z)\overline{q}(z) = 0 \tag{18}$$

such that

$$x(z,w) = \frac{q(z)\overline{p}(w) - p(z)\overline{q}(w)}{z-w}. \tag{19}$$

We now show how the polynomial Sylvester equation helps to solve matrix Sylvester equations. Given the polynomials $q(z) = z^n + q_{n-1}z^{n-1} + \cdots + q_0$ and $\overline{q}(z) = z^{\overline{n}} + \overline{q}_{n-1}z^{\overline{n}-1} + \cdots + \overline{q}_0$, we define the companion matrices, using Kalman's notation, by

$$C_q^\sharp = \begin{pmatrix} 0 & & & -q_0 \\ 1 & & & \cdot \\ & \cdot & & \cdot \\ & & \cdot & \cdot \\ & & 1 & -q_{n-1} \end{pmatrix}, \qquad C_q^\flat = \begin{pmatrix} 0 & 1 & & & \\ & & \cdot & & \cdot \\ & & & \cdot & \\ & & & & 1 \\ -q_0 & \cdot & \cdot & \cdot & -q_{n-1} \end{pmatrix}. \tag{20}$$

The companion matrices $C_{\overline{q}}^\sharp$ and $C_{\overline{q}}^\flat$ are similarly defined. The above arguments lead to the following result.

Theorem 1. *Let $n \times n$ and $\overline{n} \times \overline{n}$ matrices $A := C_q^\sharp$ and $B := C_{\overline{q}}^\flat$ be given, and $C \in \mathbb{F}^{n\times\overline{n}}$. Let $c(z,w) = \sum_{i=1}^{n}\sum_{j=1}^{\overline{n}} c_{ij}z^{i-1}w^{j-1}$. Then $X = (x_{ij}) \in \mathbb{F}^{n\times\overline{n}}$ is a solution of the matrix Sylvester equation*

$$AX - XB = C$$

if and only if $x(z,w) := \sum_{i=1}^{n}\sum_{j=1}^{\overline{n}} x_{ij}z^{i-1}w^{j-1}$ solves the polynomial Sylvester equation

$$\mathscr{S}x(z,w) = \pi_{q(z)\overline{q}(w)}(z-w)x(z,w) = c(z,w)$$

Thus, for A,B defined by $C_q^\sharp, C_{\overline{q}}^\flat$, the solutions to the polynomial Sylvester equation (16) yield a complete parametrization of all solutions to the above matrix Sylvester equation. Note, that the computational complexity of solving the polynomial Sylvester equation (16) is much lower than that of the matrix Sylvester equation, i.e. $< (n\overline{n})^3$. Thus, the special structure of the companion matrices A,B is effectively encoded into (16). For arbitrary matrices A,B, the connection with the classical Sylvester equation is a bit more complicated and requires a generalization of the above construction to matrix Kronecker product polynomial model of $D_1(z) := zI - A$ and $D_2(w) - B$. For the details, which we omit, we refer to Fuhrmann and Helmke [2009].

The analysis of two-variable polynomial model spaces reduces the analysis of the general Sylvester equation to a polynomial equation of Bezout type. This extends the method for the analysis of the Lyapunov equation which was motivated by Kalman [19] and introduced in Willems and Fuhrmann [24]. A special case is of course the homogeneous Sylvester equation which has a direct connection to the theory of Bezoutians. In case, as often arises in practice, that the characteristic polynomials of $\overline{q}(z)$ and $q(z)$ are coprime, the polynomial Sylvester equation can be solved by solving a scalar Bezout equation, which can be easily done using the Euclidean algorithm. We note in passing, that a systematic use of two-variable polynomial equations in stability theory has first appeared in Willems and Fuhrmann [24]; see also Willems and Trentelman [25] for related material using the language of quadratic differential forms. A rather non-systematic and adhoc approach without any reference to functional models has been recently given by Peeters and Rapisarda [20]; the connection with tensor products of polynomial models has been apparently overlooked by these authors.

Note that the polynomial Sylvester equation (16) can be interpreted on several different levels. First, it is a linear equation in the vector space $X_{q(z)\overline{q}(w)}$. However, any element $c(z,w) \in X_{q(z)\overline{q}(w)}$ can be interpreted as an element $\mathscr{C} \in \mathrm{Hom}_{\mathbb{F}}(X_{\overline{q}}, X_q)$, i.e. as a linear transformation from $X_{\overline{q}}$ to X_q defined by $\mathscr{C}g = \langle g, c(z,\cdot)\rangle$, thus the polynomial Sylvester equation becomes an operator equation. Finally, taking matrix representations with respect to bases of choice in $X_{\overline{q}}$ and X_q, the Sylvester equation becomes a matrix equation in $\mathbb{F}^{n\times\overline{n}}$.

3 Model Reduction

In this section we apply the polynomial Sylvester equation to the problem of model reduction, i.e. to the approximation of a large scale linear system by a lower order, easier to compute with but such that the approximation error is kept under control. We will treat only the scalar case, i.e. the case of a SISO system with its transfer function given by $g(z) = p(z)/q(z)$, where we assume that the polynomials $p(z), q(z)$ are coprime. Furthermore, we assume $g(z)$ to be strictly proper, i.e. that $\deg p < \deg q$. Our aim is to show how the polynomial Sylvester equation can be used directly for model reduction purposes both using the interpolation method or the projection method. We will show also how the important methods of Hankel norm approximation and balanced truncation relate to the polynomial Sylvester equation.

3.1 The Sylvester Equation and Interpolation

The simplest case of model reduction by interpolation is that of approximation at ∞ up to a certain order. Given a strictly proper rational function $g(z)$ having the expansion $g(z) = \sum_{j=1}^{\infty} \frac{g_j}{z^j}$ at ∞, we look for a lower degree function $\overline{g}(z)$ that matches the Markov parameters of $g(z)$, i.e.coefficients $\{g_j\}_{j=1}^{\infty}$, up to a certain order. A realization of $\overline{g}(z)$ is called a partial realization of $g(z)$.

Theorem 2. *Let $g(z) = p(z)/q(z)$ be strictly proper with $q(z), p(z) \in \mathbb{F}[z]$ coprime polynomials. Let $\overline{q}(z), \overline{p}(z) \in \mathbb{F}[z]$ be the unique solution of the polynomial Bezout-Sylvester equation*

$$q(z)\overline{p}(z) - p(z)\overline{q}(z) + 1 = 0 \tag{21}$$

that satisfies the degree constraints

$$\deg \overline{q} < \deg q, \quad \deg \overline{p} < \deg p. \tag{22}$$

Then $\overline{g}(z) = \overline{p}(z)/\overline{q}(z)$ *is a partial realization of* $g(z)$ *that approximates* $g(z)$ *at* ∞ *up to order* $n+\overline{n}$.

It is well known from antiquity that the Bezout-Sylvester equation can be solved by using the Euclidean algorithm, i.e. by recursively using the division rule of polynomials. The approximant can be obtained by use of the Lanczos polynomials which are orthogonal polynomials corresponding to a Hankel matrix associated with $g(z)$. We shall not dwell on this and refer to Gragg and Lindquist [16] or Fuhrmann [8].

So far, we used only the Bezout-Sylvester equation. If we replace it by the polynomial Sylvester equation

$$q(z)\overline{p}(z) - p(z)\overline{q}(z) + r(z) = 0, \tag{23}$$

where $\rho = \deg r < \deg q$, then again there exists a unique solution that satisfies the degree constraints (22). Thus for the error transfer function we obtain

$$e(z) = g(z) - \overline{g}(z) = \frac{p(z)}{q(z)} - \frac{\overline{p}(z)}{\overline{q}(z)} = \frac{r(z)}{q(z)\overline{q}(z)}.$$

This shows that $\overline{g}(z)$ can be obtained by interpolating the values of $g(z)$, and its derivatives up to appropriate orders, as well as at ∞ to the order of $n+\overline{n}-\rho$.

3.2 The Sylvester Equation and the Projection Method

Assume we have a SISO system with a transfer function $g(z)$ given by the minimal realization, in the state space $\mathscr{X}$, $g(z) = D + C(zI-A)^{-1}B$ which we want to approximate by a transfer function having the minimal realization, in the state space $\overline{\mathscr{X}}$, $\overline{g}(z) = \overline{D} + \overline{C}(zI-\overline{A})^{-1}\overline{B}$ and which has a lower McMillan degree. Let $\mathscr{Y}, \mathscr{W} : \overline{\mathscr{X}} \longrightarrow \mathscr{X}$ be injective linear maps with $\mathscr{W}^* : \mathscr{X} \longrightarrow \overline{\mathscr{X}}$ a suitable adjoint, for which we have

$$\mathscr{W}^*\mathscr{Y} = I, \quad \overline{A} = \mathscr{W}^*A\mathscr{Y}, \quad \overline{B} = \mathscr{W}^*B, \quad \overline{C} = C\mathscr{Y}, \quad \overline{D} = D. \tag{24}$$

The assumption that $\mathscr{W}^*\mathscr{Y} = I$ shows that we have the direct sum decomposition $\mathscr{X} = \operatorname{Im}\mathscr{Y} \oplus \operatorname{Ker}\mathscr{W}^*$. With respect to this direct sum, and choosing appropriate bases, we have the matrix representations

$$\begin{aligned} &A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}, \quad C = \begin{pmatrix} C_1 & C_2 \end{pmatrix} \\ &\overline{A} = A_{11}, \quad \overline{B} = B_1, \quad \overline{C} = C_1. \end{aligned} \tag{25}$$

This shows that the approximating system is a projection of the original system. It goes without saying that the properties of the approximating system depend on the choice of subspaces.

Our principal result in this section is the following theorem which shows that solving the polynomial Sylvester equation leads to a reduced model obtained by the projection method. With $g(z) = p(z)/q(z)$ we associate the shift realization, in the state space X_q,

$$\Sigma_{q^{-1}p} := \begin{cases} A = S_q \\ B\alpha = p\alpha, \qquad \alpha \in \mathbb{F} \\ Cf = (q^{-1}f)_{-1}, \end{cases} \tag{26}$$

whereas with $\overline{g}(z) = \overline{p}(z)/\overline{q}(z)$ we associate the shift realization

$$\Sigma_{\overline{pq}^{-1}} := \begin{cases} \overline{A} = S_{\overline{q}} \\ \overline{B}\alpha = \alpha, \qquad \alpha \in \mathbb{F} \\ \overline{C}g = (\overline{pq}^{-1}g)_{-1}. \end{cases} \tag{27}$$

Theorem 3. *Let $q(z), p(z) \in \mathbb{F}[z]$ be coprime polynomials. Let $\overline{q}(z), \overline{p}(z) \in \mathbb{F}[z]$ be the unique solution of the polynomial Sylvester equation (23) that satisfies the degree constraints (22). For the realizations defined by (26) and (27), define linear maps $\mathscr{Y} : X_{\overline{q}} \longrightarrow X_q$, $\mathscr{W} : X_{\overline{q}} \longrightarrow X_q$ and $\mathscr{W}^* : X_q \longrightarrow X_{\overline{q}}$ as follows:*

$$\begin{aligned} \mathscr{Y} &= p(S_q)\pi_q \\ \mathscr{W} &= \overline{q}(S_q)q\pi^q\overline{q}^{-1} \\ \mathscr{W}^* &= \pi_{\overline{q}}\overline{q}(S_q). \end{aligned} \tag{28}$$

With such choices, equations (24) are satisfied.

3.3 Hankel Norm Approximation

Since the fundamental work of Adamjan, Arov and Krein, and in particular the influential paper by Glover [15], Hankel norm approximations became a standard tool of modern control theory. In Fuhrmann [6, 7], precise results on the geometry of the approximation problem were derived. For the full details of the short description below, we refer to these two papers. Our aim in this section is to show how Hankel norm approximation is related to a polynomial Sylvester equation and the optimal approximant determined by interpolation. To do this, we have to depart from the convention adopted before of taking the solution of the Sylvester equation that satisfies the degree constraints (22). The reason for this is the following. It has been stated in the literature, see for example Sorensen and Antoulas [22] or Gallivan, Vandendorpe and Van Dooren [12], that for model reduction we may assume that if the transfer function of the system is strictly proper then so is the approximant. Our analysis of Hankel norm approximation shows that this is obviously false when norm bounds for the error function are involved. In fact, in the Hankel norm approximation case, the error function turns out to be a constant multiple of an all-pass function, so the optimal approximant cannot be strictly proper. A simple example is $g(z) = 2(z-1)^{-1} \in H_{-}^{\infty}$ with $\|g\|_{\infty} = 2$. A 0-th order approximant then is -1 with $\|e\|_{\infty} = 1$.

We consider a scalar rational, strictly proper and, antistable transfer function $\phi = \frac{n}{d} \in H_{-}^{\infty}$, with n, d coprime polynomials and H_{-}^{∞} the Hardy space of bounded analytic functions in the left half plane. The corresponding Hankel operator $H_{\phi} : H_{+}^{2} \longrightarrow H_{-}^{2}$ is defined by $H_{\phi} f = P_{-} \phi f$, where P_{-} is the orthogonal projection of L^2 onto H_{-}^{2}. The assumed rationality of ϕ implies that H_{ϕ} has finite rank, i.e. that $\dim H_{\phi} = \deg d < \infty$. Note that by the definition of the Hankel operator, we have $||H_{\phi}|| \leq ||\phi||_{\infty}$. It is well known that approximating a compact, and in particular a finite rank operator, by operators of lower rank is related to its singular values, i.e. to the eigenvalues of A^*A and AA^*. The spectral analysis of these, self adjoint, operators hinges on the Schmidt pairs. It has been shown that the Schmidt pair of the Hankel operator H_{ϕ} corresponding to the singular value μ has the representation $\{\frac{p}{d}, \frac{p^*}{d^*}\}$, where the polynomial p is a solution of the fundamental polynomial equation

$$np = \lambda d^* p^* + d\pi, \tag{29}$$

with λ real and $|\lambda| = \mu$. Specializing the results of Fuhrmann [1994] to the generic case of distinct singular values, then the number of antistable zeros of p_k is $k-1$. In particular, p_n is antistable. From (29), we obtain

$$\frac{n}{d} - \frac{\pi_k}{p_k} = \lambda_k \frac{d^* p_k^*}{d p_k}, \tag{30}$$

whichimplies that $||H_{\frac{n}{d}} - H_{\frac{\pi_k}{p_k}}|| = \mu_k$. Clearly, $\operatorname{rank} H_{\frac{\pi_k}{p_k}} \leq k-1$.

For the present purposes, the point to note is that from (29) it follows that $\frac{\pi_k}{p_k}$ interpolates the values of $\phi = \frac{n}{d}$ at the $2n-1$ zeros of $d^* p_k^*$. In the special case of $k = n$, the polynomials d^*, p_k^* are both stable, so all interpolation points are in the open left half plane.

We wish to point out that the fundamental polynomial equation (29) is not, in spite of the similarity, a polynomial Sylvester equation. Thus it cannot be solved using the Euclidean algorithm. Instead, one has to solve a generalized eigenvalue problem. This is to be expected as an operator norm is involved. Even so, for the case $k = n$, equation (30) is equivalent to the Bezout equation, but over H_{+}^{∞},

$$\frac{n}{d^*}\left(\frac{1}{\lambda_n}\frac{p_n}{p_n^*}\right) - \frac{d}{d^*}\left(\frac{1}{\lambda_n}\frac{\pi_n}{p_n^*}\right) = 1. \tag{31}$$

References

1. Antoulas, A.C.: Approximation of Large Scale Dynamical Systems. SIAM, Philadelphia (1995)
2. Fanizza, G., Karlsson, J., Lindquist, A., Nagamune, R.: Passivity-preserving model reduction by analytic interpolation. Linear Algebra and its Applications 425, 608–633 (2007)

3. Fuhrmann, P.A.: Algebraic system theory: An analyst's point of view. J. Franklin Inst. 301, 521–540 (1976)
4. Fuhrmann, P.A.: Duality in polynomial models with some applications to geometric control theory. IEEE Trans. Autom. Control 26, 284–295 (1981)
5. Fuhrmann, P.A.: Polynomial models and algebraic stability criteria. In: Proc. Joint Workshop on Synthesis of Linear and Nonlinear Systems, Bielefeld, Germany, June 1981, pp. 78–90 (1981)
6. Fuhrmann, P.A.: A polynomial approach to Hankel norm and balanced approximations. Linear Algebra and its Appl. 146, 133–220 (1991)
7. Fuhrmann, P.A.: An algebraic approach to Hankel norm approximation problems. In: Markus Festschrift, L., Dekker, M. (eds.) Lecture Notes in Pure and Applied Mathematics, vol. 152, pp. 523–549 (1994)
8. Fuhrmann, P.A.: A Polynomial Approach to Linear Algebra. Springer, New York (1996)
9. Fuhrmann, P.A., Gombani, A.: On the Lyapunov equation, coinvariant subspaces and partial ordering of inner functions. Int. J. Control 73, 1129–1159 (2000)
10. Fuhrmann, P.A., Helmke, U.: Tensored polynomial models. Submitted to Linear Algebra and its Appl. (2009)
11. Fuhrmann, P.A., Ober, R.: A functional approach to LQG balancing, model reduction and robust control. Int. J. Control 57, 627–741 (1993)
12. Gallivan, K., Vandendorpe, A., Van Dooren, P.: Model reduction via truncation: an interpolation point of view. Linear Algebra and its Appl. 375, 115–134 (2003)
13. Gallivan, K., Vandendorpe, A., Van Dooren, P.: Sylvester equations and projection-based model reduction. J. Comp. Appl. Math. 162, 213–229 (2004)
14. Genin, Y., Vandendorpe, A.: On the embedding of state space realizations. Math. Control, Signals, and Systems 19, 123–149 (2007)
15. Glover, K.: All optimal Hankel-norm approximations and their L^∞-error bounds. Int. J. Control 39, 1115–1193 (1984)
16. Gragg, W.B., Lindquist, A.: On the partial realization problem. Linear Algebra and its Appl., Special Issue in Linear Systems and Control 50, 277–319 (1983)
17. Heinig, G., Rost, K.: Algebraic Methods for Toeplitz-like Matrices and Operators. Akademie-Verlag, Berlin (1984)
18. Helmke, U., Fuhrmann, P.A.: Bezoutians. Linear Algebra and its Appl. 122-124, 1039–1097 (1989)
19. Kalman, R.E.: Algebraic characterization of polynomials whose zeros lie in algebraic domains. Proc. Nat. Acad. Sci. 64, 818–823 (1969)
20. Peeters, R., Rapisarda, P.: Solution of polynomial Lyapunov and Sylvester equations. In: Hanzon, B., Hazewinkel, M. (eds.) Constructive Algebra and Systems Theory, Royal Netherlands Academy of Arts and Sciences, pp. 151–166 (2006)
21. Sorensen, D.C.: Passivity preserving model reduction via interpolation of spectral zeros. Systems & Control Lett. 54, 347–360 (2005)
22. Sorensen, D.C., Antoulas, A.C.: The Sylvester equation and approximate balanced reduction. Linear Algebra and its Appl. 352, 671–700 (2002)
23. de Souza, E., Bhattacharyya, S.P.: Controllability, observabiliy and the solution of $AX - XB = C$. Linear Algebra and its Appl. 39, 167–181 (1981)
24. Willems, J.C., Fuhrmann, P.A.: Stability theory for high order systems. Linear Algebra and its Appl. 167, 131–149 (1992)
25. Willems, J.C., Trentelman, H.L.: On quadratic differential forms. SIAM J. Control and Optim. 36, 1703–1749 (1998)

Quadratic Performance Verification for Boundary Value Systems

Hisaya Fujioka

Abstract. A quadratic performance of boundary value systems is considered. The problem of positive-definiteness test for a self-adjoint operator defined by a quadratic form is reduced to that for a finite matrix without approximations.

1 Introduction

It is well-known that the compression operator $\mathscr{G}_0$ of the form

$$\mathscr{G}_0 : \begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}x_0(t) = A_0x_0(t) + B_0u_0(t), \quad x_0(0) = 0, \\ \quad y_0(t) = C_0x_0(t) + D_0u_0(t) \end{cases}$$

defined on $\mathbf{L}_2[0, h]$ for $h > 0$ plays an important role in system and control theory. In particular, by lifting signals [13], a sampled-data system is transformed to a discrete-time time-invariant system with compression and related operators as coefficients of the state space representation. When the singular values of $\mathscr{G}_0$ are of interest, we face with $\mathscr{G}_0^*\mathscr{G}_0$ which is a special case of the boundary value system [9] of the form

$$\mathscr{G} : \begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}x(t) = Ax(t) + Bu(t), \quad \Omega x(0) + \Upsilon x(h) = 0, \\ \quad y(t) = Cx(t) + Du(t). \end{cases} \tag{1}$$

The study of boundary value systems has started in 80's. See, e.g., [9, 6, 7]. Applications of boundary value systems include a class of spatio-temporal systems [8] and lifted sampled-data systems [14, 10].

In this article we consider a quadratic performance of boundary value systems and derive a numerically tractable condition to verify it. The quadratic performance is a generalization of contractiveness studied in [3], and is defined by a quadratic form in the next section. A preliminary version of this article is found in conference proceedings [2] where a related optimization problem is also studied.

Notation: The number of elements of a finite set S is denoted by $|S|$.

Hisaya Fujioka
Kyoto Univeristy, Kyoto 606-8501, Japan
e-mail: fujioka@i.kyoto-u.ac.jp

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 189–196.
springerlink.com

2 Problem Statement

Consider a boundary value system

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = Ax(t) + Bu(t), \quad \Omega x(0) + \Upsilon x(h) = 0 \tag{2}$$

satisfying the well-posedness condition

$$\det(\Xi) \neq 0, \quad \Xi := \Omega + \Upsilon \mathrm{e}^{Ah}, \tag{3}$$

and the associated quadratic form ϕ: $\mathbf{L}_2[0, h] \to \mathbb{R}$

$$\phi(u) = \int_0^h \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}^* \begin{bmatrix} Q & S \\ S^* & R \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} \mathrm{d}t \tag{4}$$

where $h > 0$, and A, B, Ω, Υ, $Q = Q^*$, $R = R^*$, and S are matrices of compatible dimensions. In particular Ω and Υ are square. Note that (2) has a unique solution $x \in \mathbf{L}_2[0, h]$ for a given $u \in \mathbf{L}_2[0, h]$ if and only if (3) is satisfied [9].

The purpose of this article is to develop a numerically tractable method to check whether the following condition holds:

Condition 1. *There exists an* $\varepsilon > 0$ *satisfying*

$$\phi(u) \geq \varepsilon \|u\|_2^2$$

for all $u \in \mathbf{L}_2[0, h]$.

One can verify that $\|\mathscr{G}\| < \gamma$ for $\mathscr{G}$ in (1) if and only if Condition 1 holds with

$$\begin{bmatrix} Q & S \\ S^* & R \end{bmatrix} = \begin{bmatrix} C & D \\ 0 & I \end{bmatrix}^* \begin{bmatrix} -I & 0 \\ 0 & \gamma^2 I \end{bmatrix} \begin{bmatrix} C & D \\ 0 & I \end{bmatrix}.$$

It is required to verify Condition 1 with

$$\begin{bmatrix} Q & S \\ S^* & R \end{bmatrix} = \begin{bmatrix} C & D \\ 0 & I \end{bmatrix}^* \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} \begin{bmatrix} C & D \\ 0 & I \end{bmatrix}$$

in [5, Theorem 2] for an $\mathbf{H}_\infty$ design of PWM systems. Several other examples can be found in [4].

Remark 1. As opposed to the study of related properties to Condition 1 for infinite-horizon state-space systems such as disipativeness [12], no assumptions are made on problem data except (3). In particular, controllability of (A, B) is not assumed. Note that (A, B) can be uncontrollable even if (2) is minimal. See [7] for detail.

3 Main Results

We first point out the following necessary condition. The proof is obvious so it is omitted.

Property 1. Condition 1 is satisfied only if $R > 0$.

In order to state the main result of this article, we use the following matrices:

$$\hat{A} := \begin{bmatrix} A & 0 \\ -Q & -A^* \end{bmatrix}, \quad \hat{B} := \begin{bmatrix} B \\ -S \end{bmatrix}, \quad \hat{C} := \begin{bmatrix} S^* & B^* \end{bmatrix},$$

$$\hat{\Omega} := \mathrm{diag}(\Omega, \mathrm{e}^{A^*h}\Upsilon^*\Xi^{-*}), \quad \hat{\Upsilon} := \mathrm{diag}(\Upsilon, \Omega^*\Xi^{-*}\mathrm{e}^{A^*h}).$$

Hamiltonian matrices H and $H_{\min}$ are also defined by

$$H := \hat{A} - \hat{B}R^{-1}\hat{C}, \quad H_{\min} := \hat{A}_{\min} - \hat{B}_{\min}R^{-1}\hat{C}_{\min}$$

supposing that $R > 0$, where $(\hat{A}_{\min}, \hat{B}_{\min}, \hat{C}_{\min})$ is a minimal realization of the related transfer function P defined by:

$$P(s) := \hat{C}(sI - \hat{A})^{-1}\hat{B} + R.$$

The following theorem is the main result of this article:

Theorem 1. *Suppose that $R > 0$. Let $\theta \in [0, 2\pi)$ satisfy*

$$\mathrm{e}^{\mathrm{j}\theta} \notin \mathrm{eig}(\mathrm{e}^{Ah}), \quad \mathrm{e}^{\mathrm{j}\theta} \notin \mathrm{eig}(\mathrm{e}^{Hh}) \tag{5}$$

and define $Z \subset \mathbb{Z}$ by

$$Z := \{ i \in \mathbb{Z} : \ \exists\lambda_1, \lambda_2 \in \mathbb{R}, \mathrm{j}\lambda_1, \mathrm{j}\lambda_2 \in \mathrm{eig}(H_{\min}), \lambda_1 \le \omega_i \le \lambda_2 \}, \quad \omega_i := \frac{2\pi i + \theta}{h}.$$

Then Condition 1 *is satisfied if and only if*

$$\begin{bmatrix} K & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} L \\ V \end{bmatrix} M \begin{bmatrix} L^* & V^* \end{bmatrix} > 0$$

where matrices K, L, and M are defined by

$$K := \begin{bmatrix} P(\mathrm{j}\omega_{i_1}) & & 0 \\ & \ddots & \\ 0 & & P(\mathrm{j}\omega_{i_{|Z|}}) \end{bmatrix}, \quad L := \begin{bmatrix} \hat{C}(\mathrm{j}\omega_{i_1}I - \hat{A})^{-1} \\ \vdots \\ \hat{C}(\mathrm{j}\omega_{i_{|Z|}}I - \hat{A})^{-1} \end{bmatrix},$$

$$M := -\frac{1}{h}(\mathrm{e}^{\mathrm{j}\theta}I - \mathrm{e}^{\hat{A}h})(\hat{\Omega} + \Upsilon\mathrm{e}^{\hat{A}h})^{-1}(\mathrm{e}^{-\mathrm{j}\theta}\hat{\Omega} + \hat{\Upsilon})J, \quad J := \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}.$$

Here the j-th element of Z is denoted by i_j. The matrix V is given by the following factorization:

$$V^*V = W_{\mathbb{Z}} - W_Z, \quad W_Z := J\sum_{i\in Z}\left((\mathrm{j}\omega_i I - \hat{A})^{-1} - (\mathrm{j}\omega_i I - H)^{-1}\right),$$

$$W_{\mathbb{Z}} := \frac{1}{2}J\left((\mathrm{e}^{\mathrm{j}\theta}I - \mathrm{e}^{\hat{A}h})^{-1}(\mathrm{e}^{\mathrm{j}\theta}I + \mathrm{e}^{\hat{A}h}) - (\mathrm{e}^{\mathrm{j}\theta}I - \mathrm{e}^{Hh})^{-1}(\mathrm{e}^{\mathrm{j}\theta}I + \mathrm{e}^{Hh})\right).$$

The proof is found in the next section.

Remark 2. One can verify that M, $W_{\mathbb{Z}}$, and W_Z are all Hermitian, and $W_{\mathbb{Z}} - W_Z \geq 0$. See the proof for detail.

The following property provides a systematic way to determine θ satisfying (5). The proof is straightforward so it is omitted.

Property 2. Let θ_0 be defined by

$$\theta_0 := \arg\max_{\theta\in\Theta}\left(\min_{\lambda\in\Lambda}|\theta-\lambda|\right),$$

$$\Lambda := \left\{\lambda : \lambda \in \mathrm{eig}(\mathrm{e}^{Ah}) \cup \mathrm{eig}(\mathrm{e}^{Hh}),\ 1-\frac{1}{\sqrt{2}} \leq |\lambda| \leq 1+\frac{1}{\sqrt{2}}\right\},$$

$$\Theta := \left\{0, \frac{2\pi}{|\Lambda|+1}, \frac{2\pi\cdot 2}{|\Lambda|+1}, \ldots, \frac{2\pi|\Lambda|}{|\Lambda|+1}\right\}.$$

Then θ_0 satisfies

$$|\theta_0 - \lambda| \geq \min\left(\sin\frac{\pi}{|\Lambda|+1}, \frac{1}{\sqrt{2}}\right)$$

for all $\lambda \in \mathrm{eig}(\mathrm{e}^{Ah}) \cup \mathrm{eig}(\mathrm{e}^{Hh})$.

Note that θ_0 can be easily obtained by direct computation and comparison since $|\Theta| = |\Lambda| + 1 \leq |\mathrm{eig}(\mathrm{e}^{Ah}) \cup \mathrm{eig}(\mathrm{e}^{Hh})| + 1$. When all the matrices of the problem data are real-valued, one can find $\theta \in [0, \pi]$ satisfying (5) and Property 2 can be simplified.

4 Proof of Theorem 1

In this section the following notation will be used: For an operator $\mathscr{O}$ on a Hilbert space X we write $\mathscr{O} > 0$ if there exists a positive scalar $\varepsilon > 0$ satisfying

$$\langle \mathscr{O}x, x\rangle \geq \varepsilon\|x\|^2$$

for all $x \in X$. The boundary value system (1) is denoted by

$$\mathscr{G} \overset{\mathrm{BVS}}{=} (A, B, C, D; \Omega, \Upsilon).$$

Invoking formulas in [9, 6], one can verify that

$$\phi(u) = u^* \mathscr{P} u, \quad \mathscr{P} \overset{\mathrm{BVS}}{:=} (\hat{A}, \hat{B}, \hat{C}, R; \hat{\Omega}, \hat{\Upsilon}).$$

It is obvious that Condition 1 is equivalent to $\mathscr{P} > 0$ and $\mathscr{P}$ is self-adjoint by the construction of ϕ.

The basic procedure of the proof is a generalization of that in [1, 3]: Suppose that there exist $\tilde{n}, \tilde{m} \in \mathbb{N}$ and a unitary operator $\tilde{\mathscr{U}} \colon \mathbf{L}_2[0, h] \to \mathbb{C}^{\tilde{n}} \oplus \tilde{X}$ for a Hilbert space $\tilde{X}$ satisfying that

$$\tilde{\mathscr{U}} \mathscr{P} \tilde{\mathscr{U}}^* = \begin{bmatrix} \tilde{K} & 0 \\ 0 & \tilde{\mathscr{K}} \end{bmatrix} + \begin{bmatrix} \tilde{L} \\ \tilde{\mathscr{L}} \end{bmatrix} \tilde{M} \begin{bmatrix} \tilde{L}^* & \tilde{\mathscr{L}}^* \end{bmatrix}, \quad \tilde{\mathscr{K}} > 0 \tag{6}$$

where $\tilde{K} = \tilde{K}^* \colon \mathbb{C}^{\tilde{n}} \to \mathbb{C}^{\tilde{n}}$, $\tilde{\mathscr{K}} = \tilde{\mathscr{K}}^* \colon \tilde{X} \to \tilde{X}$, $\tilde{M} = \tilde{M}^* \colon \mathbb{C}^{\tilde{m}} \to \mathbb{C}^{\tilde{m}}$, $\tilde{L} \colon \mathbb{C}^{\tilde{n}} \to \mathbb{C}^{\tilde{m}}$, and $\tilde{\mathscr{L}} \colon \tilde{X} \to \mathbb{C}^{\tilde{m}}$. Then $\mathscr{P} > 0$ is equivalent to

$$\begin{bmatrix} \tilde{K} & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} \tilde{L} \\ \tilde{\mathscr{V}} \end{bmatrix} \tilde{M} \begin{bmatrix} \tilde{L}^* & \tilde{\mathscr{V}}^* \end{bmatrix} > 0, \quad \tilde{\mathscr{V}} := \tilde{\mathscr{K}}^{-\frac{1}{2}} \tilde{\mathscr{L}}. \tag{7}$$

By decomposing $\tilde{X}$ into $\ker(\tilde{\mathscr{V}}^*)$ and its orthogonal complement, (7) turns to

$$\begin{bmatrix} \tilde{K} & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} + \begin{bmatrix} \tilde{L} \\ 0 \\ \tilde{V} \end{bmatrix} \tilde{M} \begin{bmatrix} \tilde{L}^* & 0 & \tilde{V}^* \end{bmatrix} > 0$$

which is equivalent to

$$\begin{bmatrix} \tilde{K} & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} \tilde{L} \\ \tilde{V} \end{bmatrix} \tilde{M} \begin{bmatrix} \tilde{L}^* & \tilde{V}^* \end{bmatrix} > 0 \tag{8}$$

where $\tilde{V}$ is defined by the full rank factorization:

$$\tilde{V}^* \tilde{V} = \tilde{W}, \quad \tilde{W} := \tilde{\mathscr{L}}^* \tilde{\mathscr{K}}^{-1} \tilde{\mathscr{L}}.$$

Noting that $\tilde{W} \colon \mathbb{C}^{\tilde{m}} \to \mathbb{C}^{\tilde{m}}$, (8) is a finite dimensional condition. It is not difficult to see that one can replace $\tilde{V}$ by any matrix factorization of $\tilde{W}$.

In the sequel we will show that one can set $\tilde{K} = K$, $\tilde{L} = L$, $\tilde{M} = M$, and $\tilde{W} = W_{\mathbb{Z}} - W_Z$. Note that one can confirm that $P(\mathrm{j}\omega_i)$ and M are Hermitian.

Following [1], let us consider $\mathscr{U} \colon \mathbf{L}_2[0, h] \to \ell_2(\mathbb{Z})$ mapping $f \mapsto \varphi$ defined by

$$\varphi_i := \frac{1}{\sqrt{h}} \int_0^h \mathrm{e}^{-\mathrm{j}\omega_i t} f(t) \, \mathrm{d}t$$

as a candidate of $\tilde{\mathscr{U}}$. It is obvious that $\mathscr{U}$ is unitary, and hence the following lemma shows that $\mathscr{U}$ satisfies the first condition of (6), i.e., $\mathscr{U} \mathscr{P} \mathscr{U}^*$ can be expressed as a sum of a block diagonal and a finite rank operators:

Lemma 1. *For $v \in \ell_2(\mathbb{Z})$, one has*

$$(\mathscr{U} \mathscr{P} \mathscr{U}^* v)_i = P(\mathrm{j}\omega_i) v_i + \hat{C}(\mathrm{j}\omega_i I - \hat{A})^{-1} M \sum_{j \in \mathbb{Z}} (\mathrm{j}\omega_j I - \hat{A})^{-*} \hat{C}^* v_j.$$

Proof. Consider a boundary value system $\mathscr{F}_{i,j}$ defined by

$$\mathscr{F}_{i,j} := \mathscr{E}_i^* \mathscr{P} \mathscr{E}_j, \quad \mathscr{E}_i :\overset{\mathrm{BVS}}{=} (\mathrm{j}\omega_i I, I, I, 0; I, 0).$$

Invoking formulas in [9, 6], one can verify that

$$\mathscr{F}_{i,j} \overset{\mathrm{BVS}}{=} (\bar{A}_{i,j}, \bar{B}, \bar{C}, 0; \bar{\Omega}, \bar{\Upsilon}), \quad \bar{\Omega} := \mathrm{diag}(0, \hat{\Omega}, I), \quad \bar{\Upsilon} := \mathrm{diag}(I, \hat{\Upsilon}, 0),$$

$$\bar{A}_{i,j} := \begin{bmatrix} \mathrm{j}\omega_i I & \hat{C} & R \\ 0 & \hat{A} & \hat{B} \\ 0 & 0 & \mathrm{j}\omega_j I \end{bmatrix}, \quad \bar{B} := \begin{bmatrix} 0 \\ 0 \\ I \end{bmatrix}, \quad \bar{C} := \begin{bmatrix} -I & 0 & 0 \end{bmatrix}.$$

Noting that the impulse response to the j-th input channel of $\mathscr{E}_i$ is given by

$$u_i(t) = \mathrm{e}^{\mathrm{j}\omega_i t} e_j$$

where e_j denote the j-th standard basis,

$$(\mathscr{U} \mathscr{P} \mathscr{U}^* v)_i = \frac{1}{h} \sum_{j \in \mathbb{Z}} F_{i,j}(0,0) v_j.$$

Here $F_{i,j}$ is the kernel of $\mathscr{F}_{i,j}$:

$$(\mathscr{F}_{i,j} v)(t) = \int_0^h F_{i,j}(t,\tau) v(\tau)\, \mathrm{d}\tau.$$

From the formula in [9, Eq. (2.7)], $F_{i,j}(0,0)$ is given by

$$F_{i,j}(0,0) = \bar{C}(\bar{\Omega} + \bar{\Upsilon} \mathrm{e}^{\bar{A}_{i,j} h})^{-1} \bar{\Omega} \bar{B}.$$

After manipulations with the equality

$$(\mathrm{j}\omega_j I - \hat{A})^{-1} \hat{B} = J (\mathrm{j}\omega_j I - \hat{A})^{-*} \hat{C}^*,$$

one finally gets

$$F_{i,j}(0,0) = h \left(\delta_{i,j} P(\mathrm{j}\omega_i) + \hat{C}(\mathrm{j}\omega_i I - \hat{A})^{-1} M (\mathrm{j}\omega_j I - \hat{A})^{-*} \hat{C}^* \right).$$

This completes the proof. □

The next property shows that $\mathscr{U}$ satisfies the second condition of (6) as well:

Property 3. Suppose $R > 0$ and define ω_i and Z as in Theorem 1. One has

$$P(\mathrm{j}\omega_i) > 0$$

for all $i \in \mathbb{Z} \setminus Z$.

Proof. One can see that $P(\mathrm{j}\omega) > 0$ for $\omega \in \mathbb{R}$ of sufficiently large absolute values since $R > 0$. Hence attentions should be paid for $\omega \in \mathbb{R}$ with which $P(\mathrm{j}\omega)$ is singular.

Non-singularity of R implies that $\mathrm{j}\omega \in \mathrm{eig}(H_{\min})$ is satisfied for $\omega \in \mathbb{R}$ if $P(\mathrm{j}\omega)$ is singular. See, e.g., [16, Lemma 13.15]. Hence the statement is trivial if $H_{\min}$ does not have eigenvalues on the imaginary axis. Otherwise let $\bar{\omega}_{\min}$, $\bar{\omega}_{\max} \in \mathbb{R}$ be the minimal and the maximal element of the set

$$\{\lambda \in \mathbb{R} : \mathrm{j}\lambda \in \mathrm{eig}(H_{\min})\}.$$

Then $P(\mathrm{j}\omega) > 0$ is satisfied for all $\omega \in \mathbb{R}$ satisfying either $\omega < \bar{\omega}_{\min}$ or $\omega > \bar{\omega}_{\max}$. Since $\bar{\omega}_{\min} \leq \omega_i \leq \bar{\omega}_{\max}$ if $i \in Z$, the statement is satisfied. □

Thus we can take K, L, M, and $\ell_2(\mathbb{Z} \setminus Z)$ as $\tilde{K}$, $\tilde{L}$, $\tilde{M}$, and $\tilde{X}$ respectively. Finally we see that

$$\begin{aligned}\tilde{W} &= \sum_{i\in\mathbb{Z}\setminus Z} (\mathrm{j}\omega_i I - \hat{A})^{-*}\hat{C}^*(P(\mathrm{j}\omega_i))^{-1}\hat{C}(\mathrm{j}\omega_i I - \hat{A})^{-1} \\ &= -J \sum_{i\in\mathbb{Z}\setminus Z} (\mathrm{j}\omega_i I - \hat{A})^{-1}\hat{B}(P(\mathrm{j}\omega_i))^{-1}\hat{C}(\mathrm{j}\omega_i I - \hat{A})^{-1} \\ &= J \sum_{i\in\mathbb{Z}\setminus Z} ((\mathrm{j}\omega_i I - H)^{-1} - (\mathrm{j}\omega_i I - \hat{A})^{-1}) \\ &= J \sum_{i\in\mathbb{Z}} ((\mathrm{j}\omega_i I - H)^{-1} - (\mathrm{j}\omega_i I - \hat{A})^{-1}) - W_Z\end{aligned}$$

The first term turns to $W_{\mathbb{Z}}$ by invoking Proposition 5 in [1]. This completes the proof of Theorem 1.

5 Concluding Remarks

A quadratic performance of boundary value systems has been considered. It has been shown that the quadratic performance verification can be reduced to checking eigenvalues of a matrix without approximations.

Related open issues include derivation of computationally cheaper sufficient conditions and applications to analysis and design problems, e.g., gain-scheduled control synthesis for sampled-data systems as partially discussed in [4].

References

1. Dullerud, G.E.: Computing the L_2-induced norm of a compression operator. Systems & Control Letters 37, 87–91 (1999)
2. Fujioka, H.: Quadratic performance analysis for finite-horizon systems. In: Proc. 16th IFAC World Congress (2005)
3. Fujioka, H.: Computing $\mathbf{L}_2$-gain of finite-horizon systems with boundary conditions. IEEE Trans. Autom. Control 52, 697–702 (2007)
4. Fujioka, H., Jönsson, U.T.: Characterizing uncertain time-varying parameters with periodic reset. In: Proc. 2009 American Control Conf., pp. 3754–3756 (2009)
5. Fujioka, H., Kao, C.-Y., Almér, S., Jönsson, U.T.: Robust tracking with $\mathbf{H}_\infty$ performance for PWM systems. Automatica 45, 1808–1818 (2009)

6. Gohberg, I., Kaashoek, M.A.: Time varying linear systems with boundary conditions and integral operators — I: the transfer operator and its properties. Integral Equations and Operator Theory 7, 325–391 (1984)
7. Gohberg, I., Kaashoek, M.A., Lerer, L.: Minimality and irreducibility of time-invariant linear boundary-value systems. Int. J. Control 44, 363–379 (1986)
8. Jovanović, M.R., Bamieh, B.: A formula for frequency responses of distributed systems with one spatial variable. Systems & Control Letters 55, 27–37 (2006)
9. Krener, A.J.: Boundary value systems. Astérisque 75-76, 149–165 (1980)
10. Mirkin, L.: Transfer function of sampled-data systems in the lifted domain. In: Proc. 44th IEEE Conf. Decision and Control and European Control Conf., pp. 5180–5185 (2005)
11. Mirkin, L., Palmor, Z.J.: A new representation of the parameters of lifted systems. IEEE Trans. Autom. Control 44, 833–840 (1999)
12. Willems, J.C.: Dissipative dynamical systems, Part II: linear systems with quadratic supply rates. Arch. Rat. Mech. Anal. 45, 352–393 (1972)
13. Yamamoto, Y.: A function space approach to sampled-data control systems and tracking problems. IEEE Trans. Autom. Control 44, 703–712 (1994)
14. Yamamoto, Y., Khargonekar, P.P.: Frequency response of sampled-data systems. IEEE Trans. Autom. Control 41, 166–176 (1996)
15. Zhou, K., Khargonekar, P.P.: On the weighted sensitivity minimization problem for delay systems. Systems & Control Letters 8, 307–312 (1987)
16. Zhou, K., Doyle, J.C., Glover, K.: Robust and Optimal Control. Prentice-Hall, Englewood Cliffs (1996)

Lyapunov Stability Analysis of Higher-Order 2-D Systems

Chiaki Kojima, Paolo Rapisarda, and Kiyotsugu Takaba

Abstract. We prove a necessary and sufficient condition for the asymptotic stability of a 2-D system described by a system of higher-order linear partial difference equations. We use the definition of asymptotic stability given by Valcher in "Characteristic Cones and Stability Properties of Two-Dimensional Autonomous Behaviors", *IEEE Trans. Circ. Syst. — Part I: Fundamental Theory and Applications*, vol. 47, no. 3, pp. 290–302, 2000. This property is shown to be equivalent to the existence of a vector Lyapunov functional satisfying certain positivity conditions together with its divergence along the system trajectories. We use the behavioral framework and the calculus of quadratic difference forms based on four variable polynomial algebra.

1 Introduction

Discrete- and continuous-time two-dimensional (in the following abbreviated as 2-D) systems have application in all those situations when the evolution of the to-be-modeled system depends on two independent variables. In this paper we adopt the behavioral framework pioneered by J. C. Willems in the 1-D case (see [13]),

Chiaki Kojima
Department of Information Physics and Computing, Graduate School of Information Science and Technology, The University of Tokyo, Hongo, Bunkyo-Ku, Tokyo 113-0033, Japan
e-mail: chiaki_kojima@ipc.i.u-tokyo.ac.jp

Paolo Rapisarda
Information: Signals, Images, Systems group, School of Electronics and Computer Science, University of Southampton, SO171BJ Southampton, United Kingdom
e-mail: pr3@ecs.soton.ac.uk

Kiyotsugu Takaba
Dept. of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University, Kyoto 606–8501, Japan
e-mail: takaba@amp.i.kyoto-u.ac.jp

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 197–206.
springerlink.com

and extended to the 2-D case by P. Rocha (see [14]) and other authors. In this setting the main object of study is the behavior, the set consisting of all the trajectories admissible by the physical laws describing the system trajectories.

The notion of stability, because of its important consequences in the analysis and design of control systems and of filters, has attracted considerable interest also in the case of 2-D systems. The issue of what the correct definition of stability is for this situation presents first and foremost the difficulty of extending the notion of "past" and "future", self-evident in the 1-D framework, to the case of two independent variables, where there is no obvious such splitting of the independent variables domain. An eminently reasonable position is to let the laws describing the physical phenomenon themselves dictate what the direction is of the evolution of the system. This is the approach pioneered by M. E. Valcher in [15] and followed in this paper.

In this paper we present a necessary and sufficient condition for the asymptotic stability of 2-D systems based on Lyapunov functions. This idea is by no means original, having been applied already in [10, 1]; however, those approaches relied entirely on a specific ("state-space") type of representation of the system, while we deal with systems described in a general form, namely as the solutions of a system of partial difference equations. Moreover, the "generalized Bézoutian" introduced in [4] is shown in this paper to be the scalar version of a generalized Bézoutian arising naturally as a Lyapunov function for 2-D systems.

Sections 2 and 3 of this paper contain background material on 2-D systems and quadratic difference forms, respectively. Section 4 contains the main result of this paper, namely a stability criterion for higher-order systems of difference equations based on Lyapunov analysis.

In this paper, the concepts and tools of the behavioral approach, and of quadratic difference forms will be put to strenuous use. The reader not familiar with them is referred to [12, 13, 14, 16] for a thorough exposition.

Notation: We denote with $\mathbb{R}^{\mathtt{r}\times\mathtt{w}}[\xi_1,\xi_2]$ (respectively, $\mathbb{R}^{\mathtt{r}\times\mathtt{w}}[\xi_1,\xi_2,\xi_1^{-1},\xi_2^{-1}]$) the set of all $\mathtt{r}\times\mathtt{w}$ matrices with entries in the ring $\mathbb{R}[\xi_1,\xi_2]$ of polynomials in 2 indeterminates, with real coefficients (respectively in the ring $\mathbb{R}[\xi_1,\xi_2,\xi_1^{-1},\xi_2^{-1}]$ of Laurent polynomials in 2 indeterminates with real coefficients). Given a nonzero Laurent polynomial $p(\xi_1,\xi_2)=\sum_{\ell,m}p_{\ell,m}\xi_1^\ell\xi_2^m\in\mathbb{R}[\xi_1,\xi_2,\xi_1^{-1},\xi_2^{-1}]$, the *Laurent variety* of p is defined as

$$\mathscr{V}_L(p):=\{(\alpha,\beta)\in\mathbb{C}\times\mathbb{C}\mid\alpha\beta\neq0,p(\alpha,\beta)=0\}$$

This definition extends to sets $\mathscr{I}$ of Laurent polynomials, with $\mathscr{V}(\mathscr{I})$ being the intersection of the Laurent varieties of all polynomials in the set. For $R\in\mathbb{R}^{\mathtt{r}\times\mathtt{w}}[\xi_1,\xi_2,\xi_1^{-1},\xi_2^{-1}]$, the *characteristic ideal* is the ideal of $\mathbb{R}[\xi_1,\xi_2]$ generated by the determinants of all $\mathtt{w}\times\mathtt{w}$ minors of R, and the *characteristic variety* is the set of roots common to all polynomials in the ideal. Further properties and definitions can be found in [3].

A set $\mathscr{K}\subset\mathbb{R}\times\mathbb{R}$ is called a *cone* if $\alpha\mathscr{K}\subset\mathscr{K}$ for all $\alpha\geq0$. A cone is *solid* if it contains an open ball in $\mathbb{R}\times\mathbb{R}$, and *pointed* if $\mathscr{K}\cap-\mathscr{K}=\{(0,0)\}$. A cone is *proper* if it is closed, pointed, solid, and convex. It is easy to see that a proper cone is uniquely identified as the set of nonnegative linear combinations of two linearly independent vectors $v_1, v_2 \in \mathbb{R}^2$. In the following we will often consider the intersection of a cone $\mathscr{K}$ with $\mathbb{Z} \times \mathbb{Z}$; whenever it will be clear from the context, we will be denoting this set with $\mathscr{K}$ instead of with $\mathscr{K} \cap \mathbb{Z} \times \mathbb{Z}$.

We denote with $\overline{\mathscr{P}_1}$ the closed unit polydisk:

$$\overline{\mathscr{P}_1} := \{(\alpha, \beta) \in \mathbb{C} \times \mathbb{C} \mid |\alpha| \leq 1, |\beta| \leq 1\}$$

Given a set $\mathscr{S} \subset \mathbb{Z} \times \mathbb{Z}$, its *(discrete) convex hull* is the intersection of the convex hull of $\mathscr{S}$ (seen as a subset of $\mathbb{R} \times \mathbb{R}$) and of $\mathbb{Z} \times \mathbb{Z}$. In the following we will also refer to the (discrete) convex hull associated with an element $p \in \mathbb{R}[\xi_1, \xi_2, \xi_1^{-1}, \xi_2^{-1}]$, meaning the (discrete) convex hull of the *support of* p, i.e. the set

$$\begin{aligned}\operatorname{supp}(p) := \{(x_1, x_2) \in \mathbb{Z} \times \mathbb{Z} \mid \text{ the coefficient of } \xi_1^h \xi_2^k \\ \text{in } p(\xi_1, \xi_2, \xi_1^{-1}, \xi_2^{-1}) \text{ is } \neq 0\}\end{aligned}$$

We denote with $\mathbb{W}^{\mathbb{T}}$ the set consisting of all trajectories from $\mathbb{T}$ to $\mathbb{W}$. We denote with σ_1, σ_2 the *shift operators* defined as

$$\begin{aligned}\sigma_i : (\mathbb{R}^{\mathtt{w}})^{\mathbb{Z}^2} &\to (\mathbb{R}^{\mathtt{w}})^{\mathbb{Z}^2} \quad i = 1, 2 \\ (\sigma_1 w)(x_1, x_2) &:= w(x_1 - 1, x_2) \\ (\sigma_2 w)(x_1, x_2) &:= w(x_1, x_2 - 1)\end{aligned}$$

2 2-D Behaviors

We call $\mathfrak{B}$ a *linear discrete-time complete 2-D behavior* if it is the subset of $(\mathbb{R}^{\mathtt{w}})^{\mathbb{Z}^2}$ consisting of all solutions to

$$R(\sigma_1, \sigma_2) w = 0 \tag{1}$$

where $R \in \mathbb{R}^{\mathtt{r} \times \mathtt{w}}[\xi_1, \xi_2, \xi_1^{-1}, \xi_2^{-1}]$. We call (1) a *kernel representation* of $\mathfrak{B}$. The set of all such behaviors is denoted with $\mathscr{L}_2^{\mathtt{w}}$.

$\mathfrak{B} \in \mathscr{L}_2^{\mathtt{w}}$ is *autonomous* if there exists a proper cone $\mathscr{K} \subset \mathbb{R} \times \mathbb{R}$ such that

$$\left[w_1, w_2 \in \mathfrak{B} \text{ and } w_{1|\mathscr{K} \cap \mathbb{Z} \times \mathbb{Z}} = w_{2|\mathscr{K} \cap \mathbb{Z} \times \mathbb{Z}}\right] \Longrightarrow [w_1 = w_2]$$

Such a cone $\mathscr{K} \cap \mathbb{Z} \times \mathbb{Z}$ will be called a proper *characteristic* cone for $\mathfrak{B}$. Observe that if $w \in \mathfrak{B}$ is such that $w_{|\mathscr{K} \cap \mathbb{Z} \times \mathbb{Z}} = 0$, then $w = 0$. The following result holds.

Theorem 1. *Let $\mathfrak{B} \in \mathscr{L}_2^{\mathtt{w}}$ be autonomous, and let $\mathfrak{B} = \ker\ R(\sigma_1, \sigma_2)$ for some $R \in \mathbb{R}^{\mathtt{r} \times \mathtt{w}}[\xi_1, \xi_2, \xi_1^{-1}, \xi_2^{-1}]$. Then there exist $H \in \mathbb{R}^{\bullet \times \bullet}[\xi_1, \xi_2, \xi_1^{-1}, \xi_2^{-1}]$ right factor prime, and $\Delta \in \mathbb{R}^{\mathtt{w} \times \mathtt{w}}[\xi_1, \xi_2, \xi_1^{-1}, \xi_2^{-1}]$ nonsingular, such that $R = H \cdot \Delta$.*

Moreover, denote $\delta := \det(\Delta) \in \mathbb{R}[\xi_1, \xi_2, \xi_1^{-1}, \xi_2^{-1}]$. The following statements are equivalent:

1. *The proper cone $\mathscr{K}$ is characteristic for $\mathfrak{B}$;*
2. *The proper cone $\mathscr{K}$ is characteristic for $\ker\ \Delta(\sigma_1, \sigma_2)$;*

3. *The proper cone $\mathscr{K}$ is characteristic for* ker $\delta(\sigma_1,\sigma_2)$*;*
4. *The discrete convex hull $\mathscr{H}_\delta$ of δ satisfies the following two conditions:*

   4*a.* $-\mathscr{H}_\delta \subset \mathscr{K}$*;*
   4*.b.* $-\mathscr{H}_\delta \subset \mathscr{K}$ *intersects the generating lines of $\mathscr{K}$ only in* $(0,0)$.

It can be shown (see [2]) that if $\mathfrak{B} \in \mathscr{L}_2^{\mathtt{w}}$ is such that $\mathfrak{B} = \ker\ R(\sigma_1,\sigma_2)$ for some right factor prime matrix $R \in \mathbb{R}^{\mathtt{r}\times\mathtt{w}}[\xi_1,\xi_2,\xi_1^{-1},\xi_2^{-1}]$, then $\mathfrak{B}$ is autonomous and finite-dimensional; then (see Lemma 2.4 of [15]) every proper cone is characteristic.

If $\mathfrak{B}$ is autonomous, and $\mathfrak{B} = \ker\ R(\sigma_1,\sigma_2)$ for some nonsingular Laurent matrix R, then $\mathfrak{B}$ is called a *square autonomous behavior*. Observe that Theorem 1 shows that for any autonomous behavior $\mathfrak{B}$ whose kernel representation can be factored as $H\Delta$ with H right factor prime and Δ nonsingular, the characteristic cone is determined by its "square autonomous part" ker $\Delta(\sigma_1,\sigma_2)$.

We now introduce the concept of stability introduced by Valcher in [15]. We examine the finite-dimensional case first.

Definition 1. Let $\mathfrak{B} \in \mathscr{L}_2^{\mathtt{w}}$ be autonomous and finite-dimensional, and let $\mathscr{K}$ be any proper cone of $\mathbb{Z}\times\mathbb{Z}$. $\mathfrak{B}$ is *$\mathscr{K}$-stable* if

$$[w \in \mathfrak{B}] \Longrightarrow \left[\lim_{\substack{(i,j)\in\mathscr{K} \\ |i|+|j|\to+\infty}} \|w(i,j)\| = 0 \right]$$

The following algebraic characterization of finite-dimensional stable behaviors (see [15, Theorem 3.3, p. 297]) holds. In order to avoid cumbersome details, we follow [15], and only consider proper cones generated by unimodular integer matrices, which are then isomorphic to the first orthant of $\mathbb{Z}\times\mathbb{Z}$, in the sense that there exists a (linear, bijective) transformation $T : \mathbb{Z}^2 \to Z^2$ such that $T(\mathscr{K})$ is the first orthant.

Theorem 2. *Let* $\mathfrak{B} = \ker\ H(\sigma_1,\sigma_2)$*, with*

$$H(\xi_1,\xi_2) = \sum_{\ell,m} H_{\ell,m}\xi_1^{\ell}\xi_2^{m} \in \mathbb{R}^{\bullet\times\mathtt{w}}[\xi_1,\xi_2,\xi_1^{-1},\xi_2^{-1}]$$

right factor prime (see [3] *for the definition), and let $\mathscr{K}$ be a proper cone isomorphic to the first orthant. Denote with T the transformation mapping $\mathscr{K}$ to the first orthant, and denote with $(t_1(\ell,m),t_2(\ell,m))$ the image of $(\ell,m) \in \mathbb{Z}\times\mathbb{Z}$ under T. Define*

$$H_T(\xi_1,\xi_2) := \sum_{\ell,m} H_{\ell,m}\xi_1^{t_2(\ell,m)}\xi_2^{t_1(\ell,m)}$$

Then the following two statements are equivalent:

1. $\mathfrak{B}$ *is $\mathscr{K}$-stable;*
2. *Every (α,β) in the Laurent variety of the maximal order minors of H_T satisfies* $|\alpha| > 1$ *and* $|\beta| > 1$.

In order to state the definition of stability for the square case, we need to introduce the following notation: given a proper cone $\mathscr{C}$, we denote with $\delta(\mathscr{C})$ its *boundary*, i.e. the generating lines of $\mathscr{C}$. We denote with $(\delta(\mathscr{C}))^n$ the set consisting of the points of $\mathbb{Z} \times \mathbb{Z}$ whose distance from $\delta(\mathscr{C})$ is less than n:

$$(\delta(\mathscr{C}))^n := \{(i,j) \in \mathbb{Z} \times \mathbb{Z} \mid \min\{|i-h|+|j-k| \text{ with } (x_1,x_2) \in \delta(\mathscr{C})\} \leq n\}$$

Definition 2. Let $\mathscr{K}$ be a proper cone such that $-\mathscr{K}$ is characteristic for a square autonomous behavior $\mathfrak{B} \in \mathscr{L}_2^{\mathtt{w}}$. $\mathfrak{B}$ is $\mathscr{K}$-stable if there exists $n \in \mathbb{N}$, $n > 0$ such that

$$[w \in \mathfrak{B}, w \text{ bounded in } (\delta(-\mathscr{K}))^n] \Longrightarrow \left[\lim_{\substack{(i,j) \in \mathscr{K} \\ |i|+|j| \to +\infty}} \|w(i,j)\| = 0 \right]$$

The following is Theorem 3.6 of [15].

Theorem 3. *Let* $\mathfrak{B} = \ker \Delta(\sigma_1, \sigma_2)$ *be a square autonomous behavior, and let* $\mathscr{K}$ *be a proper cone for* $\mathfrak{B}$ *which is* T*-isomorphic to the first orthant. Denote* $\delta := \det(\Delta)$*, and assume without loss of generality that* $\mathscr{H}_\delta \subset \mathscr{K}$ *and that* $\mathscr{H}_\delta \cap \delta\mathscr{K} = \{(0,0)\}$*. Denote with* $(t_1(\ell,m), t_2(\ell,m))$ *the image of* $(\ell,m) \in \mathbb{Z} \times \mathbb{Z}$ *under* T*. Define*

$$\Delta_T(\xi_1,\xi_2) := \sum_{\ell,m} \Delta_{\ell,m} \xi_1^{t_1(\ell,m)} \xi_2^{t_2(\ell,m)}$$

Then the following two statements are equivalent:

1. $\mathfrak{B}$ *is* $\mathscr{K}$*-stable;*
2. *The Laurent variety of* $\det \Delta_T$ *does not intersect the closed unit polydisk* $\overline{\mathscr{P}_1}$.

3 Bilinear and Quadratic Difference Forms for 2-D Systems

In order to represent bilinear and quadratic functionals of the variables of continuous-time 2-D systems, 4-variable polynomial matrices are used (see [12]). We now illustrate quadratic difference forms for 2-D discrete systems; some preliminary results are in [9].

In order to simplify the notation, define the multi-indices $\mathbf{k} := (k_1,k_2)$, $\mathbf{l} := (l_1,l_2)$, and the notation $\zeta := (\zeta_1,\zeta_2)$ and $\eta := (\eta_1,\eta_2)$, and $\zeta^{\mathbf{k}}\eta^{\mathbf{l}} := \zeta_1^{k_1}\zeta_2^{k_2}\eta_1^{l_1}\eta_2^{l_2}$. Let $\mathbb{R}^{\mathtt{w}_1 \times \mathtt{w}_2}[\zeta,\eta]$ denote the set of real polynomial $\mathtt{w}_1 \times \mathtt{w}_2$ matrices in the 4 indeterminates ζ_i and η_i, $i = 1,2$; that is, an element of $\mathbb{R}^{\mathtt{w}_1 \times \mathtt{w}_2}[\zeta,\eta]$ is of the form

$$\Phi(\zeta,\eta) = \sum_{\mathbf{k},\mathbf{l}} \Phi_{\mathbf{k},\mathbf{l}} \zeta^{\mathbf{k}} \eta^{\mathbf{l}}$$

where $\Phi_{\mathbf{k},\mathbf{l}} \in \mathbb{R}^{\mathtt{w}_1 \times \mathtt{w}_2}$; the sum ranges over the nonnegative multi-indices $\mathbf{k}$ and $\mathbf{l}$, and is assumed to be finite. This matrix induces a *bilinear difference form* (*BDF* in the following) L_Φ

$$L_\Phi : (\mathbb{R}^{\mathtt{w}_1})^{\mathbb{Z}^2} \times (\mathbb{R}^{\mathtt{w}_2})^{\mathbb{Z}^2} \longrightarrow (\mathbb{R})^{\mathbb{Z}^2}$$
$$L_\Phi(v,w) := \sum_{\mathbf{k},\mathbf{l}} (\sigma^{\mathbf{k}} v)^\top \Phi_{\mathbf{k},\mathbf{l}} (\sigma^{\mathbf{l}} w)$$

where the **k**-th shift operator $\sigma^{\mathbf{k}}$ is defined as $\sigma^{\mathbf{k}} := \sigma_1^{k_1}\sigma_2^{k_2}$, and analogously for $\sigma^{\mathbf{l}}$.

The 4-variable polynomial matrix $\Phi(\zeta_1,\zeta_2,\eta_1,\eta_2)$ is called *symmetric* if $\mathtt{w}_1 = \mathtt{w}_2 =: \mathtt{w}$ and $\Phi(\zeta_1,\zeta_2,\eta_1,\eta_2) = \Phi(\eta_1,\eta_2,\zeta_1,\zeta_2)^\top$, concisely written as $\Phi(\zeta,\eta) = \Phi(\eta,\zeta)^\top$. In this case, Φ induces also a quadratic functional

$$Q_\Phi : (\mathbb{R}^{\mathtt{w}})^{\mathbb{Z}^2} \longrightarrow (\mathbb{R})^{\mathbb{Z}^2}$$
$$Q_\Phi(w) := L_\Phi(w,w)$$

We will call Q_Φ the *quadratic difference form* (in the following abbreviated with QDF) associated with the four-variable polynomial matrix Φ.

In this paper we also consider vectors $\Psi \in (\mathbb{R}^{\mathtt{w}_1\times\mathtt{w}_2}[\zeta,\eta])^2$, i.e.

$$\Psi(\zeta,\eta) = \begin{bmatrix} \Psi_1(\zeta,\eta) \\ \Psi_2(\zeta,\eta) \end{bmatrix} =: \mathrm{col}(\Psi_i(\zeta,\eta))_{i=1,2}$$

with $\Psi_i \in \mathbb{R}^{\mathtt{w}_1\times\mathtt{w}_2}[\zeta,\eta]$ and with $\mathrm{col}(A_i)_{i=1,2}$ the matrix obtained by stacking the two matrices A_i, both with the same number of columns, on top of each other. Such Ψ induces a vector bilinear difference form (abbreviated *VBDF*), defined as

$$L_\Psi : (\mathbb{R}^{\mathtt{w}_1})^{\mathbb{Z}^2} \times (\mathbb{R}^{\mathtt{w}_2})^{\mathbb{Z}^2} \longrightarrow (\mathbb{R}^2)^{\mathbb{Z}^2}$$
$$L_\Psi(v,w) := \begin{bmatrix} L_{\Psi_1}(v,w) \\ L_{\Psi_2}(v,w) \end{bmatrix}.$$

Finally, we introduce the notion of (discrete) divergence of a VBDF. Given a VBDF $L_\Psi = \mathrm{col}(L_{\Psi_1}, L_{\Psi_2})^\top$, we define its *divergence* as the BDF defined by

$$\begin{aligned}(\mathrm{div}\, L_\Psi)(w_1,w_2) :=\ & \big(L_{\Psi_1}(w_1,w_2) - \sigma_1(L_{\Psi_1}(w_1,w_2))\big) \\ +\ & \big(L_{\Psi_2}(w_1,w_2) - \sigma_2(L_{\Psi_2}(w_1,w_2))\big)\end{aligned} \qquad (2)$$

for all w_1,w_2. It is straightforward to verify that in terms of the 4-variable polynomial matrices associated with the BDF's, the relationship between a VBDF L_Ψ and its divergence $L_\Phi = \mathrm{div}\, L_\Psi$ is expressed as

$$\Phi(\zeta_1,\zeta_2,\eta_1,\eta_2) = (1-\zeta_1\eta_1)\Psi_1(\zeta_1,\zeta_2,\eta_1,\eta_2) + (1-\zeta_2\eta_2)\Psi_2(\zeta_1,\zeta_2,\eta_1,\eta_2)$$

In order to characterize those BDFs which are the divergence of some VBDF, we need to introduce the "del" operator, defined as

$$\partial : \mathbb{R}^{\mathtt{w}_1\times\mathtt{w}_2}[\zeta_1,\zeta_2,\eta_1,\eta_2] \longrightarrow \mathbb{R}^{\mathtt{w}_1\times\mathtt{w}_2}[\xi_1,\xi_2,\xi_1^{-1},\xi_2^{-1}]$$
$$\partial\Phi(\xi_1,\xi_2) := \Phi(\xi_1^{-1},\xi_2^{-1},\xi_1,\xi_2)$$

The following result holds true.

Proposition 1. *A BDF L_Φ is the divergence of some VBDF L_Ψ if and only if $\partial\Phi(\xi_1,\xi_2)=0$.*

Proof. Necessity is straightforward. Sufficiency can be proved using a Gröbner basis argument, which can be extended entrywise to polynomial matrices. □

The definition and properties described above can be adapted to a vector quadratic difference form (VQDF) in a obvious manner.

A QDF Q_Δ induced by $\Delta\in\mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta_1,\zeta_2,\eta_1,\eta_2]$ is *nonnegative* if $Q_\Delta(w(x_1,x_2))\geq 0\ \forall(x_1,x_2)\in\mathbb{Z}^2$ and $\forall\, w\in(\mathbb{R}^{\mathtt{w}})^{\mathbb{Z}^2}$. This will be denoted with $Q_\Delta\geq 0$ or $\Delta(\zeta,\eta)\geq 0$. We call Q_Δ positive, denoted $Q_\Delta>0$ or $\Delta(\zeta,\eta)>0$, if $Q_\Delta\geq 0$ and $Q_\Delta(w(x_1,x_2))=0\ \forall(x_1,x_2)\in\mathbb{Z}^2$ implies $w=0$. Often in the following we will also consider QDFs induced by matrices of the form $\Delta(e^{-i\omega},\zeta_2,e^{i\omega},\eta_2)$, i.e. matrices in the indeterminates ζ_2,η_2 with coefficients being polynomials in $e^{i\omega}$ for some $\omega\in\mathbb{R}$. The definition of nonnegativity and positivity in this case is readily adapted from above.

4 Necessary and Sufficient Lyapunov Conditions for Stability of 2-D Systems

Using Theorem 3, we now concentrate on stability with respect to the proper cone consisting of the first orthant of $\mathbb{Z}\times\mathbb{Z}$; we denote this set with $\mathscr{K}_0$ in the following. Moreover, we only consider the case of square autonomous systems. We begin this section with a straightforward but important refinement of Proposition 3.5 of [15].

Proposition 2. *Let $\mathfrak{B}\in\mathscr{L}_2^{\mathtt{w}}$ be square and autonomous, and let $\Delta\in\mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\xi_1,\xi_2]$ nonsingular be such that $\mathfrak{B}=\ker\,\Delta(\sigma_1,\sigma_2)$. Assume that $\delta:=\det\,\Delta$ is such that $\mathscr{H}_\delta$ is a subset of $\mathscr{K}_0$, the first orthant of $\mathbb{Z}\times\mathbb{Z}$, that intersects the coordinate axes only in the origin. Then the following statements are equivalent:*

1. *$\mathfrak{B}$ is $\mathscr{K}_0$-stable;*
2. *For all $\omega\in\mathbb{R}$, the polynomial $\delta(e^{j\omega},\xi_2)$ has all its roots outside of the closed unit disk $\{z_2\in\mathbb{C}\mid |z_2|\geq 1\}$, and the polynomial $\delta(\xi_1,e^{j\omega})$ has all its roots outside of the closed unit disk $\{z_1\in\mathbb{C}\mid |z_1|\geq 1\}$.*

Proof. The proof follows from Theorem 3 and from the equivalence of statements *i*) and *iv*) in Proposition 3.1 of [5]. □

In order to state the main result of this paper we need some notation; we denote with $\mathrm{Per}_2\subset(\mathbb{R}^{\mathtt{w}})^{\mathbb{Z}^2}$ the set consisting of all trajectories $v\in(\mathbb{R}^{\mathtt{w}})^{\mathbb{Z}^2}$ such that the restriction of v to the lines $\{(x_1,x_2)\mid x_2\in\mathbb{Z}\}$ is periodic for all fixed $x_1\in\mathbb{Z}$, i.e.

$$\mathrm{Per}_2:=\left\{v\in(\mathbb{R}^{\mathtt{w}})^{\mathbb{Z}^2}\mid v(x_1,\cdot)\in(\mathbb{R}^{\mathtt{w}})^{\mathbb{R}}\text{ is periodic for all fixed }x_1\in\mathbb{Z}\right\}\ ;$$

analogously

$$\mathrm{Per}_1:=\left\{v\in(\mathbb{R}^{\mathtt{w}})^{\mathbb{Z}^2}\mid v(\cdot,x_2)\in(\mathbb{R}^{\mathtt{w}})^{\mathbb{R}}\text{ is periodic for all fixed }x_2\in\mathbb{Z}\right\}.$$

Theorem 4. *Let $\mathfrak{B} \in \mathscr{L}_2^{\mathtt{w}}$ be square and autonomous, and $R \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\xi_1,\xi_2]$ nonsingular be such that $\mathfrak{B} = \ker\ R(\sigma_1,\sigma_2)$. The following statements are equivalent:*

(1) $\mathfrak{B}$ is $\mathscr{K}_0$-stable;

(2) There exists a VQDF $Q_\Phi = \mathrm{col}(Q_{\Phi_1}, Q_{\Phi_2})$ and a QDF Q_Δ such that

(2a) *div* $Q_\Phi \overset{\mathfrak{B}}{=} -Q_\Delta$;

(2b) $Q_{\Phi_1}(w), Q_\Delta(w) > 0$ *for all* $w \in \mathfrak{B} \cap \mathrm{Per}_2$, *and* $Q_{\Phi_2}(w), Q_\Delta(w) > 0$ *for all* $w \in \mathfrak{B} \cap \mathrm{Per}_1$.

(3) There exist $\Phi = \mathrm{col}(\Phi_1,\Phi_2)$ and Δ, with Φ_1, Φ_2, $Y \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta_1,\zeta_2,\eta_1,\eta_2]$, $\Delta \in \mathbb{R}_s^{\mathtt{w}\times\mathtt{w}}[\zeta_1,\zeta_2,\eta_1,\eta_2]$ such that

(3a)
$$\begin{aligned}(1-\zeta_1\eta_1)\Phi_1(\zeta_1,\zeta_2,\eta_1,\eta_2) + (1-\zeta_2\eta_2)\Phi_2(\zeta_1,\zeta_2,\eta_1,\eta_2)\\ = -\Delta(\zeta_1,\zeta_2,\eta_1,\eta_2)\\ +R(\zeta_1,\zeta_2)^\top Y(\zeta_1,\zeta_2,\eta_1,\eta_2) + Y(\eta_1,\eta_2,\zeta_1,\zeta_2)^\top R(\eta_1,\eta_2);\end{aligned}$$

(3b) $\Phi_1(\zeta_1,\zeta_2,\eta_1,\eta_2) \overset{\mathfrak{B}\cap\mathrm{Per}_2}{>} 0$, $\Phi_2(\zeta_1,\zeta_2,\eta_1,\eta_2) \overset{\mathfrak{B}\cap\mathrm{Per}_1}{>} 0$, *and* $\Delta(\zeta_1,\zeta_2,\eta_1,\eta_2) \overset{\mathfrak{B}\cap\mathrm{Per}_i}{>} 0$, $i = 1,2$.

We refer to a VQDF Q_Φ satisfying (2a) *and* (2b) *as a Lyapunov function for* $\mathfrak{B}$.

Proof. The equivalence of statements (2) and (3) is straightforward.

We now prove the implication (3) ⇒ (1). Consider any trajectory in $\mathfrak{B}$ of the form $w(t_1,t_2) = v\,\lambda^{t_1}\,\mu^{t_2}$ for some $v \in \mathbb{C}^{\mathtt{w}}$ and $\lambda,\mu \in \mathbb{C}$. We now prove that if μ lies on the unit circle, i.e. $\mu = e^{i\omega}$ for some $\omega \in \mathbb{R}$, then $|\lambda| > 1$. Once this will have been established, statement (1) follows from Proposition 2.

Let $\zeta_1 = \overline{\lambda}$, $\eta_1 = \lambda$, $\zeta_2 = \overline{\mu} = e^{-i\omega}$, $\eta_2 = \mu = e^{i\omega}$ in (3a):

$$(1-\overline{\lambda}\lambda)\ v^\top \Phi_1(\overline{\lambda}, e^{-i\omega}, \lambda, e^{i\omega})v = -v^\top \Delta(\overline{\lambda}, e^{-i\omega}, \lambda, e^{i\omega})v$$

The right-hand side of this equation is strictly negative; on the left-hand side $v^\top \Phi_1(\overline{\lambda}, e^{-i\omega}, \lambda, e^{i\omega})v > 0$, and consequently it follows that $1-\overline{\lambda}\lambda < 0$. An analogous argument is used when $w(t_1,t_2) = v\, e^{i\omega t_1}\,\mu^{t_2}$. This proves the claim.

The proof of implication (1) ⇒ (3) is established by producing matrices $\Phi_i \in \mathbb{R}_S^{\mathtt{w}\times\mathtt{w}}[\zeta_1,\zeta_2,\eta_1,\eta_2]$, $i=1,2$, and $\Delta \in \mathbb{R}_S^{\mathtt{w}\times\mathtt{w}}[\zeta_1,\zeta_2,\eta_1,\eta_2]$ such that (3a)–(3b) hold.

Write $R(\xi_1,\xi_2) = \sum_{i=0}^{L_1} R_i(\xi_2)\xi_1^{L_1} = \sum_{i=0}^{L_2} R_i'(\xi_1)\xi_2^{L_2}$, where L_i is the highest power of ξ_i in R, $i = 1,2$. Define the four-variable polynomial matrix

$$H(\zeta_1,\zeta_2,\eta_1,\eta_2) := R(\zeta_1,\zeta_2)^\top R(\eta_1,\eta_2) - \zeta_1^{L_1}\zeta_2^{L_2}\eta_1^{L_1}\eta_2^{L_2} R(\eta_1^{-1},\eta_2^{-1})^\top R(\zeta_1^{-1},\zeta_2^{-1})\,. \quad (3)$$

Observe that $\partial H = 0$; conclude from Proposition 1 that there exists $\Phi = \mathrm{col}(\Phi_1,\Phi_2) \in \mathbb{R}^{2\mathtt{w}\times\mathtt{w}}[\zeta_1,\zeta_2,\eta_1,\eta_2]$ such that div $\Phi(\zeta_1,\zeta_2,\eta_1,\eta_2) = H(\zeta_1,\zeta_2,\eta_1,\eta_2)$. Moreover, it is easy to see using Proposition 3.2 of [6] that

$$\Phi_1(\zeta_1,e^{-i\omega},\eta_1,e^{i\omega}) = \frac{R(\zeta_1,e^{-i\omega})^\top R(\eta_1,e^{i\omega}) - \zeta_1^{L_1}\eta_1^{L_1} R(\eta_1^{-1},e^{-i\omega})^\top R(\zeta_1^{-1},e^{i\omega})}{1-\zeta_1\eta_1}\,. \quad (4)$$

From Proposition 2 it follows that since $\mathfrak{B}$ is $\mathscr{K}_0$-stable the polynomial det $\left(R(\xi_1,e^{i\omega})\right)$ is anti-Schur (meaning all its roots have modulus greater than one) for all $\omega \in \mathbb{R}$. It follows from Corollary 1 of [8] that for all $\omega \in \mathbb{R}$ $\Phi_1(\zeta_1, e^{-i\omega}, \eta_1, e^{i\omega}) >$ 0, since (4) is equivalent with Φ_1 being the R-canonical solution of an ω-dependent polynomial Lyapunov equation in two variables (see equation (4) of [8]) for the behavior described in kernel form by $R(\xi_1, e^{i\omega})$. From this it follows that $\Phi_1 \overset{\mathfrak{B}\cap\mathrm{Per}_2}{>} 0$.

An analogous argument based on the same considerations and on the fact that $R(e^{i\omega},\xi_2)$ is anti-Schur for all $\omega \in \mathbb{R}$, shows that $\Phi_2(e^{-i\omega},\zeta_2,e^{i\omega},\eta_2) > 0$ for all $\omega \in \mathbb{R}$.

In order to conclude the proof, define

$$\begin{aligned} Y(\xi_1,\xi_2) &:= \frac{1}{2}R(\xi_1,\xi_2) \\ \Delta(\zeta_1,\zeta_2,\eta_1,\eta_2) &:= \zeta_1^{L_1}\eta_1^{L_1}\zeta_2^{L_2}\eta_2^{L_2}R(\eta_1^{-1},\eta_2^{-1})^\top R(\zeta_1^{-1},\zeta_2^{-1}) \end{aligned}$$

The fact that $\Delta(\zeta_1,e^{-i\omega},\eta_1,e^{i\omega}) > 0$ and $\Delta(e^{-i\omega},\zeta_2,e^{i\omega},\eta_2) > 0$ for all $\omega \in \mathbb{R}$ follows from the $\mathscr{K}_0$-stability of $\mathfrak{B}$, which implies for all $\omega \in \mathbb{R}$ that $R(\xi_1,e^{i\omega})$ and $R(e^{i\omega},\xi_2)$ are anti-Schur. □

Remark 1. The 4-variable polynomial matrices $\Phi = \mathrm{col}(\Phi_1,\Phi_2)$ and Δ given in the proof of Theorem 4 are germane to the multivariable Bézoutian

$$\frac{R(\zeta)^\top R(\eta) - R(-\eta)^\top R(-\zeta)}{\zeta+\eta}$$

used in analyzing stability of 1-D continuous-time systems (see section 3 of [16]). In the 2-D single-variable ($\mathtt{w} = 1$) case, stability conditions based on the positivity of the coefficient matrix of an ω-dependent Bézoutian have been obtained in [4, 5].

Of course, there are more Lyapunov functions than the Bézoutian. The computation of Lyapunov functions via a (4-variable) polynomial Lyapunov equation as in [11, 16] is the subject of an ongoing investigation.

Acknowledgements. The second author gratefully acknowledges the financial support of *the Great Britain Sasakawa Foundation* for financially supporting the visit to Kyoto University during which the results presented in this paper were obtained.

References

1. Fornasini, E., Marchesini, G.: Stability Analysis of 2-D Systems. IEEE Trans. Circ. Syst. 27(12), 1210–1217 (1980)
2. Fornasini, E., Rocha, P., Zampieri, S.: State-space realization of 2-D finite-dimensional behaviors. SIAM J. Control and Optim. 31, 1502–1517 (1993)
3. Fornasini, E., Valcher, M.E.: *n*D polynomial matrices with applications to multidimensional signal analysis. Multidimensional Syst. Sign. Proc. 8, 387–407 (1997)
4. Geronimo, J.S., Woerdeman, H.J.: Positive extensions, Fejér-Riesz factorization and autoregressive filters in two-variables. Ann. Math. 160, 839–906 (2004)

5. Geronimo, J.S., Woerdeman, H.J.: Two-Variable Polynomials: Intersecting Zeros and Stability. IEEE Trans. Circ. Syst., Part I: Regular Papers 53(5), 1130–1139 (2006)
6. Kaneko, O., Fujii, T.: Discrete-time average positivity and spectral factorization in a behavioral framework. Systems & Control Letters 39, 31–44 (2000)
7. Kojima, C., Rapisarda, P., Takaba, K.: Canonical forms for polynomial and quadratic differential operators. Systems & Control Letters 56, 678–684 (2007)
8. Kojima, C., Takaba, K.: A generalized Lyapunov stability theorem for discrete-time systems based on quadratic difference forms. In: Proc. 44th IEEE Conference on Decision and Control and the European Control Conference, Seville, Spain, pp. 2911–2916 (2005)
9. Kojima, C., Takaba, K.: A Lyapunov stability analysis of 2-D discrete-time behaviors. In: Proc. 17th MTNS, Kyoto, Japan, pp. 2504–2512 (2006)
10. Lu, W.-S., Lee, E.B.: Stability analysis of two-dimensional systems via a Lyapunov approach. IEEE Trans. Circ. Syst. 32(1), 61–68 (1985)
11. Peeters, R., Rapisarda, P.: A two-variable approach to solve the polynomial Lyapunov equation. System & Control Letters 42, 117–126 (2001)
12. Pillai, H.K., Willems, J.C.: Lossless and dissipative distributed systems. SIAM J. Control Optim. 40, 1406–1430 (2002)
13. Polderman, J.W., Willems, J.C.: Introduction to Mathematical System Theory: A Behavioral Approach. Springer, Berlin (1997)
14. Rocha, P.: Structure and representation of 2-D systems. Ph.D. thesis, Univ. of Groningen, The Netherlands (1990)
15. Valcher, M.E.: Characteristic cones and stability properties of two-dimensional autonomous behaviors. IEEE Trans. Circ. Syst., Part I: Fundamental Theory and Applications 47(3), 290–302 (2000)
16. Willems, J.C., Trentelman, H.L.: On quadratic differential forms. SIAM J. Control Optim. 36(5), 1703–1749 (1998)

From Lifting to System Transformation

Yoshito Ohta

Abstract. This paper studies two kinds of generalizations of the lifting technique originally introduced for the sampled-data control theory. First, the lifting technique is extended when arbitrary inner function is used. Another direction for the extension is the system transformation for stochastic systems. The latter can be applied to the continuous-time system identification problem. It turns out that the PO-MOESP algorithm can be exploited to identify the coefficient matrices.

1 Introduction

The lifting technique introduced in [11] paved the way for the study sampled-data control systems. A continuous-time system is represented by an equivalent discrete-time system whose input and output spaces are functional spaces. This turns out a right tool for the sampled-data H^∞ control problems. This paper tries to extend the technique.

In the original lifting technique, the space decomposition for the input and output signals is based on the orthogonal complement of a shift invariant subspace. This suggests that for any inner function, using the orthogonal complement of the shift-invariant subspace, we can define a lifted system [6]. This framework was useful in studying Hankel singular values of a class of infinite-dimensional systems.

When the inner function is rational, the orthogonal complement of the shift-invariant subspace is finite-dimensional. When a particular basis is used, the lifted system becomes transformed system. This transformation is called system Hambo transform, which was studied in [1, 2, 3, 4]. In [7], the relation of the lifting technique and the system transformation was discussed.

The system transformation for stochastic systems is considered, and it is shown that the solutions to the continuous-time and transformed systems are equivalent.

Yoshito Ohta
Department of Applied Mathematics and Physics, Kyoto University, Yoshida-honmachi,
Kyoto 606-8501, Japan
e-mail: yoshito_ohta@i.kyoto-u.ac.jp

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 207–216.
springerlink.com

Furthermore, an application for the subspace identification method is discussed. The preliminary versions of this subject were published in [8, 9].

2 Preliminary

Let $L^2(0,\infty)$ be a space of square integrable functions of time $0<t<\infty$. Let $H^2(\mathbb{C}_+)$ be a space of analytic functions on the open right half plane such that

$$\|\hat{u}\| = \sup_{\nu>0}\left(\frac{1}{2\pi}\int_{-\infty}^{\infty}|\hat{u}(\nu+j\omega)|^2 d\omega\right)^{1/2} < \infty.$$

The Plancherel's theorem implies that the spaces $L^2(0,\infty)$ and $H^2(\mathbb{C}_+)$ are isomorphic by the Fourier transform. Similarly, the spaces $L^2(-\infty,0)$ and $H^2(\mathbb{C}_-)$ are defined, and they are isomorphic. Then the space $L^2(-\infty,\infty)$ of square integrable functions of time $-\infty<t<\infty$ is regarded as the direct sum $L^2(-\infty,\infty) = L^2(-\infty,0)\oplus L^2(0,\infty)$. This relation is carried over into the frequency domain by the Fourier transform. Let $L^2(j\mathbb{R})$ be the space of square integrable functions of frequency where the measure on the imaginary axis is $d\omega/(2\pi)$. Then we have $L^2(j\mathbb{R}) = H^2(\mathbb{C}_-)\oplus H^2(\mathbb{C}_+)$.

An inner function ϕ is a bounded analytic function on the open right half plane such that $|\phi(j\omega)| = 1$ or $\phi^\sim(j\omega)\phi(j\omega) = 1$ almost everywhere on the imaginary axis, where $\phi^\sim(s) = \overline{\phi(-\overline{s})}$ is the para-conjugate. Unless ϕ is constant, the space $\phi H^2(\mathbb{C}_+)$ is a proper closed subspace of $H^2(\mathbb{C}_+)$, and hence the orthogonal complement of the shift-invariant subspace $\phi H^2(\mathbb{C}_+)$, *i.e.*, $S = H^2(\mathbb{C}_+)\ominus\phi H^2(\mathbb{C}_+)$, is not a zero subspace.

Let Λ_ϕ be the multiplicative operator on $L^2(-\infty,\infty)$

$$\Lambda_\phi u = \mathscr{F}^{-1}\phi\mathscr{F}u,$$

where $\mathscr{F}$ is the Fourier transform. For simplicity, we write $S = L^2(0,\infty)\ominus\Lambda_\phi L^2(0,\infty)$.

Using the subspace S, the spaces $L^2(j\mathbb{R})$ and $H^2(\mathbb{C}_+)$ are written as

$$L^2(j\mathbb{R}) = \bigoplus_{k=-\infty}^{\infty}\phi^k S, \quad H^2(\mathbb{C}_+) = \bigoplus_{k=0}^{\infty}\phi^k S, \tag{1}$$

respectively, where $\phi^k = (\phi^\sim)^{-k}$ for $k<0$. Similarly, we have

$$L^2(-\infty,\infty) = \bigoplus_{k=-\infty}^{\infty}\Lambda_\phi^k S, \quad L^2(0,\infty) = \bigoplus_{k=0}^{\infty}\Lambda_\phi^k S. \tag{2}$$

From (2), any $u\in L^2(-\infty,\infty)$ can be represented as $u = \sum_{k=-\infty}^{\infty}\Lambda_\phi^k u_k$, where $u_k\in S$. Hence the map

$$u \mapsto \{\cdots, u_{-1}, u_0, u_1, \cdots\} \tag{3}$$

from $L^2(-\infty,\infty)$ to the space of square summable sequence in S is a norm-preserving bijection. The Fourier transform induces a frequency domain counter-part of (3):

$$\hat{u} \mapsto \{\cdots, \hat{u}_{-1}, \hat{u}_0, \hat{u}_1, \cdots\}. \tag{4}$$

Example 1. When $\phi(s) = e^{-sh}$, the space $S = L^2(0,h)$. Hence (2) becomes

$$L^2(-\infty,\infty) = \bigoplus_{k=-\infty}^{\infty} L^2(kh,(k+1)h).$$

This was used in the context of sampled-data control [11].

3 Lifted System

Consider a continuous-time linear system described by

$$\frac{dx}{dt} = Ax + Bu, \quad y = Cx + Du. \tag{5}$$

When A does not have eigenvalues on the imaginary axis, by decomposing the system into stable and anti-stable parts and solving the anti-stable part backward in time, the system (5) defines an input-output map $L^2(-\infty,\infty) \to L^2(-\infty,\infty) : u \mapsto y$. By the Fourier transform, the system (5) also defines a map $L^2(j\mathbb{R}) \to L^2(j\mathbb{R}) : \hat{u} \mapsto \hat{y}$. Let $H(s) = D + C(sI - A)^{-1}B$ be the transfer function of the system (5), which is bounded on the imaginary axis. Then the input-output map of (5) is nothing but the multiplication by H, or $\hat{y} = H(s)\hat{u}$.

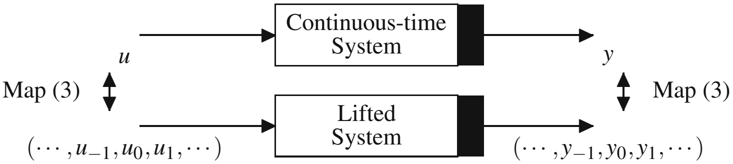

Continuous-time System
Lifted System
u
y
Map (3)
Map (3)
$(\cdots, u_{-1}, u_0, u_1, \cdots)$
$(\cdots, y_{-1}, y_0, y_1, \cdots)$

Fig. 1 Continuous-time and lifted systems

Given an inner function ϕ, the signal transformation map (3) is well-defined. Consider the diagram shown in Fig. 1. The input-output map defined by the continuous-time system (5) together with the signal transformation map (3) defines a so-called lifted system. Furthermore, it can be shown that this system is time-invariant. Assume that the system (5) is stable for the sake of simplicity. Define operators, $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{B} : S \to \mathbb{R}^n$, $\mathbf{C} : \mathbb{R}^n \to S$, and $\mathbf{D} : S \to S$ as follows:

$$\begin{aligned}
\mathbf{A}\xi &= \phi^{\sim}(A)\xi \\
\mathbf{B}u_k &= \int_{-\infty}^{0} e^{-A\tau} B\left(\Lambda_{\phi^{\sim}} u_k\right)(\tau)d\tau \\
\mathbf{C}\xi &= \mathscr{F}^{-1}\left\{C(sI-A)^{-1}\xi - \phi(s)C(sI-A)^{-1}\phi^{\sim}(A)\xi\right\}, \\
\mathbf{D}u_k &= \mathscr{F}^{-1}\left\{h(s)\mathscr{F}u_k - \phi(s)C(sI-A)^{-1}\mathbf{B}u_k\right\}.
\end{aligned}$$

Then the discrete-time system

$$\xi(k+1) = \mathbf{A}\xi(k) + \mathbf{B}u_k, \quad y_k = \mathbf{C}\xi(k) + \mathbf{D}u_k \tag{6}$$

is a state-space representation of the lifted system. When A is anti-stable, we assume that ϕ is analytic at the spectrum of $-A$. Then the operators **A**, **B**, **C**, and **D** can be defined. See [6] for detail.

Example 2. When $\phi(s) = e^{-sh}$, the signal space can be written as in Example 1. Then the operators for the system (6) are

$$\begin{aligned}
\mathbf{A}\xi &= e^{Ah}\xi \\
\mathbf{B}u_k &= \int_{-h}^{0} e^{-A\tau} B u_k(h+\tau)d\tau \\
(\mathbf{C}\xi)(t) &= Ce^{At}\xi, \ t \in (0,h), \\
(\mathbf{D}u_k)(t) &= Du_k(t) + C\int_0^t e^{A(t-\tau)}Bu_k(\tau)d\tau.
\end{aligned}$$

This is the lifted system considered in [11] to solve sampled-data control problems.

4 Rational Inner Function

When ϕ is a rational inner function, the space $S = H^2(\mathbb{C}_+) \ominus \phi H^2(\mathbb{C}_+)$ is finite dimensional. If we introduce a basis of S, then the operators of the lifted system (6) are matrices, and hence we have an ordinary discrete-time system.

The following is a way to introduce an orthonormal basis of S. Suppose that

$$\phi(s) = D_\phi + C_\phi\left(sI - A_\phi\right)^{-1}B_\phi \tag{7}$$

is a balanced realization with order n_ϕ. Define

$$\hat{v}(s) = \left[\hat{v}_1(s)\ \hat{v}_2(s)\ \cdots\ \hat{v}_{n_\phi}(s)\right] = C_\phi\left(sI - A_\phi\right)^{-1}. \tag{8}$$

Then it turns out that $\left\{\hat{v}_1, \cdots, \hat{v}_{n_\phi}\right\}$ is an orthonormal basis of $S \subset H^2(\mathbf{C}_+)$, or, to put is shortly, $\hat{v}$ is an orthonormal basis. For the time domain,

$$v(t) := \mathscr{F}^{-1}\hat{v}\left[v_1(t)\ v_2(t)\ \cdots\ v_{n_\phi}(t)\right] = C_\phi e^{A_\phi t} \tag{9}$$

is an orthonormal basis of $S \subset L^2(0,\infty)$.

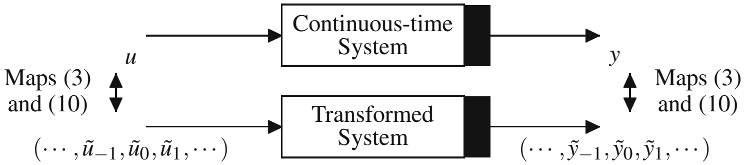

Continuous-time System
Transformed System
u
y
Maps (3) and (10)
Maps (3) and (10)
$(\cdots,\tilde{u}_{-1},\tilde{u}_0,\tilde{u}_1,\cdots)$
$(\cdots,\tilde{y}_{-1},\tilde{y}_0,\tilde{y}_1,\cdots)$

Fig. 2 Continuous-time and transformed systems

Multiplicating ϕ successively to (8), we have an orthonormal basis of $H^2(\mathbb{C}_+)$, $\{\hat{v},\phi\hat{v},\phi^2\hat{v},\cdots\}$. Including negative powers of ϕ, we obtain an orthonormal basis of $L^2(j\mathbb{R})$, $\{\cdots,\phi^{-1}\hat{v},\hat{v},\phi\hat{v},\cdots\}$. By the inverse Fourier transform, we have an orthonormal basis of $L^2(0,\infty)$ and $L^2(-\infty,\infty)$, respectively. These bases are called generalized orthonormal basis functions.

Example 3. Let $\phi(s)=(p-s)/(p+s)$, $p>0$ be a first order inner function. Then the space S is one dimensional space spanned by $\hat{v}(s)=\sqrt{2p}/(p+s)$. The set $\{\hat{v},\phi\hat{v},\phi^2\hat{v},\cdots\}$ is an orthonormal basis of $H^2(\mathbb{C}_+)$, and called the Laguerre basis.

We assume that the system (5) is single-input single-output for the sake of simplicity. Any element $u_k\in S$ is a linear combination of the basis, and hence there exists $\tilde{u}_k\in\mathbf{R}^{n_\phi}$ such that $u_k=v\tilde{u}_k$. Hence the map

$$\{\cdots,u_{-1},u_0,u_1,\cdots\}\mapsto\{\cdots,\tilde{u}_{-1},\tilde{u}_0,\tilde{u}_1,\cdots\} \tag{10}$$

is a norm-preserving bijection.

Using the maps (3) and (10), we see that the input-output map of the continuous-time system (5) defines a map of the transformed system as is shown in Fig.2. For this let X and Y be the solution to the Sylvester equations,

$$AX+XA_\phi^{\mathrm{T}}+BB_\phi^{\mathrm{T}}=0,\quad A_\phi^{\mathrm{T}}Y+YA+C_\phi^{\mathrm{T}}C=0,$$

respectively. Define

$$\tilde{A}=\phi^{\sim}(A),\quad \tilde{B}=X,\quad \tilde{C}=Y,\quad \tilde{D}=H^{\sim}(A_\phi^{\mathrm{T}}). \tag{11}$$

Then the transformed system has the state-space representation

$$\xi(k+1)=\tilde{A}\xi(k)+\tilde{B}\tilde{u}_k,\quad \tilde{y}_k=\tilde{C}\xi(k)+\tilde{D}\tilde{u}_k. \tag{12}$$

The system (12) is called system Hambo transform, and studied extensively in [4].

5 Stochastic System Transformation

Consider a continuous-time linear stochastic system

$$dx=Axdt+Bdw,\quad x(0)=0,\quad d\eta=Cxdt+Ddw, \tag{13}$$

where w is a Wiener process, and η is the measurement process. We assume that A is stable. We say that (w,η) is a solution of (13) if it satisfies:

$$x(t) = \int_0^t e^{A(t-\tau)} B dw(\tau), \quad \eta(t) = \int_0^t Cx(\tau)d\tau + Dw(t). \tag{14}$$

Let ϕ be a rational inner function having a balanced realization (7). Define matrices $\tilde{A}$, $\tilde{B}$, $\tilde{C}$, and $\tilde{D}$ as in (11). Consider a discrete-time stochastic system

$$\xi(k+1) = \tilde{A}\xi(k) + \tilde{B}\tilde{w}_k, \quad \xi(0) = 0, \quad \tilde{y}_k = \tilde{C}\xi(k) + \tilde{D}\tilde{w}_k, \tag{15}$$

where $\tilde{w}$ is a discrete-time white Gaussian process. We say that $(\tilde{w},\tilde{y})$ is a solution of (15) if it satisfies

$$\tilde{y}_k = \tilde{D}\tilde{w}_k + \sum_{i=0}^{k-1} \tilde{C}\tilde{A}^{k-i-1}\tilde{B}\tilde{w}_i.$$

Theorem 1. *Suppose that* (w,η) *is a solution of the continuous-time stochastic system* (13)*. Define*

$$\tilde{w}(k) = \int_0^\infty \Lambda_\phi^k v(t)^{\mathrm{T}} dw(t), \quad \tilde{y}(k) = \int_0^\infty \Lambda_\phi^k v(t)^{\mathrm{T}} d\eta(t). \tag{16}$$

Then the integral (16) *exists,* $\tilde{w}$ *is a discrete-time white Gaussian process, and* $(\tilde{w},\tilde{y})$ *is a solution of the discrete-time stochastic system* (15)*.*

To prove the converse direction, we need a couple of definitions. A stochastic process α has independent increments if $\alpha(t_1) - \alpha(t_0)$, $\alpha(t_2) - \alpha(t_1)$, ..., $\alpha(t_k) - \alpha(t_{k-1})$ are independent for any $0 \le t_0 < t_1 < \cdots < t_k$. A stochastic process α is time-homogeneous if $\alpha(t+h) - \alpha(s+h)$ and $\alpha(t) - \alpha(s)$ have the same distribution for any t, s, and $h > 0$. Consider stochastic processes α and β. We say β is a version of α if $\mathbb{P}\{\alpha(t) = \beta(t)\} = 1$ for any t. Consider a function f on the closed interval $J = [\underline{t}, \bar{t}]$. A function f is said to be Hölder continuous with exponent d if

$$\sup_{a,b\in J,\, a\neq b} \frac{|f(a) - f(b)|}{|a-b|^d}$$

is finite.

Theorem 2. *Consider a discrete-time white Gaussian process* $\tilde{w}$*. Suppose that* $(\tilde{w},\tilde{y})$ *is a solution of the discrete-time system* (15)*. Define*

$$c_k(t) = \int_0^t \Lambda_\phi^k v(\tau) d\tau, \quad k = 0,1,2,\cdots, \tag{17}$$

$$\check{w}(t) = \sum_{k=0}^\infty c_k(t)\tilde{w}(k) \tag{18}$$

$$\check{\eta}(t) = \sum_{k=0}^\infty c_k(t)\tilde{y}(k). \tag{19}$$

Then the processes (18) *and* (19) *are well-defined, and the sum converges in the mean square sense on any bounded interval* $[0,\bar{t}]$. *The process* $\check{w}$ *is time-homogeneous, has independent increments, and has a version that is Hölder continuous with exponent* $0 < d < 1/2$ *almost surely for any* $[0,\bar{t}]$. *Furthermore,* $(\check{w},\check{\eta})$ *is a solution of the continuous-time system* (13).

Remark 1. Theorems 1 and 2 establish the equivalence of the continuous-time stochastic system and the transformed discrete-time stochastic system in terms of solutions. Preliminary results were originally presented in [8].

6 Applications

In this section, we briefly describe applications of aforementioned frameworks to control systems theory.

6.1 Hankel Singular Values

Let ψ be a bounded function on the imaginary axis. Define $\Gamma_\psi : H^2(\mathbb{C}_-) \to H^2(\mathbb{C}_+)$, $\hat{f} \mapsto \Pi_+ \psi \hat{f}$, where Π_+ is the projection from $L^2(j\mathbb{R})$ to $H^2(\mathbb{C}_+)$. This operator is called the Hankel operator with the symbol ψ. The importance of Hankel operators is found in the area of model reduction and the H^∞ sensitivity minimization problem. The H^∞ sensitivity minimization problem for input delay systems was studied in [12], where a Hamiltonian determinant condition for the attainable norm was derived. The Hamiltonian formula was extended to include a system having a general inner function in [5].

If $\psi = \phi H$, where ϕ is an inner function, and $H(s) = D + C(sI - A)^{-1}B$ is a stable rational function, the Hankel singular value $\sigma > \sigma_{\mathrm{ess}}$ can be characterised by Schmidt pairs

$$\Gamma_\psi \hat{f} = \sigma \hat{g}, \quad \Gamma_\psi^* \hat{g} = \sigma \hat{f}, \tag{20}$$

where σ_{ess}^2 is the essential spectral radius of $\Gamma_\psi^* \Gamma_\psi$.

Let $\tilde{A}$, $\tilde{B}$, $\tilde{C}$, and $\tilde{D}$ be as in (11). Notice that the multiplication by ϕ in the transformed domain is a unit time delay. Hence the equation (20) can be represented by the following state-space equations:

$$\begin{aligned}
\tilde{\xi}(k+1) &= \tilde{A}\tilde{\xi}(k) + \tilde{B} f_{k-1}, \; k = 0,-1,-2,\cdots \\
\sigma g_0 &= \tilde{C}\tilde{\xi}(0) + \tilde{D} f_{-1}, \quad \sigma g_k = \tilde{C}\tilde{A}^{k-1}\tilde{\xi}(1), \; k = 1,2,\cdots, \\
\tilde{\zeta}(k) &= \tilde{A}^{\mathrm{T}}\tilde{\zeta}(k+1) + \tilde{C}^{\mathrm{T}} g_{k+1}, \; k = -1,0,1,\cdots, \\
\sigma f_{-1} &= \tilde{B}^{\mathrm{T}}\tilde{\zeta}(0) + D^{\mathrm{T}} g_0, \quad \sigma f_{-k} = \tilde{B}^{\mathrm{T}}\tilde{A}^{\mathrm{T}k-2}\tilde{\zeta}(-1), \; k = 2,3,\cdots.
\end{aligned}$$

A non-trivial solution (f,g) is a Schmidt pair of the Hankel operator. The existence of such a solution is equivalent to the following Hamiltonian determinant condition:

$$\det\left(\begin{bmatrix} I & -\frac{1}{\sigma}P \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ -\frac{1}{\sigma}Q & I \end{bmatrix} \phi^{\sim}(H)\right) = 0, \tag{21}$$

where

$$H = \begin{bmatrix} A + \frac{1}{\sigma^2}B\left(I - \frac{1}{\sigma^2}D^{\mathrm{T}}D\right)^{-1}D^{\mathrm{T}}C & \frac{1}{\sigma}B\left(I - \frac{1}{\sigma^2}D^{\mathrm{T}}D\right)^{-1}B^{\mathrm{T}} \\ -\frac{1}{\sigma}C^{\mathrm{T}}\left(I - \frac{1}{\sigma^2}DD^{\mathrm{T}}\right)^{-1}C & -A^{\mathrm{T}} - \frac{1}{\sigma^2}C^{\mathrm{T}}\left(I - \frac{1}{\sigma^2}DD^{\mathrm{T}}\right)^{-1}DB^{\mathrm{T}} \end{bmatrix},$$

and P and Q are the controllability and observability Gramians:

$$AP + PA^{\mathrm{T}} + BB^{\mathrm{T}} = 0, \quad QA + A^{\mathrm{T}}Q + C^{\mathrm{T}}C = 0$$

For detail, see [6].

6.2 Continuous-Time System Identification

The system model for continuous-time system identification is

$$dx = Axdt + B_1dw + B_2udt, \quad x(0) = x_0, \quad d\eta = Cxdt + D_1dw + D_2udt, \tag{22}$$

where A is stable, w is a Wiener process, and $u \in L^\infty$ is a known signal. We assume that x_0 is zero-mean Gaussian and is independent from w. Let $0 \le t_0 < t_1 < \cdots < t_i < \cdots$ be a sequence of time instances such that $t_{i+1} - t_i \ge t_{\min}$ for some $t_{\min} > 0$.

Define

$$x_i = x(t_i) \qquad \tilde{u}_{k,i} = \int_0^\infty \Lambda_\phi^k v(t)^{\mathrm{T}} u(t+t_i)dt,$$
$$\tilde{w}_{k,i} = \int_0^\infty \Lambda_\phi^k v(t)^{\mathrm{T}} dw(t+t_i), \qquad \tilde{y}_{k,i} = \int_0^\infty \Lambda_\phi^k v(t)^{\mathrm{T}} d\eta(t+t_i).$$

For fixed integers p, q and N, define

$$X_N = \begin{bmatrix} x_0 \; x_1 \; \cdots \; x_{N-1} \end{bmatrix}, \tag{23}$$

$$W_{p,q,N} = \begin{bmatrix} \tilde{w}_{p,0} & \tilde{w}_{p,1} & \cdots & \tilde{w}_{p,N-1} \\ \tilde{w}_{p+1,0} & \tilde{w}_{p+1,1} & \cdots & \tilde{w}_{p+1,N-1} \\ \vdots & \vdots & \vdots & \vdots \\ \tilde{w}_{p+q-1,0} & \tilde{w}_{p+q-1,1} & \cdots & \tilde{w}_{p+q-1,N-1} \end{bmatrix}. \tag{24}$$

The matrices $U_{p,q,N}$, and $Y_{p,q,N}$ are similarly defined.

Let $\tilde{A}$, $\tilde{B}_1$, $\tilde{B}_2$, $\tilde{C}$, $\tilde{D}_1$, and $\tilde{D}_2$ be defined *mutatis mutandis* as in (11). Construct the matrices

$$\Gamma_q = \begin{bmatrix} \tilde{C} \\ \tilde{C}\tilde{A} \\ \vdots \\ \tilde{C}\tilde{A}^{q-1} \end{bmatrix}, \quad H_{i,q} = \begin{bmatrix} \tilde{D}_i & 0 & \cdots & 0 \\ \tilde{C}\tilde{B}_i & \tilde{D}_i & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \tilde{C}\tilde{A}^{q-2}\tilde{B}_i & \tilde{C}\tilde{A}^{q-3}\tilde{B}_i & \cdots & \tilde{D}_i \end{bmatrix}, \; i = 1,2.$$

From the results in Section 5, we see that the following algebraic equation holds:

$$Y_{p,q,N} = \Gamma_q X_N + H_{1,q} W_{p,q,N} + H_{2,q} U_{p,q,N}. \tag{25}$$

The equation (25) suggests that a discrete-time subspace identification method can be applied to estimate the coefficient matrices. A major difference is that though the columns of the matrix (24) are white sequences the row are not. However, we can prove the following.

Theorem 3. *The matrices X_N, $U_{p,q,N}$, and $W_{p,q,N}$ defined by* (23) *and* (24) *satisfy*

$$\lim_{N\to\infty} \frac{1}{N} W_{q,q,N} U_{0,q,N}^{\mathrm{T}} = 0, \qquad \lim_{N\to\infty} \frac{1}{N} W_{q,q,N} U_{q,q,N}^{\mathrm{T}} = 0,$$
$$\lim_{N\to\infty} \frac{1}{N} W_{q,q,N} X_N^{\mathrm{T}} = 0, \qquad \lim_{N\to\infty} \frac{1}{N} W_{q,q,N} W_{0,q,N}^{\mathrm{T}} = 0$$

almost surely.

Remark 2. We assume that the input u satisfies the so-called PE (persistency of excitation) condition. Then from Theorem 3, it follows that the PO-MOESP algorithm [10] yields consistent estimates of the system matrices.

7 Conclusion

This article reviewed lifting technique when the signal space is represented by the orthogonal complements of the shift invariant subspace defined by an arbitrary inner function. The lifted system is a discrete-time linear time-invariant system such that commutative diagram holds. When the inner function is rational and specific basis is introduced, the lifted system is represented by the transformed system, which is sometimes called system Hambo transform.

The system transform is extended to the system whose input is a random process. This theory can be applied to the continuous-time system identification problem. Using the transformed data, the PO-MOESP subspace identification algorithm can be utilized.

References

1. de Hoog, T.J., Szabó, Z., Heuberger, P.S.C., Van den Hof, P.M.J., Bokor, J.: Minimal partial realization from generalized orthonormal basis function expansions. Automatica 38, 655–669 (2002)

2. Heuberger, P.S.C., de Hoog, T.J., Van den Hof, P.M.J., Wahlberg, B.: Orthonormal basis functions in time and frequency domain: Hambo transform theory. SIAM Journal on Control and Optimization 42, 1347–1373 (2003)
3. Heuberger, P.S.C., Van den Hof, P.M.J., Bosgra, O.H.: A generalized orthogonal basis for linear dynamical systems. IEEE Transactions on Automatic Control 40, 451–465 (1995)
4. Heuberger, P.S.C., Van den Hof, P.M.J.: The Hambo transform: a signal and system transform induced by generalized orthogonal basis functions. In: Proc. 13th IFAC World Congress, pp. 357–362 (1996)
5. Lypchuk, T.A., Smith, M.C., Tannenbaum, A.: Weighted sensitivity minimization: general plants in H_∞ and rational weights. Linear Algebra and its Applications 109, 71–90 (1988)
6. Ohta, Y.: Hankel singular values and vectors of a class of infinite dimensional systems: exact Hamiltonian formulas for control and approximation problems. Mathematics of Control, Signals, and Systems 12(361-375) (1999)
7. Ohta, Y.: Realization of input-output maps using generalized orthonormal basis functions. Systems & Control Letters 54, 521–528 (2005)
8. Ohta, Y.: A study on stochastic system transformation using generalized orthonormal basis functions. In: Proceeding of the 46th IEEE Conference on Decision and Control, pp. 3114–3119 (2007)
9. Ohta, Y., Kawai, T.: Continuous-time subspace system identification using generalized orthonormal basis functions. In: Proceedings of the 16th International Symposium on Mathematical Theory of Networks and Systems, MTNS 2004 (2004)
10. Verhaegen, M.: Identification of the deterministic part of MIMO state space models given in innovations form from input-output data. Automatica 30, 61–74 (1994)
11. Yamamoto, Y.: A function space approach to sampled data control systems and tracking problems. IEEE Trans. Autom. Control 39, 703–714 (1994)
12. Zhou, K., Khargonekar, P.P.: On the weighted sensitivity minimization problem for delay systems. Systems & Control Letters 8, 307–312 (1987)

Contractive Systems with Inputs

Eduardo D. Sontag

Dedicated to Y. Yamamoto on the occasion of his 60th birthday

Abstract. Contraction theory provides an elegant way of analyzing the behaviors of systems subject to external inputs. Under sometimes easy to check hypotheses, systems can be shown to have the incremental stability property that all trajectories converge to a unique solution. This property is especially interesting when forcing functions are periodic (a globally attracting limit cycle results), as well as in the context of establishing synchronization results. The present paper provides a self-contained introduction to some basic results, with a focus on contractions with respect to non-Euclidean metrics.

1 Introduction

The most common approach to analyzing global stability properties of nonlinear dynamical systems is through Lyapunov functions. However, in many applications, Lyapunov functions are not always easy to find, especially if steady states are not known *a priori*. Remarkably, a stronger property than stability, namely the *contraction* (or incremental stability) requirement that all solutions should converge (exponentially) towards each other, is sometimes easier to work with. Contractive dynamics result when the logarithmic norm, or matrix measure, of the Jacobian of the vector field is uniformly negative on the state space. Different norms are appropriate to different problems, just as different Lyapunov functions have to be carefully picked. Non-Euclidean norms have been found to be useful in the study of many bio-molecular problems, see for example [13].

The study of contractions in the context of stability theory dates back at least to the work of Demidovich ([4]), who established the basic convergence results with

Eduardo D. Sontag
Department of Mathematics, Rutgers University, USA
e-mail: sontag@math.rutgers.edu

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 217–228.
springerlink.com

respect to Euclidean norms, and independently to Yoshizawa ([20, 21]); see [10] for a historical discussion. In control theory, contraction theory has been popularized and extended by Slotine and coworkers, see for instance [7], [6], [19] where applications to nonlinear control, observer problems, and synchronization and consensus problems in complex networks have been developed, as well as by Nijmejer and coworkers in the context of nonlinear regulator problems, see for example [11]. In this latter work, the authors use the phrase “convergent dynamics” to refer to property that there exists a (necessarily unique) globally asymptotically stable solution to which all other solutions converge.

This paper gives a self-contained exposition, with simple proofs, of some basic results in contraction theory, It seems difficult to find such simple proofs in the literature, particularly for contractions with respect to non-Euclidean norms. We emphasize that the presentation is expository, and no substantial new results on contraction theory are claimed.

Definitions and statements of the main results are provided in Section 2, and proofs are given in Section 3.

Section 4 briefly discusses the application of contraction theory to the synchronization of coupled identical dynamical systems, following an idea of Slotine and collaborators (“virtual systems”). Also discussed there is a minor extension in which simultaneous convergence, not merely synchronization, is achieved.

For periodically forced contractive systems, globally attracting limit cycles arise, a sort of “entrainment” property. Such a property is false for general systems that have a well-defined steady-state response to constant inputs, for which even chaotic behavior may arise under periodic forcing ([16]).

In closing this introduction, we remark that a modern approach to contractive dynamics steps away from the consideration of Jacobians, and defines contraction properties by means of “logarithmic Lipschitz constants” directly associated to the vector field. This elegant approach, nicely surveyed in [14], is powerful and intuitive, and allows immediate generalizations to infinite-dimensional problems. However, in order to verify the property for particular examples, Jacobians must still be employed.

2 Definitions and Statements of Main Results

We consider in this paper systems of ordinary differential equations, generally time-dependent:

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

defined for $t \in [0,\infty)$ and $x \in C$, where C is a subset of $\mathbb{R}^n$. It will be assumed that $f(t,x)$ is differentiable on x, and that $f(t,x)$, as well as the Jacobian of f with respect to x, denoted as $J(t,x) = \frac{\partial f}{\partial x}(t,x)$, are continuous in (t,x). In applications of the theory, it is often the case that C will be a closed set, for example given by non-negativity constraints on variables as well as linear equalities representing mass-conservation laws. For a non-open set C, differentiability in x means that the vector field $f(t,\cdot)$ can be extended as a differentiable function to some open set which includes C, and the continuity hypotheses with respect to (t,x) hold on this open set.

We denote by $\varphi(t,s,\xi)$ the value of the solution $x(t)$ at time t of the differential equation (1) with initial value $x(s) = \xi$. It is implicit in the notation that $\varphi(t,s,\xi) \in C$ ("forward invariance" of the state set C). This solution is in principle defined only on some interval $s \leq t < s+\varepsilon$, but we will assume that $\varphi(t,s,\xi)$ is defined for all $t \geq s$. Conditions which guarantee such a "forward-completeness" property are often satisfied in applications, for example whenever the set C is closed and bounded, or whenever the vector field f is bounded. (See for example Appendix C in [15] for more discussion, as well as [1] for a characterization of the forward completeness property.) Under the stated assumptions, the function φ is jointly differentiable in all its arguments (this is a standard fact on well-posedness of differential equations, see for example Appendix C in [15]).

We recall (see for instance [9] or [5]) that, given a vector norm on Euclidean space ($|\cdot|$), with its induced matrix norm $\|A\|$, the associated *matrix measure* μ is defined as the directional derivative of the matrix norm in the direction of A and evaluated at the identity matrix, that is: $\mu(A) := \lim_{h\searrow 0} \frac{1}{h}\left(\|I+hA\|-1\right)$. For example, if $|\cdot|$ is the standard Euclidean 2-norm, then $\mu(A)$ is the maximum eigenvalue of the symmetric part of A. Matrix measures, also known as "*logarithmic norms*", were independently introduced by Germund Dahlquist and Sergei Lozinskii in 1959, [3, 8]. The limit is known to exist, and the convergence is monotonic, see [17, 3].

Definition 1. The system (1), or the time-dependent vector field f, is said to be *infinitesimally contracting* on a set $C \subseteq \mathbb{R}^n$ if there exists some norm in C, with associated matrix measure μ, such that, for some constant $c > 0$ (the *contraction rate*), it holds that:

$$\mu\left(J(x,t)\right) \leq -c, \quad \forall x \in C, \quad \forall t \geq 0. \tag{2}$$

The key result is that infinitesimal contractivity implies global contractivity:

Theorem 1. *Suppose that C is a convex subset of $\mathbb{R}^n$ and that $f(t,x)$ is infinitesimally contracting with contraction rate c. Then, for every two solutions $x(t) = \varphi(t,0,\xi)$ and $z(t) = \varphi(t,0,\zeta)$ of* (1)*, it holds that:*

$$|x(t)-z(t)| \;\leq\; e^{-ct}\,|\xi-\zeta|, \qquad \forall t \geq 0. \tag{3}$$

If $\mathscr{A}$ is a non-empty forward-invariant set for the dynamics, then every solution must approach $\mathscr{A}$. Indeed, take any $\zeta \in \mathscr{A}$ and any trajectory $x(t) = \varphi(t,0,\xi)$; then, with $z(t) = \varphi(t,0,\zeta)$, $\operatorname{dist}(x(t),\mathscr{A}) \leq |x(t)-z(t)| \;\leq\; e^{-ct}\,|\xi-\zeta| \to 0$ as $t \to \infty$. In particular, if an equilibrium exists, then it must be unique and globally asymptotically stable, and the same is true for periodic orbits. More interestingly, periodic orbits are assured to exist if the vector field is periodic, as would happen for a system with inputs $\dot{x} = f(x,u)$ under a periodic input $u(\cdot)$. We discuss this next.

Given a number $T > 0$, we will say that system (1) is T*-periodic* if it holds that $f(t+T,x) = f(t,x)\ \forall t \geq 0, x \in C$. Notice that a system $\dot{x} = f(x,u(t))$ with input

$u(t)$ is T-periodic $u(t)$ is itself a periodic function of period T. The basic theoretical result about periodic orbits is as follows.

Theorem 2. *Suppose that:*

- *C is a closed convex subset of $\mathbb{R}^n$;*
- *f is infinitesimally contracting with contraction rate c;*
- *f is T-periodic.*

Then, there is a unique periodic solution $\hat{x}(t) : [0,\infty) \to C$ of (1) of period T and, for every solution $x(t)$, it holds that $|x(t) - \hat{x}(t)| \to 0$ as $t \to \infty$.

Cascades of contractive systems are again contracting. To state this fact precisely, let us consider a system of the following form:

$$\begin{aligned} \dot{x} &= f(t,x) \\ \dot{y} &= g(t,x,y) \end{aligned}$$

where $x(t) \in C_1 \subseteq \mathbb{R}^{n_1}$ and $y(t) \in C_2 \subseteq \mathbb{R}^{n_2}$ for all t. We write the Jacobian of f with respect to x as $A(t,x) = \frac{\partial f}{\partial x}(t,x)$, the Jacobian of g with respect to x as $B(t,x,y) = \frac{\partial g}{\partial x}(t,x,y)$, and the Jacobian of g with respect to y as $C(t,x,y) = \frac{\partial g}{\partial y}(t,x,y)$,

When we say that $\dot{y} = g(t,x,y)$ is *infinitesimally contracting when x is viewed as a parameter* we mean that, with respect to some norm $|\cdot|$), there is an estimate $\mu(C(t,x,y)) \le -c_2 < 0$ for all $x \in C_1$, $y \in C_2$ and all $t \ge 0$.

Theorem 3. *Suppose that:*

- *the system $\dot{x} = f(t,x)$ is infinitesimally contracting;*
- *the system $\dot{y} = g(t,x,y)$ is infinitesimally contracting when x is viewed as a parameter;*
- *the mixed Jacobian $B(t,x,y)$ is bounded.*

Then, the cascaded system is infinitesimally contracting.

The basic contraction property insures that any solutions of $\dot{x} = f(t,x)$ exponentially converge to each other. The following result provides a "robustness margin" that says that any solution of the original system and any solution of a perturbed system $\dot{x} = f(t,x) + h(t)$ also exponentially converge to each other, provided that $h(t) \to 0$ exponentially. This is a "converging-input converging output" property that provides a weak type of input-to-state stability.

Theorem 4. *Assume that the system $\dot{x} = f(t,x)$ is infinitesimally contracting. Let $h(t)$ be a vector function satisfying $|h(t)| \le Le^{-kt}$ $\forall t \ge 0$ for some $k > 0$ and $L \ge 0$, Then, there exist constants $\ell > 0$ and κ such that the following property holds: For any solution $x(t) = \varphi(t,0,\xi)$ of the system $\dot{x} = f(t,x)$, and any solution $z(t) = \varphi(t,0,\zeta)$ of the system $\dot{x} = f(t,x) + h(t)$,*

$$|x(t) - z(t)| \le e^{-\ell t} \left(\kappa + |\xi - \zeta|\right) \tag{4}$$

for all $t \ge 0$.

In general, the constant κ cannot be dropped from the estimate in Theorem 4. Indeed, consider this counterexample: compare the solutions $x(t) = 0$ and $z(t) = te^{-t}$ of $\dot{x} = -x$ and $\dot{x} = -x + e^{-t}$ with $\xi = \zeta = 0$ respectively.

Observe that any solutions of $\dot{x} = f(t,x) + h_1(t)$ and $\dot{x} = g(t,x) + h_2(t)$ will also converge to each other, if h_1 and h_2 satisfy the properties for h in Theorem 4, since they both converge to any solution of the system with no h.

3 Proofs of Main Results

Proof of Theorem 1. We give the proof in a generalized form, in which convexity is replaced by a weaker constraint on the geometry of the space, but the estimate on trajectories is potentially weaker than in the convex case.

Let $K > 0$ be any positive real number and assume that a norm in $\mathbb{R}^n$ has been chosen. We will say that a subset $C \subset \mathbb{R}^n$ is *K-reachable* if, for any two points x_0 and y_0 in C there is some continuously differentiable curve $\gamma : [0,1] \to C$ such that: $\gamma(0) = x_0$; $\gamma(1) = y_0$; $|\gamma'(r)| \leq K\,|y_0 - x_0|$ for all $r \in [0,1]$. For convex sets C, we may pick $\gamma(r) = x_0 + r(y_0 - x_0)$, so $\gamma'(r) = y_0 - x_0$ and we can take $K = 1$. Thus, convex sets are 1-reachable, and it is easy to show that the converse holds as well.

Note that a set C is K-reachable for some K if and only if the length of a minimal-length (geodesic) smooth path connecting any two points x and y in C and parametrized by arc length, is bounded by some multiple K_0 of the Euclidean norm $|y-x|_2$. Indeed, re-parametrizing to a path γ defined on $[0,1]$, we have: $|\gamma'(r)|_2 \leq K_0\,|y-x|_2$. Since in finite dimensional spaces all norms are equivalent, a suitable K as in the above estimate exists.

Lemma 1. *Suppose that C is a K-reachable subset of $\mathbb{R}^n$ and that $f(t,x)$ is infinitesimally contracting with contraction rate c. Then, for every two solutions $x(t) = \varphi(t,0,\xi)$ and $z(t) = \varphi(t,0,\zeta)$ it holds that:*

$$|x(t) - z(t)| \leq Ke^{-ct}\,|\xi - \zeta| \qquad \forall t \geq 0\,. \tag{5}$$

Observe that Theorem 1 follows trivially from Lemma 1, since for convex sets we may pick $K = 1$.

Proof. Given any two points $x(0) = \xi$ and $z(0) = \zeta$ in C, pick a smooth curve $\gamma : [0,1] \to C$, such that $\gamma(0) = \xi$ and $\gamma(1) = \zeta$. Let $\psi(t,r) = \varphi(t,0,\gamma(r))$, that is, the solution of system (1) rooted at $\psi(0,r) = \gamma(r)$, $r \in [0,1]$. Since φ and γ are continuously differentiable, also $\psi(t,r)$ is continuously differentiable in both arguments. We define $w(t,r) := \frac{\partial\psi}{\partial r}(t,r)$. It follows that

$$\frac{\partial w}{\partial t}(t,r) = \frac{\partial}{\partial t}\left(\frac{\partial\psi}{\partial r}\right) = \frac{\partial}{\partial r}\left(\frac{\partial\psi}{\partial t}\right) = \frac{\partial}{\partial r} f(\psi(t,r),t).$$

As $\frac{\partial}{\partial r} f(\psi(t,r),t) = \frac{\partial f}{\partial x}(\psi(t,r),t)\frac{\partial\psi}{\partial r}(t,r)$, $\frac{\partial w}{\partial t}(t,r) = J(\psi(t,r),t)w(t,r)$, where $J(\psi(t,r),t) = \frac{\partial f}{\partial x}(\psi(t,r),t)$. Appealing to Coppel's inequality (see e.g. [18]), we have:

$$|w(t,r)| \leq |w(0,r)|\, e^{\int_0^t \mu(J(\tau))d\tau} \leq K\,|\xi - \zeta|\, e^{-ct}, \tag{6}$$

for all $x \in C$, $t \geq 0$, and $r \in [0,1]$. By the Fundamental Theorem of Calculus, we can write $\psi(t,1) - \psi(t,0) = \int_0^1 w(t,s)ds$. Hence, we obtain $|x(t) - z(t)| \leq \int_0^1 |w(t,s)|ds$. Now, using (6), the above inequality becomes:

$$|x(t) - z(t)| \leq \int_0^1 \left(|w(0,s)|\, e^{\int_0^t \mu(J(\tau))d\tau}\right) ds \leq K\,|\xi - \zeta|\, e^{-ct}.$$

This completes the proof of the lemma.

We remark that in some cases it might be possible to prove a strict contraction ($K = 1$) even if the domain is not convex, by appealing to the deeper theory of logarithmic Lipschitz constants (see [14] for definitions and details). If the (lub) logarithmic Lipschitz constant $M[f]$ of the vector field is $-c < 0$, then an estimate (3) holds. In general, $M[f]$ is an upper bound on the supremum of $\mu(J(t,x))$, with equality to the supremum in the convex case.

Proof of Theorem 2. We assume now that the vector field f is T-periodic.

Remark 1. Periodicity implies that the initial time is only relevant modulo T. More precisely:

$$\varphi(kT + t, kT, \xi) = \varphi(t, 0, \xi) \tag{7}$$

for all positive integers k, all $t \geq 0$, and all $x \in C$. Indeed, let $z(s) = \varphi(s, kT, \xi)$, $s \geq kT$, and consider the function $x(t) = z(kT + t) = \varphi(kT + t, kT, \xi)$, for $t \geq 0$. So,

$$\dot{x}(t) = \dot{z}(kT + t) = f(kT + t, z(kT + t)) = f(kT + t, x(t)) = f(t, x(t)),$$

where the last equality follows by T-periodicity of f. Since $x(0) = z(kT) = \varphi(kT, kT, \xi) = \xi$, it follows by uniqueness of solutions that $x(t) = \varphi(t, 0, \xi) = \varphi(kT + t, kT, \xi)$, which is (7). As a corollary, we also have that

$$\varphi(kT + t, 0, \xi) = \varphi(kT + t, kT, \varphi(kT, 0, \xi)) = \varphi(t, 0, \varphi(kT, 0, \xi)) \tag{8}$$

for all positive integers k, all $t \geq 0$, and all $x \in C$, where the first equality follows from the semigroup property of solutions (see e.g. [15]), and the second one from (7) applied to $\varphi(kT, 0, \xi)$ instead of ξ.

Define now $P(\xi) = \varphi(T, 0, \xi)$, where $\xi = x(0) \in C$.

Lemma 2. $P^k(\xi) = \varphi(kT, 0, \xi)$ *for all positive integers* k *and* $\xi \in C$.

Proof. We will prove the Lemma by recursion. In particular, the statement is true by definition when $k = 1$. Inductively, assuming it true for k, we have:

$$P^{k+1}(\xi) = P(P^k(\xi)) = \varphi(T, 0, P^k(\xi)) = \varphi(T, 0, \varphi(kT, 0, \xi)) = \varphi(kT + T, 0, \xi).$$

Theorem 5. *Suppose that:*

- *C is a closed K-reachable subset of $\mathbb{R}^n$;*
- *f is infinitesimally contracting with contraction rate c;*
- *f is T-periodic;*
- *$Ke^{-cT} < 1$.*

Then, there is an (unique) periodic solution $\hat{x}(t) : [0,\infty) \to C$ of (1) having period T. Furthermore, every solution $x(t)$ converges to $\hat{x}(t)$, i.e. $|x(t) - \hat{x}(t)| \to 0$ as $t \to \infty$.

Theorem 2 is a corollary, because the assumption $Ke^{-cT} < 1$ in Theorem 5 is automatically satisfied when the set C is convex (i.e. $K = 1$) and the system is infinitesimally contracting.

Proof. Observe that P is a contraction with factor $Ke^{-cT} < 1$: $|P(\xi) - P(\zeta)| \leq Ke^{-cT} |\xi - \zeta|$ for all $\xi, \zeta \in C$, as a consequence of Theorem 1. The set C is a closed subset of $\mathbb{R}^n$ and hence complete as a metric space with respect to the distance induced by the norm being considered. Thus, by the contraction mapping theorem, there is a (unique) fixed point $\bar{\xi}$ of P. Let $\hat{x}(t) := \varphi(t, 0, \bar{\xi})$. Since $\hat{x}(T) = P(\bar{\xi}) = \bar{\xi} = \hat{x}(0)$, $\hat{x}(t)$ is a periodic orbit of period T. Moreover, again by Theorem 1, we have that $|x(t) - \hat{x}(t)| \leq Ke^{-ct} \left|\xi - \bar{\xi}\right| \to 0$. Uniqueness is clear, since two different periodic orbits would be disjoint compact subsets, and hence at positive distance from each other, contradicting convergence. This completes the proof.

Notice that, even in the non-convex case, the assumption $Ke^{-cT} < 1$ can be dropped, provided that we assert only the existence of (and global convergence to) a unique periodic orbit, whose period is kT for some integer $k > 1$. Indeed, the vector field is also kT-periodic for any integer k. Picking k large enough so that $Ke^{-ckT} < 1$, we have the conclusion that such an orbit exists, applying Theorem 5.

Proof of Theorem 3. We assume that the system $\dot{x} = f(t,x)$ is infinitesimally contracting with respect to a norm $|\cdot|_1$, with contraction rate c_1, that is, $\mu_1(A(t,x)) \leq -c_1$ for all $x \in C_1$ and all $t \geq 0$, where μ_1 is the matrix measure associated to $|\cdot|_1$, the system $\dot{y} = g(t,x,y)$ is infinitesimally contracting with respect to a norm $|\cdot|_2$ with contraction rate c_2, when x is viewed as a parameter in the second system, that is, $\mu_2(C(t,x,y)) \leq -c_2$ for all $x \in C_1$, $y \in C_2$ and all $t \geq 0$, where μ_2 is the matrix measure associated to $|\cdot|_2$, and that the mixed Jacobian $B(t,x,y)$ is bounded: $\|B(t,x,y)\| \leq k$, for all $x \in C_1$, $y \in C_2$ and all $t \geq 0$, for some real number k, where "$\|\cdot\|$" is the operator norm induced by $|\cdot|_1$ and $|\cdot|_2$ on linear operators $\mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$.

We need to show that, under these assumptions, the complete system is infinitesimally contracting. More precisely, pick any two positive numbers ρ_1 and ρ_2 such that $c_1 - \frac{\rho_2}{\rho_1}k > 0$ and let $c := \min\left\{c_1 - \frac{\rho_2}{\rho_1}k, c_2\right\}$. We will show that $\mu(J) \leq -c$, where J is the full Jacobian: $J = \left[\begin{smallmatrix} A & 0 \\ B & C \end{smallmatrix}\right]$, with respect to the matrix measure μ induced by the following norm in $\mathbb{R}^{n_1+n_2}$: $|(x_1,x_2)| = \rho_1 |x_1|_1 + \rho_2 |x_2|_2$. Since $(I + hJ)x = \left[\begin{smallmatrix} (I+hA)x_1 \\ hBx_1 + (I+hC)x_2 \end{smallmatrix}\right]$ for all h and x, we have that, for all h and x:

$$|(I + hJ)x| = \rho_1 |(I + hA)x_1| + \rho_2 |hBx_1 + (I + hC)x_2|$$

$$\leq \rho_1 \, |I+hA| \, |x_1| + \rho_2 \, |hB| \, |x_1| + \rho_2 \, |I+hC| \, |x_2| \, ,$$

where from now on we drop subscripts for norms. Pick now any $h > 0$ and a unit vector x (which depends on h) such that $\|I+hJ\| = |(I+hJ)x|$. Such a vector x exists by the definition of induced matrix norm, and we note that $1 = |x| = \rho_1 \, |x_1|_2 + \rho_2 \, |x_2|_2$, by the definition of the norm in the product space. Therefore:

$$\begin{aligned}
\frac{1}{h} \quad (\|I+hJ\|-1) \; &= \; \frac{1}{h}\left(|(I+hJ)x| - |x|\right) \\
&\leq \frac{1}{h}\left(\rho_1 \, |I+hA| \, |x_1| + \rho_2 \, |hB| \, |x_1| + \rho_2 \, |I+hC| \, |x_2| - \rho_1 \, |x_1| - \rho_2 \, |x_2|\right) \\
&= \frac{1}{h}\left(|I+hA| - 1 + \frac{\rho_2}{\rho_1} h \, |B|\right) \rho_1 \, |x_1| + \frac{1}{h}\left(|I+hC|-1\right)\rho_2 \, |x_2| \\
&\leq \max\left\{ \frac{1}{h}\left(|I+hA|-1\right) + \frac{\rho_2}{\rho_1} k \, , \, \frac{1}{h}\left(|I+hC|-1\right)\right\} \, ,
\end{aligned}$$

where the last inequality is a consequence of the fact that $\lambda_1 a_1 + \lambda_2 a_2 \leq \max\{a_1, a_2\}$ for any non-negative numbers with $\lambda_1 + \lambda_2 = 1$ (convex combination of the a_i's). Now taking limits as $h \searrow 0$, we conclude that $\mu(J) \leq \max\left\{-c_1 + \frac{\rho_2}{\rho_1} k, -c_2\right\} = -c$, as desired.

Proof of Theorem 4. We first make some general remarks about perturbed systems. Consider additive perturbations of the system (1) of the following general form:

$$\dot{x} = F(x,t) \; = \; f(t,x) + h(t,x) \tag{9}$$

where the vector field $h(t,x)$ is defined for $t \geq 0$ and $x \in C$, with values in $\mathbb{R}^n$, is differentiable on x, and $h(t,x)$ and its Jacobian $H(t,x) = \frac{\partial h}{\partial x}(t,x)$ are both continuous in (t,x). We have the following simple observation:

Lemma 3. *Assume that the system $\dot{x} = f(t,x)$ is infinitesimally contracting with contraction rate c with respect to a norm $|\cdot|$. Suppose that the Jacobian of the perturbation satisfies:*

$$\|H(t,x)\| \leq c_h < c \tag{10}$$

for all $t \geq 0$ and all $x \in C$. Then, the perturbed system (9) *is infinitesimally contracting with respect to the same norm.*

Proof. The Jacobian of the new system is $\widetilde{J}(t,x) = J(t,x) + H(t,x)$, and:

$$\mu(\widetilde{J}(x,t)) \; \leq \; \mu\left(J\left(x,t\right)\right) + \mu(\left(H\left(x,t\right)\right) \; \leq \; \widetilde{c} := -c + c_h$$

by subadditivity of matrix measures and the fact that the norm always upper-bounds the matrix measure (see for instance [5, page 31]).

Some comments regarding Lemma 3 are as follows. *(i)* Suppose that $h(t,x)$ does not depend on x. Then (10) is trivially satisfied ($c_h = 0$). *(ii)* Suppose that $H(t,x) \to 0$ as $t \to \infty$, uniformly on $x \in C$. Then the system $\dot{x} = F(t - t_0, x)$ is infinitesimally contracting. That is, for any two solutions $x(t) = \varphi(t,t_0,\xi)$ and $z(t) = \varphi(t,t_0,\zeta)$ of (9) starting at time t_0, we have that:

$$|x(t) - z(t)| \;\leq\; e^{-c(t-t_0)}\, |\xi - \zeta|\,, \qquad \forall t \geq t_0 \geq 0\,.$$

Indeed, by assumption we have that $\beta(t) := \sup_{x\in C} \|H\,(x,t)\| \to 0$, so we can pick any $t_0 > 0$ so that $c_h = \beta(t_0) < c$. *(iii)* Consider any two solutions $x(t) = \varphi(t,0,\xi)$ and $z(t) = \varphi(t,0,\zeta)$ starting at time $t = 0$. Since $x(t) = \varphi(t,t_0,x(t_0))$ and $z(t) = \varphi(t,t_0,z(t_0))$, it follows that $x(t) - z(t) \to 0$ as $t \to 0$ (but not necessarily satisfying an estimate $|x(t) - z(t)| \;\leq\; e^{-ct}\,|\xi - \zeta|$).

Lemma 4. *Assume that the system $\dot{x} = f(t,x)$ is infinitesimally contracting with contraction rate c with respect to a norm $|\cdot|$. Suppose that h and its Jacobian H are exponentially decreasing, in the sense that, for some $k > 0$: $h(t,x)\,e^{kt}$ is bounded and $\left\|H(t,x)e^{kt}\right\| \leq c_h < c\,\forall x \in C, \forall t \geq 0$. Then, there exist constants $\ell > 0$ and κ such that the following property holds: for any solution $x(t) = \varphi(t,0,\xi)$ of the system $\dot{x} = f(t,x)$, and any solution $z(t) = \varphi(t,0,\zeta)$ of the system $\dot{x} = f(t,x) + h(t,x)$, the estimate (4) is valid for all $t \geq 0$.*

In the special case that h is independent of x, this proves Theorem 4.

Proof. Consider the following auxiliary system, with $p \in [0,1]$:

$$\begin{aligned}\dot{p} &= -kp\\ \dot{x} &= F_p(t,p,x) \;=\; f(t,x) \;+\; ph(t,x)e^{kt}\end{aligned}$$

viewed as a cascade. The p-subsystem is infinitesimally contracting with respect to the standard norm in $\mathbb{R}$. The x-subsystem is infinitesimally contracting when p is viewed as a parameter. Indeed, with: $C(t,p,x) = \frac{\partial F_p}{\partial x}(t,p,x) = J(t,x) + pH(t,x)e^{kt}$, we have that $\mu(C(t,p,x)) \leq -c + c_h$, as earlier. Moreover, the mixed Jacobian $B(t,x,y) = \frac{\partial F_p}{\partial p}(t,p,x) = h(t,x)e^{kt}$ is bounded, by assumption. It follows from Theorem 3 that the auxiliary system is also infinitesimally contracting with some rate ℓ, and the proof of that result shows that this contraction can be established with respect to a norm of the form: $|(p,x)| = \rho_1\,|p|_1 + \rho_2\,|x|_2$ for some $\rho_1 > 0$ and $\rho_2 > 0$, where $|p|_1$ denotes the usual norm in $\mathbb{R}$ and $|x|_2$ denotes the original norm on x.

Consider now any solution $x(t) = \varphi(t,0,\xi)$ of the system $\dot{x} = f(t,x)$ and any solution $z(t) = \varphi(t,0,\zeta)$ of the system $\dot{x} = f(t,x) + h(t,x)$.

Introduce $X(t) := (0,x(t))$ and $Z(t) := (e^{-kt},z(t))$. It is clear that $X(t)$ and $Z(t)$ are the solutions of the auxiliary system corresponding to initial conditions $X(0) = (0,\xi)$ and $Z(0) = (1,\zeta)$ respectively. Because the auxiliary system is infinitesimally contracting, $|X(t) - Z(t)| \leq e^{-\ell t}\,|X(0) - Z(0)|$ for all $t \geq 0$, where $|X(t) - Z(t)| = \rho_1 e^{-kt} + \rho_2\,|x(t) - z(t)|_2$ and $|X(0) - Z(0)| = \rho_1 + \rho_2\,|\xi - \zeta|_2$. So $\rho_2\,|x(t) - z(t)|_2 \leq e^{-\ell t}\,(\rho_1 + \rho_2\,|\xi - \zeta|_2)$. Dividing by ρ_2 and dropping the subscript for norms, we have (4) with $\kappa = \rho_1/\rho_2$.

4 Synchronization

We remark here on the use of contraction theory to show synchronization of coupled systems, based on the introduction of "virtual dynamics" by Slotine and collaborators (see for example [12]). For simplicity of notation, we consider time-invariant dynamics, but the same considerations apply to time-dependent vector fields.

Suppose that we have two diffusion-interconnected identical systems:

$$\begin{aligned}\dot{y} &= f(y) + \gamma(z) - \gamma(y)\\ \dot{z} &= f(z) + \gamma(y) - \gamma(z)\end{aligned}$$

where we think of γ as a coupling law and assume that γ is globally Lipschitz. Typically, γ is linear, so that $\gamma(z) - \gamma(y) = D(z-y)$ for a matrix D (which is often a diagonal matrix). For example, suppose that the systems are linear: $\dot{y} = Ay + D(z-y)$ and $\dot{z} = Az + D(y-z)$. Each system is individually (when $D = 0$) asymptotically stable if and only A is a Hurwitz matrix (all eigenvalues have negative real part). Using a change of variables $(y,z) \mapsto (y-z, y+z)$, we may bring this system to a block-diagonal form with blocks $A-2D$ and A, and thus it is clear that the interconnected system is asymptotically stable if and only both A and $A-2D$ are Hurwitz matrices. Moreover, the same proof (the first block corresponds to $y-z$) shows that for synchronization ($y(t) - z(t) \to 0$) it is enough that $A-2D$ be a Hurwitz matrix.

For general, not necessarily linear systems, if the system $\dot{x} = f(x)$ is infinitesimally contracting, then the decoupled systems (obtained when $\gamma = 0$) each satisfies that all solutions converge to each other.

More interestingly, a synchronization result can be established as follows. Consider the following "virtual system":

$$\dot{x} = f(x) - 2\gamma(x) + h(t) \tag{11}$$

(a different system results for each fixed input $h(\cdot)$) and suppose that the vector field $f - 2h$ is infinitesimally contracting. Take a particular solution $(y(t), z(t))$ of the coupled system. Then, $y(t)$ and $z(t)$ are two solutions of (11), when we pick $h(t) = \gamma(y(t)) + \gamma(z(t))$. It follows that $|z(t) - y(t)| \leq e^{-ct} \to 0$ for some $c > 0$, showing that the y and z subsystems synchronize. Observe that this fact did not require the contractivity of f, but only that of $f - 2\gamma$.

Still for this solution $(y(t), z(t))$ of the coupled system, we now define $h(t) = \gamma(z(t)) - \gamma(y(t))$. Using the assumption that γ is globally Lipschitz, we have that $|w(t)| \leq M|z(t) - y(t)| \leq Me^{-ct}$, for some constant M. Now, if f is contracting, we note that the equation satisfied by y is $\dot{x} = f(x) + h(t)$. As $h(t)$ is exponentially convergent to zero, Theorem 4 implies that $y(t) - x(t) \to 0$ as $t \to \infty$ for every solution of the system $\dot{x} = f(x)$. Pick any one particular such solution $x_0(\cdot)$. Then, $y(t) - x_0(t) \to 0$. We may repeat this argument for an arbitrary $(y(t), z(t))$, always comparing to the same $x_0(\cdot)$. In summary, we have the following conclusion: if both f and $f - 2\eta$ are infinitesimally contracting (not necessarily with respect to the same norm), then all solutions of the coupled system converge to the diagonal solution $(x_0(t), x_0(t))$.

The preceding considerations make the following question natural: when does contractivity of f (which is sufficient to provide a stability property for the isolated systems) already imply contractivity of $f - 2\gamma$ (so that synchronization to the uncoupled solutions occurs)?

We provide next a condition for the case when every Jacobian $D = D(x)$ of $\gamma(x)$ is a diagonal non-negative definite matrix. The question is, then, for the Jacobians $A = J(x)$: when does $\mu(A) \leq c$ imply that also $\mu(A - 2D) \leq c$?

Recall that a norm on $\mathbb{R}^n$ is said to be monotonic or "axis oriented" if the following property holds for any two vectors in $\mathbb{R}^n$: $|y_i| \leq |x_i| \Rightarrow |y| \leq |x|$. The usual norms (L^2, L^1, L^∞) are monotonic, as is any new norm of the type $|x|_P = |Px|$ for a diagonal positive definite matrix P, if $|\cdot|$ is monotonic.

Theorems 2 and 3 of [2] say that the following properties are equivalent: (1) the norm is monotonic, (2) $|x|$ depends only on the absolute values of the components of x, and (3) the associated operator norm satisfies that $\|E\| = \max_j\{E_{jj}\}$ for any diagonal matrix E. So $\|I - hD\| = \max_j\{1 - hD_{jj}\} = 1 - hd_{ii}$ for some i, which implies that $(1/h)(\|I - hD\| - 1) = -d_{ii}$ and thus $\mu(D) = d_{ii} \leq 0$. From subadditivity of matrix measures, we conclude that, for monotonic norms, $\mu(A + D) \leq \mu(A) \leq c$ and thus, for monotonic norms, we get contractivity of $f - 2\gamma$ from that of f.

References

1. Angeli, D., Sontag, E.D.: Forward completeness, unboundedness observability, and their Lyapunov characterizations. Systems and Control Letters 38, 209–217 (1999)
2. Bauer, F.L., Stoer, J., Witzgall, C.: Absolute and monotonic norms. Numerische Mathematik 3, 257–264 (1961)
3. Dahlquist, G.: Stability and error bounds in the numerical integration of ordinary differential equations. Trans. Royal Inst. Technol., Stockholm (1959)
4. Demidovich, B.P.: Dissipativity of a nonlinear system of differential equations. Vestnik Moscow State University, Ser. Mat. Mekh. 6, 19–27 (1961) (in Russian)
5. Desoer, C.A., Vidyasagar, M.: Feedback Synthesis: Input-Output Properties. SIAM, Philadelphia (2009)
6. Jouffroy, J., Slotine, J.J.E.: Methodological remarks on contraction theory. In: Proc. 42nd Conf. Decision and Control, pp. 2537–2543. IEEE Press, Los Alamitos (2004)
7. Lohmiller, W., Slotine, J.J.E.: Nonlinear process control using contraction theory. AIChe Journal 46, 588–596 (2000)
8. Lozinskii, S.M.: Error estimate for numerical integration of ordinary differential equations, I. Izv. Vtssh. Uchebn. Zaved. Mat. 5, 222 (1959)
9. Michel, A.N., Liu, D., Hou, L.: Stability of Dynamical Systems: Continuous, Discontinuous, and Discrete Systems. Springer, New York (2007)
10. Pavlov, A., Pogromvsky, A., van de Wouv, N., Nijmeijer, H.: Convergent dynamics, a tribute to Boris Pavlovich Demidovich. Systems & Control Letters 52, 257–261 (2004)
11. Pavlov, A., van de Wouw, N., Nijmeijer, H.: Uniform Output Regulation of Nonlinear Systems: A Convergent Dynamics Approach. Springer, Berlin (2005)
12. Russo, G., di Bernardo, M.: Contraction theory and the master stability function: linking two approaches to study synchronization in complex networks. IEEE Transactions on Circuit and Systems II 56, 177–181 (2009)

13. Russo, G., di Bernardo, M., Sontag, E.D.: Global entrainment of transcriptional systems to periodic inputs (submitted) (preprint, 2009), `http://arxiv.org/abs/0907.0017`
14. Söderlind, G.: The logarithmic norm: history and modern theory. BIT Numerical Mathematics 46, 631–652 (2006)
15. Sontag, E.D.: Mathematical Control Theory. In: Deterministic Finite-Dimensional Systems. Springer, New York (1998)
16. Sontag, E.D.: An observation regarding systems which converge to steady states for all constant inputs, yet become chaotic with periodic inputs. Technical report, arxiv 0906.2166 (2009)
17. Strom, T.: On logarithmic norms. SIAM J. Numer. Anal. 12, 741–753 (1975)
18. Vidyasagar, M.: Nonlinear Systems Analysis, 2nd edn. Pretice-Hall, Englewood Cliffs (1993)
19. Wang, W., Slotine, J.J.E.: On partial contraction analysis for coupled nonlinear oscillators. Biological Cybernetics 92, 38–53 (2005)
20. Yoshizawa, T.: Stability Theory by Liapunov's Second Method. The Mathematical Society of Japan, Tokyo (1966)
21. Yoshizawa, T.: Stability Theory and the Existence of Periodic Solutions and Almost Periodic Solutions. Springer, New York (1975)

Dissipativity and Stability Analysis Using Rational Quadratic Differential Forms

Kiyotsugu Takaba

Abstract. This paper is concerned with analysis of linear dynamical systems using rational quadratic differential forms (rational QDFs). The rational QDF is devised for the purpose of less conservative analysis in the behavioral system theory. We study the dissipativity of a linear system with respect to a supply rate defined by a rational QDF . Based on this analysis, a stability condition of an interconnected system is derived as a behavioral version of the passivity theorems with rational multipliers or scaled small gain theorem.

1 Introduction

We often encounter the situation of studying quadratic functionals which describe Lyapunov functions, energy functions, performance measures, etc. In particular, such quadratic functionals play crucial roles in the stability analysis of interconnected systems and robust control designs. Recent attempts in this direction in the feedback control framework include the integral quadratic constraints (IQCs) [3] and the quadratic separators [2]. These approaches utilize the quadratic functionals which involve dynamic scaling or frequency-dependent weighting matrices in order to reduce the conservativeness.

On the other hand, in the behavioral system theory pioneered by J. C. Willems (see e.g. the textbook [6]), such quadratic functionals are treated as quadratic differential forms (QDFs) that are typically characterized in terms of polynomial matrices [10]. The QDFs have been applied to a number of control problems such as LQ-optimal control [12], H^∞ control [9, 11, 1], Lyapunov stability [4], stability analysis of interconnections [5, 13, 7], etc. However, as observed at the beginning of this section, we often need a more general formulation of QDFs to obtain less conservative results.

Kiyotsugu Takaba
Dept. of Applied Mathematics and Physics, Graduate School of Informatics,
Kyoto University, Kyoto 606-8501, Japan
e-mail: takaba@amp.i.kyoto-u.ac.jp

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 229–238.
springerlink.com

In this paper, we will introduce a generalized formulation of QDFs in terms of rational matrices [8]. We call such a QDF *a rational QDF*. By using the rational QDFs, we will investigate the dissipativity of a linear time-invariant dynamical system, and then derive a condition for the stability of an interconnected system consisting of two controllable linear systems. The resulting stability condition is a generalization of the well-known small gain and passivity theorems.

2 Linear Differential Systems

We first review some preliminaries from the behavioral system theory.

Throughout this paper, we are interested in a linear time-invariant dynamical system which admits a kernel representation

$$R\left(\frac{d}{dt}\right)w=0,\;\; R(\xi)=R_0\xi+R_1\xi^2+\cdots+R_{L-1}\xi^{L-1}+R_L\xi^L\in\mathbb{R}^{\bullet\times\mathtt{w}}[\xi] \qquad (1)$$

where $w\in\mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{\mathtt{w}})$ is the system variable. The behavior $\mathfrak{B}$ is the set of all trajectories which meet the dynamic laws of the system. In the case of (1), $\mathfrak{B}$ is given by $\mathfrak{B}=\ker R(\frac{d}{dt})=\{w\in\mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{\mathtt{w}})|\; R(\frac{d}{dt})w=0\}$. We will hereafter identify a dynamical system with its behavior for simplicity of exposition. We also denote by $\mathscr{L}^{\mathtt{w}}$ the family of linear differential behaviors with $\mathtt{w}$-dimensional system variable. Of course, the kernel representation of a given $\mathfrak{B}$ is not unique. The kernel representation $R(\frac{d}{dt})w=0$ is called *minimal*, if the number of rows in $R(\xi)$ is minimal among all polynomial matrices inducing the kernel representations of the same system.

The system $\mathfrak{B}$ is said to be asymptotically stable if $\|w(t)\|\to 0$ $(t\to+\infty)$ holds for all $w\in\mathfrak{B}$. $\mathfrak{B}=\ker R(\frac{d}{dt})$ is asymptotically stable iff $R(\lambda)$ has full column rank for all $\lambda\in\mathbb{C}_+ :=\{\lambda\in\mathbb{C}:\ \mathrm{Re}\lambda\geq 0\}$.

If $\mathfrak{B}\in\mathscr{L}^{\mathtt{w}}$ is controllable, it admits *an image representation*

$$w=M(\tfrac{d}{dt})\ell,\;\; M\in\mathbb{R}^{\mathtt{w}\times\mathtt{l}}[\xi] \qquad (2)$$

In this case, $\mathfrak{B}=\mathrm{im}M(\frac{d}{dt})=\{\,w\in\mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{\mathtt{w}})\,|\;\exists\ell\in\mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{\mathtt{l}})\text{ s.t. }w=M(\frac{d}{dt})\ell\,\}$. We can always choose an observable image representation in which $M(\xi)$ is right prime, i.e. $M(\lambda)$ has full column rank for all $\lambda\in\mathbb{C}$.

In order to define a quadratic differential form using a rational matrix, we will need a rational representation of $\mathfrak{B}\in\mathscr{L}^{\mathtt{w}}$. Along the line of Willems and Yamamoto [14], we introduce the rational representation of a linear time-invariant differential behavior as follows. Let $G(\xi)$ belong to $\mathbb{R}^{\mathtt{v}\times\mathtt{w}}(\xi)$. Then, the solution of the "differential" equation $G(\frac{d}{dt})w-v=0$ is defined as follows:

$$[\![(w,v)\text{ satisfies }G(\tfrac{d}{dt})w-v=0.]\!]\;\Leftrightarrow\;[\![X(\tfrac{d}{dt})w=Y(\tfrac{d}{dt})v]\!] \qquad \text{(3a)}$$

$$\Leftrightarrow\exists z\in\mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{\mathtt{w}})\text{ s.t. }\begin{bmatrix}v\\ w\end{bmatrix}=\begin{bmatrix}N(\frac{d}{dt})\\ D(\frac{d}{dt})\end{bmatrix}z, \qquad \text{(3b)}$$

where the left and right coprime factorizations of $G(\xi)$ over $\mathbb{R}[\xi]$ are given by

$$\begin{aligned} G(\xi) &= X^{-1}(\xi)Y(\xi) = N(\xi)D^{-1}(\xi), \\ & X \in \mathbb{R}^{\mathtt{v}\times\mathtt{v}}[\xi],\ Y \in \mathbb{R}^{\mathtt{v}\times\mathtt{w}}[\xi], \\ & N \in \mathbb{R}^{\mathtt{v}\times\mathtt{w}}[\xi],\ D \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\xi]. \end{aligned} \tag{4}$$

Consider the situation where w is arbitrarily given and v is determined from $G(\frac{d}{dt})w - v = 0$. Then, $G(\frac{d}{dt})$ defines a point-to-set map as

$$G(\tfrac{d}{dt}) :\ w \mapsto \left\{ v \mid X(\tfrac{d}{dt})v = Y(\tfrac{d}{dt})w \right\} = \left\{ v = N(\tfrac{d}{dt})z \middle| D(\tfrac{d}{dt})z = w \right\} \tag{5}$$

Of course, if $G(\xi)$ is a polynomial matrix, $G(\frac{d}{dt})$ defines a point-to-point map from $\mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{\mathtt{w}})$ to $\mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{\mathtt{v}})$.

3 Quadratic Differential Form Defined by a Rational Matrix

We start this section with the original definition of a quadratic differential form given in terms of a polynomial matrix [10].

A quadratic differential form (QDF) Q_{Φ} is defined as a quadratic form of some variables and their derivatives, that is

$$Q_{\Phi} :\ \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{\mathtt{w}}) \to \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}),\ \ w \mapsto Q_{\Phi}(w) := \sum_{i=0}^{k}\sum_{j=0}^{k} \left(\frac{d^i w}{dt^i}\right)^{\top} \Phi_{ij} \left(\frac{d^j w}{dt^j}\right),$$

where $\Phi_{ij} \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}$ and $\Phi_{ji}^{\top} = \Phi_{ij}$ $(i,j = 0,1,\ldots,k)$. Obviously, Q_{Φ} has one-to-one correspondence with a symmetric two-variable polynomial matrix

$$\Phi(\zeta,\eta) = \sum_{i=0}^{k}\sum_{j=0}^{k} \zeta^i \eta^j \Phi_{ij} \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta,\eta],$$

where ζ and η correspond to the differentiation on $w^{\top}$ and w, respectively. For the detail of QDFs defined by polynomial matrices, the reader is recommended to refer to Willems and Trentelman [10].

Let $\Phi \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}[\zeta,\eta]$ be factored as

$$\Phi(\zeta,\eta) = M^{\top}(\zeta)\Sigma M(\eta),\ M \in \mathbb{R}^{\mathtt{v}\times\mathtt{w}}[\xi],\ \Sigma \in \mathbb{R}^{\mathtt{v}\times\mathtt{v}},\ \det\Sigma \neq 0. \tag{6}$$

Then, the QDF Q_{Φ} is equivalently expressed as

$$Q_{\phi}(w) = v^{\top}\Sigma v,\ \ v = M(\tfrac{d}{dt})w. \tag{7}$$

In the similar way to (6) and (7), we generalize the definition of a QDF in terms of a symmetric two-variable rational matrix. We call a symmetric two-variable rational matrix $\Phi \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}(\zeta,\eta)$ *factorizable* if there exist a rational matrix $G \in \mathbb{R}^{\mathtt{v}\times\mathtt{w}}(\xi)$, a symmetric constant matrix $\Sigma \in \mathbb{R}^{\mathtt{v}\times\mathtt{v}}$ and $\mathtt{v} \in \mathbb{N}$ such that

$$\Phi(\zeta,\eta) = G^{\top}(\zeta)\Sigma G(\eta), \ \ \det\Sigma \neq 0. \tag{8}$$

We make the following assumption throughout this section.

Assumption 1. *The two-variable rational matrix $\Phi \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}(\zeta,\eta)$ is symmetric, factorizable and admits a factorization of* (8).

Definition 1. [8] Under Assumption 1, *the rational QDF Q_Φ induced by $\Phi \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}(\zeta,\eta)$* is the point-to-set map defined by

$$Q_\Phi : \ w \mapsto \ Q_\Phi(w) := \left\{ s \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}) \,\middle|\, \exists v \in G(\tfrac{d}{dt})w \text{ s.t. } s = v^\top \Sigma v \right\} \tag{9}$$

where the map $G(\frac{d}{dt})$ is defined as in (5).

It can be verified that the rational QDF Q_Φ is well-defined in the sense that it is uniquely defined regardless of the choice of the factorization (8).

Let the right coprime factorization of $G(\xi)$ over $\mathbb{R}[\xi]$ be given by $G(\xi) = N(\xi)D^{-1}(\xi)$. Then, we see from (3b) that $v \in G(\frac{d}{dt})w$ is equivalent to the existence of $z \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{\mathtt{w}})$ satisfying $v = N(\frac{d}{dt})z$ and $w = D(\frac{d}{dt})z$. Hence, the rational QDF Q_Φ can be rewritten as

$$Q_\Phi(w) = \left\{ Q_\Xi(z) \,\middle|\, \ w = D(\tfrac{d}{dt})z \right\}, \tag{10}$$

where Q_Ξ is the polynomial QDF induced by $\Xi(\zeta,\eta) = N(\zeta)^\top \Sigma N(\eta)$.

In the following, we summarize the calculus of rational QDFs [8].

For two factorizable symmetric matrices Φ_1, $\Phi_2 \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}(\zeta,\eta)$, there holds

$$Q_{\Phi_1+\Phi_2}(w) \subset Q_{\Phi_1}(w) + Q_{\Phi_2}(w) \ \ \forall w \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{\mathtt{w}}). \tag{11}$$

We introduce the notation $\overset{\bullet}{\Phi}(\zeta,\eta) := (\zeta+\eta)\Phi(\zeta,\eta)$. Then, the relation between Q_Φ and $Q_{\overset{\bullet}{\Phi}}$ is given by

$$Q_{\overset{\bullet}{\Phi}}(w) = \left\{ r \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}) \,\middle|\, \exists s \in Q_\Phi(w) \text{ s.t. } r = \tfrac{d}{dt}s \right\}. \tag{12}$$

We say that Q_Φ or $\Phi(\zeta,\eta)$ is *nonnegative*, denoted by $\Phi \geq 0$, if

$$s(t) \geq 0 \ \ \forall t \in \mathbb{R}, \forall s \in Q_\Phi(w), \forall w \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{\mathtt{w}}).$$

A necessary and sufficient condition for the nonnegativity of the rational QDF Q_Φ is that there exists a $K \in \mathbb{R}^{\bullet\times\mathtt{w}}(\xi)$ satisfying

$$\Phi(\zeta,\eta) = K^\top(\zeta)K(\eta).$$

Definition 2. [8] (i) Q_Φ is called *average nonnegative* if

$$\int_{-\infty}^{\infty} s(t)dt \geq 0 \ \forall s \in Q_\Phi(w) \cap \mathfrak{D}, \ \forall w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathtt{w}})..$$

(ii) Q_Φ is called *average nonnegative on* $\mathbb{R}_-$ if

$$\int_{-\infty}^{0} s(t)dt \geq 0 \ \forall s \in Q_\Phi(w) \cap \mathfrak{D}, \ \forall w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathtt{w}}).$$

Necessary and sufficient conditions for the average nonnegativity of Q_Φ is given by the following theorem.

Theorem 1. [8] *Let* $\Phi \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}(\zeta, \eta)$ *satisfy Assumption* 1. *Then, the following statements are equivalent.*

(i) Q_Φ *is average nonnegative (on* $\mathbb{R}_-$*).*
(ii) There exists a symmetric factorizable rational matrix $\Psi \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}(\zeta, \eta)$ *such that*

$$\overset{\bullet}{\Psi} - \Phi \leq 0 \quad (\mathit{and}\, \Psi \geq 0). \tag{13}$$

(iii) There exist a (nonnegative) symmetric factorizable rational matrix $\Psi \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}(\zeta, \eta)$ *and a rational matrix* $F \in \mathbb{R}^{\bullet\times\mathtt{w}}(\xi)$ *such that*

$$\Phi(\zeta, \eta) = (\zeta + \eta)\Psi(\zeta, \eta) + F^\top(\zeta)F(\eta). \tag{14}$$

Moreover, among the solutions to (13), (14)*, there exists* (Ψ, F) *such that* $F(\xi)D(\xi)$ *and* $D^\top(\zeta)\Psi(\zeta, \eta)D(\eta)$ *are polynomial matrices for the right coprime factors* (N, D) *of* $G(\xi)$ *in* (8).

4 Dissipativity Analysis

In this section, we define the dissipativity of a linear differential system in terms of rational QDFs, and derive necessary and sufficient conditions for the dissipativity.

Definition 3. A linear behavior $\mathfrak{B} \in \mathscr{L}^{\mathtt{w}}$ is called *dissipative with respect to* Q_Φ if

$$\int_{-\infty}^{\infty} s(t)dt \geq 0 \, \forall s \in Q_\Phi(w) \cap \mathfrak{D}, \ \forall w \in \mathfrak{B}.$$

$\mathfrak{B}$ is called *dissipative on* $\mathbb{R}_-$ *with respect to* Q_Φ if

$$\int_{-\infty}^{0} s(t)dt \geq 0 \, \forall s \in Q_\Phi(w) \cap \mathfrak{D}, \ \forall w \in \mathfrak{B}.$$

If we view $s(t)$ as *the supply rate* (instantaneous energy flow) into the system $\mathfrak{B}$, the above definition describes the situation that the net energy supplied to $\mathfrak{B}$ from its environment through the "multiplier" $G(\frac{d}{dt})$ over $\mathbb{R}$ ($\mathbb{R}_-$) is nonnegative, namely the system dissipates the supplied energy. It should be noted that, since the supply rate $s(t)$ depends on the initial state of the multiplier $G(\frac{d}{dt})$, $s \in Q_{\Phi}(w)$ is not unique for a given trajectory $w \in \mathfrak{B}$.

Theorem 2. *Let $\Phi \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}(\zeta,\eta)$ be symmetric and factorizable. Let $\mathfrak{B} \in \mathscr{L}^{\mathtt{w}}$ be a controllable behavior whose observable image representation is given by $w = M(\frac{d}{dt})\ell$, $M \in \mathbb{R}^{\mathtt{w}\times 1}[\xi]$. Then, the following statements are equivalent.*

(i) $\mathfrak{B}$ is dissipative (on $\mathbb{R}_-$) with respect to Q_{Φ}.
(ii) There exists a symmetric factorizable rational matrix $\Psi \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}(\zeta,\eta)$ such that

$$\overset{\bullet}{\Psi} - \Phi \overset{\mathfrak{B}}{\leq} 0 \ \ (\textit{and} \ \ \Psi \overset{\mathfrak{B}}{\geq} 0). \tag{15}$$

(iii) There exist a (nonnegative) symmetric factorizable rational matrix $\hat{\Psi} \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}(\zeta,\eta)$ and a rational matrix $\hat{F} \in \mathbb{R}^{\bullet\times\mathtt{w}}(\xi)$ such that

$$M(\zeta)^{\top}\Phi(\zeta,\eta)M(\eta) = (\zeta+\eta)\hat{\Psi}(\zeta,\eta) + \hat{F}^{\top}(\zeta)\hat{F}(\eta). \tag{16}$$

In view of the equivalence (i)⇔(iii) in the above theorem, we immediately obtain the following corollary.

Corollary 1. *Let a rational matrix $\Phi \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}(\zeta,\eta)$ be symmetric and factorizable. If $\mathfrak{B}$ is dissipative on $\mathbb{R}_-$ with respect to Q_{Φ}, then there holds*

$$M(\bar{\lambda})^{\top}\Phi(\bar{\lambda},\lambda)M(\lambda) \geq 0 \ \ \textit{for almost all } \lambda \in \mathbb{C}_+ \tag{17}$$

Proof of Theorem 2. For the limited space, we only give a proof for the dissipativity w.r.t. Q_{Φ}.

(i)⇒(iii): Since $Q_{\Phi}(w) \supseteq Q_{M^T\Phi M}(\ell)$ holds for the solution of $w = M(\frac{d}{dt})\ell$, the dissipativity w.r.t. Q_{Φ} implies the average nonnegativity of $Q_{M^{\top}\Phi M}$. It thus follows from Theorem 1 that there exist a symmetric factorizable matrix $\hat{\Psi} \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}(\zeta,\eta)$ and $\hat{F} \in \mathbb{R}^{\bullet\times\mathtt{w}}(\xi)$ satisfying (16).

(iii)⇒(ii): Suppose that there exist a symmetric factorizable matrix $\hat{\Psi}(\eta,\eta)$ and $\hat{F}(\xi)$ satisfying (16). Since $M(\xi)$ is right prime, there exist rational matrices $\Psi(\zeta,\eta)$, $F(\xi)$ and $X(\zeta,\eta)$ such that $\hat{\Psi}(\zeta,\eta) = M^{\top}(\zeta)\Psi(\zeta,\eta)M(\eta)$, $\hat{F}(\xi) = F(\xi)M(\xi)$, and (16) reduces to

$$(\zeta+\eta)\Psi(\zeta,\eta) - \Phi(\zeta,\eta) = -F(\zeta)^{\top}F(\eta) + X(\eta,\zeta)^{\top}R(\eta) + R(\zeta)^{\top}X(\zeta,\eta),$$

where $R \in \mathbb{R}^{\bullet\times\mathtt{w}}[\xi]$ induces the kernel representation of $\mathfrak{B}$ ($R(\xi)M(\xi) = 0$). Since $\mathfrak{B} = \ker R(\frac{d}{dt})$, it can be shown from the above equation that, for $w \in \mathfrak{B}$, there hold

$$Q_{\overset{\bullet}{\Psi}-\Phi}(w) = Q_{\Delta}(w), \ \ \Delta(\zeta,\eta) := -F(\zeta)^{\top}F(\eta) \leq 0.$$

This implies that $\overset{\bullet}{\Psi} - \Phi$ is nonpositive along $\mathfrak{B}$.

(ii)⇒(i): Let $\overset{\bullet}{\Psi}(\zeta,\eta)$ and $\Phi(\zeta,\eta)$ be factored as $\overset{\bullet}{\Psi}(\zeta,\eta) = G_1(\zeta)^{\top}\Sigma_1 G_1(\eta)$ and $\Phi(\zeta,\eta) = G_2(\zeta)^{\top}\Sigma_2 G_2(\eta)$. Moreover, we introduce the right coprime factorization over $\mathbb{R}[\xi]$ as $\begin{bmatrix} G_1(\xi) \\ G_2(\xi) \end{bmatrix} = \begin{bmatrix} N_1(\xi) \\ N_2(\xi) \end{bmatrix} \hat{D}(\xi)^{-1}$. Since $\Psi(\zeta,\eta)$ is factorizable and since $N_1(\zeta)^\top \Sigma_1 N_1(\eta) = (\zeta+\eta)\hat{D}(\zeta)^\top \Psi(\zeta,\eta)\hat{D}(\eta)$ is a polynomial matrix, $\Xi(\zeta,\eta) := \hat{D}(\zeta)^\top \Psi(\zeta,\eta)\hat{D}(\eta)$ is also a polynomial matrix. Thus, in the same way as (10), $Q_{\overset{\bullet}{\Psi}-\Phi}(w)$ is expressed as

$$\begin{aligned} Q_{\overset{\bullet}{\Psi}-\Phi}(w) &= \left\{ Q_{N_1^\top \Sigma_1 N_1 - N_2^\top \Sigma_2 N_2}(z) \mid \hat{D}(\tfrac{d}{dt})z = w \right\} \\ &= \left\{ Q_{(\zeta+\eta)\Xi}(z) - Q_{N_2^\top \Sigma_2 N_2}(z) \mid \hat{D}(\tfrac{d}{dt})z = w \right\} \\ &= \left\{ \frac{d}{dt} Q_\Xi(z) - Q_{N_2^\top \Sigma_2 N_2}(z) \mid \hat{D}(\tfrac{d}{dt})z = w \right\} \end{aligned}$$

This implies that (15) is equivalent to

$$\frac{d}{dt} Q_\Xi(z)(t) - Q_{N_2^\top \Sigma_2 N_2}(z)(t) \le 0 \;\; \forall t \in \mathbb{R},\; \forall z \in \hat{D}^{-1}(\tfrac{d}{dt})w,\; \forall w \in \mathfrak{B}. \tag{18}$$

Integrating this inequality from $t = -\infty$ to $+\infty$ yields

$$\int_{-\infty}^{\infty} Q_{N_2^\top \Sigma_2 N_2}(z)(t)\,dt \ge 0 \;\; \forall z \in (\hat{D}^{-1}(\tfrac{d}{dt})w) \cap \mathfrak{D},\; \forall w \in \mathfrak{B}. \tag{19}$$

Since $N_2(\xi)$ and $\hat{D}(\xi)$ are not right coprime in general, there holds

$$Q_\Phi(w) \subseteq \left\{ Q_{N_2^\top \Sigma_2 N_2}(z) \,\middle|\, \hat{D}(\tfrac{d}{dt})z = w \right\}.$$

That is, for any $s \in Q_\Phi(w)$, there exists $z \in \hat{D}^{-1}(\frac{d}{dt})w$ satisfying $s = Q_{N_2 \Sigma_2 N_2}(z)$. Therefore, we conclude that (19) implies

$$\int_{-\infty}^{\infty} s(t)\,dt \ge 0 \;\; \forall s \in Q_\Phi(w) \cap \mathfrak{D},\; \forall w \in \mathfrak{B}.$$

This completes the proof of Theorem 2. □

It should be noted that (18) is a version of *dissipation inequality*. In fact, $Q_\Xi(z)$ and $s = Q_{N_2^\top \Sigma_2 N_2}(z)$ serve as *the storage function* and *the supply rate*, respectively, and the inequality in (18) means that the rate of change of the stored energy $Q_\Xi(z)$ in the system $\mathfrak{B}$ and the multiplier $G_2(\frac{d}{dt})$ does not exceed the supply rate $Q_{N_2^\top \Sigma_2 N_2}(z)$.

5 Stability Analysis of Interconnected System

Consider the interconnection of two controllable linear systems $\mathfrak{B}$, $\mathfrak{B}' \in \mathscr{L}^{\mathtt{w}}$ depicted in Fig. 1. Suppose that the behaviors $\mathfrak{B}$ and $\mathfrak{B}'$ are defined by

$$\begin{aligned} \mathfrak{B} &= \left\{ w \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{\mathtt{w}}) \,\middle|\, \exists \ell \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{\mathtt{l}}) \text{ s.t. } w = M(\tfrac{d}{dt})\ell \right\} \\ \mathfrak{B}' &= \left\{ w' \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{\mathtt{w}}) \,\middle|\, \exists \ell' \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{\mathtt{m}}) \text{ s.t. } w' = L(\tfrac{d}{dt})\ell' \right\} \end{aligned}$$

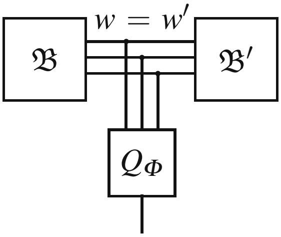


Fig. 1 Stability analysis of interconnected system

where $M \in \mathbb{R}^{\mathtt{w}\times\mathtt{l}}[\xi]$ and $L \in \mathbb{R}^{\mathtt{w}\times\mathtt{m}}[\xi]$. Without loss of generality, we assume that the image representations of both systems are observable, i.e. $M(\lambda)$ and $L(\lambda)$ have full column rank for all $\lambda \in \mathbb{C}$.

By equating $w \in \mathfrak{B}$ and $w' \in \mathfrak{B}'$, the interconnection $\mathfrak{B} \cap \mathfrak{B}'$ is defined as

$$\mathfrak{B} \cap \mathfrak{B}' = \left\{ w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathtt{w}}) \,\middle|\, \exists \ell,\, \ell' \text{ s.t. } w = M(\tfrac{d}{dt})\ell = L(\tfrac{d}{dt})\ell' \right\}. \tag{20}$$

It is easily seen that $\mathfrak{B} \cap \mathfrak{B}'$ is *asymptotically stable* , i.e. $\|w(t)\| \to 0$ $(t \to +\infty)$ for all $w \in \mathfrak{B} \cap \mathfrak{B}'$, if and only if $\begin{bmatrix} M(\lambda) & -L(\lambda) \end{bmatrix}$ has full column rank for all $\lambda \in \mathbb{C}_+$.

The next theorem is the behavioral version of the passivity theorem with rational multipliers and the scaled small gain theorem for a linear interconnected system.

Theorem 3. *Assume that a symmetric rational matrix $\Phi \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}(\zeta, \eta)$ has the factorization of* (8)*, and that $G(\xi)$ has no poles in the closed right-half plane. The interconnection $\mathfrak{B} \cap \mathfrak{B}'$ is asymptotically stable if the following conditions are satisfied.*

(i) $\mathfrak{B}$ *is dissipative on $\mathbb{R}_-$ with respect to $Q_{\Phi-\varepsilon I}$ for some constant $\varepsilon > 0$.*

(ii) $\mathfrak{B}'$ *is dissipative on $\mathbb{R}_-$ with respect to $Q_{-\Phi}$.*

Proof. From Corollary 1, the condition (i) implies

$$M(\bar{\lambda})^\top \Phi(\bar{\lambda}, \lambda) M(\lambda) \geq \varepsilon M(\bar{\lambda})^\top M(\lambda) \quad \forall \lambda \in \mathbb{C}_+. \tag{21}$$

Similarly, it follows from (ii) that

$$-L(\bar{\lambda})^\top \Phi(\bar{\lambda}, \lambda) L(\lambda) \geq 0 \quad \forall \lambda \in \mathbb{C}_+. \tag{22}$$

To prove the asymptotic stability of $\mathfrak{B} \cap \mathfrak{B}'$, we have only to show that $M(\lambda)x = L(\lambda)y$, $\lambda \in \mathbb{C}_+$ implies $(x, y) = (0, 0)$.

Multiplying (21) with x^* and x yields

$$x^* M(\bar{\lambda})^\top \Phi(\bar{\lambda}, \lambda) M(\lambda) x \geq \varepsilon \|M(\lambda)x\|^2 \tag{23}$$

Multiplying (22) with y^* and y yields

$$-y^* L(\bar{\lambda})^\top \Phi(\bar{\lambda}, \lambda) L(\lambda) y \geq 0 \tag{24}$$

Since $M(\lambda)x = L(\lambda)y$, summing up (23) and (24) yields $\varepsilon\|M(\lambda)x\|^2 = 0$. Hence, we get $M(\lambda)x = L(\lambda)y = 0$. Since both $M(\lambda)$ and $L(\lambda)$ are of full column rank by observability assumption, we obtain $(x,y) = (0,0)$. Therefore, we conclude that $\mathfrak{B} \cap \mathfrak{B}'$ is asymptotically stable. □

We conclude this section with another study of Theorem 3 in terms of the inequality (15). By Theorem 2, the conditions (i), (ii) in Theorem 3 ensure that there exist factorizable rational matrices Ψ, $\Theta \in \mathbb{R}^{\mathtt{w}\times\mathtt{w}}(\zeta,\eta)$ such that

$$\overset{\bullet}{\Psi} - \Phi + \varepsilon I \overset{\mathfrak{B}}{\leq} 0, \quad \Psi \overset{\mathfrak{B}}{\geq} 0 \tag{25}$$

$$\overset{\bullet}{\Theta} + \Phi \overset{\mathfrak{B}'}{\leq} 0, \quad \Theta \overset{\mathfrak{B}'}{\geq} 0 \tag{26}$$

Let $q \in Q_{\Psi+\Theta}(w)$ be given for $w \in \mathfrak{B} \cap \mathfrak{B}'$. Since $Q_{\Psi+\Theta}(w) \subseteq Q_{\Psi}(w) + Q_{\Theta}(w)$ from (11), it follows from the second inequalities of (25) and (26) that $Q_{\Psi+\Theta}$ is nonnegative along $\mathfrak{B} \cap \mathfrak{B}'$. Hence, $q(t)$ is nonnegative for all $t \in \mathbb{R}$.

We also see from (11) that $Q_{\overset{\bullet}{\Psi}+\overset{\bullet}{\Theta}}(w) \subseteq Q_{\overset{\bullet}{\Psi}-\Phi+\varepsilon I}(w) + Q_{\overset{\bullet}{\Theta}+\Phi}(w) + Q_{-\varepsilon I}(w)$. Hence, there exist $s \in Q_{\overset{\bullet}{\Psi}-\Phi+\varepsilon I}(w)$ and $r \in Q_{\overset{\bullet}{\Theta}+\Phi}(w)$ such that $\frac{d}{dt}q = s + r - \varepsilon\|w\|^2$, where $s(t) \leq 0$, $r(t) \leq 0$ $\forall t$ hold from the first inequalities of (25) and (26). In summary, we obtain

$$\tfrac{d}{dt}q(t) + \varepsilon\|w(t)\|^2 \leq 0, \quad q(t) \geq 0 \; \forall t \in \mathbb{R}, \; \forall q \in Q_{\Psi+\Theta}(w). \tag{27}$$

Since $q(t)$ is nonnegative for all t, by integrating the above inequality, we obtain

$$\int_0^T \|w(t)\|^2 dt \leq \varepsilon^{-1} \leq q(0).$$

Therefore, we conclude that $w \in \mathfrak{B} \cap \mathfrak{B}'$ is L^2-bounded under the conditions (i), (ii). Additionally, it turns out that $q \in Q_{\Psi+\Theta}(w)$ serves as a Lyapunov function for $\mathfrak{B}$. In fact, it is easily seen from (25), (26) that $q \in Q_{\Psi+\Theta}(w)$ represents the total energy stored in $\mathfrak{B} \cap \mathfrak{B}'$, and hence (27) means that the energy function $q \geq 0$ decays to its steady-state value as times goes to infinity.

6 Conclusions

The rational QDF allows less conservative analysis of dynamical systems in the behavioral framework. An important feature of a rational QDF is that it is a point-to-set map unlike polynomial QDFs. In this paper, we have studied the dissipativity of a linear differential system by using rational QDFs. Based on this analysis, we have derived a behavioral version of the passivity theorem with rational multipliers or the scaled small gain theorem for the stability of an interconnected system.

References

1. Belur, M.N., Trentelman, H.L.: The strict dissipativity synthesis problem and the rank of the coupling QDF. Syst. & Control Letters 51(3-4), 247–258 (2004)
2. Iwasaki, T., Hara, S.: Well-posedness of feedback systems: insights into exact robustness analysis and approximate computations. IEEE Trans. Autom. Control 43(5), 619–630 (1998)
3. Megretski, A., Rantzer, A.: System analysis via integral quadratic constraints. IEEE Trans. Autom. Control 42(6), 819–830 (1997)
4. Peeters, R., Rapisarda, P.: A two-variable approach to solve the polynomial Lyapunov equation. Syst. & Control Letters 42(2), 117–126 (2001)
5. Pendharkar, I., Pillai, H.: Systems with sector bound nonlinearities: a behavioral approach. Syst. & Control Letters 72(2), 112–122 (2007)
6. Polderman, J.W., Willems, J.C.: Introduction to Mathematical Systems Theory: A Behavioral Approach. Springer, Heidelberg (1998)
7. Takaba, K.: Robust stability analysis of uncertain interconnections. SICE Journal Control, Measurement, and System Integration 1(6), 435–442 (2008)
8. Takaba, K., Trentelman, H.L., Willems, J.C.: On rational quadratic differential forms. In: Proceedings of 17th IFAC World Congress, Seoul, pp. 1311–1318 (2008)
9. Trentelman, H.L., Willems, J.C.: H^∞ control in a behavioral context: the full information case. IEEE Trans. Autom. Control 44(3), 521–536 (1999)
10. Willems, J.C., Trentelman, H.L.: On quadratic differential forms. SIAM J. Control and Optim. 36(5), 1703–1749 (1998)
11. Willems, J.C., Trentelman, H.L.: Synthesis of dissipative systems using quadratic differential forms, Parts I and II. IEEE Trans. Autom. Control 47(1), 53–86 (2002)
12. Willems, J.C., Valcher, M.E.: Linear-quadratic control and quadratic differential forms. In: Proc. of 16th IFAC World Congress, Prague, Paper code: Mo-A15-TO/1 (2005)
13. Willems, J.C., Takaba, K.: Dissipativity and stability of interconnections. Int. J. of Robust and Nonlinear Control 17, 563–586 (2007)
14. Willems, J.C., Yamamoto, Y.: Behaviors defined by rational functions. Linear Algebra and its Applications 425(3-4), 226–244 (2007)

On Behavioral Equivalence of Rational Representations

Harry L. Trentelman

Abstract. This article deals with the equivalence of representations of behaviors of linear differential systems. In general, the behavior of a given linear differential system has many different representations. In this paper we restrict ourselves to kernel representations and image representations. Two kernel representations or image representations are called equivalent if they represent one and the same behavior. For kernel representations defined by polynomial matrices, necessary and sufficient conditions for equivalence are well-known. In this paper, we deal with the equivalence of *rational representations*, i.e. kernel and image representations that are defined in terms of *rational matrices*.

1 Introduction

It is a major pleasure and an honor to contribute this article to this book on the occasion of the sixtieth birthday of Yutaka Yamamoto. Recently, Yutaka has been working on the issue of representation of system behaviors. This has resulted in an article, together with Jan C. Willems, in which the very useful concept of *rational representation* of behaviors was introduced and studied, see [12]. It is the subject of rational representations of behaviors that will also be the topic of the present article.

Indeed, an important issue in the behavioral approach to systems and control is the issue of representation. In the behavioral approach, a system is defined in terms of its *behavior*, which is the set of all time trajectories that are compatible with the laws of the system (see [5]). In the context of linear, finite-dimensional, time-invariant systems this leads to the concept of *linear differential system*. A linear differential system is defined to be a system whose behavior is equal to the set of solutions of a finite number of higher order, linear, constant coefficient differential

Harry L. Trentelman
Research Institute of Mathematics and Computing Science, University of Groningen,
P.O. Box 800, 9700 AV Groningen, The Netherlands
e-mail: h.l.trentelman@math.rug.nl

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 239–249.
springerlink.com

equations. This set of differential equations is then called a representation of the behavior, often called a *kernel representation*. It is well known that the behavior of a given linear differential system admits many different kinds of representations. Apart from higher order linear differential equations, the behavior of a linear differential system can be represented for example in terms of finite-dimensional state space models, possibly (but not necessarily) even distinguishing between inputs and outputs (see [5], [10], [9]). Also, if it is controllable, it can be represented as the *image* of a polynomial differential operator (we then speak of an *image representation*). Traditionally, kernel and image representations of linear differential systems involve *polynomial matrices*. Recently, in [12], the concept of *rational representation* was defined and elaborated, extending the class of representations to kernel, hybrid, and image representations involving *rational* matrices.

As noted above, a given linear differential system admits many different representations. Two representations are called *equivalent* if they represent one and the same behavior. The issue of equivalence of representations of behaviors has been studied before, in an input-output framework in [6], [7], [4], [13], [2], and [1], and in a behavioral framework in [5], [10], [8], and [3]. In the present paper, we will study the equivalence of kernel representations and image representation in terms of rational matrices. In particular, we consider the question how the rational matrices appearing in equivalent rational kernel representations and rational image representations are related.

The outline of this article is as follows. In the remainder of this section we will introduce the notation, and review some basic material on polynomial and rational matrices. In Section 2 we will review linear differential systems and their polynomial and rational kernel and image representations. Section 3 deals with rational annihilators of a given behavior, and their application to the problem of equivalence of polynomial and rational kernel representations. Finally, in section 4 we will consider the equivalence of polynomial and rational image representations.

As announced, first a few words about the notation and nomenclature used. We use the standard symbols for the fields of real and complex numbers $\mathbb{R}$ and $\mathbb{C}$. $\mathbb{C}^-$ will denote the open left half complex plane. We use $\mathbb{R}^{\mathtt{n}}$, $\mathbb{R}^{\mathtt{n}\times\mathtt{m}}$, etc. for the real linear spaces of vectors and matrices with components in $\mathbb{R}$. $\mathfrak{C}^{\infty}(\mathbb{R},\mathbb{R}^{\mathtt{w}})$ denotes the set of infinitely often differentiable functions from $\mathbb{R}$ to $\mathbb{R}^{\mathtt{w}}$. We use the notation $\det(A)$, to denote the determinant of a square matrix A.

$\mathbb{R}(\xi)$ will denote the field of real rational functions in the indeterminate ξ. The following subrings of $\mathbb{R}(\xi)$ will play a role is this paper. In the first place, as usual, $\mathbb{R}[\xi]$ will denote the ring of polynomials in the indeterminate ξ with real coefficients. Then, $\mathbb{R}(\xi)_P$ will denote the subring of $\mathbb{R}(\xi)$ of all *proper* real rational functions, i.e. all real rational functions of the form n/d with $n,d \in \mathbb{R}[\xi]$ such that $\deg(n) \leq \deg(d)$. Next, $\mathbb{R}(\xi)_S$ will denote the subring of $\mathbb{R}(\xi)$ of all *stable* real rational functions, i.e. all real rational functions of the form n/d with $n,d \in \mathbb{R}[\xi]$ and d Hurwitz, i.e.. all roots of d lie in the open left half complex plane $\mathbb{C}^-$. Finally, $\mathbb{R}(\xi)_{PS}$ will denote the subring of $\mathbb{R}(\xi)$ of all *proper and stable* real rational functions, i.e. $\mathbb{R}(\xi)_{PS} := \mathbb{R}(\xi)_P \cap \mathbb{R}(\xi)_S$.

We will use $\mathbb{R}(\xi)^{\mathtt{n}}, \mathbb{R}(\xi)^{\mathtt{n}\times\mathtt{m}}, \mathbb{R}[\xi]^{\mathtt{n}}, \mathbb{R}[\xi]^{\mathtt{n}\times\mathtt{m}}$, $\mathbb{R}(\xi)^{\mathtt{n}}_{P}$, $\mathbb{R}(\xi)^{\mathtt{n}\times\mathtt{m}}_{P}$, etc. for the spaces of vectors and matrices with components in $\mathbb{R}(\xi)$, $\mathbb{R}[\xi]$, $\mathbb{R}(\xi)_P$, $\mathbb{R}(\xi)_S$ and $\mathbb{R}(\xi)_{PS}$, respectively. If one, or both, dimensions are unspecified, we will use the notation $\mathbb{R}(\xi)^{\bullet\times\mathtt{m}}$, $\mathbb{R}(\xi)^{\mathtt{n}\times\bullet}$ or $\mathbb{R}(\xi)^{\bullet\times\bullet}$, etc. Elements of $\mathbb{R}(\xi)^{\mathtt{n}\times\mathtt{m}}$ are called *real rational matrices*, elements of $\mathbb{R}[\xi]^{\mathtt{n}\times\mathtt{m}}$ are called *real polynomial matrices*.

Definition 1. A real polynomial matrix $R \in \mathbb{R}[\xi]^{\mathtt{n}\times\mathtt{m}}$ is called *left prime* over $\mathbb{R}[\xi]$ if it has a polynomial right inverse, i.e. if there exists a real polynomial matrix $M \in \mathbb{R}[\xi]^{\mathtt{m}\times\mathtt{n}}$ such that $RM = I$. $R \in \mathbb{R}[\xi]^{\mathtt{n}\times\mathtt{m}}$ is called *right prime* over $\mathbb{R}[\xi]$ if it has a polynomial left inverse, i.e. if there exists a real polynomial matrix $M \in \mathbb{R}[\xi]^{\mathtt{m}\times\mathtt{n}}$ such that $MR = I$. A square polynomial matrix $U \in \mathbb{R}[\xi]^{\mathtt{n}\times\mathtt{n}}$ is called *unimodular* if it is invertible and its inverse is again a polynomial matrix (equivalently: $\det(U)$ is a nonzero constant).

Definition 2. A proper real rational matrix $R \in \mathbb{R}(\xi)^{\mathtt{n}\times\mathtt{m}}_{P}$ is called *left prime* over $\mathbb{R}_P(\xi)$ if it has a proper right inverse, i.e. if there exists a proper real rational matrix $M \in \mathbb{R}(\xi)^{\mathtt{m}\times\mathtt{n}}_{P}$ such that $RM = I$. A stable real rational matrix $R \in \mathbb{R}(\xi)^{\mathtt{n}\times\mathtt{m}}_{S}$ is called *left prime* over $\mathbb{R}_S(\xi)$ if it has a stable right inverse, i.e. if there exists a stable real rational matrix $M \in \mathbb{R}(\xi)^{\mathtt{m}\times\mathtt{n}}_{S}$ such that $RM = I$. A proper stable real rational matrix $R \in \mathbb{R}(\xi)^{\mathtt{n}\times\mathtt{m}}_{PS}$ is called *left prime* over $\mathbb{R}(\xi)_{PS}$ if it has a proper stable right inverse, i.e. if there exists a proper real rational matrix $M \in \mathbb{R}(\xi)^{\mathtt{m}\times\mathtt{n}}_{PS}$ such that $RM = I$. In the same way, the notion of *right primeness* over these various subrings can be defined.

Equivalent characterizations of left and right primeness can be found in [12].

2 Linear Differential Systems

In this section we will review the basic material on linear differential systems and their polynomial and rational representations.

In the behavioral approach to linear systems, a dynamical system is given by a triple $\Sigma = (\mathbb{R}, \mathbb{R}^{\mathtt{w}}, \mathfrak{B})$, where $\mathbb{R}$ is the time axis, $\mathbb{R}^{\mathtt{w}}$ is the signal space, and the *behavior* $\mathfrak{B}$ is a linear subspace of $\mathfrak{C}^{\infty}(\mathbb{R}, \mathbb{R}^{\mathtt{w}})$ consisting of all solutions of a set of higher order, linear, constant coefficient differential equations. Such a triple is called a *linear differential system*. The set of all linear differential systems with $\mathtt{w}$ variables is denoted by $\mathfrak{L}^{\mathtt{w}}$.

For any linear differential system $\Sigma = (\mathbb{R}, \mathbb{R}^{\mathtt{w}}, \mathfrak{B})$ there exists a real polynomial matrix R with $\mathtt{w}$ columns, i.e. $R \in \mathbb{R}[\xi]^{\bullet\times\mathtt{w}}$, such that $\mathfrak{B}$ is equal to the space of solutions of

$$R(\tfrac{\mathrm{d}}{\mathrm{d}t})w = 0. \tag{1}$$

If a behavior $\mathfrak{B}$ is represented by $R(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ (or: $\mathfrak{B} = \ker(R)$), with $R(\xi)$ a real polynomial matrix, then we call this a *polynomial kernel representation* of $\mathfrak{B}$. If R has $\mathtt{p}$ rows, then the polynomial kernel representation is said to be *minimal* if every polynomial kernel representation of $\mathfrak{B}$ has at least $\mathtt{p}$ rows. A given polynomial kernel representation, $\mathfrak{B} = \ker(R)$, is minimal if and only if the polynomial matrix R has full row rank (see [5, Theorem 3.6.4]). The number of rows in any minimal polynomial kernel representation of $\mathfrak{B}$, denoted by $\mathtt{p}(\mathfrak{B})$, is called the *output* *cardinality* of $\mathfrak{B}$. This number corresponds to the number of outputs in any input/output representation of $\mathfrak{B}$. For a detailed exposition of polynomial representations of behaviors, we refer to [5].

Recently, in [12], representations of linear differential systems using *rational matrices* instead of polynomial matrices were introduced. In [12], a meaning was given to the equation $R(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$, where $R(\xi)$ is a given real *rational* matrix. In order to do this, we need the concept of left coprime factorization over $\mathbb{R}[\xi]$.

Definition 3. Let R be a real rational matrix. The pair of real polynomial matrices (P,Q) is called a *left coprime factorization of R over* $\mathbb{R}[\xi]$ if

1. $\det(P) \neq 0$,
2. $R = P^{-1}Q$,
3. the matrix $(P(\lambda)\ Q(\lambda))$ has full row rank for all $\lambda \in \mathbb{C}$.

A meaning to the equation

$$R(\tfrac{\mathrm{d}}{\mathrm{d}t})w = 0, \tag{2}$$

with $R(\xi)$ a real rational matrix is then given as follows: Let (P,Q) be a left coprime factorization of R over $\mathbb{R}[\xi]$. Then we define:

Definition 4. Let $w \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}^{\mathtt{w}})$. Then we define w to be a solution of (2) if it satisfies the differential equation $Q(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$.

It can be proven that the space of solutions defined in this way is independent of the particular left coprime factorization. Hence (2) represents the linear differential system $\Sigma = (\mathbb{R},\mathbb{R}^{\mathtt{w}},\ker(Q)) \in \mathfrak{L}^{\mathtt{w}}$.

Since the behavior $\mathfrak{B}$ of the system Σ is the central item, often we will speak about the system $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}$ (instead of $\Sigma \in \mathfrak{L}^{\mathtt{w}}$). If a behavior $\mathfrak{B}$ is represented by $R(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ (or: $\mathfrak{B} = \ker(R)$), with $R(\xi)$ a real rational matrix, then we call this a *rational kernel representation of* $\mathfrak{B}$. If R has $\mathtt{p}$ rows, then the rational kernel representation is called *minimal* if every rational kernel representation of $\mathfrak{B}$ has at least $\mathtt{p}$ rows. It can be shown that a given rational kernel representation $\mathfrak{B} = \ker(R)$ is minimal if and only if the rational matrix R has full row rank. As in the polynomial case, every $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}$ admits a minimal rational kernel representation. The number of rows in any minimal rational kernel representation of $\mathfrak{B}$ is equal to the number of rows in any minimal polynomial kernel representation of $\mathfrak{B}$, and therefore equal to $\mathtt{p}(\mathfrak{B})$, the output cardinality of $\mathfrak{B}$. In general, if $\mathfrak{B} = \ker(R)$ is a rational kernel representation, then $\mathtt{p}(\mathfrak{B}) = \mathrm{rank}(R)$. This follows immediately from the corresponding result for polynomial kernel representations (see [5]).

In this paper we will also use the notion of right coprime factorization over $\mathbb{R}[\xi]$:

Definition 5. Let R be a real rational matrix. The pair of matrices (M,P) is called a *right coprime factorization of R over* $\mathbb{R}[\xi]$ if

1. $\det(P) \neq 0$,
2. $R = MP^{-1}$,
3. the matrix $\begin{pmatrix} M(\lambda) \\ P(\lambda) \end{pmatrix}$ has full column rank for all $\lambda \in \mathbb{C}$.

Definition 6. A behavior $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}$ is said to be *controllable*, if for all $w_1, w_2 \in \mathfrak{B}$, there exists $T \geq 0$, and $w \in \mathfrak{B}$, such that $w(t) = w_1(t)$ for $t \leq 0$, and $w(t) = w_2(t-T)$ for $t \geq T$. It is *stabilizable*, if for every $w \in \mathfrak{B}$, there exists $w' \in \mathfrak{B}$ such that $w'(t) = w(t)$ for $t \leq 0$, and $\lim_{t \to \infty} w'(t) = 0$.

The subset of $\mathfrak{L}^{\mathtt{w}}$ of all controllable behaviors is denoted by $\mathfrak{L}^{\mathtt{w}}_{\text{cont}}$. Clearly, if $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}_{\text{cont}}$ then it is stabilizable.

It is well-known that a behavior $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}$ is controllable if and only if there exists a real polynomial matrix $M \in \mathbb{R}[\xi]^{\mathtt{w} \times \bullet}$ such that

$$\mathfrak{B} = \{ w \in \mathfrak{C}^{\infty}(\mathbb{R}, \mathbb{R}^{\mathtt{w}}) \mid \text{ there exists } \ell \in \mathfrak{C}^{\infty}(\mathbb{R}, \mathbb{R}^{\bullet}) \text{ such that } w = M(\tfrac{\mathrm{d}}{\mathrm{d}t})\ell \} \tag{3}$$

The representation (3) is called a *polynomial image representation* of $\mathfrak{B}$, and we will write $\mathfrak{B} = \operatorname{im}(M)$. It can be shown that the polynomial matrix M can be chosen of full column rank. Even more, M can be chosen to be right prime over $\mathbb{R}[\xi]$, equivalently, $M(\lambda)$ has full column rank for all $\lambda \in \mathbb{C}$. In that case, in (3) the latent variable ℓ is uniquely determined by the manifest variable w, and the image representation is called *observable*.

In [12], also the concept of *rational image representation* was introduced. We will give a brief review here. Let $H(\xi)$ be a real rational matrix , and consider the equation

$$w = H(\tfrac{\mathrm{d}}{\mathrm{d}t})\ell. \tag{4}$$

Of course (4) should be interpreted as

$$\begin{pmatrix} I & -H(\tfrac{\mathrm{d}}{\mathrm{d}t}) \end{pmatrix} \begin{pmatrix} w \\ \ell \end{pmatrix} = 0,$$

in the context of (2). If $H = D^{-1}N$ is a left coprime factorization over $\mathbb{R}[\xi]$ then $D^{-1}\begin{pmatrix} D & -N \end{pmatrix}$ is a left coprime factorization of $\begin{pmatrix} I & -H \end{pmatrix}$ and therefore (w, ℓ) satisfies (4) if and only if $D(\tfrac{\mathrm{d}}{\mathrm{d}t})w = N(\tfrac{\mathrm{d}}{\mathrm{d}t})\ell$. For a given $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}$, the representation

$$\mathfrak{B} = \{ w \in \mathfrak{C}^{\infty}(\mathbb{R}, \mathbb{R}^{\mathtt{w}}) \mid \text{ there exists } \ell \in \mathfrak{C}^{\infty}(\mathbb{R}, \mathbb{R}^{\bullet}) \text{ such that } w = H(\tfrac{\mathrm{d}}{\mathrm{d}t})\ell \}, \tag{5}$$

with $H \in \mathbb{R}(\xi)^{\mathtt{w} \times \bullet}$, is called a *rational image representation*. In that case, we write $\mathfrak{B} = \operatorname{im}(H)$. It was shown in [12] that $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}$ admits a rational image representation if and only if it is controllable.

3 Equivalence of Kernel Representations

In [12], several results on the representation of linear differential systems using rational matrices over the rings $\mathbb{R}(\xi)_P$, $\mathbb{R}(\xi)_S$ and $\mathbb{R}(\xi)_{PS}$ were obtained. Given $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}$, by definition it admits a polynomial kernel representation $R(\tfrac{\mathrm{d}}{\mathrm{d}t})w = 0$, with $R \in \mathbb{R}[\xi]^{\bullet \times \mathtt{w}}$. A basic result from [12] states that any $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}$ also admits a rational kernel representation $R(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$, with R proper and stable, i.e. $R \in \mathbb{R}(\xi)_{PS}^{\bullet\times\mathtt{w}}$ (see [12, Proposition 2]). Furthermore we quote Theorem 5 from [12]:

Theorem 1. *Let $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}$. Then we have:*

1. *There exists $R \in \mathbb{R}(\xi)_{P}^{\bullet\times\mathtt{w}}$, left prime over $\mathbb{R}(\xi)_P$, such that $\mathfrak{B}$ is represented by $R(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$.*
2. *There exists $R \in \mathbb{R}(\xi)_{S}^{\bullet\times\mathtt{w}}$, left prime over $\mathbb{R}(\xi)_S$, such that $\mathfrak{B}$ is represented by $R(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ if and only if $\mathfrak{B}$ is stabilizable.*
3. *There exists $R \in \mathbb{R}(\xi)_{PS}^{\bullet\times\mathtt{w}}$, left prime over $\mathbb{R}(\xi)_{PS}$, such that $\mathfrak{B}$ is represented by $R(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ if and only if $\mathfrak{B}$ is stabilizable.*
4. *There exists $R \in \mathbb{R}[\xi]^{\bullet\times\mathtt{w}}$, left prime over $\mathbb{R}[\xi]$, such that $\mathfrak{B}$ is represented by $R(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ if and only if $\mathfrak{B}$ is controllable.*

A basic issue in the behavioral theory is the *equivalence of representations*. Two representations are called equivalent if they represent one and the same behavior. Conditions for two minimal polynomial kernel representations to be equivalent are well known (see [5, Theorem 3.6.4]):

Proposition 1. *Let $\mathfrak{B}_1, \mathfrak{B}_2 \in \mathfrak{L}^{\mathtt{w}}$. Let $R_1, R_2 \in \mathbb{R}[\xi]^{\bullet\times\mathtt{w}}$ be such that $R_1(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ and $R_2(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ are minimal polynomial kernel representations of $\mathfrak{B}_1$ and $\mathfrak{B}_2$ respectively. Then $\mathfrak{B}_1 = \mathfrak{B}_2$ if and only if there exists a unimodular polynomial matrix U such that $R_2 = UR_1$.*

In the sequel we want to address the question how this result can be extended to *rational* kernel representations.

First, from [12, Section 7], we recall the concepts of polynomial and rational annihilators of a given behavior. Here, we introduce the notion of $\mathfrak{R}$-annihilator, where $\mathfrak{R}$ is any of the three subrings $\mathbb{R}(\xi)_{PS}$, $\mathbb{R}(\xi)_P$ or $\mathbb{R}(\xi)_S$ of $\mathbb{R}(\xi)$:

Definition 7. Let $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}$.

1. $n \in \mathbb{R}[\xi]^{1\times\mathtt{w}}$ is called a *polynomial annihilator* of $\mathfrak{B}$ if $n(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ for all $w \in \mathfrak{B}$.
2. $n \in \mathbb{R}(\xi)^{1\times\mathtt{w}}$ is called a *rational annihilator* of $\mathfrak{B}$ if $n(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ for all $w \in \mathfrak{B}$.
3. Let $\mathfrak{R}$ denote any of the subrings $\mathbb{R}(\xi)_{PS}$, $\mathbb{R}(\xi)_P$ or $\mathbb{R}(\xi)_S$ of $\mathbb{R}(\xi)$. Any rational annihilator $n \in \mathfrak{R}^{1\times\mathtt{w}}$ of $\mathfrak{B}$ is called an $\mathfrak{R}$-*annihilator* of $\mathfrak{B}$.

We denote the set of polynomial annihilators of $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}$ by $\mathfrak{B}^{\perp_{\mathbb{R}[\xi]}}$. The set of rational anihilators of $\mathfrak{B}$ is denoted by $\mathfrak{B}^{\perp_{\mathbb{R}(\xi)}}$. For any of the subrings $\mathfrak{R} = \mathbb{R}(\xi)_{PS}$, $\mathbb{R}(\xi)_P$ or $\mathbb{R}(\xi)_S$, the set of $\mathfrak{R}$-annihilators is denoted by $\mathfrak{B}^{\perp_{\mathfrak{R}}}$. It is a well-known result that for $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}$, $\mathfrak{B}^{\perp_{\mathbb{R}[\xi]}}$ is a finitely generated submodule of the $\mathbb{R}[\xi]$-module $\mathbb{R}[\xi]^{1\times\mathtt{w}}$. Moreover, if $\mathfrak{B} = \ker(R)$ is a minimal polynomial kernel representation, then this submodule is generated by the rows of R. In the context of rational representations one needs to impose controllability:

Theorem 2. *Let $\mathfrak{R}$ denote any of the subrings $\mathbb{R}(\xi)_{PS}$, $\mathbb{R}(\xi)_P$ or $\mathbb{R}(\xi)_S$.*

1. *Let $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}$. Then $\mathfrak{B}^{\perp_{\mathbb{R}(\xi)}}$ is a subspace of the $\mathbb{R}(\xi)$-linear vector space $\mathbb{R}(\xi)^{1\times\mathtt{w}}$ if and only if $\mathfrak{B}$ is controllable. In that case, if $R(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ is a minimal rational kernel representation of $\mathfrak{B}$, then the rows of R form a basis of $B^{\perp_{\mathbb{R}(\xi)}}$.*
2. *Let $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}$ be controllable, and represented by $R(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$, where $R \in \mathfrak{R}^{\bullet\times\mathtt{w}}$ is left prime over $\mathfrak{R}$. Then $\mathfrak{B}^{\perp_{\mathfrak{R}}}$ is a submodule of the $\mathfrak{R}$-module $\mathfrak{R}^{1\times\mathtt{w}}$, and the rows of R form a basis of $\mathfrak{B}^{\perp_{\mathfrak{R}}}$.*

Proof. 1.) The first statement is the content of statement 1 of theorem 11 in [12]. Let $R = P^{-1}Q$ be a left coprime factorization over $\mathbb{R}[\xi]$ of R. Then $\mathfrak{B} = \ker(Q)$ is a minimal polynomial kernel representation. Let $n \in \mathfrak{B}^{\perp_{\mathbb{R}(\xi)}}$. Then by Def. 7, $n(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ for all $w \in \mathfrak{B}$. Let $n = u^{-1}v$ be a left coprime factorization of n over $\mathbb{R}[\xi]$. Then by definition we have $n(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ for all $w \in \mathfrak{B}$ if and only if $v(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ for all $w \in \mathfrak{B}$. Thus, by Def. 7, $v \in \mathfrak{B}^{\perp_{\mathbb{R}[\xi]}}$. Consequently, there exists a $l \in \mathbb{R}[\xi]^{1\times\bullet}$ such that $v = lQ$. Hence $n = u^{-1}v = u^{-1}lQ = (u^{-1}lP)(P^{-1}Q) = (u^{-1}lP)R$. Define $m := u^{-1}lP$. Then we have $n = mR$. Hence the rows of R span the subspace $B^{\perp_{\mathbb{R}(\xi)}}$ of the $\mathbb{R}(\xi)$-linear vector space $\mathbb{R}(\xi)^{1\times\mathtt{w}}$. Finally, as $\mathfrak{B} = \ker(R)$ is a minimal rational kernel representation, the rows of R are linearly independent over $\mathbb{R}(\xi)$. We conclude then that these rows form a basis of $\mathfrak{B}^{\perp_{\mathbb{R}(\xi)}}$.

2.) If $\mathfrak{B}$ is controllable, then $\mathfrak{B}^{\perp_{\mathfrak{R}}}$ forms a submodule of the $\mathfrak{R}$-module $\mathfrak{R}^{1\times\mathtt{w}}$. This can be proven along the same lines as the proof of Theorem 11 in [12].

Let $R = P^{-1}Q$ be a left coprime factorization over $\mathbb{R}[\xi]$ of R. Then $\mathfrak{B} = \ker(Q)$ is a minimal polynomial kernel representation. Let $n \in \mathfrak{B}^{\perp_{\mathfrak{R}}}$. Then by Def. 7, $n(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ for all $w \in \mathfrak{B}$. Let $n = u^{-1}v$ be a left coprime factorization of n over $\mathbb{R}[\xi]$. Then $v(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ for all $w \in \mathfrak{B}$. Thus, by Def. 7, $v \in \mathfrak{B}^{\perp_{\mathbb{R}[\xi]}}$. Consequently, there exists a $l \in \mathbb{R}[\xi]^{1\times\bullet}$ such that $v = lQ$. Hence $n = u^{-1}v = u^{-1}lQ = (u^{-1}lP)(P^{-1}Q) = (u^{-1}lP)R$. Define $m := u^{-1}lP$. Then we have

$$n = mR. \tag{6}$$

As R is left prime over $\mathfrak{R}$, there exists $M \in \mathfrak{R}^{\mathtt{w}\times\bullet}$ such that $RM = I$. Multiplying (6) on both sides with M we obtain $nM = mRM = m$. As $n \in \mathfrak{R}^{1\times\mathtt{w}}$ and $M \in \mathfrak{R}^{\mathtt{w}\times\bullet}$, we conclude that $m \in \mathfrak{R}^{1\times\bullet}$. Hence the rows of R span the submodule $\mathfrak{B}^{\perp_{\mathfrak{R}}}$. Finally, as $\mathfrak{B} = \ker(R)$ is a minimal rational kernel representation, the rows of R are linearly independent over $\mathbb{R}(\xi)$ so also over $\mathfrak{R}$. We conclude then that these rows form a basis of $\mathfrak{B}^{\perp_{\mathfrak{R}}}$. □

The following theorem addresses the question under what conditions two minimal rational kernel representations represent the same controllable behavior:

Theorem 3. *Let $\mathfrak{B}_1, \mathfrak{B}_2 \in \mathfrak{L}^{\mathtt{w}}_{\text{cont}}$. Let $R_1, R_2 \in \mathbb{R}(\xi)^{\bullet\times\mathtt{w}}$ be such that $R_1(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ and $R_2(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ are minimal rational kernel representations of $\mathfrak{B}_1$ and $\mathfrak{B}_2$*

respectively. Then $\mathfrak{B}_1 = \mathfrak{B}_2$ *if and only if there exists a square, nonsingular rational matrix* W *such that* $R_1 = WR_2$*.*

Proof. As $\mathfrak{B}_1 = \mathfrak{B}_2$ we have $\mathfrak{B}_1^{\perp_{\mathbb{R}(\xi)}} = \mathfrak{B}_2^{\perp_{\mathbb{R}(\xi)}} =: \mathfrak{L}$. From Lemma 2, the rows of R_1 and R_2 both form a basis for the subspace $\mathfrak{L}$ of $\mathbb{R}(\xi)^{1\times \mathtt{w}}$. Then, from basic linear algebra, there exists a square, nonsingular rational matrix W such that $R_1 = WR_2$.

Conversely, let $R_1 = P_1^{-1}Q_1,\ R_2 = P_2^{-1}Q_2$ be left coprime factorizations over $\mathbb{R}[\xi]$ of R_1 and R_2. Let $W = LM^{-1}$ be a right coprime factorization over $\mathbb{R}[\xi]$ of W. Then both L and M are nonsingular. By definition we have $\mathfrak{B}_1 = \ker(Q_1)$ and $\mathfrak{B}_2 = \ker(Q_2)$. Then,

$$\begin{aligned} R_1 = WR_2 &\iff P_1^{-1}Q_1 = LM^{-1}P_2^{-1}Q_2 \\ &\iff L^{-1}P_1^{-1}Q_1 = M^{-1}P_2^{-1}Q_2 \\ &\iff (P_1L)^{-1}Q_1 = (P_2M)^{-1}Q_2. \end{aligned}$$

Since $\mathfrak{B}_1$ and $\mathfrak{B}_2$ are controllable behaviors both $Q_1(\lambda)$ and $Q_2(\lambda)$ have full row rank for all $\lambda \in \mathbb{C}$. This implies that $(P_1L \ \ Q_1)(\lambda)$ and $(P_2M \ \ Q_2)(\lambda)$ have full row rank for all $\lambda \in \mathbb{C}$. Define $\tilde{R} := (P_1L)^{-1}Q_1 = (P_2M)^{-1}Q_2$. This displays two left coprime factorizations of $\tilde{R}$, so $\mathfrak{B}_1 = \ker(Q_1) = \ker(Q_2) = \mathfrak{B}_2$. □

Next, we address the question under which conditions two minimal rational kernel representations over any of the subrings $\mathfrak{R} = \mathbb{R}(\xi)_{PS}$, $\mathbb{R}(\xi)_P$ or $\mathbb{R}(\xi)_S$ represent one and the same controllable behavior. In this case we have:

Theorem 4. *Let* $\mathfrak{B}_1, \mathfrak{B}_2 \in \mathfrak{L}^{\mathtt{w}}_{\mathrm{cont}}$. *Let* $R_1, R_2 \in \mathfrak{R}^{\bullet\times \mathtt{w}}$ *be left prime over* $\mathfrak{R}$*, and such that* $R_1(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ *and* $R_2(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ *are minimal rational kernel representations of* $\mathfrak{B}_1$ *and* $\mathfrak{B}_2$ *respectively. Then* $\mathfrak{B}_1 = \mathfrak{B}_2$ *if and only if there exists a square, nonsingular real rational matrix* $W \in \mathfrak{R}^{\bullet\times\bullet}$*, with* $W^{-1} \in \mathfrak{R}^{\bullet\times\bullet}$ *such that* $R_1 = WR_2$*.*

Note that $W \in \mathfrak{R}^{\bullet\times\bullet}$ in the above theorem should hence have an inverse which is again an element of $\mathfrak{R}^{\bullet\times\bullet}$. This condition can be restated as saying that W should be a *unimodular* element of $\mathfrak{R}^{\bullet\times\bullet}$. In particular, for the ring $\mathbb{R}(\xi)_{PS}$, W should be bi-proper and bi-stable, for the ring $\mathbb{R}(\xi)_P$, W should be bi-proper, and for the ring $\mathbb{R}(\xi)_S$, W should be bi-stable.

Proof. As $\mathfrak{B}_1 = \mathfrak{B}_2$ we have $\mathfrak{B}_1^{\perp_{\mathfrak{R}}} = \mathfrak{B}_2^{\perp_{\mathfrak{R}}} =: \mathfrak{M}$. From Lemma 2, the rows of R_1 and R_2 both form a basis for the module $\mathfrak{M}$. Then from the theory of modules we conclude that there exists a square, nonsingular real rational matrix $W \in \mathfrak{R}^{\bullet\times\bullet}$ with $W^{-1} \in \mathfrak{R}^{\bullet\times\bullet}$, such that $R_1 = WR_2$.

The proof of the converse is similar to the corresponding part of the proof of Theorem 3. □

4 Equivalence of Image Representations

As noted in Section 2, a given behavior $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}$ admits a polynomial image representation if and only it is controllable. In fact, we quote Theorem 9 from [12]:

Theorem 5. *Let $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}$. Let $\mathfrak{R}$ be any of the three subrings $\mathbb{R}(\xi)_{PS}$, $\mathbb{R}(\xi)_P$ or $\mathbb{R}(\xi)_S$ of $\mathbb{R}(\xi)$. Then the following statements are equivalent*

1. *$\mathfrak{B}$ is controllable,*
2. *$\mathfrak{B}$ admits a polynomial image representation,*
3. *$\mathfrak{B}$ admits a polynomial image representation $\mathfrak{B} = \mathrm{im}(M)$ with $M \in \mathbb{R}[\xi]^{\mathtt{w}\times\bullet}$ right prime over $\mathbb{R}[\xi]$,*
4. *$\mathfrak{B}$ admits a rational image representation,*
5. *$\mathfrak{B}$ admits a rational image representation $\mathfrak{B} = \mathrm{im}(M)$ with $M \in \mathfrak{R}^{\mathtt{w}\times\bullet}$ right prime over $\mathfrak{R}$.*

We will now study the problem of equivalence of image representations. For this, the following result will be useful. The result states that right coprime factorization of a rational image representation leads to a polynomial image representation.

Lemma 1. *Let $\mathfrak{B} \in \mathfrak{L}^{\mathtt{w}}_{\mathrm{cont}}$. Let $H \in R(\xi)^{\mathtt{w}\times\bullet}$ be such that $\mathfrak{B} = \mathrm{im}(H)$. Let $H = MP^{-1}$ be a right coprime factorization over $\mathbb{R}[\xi]$. Then $\mathfrak{B}$ =*im*(M).*

Proof. Let $H = D^{-1}N$ be a left coprime factorization over $\mathbb{R}[\xi]$. It is well-known that $\ker\begin{pmatrix} D & N\end{pmatrix} = \mathrm{im}\begin{pmatrix} M \\ -P \end{pmatrix}$. Thus we obtain

$$\begin{aligned}\mathfrak{B} &= \{w \in \mathfrak{C}^{\infty}(\mathbb{R},\mathbb{R}^{\mathtt{w}}) \mid \exists \ell \text{ such that } D(\tfrac{\mathrm{d}}{\mathrm{d}t})w = N(\tfrac{\mathrm{d}}{\mathrm{d}t})\ell\} \\ &= \{w \in \mathfrak{C}^{\infty}(\mathbb{R},\mathbb{R}^{\mathtt{w}}) \mid \exists \ell,\ell' \text{ such that } \begin{pmatrix} w \\ \ell \end{pmatrix} = \begin{pmatrix} M(\frac{\mathrm{d}}{\mathrm{d}t}) \\ -P(\frac{\mathrm{d}}{\mathrm{d}t}) \end{pmatrix} \ell'\} \\ &= \{w \in \mathfrak{C}^{\infty}(\mathbb{R},\mathbb{R}^{\mathtt{w}}) \mid \exists \ell' \text{ such that } w = M(\tfrac{\mathrm{d}}{\mathrm{d}t})\ell'\}.\end{aligned}$$ □

We will now first study the question under which conditions two polynomial image representations are equivalent, i.e. represent the same behavior.

Theorem 6. *1. Let $\mathfrak{B}_1, \mathfrak{B}_2 \in \mathfrak{L}^{\mathtt{w}}_{\mathrm{cont}}$. Let $M_1, M_2 \in \mathbb{R}[\xi]^{\mathtt{w}\times\bullet}$ have full column rank, and be such that $\mathfrak{B}_1 = \mathrm{im}(M_1)$ and $\mathfrak{B}_2 = \mathrm{im}(M_2)$. Then $\mathfrak{B}_1 = \mathfrak{B}_2$ if and only if there exists a nonsingular rational matrix R such that $M_2 = M_1 R$.*

2. Let $\mathfrak{B}_1, \mathfrak{B}_2 \in \mathfrak{L}^{\mathtt{w}}_{\mathrm{cont}}$. Let $M_1, M_2 \in \mathbb{R}[\xi]^{\mathtt{w}\times\bullet}$ be right prime over $\mathbb{R}[\xi]$, and such that $\mathfrak{B}_1 = \mathrm{im}(M_1)$ and $\mathfrak{B}_2 = \mathrm{im}(M_2)$. Then $\mathfrak{B}_1 = \mathfrak{B}_2$ if and only if there exists a unimodular polynomial matrix U such that $M_2 = M_1 U$.

Proof. We first prove the 'only if' part of statement 2. By right primeness, both $M_1(\lambda)$ and $M_2(\lambda)$ have full column rank for all $\lambda \in \mathbb{C}$, so correspond to observable image representations. From $\mathfrak{B}_1 = \mathfrak{B}_2$ it follows that also the orthogonal complements coincide, i.e. $\mathfrak{B}_1^{\perp} = \mathfrak{B}_2^{\perp}$ (see [11]). By observability we have $\mathfrak{B}_i^{\perp} = \ker(M_i^{\sim})$, where $M_i^{\sim}(\xi) := M_i^{\top}(-\xi)$ $(i = 1,2)$. By Proposition 1 there exists a unimodular polynomial matrix V such that $M_2^{\sim} = VM_1^{\sim}$. This implies $M_2 = M_1U$, with $U := V^{\sim}$ again unimodular.

Next, we prove the 'only if' part of statement 1. Both M_1 and M_2 have full column rank. Hence, we can factorize $M_i = \overline{M}_iR_i$, with $\overline{M_i}$ right prime over $\mathbb{R}[\xi]$ and R_i a nonsingular polynomial matrix $(i = 1,2)$. By nonsingularity, $R_i(\frac{\mathrm{d}}{\mathrm{d}t})$ is surjective, and therefore $\mathrm{im}(M_i) = \mathrm{im}(\overline{M}_i)$ $(i = 1,2)$. Consequently, $\mathfrak{B}_1 = \mathfrak{B}_2$ implies $\mathrm{im}(\overline{M}_1) = \mathrm{im}(\overline{M}_2)$. Then, by the 'only if' part of statement 2, there exists a unimodular polynomial matrix U such that $\overline{M}_2 = \overline{M}_1U$. This implies $M_2 = M_1R$, with $R := R_1^{-1}UR_2$.

Finally we prove the 'if' part of statement 1. Assume that $M_2 = M_1R$ with R a nonsingular rational matrix. Let $R = KL^{-1}$ be a right coprime factorization of R over $\mathbb{R}[\xi]$. Then we have $M_2L = M_1K$, with K and L nonsingular polynomial matrices. Again by surjectivity of $L(\frac{\mathrm{d}}{\mathrm{d}t})$ and $K(\frac{\mathrm{d}}{\mathrm{d}t})$, we obtain $\mathfrak{B}_1 = \mathrm{im}(M_1) = \mathrm{im}(M_1K) = \mathrm{im}(M_2L) = \mathrm{im}(M_2) = \mathfrak{B}_2$. This also proves the 'if' part of statement 2. □

Next, we consider controllable behaviors represented by rational image representations.

Theorem 7. *Let $\mathfrak{B}_1, \mathfrak{B}_2 \in \mathfrak{L}^{\mathtt{w}}_{\mathrm{cont}}$. Let $H_1, H_2 \in \mathbb{R}(\xi)^{\mathtt{w}\times\bullet}$ have full column rank, and be such that $\mathfrak{B}_1 = \mathrm{im}(H_1)$ and $\mathfrak{B}_2 = \mathrm{im}(H_2)$. Then $\mathfrak{B}_1 = \mathfrak{B}_2$ if and only if there exists a nonsingular rational matrix R such that $H_2 = H_1R$.*

Proof. Let $H_i = M_iP_i^{-1}$ be a right coprime factorization over $\mathbb{R}[\xi]$. Then by Lemma 1, $\mathfrak{B}_i = \mathrm{im}(M_i)$ $(i = 1,2)$. By Theorem 6, $\mathfrak{B}_1 = \mathfrak{B}_2$ implies that there exists a nonsingular rational matrix $\overline{R}$ such that $M_2 = M_1\overline{R}$. Thus $H_2 = H_1R$, with $R := P_1\overline{R}P_2^{-1}$ nonsingular. Conversely, if $H_2 = H_1R$ then $M_2 = M_1P_1^{-1}RP_2$. Then, by Theorem 6, $\mathrm{im}(M_1) = \mathrm{im}(M_2)$ so $\mathfrak{B}_1 = \mathfrak{B}_2$. □

The above results immediately yield the following:

Theorem 8. *Let $\mathfrak{R}$ be any of the subrings $\mathbb{R}(\xi)_{PS}$, $\mathbb{R}(\xi)_P$ or $\mathbb{R}(\xi)_S$ of $\mathbb{R}(\xi)$ Let $\mathfrak{B}_1, \mathfrak{B}_2 \in \mathfrak{L}^{\mathtt{w}}_{\mathrm{cont}}$. Let $H_1, H_2 \in \mathfrak{R}^{\mathtt{w}\times\bullet}$ be right prime over $\mathfrak{R}$, and such that $\mathfrak{B}_1 = \mathrm{im}(H_1)$ and $\mathfrak{B}_2 = \mathrm{im}(H_2)$. Then $\mathfrak{B}_1 = \mathfrak{B}_2$ if and only if there exists a nonsingular rational matrix $R \in \mathfrak{R}^{\bullet\times\bullet}$, with $R^{-1} \in \mathfrak{R}^{\bullet\times\bullet}$, such that $H_2 = H_1R$.*

Proof. Assume that $\mathfrak{B}_1 = \mathfrak{B}_2$. By Theorem 7 there exists a nonsingular rational matrix R such that $H_2 = H_1 R$, so also $H_1 = H_2 R^{-1}$. There exist left inverses $H_1^+, H_2^+ \in \mathfrak{R}^{\bullet \times \mathtt{w}}$ of H_1 and H_2, respectively. This yields $R = H_1^+ H_2$ and $R^{-1} = H_2^+ H_1$, both in $\mathfrak{R}^{\bullet \times \bullet}$. □

References

1. Blomberg, H., Ylinen, R.: Algebraic Theory for Multivariable Linear Systems. Academic, London (1983)
2. Fuhrmann, P.: On strict system equivalence and similarity. International J. Control 25, 5–10 (1977)
3. Kuiper, M.: First-Order Representations of Linear Systems. Birkhäuser, Basel (1994)
4. Pernebo, L.: Notes on strict system equivalence. International J. Control 25, 21–38 (1977)
5. Polderman, J.W., Willems, J.C.: Introduction to Mathematical Systems Theory: A Behavioral Approach. Springer, Berlin (1997)
6. Rosenbrock, H.H.: State Space and Multivariable Theory. Wiley, New York (1970)
7. Rosenbrock, H.H.: The transformation of strict system equivalence. International J. Control 25, 11–19 (1977)
8. Schumacher, J.M.: Linear systems under external equivalence. Linear Algebra and its Applications 102, 1–33 (1988)
9. Willems, J.C.: Paradigms and puzzles in the theory of dynamical systems. IEEE Trans. Autom. Control 36, 259–294 (1991)
10. Willems, J.C.: Input-output and state-space representations of finite-dimensional linear time-invariant systems. Linear Algebra and its Applications 50, 581–608 (1983)
11. Willems, J.C., Trentelman, H.L.: On quadratic differential forms. SIAM Journal on Control and Optimization 36, 1702–1749 (1998)
12. Willems, J.C., Yamamoto, Y.: Behaviors defined by rational functions. Linear Algebra and its Applications 425, 226–241 (2007)
13. Wolovich, W.A.: Linear Multivariable Systems. Springer, New York (1974)

Modeling and Stability Analysis of Controlled Passive Walking

Kentaro Hirata

Abstract. In this article, modeling and stability analysis issues of controlled passive walking are discussed. On this memorable occasion of Professor Yamamoto's 60th birthday, it is shown that this research is deeply influenced by his pioneering works on various aspects of infinite-dimensional systems theory.

1 Introduction

As the state-of-art biped walking by robots are widely recognised, research interests for more advanced energy-efficient walking are growing [1, 4, 15, 24, 25]. One possible approach is to investigate a phenomenon called as Passive Dynamic Walking (PDW).

Although decades have passed since the pioneering work by McGeer [17], still the fact that such simple mechanical links without any control can walk down the slope like a human has been fascinating us. In terms of their experiments, the authors of [20] employed a novel non-invasive feedback control scheme to reduce the sensitivity to the initial values and to realize the original steady-state periodic motion "as is" at the same time. What they did is to apply a constant torque during the current step based on the difference of the past two successive step lengths. Based on the observation that this is a discrete-time version of so-called delayed feedback control [19], we started to investigate its continuous-time counterpart.

Note that its dynamical behaviour is highly complex from nonlinear, hybrid and infinite-dimensional nature of the phenomenon considered. However, by introducing a Piecewise Affine (PWA) approximation of the walker dynamics during the single support phase, the error dynamics against the original passive walking pattern is represented by a kind of convolution operator in which the effects of initial values and discontinuous state jumps are taken into consideration.

Kentaro Hirata
Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara, Japan
e-mail: kent@is.naist.jp

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 253–262.
springerlink.com

Since it is shown that the spectral radius of this operator determines the stability of the periodic orbit, the next natural question is how to compute it via numerical procedures. We justify the computation via he fast sample and fast hold approximation by using the basic facts from the perturbation theory.

2 Piecewise Affine Dynamical Model for Passive Walking

For example, consider a compass walker on a slope as depicted in Fig. 1. Let the point masses at the hip and the toes, the length of the leg, and the slope angle are denoted by M, m, ℓ, and γ, respectively. The angles between the stance leg and the normal vector of the slope and between two legs are denoted by θ and ϕ, respectively. Let τ_ϕ be the torque applied at the ankle. By assuming that the mass of each link and the friction at the hip joint are negligible, the dynamics during the single support phase is given by

$$M_\phi \begin{bmatrix} \ddot\theta \\ \ddot\phi \end{bmatrix} + \begin{bmatrix} ml^2 \sin\phi(2\dot\phi - \dot\theta)\dot\phi \\ -ml^2 \sin\phi\ \dot\theta^2 \end{bmatrix} + \begin{bmatrix} g_\theta \\ g_\phi \end{bmatrix} = \begin{bmatrix} 0 \\ \tau_\phi \end{bmatrix}, \tag{1}$$

$$M_\phi = \begin{bmatrix} Ml^2 + 2ml^2(1-\cos\phi) & -ml^2(1-\cos\phi) \\ -ml^2(1-\cos\phi) & ml^2 \end{bmatrix},$$

$$g_\phi = -mgl\sin(\theta - \phi - \gamma),$$

$$g_\theta = -Mgl\sin(\theta-\gamma) - \tau_\phi - \sin(\theta-\gamma)\}.$$

Following [3], we simplify this model through the normalization ($g = \ell = 1$) and an assumption on idealized mass ratio ($m/M \to 0$). Further, by assuming that the angles and the angular velocities involved are small, we linearize it. With the state vector

$$x = \begin{bmatrix} \theta & \dot\theta & \phi & \dot\phi \end{bmatrix}^{\mathrm{T}},$$

and the control input $u = \tau_\phi$, the dynamics of the single support phase (1) is expressed as

$$\dot x(t) = Ax(t) + Bu(t) + b, \quad x(0+) = x_0, \tag{2}$$

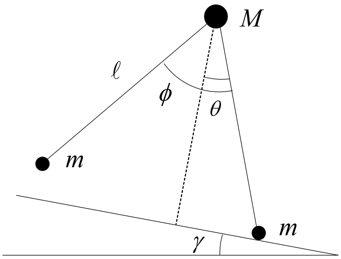


Fig. 1 Walker configuration

$$A=\begin{bmatrix}0\,1&0&0\\1\,0&0&0\\0\,0&0&1\\1\,0&-1&0\end{bmatrix},\quad B=\begin{bmatrix}0\\-1\\0\\-1\end{bmatrix},\quad b=\begin{bmatrix}0\\-\gamma\\0\\-\gamma\end{bmatrix}. \tag{3}$$

For the time being, let $u(t)\equiv 0$. The contact condition of the swing leg against the ground is given by

$$cx(t)=0, \tag{4}$$

with

$$c=\begin{bmatrix}-2\;0\;1\;0\end{bmatrix}. \tag{5}$$

Let τ_0 be the time when (4) is satisfied[1] and set $x_f=x(\tau_0)$, $\phi_f=\phi(\tau_0)$. The conservation of the angular momentum just before and after the collision yields the following instantaneous state transition rule[2]:

$$x(\tau_0+)=R(x_f)x(\tau_0), \tag{6}$$

where $R(\cdot)$ is a matrix valued function of $\alpha=[\alpha_1\;\alpha_2\;\alpha_3\;\alpha_4]^T$ given by

$$R(\alpha)=\begin{bmatrix}-1&0&0\,0\\0&\cos\alpha_3&0\,0\\-2&0&0\,0\\0&1-\cos\alpha_3&0\,0\end{bmatrix}. \tag{7}$$

Let $S(\alpha)=R(\alpha)\alpha$. Since (6) implies

$$R(x_f)x(\tau_0)=[R(\alpha)\alpha]_{\alpha=x_f}=[S(\alpha)]_{\alpha=x_f},$$

its Taylor expansion around x_f is given by

$$S(x_f+\Delta x_f)\simeq R(x_f)x_f+R(x_f)\Delta x_f+\left[\frac{\partial R(\alpha)x_f}{\partial\alpha}\right]_{\alpha=x_f}\Delta x_f. \tag{8}$$

Thus

$$S(x_f+\Delta x_f)=x(\tau_0+)+\bar{R}\Delta x_f, \tag{9}$$

$$\bar{R}=R(x_f)+\left[\frac{\partial R(\alpha)x_f}{\partial\alpha}\right]_{\alpha=x_f}, \tag{10}$$

[1] Implicitly, the case that the next collision never happen, such as the walker falls down, is excluded. Also as usual, ”the scuffing” is ignored [3, 4, 17].

[2] By this transition, the angular velocities become discontinuous. One should note that the angles are also discontinuous in this model since the stance leg and the swing leg exchange. We can avoid the latter discontinuity by taking the angles of left and right legs as the state variables. However, in that case, we must have two dynamical equations for the single support phase in contrast to the current situation with only (1).

can be used for the linearized perturbation analysis while (6) is required for the analysis of the equilibrium point and the dynamical simulations. From (7), (10) is given as

$$\bar{R} = \begin{bmatrix} -1 & 0 & 0\,0 \\ 0 & \beta & 0\,0 \\ -2 & 0 & 0\,0 \\ 0 & 1-\beta & 0\,0 \end{bmatrix}, \tag{11}$$

where $\beta = \cos\phi_f - \dot{\theta}_f \sin\phi_f$.

Thus we have a set of equations describing the motion of PDW as

$$\begin{aligned} &\dot{x}(t) = Ax(t) + b, \quad x(0+) = x_0, \\ &cx(\tau_0) = 0, \\ &x(\tau_0+) = R(x(\tau_0))x(\tau_0). \end{aligned} \tag{12}$$

The existence of a period-one orbit is equivalent to that of an initial value x_0^* results in

$$x(\tau_0+) = x_0^*. \tag{13}$$

Suppose that there is an initial condition x_0^* corresponds to a period-one orbit and $cv \neq 0$ for

$$v = Ax_f + b. \tag{14}$$

The latter one is the same as the transversality condition for more general case [4]. Now $x_f = x(\tau_0)$ satisfying $x_0 = R(x_f)x_f$ is an equilibrium point on the contact plane $cx = 0$. A perturbation Δx added to x_f at (just before) a collision results in a deviation $\Delta x_{+1}$from x_f at the next collision. This is the so-called Poincaré map. A linearized Poincaré map associated with PWA model of PDW is given by

$$P = \left(I - \frac{vc}{cv}\right) e^{A\tau_0} \bar{R}. \tag{15}$$

The structure is similar to the one obtained in [23]. See [12] for details.

For $\gamma = 0.009$, there exists a period-one orbit with $\tau_0 = 3.768$ and

$$x_f = \begin{bmatrix} -0.2 \ -0.218 \ -0.4 \ 0.002 \end{bmatrix}^T. \tag{16}$$

Corresponding linearized Poincaré map (15) is given by

$$P = \begin{bmatrix} -4.88 & 5.16 & -0.50 & 0 \\ -5.56 & 5.80 & -0.56 & 0 \\ -9.77 & 10.3 & -1.01 & 0 \\ -26.4 & 23.2 & -2.19 & 0 \end{bmatrix}.$$

Since P is contractive, this period-one orbit is locally stable.

Recently, the internal stabilizing mechanism of the passive dynamic walking is also investigated based on the model description (12) [11, 21, 22].

3 Delayed Feedback Control

Consider a nonlinear system

$$\begin{aligned} \dot{x}(t) &= f(x(t)) + u(t), \\ y(t) &= cx(t), \end{aligned} \tag{17}$$

and suppose that (17) with $u \equiv 0$ has an unstable τ-periodic orbit, say y^*. If y^* is known explicitly, the control law

$$u(t) = K\left[y(t) - y^*(t)\right], \tag{18}$$

might work for its stabilization. However, since y^* is an *unstable* periodic orbit, it is difficult to obtain it a-priori in general. Pyragas proposed the following alternative controller structure in [19] for this stabilization:

$$u(t) = K\left[y(t) - y(t-\tau)\right]. \tag{19}$$

From (19), it is obvious that $u \longrightarrow 0$ as $y \longrightarrow y^*$. Thus it is an ideal "non-invasive" way of stabilization. This characteristic is also suitable when we intend to enhance the stability of nominal periodic motions of the passive walker.

Note 1. The Laplace transform of the controller (19) contains $1 - e^{-\tau s}$, the reciprocal of the internal model $\frac{1}{1-e^{-\tau s}}$ which appears in the repetitive controller. Thus the role of poles and zeros on the unit circle are alternating in each control structure. This symmetric property may be interesting besides the fact that both are using the time delay in the control structure effectively for the compensation of periodicity. Through the joint works such as [5, 6, 7, 18, 29], Professor Yamamoto contributed to the development of the repetitive control theory.

Now consider the trajectory of the passive walker starting from x_0^*. Let $x^*(t)$, $t \in [0, \tau_0]$ be the nominal trajectory. We are interested in the effect of the perturbations against $x^*(\tau_0)$ and $x^*(\cdot)$ on the future trajectory. Let the time of the k-th state jump be t_k ($t_0 = 0$), $\tau_k := t_k - t_{k-1}$ and

$$x_1^k = x(t_k), \quad x_2^k(\xi) = x(t_{k-1} + \xi), \quad \xi \in [0, \tau_k]. \tag{20}$$

Define the deviation from the nominal trajectory by

$$\Delta x_1^k = x_1^k - x^*(\tau_0),\ \Delta x_2^k(\xi) = x_2^k(\xi) - x^*(\xi),\ \zeta \in [0, \tau_k].$$

According to the division of $x(\cdot)$, the control input $u(\cdot)$ is also divided as

$$u^k(\xi) := u(t_{k-1} + \xi), \quad \xi \in [0, \tau_k]. \tag{21}$$

If we apply the following DFC-like feedback control[3]

$$u^k(\xi) = K[x^k(\xi) - x^{k-1}(\xi)],\ \xi \in [0, \tau_k], \tag{22}$$

then the corresponding local error dynamics is governed by

$$\begin{bmatrix} \Delta x_1^{k+1} \\ \Delta x_2^{k+1}(\cdot) \end{bmatrix} = \mathscr{F} \begin{bmatrix} \Delta x_1^{k} \\ \Delta x_2^{k}(\cdot) \end{bmatrix},\ k \geq 0, \tag{23}$$

where the operator $\mathscr{F}$ on $\mathscr{Z} = \mathbf{R}^n \oplus \{\mathbf{L}^2[0, \tau_0]\}^n$ is defined as

$$\mathscr{F} z = \begin{bmatrix} Qe^{\bar{A}\tau_0}\left\{\bar{R}z_1 - \int_0^{\tau_0} e^{-\bar{A}\xi}\bar{B}z_2(\xi)d\xi\right\} \\ e^{\bar{A}\cdot}\bar{R}z_1 - \int_0^{\cdot} e^{\bar{A}(\cdot-\xi)}\bar{B}z_2(\xi)d\xi \end{bmatrix}, \quad z = \begin{bmatrix} z_1 \\ z_2(\cdot) \end{bmatrix}, \tag{24}$$

with $Q = I - \frac{vc}{cv}$, $\bar{A} = A + BK$ and $\bar{B} = BK$. (See [13] for details.)

Note 2. The operation in (20) which converts $x(\cdot)$ into $\{x_2^k(\cdot)\}$ is nothing but a lifting. As is well understood, now lifting is a standard and familiar mathematical tool commonly used in our field for the analysis of periodically time-varying systems. Undoubtedly it is a major influence of the seminal work [28] by Professor Yamamoto, a function space approach to sampled-data control systems.

If we remove the effect of the state jump from (24), i.e., set Q and $\bar{R}$ to be identity matrices, then the operator $\mathscr{F}$ represents the state transition of retarded delay-differential equation (DDE)

$$\dot{x}(t) = F_0 x(t) - F_1 x(t - r), \tag{25}$$

where $F_0 = \bar{A}$, $F_1 = \bar{B}$, and $r = \tau_0$. By introducing the following artificial input and output as

$$\dot{x}(t) = F_0 x(t) - F_1 x(t - r) + u(t),\ y(t) = x(t), \tag{26}$$

one can consider an input/output map related to DDE (25). Since this input/output map is contained in the class of pseudo-rational, this realization problem is within the scope of [26, 27, 30]. While it gives the continuous-time abstract differential equation on M_2 space (see e.g., [2]), the state transition operator (24) is its discrete-time counterpart.

4 Stability Analysis and Numerical Computation

Let $\sigma(T)$ and $r_\sigma(T)$ denote the spectrum and the spectral radius of the operator T, respectively. It can be shown that the zero solution of

$$z_{k+1} = \mathscr{F} z_k$$

or, alternatively, the nominal periodic orbit $x^*(\cdot)$ under DFC is stable if and only if $r_\sigma(\mathscr{F}) < 1$ [10]. Thus we need to compute $r_\sigma(\mathscr{F})$. Let us partition $\mathscr{F}$ compatibly with z as

[3] It is assumed that we extrapolate $x^{k-1}(\cdot)$ when $\tau_{k-1} < \tau_k$.

$$\mathscr{F}z = \begin{bmatrix} \mathscr{F}_{11} & \mathscr{F}_{12} \\ \mathscr{F}_{21} & \mathscr{F}_{22} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2(\cdot) \end{bmatrix}. \tag{27}$$

We see that $\begin{bmatrix} \mathscr{F}_{11} & \mathscr{F}_{12} \end{bmatrix}$ is compact because its range is a finite-dimensional space. Also $\begin{bmatrix} \mathscr{F}_{21} & \mathscr{F}_{22} \end{bmatrix}$ is, since it is a finite-rank perturbation of a Volterra operator on $\mathbf{L}^2[0,\tau_0]$. Consequently, $\mathscr{F}$ becomes compact and every nonzero $\lambda \in \sigma(\mathscr{F})$ is an eigenvalue of $\mathscr{F}$. As shown in [13], the eigenvalues of $\mathscr{F}$ correspond to the roots of transcendental equation

$$f(\lambda) := \left| \bar{R}Qe^{\{A-(1-1/\lambda)BK\}\tau_0} - \lambda I \right| = 0. \tag{28}$$

Then numerical methods for root finding based on the local search (such as Newton's method) can be used to find an eigenvalue, but not the whole set of eigenvalues. This is why we attempt to discretize $\mathscr{F}$ via a certain finite-dimensional approximation to obtain estimated distribution of eigenvalues.

Given $N \in \mathbf{N}$, let $\mathscr{S}_N$ and $\mathscr{H}_N$ denote the sampling and the ZOH operators with the sampling period τ_0/N, respectively. Then, a rectangular approximation of $z_2(\cdot)$ part in the output of $\mathscr{F}$ yields

$$\tilde{\mathscr{F}}_N = \tilde{\mathscr{H}}_N \tilde{\mathscr{S}}_N \mathscr{F} \tag{29}$$

where

$$\tilde{\mathscr{H}}_N = \begin{bmatrix} \mathscr{I} & 0 \\ 0 & \mathscr{H}_N \end{bmatrix}, \ \tilde{\mathscr{S}}_N = \begin{bmatrix} \mathscr{I} & 0 \\ 0 & \mathscr{S}_N \end{bmatrix}.$$

This is the fast sampling and fast hold (FSFH) approximation of $\mathscr{F}$. One can give a mathematical justification for the spectral computation based on this discretization scheme similarly to [31], where the numerical computation of the frequency response for sampled-data systems is considered.

Let Δ_N denote the difference between $\mathscr{F}$ and $\tilde{\mathscr{F}}_N$, i.e.,

$$\tilde{\mathscr{F}}_N = \mathscr{F} + \Delta_N,$$

and $D_\delta = \{s \in \mathbf{C}; |s| > \delta\}$. For $\delta > 0$, the spectrum of $\mathscr{F}$ outside of D_δ constitute a finite system of eigenvalues in the sense of [16] and is continuous against the perturbation Δ_N if it is generalized convergent. Consider the family of functions $\Phi_z := \{\phi_z |\, z \in U\}$ with

$$\phi_z = \begin{bmatrix} \mathscr{F}_{21} & \mathscr{F}_{22} \end{bmatrix} z,$$
$$U = \{z |\, z \in \mathscr{Z}, \|z\|_{\mathscr{Z}} = 1\}, \ \|z\|_{\mathscr{Z}} = \|z_1\|_{\mathbf{R}^n} + \|z_2(\cdot)\|_{\{\mathbf{L}^2[0,h]\}^n}.$$

As in [31], uniform equicontinuity of Φ_z is essential to prove that $\|\Delta_N\| \to 0$ as $N \to \infty$ [10]. By exchanging the order of the operators in (29), one can derive a matrix representation for $\mathscr{F}_N = \tilde{\mathscr{S}}_N \mathscr{F} \tilde{\mathscr{H}}_N$ as

$$\mathscr{F}_N = \begin{bmatrix} QW^N R & QW^{N-1}V & \cdots & \cdots & QV \\ \bar{R} & 0 & \cdots & \cdots & 0 \\ W\bar{R} & V & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ W^{N-1}\bar{R} & W^{N-2}V & \cdots & V & 0 \end{bmatrix}, \tag{30}$$

where

$$W = e^{\bar{A}\tau_0}, \quad V = -\int_0^{\tau_0} e^{\bar{A}(\tau-\xi)}\bar{B}d\xi. \tag{31}$$

In [14], the numerical procedure described above is applied to the stability analysis of experiments with real robot [20].

Example 1. As mentioned earlier, $\mathscr{F}$ with $Q = \bar{R} = I$ represents the state evolution of DDE of type (25). Let us consider the stability of

$$\dot{x}(t) = -1/2x(t) - x(t-h). \tag{32}$$

It corresponds the closed-loop stability of the block diagram in Fig. 2 with $G(s) = 2/(2s+1)$. Since the phase margin of $G(j\omega)$ is $\pi/3$ at $\omega = \sqrt{3}/2$, the closed-loop system is stable for $h \in [0, 4\sqrt{3}\pi/9]$. We verify the numerical accuracy of the computation of $\sigma(\mathscr{F})$ via the eigenvalues of $\mathscr{F}_N$ with $\bar{A} = -1/2, \bar{B} = -1$, $\tau_0 = 4\sqrt{3}\pi/9$. Table 1 shows the first 6 eigenvalues of $\mathscr{F}_N$ for each N. Note that the eigenvalues are sorted by the modulus in descending order. As guaranteed mathematically, each eigenvalue approaches to a point in $\sigma(\mathscr{F})$ as $N \to \infty$.

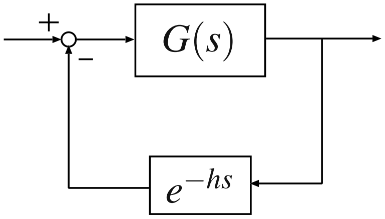


Fig. 2 Retarded delay-differential system

Table 1 Eigenvalues of $\mathscr{F}_N$

	z_1, z_2	z_3, z_4	z_5, z_6	$z_7, z_8, \cdots$
$N = 100$	$-0.494 \pm 0.872j$	$0.011 \pm 0.311j$	$0.019 \pm 0.173j$	$\cdots$
$N = 200$	$-0.497 \pm 0.869j$	$0.005 \pm 0.309j$	$0.013 \pm 0.172j$	$\cdots$
$N = 400$	$-0.499 \pm 0.868j$	$0.002 \pm 0.309j$	$0.010 \pm 0.172j$	$\cdots$
$N = 800$	$-0.499 \pm 0.867j$	$0.000 \pm 0.308j$	$0.008 \pm 0.172j$	$\cdots$
$N = 1600$	$-0.500 \pm 0.866j$	$-0.001 \pm 0.308j$	$0.008 \pm 0.171j$	$\cdots$
true value	$-1/2 \pm \sqrt{3}/2j$	$-0.001 \pm 0.308j$	$0.007 \pm 0.171j$	$\cdots$

However, in contrast to the frequency response, i.e., the singular value case, the convergence of the spectrum computation in terms of the number of the division N is relatively slow. This fact motivates the following subsequent works; application of modified FSFH approximation [8], relaxation of causality of the hold operator [9].

5 Conclusions

This article illustrates modeling and stability analysis issues of controlled passive walking via delayed feedback. The author received his Ph.D. degree under the supervision of Professor Yutaka Yamamoto twelve years ago. I would like to express my sincere gratitude for his passionate education and life-long influence. I hope I can learn even a little about the scholarship from my Master Y, as a padawan learns a lot about Jedi from Master Yoda in *Star Wars*.

References

1. Collins, S., Ruina, A., Tedrake, R., Wisse, M.: Efficient bipedal robots based on passive-dynamic Walkers. Science 307, 1082–1085 (2005)
2. Curtain, R.F., Zwart, H.J.: An Introduction to Infinite-Dimensional Linear Systems Theory. Springer, Heidelberg (1995)
3. Garcia, M., Chatterjee, A., Ruina, A., Coleman, M.: The simplest walking model: stability, complexity, and scaling. ASME Journal of Biomech. Eng. 120(2), 281–288 (1998)
4. Grizzle, J., Abba, G., Plenstan, F.: Asymptotically stable walking for biped robots: analysis via systems with impulse effects. IEEE Trans. Automat. Control 46, 51–64 (2001)
5. Hara, S., Yamamoto, Y.: Stability of repetitive control systems. In: Proc. 24th IEEE CDC, pp. 326–327 (1985)
6. Hara, S., Yamamoto, Y.: Stability of multivariable repetitive control systems — stability condition and characterization of stabilizing controllers. Trans. SICE 22, 1256–1261 (1986) (in Japanese)
7. Hara, S., Yamamoto, Y., Omata, T., Nakano, M.: Repetitive control system — a new-type servo system. IEEE Trans. Autom. Control 33, 659–668 (1988)
8. Hagiwara, T., Hirata, K.: Fast-lifting approach to the computation of the spectrum of retarded time-delay systems. In: Proc. ECC 2009 (2009)
9. Hirata, K., Itokazu, A., Hagiwara, T.: On numerical computation of the spectrum of a class of integral operators via non-causal hold discretization. In: Proc. of IFAC-TDS 2009 (2009)
10. Hirata, K.: On numerical computation of the spectrum of a class of convolution operators related to delay systems. Trans. ISCIE 21(3), 82–88 (2008) (in Japanese)
11. Hirata, K.: On Internal stabilizing mechanism of passive dynamic walking. In: Proc. SICE SI Division Annual Conf. (2008)
12. Hirata, K., Kokame, H.: Stability analysis of linear systems with state jump and effect of feedback control. Trans. ISCIE 17(12), 553–560 (2004) (in Japanese)
13. Hirata, K., Kokame, H.: Delayed feedback control of linear systems with state jump. Trans. ISCIE 18(3), 118–125 (2005) (in Japanese)
14. Hirata, K., et al.: Stability theory of delay systems revisited — from classics to DFC of passive walker— Part I, II. Journal of SICE 45 (2006) (in Japanese)

15. Hiskens, I.A.: Stability of hybrid system limit cycles: application to the compass gait biped robot. In: Proc. 40th IEEE CDC, pp. 774–779 (2001)
16. Kato, T.: Perturbation Theory for Lienar Operators. Springer, Heidelberg (1980)
17. McGeer, T.: Passive dynamic walking. Int. J. Robotics Research 9(2), 62–82 (1990)
18. Nakano, M., Inoue, T., Yamamoto, Y., Hara, S.: Repetitive Control. SICE (1989) (in Japanese)
19. Pyragas, K.: Continuous control of chaos by self-conbtrolling feedback. Physical Letters A 170, 421–428 (1992)
20. Sugimoto, Y., Osuka, K.: Walking control of quasi-passive-dynamic walking robot quartet III based on delayed feedback control. In: Proc. of the 5th Int. Conf. on CLWAR, pp. 123–130 (2002)
21. Sugimoto, Y., Osuka, K.: Stability analysis of passive dynamic walking — an approach via interpretation of Poincare map's structure. Trans. ISCIE 18(7), 19–24 (2005) (in Japanese)
22. Sugimoto, Y., Osuka, K.: Hierarchical implicite feedback structure in passive dynamic walking. Journal of Robotics and Mechatronics 20(4), 1–8 (2008)
23. Varigonda, S., Georgiou, T.T.: Dyanamics of relay relaxation oscillators. IEEE Trans. Automat. Control 46(1), 65–77 (2001)
24. Westervelt, E., Grizzle, J., Koditschek, D.: Hybrid zero dynamics of planar biped walkers. IEEE Trans. Automat. Control 48, 42–56 (2003)
25. Westervelt, E., Grizzle, J., Chevallereau, C., Choi, J., Morris, B.: Feedback Control of Dynamic Bipedal Robot Locomotion. CRC Press, Boca Raton (2007)
26. Yamamoto, Y.: Realization of pseudo-rational input/output maps and its spectral properties. Mem. Fac. Eng. Kyoto Univ. 47, 221–239 (1985)
27. Yamamoto, Y.: Pseudo-rational input/output maps and their realizations: a fractional representation approach to infinite-dimensional systems. SIAM J. Control and Optimization 26, 1415–1430 (1988)
28. Yamamoto, Y.: A function space approach to sampled-data control systems and tracking problems. IEEE Trans. Automat. Control 39, 703–712 (1994)
29. Yamamoto, Y., Hara, S.: The internal model principle and stabilizability of repetitive control systems. Trans. SICE 22, 830–834 (1986) (in Japanese)
30. Yamamoto, Y., Ueshima, S.: A new model for neutral delay-differential systems. Int. J. Control 43, 465–472 (1986)
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)

An Optimization Approach to Weak Approximation of Lévy-Driven Stochastic Differential Equations

Kenji Kashima and Reiichiro Kawai

Abstract. We propose an optimization approach to weak approximation of Lévy-driven stochastic differential equations. We employ a mathematical programming framework to obtain numerically upper and lower bound estimates of the target expectation, where the optimization procedure ends up with a polynomial programming problem. An advantage of our approach is that all we need is a closed form of the Lévy measure, not the exact simulation knowledge of the increments or of a shot noise representation for the time discretization approximation. We also investigate methods for approximation at some different intermediate time points simultaneously.

Preface

I[1] have received my Bachelor, Master and Ph.D. degrees from Kyoto University under the supervision of Yamamoto-Sensei. As far as I know, my Bachelor thesis was the first trial to apply "H^∞ signal reconstruction via sampled-data control theory" [6, 17] to *actual* audio data processing. I was extremely impressed to see the transdisiplinary aspect of system control theory. This invaluable experience is a source of my desire to contribute to other fields outside of the control community. Fortunately, I became acquainted with the second author whose speciality is probability theory and mathematical finance. In this article, we present our ongoing joint work that aspires to be as successful in the mathematical finance community as the "YY-filter" [10] is in the signal processing world.

Kenji Kashima
Tokyo Institute of Technology, 2-12-1, Oh-okayama, Meguro-ku, Tokyo 152-8552, Japan
e-mail: kashima@mei.titech.ac.jp

Reiichiro Kawai
University of Leicester, Leicester LE1 7RH, UK
e-mail: reiichiro.kawai@gmail.com

[1] In this section, "I" refers to the first author, Kenji Kashima.

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 263–272.
springerlink.com

1 Introduction

Stochastic differential equations have long been used to build realistic models in economics, finance, biology, the social sciences, chemistry, physics and other fields. In most active fields of application, dynamics with possible sudden shift have become more and more important. To model such shifts, one would like to employ stochastic differential equations where the underlying randomness contains jumps. For this purpose, the diffusion process is not sufficient since its sample paths are almost surely continuous. On the other hand, Lévy-driven stochastic differential equations, which contain diffusion as a special case, can formulate stochastic behavior with jumps. Regardless of its practical importance, however, the theory and the computational techniques of the Lévy processes have not been developed thoroughly as in the diffusion case. As nice references on the subject, we refer to Applebaum [1].

From a practical point of view, the sample paths approximation of stochastic differential equations has been a central issue for the purpose of numerical evaluation and simulation on the computer. There are two notions of the approximation; strong and weak approximations. The strong approximation schemes provide pathwise approximations which can be employed in scenario analysis, filtering or hedge simulation. For applications such as derivative pricing, the computation of moments or expected utilities, the so-called weak approximations are sufficient, that is, we need to estimate the expected value of a function. Other applications of the weak approximation include the computation of functional integrals, invariant measures, and Lyapunov exponents.

The theoretical properties of time discretization schemes are mostly studied for the diffusion case. See [7] for detailed investigation. In fact, the weak approximation of the Lévy-driven stochastic differential equations via Monte Carlo type methods is still very difficult. Moreover, the other existing methods are applicable only to some of the simplest Lévy processes. The main purpose of this paper is to propose a new approach to weak approximation of Lévy-driven stochastic differential equations. Unlike Monte Carlo simulation with the time discretization approximation of sample paths, we employ a mathematical programming framework to obtain numerically upper and lower bounds of the target expectation.

To this end, we follow the methodologies investigated in various fields of application by several authors, for example, Bertsimas, Popescu and Sethuraman [2], Helmes, Röhl and Stockbridge [4], Lasserre, Prieto-Rumeau and Zervos [9], to mention just a few[2]. Note that these results deal only with the pure diffusion case (i.e., without jump component) for which standard Monte Carlo methods are sufficient. In this sense, it should be emphasized that our result is not a trivial extension. The main drawback is the complexity of the Ito formula for Lévy-driven stochastic differential equations. As such, we need to carefully examine whether or not the resulting

[2] It is known that there exist two dual formulation of this framework, both of which arrive at a semi-definite programming in the end. One is the so-called generalized moment problem that makes use of the semi-definiteness of (localizing) moment matrices. The other is a polynomial optimization approach for which sum-of-squares relaxation efficiently works. In this paper, our discussion is based on the latter formulation.

optimization problems are practically solvable. Fortunately, as we show in the following sections, our approach covers various practically important Lévy-driven stochastic differential equations.

The rest of this paper is organized as follows. Section 2 gives mathematical definition of Lévy-driven stochastic differential equations. Section 3 introduces and studies our optimization approach to the weak approximation. Section 4 provides a numerical example to illustrate that our method is able to efficiently capture the marginal distributions of Lévy-driven stochastic differential equations. Finally, Section 5 concludes this paper.

2 Problem Formulation

Let us begin this section with general notations which will be used throughout the text. For $k \in \mathbb{N}$, ∂_k indicates the partial derivative with respect to k-th argument. We denote by C^{k_1,k_2} the class of continuous functions with continuous differentiability of k_1-time for the first argument and of k_2-time for the second argument.

Let X_0 be given in $\mathbb{R}$ and let $T > 0$. Consider a one-dimensional stochastic differential equation

$$dX_t = a_0(t,X_t)\,dt + a_1(t,X_t)\,dW_t + \int_{\mathbb{R}_0} b(t,X_{t-},z)\,(\mu-\nu)(dz,dt), \quad t \in [0,T],$$

where $\{W_t : t \geq 0\}$ is a standard Brownian motion and where μ is a Poisson random measure on $\mathbb{R}_0$ whose compensator is given by the Lévy measure ν satisfying $\int_{|z|>1} |z|\nu(dz) < +\infty$ and $\int_{\mathbb{R}_0}(|z|^2 \wedge 1)\nu(dz) < +\infty$. In order for the solution of (1) to be well defined, we impose the usual Lipschitz conditions and linear growth conditions on a_0, a_1 and b. We henceforth equip our underlying probability space with the natural filtration $(\mathscr{F}_t)_{t\in[0,T]}$ generated by $\{X_t : t \in [0,T]\}$. Moreover, throughout this study, we assume that $b(t,x,z) \neq 0$ and $\nu \neq 0$ to avoid triviality.

Our interest throughout this study is in approximating the expectation

$$\mathbb{E}\left[V(X_T)\right], \tag{1}$$

with the given V satisfying $\mathbb{E}[|V(X_T)|] < +\infty$. Note that the function V may have discontinuities. For the computation of $\mathbb{E}[V(X_T)]$, standard techniques include the Monte Carlo simulation of sample paths through the time discretization of stochastic differential equations, or even some exact knowledge of sample paths such as series representation of the Poisson jump component.

3 Optimization Approach to Weak Approximation

3.1 Ito Formula and Supermartingale

We are now in a position to introduce our optimization approach to the weak approximation. Let $\mathscr{X}(\subseteq \mathbb{R})$ be a support of $\{X_t : t \in [0,T]\}$ defined in (1). For $f \in C^{1,2}([0,T] \times \mathscr{X}; \mathbb{R})$, the Ito formula yields

$$\begin{aligned} df(t,X_t) =& \mathscr{A} f(t,X_t)dt + \partial_2 f(t,X_t)a_1(X_t)dW_t \\ &+ \int_{\mathbb{R}_0} B_z f\left(t,X_{t-}\right)(\mu-\nu)(dz,dt), \quad a.s., \end{aligned}$$

where

$$\begin{aligned} \mathscr{A} f(t,x) :=& \partial_1 f(t,x) + \partial_2 f(t,x)a_0(x) + \frac{1}{2}\partial_2^2 f(t,x)a_1(x)^2 \\ &+ \int_{\mathbb{R}_0} \left(B_z f(t,x) - \partial_2 f(t,x)b(x,z)\right)\nu(dz). \end{aligned}$$

and for $z \in \mathbb{R}_0$,

$$B_z f(t,x) := f\left(t, x + b\left(x,z\right)\right) - f\left(t,x\right).$$

Here, if

$$\mathbb{E}\left[\int_0^T \left(\partial_2 f(t,X_t)a_1(t,X_t)\right)^2 dt\right] < +\infty,$$

and if

$$\mathbb{E}\left[\int_0^T \int_{\mathbb{R}_0} \left(B_z f(t,X_t)a_1(t,X_t)\right)^2 \nu(dz)dt\right] < +\infty, \tag{2}$$

then the stochastic process

$$\left\{ f(t,X_t) - f(0,X_0) - \int_0^t \mathscr{A} f(s,X_s)ds : t \in [0,T] \right\}$$

is a square-integrable martingale with respect to the filtration. We can then derive one of important building blocks of our approach, the so-called Dynkin formula:

$$\mathbb{E}\left[f(T,X_T)\right] - f(0,X_0) = \mathbb{E}\left[\int_0^T \mathscr{A} f(s,X_s)ds\right]. \tag{3}$$

Hence, as soon as one finds an $f \in C^{1,2}([0,T] \times \mathscr{X}; \mathbb{R})$ such that

$$\begin{cases} \mathscr{A} f(t,x) \leq 0, \ (t,x) \in [0,T] \times \mathscr{X}, \\ \quad f(t,x) \geq V(x), \ x \in \mathscr{X}, \end{cases} \tag{4}$$

it follows

$$\mathbb{E}\left[V(X_T)\right] \leq \mathbb{E}\left[f(T,X_T)\right] \leq f(0,X_0). \tag{5}$$

Clearly, $f(0,X_0)$ serves as an upper bound of $\mathbb{E}[V(X_T)]$. To minimize the upper bound $f(0,X_0)$, we now turn to the optimization problem

$$\left| \begin{array}{ll} \min & f(0,X_0) \\ \text{s.t.} & f(t,x) \geq V(x), \ x \in \mathscr{X}, \\ & \mathscr{A} f(t,x) \leq 0, \ (t,x) \in [0,T] \times \mathscr{X}, \\ & f \in C^{1,2}([0,T] \times \mathscr{X}; \mathbb{R}). \end{array} \right.$$

3.2 Main Result

This optimization problem is very difficult to deal with since the class definitions of the functions f and V are too broad. To ease the above optimization problem, we restrict the class of the function f to be a polynomial both in t and x, that is, in the form

$$f(t,x) = \sum_{\{0\le k_1\le K_1, 0\le k_2\le K_2\}} c_{k_1,k_2} t^{k_1} x^{k_2}, \tag{6}$$

for some natural numbers K_1 and K_2 and for a sequence $\{c_{k_1,k_2}\}_{k_1\le K_1,k_2\le K_2}$ of constants. For convenience in notation, we henceforth denote by C_p the class of polynomial functions in the form (6). We also need to set V to be a *piecewise* polynomial. Moreover, we assume that both a_0 and a_1 are polynomials. We are then instead to solve the following optimization problem

$$\left|\begin{array}{ll} \min & f(0,X_0) \\ \text{s.t.} & f(t,x) \ge V(x),\ x\in\mathscr{X}, \\ & \mathscr{A}f(t,x)\le 0,\ (t,x)\in[0,T]\times\mathscr{X}, \\ & f\in C_p. \end{array}\right. \tag{7}$$

For the purpose of comparison, suppose that there is no jump in (1), that is, $b\equiv 0$ as in [13]. This assumption clearly makes $\mathscr{A}f$ a polynomial, and consequently (7) is a polynomial optimization problem. This is the main reason that the pure diffusion case is easier to deal with in this framework. In general, polynomial optimization problems are still NP hard. However, if the degrees of f are fixed, sums of squares relaxation enables us to solve the problem efficiently. For details, we refer to Parrilo [11]. On the other hand, this technique is not directly applicable to the model with general stochastic jumps. This is because $\mathscr{A}f$ is not necessarily a polynomial due to the additional integral term.

To circumvent this difficulty, we decompose the function b as follows:

Assumption 1. *Functions a_0 and a_1 are polynomials, and b is decomposed as*

$$b(t,x,z) = b_1(t,x)b_2(z),$$

where b_1 is a polynomial and where $b_2 : \mathbb{R}_0 \mapsto \mathbb{R}$ such that

$$\int_{\mathbb{R}_0} |b_2(z)|^k \nu(dz) < +\infty, \quad k = 2,\ldots,K_2. \qquad \square$$

Theorem 1. *Under Assumption 1, for any $f\in C_p$, $\mathscr{A}f$ is a polynomial in t and x. Moreover, the coefficients of $\mathscr{A}f$ are affine with respect to those of f.*

Proof. A simple algebra yields

$$\mathscr{A}f(t,x) = \partial_1 f(t,x) + \partial_2 f(t,x)a_0(t,x) + \frac{1}{2}\partial_2^2 f(t,x)a_1(t,x)^2$$
$$+ \sum_{\{0\le k_1\le K_1, 2\le k_2\le K_2\}} c_{k_1,k_2} t^{k_1} \sum_{k=0}^{k_2-2} {}_{k_2}C_k x^k b_1(t,x)^{k_2-k} M_{k_2-k}$$

where

$$M_l := \int_{\mathbb{R}_0} b_2(z)^l \nu(dz),\ l = 2,\dots,K_2.$$

This completes the proof. □

Clearly, the optimization (7) is now a polynomial programming problem. To be more precise, this problem is numerically tractable for any piecewise polynomial V. Finally, to obtain a lower bound for $\mathbb{E}[V(X_T)]$, we are to find a $g \in C_p$ via the polynomial programming

$$\left|\begin{array}{ll} \max & g(0,X_0) \\ \text{s.t.} & g(t,x) \le V(x),\ x \in \mathscr{X}, \\ & \mathscr{A} g(t,x) \ge 0,\ (t,x) \in [0,T] \times \mathscr{X}, \\ & g \in C_p. \end{array}\right. \tag{8}$$

Notice that our optimization approach does not require the sample paths simulation at all for the computation of the expectation $\mathbb{E}[V(X_T)]$. It is a great advantage of our approach that all we need is the Lévy measure in closed form, not the exact knowledge of the increments or of a shot noise representation for sample paths simulation for the weak approximation with the sample paths discretization.

3.3 Simultaneous Approximation for Homogeneous Process

In this section, we show that the optimal solution obtained through our approach provides some additional information, that are of direct practical use.

Firstly, note that the initial value X_0 does not appear in the constraints (4) in the previous section. Therefore, if f satisfies (4), $f(0,\tilde{x})$ automatically gives upper bounds for $\mathbb{E}_{\tilde{x}}[V(X_T)]$, where the notation $\mathbb{E}_x$ denotes the expectation taken under which the initial state of the stochastic differential equation (1) is given deterministically by $X_0 = x$.

The next theorem indicates that functions satisfying (4) can also serve as bounds at arbitrary intermediate time points.

Theorem 2. *Assume that (1) is time-homogeneous, i.e., a_1, a_2, and b are independent of t. Suppose that $f \in C^{1,2}$ satisfies (4). Then, for every $\bar{T} \in [0,T]$*

$$\mathbb{E}[V(X_{\bar{T}})] \le f(T - \bar{T}, X_0). \tag{9}$$

Proof. Define

$$f^\circ(t,x) := f(t + (T - \bar{T}), x).$$

Due to the time homogeneity, we have

$$\mathscr{A} f^\circ(t,x) = \mathscr{A} f(t + (T - \bar{T}), x) \le 0,\ (t,x) \in [0,\bar{T}] \times \mathscr{X}.$$

We also have

$$f^\circ(\bar{T},x) = f(T,x) \le V(x),\ \text{in } \mathscr{X}.$$

By combining these inequalities and Dynkin formula, we obtain

$$\begin{aligned}\mathbb{E}[V(X_{\bar{T}})] &\leq \mathbb{E}[f^{\circ}(\bar{T},X_{\bar{T}})]\\ &= f^{\circ}(0,X_0)+\mathbb{E}\left[\int_0^{\bar{T}} \mathscr{A}f^{\circ}(s,X_s)ds\right]\\ &\leq f(T-\bar{T},X_0).\end{aligned}$$

This completes the proof. □

We here make a brief comment on the choice of the cost function in the optimization problem. When we attempt to find as tight bounds for (1) as possible, we should solve (7) and (8). However, we need to approximate $\mathbb{E}[V(X_{\bar{T}})]$ for some different time points $\bar{T} \in [0,T]$ and also different initial value X_0, it is useful to suitably change the cost function. Fortunately, for suitable measure ϕ on $[0,T]\times\mathbb{R}$, we can similarly optimize

$$\int f(t,s)\phi(dt,ds), \int g(t,s)\phi(dt,ds),$$

since these are linear combination of the decision variables (the coefficients of f and g).

4 Numerical Examples

In this section we give some approximation examples. In the numerical examples presented hereafter, we utilize MATLAB SOSTOOLS combined with SeDuMi [12, 16], using a computer with a Pentium 4 3.2GHz processor and 2 GB memory.

4.1 Ornstein-Uhlenbeck-Type Process with Gamma Stationary Distribution

Let ν be a Lévy measure on $\mathbb{R}_+$ such that $\int_{\mathbb{R}_+} z\nu(dz) < +\infty$. Set $a_0(t,x) = -\lambda x + \int_{\mathbb{R}_+} z\nu(dz)$ for some $\lambda > 0$, $a_1(t,x) = 0$, $b_1(t,x) = 1$, $b_2(z) = z$, and X_0 is independent of μ. Then, the stochastic differential equation (1) reduces to

$$dX_t = -\lambda X_t dt + \int_{\mathbb{R}_+} z\mu(dz,dt),$$

which is called an Ornstein-Uhlenbeck-type process. (See, for example, Sato [15] for its details.) Its solution is given by

$$X_t = e^{-\lambda t}X_0 + \int_0^t \int_{\mathbb{R}_+} e^{-\lambda(t-s)} z\mu(dz,ds).$$

For simplicity, we further fix $X_0 = 0$, $\lambda = 1$ and $\nu(dz) = bae^{-bz}dz$, where $a > 0$ and $b > 0$. Then, we can prove that the stationary distribution of $\{X_t : t \geq 0\}$ is gamma with density $p(x) = b^a/\Gamma(a)x^{a-1}e^{-bx}$, $x \in \mathbb{R}_+$.

Here, we investigate the distribution transition via the moment estimations of $\mathbb{E}[X_t] = (1-e^{-t})a/b$, $\mathbb{E}[X_t^2] = (1-e^{-2t})a/b^2 + (1-e^{-t})^2a^2/b^2$, and $\lim_{t\uparrow+\infty}\mathbb{E}[X_t^k] = \Gamma(a+k)/(b^k\Gamma(a))$, for $k \in \mathbb{N}$. Note that $\mathscr{X} = \mathbb{R}_+$ and that $\int_{\mathbb{R}_+} z^k\nu(dz) = ak!/b^k$ for $k \in \mathbb{N}$. For $f \in C_p([0,T]\times\mathbb{R}_+;\mathbb{R})$, we have

$$\begin{aligned}\mathscr{A}f(t,x) = &\sum_{\{1\leq k_1\leq K_1, 0\leq k_2\leq K_2\}} c_{k_1,k_2}k_1t^{k_1-1}x^{k_2}\\ &+\left(-x+\frac{a}{b}\right)\sum_{\{0\leq k_1\leq K_1, 1\leq k_2\leq K_2\}} c_{k_1,k_2}t^{k_1}k_2x^{k_2-1}\\ &+\sum_{\{0\leq k_1\leq K_1, 2\leq k_2\leq K_2\}} c_{k_1,k_2}t^{k_1}\sum_{k=0}^{k_2-2} {}_{k_2}C_kx^k\frac{a(k_2-k)!}{b^{k_2-k}}.\end{aligned}$$

The condition (2) holds for each K_1 and K_2, since

$$\int_0^t\int_{\mathbb{R}_+} e^{-\lambda(t-s)}z\mu(dz,ds) \leq \int_0^T\int_{\mathbb{R}_+} z\mu(dz,ds), \quad a.s.,$$

where the right hand side is an infinitely divisible random variable, whose Lévy measure has an exponential decay at infinity.

We present numerical results in Table 1. We set $K_1 = p$ for the estimation of the p-th moment. It is known that the computational burden for solving the polynomial optimization via sum of squares decomposition significantly increases as the degree of the polynomial becomes larger. In view of this, we choose large $K_2 = 10$. Even in this case, however, computation time is at most 2 seconds. For comparison with Monte Carlo methods, we also provide 99%-confidence interval with 1000000 iid samples. As can be observed, even with the extraordinarily large number of samples,

Table 1 Moment transition with $X_0 = 0$ and $(a,b) = (0.1, 1.5)$. The numbers in parentheses indicate theoretical value. The intervals are 99%-confidence interval with 1000000 independent samples.

	$t = 1$	$t = 2$	$t = 3$	$t \uparrow +\infty$
$\mathbb{E}[X_t]$	0.042141 – 0.042141 (0.042141) [0.0417327, 0.0427467]	0.057644 – 0.057644 (0.057644) [0.0569853, 0.0580576]	0.063347 – 0.063348 (0.063348) [0.0628788, 0.0639666]	(0.06667)
$\mathbb{E}\left[X_t^2\right]$	0.040205 – 0.040205 (0.040205) [0.0396179, 0.0414835]	0.046952 – 0.046955 (0.046953) [0.0457206, 0.0476003]	0.048331 – 0.048347 (0.048347) [0.0476659, 0.0496044]	(0.04889)
$\mathbb{E}\left[X_t^3\right]$	0.061217 – 0.061268 (n/a) [0.0591606, 0.0649658]	0.066812 – 0.066886 (n/a) [0.0633623, 0.0688394]	0.068009 – 0.068051 (n/a) [0.0660050, 0.0719257]	(0.06844)

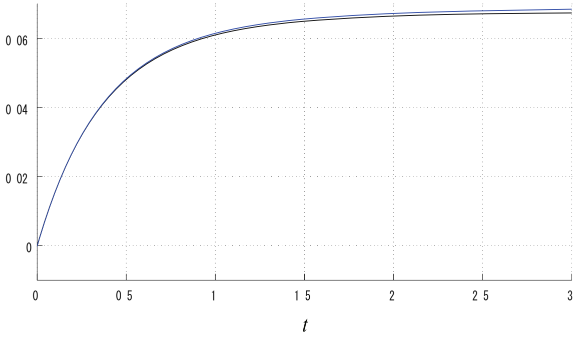


Fig. 1 Lower and upper bounds for $\mathbb{E}[X_t^3]$ at intermediate time points $t \in [0,3]$. The detailed values are 0.060992 – 0.061467 ($t = 1$) and 0.066461 – 0.067219 ($t = 2$).

the 99%-confidence intervals are far from being comparable with our results. Note that any large sample size can never be in competition with our results since the upper and lower bounds obtained through our method form nothing but the 100%-confidence interval.

Recall that the current model is time-homogeneous. Hence, according to Theorem 2, the obtained bounding functions also give upper and lower bounds for intermediate time points without solving other optimization problem. For example, as a byproduct of the computation of the bounds for $\mathbb{E}[X_3^3]$, we can provide a parametric bounds for $\mathbb{E}[X_t^3]$ for every $t \in [0,3]$; see Fig. 1. In this case, the accuracy is close to the pointwise optimization result in Table 1. Actually, the gap is smaller than the case of Monte Carlo methods in Table 1.

5 Conclusion

In this paper, we have developed a new approach to the weak approximation of Lévy-driven stochastic differential equations via an optimization problem yielding upper and lower bounds on the target expectation. The advantage of our approach is that all we need is the Lévy measure in closed form. We need neither the exact knowledge of the increments nor a shot noise representation for sample path simulation for the weak approximation with the sample path discretization. We have also investigated how we can obtain accurate approximation at transient times.

The most important remaining work is the improvement of the approximation accuracy. It is a good direction to pursue to use exponentially tempered polynomials [5]. Other currently ongoing work is application to calibration in finance.

References

1. Applebaum, D.: Lévy Processes and Stochastic Calculus. Cambridge University Press, Cambridge (2004)
2. Bertsimas, D., Popescu, I., Sethuraman, J.: Moment problems and semidefinite programming. In: Wolkovitz, H. (ed.) Handbook on Semidefinite Programming: Theory, Algorithms, and Applications, pp. 469–509 (2000)

3. Eriksson, B., Pistorius, M.: A method of moments approach to pricing double barrier contracts driven by a general class of jump diffusions (2008), arXiv:0812.4548v1
4. Helmes, K., Röhl, S., Stockbridge, R.H.: Computing moments of the exit time distribution for Markov processes by linear programming. Operations Research 49(4), 516–530 (2001)
5. Kashima, K., Kawai, R.: A weak approximation of stochastic differential equations with jumps through tempered polynomial programming (submitted, 2009)
6. Khargonekar, P.P., Yamamoto, Y.: Delayed signal reconstruction using sampled-data control. In: Proc. 35th Conf. Decision and Control, pp. 1259–1263 (1996)
7. Kloeden, P.E., Platen, E.: Numerical Solution of Stochastic Differential Equations, 3rd edn. Springer, Berlin (1999)
8. Lasserre, J.B., Prieto-Rumeau, T.: SDP vs. LP relaxations for the moment approach in some performance evaluation problems. Stochastic Models 20(4), 439–456 (2004)
9. Lasserre, J.B., Prieto-Rumeau, T., Zervos, M.: Pricing a class of exotic via moments and SDP relaxations. Mathematical Finance 16(3), 469–494 (2006)
10. Nagahara, M.: YY filter — a paradigm of digital signal processing. To appear in Perspectives in Mathematical System Theory. Springer, Heidelberg (2010)
11. Parrilo, P.A.: Semidefinite programming relaxations for semialgebraic problems. Mathematical Programming, Series B 96(2), 293–320 (2003)
12. Prajna, S., Papachristodoulou, A., Seiler, P., Parrilo, P.A.: SOS-TOOLS: Sum of Squares Optimization Toolbox for MATLAB (2004)
13. Primbs, J.A.: Optimization based option pricing bounds via piecewise polynomial super- and sub-martingales. In: Proc. 2008 American Control Conf., pp. 363–368 (2008)
14. Protter, P., Talay, D.: The Euler scheme for Lévy driven stochastic differential equations. Annals of Probability 25, 393–423 (1997)
15. Sato, K.: Lévy Processes and Infinitely Divisible Distributions. Cambridge Univ. Press, Cambridge (1999)
16. Sturm, J.: SeDuMi, version 1.1 (2006)
17. Yamamoto, Y.: A function space approach to sampled-data control systems and tracking problems. IEEE Trans. Atutom. Control 39, 703–712 (1994)

Compound Control: Capturing Multivariable Nature of Biological Control

Hidenori Kimura, Shingo Shimoda, and Reiko J. Tanaka

Abstract. Multivariable extension of design theory was a historic milestone that innovated the basic framework of control theory fundamentally. Analogous situation is now coming into the stage of biological control research. This paper addresses the present state-of-arts of this issue, introducing the multivariable nature of biological control, the formulation of the problem under specific assumptions and preliminary results obtained. The paper suggests the possible paradigm change in future biological control research.

1 Introduction

Emergence of state space theory in early 1960's was really a *revolution* in the history of control engineering. State space framework renovated the fundamental framework of control theory and elucidated the deep and rich structures of control systems. It not only enlarged our scope of control systems, but also gave new lights on the fundamental notions of control such as feedback, feedforward, servo, stability, identification and so on.

Kalman who initiated the state space theory said that his underlying motivation to propose state space theory was to establish the rigorous and scientific foundation of control theory, just like Shannon's communication theory [1]. Indeed, his objective has been achieved with great success in this respect. State-space method gave mathematically rigorous and practically versatile framework for control systems design. It should be noted that state-space framework also gave powerful tool for dealing

Hidenori Kimura and Shingo Shimoda
RIKEN BSI-TOYOTA Collaboration Center, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan
e-mail: {hkimura,shimoda}@brain.riken.jp

Reiko J. Tanaka
Department of Bioengineering and Institute for Mathematical Sciences, Imperial College London, SW7 2PG, UK
e-mail: r.tanaka@imperial.ac.uk

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 273–283.
springerlink.com

with multi-input multi-output (MIMO) plants. Towards the end of 1950's, pressing needs arose to control plants with multiple inputs that were coupled together to generate multiple outputs. At that time, the design methods were only available for single-input single-output (SISO) plants, and the only way to deal with such plants was to decouple the input-output relations and design the controller for each decoupled loop. This way of decoupling was not satisfactory in many reasons because it actually reduced the problem to finding inverse systems. The method generated usually excessively complicated controllers and even unstable ones.

In terms of traditional transfer function approach, the extension of design methods for SISO plants to MIMO plants is just the matter of extending the results on transfer functions to transfer function *matrices*. This extension, however, is far from trivial. For instance, while the notions of gain and phase have clear meaning in terms of transfer functions, their meanings are not clear at all in transfer function matrices. It was desired to create some new theoretical framework to deal with multivariable control. State-space method was just the appropriate one to fulfill this need. It can deal with MIMO plants very naturally and the relationship between state-space framework and classical transfer function approach has been fully exploited.

Thus, it is the common retrospect that the extension of SISO control to MIMO control was not at all trivial and required a great paradigm shift in control engineering which renovated the fundamental framework in almost all aspects of control theory. In some sense, the real design theory of control systems started when multivariable plants became a feasible target of theory. This paper states that an analogous paradigm shift from SISO to MIMO is now taking place in biological control.

Control mechanisms play crucial roles in all species, all functions and all levels of organs in the living organism. Therefore, control in living organisms has been an important research issue in biology. An exposition of the biology/control interplay is found in a review paper from the start of modern biology at the beginning of 20th century [2]. As in the case of design theory for engineering systems, the analysis of biological control systems has been based on the framework of SISO systems. Now, it faces with multivariable situations in many respects. The nature of multivariable features in biological control, however, is significantly different from multivariable control of engineering systems. In engineering control, multivariable nature was brought in by the MIMO plants, whereas in biological control, it was brought in by MIMO *controllers*. In biological control, the controller must be multivariable from the nature of environment that has enormous number of factors. This is one of intrinsic differences between engineering control and biological control. Also, the multivariable nature of controllers is significantly different from engineering systems. It is essentially combinatorial in biological control, while it is essentially interaction in engineering. Biological control in face of huge number of possible environmental changes must use this feature in its daily operation. Our long-term challenge to find some universal principles of biological control has produced the notion of *compound control* that utilizes the compound nature of environmental change [3]. The detail will be exposed in the text.

In Section 2, the intrinsic nature of multivariable aspects in biological control is explained. Section 3 gives examples of compound control.

2 Compound Control

In neuroscience, human motions have been divided into the two categories, voluntary and involuntary. The border distinguishing these categories is not clear because the notion of consciousness or awareness is unclear and ambiguous, and still very controversial. Sherrington, one of the founder of neurosciences, identified the involuntary movement with *reflex*, which is a simple body response to exogeneous disturbance. He thought that all the human motions were just combinations of simple basic reflexes which were built in human neural systems innately. Although his idea was abandoned after discoveries of sophisticated voluntary motions, we were very much influenced by his view of human body in establishing the notion of compound control. Now, in the recent progress of neuroscience, reflex itself is found to be a very complex process of integration of various controls with rich structure, and it is closely related to cognitive capability of brain though it is done unconsciously. Reflex is a body movement done unconsciously, but it is an intelligent process that encompasses body motion, power consumption and physical interactions to environment. Reflex is really a symbolic issue of the embodiment of intelligence and the compound control at an individual level addresses this issue from the control point of view.

If humans encounter some unknown environmental situations in his/her life, it is required to react to the situations by creating a new pattern of behaviors that may suit the situations. This ability of adaptation is actually the essence of intelligence without which humans cannot survive. In short, *the intelligence is an ability to overcome unknowns to adapt to new environment.*

If all the possible environmental situations we may come across can be listed up before actions or decisions, we can prepare appropriate behavioral response to each situation in advance. In such cases, unknowns are out of sight and no intelligence is needed. Of course it is not the case. The number of all possible environmental situations is so enormous that we cannot prepare for all the situations. The so-called *frame problem* is still a big issue both in artificial intelligence, and in robotics. How do we manage to create appropriate behavioral patterns in new environmental situations that we have not experienced in the past? The compound control takes a special stance on this issue.

It is true that we encounter unknown environmental situations in daily life, but the unknowns do not come randomly. They come along under some specific context of current behaviors and tasks. Therefore, majority of unknowns is related to knowns or some unknown combinations of knowns. Sometimes, we can predict what sort of unknown we will have in the next stage of behaviors. Our decision or selections on the next movements are somehow determined by the context we are aware of, or by the knowledge how the current unknown situations are composed of known situations. It is most important, and therefore most difficult, to explicitly formulate the above procedure of reducing unknowns to predictable combinations of knowns. We call this process *tacit learning* [4].

To sum up, our approach to biological control is based on the following assumptions:

1. Majority of unknowns is the combination of some basic knowns to which our responses have been well established.
2. The basic knowns are composed not only spacially but also temporally, to execute a new task.
3. To understand how the basic knowns are combined to create unknowns directly leads to generate appropriate behavioral responses. This process of understanding is actually termed to be *learning*.

The notion of compound control is based essentially on the above assumptions.

3 Examples of Compound Control

To justify our approach, we demonstrate three examples of compound control. The first one is concerned with the genetic control at the cell level. We show the most basic combinatorial logic of compound control against environmental changes. The second one shows that the temporal combination of basic environmental changes is the key issue in carrying out tasks of individuals. The result demonstrated that the compound control approach led to a remarkably smooth biped walking by robot. The third one considers the postural reflex as a typical example of compound control which we are now trying to formulate from compound control viewpoint.

3.1 Intracellular Genetic Control

Combinational characteristics is the most basic feature of compound control which are found in intracellular genetic control in a clear way. A classical example of combinational compound control is the *lac operon* which was extensively studied by Monod in early 1960's [5]. Lac operon is a collection of genes of bacteria that codes enzymes to metaborize lactose, which expresses when lactose is taken by the cell to be catabolized to supply energy. However, if glucose is simultaneously taken, then it does not express because glucose is more effective than lactose as the energy resource. This is an example of *catabolite repressions* in metabolic networks.

In this case, the environment of lac operon has two factors, namely, lactose and glucose concentrations. The response of the lac operon to the environmental change can be roughly represented by a logical function, if we use the binary representation of the concentrations of lactose and glucose by logic variables x and y, respectively. Assume that $x = true(or\ 1)$ implies "high" lactose concentration, while $x = false(or\ 0)$ implies "low" lactose consentration. The same assignment applies to y. The response r of lac operon is represented by

$$r = x \wedge \overline{y}, \tag{1}$$

where overbar denotes the logical negation. Logical representation of genetic switch systems like (1) is found to be useful recently. Our purpose here is to classify logical functions in order to clarify the intrinsic nature of logical expressions [6].

Consider a logical function of two variables

$$r = f(x,y).$$

Even if the environmental cue x is replaced by its negation $\overline{x}$, the intrinsic nature of the compound mechanism is not alfered, because it is the matter of replacing inhibition by potentiation. Therefore, the function $f(x,y)$ and $f(\overline{x},y)$ are considered to bear similar compound structure and belong in the same class. The same reasoning applies to the function $f(x,\overline{y})$. Also, the interchange between x and y do not affect the compound nature of interactions. So, we regard $f(x,y)$ and $f(y,x)$ in the same class.

The above rules yield a binary relations. For a given two logical functions f and g, we say that $f \sim g$ iff $g(x,y)$ is one of $f(x,y)$, $f(\overline{x},y)$, $f(x,\overline{y})$ and $f(y,x)$. It is trivial to see that this binary relation satisfies

$$\begin{aligned} & f \sim f && \text{(reflexivity)} \\ & f \sim g \quad \text{implies} \quad g \sim f && \text{(symmetry)} \\ & f \sim g,\ g \sim h \quad \text{implies} \quad f \sim h && \text{(transitivity)} \end{aligned}$$

Therefore, the classification based on the above binary relation yields equivalence classes. Any logical function belongs to one and only one class in this classification. Except trivial functions, there are 14 logical functions with two variables. It can be shown that they are divided into four classes under the above binary relation [6]. Representative functions of the four classes are given as follows

$$\begin{aligned} f_p(x,y) &= x \\ f_i(x,y) &= x \vee y \\ f_d(x,y) &= x \wedge \overline{y} \\ f_t(x,y) &= (x \wedge \overline{y}) \wedge (\overline{x} \wedge y) \end{aligned}$$

These four functions represent *parallel, induction, dismissal* and *tirgger* classes, respectively. Specifically, the term parallel implies that two environmental factors are not compound. They are independent each other. Other names came from the biological functions they perform in genetic regulations. Figure 1 indicates these classes in terms of the truth table. It is easy to see that identifying $f(x,y)$ with $f(\overline{x},y)$ means that the class is invariant with respect to the permutation of the columns of the truth table, while identification of $f(x,y)$ with $f(x,\overline{y})$ means the invariance with respect to permutation of rows. The identification of $f(x,y)$ with $f(y,x)$ is the invariance with respect to both column and row permutations.

To represent more rigorously the above reasoning, let σ_1 and σ_2 represent the permutations of columns and rows of the truth table of the logical function, respectively. We consider the commutative group Σ composed of four operations

$$\Sigma = \{e, \sigma_1, \sigma_2, \sigma_1\sigma_2\},$$

where e denotes the identity operation. Note that $\sigma_i^2 = e,\ i = 1,2$. Now, our equivalence class is induced by the above group, i.e., functions f_1 and f_2 are said to be in the same class if and only if $f_1 = \sigma f_2$ for some $\sigma \in \Sigma$.

The lac operon belongs to the class of dismissal which implies that the expression of the lactose enzyme must be *dismissed* by the existence of rich glucose. In [6], examples of genetic regulations corresponding to induction and trigger are demonstrated.

Extension of the above results to general n variable logical functions is a difficult problem. We have got the result for $n = 3$. There are 254 non-trivial logical functions with 3 variables. The number of equivalence classes in this case is 12 [6].

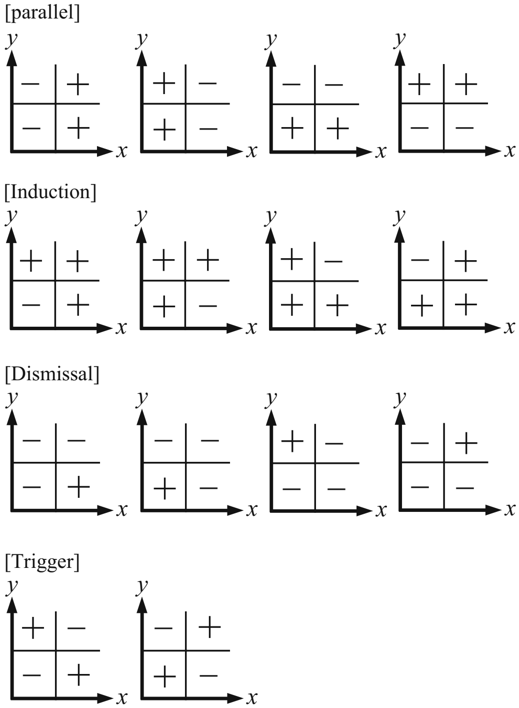


Fig. 1 Equivalence Classes of Logical Functions. + implies that output is true, – implies it is false

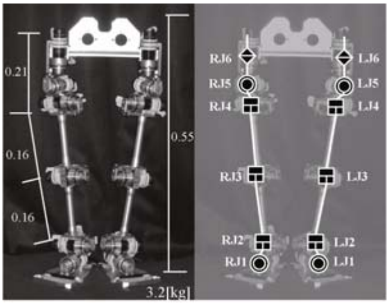
3.2 *Biped Locomotion*

Now we jump up from intracellular regulation to movements of individuals. Human being is the only animal that walks biped in daily life. To mimic biped locomotion in robots has attracted many researchers in the area of robotics, brain sciences and control. There have been considerable amount of work concerning biped robot using variety of approaches [7], [8]. Theoretical studies have also been done from many points of view (e.g. [9], [10]). It is of great interest specially from control theory for many reasons. First, it is a combination of postures (statics) and their transitions (dynamics). Second, it needs a delicate integration of locomotion and balance. Third, it embodies both innate ability and acquired skill.

We now regard the control issue of biped walking from a point of view of compound control. Biped locomotion is actually composed of several postures and transitions from one posture to another. Since we view biped walking as consecutive repetitions of shifts from one posture to another, it is of supremum importance to specify desired postures to be attained. In this case, the configurations of body muscle-skelton system is considered to be the environment. The desired postures for biped locomotion are picked up among enormous number of possible body configurations. The control is just to carry out the transitions from one posture to the next. The selected postures constitute fundamental set of environmental changes and their combination are carried out in an appropriate time frame. This is the view of biped locomotion based on compound control.

In order to justify our view, we have constructed a biped robot and designed a neural controller to control it. The robot we have constructed is shown in Fig. 2. It has 14 DOF with height 0.5m and weight approximately 3.5kg. Each joint is controlled by independent drivers. The locomotion is carried out by control commands to shift from the present posture to the next one. We picked up eight postures (four postures to each leg) described in Fig. 3, where ρ_i denotes the i-th joint.

Fig. 2 Overview of 12DOF Biped Robot



To specify each posture as a *snapshot*, we only use a few joints as is shown in Fig. 3 which were regulated by neural controllers [4]. No specific command were given to the rest of the joints. Instead, they were asked to *choose their own command signals* by themselves by a self-reference controller [11] which was prepared for our robot based on neural computations. We did not use any model for control, nor trajectory design for guaranteeing stability. We just implement output regulators and self-reference regulators to each joint and switched from output regulation to self-reference regulation and vise versa, depending upon what sort of posture transitions were required at each time. It is worth noting that at each trial, the states of the integrators were memorized and used in the next trial when the robot came to the same posture. This procedure seems to act as a sort of learning.

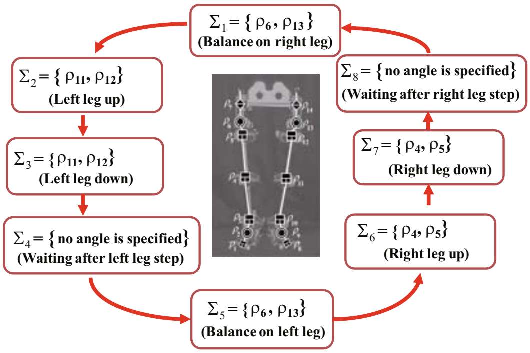

$\Sigma_1 = \{\rho_6, \rho_{13}\}$
(Balance on right leg)
$\Sigma_2 = \{\rho_{11}, \rho_{12}\}$
(Left leg up)
$\Sigma_3 = \{\rho_{11}, \rho_{12}\}$
(Left leg down)
$\Sigma_4 = \{$no angle is specified$\}$
(Waiting after left leg step)
$\Sigma_5 = \{\rho_6, \rho_{13}\}$
(Balance on left leg)
$\Sigma_6 = \{\rho_4, \rho_5\}$
(Right leg up)
$\Sigma_7 = \{\rho_4, \rho_5\}$
(Right leg down)
$\Sigma_8 = \{$no angle is specified$\}$
(Waiting after right leg step)

Fig. 3 ρ_i denotes the angle of the i-th joint. Constraint Conditions for Bipedal Walking: $\Sigma_i = \{$Constraint Conditions (specified angle in Task i)$\}$

We summarize our design principles:

1. Snapshot postures are assigned.
2. Each time when the robot configuration reaches a snapshot, the transition to the next snapshot is ordered through the controllers.
3. Those joints which are involved in the specification of the next snapshot are given desired commands, and all the rest of the joints are set free.
4. If the robot falled down, it was raised up manually by human being, but the controllers were still working during this *salvation*.
5. No model nor trajectory was given.

Fig. 4 shows a process of learning through the time profile of leg angle and the trajectory of center of mass. It should be noted that the robot chose its own pace and gait through learning that fit his/her body structure. The pace will change if the robot carries a tip load.

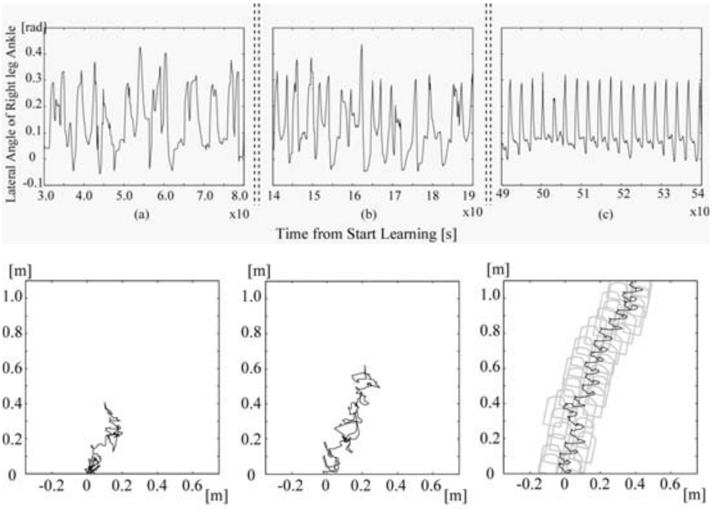
It is surprising to notice that our robot is also very efficient. Fig. 5 shows that after learning our robot walked very efficiently under the efficiency index

$$I = \frac{(Consumed\ Energy)}{(Mass)\cdot(Walking\ Distance)}$$

compared with existing humanoid robots. We do not know where this efficiency comes from.

3.3 Postural Reflex

The third example we would like to bring forward is postural reflex, which is a class of reflex necessary for maintaining body balance against gravity. If we are forced to walk through a narrow path, we unconsciously lift our hands horizontally to keep balance in order to increase the body inertia around the center of mass. This is a typical example of postural reflex which is built in our nervous system innately. Postural reflex is also closely related to biped walking we have discussed in the previous subsection. Its importance in our daily life is unquestionable. The recent



(a)30 80[s] (b)140 190[s] (c)490 540[s]

Fig. 4 Experimental Results

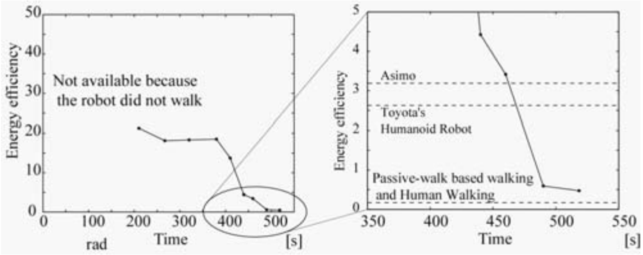


Fig. 5 Energy Efficiency

progress of neurosciences has disclosed that it is a process of complex integration of sensors, brain and musculo-skeleton systems.

We have been interested in neural mechanisms of balance maintenance (e.g.[12]) and now we are extending our scope to that of postural reflex from the viewpoint of compound control. According to our notion of compound control, postural reflex is assumed to be composed of several basic disturbance/response pairs and specific reflex is generated through a combination of these basic pairs. Disturbance in the reflex may not be purely exogeneous. It may be created via our voluntary movements which is endogenous and predictable. For instance, when we bend the upper body forward, our hip is automatically pulled backward in order to keep the location of the center of mass invariant. This is also a postural reflex.

Although we do not have any detailed picture of the compound control mechanism of postural reflex, and we do not have identified yet the list of basic disturbance/response pairs for postural reflex similar to what we had for biped locomotion, there are some novel approaches to justify our view. Alexandrov and his colleagues proposed the notion of *eigenmovements* for the regulation of human standing [13, 14]. It is based on the linearized model of robot dynamics with 3 Degree-of-Freedom links. They computed 3 eigenmovements which correspond to classical *hip, knee* and *ankle strategies* that were extensively used in rehabilitation engineering. They thought that the recovering stability from the effect of disturbance is achieved by combining the basic disturbance/response pairs. More specifically, the reflex is called up as a time profile of combination of these basic pairs based on the analysis of how the basic patterns of disturbance are combined to yield the current disturbance.

4 Discussions and Conclusions

Living organisms spend their daily life interacting to changing environments that has huge varieties of possible inputs and disturbances to them. Thinking of this obvious fact, and taking into account the fact that each behavioral component needs to be controlled, it is natural to think that biological control bears essentially multivariable features. More precisely, some control mechanisms to integrate several environmental factors are necessary to produce efficient control actions. Therefore, it is important to exploit multivariable features of biological control and formulate them in terms of control. Some essences of multivariable features have been reported in this paper.

In genetic regulation, combinatorial aspects of environmental changes were emphasized and possible logical patterns of compound control were clarified. In intracellular control research, main paradigm currently assumes that single environmental factor produces single response. To confirm the dominance of this picture, we only see the metabolic map of the cell in which almost all parts of the diagram are occupied by single-input single-output processes. We are sure that the situation will be changed drastically in future and theoretical, as well as experimental, paradigm will be shifted from SISO to MIMO.

We have also presented control of biped locomotion as an example of time-ordered compound control. It was unexpectedly successful in the sense that the balance which was not at all considered in control design emerged spontaneously with extremely good energy performance. Unfortunately, we have not yet succeeded to explain the reason why our approach based on compound control, or successive connections of snapshot postures, was able to keep balance (stability). Some theoretical study disclosed the implication of control strategy from design points of view [11]. There, it was pointed out that the novelty of control scheme is boiled down to variable gain feedback and adaptation of integrators. The precise analysis, however, was far from complete. For detail, see [11].

We have also discussed about posture reflex which is now getting renewed interest from the viewpoint of compound control.

Although our position of exposing the results on compound control is still at a preliminary stage, we are very much confident that compound control has captured the essential multivariable features of biological control.

We acknowledge Drs. Iwao Maeda and Takashi Yamamoto at Toyota Motor Co. for their support and encouragement.

References

1. Kalman, R.E.: On the general theory of control systems. In: Proc. 1st IFAC World Congress, Moscow, vol. 1, pp. 481–492 (1960)
2. Kimura, H.: Control issues in life sciences. In: Bittanti, S. (ed.) Proc. 2nd Convegno Internazionale sui Problemi dell'Automatismo, Consiglio Nozionale delle Ricerche (2006)
3. Tanaka, R.J., et al.: Compound control — adaptation to multiple environmental changes. To appear in Proc. 48th IEEE Conf. on Decision and Control (2009)
4. Shimoda, S., Kimura, H.: Bio-mimetic approach to tacit learning based on compound control. IEEE Trans. Systems, Man and Cybern. (2009)
5. Monod, J., et al.: Genetic regulatory mechanisms in the synthesis of proteins. J. Molecular Biol. 3, 318–356 (1961)
6. Tanaka, R.J., Kimura, H.: Mathematical classification of regulatory logics for compound environmental changes. J. Theor. Biol. 251, 363–379 (2008)
7. Raibert, M.H.: Legged Robots that Balance. MIT Press, Cambridge (1986)
8. McGeer, T.: Passive dynamic walking. Int. J. Robotics Research 9, 62–82 (1990)
9. Vukobratovic, M., et al.: Biped Locomotion – Dynamics, Stability, Control and Application. Springer, Heidelberg (1990)
10. Grizzel, J.W., et al.: Asymptotically stable walking for biped robots: analysis viasystems with impulse effects. IEEE Trans. Automat. Control 46, 51–64 (2001)
11. Kimura, H., Shimoda, S.: Reflex-type control of biped locomotion (to appear)
12. Kimura, H., Jiang, Y.: A PID model of human balance keeping. IEEE Control Systems Magazine 26(6), 18–23 (2006)
13. Alexandrov, A.V., et al.: Biomechanical analysis of movement strategies in human forward trunk bending, I. Modeling. Biological Cybernetics 84, 425–434 (2001)
14. Alexandrov, A.V., et al.: Feedback equilibrium control during human standing. Biological Cybernetics 93, 309–322 (2005)

Law of Large Numbers, Heavy-Tailed Distributions, and the Recent Financial Crisis

Mathukumalli Vidyasagar

1 Introduction

In this paper we examine some possible inter-relationships between, on the one hand, the law of large numbers as it applies to cumulative sums of independent and identically distributed random variables *where the distribution of each random variable is heavy-tailed*, and on the other hand, the recent financial crisis. For the purposes of the present article, a distribution function $\Phi_X(\cdot)$ of a random variable X is said to be 'heavy-tailed' if X has a finite first moment (i.e. an expected value) but not a finite second moment. It is of course possible to construct distribution functions where even the first moment fails to exist. But this case is of no interest to us, because in such a situation it would not make sense to talk about 'law of large numbers'.

The basic conclusion of the discussion below can be stated very simply: Suppose we know $\Phi_X(x)$ only for x belonging to some compact subset of the real numbers (think of a finite interval $[a,b]$), which we can label as the 'observable universe'. Then it is clearly not possible to deduce whether the associated random variable X does, or does not, have finite moments of various orders. The finiteness of the moments of various orders depends on the behavior of $\Phi_X(x)$ as $|x| \to \infty$, that is, when x lies *outside* the observable universe. In such a case, one should impose whatever assumptions one wishes to on Φ_X with a view to making one's life easier, *provided* that the consequences of imposing these assumptions are not visible over the observable universe.

One of the more dramatic properties of cumulative sums of heavy-tailed random variables is that the 'tail probability estimates' of such sums are *qualitatively different* from those of random variables with finite second moments. Specifically, suppose $\{Y_i\}_{i\geq 1}$ is an i.i.d. process. Let us define the cumulative sum $C_l = \sum_{i=1}^{l} Y_i$, and

Mathukumalli Vidyasagar
Cecil & Ida Green (II) Professor, Erik Jonsson School of Engineering & Computer Science,
University of Texas at Dallas, 800 W. Campbell Road, Richardson, TX 75080, USA
e-mail: M.Vidyasagar@utdallas.edu

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 285–295.
springerlink.com

examine the behavior of the tail probability $r(l,\varepsilon) := \Pr\{C_l \geq l\varepsilon\}$. Clearly $r(l,\varepsilon)$ is also the probability of the cumulative *average* exceeding the threshold ε. If Y_1 has a finite second moment, then one can invoke the central limit theorem and show that the tail probability $r(l,\varepsilon)$ approaches the tail probability of the Gaussian distribution. This means that, if Y_1 has finite second moment, then the rare event of C_l significantly exceeding its expected value $lE(Y_1)$ occurs only via *each* of the individual variables $Y_1,\ldots,Y_l$ significantly exceeding its expected value. In contrast, if Y_1 does not have finite second moment but has only finite first moment, then under mild conditions the tail probability $r(l,\varepsilon)$ decays at a *linear rate*, not an exponential rate. See [8] for a good survey of such results, especially Equation (10). As a result, in this case it is just as likely that C_l has a rare event by *just one* of the individual random variables Y_i having a humongous excursion beyond its expected value. This statement is made precise in Theorem 5 below. In the language of large deviation theory, the 'scale' at which the tail probability of the cumulative sum C_l decays is drastically different in the two cases (exponential when Y_1 has finite second moment, and linear otherwise). This property of heavy-tailed random variables was put on a sound mathematical foundation only during the 1980s and 1990s; see for example [12].

On the basis of empirical studies of financial asset prices during the past fifty years or so, it has been observed that '10σ' events and '15σ' events have been observed far more frequently than the Gaussian distribution would predict, and often in quick succession. For instance:

- The standard deviation of the S&P average is about 1.2%, based on very long-term studies. Therefore the 21% decline in the Dow Jones average in October 1987 was a 20σ event. The 8% selloff in 1989 was a 7σ event.
- Such movements are not limited only to stocks. On 24 February 2003, the price of natural gas changed by 42% in one day, representing a 12σ event.

In reality the Gaussian distribution would tell us that such events should take place at most once within the known age of the universe.

Another important point is that, over a long period of time, most of the cumulative change in prices takes place over just a few days. For instance:

- The share price of Akamai from November 2001 until now increased by a factor of 12.10, for a return of 1110%. However, out of this huge return, *just three days* account for 694%, while the remaining hundreds of days account for 416%.
- One may be tempted to suppose that an aggregated quantity like the S&P average will smooth out the variations in individual stock prices, and therefore may not exhibit such behavior. However, the reality is that this phenomenon persists even at the aggregate level. Between 1955 and 2004, the S&P average moved up by a factor of 180. However, if we remove the ten largest movements (most of which were negative), then the return would be a factor of 350.

Thus the empirical evidence suggests that the tail probability of the cumulated sum may possibly exhibit decay over a linear time scale and not an exponential time scale. In other words, the qualitative differences between the behavior of heavy-tailed random variables and those with finite second moment are most definitely 'visible within the observable universe'. Therefore, in order to explain these observed strange behavioral patterns of asset prices, it may be unavoidable to assume that one-period returns on asset prices are heavy-tailed random variables. In other words, the issue is not so much that 10σ or 20σ events occur more frequently than a Gaussian distribution would predict; rather, the issue may be that there is no σ to begin with! The consequences of such an assumption for mathematical finance are quite significant: A major reworking of the theory is needed. The directions that such a theory would need to take are discussed here.

2 Modelling Asset Prices

In this section we give an extremely brief overview of current theories of asset price models. There is more than one way to reach the various conclusions below; moreover, due to length limitations, we do not always state the hypotheses and conclusions with absolute precision, and prefer to indicate the general ideas. In the process, we will definitely violate Einstein's maxim that "Things should be made as simple as possible, but not simpler."

During the past 150 years or so, several deep thinkers have come up with various models of asset price movements as a function of time. Initially (that is to say, until about 50 years ago), the models were *empirical*, that is, they attempted to explain actually observed price movements. During the past 50 years or so, the trend has been somewhat reversed, and modern finance theory is more 'axiomatic' than 'empirical'.

Let us consider an asset whose price is uncertain, usually referred to as a 'stock', denoted by $\{S_t\}$. Rather than model S_t itself, it is more natural to model the *return* on the asset, that is, the ratio $\{S_t/S_0\}$. Let us denote this process by $\{\lambda_t\}$. One of the fundamental assumptions of modern finance theory is that of an 'efficient market'. This assumption means that every player in the marketplace has access to exactly the same information. As a result, if *everyone* in the marketplace were to believe that the share price of IBM will move to \$ 100 tomorrow, then in fact it will move to \$ 100 today itself. Thus today's price has already factored in all that is known and predictable about the future price movements of the asset. By a slight stretch, it follows that the change in the asset price that takes place between today and tomorrow *cannot be predicted* on the basis of all that is known today. Thus, viewed as a stochastic process, $\{\lambda_t\}$ has *independent increments*. Moreover, if we take out seasonal factors (which are presumably known ahead of time and thus computable), the returns process $\{\lambda_t\}$ is also *stationary*. Now a stationary process with independent increments is called a Lévy process. At some risk of imprecision, we can think of a Lévy process as a combination of a countable number of Poisson (jump) processes and a Wiener process. If we add the assumption that the rate process $\{\lambda_t\}$ has *continuous sample paths*, then no jumps are possible and the rate process *must be* a Wiener process with drift. This is the rationale behind the standard assumption that asset prices follow a geometric Brownian motion (GBM) model, i.e., that

$$S_t = S_0 \exp\left[\left(\mu - \frac{1}{2}\sigma^2\right)t + \sigma W_t\right], \quad t \in [0,T],$$

where $\{W_t\}$ is Brownian motion or a Wiener process, the mean μ is called the "drift" of the geometric Brownian motion, and the variance σ is called the "volatility."

In Chapter 1 of the highly readable book by Davis and Etheridge [5], it is mentioned that as far back as 1853, Jules Regnault postulated that, in the absence of any new information, the price of an asset will either go up or go down by a fixed amount, and with equal probability. Moreover, he also observed that the spread in prices is proportional to the *square root* of the time interval. 'Modern' readers will have no difficulty recognizing that what Regnault was describing is a random walk, where the variance of the movement is proportional to time. Moreover, as time is quantized over ever-smaller intervals, the random walk approaches Brownian motion. Consequently the asset prices follow a GBM model.

The above justification for the GBM model is very 'frugal' in the sense that it is based on a minimal set of assumptions. There is only one thing wrong with it: *Actual asset price movements do not follow the GBM model.* See for example Figures 2.6 and 2.7 of [2]. To address this situation while still retaining the tractability of the problem, several authors have proposed various alternatives to the GBM model, such as NIG (Normal Inverse Gaussian), Variance Gamma (VG), CGMY (Car-Geman-Madan-Yor), etc. The book by Benth [2] has a discussion of NIG processes, while the compendium [7] contains several papers that discuss VG and CGMY processes. However, for what can only be called purely 'ideological' reasons, all of these authors cling tenaciously to the assumption that the return process *must have finite variance*. The use of models wherein the return process has infinite second moment is simply dismissed out of hand. Other authors advocate the use of distribution functions with kurtosis different from that of the Gaussian. Recall that the kurtosis κ of a distribution Φ_Y is defined as the ratio $E[Y^4]/(E[Y^2])^2$. For simplicity we are assuming that $E[Y] = 0$ and the correction to the formula for the case of a non-zero mean is obvious. The Gaussian distribution has a kurtosis $\kappa = 3$. However, the very use of the concept of kurtosis implies that one-period asset returns *must have finite variance*, which in turn implies that the central limit theorem holds, which in turn is inconsistent with empirically observed data. So kurtosis is not a panacea either.

3 Stable Distributions

An entirely different approach to modelling the asset return process can be achieved using the notion of stable distributions. Suppose $\{Y_i\}_{i\geq 1}$ is a stationary i.i.d. stochastic process, and let us define the cumulative sum and average processes as follows:

$$C_l := \sum_{i=1}^{l} Y_i, A_l := \frac{C_l}{l} = \frac{1}{l}\sum_{i=1}^{l} Y_i.$$

The widely (mis-)quoted central limit theorem states the following: Suppose each Y_i has finite mean μ and finite variance σ. Then the random variable $(A_l - \mu)/(\sigma l^{1/2})$

converges in distribution to a normal (Gaussian) random variable $N(0,1)$. See for example [4, pp. 167–168] for a precise statement and proof. As an aside, it is interesting to note that, while the central limit theorem was already known in the early 19th century for the case when $\{Y_i\}$ is a Bernoulli process (two possible outcomes), the more general result was proved only in 1901 by A. M. Lyapunov [9], who is somewhat better known in the controls community for his fundamental work on stability theory.

The central limit theorem as stated above applies *only* when each Y_i has both finite mean as well as finite variance. What happens if Y_i has finite mean but not finite variance? Suppose the one-period return process $\{Y_i\}$ is stationary and has a distribution function Φ_Y. What if anything does the distribution function of the cumulative average process $\{A_l\}$ look like as $l \to \infty$? The behavior of the distribution of A_l was worked out in the 1920s by Paul Lévy in a series of papers, and led to the theory of 'stable' distributions. As these are quite germane to the present discussion, we take a brief detour to describe this theory. Chapter 9 of [4] gives a comprehensive treatment of this theory, so our discussion is rather cursory.

Let us make the problem formulation more general and ask: What are the possible limits (in distribution) laws of the random variable

$$S_l := \frac{1}{a_l} \sum_{i=1}^{l} Y_i - b_l, \tag{1}$$

where $a_l \to \infty, b_l/l \to 0$ as $l \to \infty$? If $E(Y_1^2) < \infty$, then the only possible nonzero limit occurs when $a_l \sim l^{1/2}$ and the limit is a Gaussian random variable. So the interesting situation is when $E(Y_1^2) = \infty$. This leads naturally to the notion of a stable distribution. A distribution Φ_X of a random variable X is said to be **stable** if the following statement is true: Suppose Y, Z are independent and identical copies of X. Then for each pair of real numbers a, b, there exist two other real numbers c, d such that $aY + bZ$ is distributionally equal to $cX + d$.[1] The distribution Φ_X is said to be **strictly stable** if $d = 0$, or in other words, $aY + bZ$ is distributionally equal to cX. Finally, the distribution Φ_X is said to be p-**strictly stable** if $c = (a^p + b^p)^{1/p}$. It is easy to verify that every Gaussian distribution is stable, whereas every *zero mean* Gaussian distribution is 2-strictly stable.

Now we quote, without proof, some of the useful properties of stable distributions.

Theorem 1. ([4, Prop. 9.25]) *A distribution Φ can be the limit in distribution of an average of the form* (1) *if and only Φ is a stable distribution.*

This theorem means that, irrespective of what the one-period asset return process looks like, *the cumulative average return process must have a stable distribution.*

Given a random variable X with the distribution function Φ, recall that its **characteristic function** is denoted by $\hat{\Phi}$ and is defined by

[1] This means that $aY + bZ$ and $cX + d$ have the same distribution.

$$\hat{\Phi}(u) := E[\exp(\mathbf{i}uX)] = \int_{-\infty}^{\infty} e^{\mathbf{i}ux} \Phi(dx),$$

where $\mathbf{i}$ denotes $\sqrt{-1}$. With this definition, it is possible to give an explicit characterization of all possible stable distributions.

Theorem 2. ([4, Theorem 9.27]) *Suppose Φ is a stable distribution. Then either Φ is Gaussian, or else there exist a number $\alpha \in (0,2)$ called the 'exponent' of the stable distribution, and constants $m_1, m_2 \geq 0$ and β such that*

$$\log \hat{\Phi}(u) = \mathbf{i}u\beta \, + m_1 \int_0^{\infty} \left(e^{\mathbf{i}ux} - 1 - \frac{\mathbf{i}ux}{1+x^2} \right) \frac{dx}{x^{1+\alpha}} + m_2 \int_{-\infty}^{0} \left(e^{\mathbf{i}ux} - 1 - \frac{\mathbf{i}ux}{1+x^2} \right) \frac{dx}{|x|^{1+\alpha}}. \tag{2}$$

An alternate description of stable distributions is given next.

Theorem 3. ([4, Theorem 9.32]) *A function $f(u)$ is the characteristic function of a stable distribution with exponent $\alpha \in (0,1) \cup (1,2)$ if and only if $f(u) = \exp(\psi(u))$ where $\psi(u)$ is of the form*

$$\psi(u) = \mathbf{i}uc - d|u|^{\alpha} \left(1 + \mathbf{i}\theta \frac{u}{|u|} \tan \frac{\alpha\pi}{2} \right), \tag{3}$$

where c is real, d is real and positive, and θ is real with $|\theta| \leq 1$.

A distribution Φ is said to be in the **domain of attraction** of a stable distribution with exponent $\alpha < 2$ if there is a sequence of constants $\{a_l\}, \{b_l\}$ such that $(1/a_l)\sum_{i=1}^{l} Y_i - b_l \to X$ in distribution, where X has a stable distribution with exponent α, and the Y_i are independent and identical copies of a random variable with distribution Φ. The next result states what kinds of distributions are in the domain of a stable distribution with exponent α.

Theorem 4. ([4, Theorem 9.34]) *The distribution Φ is in the domain of attraction of a stable distribution with exponent $\alpha < 2$ if and only if there exist two nonnegative constants M_-, M_+, not both zero, such that*

$$\lim_{y \to \infty} \frac{\Phi(-y)}{1 - \Phi(y)} = \frac{M_-}{M_+}, \tag{4}$$

and for every $\xi > 0$,

$$\lim_{y \to \infty} \frac{1 - \Phi(\xi y)}{1 - \Phi(y)} = \frac{1}{\xi^{\alpha}} \text{ if } M_+ > 0, \lim_{y \to \infty} \frac{\Phi(-\xi y)}{\Phi(-y)} = \frac{1}{\xi^{\alpha}} \text{ if } M_- > 0. \tag{5}$$

The significance of Theorem 4 is the following: Suppose we have observed that the cumulative average return process is a stable distribution with exponent $\alpha < 2$. Then the distribution of the one-period return process (which need not be stable of course) *must satisfy the conditions in (4) and (5).* Conversely, suppose that one-period returns satisfy (4) and (5) with some exponent α. Then the cumulative return process must have a stable distribution *with the same exponent.*

Note that a distribution that satisfies the two properties in (5) is referred to as being 'Paretan' after the Italian economist Vilfredo Pareto.

Next, let us study the behavior of the tail probabilities of sums (or averages) of heavy-tailed random variables. As before, suppose $\{Y_i\}_{i \geq 1}$ is an i.i.d. process, and define C_l to be the l-th cumulative sum, and A_l to be the l-th cumulative average. For simplicity assume that all Y_i are nonnegative-valued random variables with finite mean μ. Now let us consider two distinct quantities:

$$\gamma_l(\varepsilon) := \Pr\{C_l \geq l\varepsilon\} = \Pr\{A_l \geq \varepsilon\}, \delta_l(\varepsilon) := \Pr\{\max\{Y_1, \ldots, Y_l\} \geq l(\varepsilon - \mu)\}.$$

Thus $\gamma_l(\varepsilon)$ is the probability that the probability that the cumulative sum C_l exceeds the threshold $l\varepsilon$, or equivalently, the probability that the cumulative average A_l exceeds the threshold ε, whereas $\delta_l(\varepsilon)$ is the probability that *any one* of the l random variables $Y_1, \ldots, Y_l$ exceeds l times the threshold $\varepsilon - \mu$. For this situation, we have the following statement:

Theorem 5. ([12], [8, Eq. (10)]) *Suppose each Y_i is nonnegative valued, and define $\alpha := \inf\{p : E[Y_1^p] < \infty\}$. Suppose $\alpha \in (1,2)$, and suppose $\Phi(y) \geq f(y)/y^\alpha$ for some 'slowly varying' function $f(\cdot)$.*[2] *Then $\gamma_l(\varepsilon) \sim \delta_l(\varepsilon)$ for every ε.*

Theorem 5 states that there is a fundamental difference between the way the tail probabilities of A_l behave when each Y_i has finite second moment, and when it does not. Suppose each Y_i has finite second moment. Then $\gamma_l(\varepsilon)$ looks like the Gaussian tail. As a result, the 'rare event' of the cumulative sum C_l exceeding the threshold $l\varepsilon$ occurs much more likely via each individual Y_i assuming a somewhat larger than normal value. This rare event is thus an accumulation of several 'mini-rare events'. In contrast, if Y_i has a stable distribution with exponent $\alpha < 2$, then $\gamma_l(\varepsilon) \sim \delta_l(\varepsilon)$ for every ε. This means that the rare event of the cumulative sum C_l exceeding the threshold $l\varepsilon$ is just as likely to occur because one or two of the individual variables Y_i assume truly humongous values larger than $l(\varepsilon - \mu)$ (remember that l approaches infinity!) as because every individual variable Y_i slightly exceeds the threshold ε.

As mentioned in the introduction, it has already been observed empirically that real asset prices demonstrate precisely this kind of behavior: A significant part of the cumulative change in asset prices over a very long period actually takes place over just a few days. In a very early paper [10], Mandelbrot observed that cotton prices over a century show this kind of behavior, and used this observation as an argument in favor of modeling asset returns by Paretan distributions. However, the fact that this kind of behavior is symptomatic of *any* heavy-tailed asset returns seems to have been fully established only by the 1990s.

[2] A function $f : \mathbb{R} \to \mathbb{R}$ is said to be **slowly varying** if $\lim_{y \to \infty} f(ry)/f(y) = 1$ for every $r > 1$.

4 Implications to Option Pricing, Hedging Strategies Etc.

In this section, we discuss the implications of modelling asset returns by heavy-tailed distributions on option pricing, and hedging. Recall that a 'call' option is an instrument that gives the buyer the right, but not the obligation, to buy the asset at a prespecified price known as the 'strike' price, which is denoted by K. Similarly, a 'put' option gives the buyer the right, but not the obligation, to sell an asset at the strike price. There is also a time of expiry of the option, denoted by T. In a European option, the buyer can exercise his claim only *at* time T, whereas in an American option, the buyer can exercise his option at any time *up to* time T. To simplify the discussion, we restrict ourselves here to European options.

The idea of using a combination of options and the underlying stock to limit one's risk was well-known long before the seminal work of Black-Scholes [3] and Merton [11], The Ph.D. thesis work of Bachelier [1] on computing the 'right' price for an option predates Einstein's work on Brownian motion by a full five years. Strategies such as buying an asset and simultaneously buying a put option, or simultaneously selling a put and buying a call option (or vice versa), have been widely practiced for a few centuries at least. One of the important contributions of Black-Scholes was the idea of 'dynamic hedging', whereby the seller of an option can *continuously adjust* his portfolio between the stock and the bond. In order to have a tractable problem, they had to make some simplifying assumptions and create an idealized situation. The 'unrealism' of their model does not in any way detract from the significance of their contribution.

To simplify the discussion, suppose there is only one asset whose future price movements are uncertain (which is commonly referred to as the 'stock'), and another asset whose future price movement is purely deterministic (which is commonly referred to as the 'bond'). If we discount future prices of the stock by dividing by the price of the bond at that time instant, one effectively gets the stock price *measured in constant (or risk-free) currency*. Let S_t denote the price of the stock at time t, and let the value of the bond equal 1 at all times. Let K denote the strike price of the option, and T the maturity (or expiry) time. Thus, at time T, the (European) option has the value $\{S_T - K, 0\}_+$, where $(\cdot)_+$ denotes the positive part. Thus if the final price S_T exceeds the strike price, then the seller of the option has to incur an expenditure of $S_T - K$ to procure the stock at the prevailing price, and then give it to the buyer of the option – *unless* he already possesses some quantity of the stock. If $S_T \leq K$, then the option is not exercised, and the seller of the option does not incur any expenditure. Now the option-pricing question is merely this: What is the price that the seller of the option should charge at time $t = 0$ for the option?

Since the seller of the option has the 'option' of investing some of the proceeds into the underlying stock, and continually changing his investment portfolio, let us introduce another term, which may be called the 'trading strategy'. Thus the strategy is (loosely speaking) allocating the available money to buy a_t shares and b_t bonds at time t. Clearly the trading strategy has to nonanticipatory; thus a_t must be measurable with respect to the σ-algebra generated by $\{S_\tau\}_{0 \leq \tau \leq t}$. Also, the strategy must be 'self-financing.' This notion can be explained easily in the discrete-time context, and then generalized to continuous-time. Suppose both the stock price S_t and the portfolio (a_t, b_t) are changed only at discrete instants of time. At time $t = t_i$, the portfolio (a_t, b_t) has a value of $a_{t_i} S_{t_i} + b_{t_i}$. At time $t = t_{i+1}$, the value of the stock changes to $S_{t_{i+1}}$, and the value of the portfolio changes to $a_{t_i} S_{t_{i+1}} + b_{t_i}$. (Recall that the value of the bond is constant with time.) Thus the self-financing requirement can be stated as

$$a_{t_{i+1}} S_{t_{i+1}} + b_{t_{i+1}} = a_{t_i} S_{t_{i+1}} + b_{t_i}.$$

In the continuous-time case, the above relation gets replaced by a stochastic integral.

Once a trading strategy has been chosen, the final value of the portfolio will be $X_T = a_T S_T + b_T$. Naturally, X_T is a random number. The quantity $V_T = X_T - \{S_T - K, 0\}_+$ is called the 'hedging error'. It is fairly obvious that the *expected value* of the hedging error $E(V_T)$ has to equal zero. However, the great achievement of Black-Scholes was to show that through dynamic hedging, it is also possible to make the *variance* of V_T equal to zero, thus in principle eliminating all risk. It is necessary to specify the probability measure under which the variance equals zero. In the Black-Scholes formalism, one replaces the 'real-world' measure $\tilde{P}$ by another equivalent measure $\tilde{Q}$ under which the stock price process $\{S_t\}$ becomes a martingale. Actually the variance of V_T is zero under $\tilde{Q}$, which means that $\tilde{Q}$-almost surely $V_T = 0$. But since $\tilde{Q}$ is equivalent to $\tilde{P}$, this means that the hedging error is almost surely equal to zero under the real-world measure $\tilde{P}$ as well. Of course, this conclusion holds under very idealized conditions, such as the asset prices following a GBM model, no transaction costs, assets being available for sale in unlimited quantities and infinitely divisible (no quantization), etc. Nevertheless, there is no doubt that the theory has influenced an entire generation of traders and investors.

If the asset prices do not follow a GBM model, then in general it is not possible to ensure that V_T will turn into a deterministic constant, namely zero. One approach that has been studied in the literature is that of *minimum variance hedging*. Ideally one would like to minimize $E[V_T^2, \tilde{P}]$, but this problem has proved to be intractable. So instead one tries to minimize $E[V_T^2, \tilde{Q}]$ where $\tilde{Q}]$ is a martingale measure equivalent to $\tilde{P}$. Since in general there can be infinitely many equivalent martingale measures, the various hedging strategies are not strictly comparable, thus leading to a most unsatisfactory situation. We do not pursue that topic here for want of space.

Instead, we will raise an entirely different question, namely: What happens if the stock returns process $\{\lambda_t\}$ is a heavy-tailed process, so that X_T may not have finite variance? While we do not have any answers, in the remainder of this section we take a stab at formulating some interesting questions.

Choice of Trading Strategies to Achieve Finite Variance for the Hedging Error: The fact that the returns process $\{\lambda_t\}$ and the final stock price S_T are heavy-tailed does not *necessarily* imply that the hedging error V_T is heavy-tailed (though that would seem plausible). Is it possible to prove rigorously that *no trading strategy exists* that would ensure that V_T has finite variance, or equivalently, that V_T does not have finite second moment under any trading strategy?

Choice of Criterion to Choose a 'Good' or Even 'the Best' Trading Strategy: If indeed V_T has infinite variance, then approaches such as minimum variance hedging have no meaning. Instead one would have to adopt something like the much-maligned VAR (Value at Risk) criterion. One should fix an acceptable level of risk, call it δ, and define the value at risk $v(\delta)$ as

$$v(\delta) := \inf\{v : 1 - \Phi_{V_T}(v) \leq \delta\}.$$

If Φ_{V_T} is continuous and monotonic, the above expression can be simplified to: $v(\delta) = (1 - \Phi_{V_T})^{-1}(\delta)$. Then one can choose the trading strategy so as to minimize $v(\delta)$. An alternative is to compute the 'expected amount at risk', which is the expectation of V_T conditioned on V_T exceeding $v(\delta)$. While these concepts may be appealing, the major drawback is that there is absolutely no hope of getting anything remotely similar to 'closed form formulas', which are one of the most appealing features of Black-Scholes theory.

And finally, we leave the reader with the following question: *Do averaged asset returns really exhibit heavy-tails and self-similarity?* All of the empirical evidence cited in the Introduction really pertains only to *one-period returns*. Thus this empirical evidence could, in principle, be explained by one-period returns exhibiting heavy kurtosis. The one-period returns process can be anything under the sun, and it is only the cumulated average returns process that needs to exhibit a stable distribution. Thus, in order to make the case persuasive, one would need to marshal enough empirical evidence to show that the averaged returns process exhibits the behavior described in Theorem 5.

Acknowledgements. The author thanks Sanjay Bhat and Vijaysekhar Chellaboina for their careful reading of and comments on this manuscript.

References

1. Bachelier, L.: Théorie de la spéculation. Annales Scientifiques de lÉcole Normale Supérieure 3(17), 2186 (1900)
2. Benth, F.E.: Option Theory with Stochastic Analysis. Springer, Berlin (2004)
3. Black, F., Scholes, M.: The theory of options and corporate liabilities. Journal of Political Economy 81, 637–654 (1973)
4. Breiman, L.: Probability. SIAM Publications, Philadelphia (1992)
5. Davis, M.H.A., Etheridge, A.: Louis Bachelier's Theory of $peculation: The Origins of Modern Finance. Princeton University Press, Princeton (2006)
6. Dembo, A., Zeitouni, O.: Large Deviation Techniques and Applications. Springer, Berlin (1998)
7. Fu, M.C., et al. (eds.): Advances in Mathematical Finance. Birkhäuser, Boston (2007)
8. Gantert, N.: A note on logarithmic tail asymptotics and mixing. Statistics and Probability Letters 49, 113–118 (2000)
9. Lyapunov, A.M.: Nouvelle forme du théorème sur la limite de probabilités. Mémoires de l'Académie Impŕiale des Sciences de St. Pétersbourg, ser. 8, 12(5), 1–24 (1901)

10. Mandelbrot, B.: The variation of certain speculative prices. The Journal of Business 36(4), 394–419 (1963)
11. Merton, R.: Theory of rational option pricing. Bell Journal of Economics and Management Science 4, 141–183 (1973)
12. Vinogradov, V.: Refined Large Deviation Limit Theorems. Pitman Research Notes in Mathematics, No. 315 (1994)

Markov Models for Coherent Signals: Extrapolation in the Frequency Domain[*]

Roger W. Brockett

1 Introduction

In this paper we describe a Markov model for a class of stochastic processes whose properties relate to a wide variety of situations in which time-frequency analysis is relevant. These include the production of sound and, in particular, music, frequency modulated digital communication systems, stimulated emission in atomic physics and other aspects of coherent light. More specifically, it provides a new point of view on some ideas explored over the years by Yutaka Yamamoto [1, 2, 3] and others [4] relating to the regeneration of sound from compressed versions obtained by band limiting and sampling. We frame this particular question as one of extrapolation in the frequency domain and argue that analytically tractable signal models can play a significant role in facilitating the reconstruction process. We also touch briefly on the data needed for parameter selection and model validation.

Optimal causal extrapolation for stationary time series was treated already in the earliest days of time series analysis. It is obviously of great interest in fields as diverse as weather prediction, finance and fire control. Extrapolation in the frequency domain is a more recent idea, largely motivated by technologies related to data compression and psychoacoustics but also related to the generation and description of coherent light [5] and other problems in quantum electrodydamics, including quantum computing. In fact, the famous $A-B$ coefficients in Einstein's model for blackbody radiation can be seen as elements of a special case. The purpose of this paper is to describe a class of models which allow one to formulate and resolve questions about *statistically justifiable* extrapolation in the frequency domain. Our analysis centers around three questions:

Roger W. Brockett
School of Engineering and Applied Sciences, Harvard University, USA

[*] This work was supported in part by the US Army Research Office under grant DAAG 55 97 1 0114.

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 299–307.
springerlink.com

1. model selection
2. data collection for model verification
3. extrapolation algorithms

As is well known, the characterization of a wide sense stationary process in terms of its power spectrum does not carry information about the short term fluctuations in energies, or the relative phases, of the signals present in the spectral decomposition. This may seen as a consequence of the fact that in assuming that the process is wide sense stationary such information is necessarily averaged out. Thus the data required to support realistic frequency domain extrapolation is simply not present in the second order statistics of a stationary process. Even so, the assumption of wide sense stationarity is natural and in some applications a virtual necessity for technological reasons. What we will show here is that there are reasonably tractable models for signals, which shape the higher order statistics in relevant ways; these are wide sense stationary processes which capture short time correlations between the various contributions to the power spectrum.

Our models are based on stochastic differential equations of the Itô type involving a finite state part x and a type of random sawtooth wave τ. If N_{ij} are Poisson counters and if e_i denotes the standard basis vectors in $\mathbb{R}^n$ then one way to describe the systems studied here is via

$$dx = \sum (e_j - e_i)e_i^T x dN_{ij} \;\; ; \;\; x \in \{e_1, e_2, ..., e_n\}$$

$$d\tau = dt - \tau \sum e_i^T x dN_{ij}$$

$$y(t) = e^{H\tau}Bx - m \;\; ; \;\; y(t) \in \mathbb{R}^m$$

The rates of the Poisson counters N_{ij} determine the transition rates from x_j to x_i. If the eigenvalues of H are purely imaginary then the components of y will be sums of sinusoids that are restarted at random times. Depending on further details to be introduced later, various phase relationships can be modeled. This class of models does not seem to have been studied before and our results on computing the statistical properties of the solutions seem to be new and to have a number of interesting applications.

Questions centering on the realization of a stationary autocorrelation function via finite state Markov processes have studied for many years and it is known that a given autocorrelation function can be approximated with arbitrary accuracy by a linear function of a time invariant finite state Markov process [6]. Less explored, is the idea of realizing a stationary statistics with time varying systems. This possibility exists for both realizations based on Wiener processes as in

$$dx = Axdt + Bdw \;\; ; \;\; y = c(t)x$$

and for those based on finite state Markov processes. In the situation for which $c(t) = c_0 e^{Ht}$ we have

$$\lim_{t \longrightarrow \infty} \mathscr{E} y(t)y(t+\tau) = c_0 e^{Ht} \Sigma e^{A\tau} e^{H^T(t+\tau)} c_0^T$$

where

$$A\Sigma + \Sigma A^T = -BB^T$$

In this case stationarity requires that there be enough shared structure between H,A and B so as to make

$$c_0 e^{Ht} \Sigma e^{A\tau} e^{H^T(t+\tau)} c_0^T = c_0 \Sigma e^{A\tau} e^{H^T \tau} c_0^T$$

We turn now to the main subject of this paper, the use of finite state models in connection with fixed forms of continuous signals. Consider the (x,τ,y) system introduced above. Even though y is a function of a Markov process whose state space is the product of a finite set and the half-line $[0,\infty)$, we find it convenient to suppress the reference to τ and recast the situation in terms of an evolution equation for (x,y) .

$$dx = \sum (e_j - e_i) e_i^T x dN_{ij} \;\; ; \;\; x \in \{e_1, e_2, ..., e_n\}$$

$$dy = Hy - yB \left(\sum_i e_i^T x dN_{ij} \right) + B \sum_i \sum_j e_i e_j^T x dN_{ij}$$

The details about how this model works and how to make calculations with it will emerge in the next few sections but the following example is suggestive.

Example 1. Consider a four dimensional vector $x \in \{e_1, e_2\} \times \{e_3, e_4\}$ and a model

$$d \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \sum_{i=1}^{4} \begin{bmatrix} -1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} dN_i$$

with

$$y(t) = \sin \omega t (x_1 - x_2) + \cos \omega t (x_3 - x_4)$$

Clearly the expected value of y in steady state is zero. The expression for the autocorrelation function of y simplifies using $\cos t \cos(t+\sigma) + \sin t \sin(t+\sigma) = \cos \sigma$ to

$$\lim_{t \to \infty} y(t) y(t+\sigma) = \cos \omega \sigma e^{-|\sigma|}$$

This example shows that time varying models can be wide sense stationary. However, it does not incorporate any correlation between activity at different frequencies. Later we will describe simple systems which do have this property.

2 An Example

Because the model we will be working with has a number of unusual aspects, it seems appropriate to illustrate some of its properties in a relatively simple setting. For this reason we begin with a detailed examination of a specific example.

Example 2. Let N_{12} and N_{21} be a standard Poisson counters of rates λ_{12} and λ_{21}, respectively. Let x satisfy the Itô equation

$$d\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} x_1 dN_{12} + \begin{bmatrix} 1 \\ -1 \end{bmatrix} x_2 dN_{21}$$

We assume that $x(0) \in \{e_1, e_2\} = E$ and note that from the structure of the equation x evolves in this set. Let τ satisfy the Itô equation

$$d\tau = dt - \tau(x_1 dN_{12} + x_2 dN_{21}) \; ; \; \tau(0) = 0$$

The sample paths of τ are sawtooth-like, evolving with a slope of one but being reset to zero each time x changes it value. Given the counting rates λ_{12} and λ_{21}, it is not difficult to see that in steady state the expected values of x_1 and x_2 are $\lambda_{21}/(\lambda_{12}+\lambda_{21})$ and $\lambda_{12}/(\lambda_{12}+\lambda_{21})$, respectively. From this we see that in steady state the expected number of transitions for x per unit time is $2\lambda_{12}\lambda_{21}/(\lambda_{12}+\lambda_{21})$ and that the expected value of τ is the reciprocal of this number; larger values of the counting rates means the τ is reset more often and consequently will have a smaller average value. To complete the definitions, we let y be given by

$$y(t) = e^{H_1 \tau(t)} x(t) - m$$

with m a constant and

$$H = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

Thus y is a two component vector whose sample paths consist of segments of sine waves with a fixed amplitude and a fixed relative phase, offset by m. In general we will choose m to make the average value of y zero but in some of the more interesting cases this is achieved with $m = 0$.

We turn now to the statistical properties of these variables with the goal of showing that as time evolves their second order statistical properties approach those of a wide sense stationary process with a rational power spectrum but that the higher order statistics s carry phase information, giving the signals "coherence". Taking expectations of both sides of the equation for x we see that

$$\frac{d}{dt}\mathscr{E}x = \begin{bmatrix} -\lambda_{12} & \lambda_{21} \\ \lambda_{12} & -\lambda_{21} \end{bmatrix} \mathscr{E} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = A\mathscr{E}x$$

so that $\mathscr{E}x$ approaches the steady state value given above. To compute the statistical properties of y we observe that y satisfies the the Itô equation

$$dy = Hydt - y(x_1 dN_{12} + x_2 dN_{21}) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} x_1 dN_{12} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} x_2 dN_{21}$$

and that by working with y we can eliminate the direct involvement of τ. Moreover, the tensor product of x and y satisfies

$$d(yx_1) = Hyx_1dt - yx_1dN_{12} + e_1x_2dN_{21}$$

$$d(yx_2) = Hyx_2dt - yx_2dN_{21} + e_2x_1dN_{12}$$

To bring out the linear structure of the evolution we set up the system

$$d\begin{bmatrix} x \\ yx_1 \\ yx_2 \end{bmatrix} = \begin{bmatrix} e_2e_1^TdN_{12} + e_1e_2^TdN_{21} & 0 & 0 \\ e_1e_2^TdN_{21} & Hdt - IdN_{12} & 0 \\ e_2e_1^TdN_{12} & 0 & Hdt - IdN_{21} \end{bmatrix}\begin{bmatrix} x \\ yx_1 \\ yx_2 \end{bmatrix}$$

Taking expectations, and recalling the previously given definition of A we get the lower triangular system

$$\frac{d}{dt}\mathscr{E}\begin{bmatrix} x \\ yx_1 \\ yx_2 \end{bmatrix} = \begin{bmatrix} A & 0 & 0 \\ e_1e_2^T\lambda_{21} & H - I\lambda_{12} & 0 \\ e_2e_1^T\lambda_{12} & 0 & H - I\lambda_{21} \end{bmatrix}\mathscr{E}\begin{bmatrix} x \\ yx_1 \\ yx_2 \end{bmatrix}$$

Of course $\mathscr{E}y$ can be obtained as $\mathscr{E}y = \mathscr{E}yx_1\mathscr{E}x_1 + \mathscr{E}yx_2\mathscr{E}x_2$. From this expression we see that the exponentials describing the rates of decay present in the expression for transient terms in the expectation of y consist of pairwise sums of the eigenvalues of finite state transition matrix A and the matrices $H - \lambda_{ij}I$. Thus A and $H - \lambda_{ij}I$ both play a role in shaping the power spectrum but if H is large and A is small, the eigenvalues of H will dominate. Applying the ordinary version of the chain rule together with the previously given results for various expectations one can arrive at a differential equation for the expected value of y if that is needed.

From the previous equation we see that the steady state averages satisfy

$$\begin{bmatrix} H - I\lambda_{12} & 0 \\ 0 & H - I\lambda_{21} \end{bmatrix}\begin{bmatrix} \mathscr{E}yx_1 \\ \mathscr{E}yx_2 \end{bmatrix} = \begin{bmatrix} e_2\frac{\lambda_{21}}{\lambda_{12}+\lambda_{21}} \\ e_1\frac{\lambda_{12}}{\lambda_{12}+\lambda_{21}} \end{bmatrix}$$

so that in steady state the expression $\mathscr{E}y = \mathscr{E}yx_1\mathscr{E}x_1 + \mathscr{E}yx_2\mathscr{E}x_2$ becomes

$$\mathscr{E}y = (H - I\lambda_{12})^{-1}e_1\left(\frac{\lambda_{21}}{\lambda_{12} + \lambda_{21}}\right)^2 + (H - I\lambda_{21})^{-1}e_2\left(\frac{\lambda_{12}}{\lambda_{12} + \lambda_{21}}\right)^2$$

Note that if $\lambda_{12} = \lambda_{21} = \lambda$ this simplifies to

$$\mathscr{E}y = \frac{1}{2}(H - I\lambda)^{-1}(e_1 + e_2)$$

Equally important for our purposes are the second order statistics of y. Just as we needed to work with yx_i to get the mean for y, we need to work with yy^Tx_i to compute the second order statistics of y. This requires that we set up equations for the expected value of

$$M = \begin{bmatrix} x \\ yx_1 \\ yx_2 \end{bmatrix}\begin{bmatrix} x^T & y^Tx_1 & y^Tx_2 \end{bmatrix}$$

A number of simplifications occur here because $x_i^2 = x_i$ and $x_i x_j = 0$ if $i \neq j$. The only nontrivial calculations are

$$dyy^T x_i^2 = Hyy^T x_i dt + x_i yy^T H^T dt - \sum_{j=1}^{n} yy^T x_i dN_{ij} + \sum_{j=1}^{n} e_i e_j^T x_2 dN_{ji}$$

Thus we have

$$d\begin{bmatrix} x \\ yy^T x_1 \\ yy^T x_2 \end{bmatrix} = \begin{bmatrix} e_2 e_1^T x dN_{12} + e_1 e_2^T x dN_{21} \\ (Hyy^T x_1 + yy^T x_1 H^T)dt - yy^T x_1 dN_{12} + e_1 e_1^T dN_{21} \\ (Hyy^T x_2 + yy^T x_2 H^T)dt - yy^T x_2 dN_{21} + e_2 e_2^T dN_{12} \end{bmatrix}$$

Taking expectations gives a set of three equations for the variances.

$$\frac{d}{dt}\mathscr{E}\begin{bmatrix} x \\ yy^T x_1 \\ yy^T x_2 \end{bmatrix} = \begin{bmatrix} e_2 e_1^T \mathscr{E} x\lambda_{12} + e_1 e_2^T \mathscr{E} x\lambda_{21} \\ \mathscr{E}(Hyy^T x_1 + yy^T x_1 H^T) - \mathscr{E} yy^T x_1 \lambda_{12} + e_1 e_1^T \lambda_{21} \\ \mathscr{E}(Hyy^T x_2 + yy^T x_2 H^T) - \mathscr{E} yy^T x_2 \lambda_{21} + e_2 e_2^T \lambda_{12} \end{bmatrix}$$

The steady state value of the autocorrelation function can be obtained from the steady state variance and the differential equation for the means in accordance with the usual differential equation for the covariance

$$\frac{d}{d\sigma}\mathscr{E}z(t)z^T(+\sigma) = \mathscr{E}z(t)z^T(t+\sigma)F^T$$

where Fz is the right-hand side of the equation for the evolution of the mean.

Aspects of this model that we wish to emphasize include:

1. The model admits a wide sense stationary steady state. whose correlation function and power spectrum can be directly related to the parameters of the model.
2. The wave forms are segments of sinusoids or damped sinusoids, with fixed amplitude and phase relationships, reset at random points in time. These give a coherence to the signals not present in the standard Gauss-Markov models.
3. All moments can be computed as solutions to linear differential equations, allowing the analysis of high order correlations.

3 The General Model

The example of the previous section will now be generalized in two significant ways. However, even with this added generality it will still be possible to make the same calculations relating the model parameters to the statistical properties of the solutions.

First of all we replace the two state model for x used in the example above with a model involving an arbitrary (finite) number of states. The sample path equation takes the form

$$dx = \sum (e_j - e_i)e_i^T x dN_{ij} \ ; \ x \in \{e_1, e_2, ..., e_n\}$$

The choices for the rates of the counters provide the same flexibility as the usual choices for the transition probabilities in a finite Markov chain. The model for y allows for an arbitrary (finite) set of sinusoids or damped sinusoids. It could be written as

$$y(t) = Ce^{H\tau}Bx(t) \ ; \ y(t) \in \mathbb{R}^p$$

but we prefer to eliminate τ and describe y in differential form

$$dy = CHydt - CyB \left(\sum_i e_i^T x dN_{ij} \right) + CB \sum_i \sum_j e_i e_j^T x dN_{ij}$$

where H is an arbitrary m-by-m matrix, and B and C are rectangular matrices of the appropriate sizes. The flexibility provided by these extensions can be used in the following ways:

1. If H has several rationally related imaginary eigenvalues then randomly generated multi-frequency segments with fixed or random phase relationships can be modeled. This is useful for frequency domain extrapolation. In particular in this situation the relative sizes of the entries in the columns of B determine the relative strength of the various harmonics present in y.
2. The amplitude of a harmonic present at one frequency can be correlated to the amplitude of a harmonic present at a second frequency. This is never the case for Gauss-Markov processes generated by $dx = Axdt + Bdw$; $y = cx$
3. One can shape the probability distribution of the phase difference between two components of y.
4. Higher order statistics can be computed using the higher order moment equations as in [7].

We will illustrate these points with a series of examples.

Example 3. Let the x process be as above with the λ_{ij} all equal to λ. Let H and B be given by

$$H = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & -3 & 0 \end{bmatrix} \ ; \ B = \begin{bmatrix} 2 & 1 \\ 0 & 0 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}$$

and let $c = e_1^T + e_3^T$. Thus when x switches from e_1 to e_2, y switches from $y = 2\cos(t-t_k) + \cos 3(t-t_k)$ to $y = \cos t(t-t_k)$. If $\lambda << 1$ so that x will typically spend many periods in a state before switching, the expected power in y is concentrated near $\omega = 1$ and $\omega = 3$. Without doing a detailed calculations, we can see that a moving average approximation to the power near $\omega = 1$ such as

$$p(1) = \frac{1}{T^2}\mathscr{E}\left(\int_{-T}^{T} e^{-.5|t-\sigma|} cos(t-\sigma)y(t-\sigma)d\sigma\right)^2$$

is more strongly correlated to a similar moving average for the power near $\omega = 3$

$$p(3) = \frac{1}{T^2}\mathscr{E}\left(\int_{-T}^{T} e^{-.5|t-\sigma|} cos(3t-3\sigma)y(t-\sigma)d\sigma\right)^2$$

than would be the case for a Gauss-Markov model with the same power spectrum.

For carrying out calculations of the type done in example one it is better to postpone the consideration of the effect of C until the very end, working instead with a differential equation for the quantity $e^{H\tau}Bx$.

Finally, If F has eigenvalues lying on the imaginary axis the system can be viewed as modeling a frequency selection process. As such, it would provide a basis for extrapolation in frequency domain.

4 Data

Given a stochastic process y we can let it excite a non stationary dynamical system with the view of extracting information about the harmonic content of y. For example the solution of the equations

$$\frac{d}{dt}\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} \sin t \\ \cos t \end{bmatrix} y(t) = -z + fy$$

will be such that $z_1^2 + z_2^2$ is a measure of the power y has in the vicinity of $\omega = 1$. One way to make this precise is to assume that y is generated by a Gauss-Markov process

$$dx = Axdt + Bdw \; ; \; y = cx$$

and to calculate the variance of z as a function of the power spectrum of y. In terms of the notation

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = \mathscr{E}\begin{bmatrix} xx^T & xz^T \\ zx^T & zz^T \end{bmatrix}$$

we have

$$\dot{\Sigma} = \begin{bmatrix} A & 0 \\ fc & -I \end{bmatrix}\Sigma + \Sigma\begin{bmatrix} A^T & c^T f^T \\ 0 & -I \end{bmatrix} + \mathscr{E}\begin{bmatrix} BB^T & 0 \\ 0 & 0 \end{bmatrix}$$

and so in steady state Σ_{11} is a constant. The differential equations for the remaining quantities are

$$\dot{\Sigma}_{12} = (A-I)\Sigma_{12} + \Sigma_{11}c^T f^T$$
$$\dot{\Sigma}_{22} = -2\Sigma_{22} + fc\Sigma_{12} + \Sigma_{21}c^T f^T$$

The equation for Σ_{12} shows clearly that if $A - I$ has a an eigenvalue near i then f will couple strongly with this mode and $fc\Sigma_{12} + \Sigma_{12}c^T f^T$ will, in turn, generate a large response in z. Not only does the size of $||z||$ reflect the power near $\omega = 1$ but the values of z_1, z_2 and $z_1 z_2$ track phase information as time evolves.

Elaborating on this idea, if we replace the z system by a bank of such filters having f's with different frequencies, then one can collect not only information on the power at a large number of frequencies $\omega_1, \omega_2, \omega_3, ...\omega_k$, but also, through the formation of fourth order averages such as $\mathscr{E}(z_1^2 + z_2^2)(z_3^2 + z_4^2)$, it is possible to evaluate the correlation between the power at different frequencies. These data provided a basis for extrapolation in the frequency domain.

In view of the orthogonality properties of sinusoids it is to be expected that the value of integrals such as

$$\eta = \int_{-\infty}^{\infty} e^{-|t|} \cos t \cos(3t + \phi) dt$$

will be small even if $\cos t$ and $\cos 3t$ are strongly represented in the signal. On the other hand, and integral such as

$$\eta = \int_{-\infty}^{\infty} e^{-|t|} z_1^2 z_3^2 dt$$

will reflect the extent to which two different sin functions are present simultaneously, although it is insensitive to their signs. Thus we see that these particular fourth order statistics can form the basis for frequency extrapolation.

References

1. Yamamoto, Y., Hara, S.: Sampled-data control systems I Their representations. Systems/Information/Control 43(8), 436–443 (1999)
2. Yamamoto, Y., Hara, S.: Sampled-data control systems II Frequency response and its computation. Systems/Information/Control 43(10), 561–568 (1999)
3. Hara, S., Yamamoto, Y.: Sampled-data control systems III Optimal control problems and their solutions. Systems/Information/Control 43(12), 660–668 (1999)
4. Budsabathon, C., Nishihara, A.: Bandwidth Extension with Hybrid Signal Extrapolation for Audio Coding. IEICE Trans. Fundamentals E90-A, 1564–1569 (2007)
5. Loudon, R.: The Quantum Theory of Light. Oxford University Press, Oxford (2000)
6. Brockett, R.W.: Stationary Covariance Generation with Finite State Markov Processes. In: Proc. 1977 Joint Automatic Control Conference, pp. 1057–1060 (1977)
7. Brockett, R.W.: Parametrically Stochastic Linear Systems. In: Wets, R.J.B. (ed.) Stochastic Systems: Modeling, Identication, and Optimization. Studies in Nonlinear Programming, vol. 5, pp. 8–21. North Holland Publishers, The Netherlands (1976)

Digital Signal Processing and the YY Filter

Bruce Francis

Abstract. One of Professor Yamamoto's important research contributions has been in audio signal processing. His YY filter is a digital-to-analog converter for the playback of an audio file. This tutorial paper begins by reviewing some topics in audio signal processing, including the YY filter. This is followed by a walk-through of some concepts and results in digital signal processing using the framework of linear operators on Hilbert space, a subject pioneered by Professor Yamamoto.

1 Introduction

This paper is intended as a tribute to Professor Yamamoto's contributions to digital signal processing, in particular the YY filter for audio playback [10]. It is a tutorial on multirate digital signal processing, audio signal processing in general, and the YY filter specifically. Professor Yamamoto pioneered the use of operator theory in signal processing and this paper follows in his very elegant framework.

2 Digital Storage of Audio Signals

An iPod is a beautiful piece of engineering. It is elegant in physical design, intuitive to use, and amazingly small. In addition, under the hood is a wealth of signals and systems theory. This section is a tutorial on how audio signals are stored on media (CDs, DVDs, flash memories). Popular references are [5, 6].

The human ear is not a perfect filter, but it can hear a pure tone below about 20 kHz and above about 20 Hz. By the sampling theorem, an audio signal should therefore be sampled at greater than 40 kHz. For practical reasons, the standard of

Bruce Francis
Department of Electrical and Computer Engineering, University of Toronto,
10 King's College Rd., Toronto, Ontario M5S 3G4, Canada
e-mail: `Bruce.Francis@utoronto.ca`

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 309–320.
springerlink.com

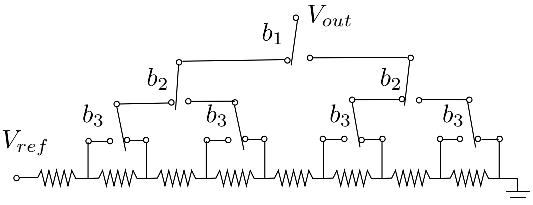


Fig. 1 A 3-bit DAC.

44.1 kHz was adopted. So to digitally store an audio signal one would initially think of merely sampling at 44.1 kHz and then quantizing each sample with sufficiently many bits, say 18 or 20. This wouldn't work, for two main, interesting reasons.

First, music signals have lots of high frequency harmonics. So to avoid aliasing from the sampling process the audio signal must be lowpass filtered. But it's difficult and expensive to build an analog filter with a sharp cutoff at 40 kHz and no phase distortion up to that frequency. So audio signals are oversampled by a large integer factor, say 64 (but it could be up to 512), and filtered by an anti–aliasing filter that has negligible distortion up to 40 kHz and gradually tapers off until 64 × 40 kHz. After sampling, the signal is lowpass filtered in the discrete-time domain and then downsampled to 44.1 kHz ; filters with sharp cutoff are cheaper and easier to build in the digital domain.

The second reason relates to the problem of building reliable systems out of faulty components. It's not possible to build a simple quantizer of 18 bits. The easiest way to see this is with the reverse direction of a digital-to-analog converter (DAC), as in Figure 1. The filter ladder converts a 3-bit word $b_1b_2b_3$, where $b_i = 0$ or 1, to the analog voltage V_{out}:

$$V_{out} = V_{ref}(b_1 2^{-1} + b_2 2^{-2} + b_3 2^{-3}).$$

Consider an N-bit DAC like this one and suppose the resistors are not perfect but have some manufacturing tolerance so that their values in ohms are $R(1 \pm \delta)$, where δ is, say, 10^{-3}, that is, 0.1%. Suppose we want to convert the N-bit binary number $10\cdots0$ to analog without any error. In the resistor ladder, we need to generate half the voltage V_{ref} by selecting the right half of the resistors. There are 2^N resistors, and the voltage divider rule gives that our goal is

$$\frac{\text{sum of right-half resistances}}{\text{sum of all resistances}} = \frac{1}{2}.$$

We will make the smallest error possible of 1 bit when the ratio on the left equals $2^{-1} + 2^{-N}$ instead of 2^{-1}, that is, it corresponds to $100\cdots01$ instead of $100\cdots0$. Suppose the right half resistors are high at $R(1+\delta)$ ohms and the left half are low at $R(1-\delta)$ ohms. Then the equation becomes

$$\frac{\frac{1}{2}2^N R(1+\delta)}{\frac{1}{2}2^N R(1+\delta)+\frac{1}{2}2^N R(1-\delta)} = \frac{1}{2}+\frac{1}{2^N}.$$

Since $\delta = 10^{-3}$, this simplifies to $10^3 = 2^{N-1}$. Thus $N = 1+3\log_2(10) = 11$. To recap, the converter may make a 1-bit error if $N = 11$ and the resistors have an accuracy of 0.1%. Thus the number of obtainable bits without error is only $N = 10$. But we need 18 or 20 to avoid perception by the ear.

Figure 2 shows the basic idea behind audio digital storage. The analog input is $x_c(t)$. The input lowpass filter has negligible distortion up to 44.1 kHz and reaches close to zero gain at frequency 64 × 44.1 kHz. The feedback loop with an integrator is a $\Sigma-\Delta$ system: Σ signifies integration (summation) and Δ signifies negative feedback (difference). The block Q is a 1-bit quantizer, i.e., the output equals $+1$ if the input is non-negative and equals -1 if the input is negative. The other components are S, the continuous-time sampler, a zero-order hold (ZOH), a discrete-time lowpass filter, with cutoff frequency $\pi/64$, and a downsampler by the factor 64, whose nth output equals its $64n$th input. To see how this system works, first note that, because of the analog LPF, Figure 2 is equivalent to Figure 3, which in turn is equivalent to Figure 4. This last system has a discrete integrator in a feedback loop. The integrator has infinite DC gain, without the requirement of accurate circuit component R and C values. How can we get good analog-to-digital conversion

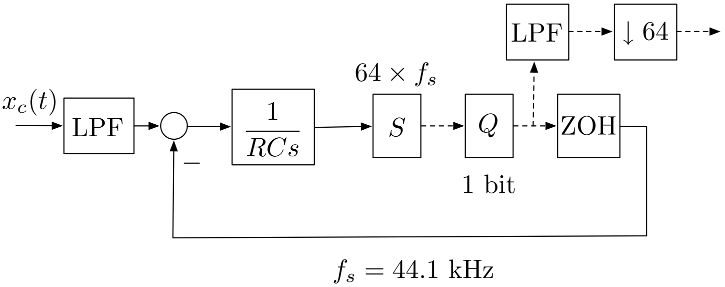


Fig. 2 A $\Sigma-\Delta$ oversampled analog-to-digital converter; one channel.

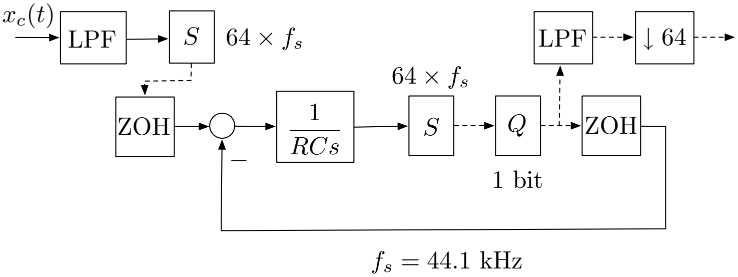


Fig. 3 Equivalent since the input is oversampled.

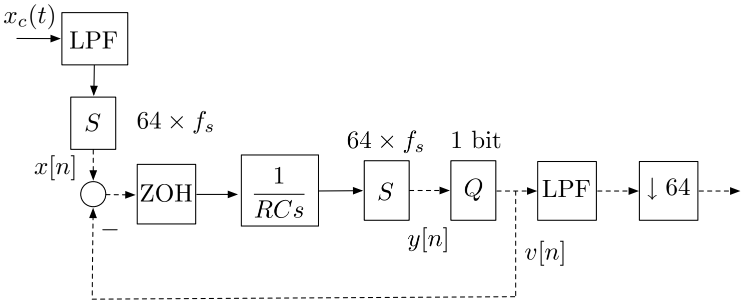


Fig. 4 Equivalent by block diagram manipulation.

with a quantizer with only 1 bit? Because the average of $v[n]$ over a long window equals the average of $x[n]$ over the same window. To see this, let's say the window is from $n = 0$ to $n = N$. We have

$$y[n+1] = y[n] + k(x[n] - v[n]), \quad k = T/RC.$$

This yields

$$y[N] = y[0] + k \sum_{n=0}^{N-1} (x[n] - v[n])$$

which in turn leads to

$$\frac{y[N] - y[0]}{N} = k \left\{ \text{avg}_{0 \leq k \leq N-1} x[n] - \text{avg}_{0 \leq k \leq N-1} v[n] \right\}.$$

Under the assumption that the loop is stable, so that $y[n]$ is bounded, the left-hand side is small for N large, and therefore

$$\text{avg}_{0 \leq k \leq N-1} x[n] \approx \text{avg}_{0 \leq k \leq N-1} v[n].$$

But, being highly oversampled, $x[n]$ is nearly constant over the window. Consequently, the final output in Figure 4 is a very close approximation to $x_c(t)$ sampled at 44.1 kHz. This signal, after digital encoding, is the stored signal.

3 Digital Playback of Audio Signals

The playback of an audio signal is much simpler in principle: decode, zero-order hold, and lowpass filter. Figure 5 shows the architecture for playback via the YY filter. The input to the diagram is the signal stored at 44.1 kHz. The commercial playback system by SANYO is proprietary, but the basic idea is to upsample (typically $L = 8$), then filter, as in the diagram.

The filter G is designed by formulating the error system in Figure 6 and computing G to minimize the $\mathscr{L}_2$-induced norm of the error system.

The YY filter G has to be implemented somehow. There are many ways to do it. In the next section we look at one way, its choice being related primarily to esthetics—it's quite cute.

4 Digital Filters in an Operator Framework

Signal processing is largely a linear systems subject and therefore the appropriate framework is linear operator theory. Many of the systems have signals running at different data rates, for example, sample-rate changers, oversampled analog-to-digital converters, and filter banks for subband coding. Consequently the systems are time varying. Standard references are [4, 8], although they don't use the language of operators. This section introduces the framework of operator theory for signal processing, and looks at some example structures and basic questions.

Definitions and Notation

A discrete-time signal will be denoted, for example, by $x[n]$, where n is an integer. Thus $n \in \mathbb{Z}$ and $x[n] \in \mathbb{R}$, or x is a mapping from $\mathbb{Z}$ to $\mathbb{R}$. Signal space will be the Hilbert space ℓ_2 of square-summable sequences. Then a digital filter is a bounded linear operator on ℓ_2, denoted for example by H.

The unit delay, mapping $x[n]$ to $x[n-1]$, is denoted by U, and H is time invariant (or Toeplitz) iff H commutes with U. A causal finite impulse response (FIR) filter is a polynomial in U. For example,

$$\frac{1}{3}(I+U+U^2)$$

is the filter that averages the most recent three input values. For a positive integer L, the downsampler D_L maps $x[n]$ to $x[Ln]$. We drop the subscript L to avoid clutter.

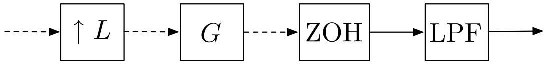


Fig. 5 Playback of an audio signal. The YY filter is G.

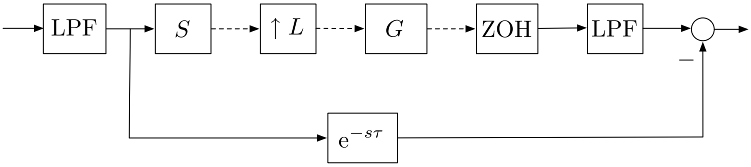


Fig. 6 The YY filter is designed by optimizing the error system.

Adjoint operators will be denoted by H^*; for example U^* is forward advance and D^* is the upsampler, whose action (for integer L) is

$$x[n] \mapsto \begin{cases} x[n/L], & \text{if } L \text{ divides } n \\ 0, & \text{if not.} \end{cases}$$

Some digital filters are multi-input and/or multi-output, in which case they act on the vector-valued space ℓ_2^m for some positive integer m. A memoryless LTI operator is a pure gain: $y[n] = ax[n]$. A memoryless MIMO LTI operator on ℓ_2^m is a pure gain matrix: $y[n] = Ax[n], A \in \mathbb{R}^{m\times m}$.

Block Signal Processing

Let D denote downsample by 2 and define

$$B = \begin{bmatrix} D \\ DU^* \end{bmatrix}.$$

Thus $B : \ell_2 \longrightarrow \ell_2^2$ and its action is

$$x[n] \mapsto \begin{bmatrix} x[2n] \\ x[2n+1] \end{bmatrix}.$$

Clearly B is the blocking operator that takes a 1-D signal and partitions it into blocks of length 2. The operator B is unitary: $B^*B = I$. Written out, this equation is

$$D^*D + UD^*DU^* = I. \tag{1}$$

The simplest structure using block processing is shown in Figure 7. The input x is broken up into blocks or segments of length 2, then passed through the MIMO system G, and the two outputs are combined to produce y. This structure was studied by Vaidyanathan and Mitra [9], who asked the question, when is this system linear time-invariant (LTI), or alias free in their terminology? To illustrate the elegance and power of the operator notation, let's answer that question very simply.

The system from x to y is B^*GB. It is LTI iff it commutes with U:

$$U(B^*GB) = (B^*GB)U.$$

This equation is equivalent to

$$(BUB^*)G = G(BUB^*),$$

i.e., G commutes with BUB^*. But

$$BUB^* = \begin{bmatrix} D \\ DU^* \end{bmatrix} U \begin{bmatrix} D^* & UD^* \end{bmatrix} = \begin{bmatrix} 0 & U \\ I & 0 \end{bmatrix}.$$

However

$$G=\begin{bmatrix}G_{11} & G_{12}\\ G_{21} & G_{22}\end{bmatrix} \text{ and } \begin{bmatrix}0 & U\\ I & 0\end{bmatrix}$$

commute iff G has the form

$$G=\begin{bmatrix}G_1 & UG_2\\ G_2 & G_1\end{bmatrix},$$

where G_1, G_2 are LTI. Such a matrix is called pseudocirculant in [9].

Block Filtering

Most digital filters in applications are FIR. The YY filter, for example, is designed using state-space methods, and then approximated by an FIR filter for implementation, the order being roughly 100 to 150, depending on the oversampling ratio.

So let us turn to the problem of implementing a causal FIR filter H. The direct way to implement H is by convolution of the impulse response function $h[n]$ with the input $x[n]$ to produce the output $y[n]$. The simplest block processing implementation of H would have G a memoryless MIMO system: The blocks are processed one after another. The identity

$$H = B^*(BHB^*)B$$

suggests we might try $G = BHB^*$. However it turns out that BHB^* is memoryless iff H is itself memoryless. Hence this naive approach doesn't work. As we shall see, appending zeros to the input blocks is the way forward.

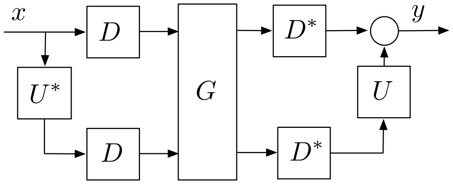


Fig. 7 A naive attempt at block filtering.

Consider the very simple first-order case $H = U$, that is, the transfer function is z^{-1}. The order of the filter is 2, because, as a polynomial in U, H has degree 1 and hence two coefficients. From (1) we have

$$U = U(D^*D + UD^*DU^*) = \begin{bmatrix}D\\ DU^*\\ DU^{*2}\end{bmatrix}^* \begin{bmatrix}0 & 0 & 1\\ 1 & 0 & 0\\ 0 & 1 & 0\end{bmatrix} \begin{bmatrix}D\\ DU^*\\ 0\end{bmatrix}.$$

Likewise, the general order 2 filter factorizes as

$$h[0]I + h[1]U = \begin{bmatrix}D\\ DU^*\\ DU^{*2}\end{bmatrix}^* \begin{bmatrix}h[0] & 0 & h[1]\\ h[1] & h[0] & 0\\ 0 & h[1] & h[0]\end{bmatrix} \begin{bmatrix}D\\ DU^*\\ 0\end{bmatrix}.$$

The form is $H = B_e^* G B_z$. Let's look at these three factors. Moving from right to left we have first the blocking operator appended with a zero block:

$$B_z = \begin{bmatrix} D \\ DU^* \\ 0 \end{bmatrix} : \ell_2 \longrightarrow \ell_2^3.$$

In the middle is $G : \ell_2^3 \longrightarrow \ell_2^3$ that acts like this:

$$w = Gv, \;\; w[n] = Cv[n], \;\; C = \begin{bmatrix} h[0] & 0 & h[1] \\ h[1] & h[0] & 0 \\ 0 & h[1] & h[0] \end{bmatrix}, \;\; w, v \in \ell_2^3.$$

This is a memoryless operator of multiplying by a circulant matrix. Finally, we have the adjoint of the extended blocking operator:

$$B_e = \begin{bmatrix} D \\ DU^* \\ DU^{*2} \end{bmatrix}.$$

The block diagram corresponding to this factorization is Figure 8. All arrows represent signals in ℓ_2 and all boxes stand for operators.

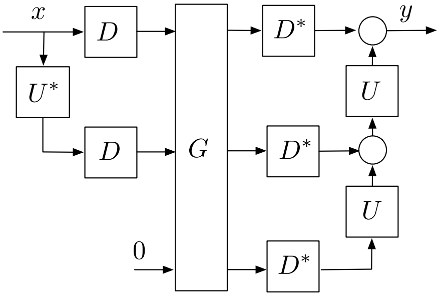


Fig. 8 Implementation of order-2 FIR operator by a memoryless circulant operator.

Figure 9 is an equivalent block diagram using more conventional notation.

Of course, the implementation in Figure 8 is more complicated than the original $h[0]I + h[1]U$, so to see the point of this factorization we turn now to the general case.

For a general positive integer N, D denotes downsample by N and B denotes the blocking operator $\ell_2 \longrightarrow \ell_2^N$. Append by $N-1$ zeros to get $B_z : \ell_2 \longrightarrow \ell_2^{2N-1}$, and extend by $N-1$ terms $DU^{*N}, DU^{*(N+1)}, \ldots$ to get $B_e : \ell_2 \longrightarrow \ell_2^{2N-1}$. Finally, P denotes the canonical $(2N-1) \times (2N-1)$ permutation matrix

$$P = \begin{bmatrix} 0 & 1 \\ I & 0 \end{bmatrix}.$$

Every $(2N-1) \times (2N-1)$ circulant matrix is a polynomial in P.

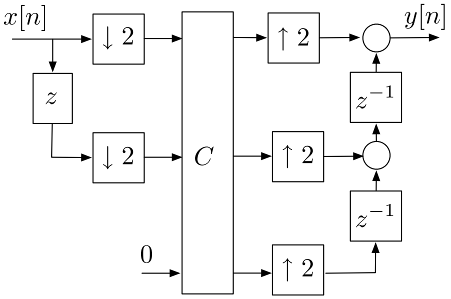


Fig. 9 Conventional block diagram.

Theorem 1. *Let H be a polynomial in U of order N:*

$$H = h[0]I + \cdots + h[N-1]U^{N-1}.$$

Then $H = B_e^* G B_z$, where $G : \ell_2^{2N-1} \longrightarrow \ell_2^{2N-1}$ is the memoryless operator of multiplication by the circulant matrix

$$C = h[0]I + h[1]P + \cdots + h[N-1]P^{N-1}.$$

Block Filtering and FFT

It is a beautiful fact, fundamental in signal processing, that every circulant matrix is diagonalized by the Fourier matrix [2]. Let C be as in the theorem. Let Δ be the $(2N-1) \times (2N-1)$ diagonal matrix, the diagonal elements being the discrete Fourier transform (DFT) of $(h[0], \ldots, h[N-1], 0 \ldots, 0)$. These are the eigenvalues of C. And let F be the $(2N-1) \times (2N-1)$ Fourier matrix: If the rows and columns are numbered from 0 to $2N-2$, element (m,n) of F is

$$\mathrm{e}^{-j2\pi mn/(2N-1)}.$$

Then $FCF^{-1} = \Delta$. Therefore the FIR filter can be implemented by the structure in Figure 10, shown for input blocks of length two only and where the δ's are the diagonal elements of Δ.

Of course, multiplying by F can be performed quickly using the fast Fourier transform algorithm (FFT).

Experiments have been done to study the speedup of the implementation in Figure 10 over that in Figure 9. The graph in Figure 11 taken from [7] is typical.

Aside on Orthogonal Frequency-Division Multiplexing

In the preceding a Toeplitz operator was factored into a product where one of the factors is circulant. Here we perform a dual construction.

Orthogonal frequency-division multiplexing (OFDM) is used in wireless data communication, such as the IEEE standard 802.11a. OFDM is an instance of

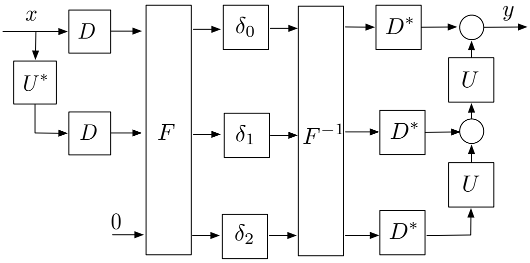


Fig. 10 Block filtering using FFT.

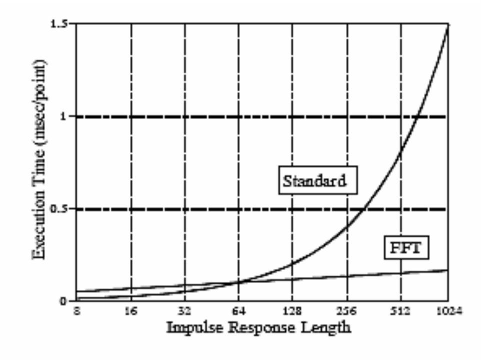


Fig. 11 Block filtering using FFT.

multicarrier modulation. It would take too long to do OFDM in detail, so the goal here is just to describe the underlying idea.

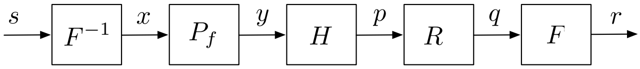


Fig. 12 OFDM.

To keep things simple, we'll do the finite matrix case. The OFDM structure to transmit a binary signal $s[n]$ of length N is illustrated in Figure 12. In this diagram, $s[n]$ and $r[n]$ are each length-N signals—the signal sent and the signal received. The signals $x[n], q[n]$ are length N too, while $y[n], p[n]$ are padded. The matrix H is a causal Toeplitz matrix representing an LTI channel. The matrices P_f and R are chosen to make RHP_f circulant. For example, if

$$H = \begin{bmatrix} h[0] & 0 & 0 & 0 \\ h[1] & h[0] & 0 & 0 \\ 0 & h[1] & h[0] & 0 \\ 0 & 0 & h[1] & h[0] \end{bmatrix}, \tag{2}$$

then

$$P_f = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad R = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Thus P_f cleverly adds a prefix; y is defined by

$$y = \begin{bmatrix} y[0] \\ y[1] \\ y[2] \\ y[3] \end{bmatrix} = \begin{bmatrix} x[2] \\ x[0] \\ x[1] \\ x[2] \end{bmatrix}.$$

And R removes the prefix from p to produce q. The matrix $C = RHP_f$ is

$$\begin{bmatrix} h[0] & 0 & h[1] \\ h[1] & h[0] & 0 \\ 0 & h[1] & h[0] \end{bmatrix}.$$

We conclude that the matrix from s to r is diagonal: $r = \Delta s$. No intersymbol interference!

5 Summary of Main Points

Digital audio processing, both storage and playback, involves several interesting techniques, some control theoretic, to overcome practical constraints: Oversampling is used because it is easier to design and build a sharp cutoff filter in discrete time; integral feedback is used because it tolerates component imperfections. As a consequence of oversampling, a 1-bit quantizer suffices.

The YY filter is designed by $\mathscr{H}_\infty$ optimization of an error system. Other digital signal processing problems have been tackled this way, for example, [3, 1].

The formalism of linear operators on Hilbert space is very applicable. Factoring a Toeplitz operator into a block circulant matrix is the key construction for fast implementation via FFT. A related result is the basis for OFDM.

Acknowledgements. I had the good fortune to teach my department's DSP course in 2008/2009. Thanks to my signal processing colleagues, especially Wei Yu and Richard Schreier, for helping me learn the subject.

End Note

It is a pleasure to acknowledge my deep respect for Yutaka, my admiration of his many contributions over the years, and my enjoyment of our friendship. Our research has overlapped on several topics—regulator theory, sampled-data control systems, multirate digital signal processing. We co-authored a CDC paper, and we both had the good fortune to supervise Hideaki Ishii. In 2004 I had the pleasure of listening to Maria Callas in Yutaka's lab. That experience was one of the motivations for this review paper.

References

1. Chen, T., Francis, B.A.: Design of multirate filter banks by H^∞ optimization. IEEE Trans. Signal Processing 43, 2822–2830 (1995)
2. Davis, P.J.: Circulant Matrices, 2nd edn. Chelsea (1994)
3. Ishii, H., Yamamoto, Y., Francis, B.A.: Sample-rate conversion via sampled-data H^∞ control. In: Proc. 38th IEEE Conf. Decision and Control, pp. 3440–3445 (1999)
4. Oppenheim, A.V., Schafer, R.W., Buck, J.R.: Discrete-Time Signal Processing, 2nd edn. Prentice-Hall, Englewood Cliffs (1999)
5. Pohlmann, K.C.: Principles of Digital Audio, 5th edn. McGraw-Hill, New York (2005)
6. Schreier, R., Temes, G.C.: Understanding Delta-Sigma Data Converters. Wiley, Chichester (2005)
7. Smith, S.W.: Digital Signal Processing: A Practical Guide for Engineers and Scientists. Newnes (2002)
8. Vaidyanathan, P.P.: Multirate Systems and Filter Banks. Prentice-Hall, Englewood Cliffs (1992)
9. Vaidyanathan, P.P., Mitra, S.K.: Polyphase networks, block digital filtering, LPTV systems, and alias-free QMF banks: a unified approach based on pseudocirculants. IEEE Trans. ASSP 36, 381–391 (1988)
10. Yamamoto, Y., Nagahara, M., Fujioka, H.: Multirate signal reconstruction and filter design via sampled-data H^∞ control. In: Proc. MTNS (2000)

Sparse Blind Source Separation via ℓ_1-Norm Optimization

Tryphon T. Georgiou and Allen Tannenbaum

Abstract. The title of the paper refers to an extension of the classical blind source separation where the mixing of unknown sources is assumed in the form of convolution with impulse response of unknown linear dynamics. A further key assumption of our approach is that source signals are considered to be sparse with respect to a known dictionary, and thereby, an ℓ_1-optimization is a natural formalism for solving the un-mixing problem. We demonstrate the effectiveness of the framework numerically.

1 Introduction

One of the most powerful tools in signal analysis which has been developed in recent years is a collection of techniques that allows sparse representations of signals. Fundamental theoretical contributions from a number of researchers [3, 4, 5, 6, 7, 8, 9, 29] has sparked this rapidly developing field which is driven by a wide spectrum of applications from robust statistics, data compression, compressed sensing, image processing, estimation, and high resolution signal analysis. The present work builds on the well-paved paradigm of sparse representations by focusing on a problem of system/source identification known as blind source separation.

Blind source separation (BSS) refers to the problem of separating sources from linear mixtures of these with unknown coefficients. For the special case where sources represent speech signals, the separation of voices corresponding to individual speakers is often referred to as the "cocktail party problem". Early work was based on the assumption that such signals are often statistically independent, and

Tryphon T. Georgiou
University of Minnesota, Minneapolis, MN, USA
e-mail: `tryphon@umn.edu`

Allen Tannenbaum
Georgia Institute of Technology, Atlanta, GA, USA, and Technion, Haifa, Israel
e-mail: `tannenba@ece.gatech.edu`

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 321–330.
springerlink.com

explored properties of second and higher order statistics. Typically, the required "unmixing matrix" was sought as a solution to a suitable optimization problem which either maximizes the distance from "Gaussianity" of the individual components or their statistical independence (e.g., see [10, 15, 16, 18, 26]). Thus, in the early work the mixing is always assumed algebraic as no dynamics of the intervening medium is taken into account.

In light of advances in the aforementioned sparse representation theory, the idea of using prior information about the sources in the form of membership in a dictionary became an attractive alternative to the postulate of statistical independence and was proposed already by Zibulevsky and Pearlmutter [32] in 2001, and more recently by Li *et al.* [22] and others, using various combinations of well-studied tools (K-means, Bayesian formalism, etc.) in combination with ℓ_1-optimization. The purpose of our work is to deal in a similar manner with the mechanism of mixing and identify system dynamics based on suitable prior information encoded in a appropriate dictionary as well.

Thus, our approach is rather direct as it assumes that the observed signals are outputs of an unknown dynamical system driven by unknown inputs, albeit both inputs as well as the impulse responses of the underlying dynamics being sparse mixtures from known dictionaries. Thus, a salient feature of our formulation is that the "mixing" of signals has a structure inherited by linear dynamics and, the "mixing matrix" has entries that are Toeplitz matrices by themselves. In general, this is an underdetermined nonlinear problem with many possible solutions. Therefore, it is both natural and meaningful to seek, beside sparse representations of signals, small complexity of the intervening dynamics. The latter can again be expressed as sparsity of the impulse response with respect to its own dictionary.

We now summarize the remainder of this note. In Section 2, we describe the basic concepts underlying sparse representations and ℓ_1-optimization theory. Section 3 gives the mathematical formulation of the problem under consideration and some numerical methods for its solution. In Section 4, we describe the ℓ_1 approach to solving the convolutive blind source separation problem. Finally, in Section 5, we study an illustrative example to elucidate the ℓ_1 methodology for blind source separation.

2 Sparse Representations and ℓ_1-Optimization

In order to motivate our methodology, we will give here some of the relevant background on sparse representations using the ℓ_1 norm. Full details may be found in [2, 3, 4, 5, 6, 7, 8, 9, 29], and the references therein.

Consider an underdetermined problem

$$\mathbf{H}x = y \tag{1a}$$

where the vector x represents the model, $\mathbf{H}$ is a linear ill-posed operator and the vector y contains data obtained from measurement. One possible way to regularize the problem is by using a Tikhonov-like regularization scheme [19]. However, it is often natural to assume that the model x is a linear combination of a small number of possible vectors that are collected into a $n \times N$ matrix B, where typically $N >> n$ (referred to as an "over-complete" basis or dictionary), and thus

$$x = Bv.$$

Therefore, the model-complexity is quantified by the number of nonzero entries of v called the sparsity $\|v\|_0$ of v. Seeking a solution with minimal number of nonzero entries can also be thought of as a form of regularization. Despite what the notation may suggest, $\|v\|_0$ is not a norm. In fact, the problem of minimizing $\|v\|_0$ subject to (1a) is combinatorial in nature and practically infeasible. However, it has been recognized for some time and, in recent years has formed the basis of the powerful theory of compressed sensing, that the ℓ_1-norm $\|\cdot\|_{\ell_1}$ can be thought of as a (convex) relaxation of $\|\cdot\|_0$ and that minimizing $\|\cdot\|_{\ell_1}$ in practice as well as in theory, for many interesting cases, leads with overwhelming probability to sparse solutions.

In practice, a more natural problem includes measurement noise ε and that equality in (1a) is not exact, that is

$$\mathbf{H}x + \varepsilon = y. \tag{1b}$$

Accordingly, it has been suggested that such problems can be effectively treated by one of the following formulations [17, 25]:

(i) $v = \operatorname{argmin}\{\|v\|_1 \mid \|\mathbf{H}Bv - y\|_r \leq \tau\}$, known as ***Basis Pursuit Denoising***,
(ii) $v = \operatorname{argmin}\{\|\mathbf{H}Bv - y\|_r \mid \|v\|_1 \leq \sigma\}$, known as ***Least Absolute Shrinkage and Selection Operator (LASSO)***, and
(iii) $v = \operatorname{argmin}\{\mu\|v\|_1 + \|\mathbf{H}Bv - y\|_r^r\}$, known as ***Relaxed Basis Pursuit***,

with r typically taken as either 1 or 2. The parameters τ, σ, μ need to be chosen so that the solution does not over-fit the data. In practice, the optimal choice for σ, τ and μ may not be obvious, especially when the noise level is not well known in advance. The interesting feature of these solutions is that they yield a sparse v, that is they produce a v with very few non-zero entries and this has been explained and justified in a series of papers, see e.g., [6, 9, 25] and the references therein.

On the numerical side of things, although such problems where already addressed in the 70's, recent work has shown dramatic improvements. Currently, there are three approaches that seem most efficient:

(a) ***Methods based on shrinkage.*** Such methods (softly) threshold the solution at each iteration [12]. This may be considered to be an expectation maximization (EM) type algorithm that is very effective especially for image deconvolution problems.
(b) ***Interior point methods.*** This is a class of algorithms based on work of Karmarkar [20] for linear programming. The technique uses a self-concordant barrier function in order encode the convex set; see [1] and the references therein.
(c) ***Methods based on reformulation and projection.*** In these methods one sets $v = p - q$ where $p, q \geq 0$ (elementwise) and solves the corresponding optimization problem by a projection method [11]. We will give details about this method

below since this we have found to be most effective for the type of problems considered in this note.

In the compressed sensing literature the matrix $\mathbf{H}$ and the dictionary B are typically known. However, formulating blind source separation in a similar setting, one needs to estimate $\mathbf{H}$ as well as recover x [22, 32]. The well-posedness of such a problem, draws on additional sets of natural assumptions. As we explain below, we are interested in the convolutive blind source problem where it is natural to assume linear time invariance, and therefore a Toeplitz, block-Toeplitz (for multi-source/multi-sensor problems, e.g., see [13]), or circulant structure for $\mathbf{H}$.

3 Numerical Aspects

Below we will see that the problem we are considering amounts to the following: Let M_i be positive semi-definite matrices and $\xi \in \mathbf{R}^n$. Then we want to find the minimum under the constraint that all components of ξ be non-negative of

$$\sum |\xi^T M_i \xi - y_i| + \sum \xi_j. \tag{2}$$

This is a special case of the following problem: Compute for $f : \mathbb{R}^n \to \mathbb{R}$ continuously differentiable

$$\min\{f(\xi) : \xi \in \Omega\},\ \ \Omega = \{\xi : a_i \leq \langle u_i, \xi \rangle \leq b_i,\ i = 1, \ldots, N\},$$

that is minimize f on a linearly constrained set. In [23], it is shown that a version of Newton's method, *truncated Newton's* will give fast convergence to the global minimum. We should note that conjugate gradients are perhaps the most used ones in the Newton approach. A typical optimization procedure has two layers of iterations: at each outer iteration an inner conjugate gradient procedure finds the Newton direction. But in many practical situations such conjugate gradient methods suffer from lengthy iterations in certain situations. Thus the author of [23], propose a ”trust region” version of the standard Newton optimization approach. We briefly describe this idea here.

Briefly, there are two basic optimization strategies. In the *line search* method, one chooses a direction and searches along this direction from the current choice $\xi^{(k)}$ for the next choice $\xi^{(k+1)}$ with a lower value. The distance to move along the direction $u^{(k)}$ is found by solving a one-dimensional problem of the form $\min_{\lambda > 0} f(\xi^{(k)} + \lambda u^{(k)})$. In line search, one just tries to approximate this minimum and works well for smaller problems.

The second technique is known as *trust region*, the information known about f is employed to construct a model function $\hat{f}_k$ whose behavior near $\xi^{(k)}$ is similar to that of f. Since $\hat{f}_k$ may not be very useful (i.e., a reasonable approximation) of f far from $\xi^{(k)}$, we only look for a minimizer of $\hat{f}_k$ in some “trust region” around $\xi^{(k)}$. This type approach employed by [23] for the type of ℓ_1 optimization problem considered in this paper has been found to be very effective.

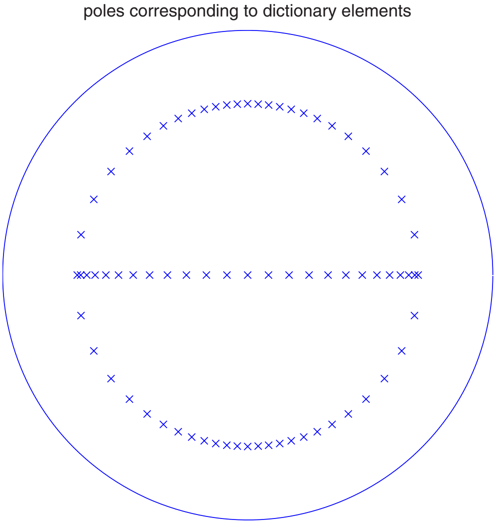


Fig. 1 Pole locations marked by a × inside the unit circle

4 BSS via ℓ_1-Minimization

We now explain our approach to identify two signals x_k and h_k, $k = 0, 1, \ldots, n-1$, from knowing only their convolution

$$y_k = h_k \star x_k := \sum_{\ell=0}^{n-1} h_\ell x_{k-\ell}, \text{ for } k = 0, 1, \ldots, n-1, \tag{3}$$

by assuming a prior in the form of membership of both x_k and h_k in suitable dictionaries. For the purposes of the present work we assume that the signals are scalar, with x_k representing the input to an unknown dynamical system with impulse response h_k. We further assume that the state of this unknown system starts at a zero initial condition. Either assumption can be relaxed by suitable reformulation of the problem along the lines of [13].

Clearly, the problem of finding x_k and y_k from (3) is both underdetermined and non-convex. The requirement that the two signals are sparsely represented with respect to dictionaries B_x and B_h can be "encouraged" by requiring that the ℓ_1 norm of respective selection vectors v_x and v_h is penalized accordingly. Thus, we propose the following optimization problem for this purpose:

$$(v_x, v_h) := \operatorname{argmin}\{\|v_x\|_1 + \|v_h\|_1 + \|y - Hx\|_r \mid x = B_x v_x, h = B_h v_h, \text{ and } h_0 = 1\} \tag{4}$$

where $r \in \{1, 2\}$, $x := (x_0, x_1, \ldots, x_{n-1})^T$, and y, h are defined similarly, B_x and B_h are known $n \times N_x$ and $n \times N_x$ matrices (i.e., suitably defined dictionaries), and

$$\mathbf{H} = \begin{bmatrix} h(0) & 0 & \dots & 0 \\ h(1) & h(0) & \dots & 0 \\ \vdots & \vdots & & \vdots \\ h(n-1) & h(n-2) & \dots & h(0) \end{bmatrix}.$$

The optimization in (4) can be cast in the form of (2) for $r = 1$ by expressing each selection vector v as a difference $v = v_+ - v_-$ of vectors with positive entries v_+, v_-. Then the part of the cost in $\|y - Hx\|_1$ can be expressed as a sum of absolute values of quadratic expressions in $\xi := (v_{x+},\, v_{x-},\, v_{h_+},\, v_{h-})^T$ as indicated in (2).

Invariably, the performance of such mathematical tools, where the solution of an optimization problem provides a possible explanation of the data, depends on the choice and relative importance place on various terms. For instance, an added weight that accentuates the contribution of the entries of v_x has as effect to improve the sparsity of the relevant signal, i.e., in this case x with respect to B_x. A multi-variable version of this problem, where x_k may represent many sources, and similarly for h_k, and thus it is vectorial, can be also cast in the same framework (see [13]). Further, additional terms in the functional can be used to "encourage" independence between the various sources in the spirit of Independent Component Analysis, or to "encourage" or "discourage" the various sources from sharing the same elements from the respective dictionaries (cf. [13]). This circle of possibilities will be explored further in future work. The main statement of the present note is that the tools of the ℓ_1/compressed sensing theory can be directly applied

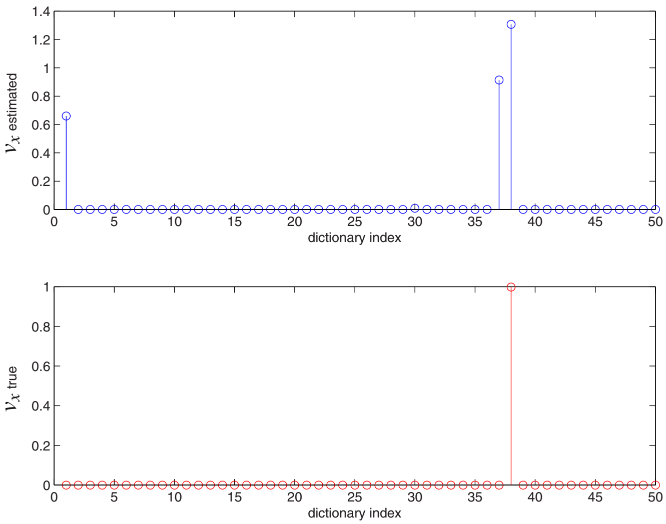


Fig. 2 Atoms v_x of the input ("true" and "estimated")

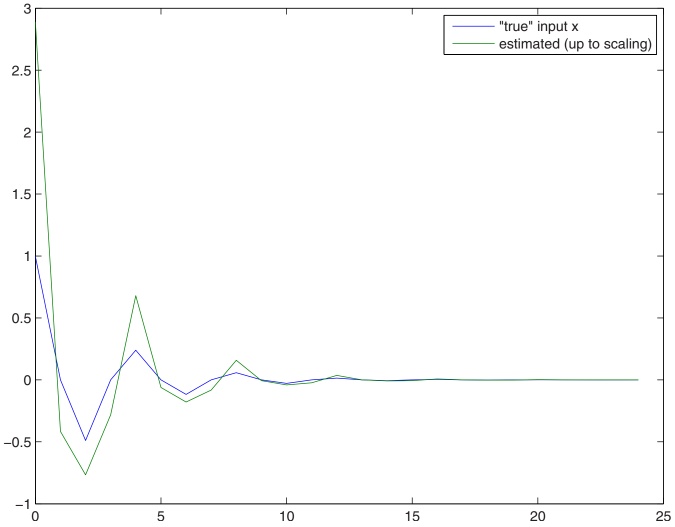


Fig. 3 Comparison of “true” vs. estimated inputs

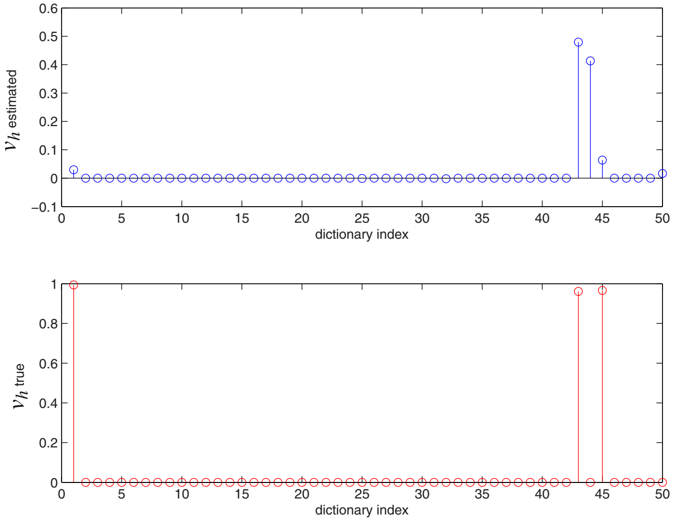


Fig. 4 Atoms v_h of the impulse response (“true” and “estimated”)

to the blind source deconvolution problem and that with relatively straightforward optimization one can obtain reasonably consistent results. This will be highlighted in the following example.

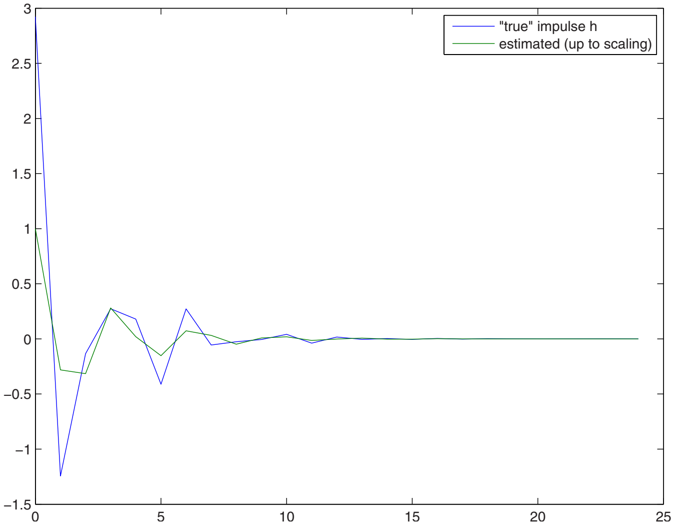


Fig. 5 Comparison of “true” vs. estimated impulse responses

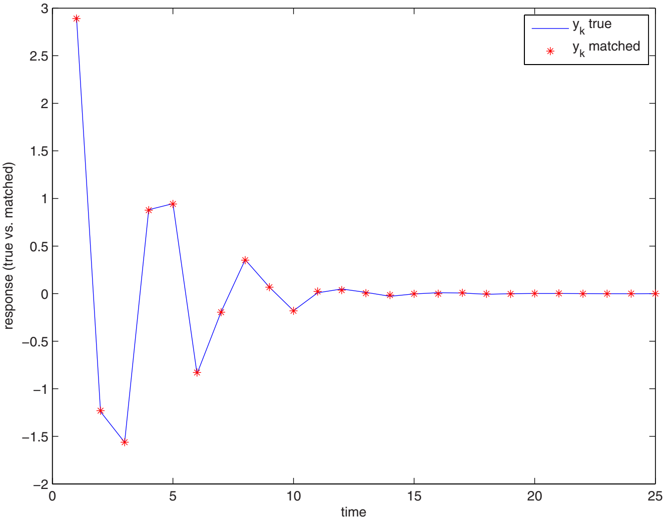


Fig. 6 Matching of the observed output

5 Example

In our example we have taken both directories B_x and B_h to be identical, containing first and second-order responses of systems with poles distributed as shown in Figure 1. A randomly selected pair of sparse vectors $\hat{v}_x$, $\hat{v}_h$ with one and three atoms, respectively, are shown in Figures 2 and 4 (in the lower subplots, respectively). The same figures, in the upper subplots, display the selection vectors v_x, v_h that minimize the functional in (4). In general, these can only be guaranteed to correspond to a local minimum. The estimated pairs of input/impulse-response are shown in Figures 3, 5, respectively. Figure 6 shows the matching between the original output y (data) and the one corresponding to the minimizing choice for the functional in (4).

Due to space restrictions we have only presented one example. It is apparent that, because the problem of blind source separation is inherently ill-posed, alternative numerical solutions are possible. This is also the case for our particular regularization of the problem, as it is hard to ensure a global minimum of (4). Indeed, as can be seen from the example, modes of input and impulse response may "affect/nudge" each other, and at times (as we have observed in other examples) may be swapped to produce a nearby local minimum. This is natural since the perceived effect at the output may often be the same in such situations. Also, the scaling of x and h is inherently uncertain since a multiplicative factor can be traded without affecting their convolution, hence the normalization $h_0 = 1$ in (4). In all instances, the estimated input/impulse-response pair reproduces quite accurately the given data y and the two are relatively sparse with respect to the corresponding dictionaries, as expected.

Acknowledgements. This work was supported in part by grants from NSF, AFOSR, ARO, as well as by NIH (NAC P41 RR-13218) through Brigham and Women's Hospital. This work is part of the National Alliance for Medical Image Computing (NAMIC), funded by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant U54 EB005149. Information on the National Centers for Biomedical Computing can be obtained from: http://nihroadmap.nih.gov/bioinformatics.

References

1. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
2. Bruckstein, A.M., Donoho, D.L., Elad, M.: From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Review 51, 34–81 (2009)
3. Candes, E., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Transactions on Information Theory 52, 489–509 (2006)
4. Candes, E., Tao, T.: Decoding by linear programming. IEEE Transactions on Information Theory 51, 4203–4215 (2005)
5. Candes, E., Tao, T.: Near optimal signal recovery from random projections: universal encoding strategies. IEEE Transactions on Information Theory 52, 5406–5425 (2006)
6. Candes, E., Braun, N., Wakin, M.: Sparse signal and image recovery from compressive samples. In: 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 976–979 (2007)
7. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM Journal on Scientific Computing 20(1), 33–61 (1998)
8. Donoho, D.L.: For most large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution. Communications on Pure and Applied Mathematics 59(6), 797–829 (2006)

9. Donoho, D.L., Elad, M.: Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization. Proc. Nat. Aca. Sci. 100(5), 2197–2202 (2003)
10. Duda, R., Hart, R., Stork, D.: Pattern Recognition. Willey, New York (2001)
11. Figueiredo, M.A.T., Nowak, R.D., Wright, S.J.: Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. IEEE Journal of Selected Topics in Signal Processing 1, 586–597 (2007)
12. Figueiredo, M., Nowak, R.: An EM algorithm for wavelet-based image restoration. IEEE Trans. Image Processing 12, 906–916 (2003)
13. Georgiou, T.T., Tannenbaum, A.: Sparse blind source deconvolution with application to high resolution frequency analysis. In: Three Decades of Progress in Systems and Control. Springer, New York (to appear, 2009)
14. Haber, E., Ascher, U.M., Oldenburg, D.: On optimization techniques for solving nonlinear inverse problems. Inverse Problems 16, 1263–1280 (2000)
15. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco (2001)
16. Haykin, S.: Blind Deconvolution. Prentice Hall, Englewood Cliffs (1994)
17. Hyvarinen, A., Hoyer, P., Oja, E.: Sparse code shrinkage: denoising of nongaussian data by maximum likelihood estimation. Neural Computation 11(7), 1739–1768 (1999)
18. Hyvarinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley and Sons, Chichester (2001)
19. Johansen, T.A.: On Tikhonov regularization, bias and variance in nonlinear system identification. Automatica 33(3), 441–446 (1997)
20. Karmarkar, N.: A new polynomial-time algorithm for linear programming. Combinatorica 4, 373–395 (1984)
21. Koh, K., Kim, S.J., Boyd, S.: An interior-point method for large-scale ℓ_1-regularized logistic regression. J. Mach. Learn. Res. 8, 1519–1555 (2007)
22. Li, Y., Cichocki, P., Amari, S.: Sparse component analysis and blind source separation of underdetermined mixtures. Neural Computation 16, 1193–1234 (2004)
23. Lin, C., More, J.: Newton's method for large bound constrained optimization problems. SIAM J. Optimization 9, 1100–1127 (1999)
24. Malioutov, D.M., Cetin, M., Willsky, A.S.: Optimal sparse representations in general overcomplete base. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (2004)
25. Mallat, S., Zhang, Z.: Matching pursuits with time-frequency dictionaries. IEEE Transactions on Signal Processing 41(12), 3397–3415 (1993)
26. Michailovich, O., Tannenbaum, A.: Blind deconvolution of medical ultrasound images: parametric inverse filtering approach. IEEE Trans. Image Process. 16, 3005–3019 (2007)
27. Nocedal, J., Wright, S.: Numerical Optimization. Springer, New York (1999)
28. Romberg, J.K.: Sparse signal recovery via ℓ_1 minimization. In: Proceedings of the 40th Annual Conference on Information Sciences and Systems, March 2006, pp. 213–215 (2006)
29. Tsaig, Y., Donoho, D.L.: Breakdown of equivalence between the minimal ℓ^1-norm solution and the sparsest solution. Signal Processing 86(3), 533–548 (2006)
30. Tuzlukov, V.: Signal Detection Theory. Springer, New York (2001)
31. Wang, J., Sacchi, M.D.: High-resolution wave-equation amplitude-variation-withray-parameter imaging with sparseness constraints. Geophysics 72(1), S11–S18 (2007)
32. Zibulevsky, M., Pearlmutter, B.A.: Blind source separation by sparse decomposition in a signal dictionary. Neural Computation 13, 863–882 (2001)

YY Filter — A Paradigm of Digital Signal Processing

Masaaki Nagahara

Abstract. YY filter, named after the founder Prof. Yutaka Yamamoto, is a digital filter designed by sampled-data control theory, which can optimize the analog performance of the signal processing system with AD/DA converters. This article discusses problems in conventional signal processing and introduces advantages of the YY filter.

Keywords: YY filter, sampled-data control, digital signal processing.

1 Introduction

YY filter is named after Prof. Yutaka Yamamoto, who is the founder of the modern sampled-data control theory. Before introducing the filter, I would like to write about him.

Prof. Yutaka Yamamoto has published a textbook on mathematics [21] in 1998. In that year, I was an undergraduate student in Kobe University, and I started studying control theory. I bought the book at that time, and found it very attractive. Affected by his book, I desired to be supervised by Prof. Yutaka Yamamoto in Kyoto University. I then luckily entered the university, and I began to study as a graduate student. I have studied sampled-data control and its application to digital signal processing. This study has started by Khargonekar and Yamamoto [6], which Prof. Yamamoto has been energetically addressing. Under his supervision, I finished my doctoral thesis titled "*Multirate Digital Signal Processing via Sampled-Data H^∞ Optimization*," [8] in 2003. This study has been of capital interest to me. I now begin the introduction of this study, YY filters.

In signal processing, signal reconstruction is a fundamental problem. For this problem, Shannon sampling theorem [14, 15] is widely used. This theorem is based

Masaaki Nagahara
Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan
e-mail: nagahara@ieee.org

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 331–340.
springerlink.com

on the assumption that the analog signal to be reconstructed is fully band-limited up to the Nyquist frequency. This assumption is however not realistic, since no real analog signals are fully band-limited. To such problems, sampled-data control theory has been applied in [6]. This is the first article of YY filter, which solves the delayed signal reconstruction problem. Based on this study, many researches have been made: multirate signal reconstruction [24], wavelet expansion [4] audio signal compression [1], fractional delay filters [10], image processing [3], adaptive filtering [11], probability density estimation [9], and repetitive control [12].

In this article, I omit discussion on these applications as space is limited, and I will concentrate on problems in Shannon's theorem (or its generalization) and advantages of the YY filter over the conventional theorem.

2 Problems in Sampling Theorem

2.1 *Shannon Sampling Theorem*

Let x be a continuous-time signal in L^2, the Lebesgue spaces consisting of the square integrable real functions on $\mathbb{R} = (-\infty, \infty)$. The problem here is to recover the original signal x from its sampled data $\{x(nh)\}_{n\in\mathbb{Z}}$, where $h > 0$ is the sampling period. This problem is however ill-posed unless there is an a priori condition on the original signal x. The sampling theorem, usually attributed to Shannon, answers this question under the hypothesis of band-limited signals [14, 15]. That is, it is assumed that the support of the Fourier transform $\hat{x}(\mathrm{j}\omega)$ of x is limited to the frequency range lower than the Nyquist frequency π/h:

Theorem 1 (Whittaker-Shannon). *Suppose that $x \in L^2$ is fully band-limited, i.e.,*

$$x \in BL^2 := \left\{x \in L^2 : \hat{x}(\mathrm{j}\omega) = 0, |\omega| \geq \pi/h\right\}. \tag{1}$$

Then the following formula uniquely determines x:

$$x(t) = \sum_{n=-\infty}^{\infty} x(nh)\phi(t - nh), \quad t \in \mathbb{R}, \tag{2}$$

where $\phi(t) := \mathrm{sinc}(t/h) := \sin(\pi t/h)/(\pi t/h)$.

The reconstruction procedure is shown in Fig. 1. In this figure, the signal $w \in L^2$ is convoluted (or filtered) by ϕ, i.e.,

$$x(t) = \int_{-\infty}^{\infty} \phi(t-\tau)w(\tau)\mathrm{d}\tau = (\phi * w)(t).$$

Then the signal x is in BL^2 (i.e., band-limited) since $\hat{\phi}(\mathrm{j}\omega) = 1$ if $\omega \in (-\pi, \pi)$ and $\hat{\phi}(\mathrm{j}\omega) = 0$ if $\omega \notin (-\pi, \pi)$. The signal x is sampled by the ideal sampler $\mathscr{S}$ with the sampling period h:

$$(\mathscr{S}x)[n] := x(nh), \quad n \in \mathbb{Z}.$$

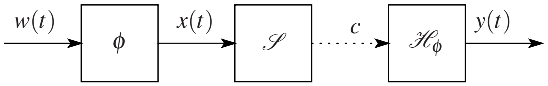


Fig. 1 Shannon sampling theorem; the signal $w \in L^2$ is band-limited by ϕ and sampled by the ideal sampler $\mathscr{S}$. Then an analog signal y is produced by the hold device $\mathscr{H}_\phi$ to reconstruct x.

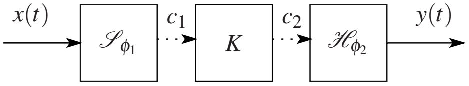


Fig. 2 Generalized sampling theorem; the signal $x \in V(\phi_2)$ is sampled by the generalized sampler $\mathscr{S}_{\phi_1}$ to become the discrete-time signal c_1. Then c_1 is filtered by K to become c_2. Finally, an analog signal y is produced by the hold device $\mathscr{H}_\phi$ to reconstruct x.

Then the discrete-time signal $c = \mathscr{S}x$ becomes an analog signal y by the hold device $\mathscr{H}_\phi$:

$$(\mathscr{H}_\phi c)(t) := \sum_{n=-\infty}^{\infty} c[n]\phi(t-nh).$$

By the sampling theorem, the reconstructed signal y is exactly equal to x (not w).

Shannon sampling theorem is a beautiful result and is the fundamental theory for the conventional digital signal processing. However we can find the following questions in real applications:

- The band-limiting assumption (1) does not hold for real signals such as audio, image, or video signals.
- The reconstruction formula (2) is hard to implement on a real device, since the sinc function has infinite support, in particular it is not causal.

2.2 Generalized Sampling Theorem

The sampling theory mentioned above has been extended to more general case [15, 17], that is, the function ϕ is not necessarily a sinc function, and the sampler is a generalized sampler $\mathscr{S}_{\phi_1}$ defined by

$$(\mathscr{S}_{\phi_1}x)[n] := \int_{-\infty}^{\infty} \phi_1(nh-\tau)x(\tau)\mathrm{d}\tau = \langle x, \phi_1(\cdot - nh)\rangle, \quad n \in \mathbb{Z}.$$

In this definition, we have $\mathscr{S}_{\phi_1}x = \mathscr{S}(\phi_1 * x)$, and hence the function ϕ_1 is considered as the impulse response of the acquisition device. Fig. 2 shows a generalized situation. In this figure, the analog input x is sampled by the generalized sampler $\mathscr{S}_{\phi_1}$. Then the sampled signal c_1 is filtered by a digital filter K, and then an analog signal y is obtained by the hold device $\mathscr{H}_{\phi_2}$. In this setting, a generalized sampling theorem is proposed by [17]. The idea is the notion of *consistency*: the output y in Fig. 2 can be perfectly reconstructed by the same system, that is, for all $n \in \mathbb{Z}$,

$$\langle x, \phi_1(\cdot - nh)\rangle = \langle y, \phi_1(\cdot - nh)\rangle.$$

This implies that the reconstruction system works as a projector. To achieve consistency, the optimal filter K which is linear and time-invariant (LTI) is constructed by the *oblique projection* of x onto $V(\phi_2)$ perpendicular to $V(\phi_1)$, where $V(\phi_1)$ and $V(\phi_2)$ are closed subspaces in L^2, which is defined by

$$V(\phi_i) := \left\{ x = \sum_{n=-\infty}^{\infty} c[n]\phi_i(t - nh), c \in \ell^2 \right\}, \quad i = 1, 2.$$

By the oblique projection, the following generalized sampling theorem is obtained [17].

Theorem 2 (Unser and Aldroubi). *Suppose that $x \in V(\phi_2)$ and the filter*

$$A_{12}(z) = \sum_{n=-\infty}^{\infty} \langle \phi_1(\cdot - nh), \phi_2\rangle z^{-n}$$

is invertible[1]*. Then the following formula uniquely determines x:*

$$x(t) = \sum_{n=-\infty}^{\infty} (c_1 * k)[n]\phi_2(t - nh), \quad t \in \mathbb{R}$$

where $c_1 := \mathscr{S}_{\phi_1} x$ and k is the impulse response of $K(z) = A_{12}(z)^{-1}$.

The assumption $x \in V(\phi_2)$ can be interpreted as a *generalized band-limiting* condition. Then we again have a problem of non-band-limited inputs, that is, $x \notin V(\phi_2)$. In this case, the reconstructed signal y can have a large error [13]. Moreover, it is possible that the optimal filter will be unstable. This problem is discussed precisely in the next subsection.

2.3 Causality and Stability

In real-time systems, *causality* is a necessary condition for signal processing. For the sake of simplicity, we assume[2] that $\phi_1(t) = \phi_2(t) = 0$ if $t < 0$. Then the causality of the reconstruction system in Fig. 2 depends on the causality of the filter K. If the impulse response $\{k[n]\}$ of the filter K satisfies $k[n] = 0$, $n < 0$, then the reconstruction system is causal. However, in many cases, the filter K may be non-causal, for example, in the case of polynomial splines [18] and exponential splines [16, 19]. This is because the filter $K(z)$ has poles outside of the unit circle in $\mathbb{C}$ [13, 18].

[1] This means that $A_{12}(z)$ has no zeros on the unit circle in $\mathbb{C}$.

[2] If the assumption does not hold, the reconstruction system can be non-causal.

In particular, it is shown [13] that high order exponential splines can produce filters with poles outside the unit circle provided that the sampling time is sufficiently small. Therefore, if the non-causal filter is realized as a causal one, the poles outside the unit circle lead to an *unstable* filter.

2.4 Summary

The problems in (generalized) sampling theorem discussed above are the followings.

Problems in (generalized) sampling theorem

1. If the input signal x is not (generalized) band limited, the reconstructed signal can show a large error. In other words, the reconstruction is *not robust* against uncertainty of input signals.
2. The reconstruction system can be non causal.
3. The causal realization of the reconstruction filter K can be unstable.

3 Sampled-Data H^∞ Optimal Signal Reconstruction — YY Filter

As we see in the previous section, (generalized) sampling theorem has three problems: robustness, causality and stability. In this section, we introduce a new signal processing, *sampled-data signal processing*, or *YY filter*, which is based on sampled-data control theory.

3.1 Problem Formulation

The main reason to adopt sampled-data control theory is that we can design a digital filter which optimizes the *intersample behavior*. In other words, we can minimize the reconstruction error for non-band-limited signals. Moreover, we adopt the H^∞ *performance index* for this optimization. By H^∞ optimization, we can gain the robustness against the input uncertainty.

In the sampling theorem, the optimal reconstruction is a projector on a subspace in L^2, in which for every input the error is minimized in L^2 sense. This means that the error depends on the input and there can be an input for which the error can be arbitrary large. On the other hand, H^∞ optimization is an optimization for the *worst case*, by which we can guarantee an error level for *any* inputs. This leads to the robustness against the input uncertainty.

It is obvious that there is no optimal filter K which minimizes the error for *all* signals in L^2, or the optimal filter can be $K = 0$. To reconstruct or interpolate the

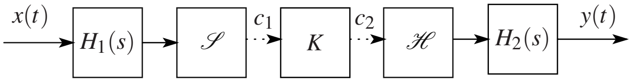


Fig. 3 Signal Reconstruction; the signal $x \in L^2$ is filtered by an analog filter $H_1(s)$ and sampled by the ideal sampler $\mathscr{S}$. Then an analog signal y is produced by the zero-order hold $\mathscr{H}$ and an analog filter $H_2(s)$.

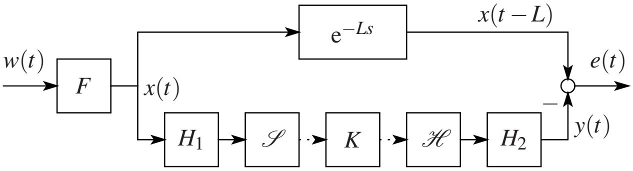


Fig. 4 Error system $\mathscr{E}(K)$

intersample data, we should assume some *a priori* information for the inputs. Therefore, we assume that the inputs are in the following subspace in L^2,

$$FL^2 := \{Fw : w \in L^2(\mathbb{R}_+)\}$$

where F is an analog filter which is stable and strictly causal, and $L^2(\mathbb{R}_+)$ is the Lebesgue spaces consisting of the square integrable real functions on $\mathbb{R}_+ = [0,\infty)$. The filter F is an analog model of the input signals. The space $L^2(\mathbb{R}_+)$ is a subspace of L^2, by which we can take account of causality and stability of the reconstruction system. Our signal subspace FL^2 is in a sense larger than BL^2 or $V(\phi_2)$ because every signal in BL^2 or $V(\phi_2)$ can be expanded by $\{\mathrm{sinc}(t-nT)\}$ or $\{\phi_2(t-nT)\}$, on the other hand, FL^2 needs $\{\phi(2^{-m}(t-nT))\}$ for some ϕ (wavelet expansion [20]). In other words, a signal in FL^2 can contain arbitrary high frequency components, the decay rate of which is governed by the filter F.

To optimize for the worst case, we consider the following performance index:

$$J(K) = \sup_{\substack{x \in FL^2 \\ x \neq 0}} \frac{\left\| \left(e^{-Ls} - H_2 \mathscr{H} K \mathscr{S} H_1\right) x \right\|_{L^2(\mathbb{R}_+)}}{\|x\|_{L^2(\mathbb{R}_+)}}. \tag{3}$$

This is equivalent to the H^∞ norm of the sampled-data error system

$$\mathscr{E}(K) := (e^{-Ls} - H_2 \mathscr{H} K \mathscr{S} H_1)F. \tag{4}$$

The block diagram of this error system is shown in Fig. 4.

3.2 Computation of YY Filter

The optimal filter K_{opt} which minimizes $J(K)$ in (3) can be obtained by numerical computation. To compute the optimal filter K_{opt}, we discretize the sampled-data

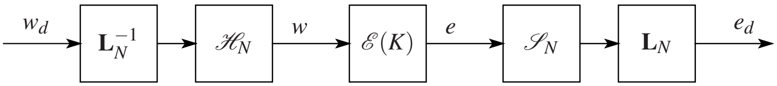


Fig. 5 Fast discretization of the sampled-data system $\mathscr{E}(K)$: $\mathbf{L}_N$ is the blocking operator, $\mathscr{S}_N$ and $\mathscr{H}_N$ are respectively the fast sampler and the fast hold with sampling period h/N.

error system $\mathscr{E}(K)$ in (4) by approximation [5, 23] or H^∞ discretization [7]. We here discuss the approximation technique for minimizing $J(K)$ in (3). We first introduce fast sampling and fast hold. Let $\mathscr{S}_N$ and $\mathscr{H}_N$ are respectively the ideal sampler and the zero-order hold with period h/N, where N is a positive integer ($N \geq 2$). Then the system $\mathscr{S}_N\mathscr{E}(K)\mathscr{H}_N$ becomes a discrete-time multi-rate system with sampling periods h and h/N. Then we introduce the *blocking operator* $\mathbf{L}_N$, or the *discrete-time lifting operator* [2, 8]:

$$\mathbf{L}_N : \{v[0], v[1], v[2], \dots\} \mapsto \left\{ \begin{bmatrix} v[0] \\ v[1] \\ \vdots \\ v[N-1] \end{bmatrix}, \begin{bmatrix} v[N] \\ v[N+1] \\ \vdots \\ v[2N-1] \end{bmatrix}, \dots \right\}.$$

This operator converts a 1-dimensional signal v into an N-dimensional signal and the sampling rate becomes N times slower. This operation makes it possible to equivalently convert multirate systems into single-rate ones, and hence the analysis and design become easier. By using this operator, the system $E_N(K)$ defined by

$$E_N(K) := \mathbf{L}_N\mathscr{S}_N\mathscr{E}(K)\mathscr{H}_N\mathbf{L}_N^{-1} \tag{5}$$

becomes a discrete-time LTI system. Moreover, we can say that for any integer $N \geq 2$ and any stable K, there exist discrete-time LTI systems $G_{1,N}$, $G_{2,N}$ and $G_{3,N}$ such that [8]

$$E_N(K) = G_{1,N} + G_{2,N}KG_{3,N},$$

and the LTI system $E_N(K)$ is approximation of $\mathscr{E}(K)$ in the sense that [23]

$$\lim_{N\to\infty} \|E_N(K)\|_\infty \to J(K) = \|\mathscr{E}(K)\|_\infty.$$

The optimization of minimizing $E_N(K)$ is easily done by using discrete-time H^∞ optimization technique. We can therefore obtain a stable and causal filter K which approximates the optimal filter K_{opt}.

3.3 Robustness

Next let us consider robustness against uncertainty of the analog signal model $F(s)$. In practice, $F(s)$ cannot be identified exactly. We therefore partially circumvent this

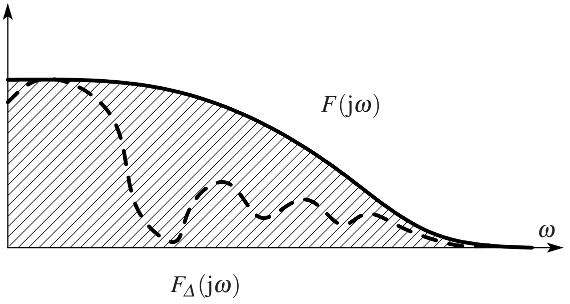


Fig. 6 Nominal filter $F(s)$ (solid) and perturbed $F_\Delta(s)$ (dash)

defect by discussing the robustness of the filter against uncertainty of $F(s)$. Let us assume the unstructured uncertainty of the following type:

$$F_\Delta(s) := F(s)(1+\Delta(s)), \quad \mathscr{E}^\Delta(K) := \left(\mathrm{e}^{-Ls} - H_2\mathscr{H}K\mathscr{S}H_1\right)F_\Delta,$$
$$\Delta \in \boldsymbol{\Delta} := \{\Delta : \|1+\Delta\|_\infty \leq \gamma\}.$$

Then we have the following proposition:

Proposition 1. *For any stable K and* $\Delta \in \boldsymbol{\Delta}$, *we have* $\|\mathscr{E}^\Delta(K)\|_\infty \leq \gamma\|\mathscr{E}(K)\|_\infty$.

By this proposition, the nominal performance $\|\mathscr{E}(K)\|_\infty$ is guaranteed against the perturbation $\Delta \in \boldsymbol{\Delta}$ if $\gamma \leq 1$. In some cases, it is possible that $\gamma = 1$, in which case the performance is bounded as illustrated in Fig. 6. This means that if we take $F(s)$ that covers all possible gain characteristics of the input analog signals, it gives a bound for the error norm. This at least partially justifies the choice of the first-order weighting $F(s)$ in the previous section.

3.4 FIR YY Filter by LMI

The error system (4) or (5 is *affine* in the filter K to be designed. By this fact, we can design the optimal FIR (finite impulse response) filter of the form

$$K(z) = \sum_{n=0}^{N} a_n z^{-n}.$$

By this, the error system (5) is affine in the design parameter $a_0, a_1, \ldots, a_N$. It follows that the optimization of minimizing $\|E_N(K)\|_\infty$ can be described by an LMI (linear matrix inequality) by the bounded real lemma or Kalman-Yakubovic-Popov lemma [22]. The optimization with an LMI can be solved easily by computer softwares.

3.5 Summary

The advantages of the YY filter discussed in this section are the followings.

Advantages of YY filter

1. the optimal filter is always causal and stable.
2. the design takes the inter-sample behavior into account.
3. the system is robust against the uncertainty of input signals.
4. the optimal FIR filter is also obtainable via an LMI.

4 Conclusions

In this article, problems in Shannon's theorem have been pointed out and the advantages of YY filter over the conventional signal processing have been introduced. In fact, YY filters are implemented in commercial MD players, silicon-audio devices, and mobile phones. One of future works is design of adaptive YY filters.

References

1. Ashida, S., Kakemizu, H., Nagahara, M., Yamamoto, Y.: Sampled-data audio signal compression with Huffman coding. In: Proc. SICE Annual Conf. 2004, pp. 972–976 (2004)
2. Chen, T., Francis, B.A.: Optimal Sampled-Data Control Systems. Springer, Heidelberg (1995)
3. Kakemizu, H., Nagahara, M., Kobayashi, A., Yamamoto, Y.: Noise reduction of JPEG images by sampled-data H^∞ optimal ε filters. In: Proc. SICE Annual Conf. 2005, pp. 1080–1085 (2005)
4. Kashima, K., Yamamoto, Y., Nagahara, M.: Optimal wavelet expansion via sampled-data control theory. IEEE Signal Process. Lett. 11, 79–82 (2004)
5. Keller, J.P., Anderson, B.D.O.: A new approach to the discretization of continuous-time controllers. IEEE Trans. Autom. Control 37, 214–223 (1992)
6. Khargonekar, P.P., Yamamoto, Y.: Delayed signal reconstruction using sampled-data control. In: Proc. 35th IEEE CDC, pp. 1259–1263 (1995)
7. Mirkin, L., Tadmor, G.: Yet another H^∞ discretization. IEEE Trans. Autom. Control 48, 891–894 (2003)
8. Nagahara, M.: Multirate digital signal processing via sampled-data H$^{?8?}$ optimization. Ph.D. thesis, Kyoto University (2003)
9. Nagahara, M., Sato, K.I., Yamamoto, Y.: H^∞ optimal nonparametric density estimation from quantized samples. In: Proc. 40th ISCIE SSS (2008)
10. Nagahara, M., Yamamoto, Y.: Optimal design of fractional delay filters. In: Proc. 42nd IEEE CDC, pp. 6539–6544 (2003)
11. Nagahara, M., Yamamoto, Y.: Hybrid design of filtered-x adaptive algorithm via sampled-data control theory. In: Proc. 2008 IEEE ICASSP, pp. 353–356 (2008)

12. Nagahara, M., Yamamoto, Y.: Robust repetitive control by sampled-data H^∞ filters? To appear in Proc. 48th IEEE CDC (2009)
13. Nagahara, M., Yamamoto, Y., Khargonekar, P.P.: Stability of signal reconstruction filters via exponential splines. In: Proc. 17th IFAC World Congress, pp. 1414–1419 (2008)
14. Shannon, C.E.: Communication in the presence of noise. Proc. IRE 37(1), 10–21 (1949)
15. Unser, M.: Sampling — 50 years after Shannon. Proc. IEEE 88(4), 569–587 (2000)
16. Unser, M.: Cardinal exponential splines: part II — think analog, act digital. IEEE Trans. Signal Process. 53(4), 1439–1449 (2005)
17. Unser, M., Aldroubi, A.: A general sampling theory for nonideal acquisition devices. IEEE Trans. Signal Process. 42(11), 2915–2925 (1994)
18. Unser, M., Aldroubi, A., Eden, M.: B-spline signal processing: part II — efficient design and applications. IEEE Trans. Signal Process. 41(2), 834–848 (1993)
19. Unser, M., Blu, T.: Cardinal exponential splines: part I — theory and filtering algorithms. IEEE Trans. Signal Process. 53(4), 1425–1438 (2005)
20. Vetterli, M., Kovaçević, J.: Wavelets and Subband Coding. Prentice Hall, Englewood Cliffs (1995)
21. Yamamoto, Y.: Mathematics for Systems and Control. Asakura Publishing (1998)
22. 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)
23. Yamamoto, Y., Madievski, A.G., Anderson, B.D.O.: Approximation of frequency response for sampled-data control systems. Automatica 35(4), 729–734 (1999)
24. Yamamoto, Y., Nagahara, M., Fujioka, H.: Multirate signal reconstruction and filter design via sampled-data H^∞ control. In: Proc. 14th MTNS (2000)

How to Sample Linear Mechanical Systems

Mattia Bruschetta, Giorgio Picci, and Alessandro Saccon

Abstract. *Variational integrators* is a a new discretization technique of the equations of motion of a mechanical system introduced by Veselov and further developed by J. Marsden an co-workers, which is now widely used by numerical analysts working in various applied fields. This discretization technique, unlike the usual discretization procedures familiar in control, e.g. zero-order-hold, can lead to simple and well-conditioned transformation formulas for the recovery of the continuous time parameters from the discretized model. We discuss variational integrators for linear second order mechanical systems and show that physically meaningful properties of the continuous-time model, like passivity, are preserved. Variational integrator discretization is also shown to provide well-conditioned models for the identification of continuous-time second-orders systems starting from measured data.

1 Introduction and Problem Statement

We are interested in linear second order models of mechanical systems of the following form:

$$M\ddot{q} + D\dot{q} + Kq = f \tag{1}$$

where M and K, both symmetric positive definite matrices in $\mathbb{R}^{n\times n}$, have the interpretation of generalized mass (or inertia) and generalized stiffness coefficient matrices respectively, while $D \in \mathbb{R}^{n\times n}, D = D^\top, D \geq 0$ is a linear (viscous) damping coefficient. The generalized forces f acting on the system are in general not

Mattia Bruschetta and Giorgio Picci
Department of Information Engineering, University of Padova,
via Gradenigo 6/B, Padova, Italy
e-mail: {bruschet,picci}@dei.unipd.it

Alessandro Saccon
Instituto de Sistemas e Robtica, Instituto Superior Tecnico, Lisboa, Portugal
e-mail: asaccon@isr.ist.utl.pt

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 343–353.
springerlink.com

independent and can be expressed as a linear function of a vector of independently assignable generalized input forces u of dimension $k \leq n$; namely

$$f = Lu$$

where the matrix L, which will be assumed to be known, describes the physical locations at which the input forces u act on the system. Without loss of generality it may be assumed that L is of full column rank; i.e.

$$\operatorname{rank} L = k\,. \tag{2}$$

For simplicity we shall assume that the whole generalized configuration vector q is a measurable quantity. This assumption is equivalent to assuming that a "full set of sensors" is available; i.e. that all n degrees of freedom are measured via a linear sensor equation of the form $y = Cq$ where C is square invertible. System (1) can also be represented in state space form; defining $x := [q,\, \dot{q}]^\top$, one gets

$$\dot{x} = \begin{bmatrix} 0 & I \\ -M^{-1}K & -M^{-1}D \end{bmatrix} x + \begin{bmatrix} 0 \\ M^{-1}L \end{bmatrix} u \tag{3}$$

which should be coupled with the output equation $q = [I\; 0]x$. The special (Hamiltonian) structure of this realization leads to the inverse second-order polynomial transfer function of the model (1). Throughout the paper we shall assume that the system (1) with input u is controllable. See [10] for a direct test of controllability/observability of second order models of the type considered in this paper. Note that under these assumptions the system is automatically controllable and observable and hence minimal. This is a necessary condition for parameter identifiability

Now, system identification deals almost exclusively with discrete-time data and discrete time models. Nevertheless in several areas of engineering, and especially in mechanical or structural engineering, the estimation of physical parameters which pertain to the underlying (physical) *continuous time model* of the type (1) is very often required. A typical example is the estimation of the proper modes of vibration of a mechanical structure. The proper modes are the eigenvalues of a linear vector second order *continuous time* system; i.e. are solutions of an algebraic equation of the form :

$$\det\left(Ms^2 + Ds + K\right) = 0$$

and it is a fact that accurate information on these proper values and on the associated proper vectors may be hard to get from an estimated discretized system, no matter how accurate the estimates may be. The reasons for this difficulty may be described as follows.

Let the discrete time index k relate to a sampling period of length h and assume we fit, by some identification algorithm, measured sampled input-output data by a discrete-time state space model of the form

$$x(k+1) = Fx(k) + Gf(k)\,, \qquad q(k) = Hx(k) + Ju(k)\,. \tag{4}$$

If h is short enough so that the input function can be well approximated by a piecewise constant function, we can naturally imagine (4) to be related to an underlying continuous state space model (not necessarily the particular realization (3)) by the standard zero-order-hold (ZOH) sampler. We may then attempt to recover the original parameters, say the matrices A and B of a $2n$ dimensional continuous time model, from estimates of the parameters (F,G) of the discrete time model (4), by inverting the relations $F = \exp Ah$, $G = \int_0^h \exp As\, ds B$. This is what is implemented in the d2c routine in MATLAB. In certain circumstances this may however turn into a very ill-conditioned problem. In particular the recovery of matrix A from the estimated F involves the computation of the logarithm of F which may be a complex matrix or, for a large sampling period, be undefined as requiring the inversion of the exponential map in a region of the complex plane where it is not invertible. A common belief is that the problem should be solvable by choosing a suitably high sampling frequency, but actually it easy to see that, even in the trivial example of a scalar F subject to a perturbation δF, the relative error incurred when computing $A + \delta A := \frac{1}{h}\log(F + \delta F)$ is

$$\frac{\delta A}{A} = \frac{1}{\log F}\frac{\delta F}{F}$$

a similar formula holding in the matrix case, see [5]. Since for $h \to 0$ $F \to I$, the condition number of computing $A = \frac{1}{h}\log F$ tends to infinity when $h \to 0$. This means that when the sampling frequency is very high, the effect of unavoidable random errors on the estimates of F (and G) could be dramatically amplified in computing A by the logarithmic transformation. See [5] and the references therein. In any case, even with a clever use of anti-aliasing filters, oversampling is well-known to bring in noise in the estimates and deteriorate the identification process. Note in addition that standard discretization recipes like ZOH do not in general preserve physical properties of the underlying continuous system such as *passivity*, which may then be impossible to be recaptured when transforming the discrete to a continuous model. In particular the equivalence to second order representations of the type (1), which is indeed a characteristic of linear models of mechanical systems (Newton law) will in general be lost. In fact, a continuous state-space realization obtained by the d2c routine from a discrete model (4) identified say by standard subspace identification methods will never possess the Hamiltonian structure which is necessary for the input-output relation of the system to have the second-order form (1) and hence to allow for the recovery of the physical parameters M, D, K. That this is not of purely academic interest is witnessed by the interest on this problem in the recent mechanical engineering literature, see e.g. [4], [12] and the references therein.

Unfortunately the literature on continuous time identification, see e.g. [7, 15] does not seem to be of help. Continuous time black-box identification seems to be still in its infancy. Most algorithms turn eventually out to relay on the MATLAB d2c routine which computes the logarithm and, for the reasons given above, should be avoided. To our knowledge, serious concrete applications of any of continuous time identification methods to real world problems seem to be missing. There seem to be substantial progress still to be made in this area.

Ideally one should be able to discretize the system (1) or (3) in such a way that F and G depend in a simple explicit way on the original physical parameters (A, B) so as to make the discrete to continuous conversion simple and well-conditioned. A naive attempt in this direction is Euler discretization by which $F = I + Ah$ and $G = Bh$. This however either requires very small h or leads to too rough a discrete-time approximation to be of practical use in most cases.

It seems that in order to solve this problem one should find a discretization procedure leading to discrete models in which the properties and the physical meaning of the continuous system are maintained in the structure and in the matrices parametrizing the model. Ideally, a discretized model should correspond to a (generalized) *discrete-time Newton law* and the model parameters be interpretable accordingly, and possibly being simply related to the parameters of the original continuous time model.

How this program can actually be carried through is discussed in the conference paper [2] where some preliminary applications of this discretization to identification are discussed. The scope of this paper is to explore the Hamiltonian structure of the discretized model and its relevance in discussing preservation of passivity.

2 The Variational Integrators Approach to Discretization

A novel twist to the discretization problem has been provided by the theory of *variational integrators*, see [18], and the recent work of J. Marsden and co-workers, see [13]. These techniques seem to be fairly well known to numerical analysts working with mechanical models but not so familiar to the system and control community. The key idea is that the discrete equations of motion should not be derived by attempting a direct discretization of the equations (1) or (3) but rather derived by paraphrasing what happens in continuous time; i.e. by making stationary a *discrete action integral* defined in terms of a suitable *discrete Lagrangian function*. The (discrete) equations of motion should then follow just like the Euler-Lagrange equations in continuous time. In short, the variational integrators paradigm is to build from scratch a theory of Lagrangian *Discrete Mechanics*.

In the standard (continuous-time) approach to Lagrangian mechanics we are given a Lagrangian function $L(q(t), \dot{q}(t))$ and external forces $f_L(q(t), \dot{q}(t), t)$ and the equations of motion follow from so-called *Lagrange-d'Alembert principle* (equivalent in the conservative case to zeroing the variation of the action functional while holding the endpoints of the curve $q(t)$ fixed). This leads to (see e.g. [13, p. 421]) the well-known forced *Euler Lagrange equations*:

$$\frac{\partial L}{\partial q}(q,\dot{q}) - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}(q,\dot{q})\right) + f_L(t) = 0\,. \tag{5}$$

For a quadratic Lagrangian,

$$L(q(t),\dot{q}(t)) = \frac{1}{2}\dot{q}^\top M\dot{q} - \frac{1}{2}q^\top Kq \tag{6}$$

and an external force composed by a dissipation force $f_D = -D\dot{q}$ and the actual (generalized) external force $f(t)$:

$$f_L(t) := -D\dot{q}(t) + f(t), \tag{7}$$

one obtains a linear second order vector differential equation of the form (1).

In order to mimic this procedure in discrete time one may first consider a curve segment between two configuration points $q_0 = q(0)$ and $q_1 = q(h)$ in the configuration space $Q \subset \mathbb{R}^n$, placed h units of time apart. The *discrete Lagrangian* increment $L_d(q_0,q_1,h)$ must contribute to the action integral along the above curve segment. One defines the *exact (forced) discrete Lagrangian* and the *exact discrete forces* on that curve segment as:

$$L_d^E(q_0,q_1,h) = \int_0^h L(q(t),\dot{q}(t))dt \tag{8}$$

$$f_d^{E+}(q_0,q_1,h) = \int_0^h f_L(q(t),\dot{q}(t),t)\cdot\frac{\partial q(t)}{\partial q_1}dt \tag{9}$$

$$f_d^{E-}(q_0,q_1,h) = \int_0^h f_L(q(t),\dot{q}(t),t)\cdot\frac{\partial q(t)}{\partial q_0}dt \tag{10}$$

where $q : [0,h] \to Q$ is the solution of the forced Euler-Lagrange equation (5) with endpoint conditions $q(0) = q_0$ and $q(h) = q_1$. See [13, p. 427] for details. If it were possible to compute the integral (8) explicitly, we would have a discrete model that describes exactly the continuous dynamic at the discrete time instants $t = kh$. In general this computation is not possible and we need to use an approximation both for the discrete Lagrangian and for the discretized external forces. These approximations we denote $L_d(q_0,q_1)$, $f_d^+(q_0,q_1,k)$, $f_d^-(q_0,q_1,k)$ without superscripts. It is remarkable that although many approximations are possible, the "stationary action" principle leads in any case to *Discrete Euler Lagrange Equations* of a standard form

$$D_2L_d(q_{k-1},q_k) + D_1L_d(q_k,q_{k+1}) + f_d^+(q_{k-1},q_k,k) + f_d^-(q_k,q_{k+1},k+1) = 0 \tag{11}$$

where D_i stands for the partial derivative operator applied to the i-th argument of the function on which it is acting. The specific form of the approximations depend on the specific discretization rule used for approximating the integrals. A simple way to approximate the Lagrangian, is to use the so-called "midpoint rule":

$$q \simeq \frac{q_0+q_1}{2}, \qquad \dot{q} \simeq \frac{q_1-q_0}{h} \tag{12}$$

which for the quadratic Lagrangian (6) leads to:

$$L_d(q_k,q_{k+1}) = h[(\frac{q_{k+1}-q_k}{h})^\top \frac{M}{2}(\frac{q_{k+1}-q_k}{h}) - (\frac{q_{k+1}+q_k}{2})^\top \frac{K}{2}(\frac{q_{k+1}+q_k}{2})]\,. \tag{13}$$

As for the external forces (7), the midpoint rule discretization of the general exact expressions (9), (10), yields

$$f_d^+(q_{k-1},q_k,k) = -D\frac{q_k-q_{k-1}}{2} + \frac{h}{4}[f(h(k-1))+f(hk))]$$
$$f_d^-(q_k,q_{k+1},k+1) = -D\frac{q_{k+1}-q_k}{2} + \frac{h}{4}[f(hk)+f(h(k+1))]\,.$$

By putting together the above with

$$D_1L_d(q_k,q_{k+1}) = -M\frac{q_{k+1}-q_k}{h} - \frac{h}{2}K\frac{q_{k+1}+q_k}{2},$$
$$D_2L_d(q_{k-1},q_k) = M\frac{q_k-q_{k-1}}{h} - \frac{h}{2}K\frac{q_k+q_{k-1}}{2},$$

and rearranging the time index, we find the *forced discrete Euler Lagrange equations* which are the discrete-time counterpart to system (1):

$$\left(\frac{M}{h}+\frac{hK}{4}+\frac{D}{2}\right)q(k) - \left(\frac{2M}{h}-\frac{hK}{2}\right)q(k-1) + \left(\frac{M}{h}+\frac{hK}{4}-\frac{D}{2}\right)q(k-2) = f_d(k) \tag{14}$$

where for typographical homogeneity the q_k's have been rewritten $q(k)$ and

$$f_d(k) := \frac{h}{4}\,[f(hk)+2f(h(k-1))+f(h(k-2))]\,, \tag{15}$$

is an equivalent discrete force. Introducing the *discrete mass, damping and stiffness matrices*,

$$M_d := \frac{M}{h}+\frac{hK}{4}+\frac{D}{2}, \quad D_d := -\left[\frac{2M}{h}-\frac{hK}{2}\right] \quad K_d := \frac{M}{h}+\frac{hK}{4}-\frac{D}{2}, \tag{16}$$

equation (14) can be rewritten in a convenient second-order form as

$$M_d q(k) + D_d q(k-1) + K_d q(k-2) = f_d(k)\,, \tag{17}$$

where

$$f_d(k) = L\frac{h}{4}\,[u(hk)+2u(h(k-1))+u(h(k-2))] := Lu_d(k)\,, \tag{18}$$

the matrix L being the same as in the continuous-time model and $u(k)$ denotes the sampled value of the input force at $t = kh$. Note that the computation of the discrete forcing function $\{f_d(k)\}$ (or $u_d(k)$) requires adjacent samples at times k, $k-1$ and $k-2$ of the sampled external force f (or u).

Finally, by inverting the relations (16) we see that the original continuous time parameters (M, D, K) can be easily recovered from the parameters of the discretized model (14) by means of the *linear relations*

$$M := \frac{h}{4}[M_d + K_d - D_d], \quad D := M_d - K_d, \quad K := \frac{1}{h}[M_d + K_d + D_d]. \tag{19}$$

This is precisely what we wanted to achieve.

Naturally, it must be kept in mind that the solution of (14) provides an *approximation* of the exact flow $t \mapsto q(t)$ sampled at $t = kh$. The approximation error for the midpoint rule is of the order of $O(h^2)$ see [13, p. 402]. More sophisticated approximation schemes than (12) can provide approximations of arbitrarily high order, see [8].

One is tempted to interpret (14) as a discrete analog of Newton's law. In this spirit, one wonders if for the discrete (approximate) system the classical conservation laws of mechanics may still hold. In particular one may wonder if there is an associated discrete Hamiltonian function playing a similar role to the classical notion of total energy in continuous-time mechanics. A similar question rephrased in a system theoretic setting is if the variational discretization *preserves passivity* i.e. if the discretized mechanical system (14) obeys a *discrete dissipation inequality* of the same kind satisfied by the continuous system (3). In the next sections we shall show that for the particular " midpoint rule" discretization considered above this is indeed the case.

3 The Midpoint Rule and the Cayley Transform

It is a remarkable fact that that discretization by the midpoint rule (12) applied to a general linear time-invariant system is equivalent to the Cayley (or Tustin) transformation. In particular this will hold for the the midpoint rule variational integrator described above. Relations with the Cayley transform seem to have been noticed before; e.g. see [1], but in a rather different context.

Let us apply the midpoint rule approximation to a general state space model

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y &= Cx + Du \end{aligned} \tag{20}$$

on the interval $[0, h]$, getting

$$x(h) - x(0) \simeq \frac{h}{2}[Ax(h) + Bu(h) + Ax(0) + Bu(0)] = hA\frac{x(h) + x(0)}{2} + hB\frac{u(h) + u(0)}{2},$$

which leads to the discrete linear equation:

$$\left(I - \frac{h}{2}A\right)\bar{x}((k+1)h) = \left(I + \frac{h}{2}A\right)\bar{x}(kh) + \frac{hB}{2}\left(u((k+1)h) + u(kh)\right). \tag{21}$$

Note that the discrete state $\bar{x}(kh)$ is only an approximation of the sampled original continuous state $x(kh)$, even if the input function is piecewise-linear on each sampling interval (in which case the integration of u by the trapezoidal rule would be exact). Now, $I - \frac{h}{2}A$ is certainly invertible if h is small enough and we can solve the equation for $\bar{x}((k+1)h)$. Defining new discrete input and output sequences by the "midpoint rule"

$$u_{\frac{1}{2}}(kh) := \frac{u((k+1)h)+u(kh)}{2}, \qquad y_{\frac{1}{2}}(kh) := \frac{y((k+1)h)+y(kh)}{2} \tag{22}$$

we can write the discretized state space model in the form:

$$\begin{cases} \bar{x}((k+1)h) &= A_{\frac{1}{2}}\bar{x}(kh) + B_{\frac{1}{2}}u_{\frac{1}{2}}(kh) \\ y_{\frac{1}{2}}(kh) &= C_{\frac{1}{2}}\bar{x}(kh) + D_{\frac{1}{2}}u_{\frac{1}{2}}(kh) \end{cases} \tag{23}$$

where

$$A_{\frac{1}{2}} = \left(I - \frac{h}{2}A\right)^{-1}\left(I + \frac{h}{2}A\right), \qquad B_{\frac{1}{2}} = h\left(I - \frac{h}{2}A\right)^{-1}B,$$

$$C_{\frac{1}{2}} = C\left(I - \frac{h}{2}A\right)^{-1}, \qquad D_{\frac{1}{2}} = \frac{h}{2}C\left(I - \frac{h}{2}A\right)^{-1}B + D. \tag{24}$$

These formulas, modulo the introduction of the factor $h/2$, define precisely the well-known *Cayley transform* of linear systems theory. Naturally the analog in terms of transfer functions is the bilinear, so-called *Tustin transform*, well-known in sampled-data control

$$s = \frac{2}{h}\frac{z-1}{z+1}. \tag{25}$$

Now, it is almost immediate to check that by applying the Tustin transform to the transfer function of the second order system (1) one in fact obtains the discrete transfer function of the second order difference equation (17) with input exactly given by the expression (18). In other words,

Theorem 1. *Variational integration by the midpoint rule* (12) *applied to a linear mechanical system produces the discretization defined by the Cayley transform. In other words,* (16) *are the input-output counterparts of the Cayley transform* (24) *applied to the natural state space realization* (3).

4 Preservation of Passivity under Midpoint Sampling

There is a sizable literature on passivity of sampled (i.e. discretized) continuous linear systems, see [3, 14, 16]. Even if there is a clear axiomatic definition of passive discrete linear (and nonlinear) systems, it is not immediately clear how to do sampling in such a way as to preserve passivity. For example, it is well known that with the standard definition of sampled input-output functions, neither Euler method nor the Zero-Order-Hold sampling in general preserve passivity, see [9].

We shall now assume that the linear system (20) is passive; i.e. $\dim y(t) = \dim u(t)$ and there exist a quadratic energy function $V(x) = \frac{1}{2}x^\top Px$ such that

$$V(x(t)) - V(x(0)) \leq \int_0^t y^\top(s)u(s)\,ds\,. \tag{26}$$

It is a basic fact of linear systems theory [19] that dissipativity is *equivalent* to the existence of symmetric positive semidefinite matrices P solutions of the linear matrix inequality (LMI)

$$\begin{bmatrix} A^\top P + PA & PB - C^\top \\ C - B^\top P & D + D^\top \end{bmatrix} \leq 0 \tag{27}$$

which is in turn equivalent to the fact that the transfer function of a passive system: $G(s) := C(sI-A)^{-1}B + D$ should be *positive-real*, see [19] .

Lossless systems are an important special case. For these systems the inequality in (26) is replaced by an equality sign. It can be shown that, under natural minimality assumptions for the realization (A,B,C), in this case the LMI (27) has a *unique* solution $P = P^\top$ which is strictly positive definite. This function is a bona-fide *total energy* of the system. Linear port-controlled Hamiltonian systems (see [17]) are a special case: they are lossless systems with an Hamiltonian structure. It is shown in [17] that the energy function of these systems is in fact the Hamiltonian function.

Passivity for discrete linear system is defined as for the continuous-time case. A discrete linear system,

$$\begin{aligned} x(k+1) &= A_d x(k) + B_d u(k) \\ y(k) &= C_d x(k) + D_d u(k) \end{aligned} \tag{28}$$

is *passive* if there exist a quadratic energy function $V(x) = \frac{1}{2}x^\top Px$ such that

$$V\left(x(k+1)\right) - V\left(x(k)\right) \leq y(k)^\top u(k)\,.$$

It is shown that a linear discrete system in the form (28) is passive if and only if the discrete linear matrix inequality (DLMI):

$$\begin{bmatrix} A_d^\top PA_d - P & A_d^\top PB_d - C_d^\top \\ B_d^\top PA_d - C_d & B_d^\top PB_d - (D_d + D_d^\top) \end{bmatrix} \leq 0 \tag{29}$$

admits symmetric positive semidefinite solution matrices P. The discrete LMI condition can be generalized to nonlinear systems as reported for example in [11].

The following fact was apparently first discovered by P. Faurre in 1973 and can be found in an unpublished INRIA report [6].

Theorem 2 (P. Faurre). *Consider a (minimal) linear system* (20) *and its discrete-time counterpart obtained by the Cayley transform formulas* (24). *Then one system* *is passive if and only if the other is, and the energy functions are the same. Moreover the set of solution of the DLMI is the same of the LMI for the continuous time model.*

Note that this statement *per se* does not tell how the inputs and outputs of the continuous system should be "sampled" in order to preserve passivity nor what relation the discrete state of the sampled system has with the continuous state. The midpoint rule interpretation of the bilinear transformation given above answers these questions.

In particular, the midpoint rule variational integrator is a passive discrete mechanical system which is conservative (lossless) if and only if the original continuous-time system was.

5 Conclusions and Related Work

A natural discretization procedure for the equations of motion of a linear mechanical system has been described which leads to a much better conditioned recovery of the continuous time mechanical parameters than the usual discretizations. Some applications to identification of linear mechanical systems are reported in [2].

References

1. Austin, M.A., Krishnaprasad, P.S., Wang, L.S.: Almost poisson integration of rigid body dynamics. J. Comput. Phys. 52, 105–117 (1993)
2. Bruschetta, M., Picci, G., Saccon, A.: Discrete mechanical systems: Second order modelling and identification. In: Proceedings of the 15th IFAC System Identification Symposium (SYSID), [CD-ROM], Saint Malo, France, pp. 456–461 (2009)
3. Costa-Castello, R., Fossas, E.: On preserving passivity in sampled data linear systems. In: Proceedings of the 2006 American Control Conference, Minneapolis, MI, pp. 4373–4378 (2006)
4. De Angelis, M., Lus, H., Betti, R., Longman, R.W.: Extracting physiscal parameters of mechanical models from identified state-space representations. ASME Trans. Journal Appl. Mech. 69, 617–625 (2002)
5. Dieci, L., Papini, A.: Conditioning and Padè approximation of the logarithm of a matrix. SIAM J. Matr. Anal. Appl. 21, 913–930 (2000)
6. Faurre, P.: Réalisations Markoviennes de processus stationnaires. INRIA, Roquencourt, France, Rapport de Recherche no. 13 (1973)
7. Garnier, H., Wang, L. (eds.): Identification of Continuous-time Models from Sampled Data. Springer, London (2008)
8. Hairer, E., Lubich, C., Wanner, G.: Geometric Numerical Integration: Structure Preserving Algorithms for Ordinary Differential Equations, 2nd edn. Springer, Berlin (2005)
9. Jiang, J.: Preservation of passivity under discretization. Journal of the Franklin Institute 330, 721–734 (1993)
10. Laub, A., Arnold, W.F.: Controllability and observability criteria for linear second-order models. IEEE Trans. Automatic Control 29, 163–165 (1984)
11. Lin, W., Byrnes, C.I.: Passivity and absolute stabilization of a class of discrete-time nonlinear systems. Automatica 31(263-267) (1995)

12. Lus, H., De Angelis, M., Betti, R., Longman, R.W.: Constructing second-order models of mechanical systems from identified state-space realizations, part I: theoretical discussions. ASCE Journal of Engineering Mech. 129, 477–488 (2002)
13. Marsden, J.E., West, M.: Discrete mechanics and variational integrators. Acta Numerica, 357–514 (2001)
14. Monaco, S., Normand-Cyrot, D., Tiefensee, F.: From passivity under sampling to a new discrete-time passivity concept. In: Proceedings of the 47th IEEE Conference on Decision and Control, Cancun Mexico, pp. 3157–3162 (2008)
15. Sinha, N.K., Rao, G.P.: Identification of Continuous-Time Systems. Kluwer Academic, Dordrecht (1991)
16. Stramigioli, S., Secchi, C., van der Schaft, A.: Sampled data system passivity and discrete port-Hamiltonian systems. IEEE Transactions on Robotics 21, 574–587 (2002)
17. van der Schaft, A.J.: L^2 Gain and Passivity Techniques in Nonlinear Control, 2nd edn. Springer, Heidelberg (2000)
18. Veselov, A.P.: Integrable discrete-time systems and difference operators. Funkts. Anal. Prilozhen (Funct. An. and Appl.) 22, 113 (1988)
19. Willems, J.C.: Dissipative dynamical systems, Part II: linear systems with quadratic supply rates. Arch. Ration. Mech. Anal. 45, 321–393 (1972)

A Note on LQ Decomposition in Stochastic Subspace Identification

Tohru Katayama

In this paper, we consider the role of LQ decomposition in the realization-based subspace identification method for discrete-time stochastic systems as a continuation of our earlier work [6] in deterministic setting. Under the assumption that the past horizon of the data matrix is infinite, we reveal a nice block lower triangular structure of a certain L-factor related to the stochastic component in the LQ decomposition. Adapting this theoretical result to finite input-output data, we derive an approximate method of identifying all the system parameters, including the steady state Kalman gain and the covariance of the innovation process, from L-factors of a single LQ decomposition of the data matrix.

1 Introduction

The LQ decomposition, together with the singular value decomposition (SVD), has extensively been used as a numerical tool in subspace system identification methods [5, 12, 14]. We have employed the LQ decomposition for a preliminary orthogonal decomposition of the output process into deterministic and stochastic components in order to develop a stochastic realization theory in the presence of exogenous input [9]. Also, we have examined the role of LQ decomposition in the subspace identification for deterministic systems [6], in which a relation between the column vectors of L-factors in the LQ decomposition and the past and future inputs-outputs used in defining a Hankel operator [4] was clarified.

It is well known that in the PO-MOESP method [13, 14], the system matrices are determined by using L-factors related to the deterministic component in the LQ decomposition of the data matrix. For combined deterministic-stochastic systems [11, 12], the steady state Kalman gain and the covariance matrix of innovation process are derived, after the identification of deterministic component, by solving an

Tohru Katayama
Faculty of Culture and Information Science, Doshisha University, KyoTanabe,
Kyoto 610-0394, Japan
e-mail: tohru_katayama@nifty.com

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 355–364.
springerlink.com

algebraic Riccati equation that arises in stochastic realization theory [1, 3], so that an L-factor related to the stochastic component has been discarded.

This fact implies that the L-factor related to the stochastic component in the LQ decomposition still remains to be utilized in subspace identification algorithms for combined deterministic-stochastic systems. Along this line, in Di Ruscio [2], a method of computing the steady state Kalman gain directly from L-factors related to the stochastic component has been derived in a different setting than standard subspace identification problems. It seems, however, that the derivation of algorithms in [2] is obscure in the sense that no proofs are provided therein.

In this paper, under the assumption that the past horizon is infinite and the number of input-output data goes to infinity, we analyze a certain L-factor in the LQ decomposition by using a technique developed in the PO-MOESP method [13], thereby showing that the L-factor has a special block lower triangular structure. Adapting this theoretical result to finite input-output data, we derive a new procedure for identifying the steady state Kalman gain and the covariance of innovation process. We therefore show that all the system parameters of a stochastic system with an innovation form are computed from L-factors only in a single LQ decomposition of finite data matrix. This result is not unexpected, in view of the fact that the LQ decomposition transforms a given data matrix into a product of a lower triangular matrix and an orthogonal matrix, where the former carries the information useful for system identification (least-squares), while the latter provides orthogonal bases of the row space of data matrix.

The rest of the paper is organized as follows. The problem is stated in Section 2. Section 3 reviews basic matrix state-input-output equations. By introducing the LQ decomposition, a subspace method of computing the deterministic component is briefly discussed in Section 4. In Section 5, we analyze a block lower triangular structure of a certain L-factor related to the stochastic component in the LQ decomposition under the assumption that the past horizon is infinite. For finite input-output data, we then derive a method of estimating the steady state Kalman gain and the covariance of innovation process. We conclude this paper in Section 6.

2 Problem Statement

Suppose that a stochastic system is given by the innovation form

$$x(t+1) = Ax(t) + Bu(t) + Ke(t) \tag{1}$$

$$y(t) = Cx(t) + Du(t) + e(t) \tag{2}$$

where $x \in \mathbb{R}^n$ is the state vector, $u \in \mathbb{R}^m$ the exogenous input, $y \in \mathbb{R}^p$ the output vector, $e \in \mathbb{R}^p$ the innovation vector, and $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $K \in \mathbb{R}^{n\times p}$, $C \in \mathbb{R}^{p\times n}$, $D \in \mathbb{R}^{p\times m}$ are constant matrices. In the following, we assume that (A,B) and (A,K) are reachable and (C,A) is observable. Also, the innovation process e is a white noise vector with mean zero and covariance matrix

$$E\{e(t)e^{\mathrm{T}}(s)\} = \Lambda\delta_{ts}, \quad \Lambda > 0, \quad \Lambda \in \mathbb{R}^{p\times p} \tag{3}$$

and is uncorrelated with the past state, i.e. $E\{e(t)x^{\mathrm{T}}(s)\} = 0, t \geq s$.

The problem in this paper is to examine the role of LQ decomposition in the subspace identification of combined deterministic-stochastic systems based on finite input-output data $\{y(t), u(t), t = 0, 1, \cdots, T\}$. We assume throughout the paper that there is no feedback from the output to the input [9] and that the input is persistently exciting (PE) with sufficiently high order. We also assume in Section 5 that the past horizon is infinite to analyze an L-factor related to the stochastic component.

In the following, we show that all the system parameters of a stochastic system with an innovation form can be identified by a single LQ decomposition of the data matrix without solving the algebraic Riccati equation; see also [10] for a related issue of the LQ decomposition in Hilbert space.

3 Matrix State-Input-Output Equation

In this section, we derive the matrix state-input-output equation, on which most subspace identification methods are based.

Let $k > n$. We define the stacked vectors as

$$y_k(t) := \begin{bmatrix} y(t) \\ y(t+1) \\ \vdots \\ y(t+k-1) \end{bmatrix}, \quad u_k(t) := \begin{bmatrix} u(t) \\ u(t+1) \\ \vdots \\ u(t+k-1) \end{bmatrix}, \quad e_k(t) := \begin{bmatrix} e(t) \\ e(t+1) \\ \vdots \\ e(t+k-1) \end{bmatrix}$$

where $y_k(t), \boldsymbol{e}_k(t) \in \mathbb{R}^{kp}$, $\boldsymbol{u}_k(t) \in \mathbb{R}^{km}$. Then, we see that the stacked vectors satisfy the well-known augmented equation

$$y_k(t) = \mathscr{O}_k x(t) + \mathscr{T}_k \boldsymbol{u}_k(t) + \mathscr{K}_k e_k(t) \tag{4}$$

where $\mathscr{O}_k$ is the extended observability matrix given by

$$\mathscr{O}_k = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{k-1} \end{bmatrix} \in \mathbb{R}^{kp\times n},$$

and where $\mathscr{T}_k$ is the block lower triangular Toeplitz matrix defined by

$$\mathscr{T}_k = \begin{bmatrix} D & & \mathbf{0} \\ CB & D & \\ \vdots & \ddots & \ddots \\ CA^{k-2}B \cdots & CB & D \end{bmatrix} \in \mathbb{R}^{kp\times km} \tag{5}$$

and $\mathscr{K}_k$ is the block lower triangular Toeplitz matrix defined by

$$\mathscr{K}_k = \begin{bmatrix} I_p & & & \mathbf{0} \\ CK & I_p & & \\ \vdots & \ddots & \ddots & \\ CA^{k-2}K & \cdots & CK & I_p \end{bmatrix} \in \mathbb{R}^{kp \times kp} \tag{6}$$

We also define block Hankel matrices

$$Y_{0|k-1} = \begin{bmatrix} y(0) & y(1) & \cdots & y(N-1) \\ y(1) & y(2) & \cdots & y(N) \\ \vdots & \vdots & \ddots & \vdots \\ y(k-1) & y(k) & \cdots & y(k+N-2) \end{bmatrix} \in \mathbb{R}^{kp \times N}$$

$$U_{0|k-1} = \begin{bmatrix} u(0) & u(1) & \cdots & u(N-1) \\ u(1) & u(2) & \cdots & u(N) \\ \vdots & \vdots & \ddots & \vdots \\ u(k-1) & u(k) & \cdots & u(k+N-2) \end{bmatrix} \in \mathbb{R}^{km \times N}$$

and

$$E_{0|k-1} = \begin{bmatrix} e(0) & e(1) & \cdots & e(N-1) \\ e(1) & e(2) & \cdots & e(N) \\ \vdots & \vdots & \ddots & \vdots \\ e(k-1) & e(k) & \cdots & e(k+N-2) \end{bmatrix} \in \mathbb{R}^{kp \times N}$$

Similarly, we define $U_{k|2k-1}, Y_{k|2k-1}, E_{k|2k-1}$. It then follows from (4) that the matrix state-input-output equations are given by

$$Y_{0|k-1} = \mathscr{O}_k X_0 + \mathscr{T}_k U_{0|k-1} + \mathscr{K}_k E_{0|k-1} \tag{7}$$

$$Y_{k|2k-1} = \mathscr{O}_k X_k + \mathscr{T}_k U_{k|2k-1} + \mathscr{K}_k E_{k|2k-1} \tag{8}$$

where

$$X_0 = [x(0)\ x(1)\ \cdots\ x(N-1)] \in \mathbb{R}^{n \times N}$$
$$X_k = [x(k)\ x(k+1)\ \cdots\ x(k+N-1)] \in \mathbb{R}^{n \times N}$$

are the initial state matrices for (7) and (8), respectively.

4 LQ Decomposition

The next step is to apply the LQ decomposition to data matrix in order to obtain some structural information about the extended observability matrix $\mathscr{O}_k$, the block Toeplitz matrices $\mathscr{T}_k$ and $\mathscr{K}_k$, from which system parameters are identified.

To make precise the assumption that the input u is PE with sufficiently high order, we assume here that $U_{0|2k-1}$ has full row rank, i.e. $\mathrm{rank}(U_{0|2k-1}) = 2km$. Also, note that the reachability of $(A,[B\ \ K])$ assures that X_k has full row rank.

Define the past input-output matrix $W_{0|k-1} = \begin{bmatrix} U_{0|k-1} \\ Y_{0|k-1} \end{bmatrix} \in \mathbb{R}^{k(m+p)\times N}$, which was used as an instrument to remove noise effects from the matrix state-input-output equation [15]. According to [13], the actual computation to remove the effect of noises from (8) is performed by the LQ decomposition

$$\begin{bmatrix} U_{k|2k-1} \\ W_{0|k-1} \\ Y_{k|2k-1} \end{bmatrix} = \begin{bmatrix} R_{11} & 0 & 0 \\ R_{21} & R_{22} & 0 \\ R_{31} & R_{32} & R_{33} \end{bmatrix} \begin{bmatrix} Q_1^{\mathrm{T}} \\ Q_2^{\mathrm{T}} \\ Q_3^{\mathrm{T}} \end{bmatrix} \tag{9}$$

where $R_{11} \in \mathbb{R}^{km\times km}$, $R_{22} \in \mathbb{R}^{k(m+p)\times k(m+p)}$, $R_{33} \in \mathbb{R}^{kp\times kp}$ are block lower triangular matrices, and $Q_1 \in \mathbb{R}^{N\times km}$, $Q_2 \in \mathbb{R}^{N\times k(m+p)}$, $Q_3 \in \mathbb{R}^{N\times kp}$ are orthogonal matrices with $Q_i^{\mathrm{T}} Q_j = I\delta_{ij}$, $i,j = 1,2,3$.

Post-multiplying (8) by Q_1 and Q_2 respectively yields

$$Y_{k|2k-1}Q_1 = \mathscr{O}_k X_k Q_1 + \mathscr{T}_k U_{k|2k-1} Q_1 + \mathscr{K}_k E_{k|2k-1} Q_1 \tag{10}$$

$$Y_{k|2k-1}Q_2 = \mathscr{O}_k X_k Q_2 + \mathscr{T}_k U_{k|2k-1} Q_2 + \mathscr{K}_k E_{k|2k-1} Q_2 \tag{11}$$

Also, from (9), we have

$$Y_{k|2k-1}Q_1 = R_{31}, \qquad U_{k|2k-1}Q_1 = R_{11} \tag{12}$$

$$Y_{k|2k-1}Q_2 = R_{32}, \qquad U_{k|2k-1}Q_2 = 0 \tag{13}$$

Thus substituting (12) and (13) into (10) and (11) respectively yields

$$R_{31} = \mathscr{O}_k X_k Q_1 + \mathscr{T}_k R_{11} + \mathscr{K}_k E_{k|2k-1} Q_1 \tag{14}$$

$$R_{32} = \mathscr{O}_k X_k Q_2 + \mathscr{K}_k E_{k|2k-1} Q_2 \tag{15}$$

It should be noted that the innovation process e is a white noise and that there is no feedback from the output to the input. Thus we see that the innovation process is uncorrelated with the past input-output $W_{0|k-1}$ and the future input $U_{k|2k-1}$. Thus, it follows that

$$\lim_{N\to\infty} \frac{1}{N} E_{k|2k-1} [U_{k|2k-1}^{\mathrm{T}}\ \ W_{0|k-1}^{\mathrm{T}}] = 0$$

Moreover, we see from (9) that

$$[U_{k|2k-1}^{\mathrm{T}}\ \ W_{0|k-1}^{\mathrm{T}}] = [Q_1\ \ Q_2] \begin{bmatrix} R_{11} & 0 \\ R_{21} & R_{22} \end{bmatrix}^{\mathrm{T}}$$

where R_{11} and R_{22} are nonsingular. Thus, we have [13]

$$\lim_{N\to\infty} \frac{1}{\sqrt{N}} E_{k|2k-1} [Q_1\ \ Q_2] = 0 \tag{16}$$

This implies that Q_1^{T} and Q_2^{T} are uncorrelated with the future noise $E_{k|2k-1}$.

Letting $N \to \infty$ in (14) and (15), we get asymptotically

$$R_{31} = \mathscr{O}_k X_k Q_1 + \mathscr{T}_k R_{11}, \qquad R_{32} = \mathscr{O}_k X_k Q_2 \tag{17}$$

where the common factor $1/\sqrt{N}$ in the above relations is suppressed. A subspace method of computing (A,B,C,D) of the deterministic component from the above equations is well known in the MOESP method; see [13, 14].

5 Asymptotic Analysis of Structure of R_{33}

Recall from (9) that

$$Y_{k|2k-1} = R_{31}Q_1^{\mathrm{T}} + R_{32}Q_2^{\mathrm{T}} + R_{33}Q_3^{\mathrm{T}} \tag{18}$$

As discussed in Section 4, the L-factors R_{31} and R_{32} contain complete information about the deterministic component (A,B,C,D) asymptotically. Hence, it is clear from (18) that the L-factor R_{33} carries information about the stochastic component; but it seems that this fact was overlooked in the literature. The analysis below is motivated by Di Ruscio [2], who has developed a method of computing the steady state Kalman gain directly from L-factors in a different setting than a standard subspace identification method; however, the structure of L-factors is yet to be clarified.

As shown in (16), the correlation between the noise $E_{k|2k-1}$ and orthogonal factors Q_1^{T} and Q_2^{T} has been evaluated to derive the basic equation (17) satisfied by the deterministic component [13]; however, the correlation between $E_{k|2k-1}$ and Q_3 has not been considered therein, which is the main topic of this section.

For the asymptotic analysis of the structure of R_{33}, we assume that y, u and e are 2nd-order stationary processes and that the past horizon is infinite. This implies that the entire past input-output $W_{-\infty|k-1}$ is available [7], so that the LQ decomposition should be understood in this context. In fact, though R_{31} and R_{33} have finite number of columns, R_{32} (and Q_2) in (9) has infinite number of columns.

It follows from (8) and (18) that

$$\mathscr{O}_k X_k + \mathscr{T}_k U_{k|2k-1} - R_{31}Q_1^{\mathrm{T}} - R_{32}Q_2^{\mathrm{T}} = R_{33}Q_3^{\mathrm{T}} - \mathscr{K}_k E_{k|2k-1} \tag{19}$$

The first term $\mathscr{O}_k X_k$ in the left-hand side of (19) is the oblique projection of $Y_{k|2k-1}$ onto $W_{-\infty|k-1}$ along $U_{k|2k-1}$ [7, 11]. This implies that the row vectors of $\mathscr{O}_k X_k$ lie in the row space of $W_{-\infty|k-1}$, so that it is expressed in terms of Q_1^{T} and Q_2^{T}. Also, from (9), $U_{k|2k-1}$ is expressed in terms of Q_1^{T}. Thus, the left-hand side of (19) is expressed as a linear combination of Q_1^{T} and Q_2^{T}. Hence, post-multiplying (19) by Q_3 yields

$$R_{33} = \mathscr{K}_k E_{k|2k-1} Q_3 \in \mathbb{R}^{kp \times kp} \tag{20}$$

On the other hand, we see from (16) that $E_{k|2k-1}$ is orthogonal to Q_1^{T} and Q_2^{T} asymptotically, so that $E_{k|2k-1}$ is expressed in terms of Q_3^{T} asymptotically. Thus, we can write

$$E_{k|2k-1} = H_k Q_3^{\mathrm{T}} + h_N, \qquad H_k \in \mathbb{R}^{kp\times kp}, \qquad h_N \in \mathbb{R}^{kp\times N}$$

where the norm of $H_k/\sqrt{N}$ is bounded, and that of $h_N/\sqrt{N}$ goes to 0 as $N \to \infty$. This implies that

$$\left\| \frac{1}{\sqrt{N}} \left(E_{k|2k-1} - H_k Q_3^{\mathrm{T}} \right) \right\|^2 \leq \frac{\|h_N\|^2}{N} \to 0$$

Now consider

$$\frac{1}{N} E_{k|2k-1} E_{k|2k-1}^{\mathrm{T}} = \frac{1}{N} \left(H_k H_k^{\mathrm{T}} + h_N Q_3 H_k^{\mathrm{T}} + H_k Q_3^{\mathrm{T}} h_N^{\mathrm{T}} + h_N h_N^{\mathrm{T}} \right)$$

Under the ergodicity assumption [8], we see from (3) that the left-hand side of the above equation converges, i.e.

$$\lim_{N\to\infty} \frac{1}{N} E_{k|2k-1} E_{k|2k-1}^{\mathrm{T}} = \begin{bmatrix} \Lambda & & \\ & \ddots & \\ & & \Lambda \end{bmatrix} = \mathscr{L} \tag{21}$$

Also, from the conditions for H_k and h_N and the fact that $\|Q_3\| = 1$, we have

$$\lim_{N\to\infty} \left\| \frac{1}{N} E_{k|2k-1} E_{k|2k-1}^{\mathrm{T}} - \frac{1}{N} H_k H_k^{\mathrm{T}} \right\| = 0 \tag{22}$$

so that from (21) and (22),

$$\lim_{N\to\infty} \frac{1}{N} H_k H_k^{\mathrm{T}} = \mathscr{L}$$

Similarly, we can show that

$$\lim_{N\to\infty} \frac{1}{N} E_{k|2k-1} Q_3 Q_3^{\mathrm{T}} E_{k|2k-1}^{\mathrm{T}} = \mathscr{L} \tag{23}$$

Thus, from (20) and (23), we have

$$\lim_{N\to\infty} \frac{1}{N} R_{33} R_{33}^{\mathrm{T}} = \mathscr{K}_k \begin{bmatrix} \Lambda & & \\ & \ddots & \\ & & \Lambda \end{bmatrix} \mathscr{K}_k^{\mathrm{T}}$$

Since $\Lambda > 0$, it follows that for $N \to \infty$, a Cholesky factor is given by

$$\frac{1}{\sqrt{N}} R_{33} = \mathscr{K}_k \begin{bmatrix} F_1 & & \\ & \ddots & \\ & & F_k \end{bmatrix} \tag{24}$$

By definition, both R_{33} and $\mathscr{K}_k$ are block lower triangular, so that we see that F_i, $i = 1, \cdots, k$ are also block lower triangular and satisfy

$$\lim_{N \to \infty} F_i F_i^{\mathrm{T}} = \Lambda, \quad i = 1, \cdots, k \tag{25}$$

Moreover, since R_{33} is unique up to post-multiplication by a signature matrix, so are F_i, $i = 1, \cdots, k$. The main result of the paper is summarized as follows.

Proposition 1. *Suppose that the past horizon is infinite. Then, for $N \to \infty$, the L-factor R_{33} has a block lower triangular structure of the form*

$$\frac{1}{\sqrt{N}} R_{33} = \begin{bmatrix} F_1 & 0 & & \\ CKF_1 & F_2 & \ddots & \\ \vdots & \vdots & \ddots & 0 \\ CA^{k-2}KF_1 & CA^{k-3}KF_2 & \cdots & F_k \end{bmatrix} \in \mathbb{R}^{kp \times kp} \tag{26}$$

Proof. A proof is immediate from (6) and (24). □

Adapting the above result to finite input-output data, we obtain a procedure of identifying K and Λ by using the finite horizon LQ decomposition of (9).

For simplicity, we write $\bar{R}_{33} = R_{33}/\sqrt{N}$. Then, the diagonal elements of (26) are expressed as $F_i = \bar{R}_{33}((i-1)p+1 : ip, (i-1)p+1 : ip)$, $i = 1, \cdots, k$. It then follows from (26) that

$$\begin{aligned} \mathscr{O}_{k-1} K F_1 &= \bar{R}_{33}(p+1 : kp, 1 : p) \\ \mathscr{O}_{k-2} K F_2 &= \bar{R}_{33}(2p+1 : kp, p+1 : 2p) \\ &\vdots \\ \mathscr{O}_1 K F_{k-1} &= \bar{R}_{33}((k-1)p+1 : kp, (k-2)p+1 : (k-1)p) \end{aligned} \tag{27}$$

Algorithm of Identifying (K, Λ)

Step 1: Rewrite the right-hand side members in (27) as

$$\mathscr{R}_i = \bar{R}_{33}((i-1)p+1 : kp,\ (i-1)p+1 : ip) F_i^{-1}, \quad i = 1, \cdots, k$$

where F_i, $i = 1, \cdots, k$ are nonsingular. Then (27) is reduced to

$$\begin{bmatrix} \mathscr{O}_{k-1} \\ \mathscr{O}_{k-2} \\ \vdots \\ \mathscr{O}_1 \end{bmatrix} K = \begin{bmatrix} \mathscr{R}_1 \\ \mathscr{R}_2 \\ \vdots \\ \mathscr{R}_{k-1} \end{bmatrix} \tag{28}$$

Step 2: Apply the least-squares method to (28) to obtain an estimate of the steady state Kalman gain K.

Step 3: It follows from (25) that an estimate of the covariance matrix Λ is given by

$$\hat{\Lambda} = \frac{1}{k} \sum_{i=1}^{k} F_i F_i^{\mathrm{T}}$$

We see that the identification method of K in the above algorithm is quite similar to that of (B,D) in the MOESP method; see [14] that exploits the block structure of the Toeplitz matrix $\mathscr{T}_k$ of (5). It should be noted that the identified feedback matrix $A - KC$ may not necessarily be stable[1].

We have thus established a realization-based subspace identification method for combined deterministic-stochastic systems based on the LQ decomposition. It should be noted that the present result can be used as a supplement to realization-based subspace identification algorithms developed in [7, 9, 11, 13].

6 Conclusions

In this paper, we have clarified a special role of the LQ decomposition in subspace identification methods for combined deterministic-stochastic systems. In particular, under the assumption that the past horizon is infinite, we have shown that R_{33} has a nice block lower triangular structure asymptotically. By adapting the theoretical result to finite input-output data, we can easily derive a procedure for computing the stochastic component. Thus, combining with the earlier result [6], the role of the LQ decomposition in realization-based N4SID methods is completely clarified.

Epilogue

It was an early spring of 1975. I traveled the US from west to east, after spending one year at System Science Department, UCLA. I still remember quite well, when I visited Yutaka at University of Gainesville, where he was a PhD student of graduate program under the guidance of Professor Kalman, whom I met for the first time in his office. Yutaka also took me to the office of Professor Popov, with whom I took a picture. To see alligators, we went out together with Tsuyoshi Matsuo, who was studying dynamical system theory there. My visit to Gainesville three decades ago, where I met many important people, was a quite memorable event during my stay in the US. I am grateful to Yutaka for his hospitality at Gainesville.

References

1. Akaike, H.: Markovian representation of stochastic processes by canonical variables. SIAM J. Control 13(1), 162–173 (1975)
2. Di Ruscio, D.: A method for identification of combined deterministic stochastic systems. In: Aoki, M., Havenner, A.M. (eds.) Applications of Computer Aided Time Series Modeling, pp. 181–205. Springer, Heidelberg (1997)

[1] A way of getting a stabilizing K is to solve a Kalman filter Riccati equation using noise covariances computed from residuals, after identifying the state vector and the deterministic component [7].

3. Faurre, P.: Stochastic realization algorithms. In: Mehra, R., Lainiotis, D. (eds.) System Identification: Advances and Case Studies, pp. 1–25 (1976)
4. Kalman, R.E., Falb, P.L., Arbib, M.A.: Topics in Mathematical System Theory. McGraw-Hill, New York (1969)
5. Katayama, T.: Subspace Methods for System Identification. Springer, Heidelberg (2005)
6. Katayama, T.: Role of LQ decomposition in subspace identification methods. In: Chiuso, A., Ferrante, A., Pinzoni, S. (eds.) Modeling, Estimation and Control — Festschrift in Honor of Giorgio Picci on the Occasion of his Sixty-Fifth Birthday. Lecture Notes in Control and Information Sciences, vol. 364, pp. 207–220. Springer, Heidelberg (2007)
7. Katayama, T., Picci, G.: Realization of stochastic systems with exogenous inputs and subspace identification methods. Automatica 35(10), 1635–1652 (1999)
8. Ljung, L.: System Identification — Theory for the User, 2nd edn. Prentice-Hall, Englewood Cliffs (1999)
9. Picci, G., Katayama, T.: Stochastic realization with exogenous inputs and ‘subspace methods’ identification. Signal Processing 52(2), 145–160 (1996)
10. Tanaka, H., Katayama, T.: A stochastic realization algorithm via block LQ decomposition in Hilbert space. Automatica 42(5), 741–746 (2006)
11. Van Overschee, P., De Moor, B.: N4SID — Subspace algorithms for the identification of combined deterministic-stochastic systems. Automatica 30(1), 75–93 (1994)
12. Van Overschee, P., De Moor, B.: Subspace Identification for Linear Systems. Kluwer Academic, Dordrecht (1996)
13. Verhaegen, M.: Identification of the deterministic part of MIMO state space models given in innovations form from input-output data. Automatica 30(1), 61–74 (1994)
14. Verhaegen, M., Verdult, V.: Filtering and System Identification — A Least Squares Approach. Cambridge University Press, Cambridge (2007)
15. Viberg, M.: Subspace-based methods for identification of linear time-invariant systems. Automatica 31(12), 1835–1851 (1995)

Modeling Systems Based on Noisy Frequency and Time Domain Measurements

Sanda Lefteriu*, Antonio C. Ionita**, and Athanasios C. Antoulas

Abstract. The Loewner matrix framework can identify the underlying system from given noise-free measurements either in the frequency, or in the time domain[1, 2]. This paper provides an analysis of the effects of noise on the performance of the SVD implementation of the Loewner matrix framework for different noise levels and proposes an improved approach which is able to identify an approximation of the original system even for high levels of noise. Moreover, for frequency domain measurements, our framework can handle systems with a large number of inputs and outputs while requiring small computational time and storage.

1 Introduction

Modeling systems based on tabulated data obtained from direct measurements is common to many engineering applications. Due to their limited capability, measurement devices can output only a certain number of digits of the measured quantity,

Sanda Lefteriu
Rice University, 6100 Main Street, MS-366, Houston, TX-77005, USA, Fraunhofer Institut für Techno- und Wirtschaftsmathematik, Fraunhofer-Platz 1, 67663 Kaiserslautern, Germany, Jacobs University Bremen, Campus Ring 1, 28725 Bremen, Germany
e-mail: Sanda.Lefteriu@rice.edu

Antonio C. Ionita
Rice University, 6100 Main Street, MS-366, Houston, TX-77005, USA
e-mail: cosmin.ionita@rice.edu

Athanasios C. Antoulas
Rice University, 6100 Main Street, MS-366, Houston, TX-77005, USA, Jacobs University Bremen, Campus Ring 1, 28725 Bremen, Germany
e-mail: aca@rice.edu

* Partially supported by the Roberto Rocca education program and DFG.

** Partially supported by NSF grant DMS-CAREER-0449973.

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 365–378.
springerlink.com

some of which may be wrong. On the other hand, noise should not affect a robust algorithm used to build the model.

In the electronics community, macromodeling is frequently used to build interconnect models from data obtained via full-wave simulations or direct measurements. For multi-port systems, currently available macromodeling techniques [3] are expensive. Thus, in [1], we propose a new approach which is based on a system-theoretic tool, the Loewner matrix pencil constructed in the context of tangential interpolation. Several implementations are possible, but they all construct models of low order and are especially designed for the case of a large number of terminals. Moreover, they allow the identification of the underlying system, rather than merely fitting the measurements. After a short review of system theoretic concepts in Sect. 2, we briefly introduce the Loewner framework in Sect. 3 and its SVD implementation in Sect. 4. In Sect. 5, we investigate its robustness through a controlled experiment and notice that, as it is, it fails to identify the underlying system for high noise values. This is due to the fact that noise affects the poles recovered by our algorithm and some of the physical poles show up only after overmodeling, so choosing an order which is too low will lead to high approximation errors. Thus, a model of dimension higher than the order of the underlying system is needed to capture all physical poles. To solve this, we take additional steps which involve computing and sorting the norms of the residues, as well as those of the dominance quantities, which determine the dominant poles, of these higher order models. Even for high noise ratios, the poles recovered using one of these criteria are perturbations, in the order of the noise level, of the physical poles. We validate these observations by an example obtained from real measurements.

2 System Theory

An LTI system Σ with m-inputs, p-outputs and n-internal variables in *descriptor-form representation* is given by a set of differential and algebraic equations:

$$\Sigma: \quad \mathbf{E}\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \;\; \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t), \tag{1}$$

where $\mathbf{x}(t)$ is an internal variable (the state, if $\mathbf{E}$ is invertible), $\mathbf{u}(t)$ is the input, $\mathbf{y}(t)$ is the corresponding output, while $\mathbf{E}, \mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$, $\mathbf{C} \in \mathbb{R}^{p\times n}$, $\mathbf{D} \in \mathbb{R}^{p\times m}$ are constant with $\mathbf{E}$ possibly singular. The *transfer function* of Σ is $\mathbf{H}(s) = \mathbf{C}\left(s\mathbf{E} - \mathbf{A}\right)^{-1}\mathbf{B} + \mathbf{D}$. The set $[\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}]$ is called a *realization* of $\mathbf{H}(s)$. The realization of a transfer function is not unique; those of the smallest possible order n are called *minimal realizations*. A realization is minimal if it is completely *controllable* and *observable*. A descriptor system with $(\mathbf{A}, \mathbf{E})$ regular is *completely controllable* [4] if $\operatorname{rank}[\mathbf{A} - \lambda\mathbf{E}, \mathbf{B}] = n$, $\forall$ finite $\lambda \in \mathbb{C}$ and $\operatorname{rank}[\mathbf{E}, \mathbf{B}] = n$. It is called *completely observable* if $\operatorname{rank}\left[\mathbf{A}^T - \lambda\mathbf{E}^T, \mathbf{C}^T\right] = n$, $\forall$ finite $\lambda \in \mathbb{C}$ and $\operatorname{rank}\left[\mathbf{E}^T, \mathbf{C}^T\right] = n$, where $(\cdot)^T$ denotes transpose. The matrix pencil $(\mathbf{A}, \mathbf{E})$ is *regular* if the matrix $\mathbf{A} - \lambda\mathbf{E}$ is nonsingular for some finite $\lambda \in \mathbb{C}$. The *poles* are given by the eigenvalues of the matrix pencil $(\mathbf{A}, \mathbf{E})$: poles of $\Sigma = \lambda(\mathbf{A}, \mathbf{E})$. Σ is *stable* if all its finite poles are in the left-half plane: Σ stable $\Leftrightarrow Re(\lambda(\mathbf{A}, \mathbf{E})) < 0$ for $|\lambda(\mathbf{A}, \mathbf{E})| < \infty$.

Problem Statement: Our technique can model any kind of frequency domain data, but here we focus on scattering parameters. An LTI system Σ models a set with k *noisy* samples of a device with p ports $(f_i, \tilde{\mathbf{S}}^{(i)} = \mathbf{S}^{(i)} + \mathbf{N}^{(i)})$, $i = 1, \cdots, k$, where $f_i \in \mathbb{R}$ (the frequency where we measure), $\mathbf{S}^{(i)}$, $\mathbf{N}^{(i)} \in \mathbb{C}^{p\times p}$ (the true S-parameter and the noise matrix, respectively), if the value of the associated transfer function evaluated at $j \cdot 2\pi f_i$ is close to the noisy measurement: $\mathbf{H}(j\cdot 2\pi f_i) \approx \tilde{\mathbf{S}}^{(i)}$, $i = 1, \ldots, k$.

3 Tangential Interpolation

Tangential interpolation data consist of *right interpolation data* $(\lambda_i, \mathbf{r}_i, \mathbf{w}_i)$ and *left interpolation data* $(\mu_j, \ell_j, \mathbf{v}_j)$, where $\lambda_i \in \mathbb{C}$, $\mathbf{r}_i \in \mathbb{C}^{p\times 1}$, $\mathbf{w}_i \in \mathbb{C}^{p\times 1}$ and $\mu_j \in \mathbb{C}$, $\ell_j \in \mathbb{C}^{1\times p}$, $\mathbf{v}_j \in \mathbb{C}^{1\times p}$, for $i = 1, \ldots, k$ and $j = 1, \ldots, h$, or, more compactly,

$$\Lambda = \operatorname{diag}[\lambda_1, \cdots, \lambda_k] \in \mathbb{C}^{k\times k}, \qquad M = \operatorname{diag}[\mu_1, \cdots, \mu_h] \in \mathbb{C}^{h\times h}, \tag{2}$$

$$\mathbf{R} = [\mathbf{r}_1, \cdots, \mathbf{r}_k] \in \mathbb{C}^{m\times k}, \qquad \mathbf{L}^T = [\ell_1, \cdots, \ell_h] \in \mathbb{C}^{h\times p}, \tag{3}$$

$$\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_k] \in \mathbb{C}^{p\times k}, \qquad \mathbf{V}^T = [\mathbf{v}_1, \cdots, \mathbf{v}_h] \in \mathbb{C}^{h\times m}. \tag{4}$$

The quantities λ_i, μ_j are points where the function is evaluated, $\mathbf{r}_i$, ℓ_j are referred to as tangential directions on the right and on the left, while $\mathbf{w}_i$, $\mathbf{v}_j$ are right and left tangential data. Tangential values may be given as above, but most often matrix data, i.e., the value of a transfer function matrix at several points, is provided. In this case, tangential data can be obtained by following Sect. 4.1.

The rational interpolation problem for tangential data aims at finding a realization $[\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}]$, such that the associated transfer function satisfies the *right and left constraints* $\mathbf{H}(\lambda_i)\mathbf{r}_i = \mathbf{w}_i$, $\ell_j\mathbf{H}(\mu_j) = \mathbf{v}_j$. The key tools for studying this problem are the *Loewner matrix* together with the *shifted Loewner matrix* associated with the data; for the material in Sect. 3.1 and 3.2, we refer to [2] for details on proofs.

3.1 The Loewner and the Shifted Loewner Matrices

Given $Z = \{z_1, \cdots, z_P\}$, points in the complex plane, and $\{\mathbf{H}(z_1), \cdots, \mathbf{H}(z_P)\}$, the evaluation of a rational matrix function $\mathbf{H}(s)$ at those points, we partition $Z = \{\lambda_1, \ldots, \lambda_k\} \cup \{\mu_1, \ldots, \mu_h\}$, $k + h = P$, and obtain tangential data (2)-(4) from matrix data by selecting $\mathbf{r}_i$ and ℓ_j. We build the Loewner and shifted Loewner matrices as:

$$\mathbb{L} = \begin{bmatrix} \frac{\mathbf{v}_1\mathbf{r}_1 - \ell_1\mathbf{w}_1}{\mu_1 - \lambda_1} & \cdots & \frac{\mathbf{v}_1\mathbf{r}_k - \ell_1\mathbf{w}_k}{\mu_1 - \lambda_k} \\ \vdots & \ddots & \vdots \\ \frac{\mathbf{v}_h\mathbf{r}_1 - \ell_h\mathbf{w}_1}{\mu_h - \lambda_1} & \cdots & \frac{\mathbf{v}_h\mathbf{r}_k - \ell_h\mathbf{w}_k}{\mu_h - \lambda_k} \end{bmatrix}, \quad \sigma\mathbb{L} = \begin{bmatrix} \frac{\mu_1\mathbf{v}_1\mathbf{r}_1 - \lambda_1\ell_1\mathbf{w}_1}{\mu_1 - \lambda_1} & \cdots & \frac{\mu_1\mathbf{v}_1\mathbf{r}_k - \lambda_k\ell_1\mathbf{w}_k}{\mu_1 - \lambda_k} \\ \vdots & \ddots & \vdots \\ \frac{\mu_h\mathbf{v}_h\mathbf{r}_1 - \lambda_1\ell_h\mathbf{w}_1}{\mu_h - \lambda_1} & \cdots & \frac{\mu_h\mathbf{v}_h\mathbf{r}_k - \lambda_k\ell_h\mathbf{w}_k}{\mu_h - \lambda_k} \end{bmatrix}. \tag{5}$$

They satisfy the Sylvester equations $\mathbb{L}\Lambda - M\mathbb{L} = \mathbf{LW} - \mathbf{VR}$ and $\sigma\mathbb{L}\Lambda - M\sigma\mathbb{L} = \mathbf{LW}\Lambda - M\mathbf{VR}$ and have a system theoretic interpretation in terms of the tangential controllability and observability matrices (see [2]). For a thorough discussion on choosing the directions $\mathbf{r}_i$ and ℓ_j to ensure system identification, see [5].

3.2 The Solution to the General Tangential Interpolation Problem in the Loewner Framework

Next we review the conditions for the solution of the general tangential interpolation problem by means of state-space matrices $[\mathbf{E},\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D}]$, as presented in [2].

Lemma 1. *Assume that* $k = h$ *and that* $\det(x\mathbb{L} - \sigma\mathbb{L}) \neq 0$, *for all* $x \in \{\lambda_i\} \cup \{\mu_j\}$ *(i.e., the matrix pencil* $(\sigma\mathbb{L},\mathbb{L})$ *is regular and* $\mu_j, \lambda_i \notin \lambda\,(\sigma\mathbb{L},\mathbb{L})$*). Then* $\mathbf{E} = -\mathbb{L}$, $\mathbf{A} = -\sigma\mathbb{L}$, $\mathbf{B} = \mathbf{V}$, $\mathbf{C} = \mathbf{W}$ *and* $\mathbf{D} = \mathbf{0}$ *is a minimal realization of an interpolant of the data. Thus, the associated transfer function*

$$\mathbf{H}(s) = \mathbf{W}(\sigma\mathbb{L} - s\mathbb{L})^{-1}\mathbf{V} \tag{6}$$

satisfies the left ($\ell_j\mathbf{H}(\mu_j) = \mathbf{v}_j$*) and right interpolation conditions (*$\mathbf{H}(\lambda_i)\mathbf{r}_i = \mathbf{w}_i$*).*

This holds if measurements are not noisy and the number of samples are not more than needed. For strictly proper systems, the rank of the Loewner and shifted Loewner matrices is the order of the system (McMillan degree).

3.3 The Loewner Matrix Pencil under Noisy Data

Noisy frequency domain data is contained in the $\mathbf{V}$ and $\mathbf{W}$ data matrices, while Λ, M, $\mathbf{R}$ and $\mathbf{L}$ are assumed not affected by noise. We denote the matrices $\mathbf{V}$ and $\mathbf{W}$ containing noisy data by $\tilde{\mathbf{V}}$ and $\tilde{\mathbf{W}}$. Thus, the Loewner matrix will be affected by noise, but will still satisfy a Sylvester equation: $\tilde{\mathbb{L}}\Lambda - M\tilde{\mathbb{L}} = \mathbf{L}\tilde{\mathbf{W}} - \tilde{\mathbf{V}}\mathbf{R}$, so

$$\underbrace{\left(\tilde{\mathbb{L}} - \mathbb{L}\right)}_{\Delta\mathbb{L}}\Lambda - M\underbrace{\left(\tilde{\mathbb{L}} - \mathbb{L}\right)}_{\Delta\mathbb{L}} = \mathbf{L}\underbrace{\left(\tilde{\mathbf{W}} - \mathbf{W}\right)}_{\Delta\mathbf{W}} - \underbrace{\left(\tilde{\mathbf{V}} - \mathbf{V}\right)}_{\Delta\mathbf{V}}\mathbf{R} \tag{7}$$

Similarly, for the shifted Loewner matrix, we have that $(\Delta\sigma\mathbb{L})\Lambda - M(\Delta\sigma\mathbb{L}) = \mathbf{L}\,(\Delta\mathbf{W})\,\Lambda - M\,(\Delta\mathbf{V})\,\mathbf{R}$, where $\Delta\mathbb{L}$ and $\Delta\sigma\mathbb{L}$ are perturbations introduced by noise. Thus, the poles of the system recovered from noisy data, which are the generalized eigenvalues of the matrix pencil $\left(\widetilde{\sigma\mathbb{L}},\tilde{\mathbb{L}}\right) = (\sigma\mathbb{L} + \Delta\sigma\mathbb{L},\mathbb{L} + \Delta\mathbb{L})$, are perturbations of the original values, which are the eigenvalues of $(\sigma\mathbb{L},\mathbb{L})$.

4 Implementation

This section reviews an implementation approach which was introduced in [1]. We assume that data sets contain k samples of the multi-port S-parameters, $\mathbf{S}^{(i)}$, at frequency points $j\omega_i$, for $i = 1,\dots,k$.

4.1 Generating Tangential Data from Matrix Data for S-Parameters

To obtain a real system, we can choose the right interpolation data as the odd samples, together with their complex conjugates, and the left interpolation data, as the even samples, with their complex conjugates:

$$\left(j\omega_{2i-1}, -j\omega_{2i-1}; \mathbf{r}_i, \mathbf{r}_i; \mathbf{w}_i = \mathbf{S}^{(2i-1)}\mathbf{r}_i, \overline{\mathbf{w}}_i = \overline{\mathbf{S}^{(2i-1)}\mathbf{r}_i}\right) \tag{8}$$

$$\left(j\omega_{2i}, -j\omega_{2i}; \ell_i, \ell_i; \mathbf{v}_i = \ell_i\mathbf{S}^{(2i)}, \overline{\mathbf{v}}_i = \ell_i\overline{\mathbf{S}^{(2i)}}\right) \tag{9}$$

where $\omega_i = 2\pi f_i \in \mathbb{R}$, for $i = 1, \cdots, \frac{k}{2}$. The directions may be chosen as $\mathbf{r}_i = \mathbf{e}_m \in \mathbb{R}^{p\times 1}$, with $m = p$ for $i = p\cdot c_1$ and $m = 1, \cdots, p-1$ for $i = p\cdot c_1 + m$, for some $c_1 \in \mathbb{Z}$, where $\mathbf{e}_m$ is the m-th column of the identity matrix $\mathbf{I}_p$, and similarly for ℓ_i. Without loss of generality, we have assumed an even number of samples. The Loewner and shifted Loewner matrices are built using (5). A change of basis is performed to ensure real matrix entries: $\hat{\Lambda} = \hat{\Pi}^*\Lambda\hat{\Pi}$, $\hat{M} = \hat{\Pi}^*M\hat{\Pi}$, $\hat{\mathbf{L}} = \hat{\Pi}^*\mathbf{L}$, $\hat{\mathbf{V}} = \hat{\Pi}^*\mathbf{V}$, $\hat{\mathbf{R}} = \mathbf{R}\hat{\Pi}$, $\hat{\mathbf{W}} = \mathbf{W}\hat{\Pi}$ $\hat{\mathbb{L}} = \hat{\Pi}^*\mathbb{L}\hat{\Pi}$, $\sigma\hat{\mathbb{L}} = \hat{\Pi}^*\sigma\mathbb{L}\hat{\Pi}$, where

$$\hat{\Pi} = \text{blkdiag}\left[\Pi, \ldots, \Pi\right] \in \mathbb{C}^{k\times k},\ \Pi = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -j \\ 1 & j \end{bmatrix}.$$

4.2 SVD Implementation Approach

The idea is to use all measurements to construct the Loewner matrix pencil as presented in Sect. 4.1. Lemma 1 ensures perfect recovery of the system when the Loewner matrix pencil is regular and the measurements are noise-free. However, when too many measurements are available, the pencil is singular, so one needs to eliminate the singular part. Under the assumption that $\forall x \in \{\lambda_i\} \cup \{\mu_i\}$

$$\text{rank}\left(x\mathbb{L} - \sigma\mathbb{L}\right) =: r, \tag{10}$$

one can perform the singular value decomposition:

$$x\mathbb{L} - \sigma\mathbb{L} = \mathbf{YSX}^*, \tag{11}$$

where $\text{rank}\left(x\mathbb{L} - \sigma\mathbb{L}\right) = \text{rank}\left(\mathbf{S}\right) =: r$, $\mathbf{Y}, \mathbf{X} \in \mathbb{C}^{k\times r}$, where r is the *dimension of the regular part* of $x\mathbb{L} - \sigma\mathbb{L}$. For strictly proper systems, r is precisely n, the order of the underlying system, while for proper systems, $r = n + rank(\mathbf{D})$. Using the singular vectors as projectors, the realization [2] is given as $\mathbf{E} = -\mathbf{Y}^*\mathbb{L}\mathbf{X}$, $\mathbf{A} = -\mathbf{Y}^*\sigma\mathbb{L}\mathbf{X}$, $\mathbf{B} = \mathbf{Y}^*\mathbf{V}$, $\mathbf{C} = \mathbf{W}\mathbf{X}$ with $\mathbf{D} = \mathbf{0}$.

Nevertheless, real-world measurements are noisy, so the singular values and eigenvalues of the Loewner pencil are corrupted by noise. Next, we analyze the effects of noise on the performance of the SVD implementation approach.

5 Numerical Example

In the controlled experiments conducted, we introduce *random noise relative* to the entries in $\mathbf{S}^{(i)}$: $\mathbf{N}^{(i)} = \mathbf{S}^{(i)}. * 10^{-\mathrm{SNR}/10}(randn(p) + j \cdot randn(p))$, where $.*$ is the entry-by-entry multiplication in MATLAB.

The accuracy was assessed using two error measures:

- the normalized $\mathscr{H}_\infty$-norm of the error system:

$$\mathscr{H}_\infty \text{ error } = \max_{i=1\ldots k} \sigma_1 \left(\mathbf{H}(j\omega_i) - \tilde{\mathbf{S}}^{(i)}\right) / \max_{i=1\ldots k} \sigma_1 \left(\tilde{\mathbf{S}}^{(i)}\right), \tag{12}$$

- the normalized $\mathscr{H}_2$-norm of the error system:

$$\mathscr{H}_2 \text{ error } = \sqrt{\sum_{i=1}^{k} \left\|\mathbf{H}(j\omega_i) - \tilde{\mathbf{S}}^{(i)}\right\|_F^2 / \sum_{i=1}^{k} \left\|\tilde{\mathbf{S}}^{(i)}\right\|_F^2}, \tag{13}$$

For comparison, we compute the $\mathscr{H}_\infty$ and $\mathscr{H}_2$-norm errors of the noise:

- the $\mathscr{H}_\infty$-norm of the noise we introduced: $\mathscr{H}_\infty$ norm $= \max_{i=1\ldots k} \sigma_1(\mathbf{N}^{(i)})$,
- the $\mathscr{H}_2$-norm of the noise we introduced: $\mathscr{H}_2$ norm $= \sqrt{\sum_{i=1}^{k} \left\|\mathbf{N}^{(i)}\right\|_F^2}$.

5.1 A-Priori Known System

We consider an explicitly given system of order 14 with $p = 2$ ports. It has a-priori known matrices and we create $k = 134$ measurements of its transfer function on the imaginary axis between 10^{-1} and 10^{1} rad/sec, for different SNR values.

5.1.1 Noise-Free Case

When we artificially create measurements of the transfer function and do not introduce any noise, we can immediately identify the order of the system from the drop of the singular values of the Loewner and shifted Loewner matrices (Fig. 1(a)). Using the singular vectors associated to the non-zero singular values as projectors, the original system is recovered (Table 1).

5.1.2 SNR$=80$

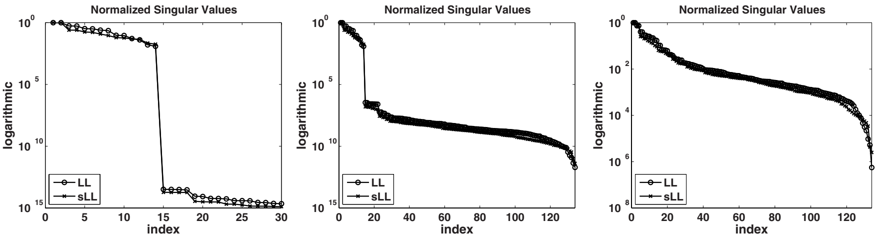
For small amounts of noise added to our measurements, we still notice a decay in the singular values of the Loewner matrix pencil (Fig. 1(b)) which suggests that the

Table 1 Results for noise-free measurements

$\mathscr{H}_\infty$ error	$\mathscr{H}_2$ error
6.9291e–14	1.6246e–14

Table 2 Results for SNR$=80$

	$\mathscr{H}_\infty$ error	$\mathscr{H}_2$ error
Model	1.8032e–7	1.5312e–7
Noise	5.5257e–8	2.7242e–7



(a) Noise-free (b) SNR= 80 (c) SNR= 20

Fig. 1 Singular value drop of the Loewner matrix pencil

underlying system is of order 14. Note that the singular values which used to be zero are now perturbed by noise. Table 2 shows the errors for the order 14 system built using the procedure in Sect. 4.2 when compared to the errors due to the noise.

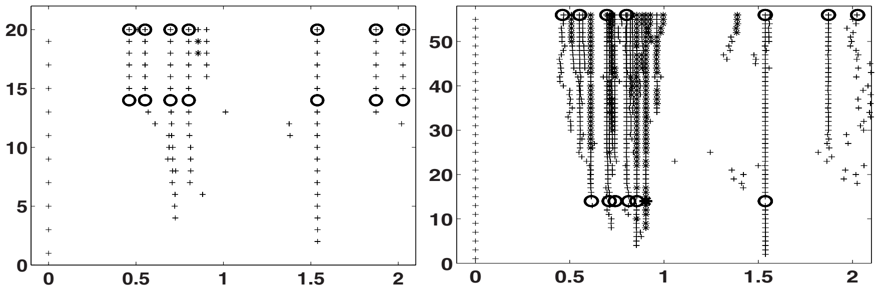
5.1.3 SNR= 20

When the added relative noise is large, the order of the underlying system can no longer be identified from the singular value decay of the Loewner pencil (Fig. 1(c)).

Thus, it is instructive to look at so called *stabilization diagrams* [6]. They show how the poles of the system evolve when incrementally increasing the order of the model (in our case, the order is given by the number of singular values retained after the SVD truncation). Fig. 2 shows the absolute value of the imaginary part of the poles (on the abscissa) versus the different truncation orders for SVD (on the ordinate). Pluses are stable poles, while stars are unstable poles. Moreover, we have highlighted as circles the poles which are obtained when truncating to order 14 (which is the true order of the system) and the original poles (the last set of circles).

For SNR= 80, the diagram (Fig. 2(a)) shows that the poles of the order 14 model approximate well the original ones. Once physical poles were identified, they are present for all subsequent orders higher than the dimension of the underlying system. For SNR= 20, the stabilization diagram in Fig. 2(b) is not as clear as before. Truncation at order 14 has occurred too early, since not all physical poles are present and it is necessary to go to order 30 and higher to obtain relatively good approximations of the original poles. This diagram suggests that it is necessary to construct a high order model from our noisy data. However, such a high-dimensional model would have unstable poles, as well as others which are non-physical, so a reduction step is necessary to eliminate the spurious part of the model.

In the civil and mechanical engineering communities, stabilization diagrams are built to reveal the physical poles as those which do not move too much in the complex plane with increasing the order of the model, as they are immune to perturbations. This is based on the intuition that noise in the data should not affect the underlying system poles. Therefore, stable poles with almost constant imaginary part for different model orders are good candidates for physical poles. However, for SNR= 20, some stable poles with imaginary part not changing are not physical.



(a) SNR= 80 (b) SNR= 20

Fig. 2 Stabilization diagrams

Thus, this criterion gives, apart from the physical poles, others which are which are spurious and try to fit the noise. This suggests looking for another criterion to detect the true poles of the underlying system.

Below, we investigate two alternatives. The first is to compute the 2-norm of the residue matrices of the high-order model and retain the poles with the largest residue norms. For descriptor systems, residues are computed as:

$$\mathrm{Res}_i = (\mathbf{C}\mathbf{x}_i)\,(\mathbf{y}^*\mathbf{E}\mathbf{x}_i)^{-1}\,(\mathbf{y}_i^*\mathbf{B}) \tag{14}$$

where $\mathbf{x}_i$ and $\mathbf{y}_i$ are the right and left eigenvectors associated to the eigenvalue λ_i of the matrix pencil $(\mathbf{A},\mathbf{E})$: $\mathbf{A}\mathbf{x}_i = \lambda_i\mathbf{E}\mathbf{x}_i$, $\mathbf{y}_i^*\mathbf{A} = \lambda_i\mathbf{y}_i^*\mathbf{E}$. The second is to preserve the most dominant poles, where dominance is measured by the quantities q_i defined as:

$$q_i = ||\mathrm{Res}_i||_2/\mathrm{Re}(\lambda_i) \tag{15}$$

Both criteria are motivated by the pole-residue expansion of the transfer function:

$$\mathbf{H}(s) = \sum_{i=1}^{n} \frac{\mathrm{Res}_i}{s-\lambda_i} \tag{16}$$

A pole with large residue norm and/or large dominance quantity contributes more to the response, while the rest do not influence it very much.

For SNR= 80, it is clear that the first 14 poles are the physical ones from the residue norms, as well as the dominance quantities (Table 3). For SNR= 20, the first 14 poles sorted according to the residue norms are approximations of the original ones, but if sorted according to the dominance criterion, this would not be the case (non-physical poles 19, 20 are more dominant than recovered poles 13, 14). Tables 3 and 4 also include a column for the minimum distance between each recovered pole and all original ones. All distances are in the range of the corresponding noise level. For SNR= 80, the shortest distances are between those poles with the largest residues (and most dominant) and the original poles. This is not true for SNR= 20, as the non-physical poles 19, 20 are closer than the recovered poles 3, 4. Table 5

Table 3 Results for SNR= 80

i	original poles λ_i	$\widetilde{\lambda}_i$	$\min_j \lvert\lambda_j - \widetilde{\lambda}_i\rvert$	$\|\mathrm{Res}_i\|_2$	q_i
1–2	–7.6155e–1±4.6137e–1i	–7.6155e–1±4.6137e–1i	3.0764e–8	1.6903e+0	2.2195e+0
3–4	–1.3389e+0±2.0269e+0i	–1.3389e+0±2.0269e+0i	9.6743e–8	1.3825e+0	1.0326e+0
5–6	–1.5742e–1±5.5226e–1i	–1.5742e–1±5.5226e–1i	3.9031e–9	2.6748e–1	1.6992e+0
7–8	–8.2952e–2±1.8726e+0i	–8.2952e–2±1.8726e+0i	1.2044e–9	1.0346e–1	1.2472e+0
9–10	–2.1923e–2±6.9690e–1i	–2.1923e–2±6.9690e–1i	3.8541e–10	3.8603e–2	1.7609e+0
11–12	–2.2294e–3±1.5371e+0i	–2.2294e–3±1.5371e+0i	1.3721e–10	1.0555e–2	4.7342e+0
13–14	–9.9863e–3±8.0143e–1i	–9.9863e–3±8.0143e–1i	2.7708e–10	8.7155e–3	8.7275e–1
15–16		1.5810e–4±6.1158e–1i	8.8127e–2	2.8435e–11	1.7985e–7
17–18		-4.5355e–5±9.0452e–1i	1.0357e–1	2.2892e–12	5.0474e–8
19–20		3.1624e–5±8.5534e–1i	5.4833e–2	8.5254e–13	2.6959e–8

Table 4 Results for SNR= 20

i	original poles λ_i	$\widetilde{\lambda}_i$	$\min_j \lvert\lambda_j - \widetilde{\lambda}_i\rvert$	$\|\mathrm{Res}_i\|_2$	q_i
1–2	–7.6155e–1±4.6137e–1i	–7.6614e–1±4.5309e–1i	9.4636e–3	1.7541e+0	2.2895e+0
3–4	–1.3389e+0±2.0269e+0i	–1.3253e+0±1.9934e+0i	3.6186e–2	1.3865e+0	1.0461e+0
5–6	–1.5742e–1±5.5226e–1i	–1.5794e–1±5.5379e–1i	1.6158e–3	2.6732e–1	1.6925e+0
7–8	–8.2952e–2±1.8726e+0i	–8.1729e–2±1.8702e+0i	2.6866e–3	1.0240e–1	1.2530e+0
9–10	–2.1923e–2±6.9690e–1i	–2.1764e–2±6.9738e–1i	5.0648e–4	3.7752e–2	1.7346e+0
11–12	–2.2294e–3±1.5371e+0i	–2.3526e–3±1.5370e+0i	1.5424e–4	1.0786e–2	4.5848e+0
13–14	–9.9863e–3±8.0143e–1i	–9.8073e–3±8.0154e–1i	2.1396e–4	8.4846e–3	8.6513e–1
15–16		–3.7258e–3±5.9846e–1i	1.0010e–1	6.4623e–4	1.7344e–1
17–18		3.9288e–3±7.6797e–1i	3.6232e–2	3.6728e–4	9.3486e–2
19–20		–1.3289e–4±7.9193e–1i	1.3682e–2	2.7105e–4	2.0397e+0

Table 5 Results for SNR= 20

Model	$\mathscr{H}_\infty$ error	$\mathscr{H}_2$ error
Order 14 after SVD truncation	5.0048e–1	4.3900e–1
Order 56 after SVD truncation	8.0286e–2	3.3064e–2
Order 14 after largest residue selection	2.4592e–2	2.4465e–2
Noise	5.1320e–2	2.6980e–1

presents the errors after the SVD truncation of order 14, as well as after the 14 poles with the largest residues were chosen from a model of order 56.

We looked at the pseudospectra of the original matrix and compared the location of the poles given by our improved algorithm for SNR= 20 to the pseudospectra bounds corresponding to perturbations in the range $10^{-1.6} - 10^{-2}$ (Fig. 3, generated using EigTool [7]). Pseudospectra can be defined as [8]

$$\Lambda_\varepsilon(\mathbf{A}) = \{z \in \mathbb{C} : z \in \Lambda(\mathbf{A}+\mathbf{P}) \text{ for some } \mathbf{P} \text{ with } ||\mathbf{P}|| \leq \varepsilon\} \qquad (17)$$

The above definition is not for matrix pencils. Still, the poles recovered by selecting those with the largest residues of a high order system are enclosed by the circle in

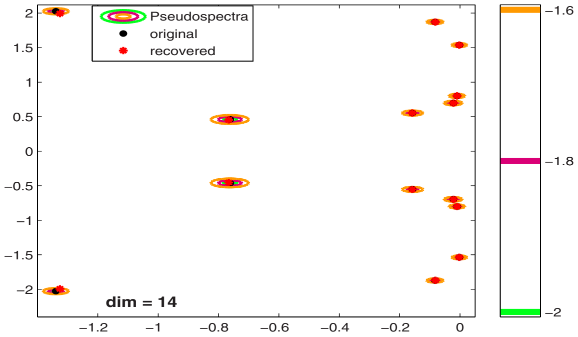
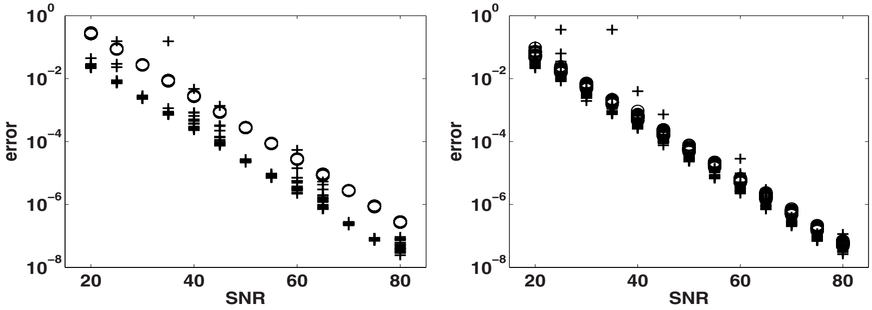


Fig. 3 Pseudospectra of the **A** matrix (contours correspond to levels of ε of $10^{-1.6}$, $10^{-1.8}$, 10^{-2})



(a) $\mathscr{H}_2$ error results

(b) $\mathscr{H}_\infty$ error results

Fig. 4 Results for different SNR values (the circles are the norms of the noise, while the pluses are the norms of our models)

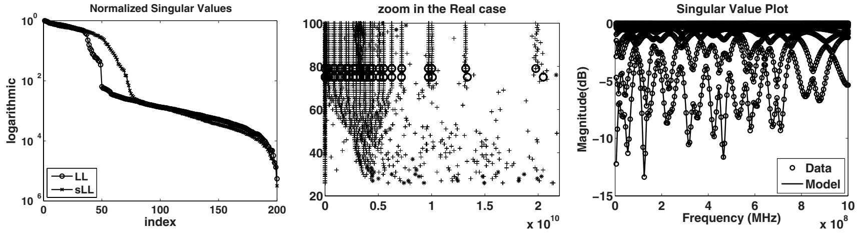
the complex plane which bounds the original poles, perturbed by adding matrices **P** with norm up to 10^{-2} to the matrix $\mathbf{E}^{-1}\mathbf{A}$.

We performed Monte Carlo analyses for 20 different random perturbations for each SNR value. The $\mathscr{H}_2$ and $\mathscr{H}_\infty$ error results are shown in Fig. 4. We notice that, with a few exceptions, the $\mathscr{H}_2$-norm errors for our models are well below the values of the noise, while the $\mathscr{H}_\infty$-norm errors are close to those of the noise.

5.2 *Example Involving Measurements*

Measurements were performed using a vector network analyzer (VNA) and were provided by CST AG in magnitude-angle format (with at most 9 significant digits for the magnitude and at most 6 significant digits for the angle). This data set contains $k = 200$ samples between 5MHz and 1GHz of a device with 26 ports.

The singular value drop of $x\mathbb{L} - \sigma\mathbb{L}$ (Fig. 5(a)), where $x = 2\pi f_1$, does not reveal the order of the underlying system. Thus, we proceed with building the stabilization



(a) SVD drop of the Loewner matrix pencil (b) Stabilization diagram (c) Frequency response

Fig. 5 Plots for the data set obtained from a device with 26 ports

diagram (Fig. 5(b)). This shows that, by order 75, all poles are approximated, but as the order is increased, the estimates start converging. Thus, the order 79 model provides good approximations of the poles, so we apply the largest residues and dominance criteria to trim the 75 physical poles from the rest. The first circles are the poles at order 75, while the second set of circles are the 75 most dominant poles of the order 79 system. Both criteria indicate that indeed, the underlying system is of order 75 (Table 6). The first 26 poles, some of which are unstable, in fact correspond to poles at infinity. Table 7 shows the error norms.

Table 6 Dominance quantities

i	$\widetilde{\lambda}_i$	$\|\text{Res}_i\|_2$	q_i
1	1.8641e+13	1.0204e+14	5.4738e+0
2	5.5859e+12	1.0101e+14	1.8083e+1
3–4	3.8530e+12±1.7468e+12i	4.6061e+13	1.1954e+1
⋮	⋮	⋮	⋮
71	-1.9983e+8	2.7662e+8	1.3843e+0
72–73	−1.8893e±8 +2.9192e+9i	2.5603e+8	1.3551e+0
74–75	−1.6546e+8±7.8937e+8i	2.1840e+8	1.3200e+0
76–77	2.5402e+7±3.8980e+9i	1.3492e+6	5.3115e–2
78–79	–3.1911e+6±3.4011e+9i	9.6859e+5	3.0353e–1

Table 7 Error norm results

Model	$\mathcal{H}_\infty$ error	$\mathcal{H}_2$ error
Order 75	1.9513e–2	1.9859e–3
Order 79	9.8777e–3	8.7537e–4
Dominant 75	7.2571e–3	8.5974e–4

6 System Identification from Noisy Time Domain Data

6.1 Noise-Free Case

In the time domain, we are interested in identifying a system given samples of its impulse response, $\mathbf{h}(t) := \mathbf{C}e^{\mathbf{A}t}\mathbf{B} + \mathbf{D}\delta(t)$. For simplicity, assume $\mathbf{E} = \mathbf{I}$ and $\mathbf{D} = 0$.

Suppose the time axis is uniformly spaced with time step Δt and the impulse response samples are $\{\mathbf{h}_i; i = 0, \ldots, 2k\}$, with $\mathbf{h}_i := \mathbf{h}(i\Delta t)$. Let the matrix $\mathbf{M} := e^{\mathbf{A}\Delta t}$, then $\mathbf{h}_i = \mathbf{C}\left(e^{\mathbf{A}\Delta t}\right)^i \mathbf{B} = \mathbf{C}\mathbf{M}^i\mathbf{B}$. Thus, the uniform sampling of the impulse response of the continuous-time system $[\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}]$ is equivalent with measuring Markov parameters of the discrete-time system $[\mathbf{M}, \mathbf{B}, \mathbf{C}, \mathbf{D}]$.

Realization from Markov parameters [9][10] relies on the Hankel matrix $\mathscr{H}$ and the shifted Hankel matrix $\sigma\mathscr{H}$, defined as

$$\mathscr{H} := \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \dots & \mathbf{h}_k \\ \mathbf{h}_2 & \mathbf{h}_3 & \cdots & \mathbf{h}_{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{h}_k & \mathbf{h}_{k+1} & \dots & \mathbf{h}_{2k-1} \end{bmatrix}, \sigma\mathscr{H} := \begin{bmatrix} \mathbf{h}_2 & \mathbf{h}_3 & \dots & \mathbf{h}_{k+1} \\ \mathbf{h}_3 & \mathbf{h}_4 & \dots & \mathbf{h}_{k+2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{h}_{k+1} & \mathbf{h}_{k+2} & \dots & \mathbf{h}_{2k} \end{bmatrix}. \tag{18}$$

In [11], it was shown that these are a special case of Loewner matrices constructed from derivatives of $\mathbf{H}(s^{-1})$ at $s = 0$, where $\mathbf{H}(s)$ is the system transfer function.

From the short SVD of the Hankel matrix, $\mathscr{H} = \mathbf{Y}\mathbf{S}\mathbf{X}^T$, with $\mathbf{Y} \in \mathbb{R}^{k\times r}$, $\mathbf{S} \in \mathbb{R}^{r\times r}$, $\mathbf{X} \in \mathbb{R}^{r\times k}$, and $r := \text{rank}\mathscr{H}$, a minimal realization can be written down as $\mathbf{E}_\mathbf{M} = \mathbf{Y}^T\mathscr{H}\mathbf{X}, \mathbf{A}_\mathbf{M} = \mathbf{Y}^T\sigma\mathscr{H}\mathbf{X}, \mathbf{B}_\mathbf{M} = \mathbf{Y}^T\mathscr{H}(:,1), \mathbf{C} = \mathscr{H}(1,:)\mathbf{X}$ and $\mathbf{D} = 0$, leading to $\mathbf{M} = \mathbf{E}_\mathbf{M}^{-1}\mathbf{A}_\mathbf{M}, \mathbf{B} = \mathbf{E}_\mathbf{M}^{-1}\mathbf{B}_\mathbf{M}, \mathbf{C}, \mathbf{D}$.

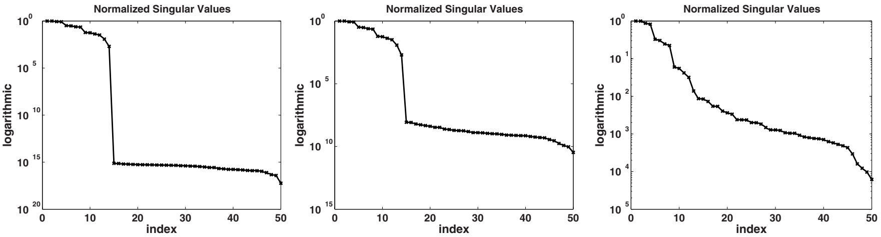
6.2 Noisy Measurements

Consider each impulse response measurement to be corrupted by noise $\{\widetilde{\mathbf{h}}_i := \mathbf{h}_i + \mathbf{n}_i; i = 0, \dots, 2k\}$ and construct the noisy Hankel matrices $\widetilde{\mathscr{H}}$ and $\sigma\widetilde{\mathscr{H}}$. A large order realization is given by $\widetilde{\mathbf{E}} = \widetilde{\mathscr{H}}, \widetilde{\mathbf{A}} = \sigma\widetilde{\mathscr{H}}, \widetilde{\mathbf{B}} = \widetilde{\mathscr{H}}(:,1), \widetilde{\mathbf{C}} = \widetilde{\mathscr{H}}(1,:)$, with poles $\widetilde{\lambda}_i, i = 1, \dots, k$. From this realization we want to keep only the n poles that best match the poles of the original system $\lambda_i, i = 1, \dots, n$.

We proceed with a numerical experiment where we control the noise level. We consider a system of order $n = 14$, compute the impulse response samples $\mathbf{h}_i$ for $i = 1, \dots, 100$, and then add noise $\mathbf{n}_i$ with different SNR values.

6.2.1 SNR $= 80$

For a low noise level (Fig. 6(b)), there is a significant drop between the 14th to 15th singular values, indicating that the order of the original system is 14. Thus,



(a) Noise-free (b) SNR= 80 (c) SNR= 20

Fig. 6 Singular value drop of the Hankel matrix

Table 8 Results for SNR= 80

i	λ_i	$\widetilde{\lambda}_i$	q_i	$\lvert\mathrm{Res}_i\rvert$	$\min_j \left\lvert\lambda_j - \widetilde{\lambda}_i\right\rvert$
1–2	–8.2952e–2 ±1.8726e+0i	–8.2952e–2 ±1.8726e+0i	1.9187e+0	1.5916e–1	7.7569e–10
3–4	–2.1923e–2 ±6.9690e–1i	–2.1923e–2 ±6.9690e–1i	1.6042e+0	3.5168e–2	6.1002e–9
5–6	–7.6155e–1 ±4.6137e–1i	–7.6155e–1 ±4.6137e–1i	1.4909e+0	1.1354e+0	4.3122e–7
7–8	–1.5742e–1 ±5.5226e–1i	–1.5742e–1 ±5.5226e–1i	1.3064e+0	2.0566e–1	1.4390e–8
9–10	–9.9863e–3 ±8.0143e–1i	–9.9863e–3 ±8.0143e–1i	1.1260e+0	1.1245e–2	7.5328e–9
11–12	–2.2294e–3 ±1.5371e+0i	–2.2294e–3 ±1.5371e+0i	8.3503e–1	1.8616e–3	1.7025e–9
13–14	–1.3389e+0 ±2.0269e+0i	–1.3389e+0 ±2.0269e+0i	6.0088e–1	8.0454e–1	4.4483e–7
15–16		–1.2608e–1 ±1.8058e+0i	3.7960e–8	4.7858e–9	7.9480e–2
17–18		–2.8248e–1 ±3.2369e+0i	3.3928e–8	9.5842e–9	1.3789e+0
19–20		–8.5192e–1 ±5.5539e+0i	2.1474e–8	1.8294e–8	3.5604e+0

Table 9 Results for SNR= 20

i	λ_i	$\widetilde{\lambda}_i$	q_i	$\lvert\mathrm{Res}_i\rvert$	$\min_j \left\lvert\lambda_j - \widetilde{\lambda}_i\right\rvert$
1–2	–9.9863e-3 ±8.0143e–1i	–4.5898e–3 ±7.9554e–1i	2.2014e+0	1.0104e–2	7.9866e–3
3–4	–8.2952e-2 ±1.8726e+0i	–8.2818e–2 ±1.8716e+0i	1.9343e+0	1.6019e–1	1.0257e–3
5–6	–1.5742e–1 ±5.5226e–1i	–1.6886e–1 ±5.6579e–1i	1.6174e+0	2.7312e–1	1.7716e–2
7–8	–2.1923e–2 ±6.9690e–1i	–2.2319e–2 ±6.8954e–1i	1.4795e+0	3.3021e–2	7.3651e–3
9–10	–7.6155e–1 ±4.6137e–1i	–5.3492e–1 ±5.0979e–1i	1.3066e+0	6.9894e–1	2.3175e–1
11–12	–2.2294e–3 ±1.5371e+0i	–2.5078e–3 ±1.5388e+0i	7.3604e–1	1.8459e–3	1.6648e–3
13–14	–1.3389e+0 ±2.0269e+0i	-1.0990e+0 ±1.8497e+0i	5.9890e–1	6.5819e–1	2.9830e–1
15–16		–1.1619e–1 ±1.8235e+0i	5.6860e–2	6.6065e–3	5.9323e–2
17–18		–2.6236e–1 ±3.2348e+0i	2.9022e–2	7.6143e–3	1.3739e+0
19–20		–8.4679e–1 ±5.5539e+0i	2.1411e–2	1.8131e–2	3.5611e+0

the system can be easily recovered from the SVD. Also, note that the zero singular values from the noise-free case are now in the order of the SNR.

Alternatively, one can recover the original system from an analysis of the dominant poles. In Table 8, the poles of the approximant $\widetilde{\lambda}_i$ are sorted according to their dominance measure q_i. Note that between q_{14} and q_{15} there is a large jump, roughly in the order of $1/\mathrm{SNR}$. This is also present between $|\mathrm{Res}_{14}|$ and $|\mathrm{Res}_{15}|$. Therefore, both q_i and the residues show that the order of the original system is $n = 14$. In the last column of Table 8 we compute the distance from $\widetilde{\lambda}_i$ to the closest λ_i and, indeed, we notice that the 14 most dominant poles $\widetilde{\lambda}_i$ are the best approximants to λ_i.

6.2.2 SNR $= 20$

As the noise level increases to SNR $= 20$, the decay of the singular values (Fig. 6(c)) does not reveal the order of the original system. Additionally, looking only at the residues for this high level of noise does not reveal as much information as before. Nevertheless, we can still say something about the approximate poles $\widetilde{\lambda}_i$. From the last column of Table 9, as in the SNR $= 80$ case, we notice that the 14 most dominant poles $\widetilde{\lambda}_i$ are also the ones closest to the original poles λ_i.

7 Conclusion

Through a controlled experiment, we investigated the robustness of the SVD implementation in the Loewner matrix framework as it was initially proposed in [1, 2] using frequency and time domain data. As it is, the procedure exhibits poor performance for high noise levels, so we gained some insight into the pole evolution with increasing model orders from the stabilization diagrams, which suggested an improvement. Sorting the poles of a system of high order (chosen such that the poles have converged on the stabilization diagram) decreasingly according to the residue norm or the dominance measure, leads to the first ones being approximations of the physical poles. The distances between the recovered poles and the original ones are in the order of the noise level introduced in the data and, moreover, the approximated poles are within appropriate pseudospectra bounds corresponding to the noise level.

References

1. Lefteriu, S., Antoulas, A.C.: A new approach to modeling multi-port systems from frequency domain data. IEEE Trans. Comput.-Aided Design Integr. Circuits Syst. 29(1), 14–27 (2010)
2. Mayo, A.J., Antoulas, A.C.: A framework for the solution of the generalized realization problem. Linear Algebra and its Applications 405, 634–662 (2007)
3. Gustavsen, B., Semlyen, A.: Rational approximation of frequency domain responses by vector fitting. IEEE Trans. Power Del. 14, 1052–1061 (1999)
4. Yip, E., Sincovec, R.: Solvability, controllability, and observability of continuous descriptor systems. IEEE Trans. Autom. Control 26, 702–707 (1981)
5. Lefteriu, S., Antoulas, A.: Topics in model order reduction with applications to circuit simulation. In: Proc. Model Reduction for Circuit Simulation Workshop. Springer, Heidelberg (2009) (under review)
6. Peeters, B., der Auweraer, H.V., Guillaume, P., Leuridan, J.: The polymax frequency-domain method: a new standard for modal parameter estimation? Sound and Vibration 11, 395–409 (2004)
7. Wright, T.G.: EigTool (2002), http://www.comlab.ox.ac.uk/projects/pseudospectra/eigtool
8. Trefethen, L.N., Embree, M.: Spectra and Pseudospectra: the Behaviour of Nonnormal Matrices and Operators. SIAM, Philadelphia (2005)
9. Ho, B., Kalman, R.: Effective construction of linear, state-variable models from input/output function. Regelungstechnik 14, 545–548 (1966)
10. Silverman, L.: Realizations of linear dynamical systems. IEEE Trans. Autom. Control 16, 554–567 (1971)
11. Antoulas, A.C., Anderson, B.D.O.: On the scalar rational interpolation problem. IMA J. of Mathematical Control and Information 3, 61–88 (1986)

Blind Identification of Polynomial Matrix Fraction with Applications

Kenji Sugimoto

Abstract. Fractional representation by polynomial matrices is a tool for describing linear dynamical systems, often providing a unique insight into the systems. It has turned out that the coefficient matrices in this representation can be identified *without knowing the input data*, under some statistic assumptions. This is an outcome by combining system theory with a recent progress in signal processing; i.e., a methodology generically called independent component analysis.

This article summarizes this technique of blind system identification. Restricted optimization for parameter estimation plays a key role. Also presented are some of its applications in various fields such as input distortion compensation, disturbance suppression, and time series prediction in financial engineering.

1 Introduction

The author of this article was the first undergraduate student supervised by Professor Yutaka Yamamoto, just after he received the PhD in Florida. Since then, he has guided the author to the world of linear system theory, particularly for MIMO (Multi-Input Multi-Output) finite-dimensional systems. The master and doctor theses were respectively about realization by polynomial matrices and about a polynomial matrix approach to the inverse problem of LQ optimal control. Even now, system representation theory continues to be a resource of various research topics to the author, as will be seen below. This article is indebted to Professor Yamamoto in this sense.

Given input/output data, MIMO system identification is now established by means of state-space representation. Yet this article raises the issue of identifying polynomial matrix fraction *without knowledge of input data*. This can be carried

Kenji Sugimoto
Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma-city,
Nara, 630-0192, Japan
e-mail: kenji@is.naist.jp

J.C. Willems et al. (Eds.): Persp. in Math. Sys. Theory, Ctrl., & Sign. Pro., LNCIS 398, pp. 379–388.
springerlink.com

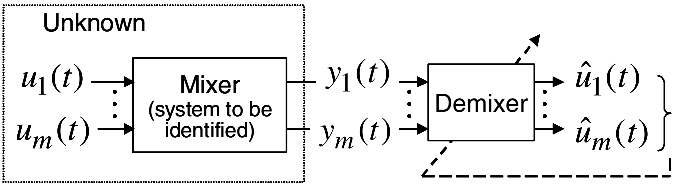


Fig. 1 Basic notion of blind signal deconvolution; y is observed and $\hat{u}$ is recovered.

out in terms of Blind Signal Deconvolution (BSD), which has recently attracted much attention in signal processing. BSD is to recover (i.e., deconvolute) multiple source signals from observation of their convolutive mixture and is applied to various fields such as acoustics, biosignal processing; see, e.g., [1, 2] and the references therein. The entire methodology is also referred to as Independent Component Analysis (ICA), since deconvolution is mainly achieved based on statistic independence of the source signals.

Fig. 1 shows a basic configuration of BSD. The source signal $u(t) := (u_1, \cdots, u_m)'$ passes through a MIMO dynamical system (called mixer) so that the components of $y(t) := (y_1, \cdots, y_m)'$ are convolutive mixtures of those of $u(t)$. A marked feature in this problem is that both the input $u(t)$ and the system (mixer) are unknown: only its output $y(t)$ is available for us. We have to adjust the demixer so as to obtain an estimate $\hat{u}(t) := (\hat{u}_1, \cdots, \hat{u}_m)'$ of $u(t)$, according to its statistic assumptions.

In signal processing, their main concern is by nature signal recovery itself, while in system theory we are more interested in the systems. When deconvolution is achieved, the demixer plays a role of inverse system of the mixer, in a sense. It is then natural to expect that the mixer is identified blindly by using the obtained knowledge. Since existing BSD methods are not necessarily efficient for this purpose, the author's group has proposed another method based on polynomial matrix fraction; first for the demixer with constant numerator matrix [3], then for a general case [4]. This BSD-based system identification can be applied to a wide variety of control issues such as input distortion compensation, disturbance suppression, change detection, and time series prediction [5, 6, 7]. This article is a short summary of these works.

The remainder of this article is organized as follows: After giving a brief review of BSD in Sect. 2, we describe the method for blind identification of polynomial matrix fraction in Sect. 3. Then we introduce some of its applications in Sect. 4. Sect. 5 concludes the article.

Transposition of vector v is denoted by v'. By I and O we denote respectively the identify matrix and zero matrix of appropriate sizes.

2 Review of BSD

We start by describing the properties of our systems and signals, and then give a brief review of conventional BSD methods.

2.1 System Description

As a mixer in Fig. 1, consider the polynomial matrix (left) fraction of the discrete-time transfer matrix $G(z) = D(z)^{-1}N(z)$, where

$$D(z) = I + D_1 z^{-1} + \cdots + D_p z^{-p} \quad \text{and} \quad N(z) = N_0 + N_1 z^{-1} + \cdots + N_q z^{-q}. \tag{1}$$

This is equivalent to the input/output relation on the time-domain:[1]

$$y(t) + \sum_{i=1}^{p} D_i y(t-i) = N_0 u(t) + \sum_{j=1}^{q} N_j u(t-j), \quad t = 0, 1, \cdots, T. \tag{2}$$

We assume that D_i and N_j are $m \times m$ constant matrices with $m > 1$ and N_0 is invertible. We also assume that both $N(z)$ and $D(z)$ are Hurwitz; i.e., $\det D(z) \neq 0$ and $\det N(z) \neq 0$ whenever z is in the instability region. Namely we treat square, stable, minimum phase, and biproper systems.

The class of the above systems is denoted as VARMA(p,q), standing for the Vector AutoRegressive Moving Average model. Our objective is to estimate the coefficients D_i and N_j for given p and q.

Remark 1. The pair $(D(z), N(z))$ in Eq. (1) is *not* necessarily left coprime, hence there remains a certain redundancy in the coefficients. This would be avoided if we took the more compact form

$$D(z) = \begin{pmatrix} d_1(z) \\ \vdots \\ d_m(z) \end{pmatrix}, \quad d_k(z) = e_k + d_k{}^1 z^{-1} + \cdots + d_k{}^{p_k} z^{-p_k}, \ \ k = 1, \cdots, m, \tag{3}$$

where e_k is the k-th row vector of the identity matrix I, and $p_1, \cdots, p_m$ are the observability indices of a minimal realization of the system [4]. The form (3) coincides with Eq. (1) iff $p_1 = \cdots = p_m = p$. But it is difficult to know the indices in advance, particularly when they are not equal, in addition to the notational complexity of Eq. (3). We thus adopt Eq. (1) in this article, at the expense of coprimeness. This does not lose any generality, but obtained parameters are not unique.

2.2 Signal Properties

We assume that the input signal $u = (u_1, \cdots, u_m)'$ is a random vector having the following properties, often called i.i.d (independent and identically distributed):

1. the components have zero mean: $\mathscr{E}[u] = 0$, where $\mathscr{E}$ stands for expectation;
2. their distributions are identical throughout the interval $[0, T]$ and at most one is Gaussian;

[1] We consider the finite time-interval $[0, T]$ for T large enough.

3. the components are spacially independent: $f(u) = \prod_{i=1}^{m} f_i(u_i)$, where f and f_i $(i = 1, \cdots, m)$ are respectively the joint and marginal probability density functions;
4. $u(t)$ is also temporally independent.

We assume that $y(t), t = 0, \cdots, T$ is available while $u(t)$ is not.

2.3 Solution for Instantaneous Mixture

We first consider the special case where $p = q = 0$; i.e., the mixer is a constant matrix $A := N_0$ (called instantaneous mixture). Let us take as the demixer a constant matrix W as well. Then, given input sequence in Eq. (2), the Kullback Leibler divergence of $\hat{u}(t) = Wy(t)$ is a function of W:

$$\mathscr{J}(\hat{u}; W) := \int f(\hat{u}) \log \frac{f(\hat{u})}{\prod_{i=1}^{m} f_i(\hat{u}_i)} d\hat{u}.$$

By definition $\mathscr{J}(\hat{u}; W) = 0$ iff $\hat{u}$ are spacially independent. In order to minimize this cost, Amari *et al.* [8] have proposed a learning law based on what they call natural gradient:

$$W \leftarrow W + \alpha \Delta W, \quad \Delta W = \left(I - \mathscr{E}[\psi(\hat{u})\hat{u}']\right) W \tag{4}$$

where $\alpha > 0$ is a small number and ψ is the vector of score functions

$$\psi(\hat{u}) = (\psi_1(\hat{u}_1), \cdots, \psi_m(\hat{u}_m))', \quad \psi_i(x) = -\frac{d \log f_i(x)}{dx} \quad \text{for } i = 1, \cdots, m. \tag{5}$$

The score functions are usually unknown but it is sufficient to approximate them by either x^3 or $\tanh x$.

It has been already shown that if $\mathscr{J} = 0$ is attained after adaptation, then the cascade satisfies

$$WA = P\Lambda \tag{6}$$

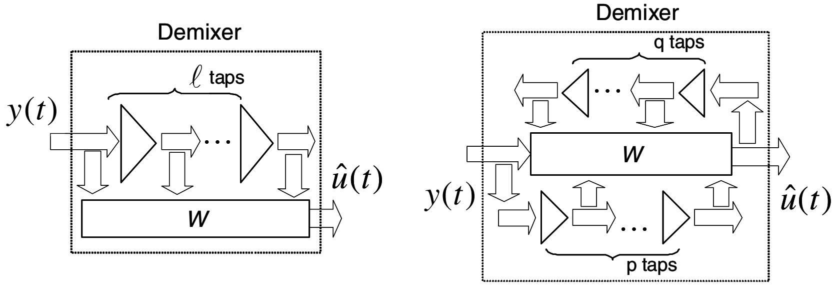
for some permutation matrix P and some diagonal matrix Λ such that $\det \Lambda \neq 0$. This means that $\hat{u}$ coincides with u up to scale and permuting indeterminacies on its components. Signal recovery in this sense is thereby achieved [1, 2].

2.4 Convolutive Mixture

Now we are ready to consider the original mixer (1) with $p > 0$, which is far more difficult than the instantaneous mixture in Sect. 2.3. A well-known method is to provide a demixer as

$$H(z) = H_0 + H_1 z^{-1} + \cdots + H_\ell z^{-\ell}, \quad W = (H_0, \cdots, H_\ell) \tag{7}$$

See the left-hand side of Fig. 2, where triangulars represent one-step time delay elements. Since the mixer is stably invertible, this achieves signal deconvolution



FIR filter (conventional)

Filter with feedback (proposed)

Fig. 2 BSD methods: conventional and proposed

in satisfactory accuracy for ℓ large enough (called FIR filter approximation). For our purpose of blind identification, however, the demixer often becomes too high-dimensional in this method. A frequency domain method is also well known for signal recovery. We will give yet another method in the next section.

3 Proposed Method (VARMA-ICA)

Instead of Eq. (7), let us consider $H(z) = \tilde{N}(z)^{-1}\tilde{D}(z)$, where

$$\tilde{N}(z) = I + \tilde{N}_1 z^{-1} + \cdots + \tilde{N}_q z^{-q} \quad \text{and} \quad \tilde{D}(z) = \tilde{D}_0 + \tilde{D}_1 z^{-1} + \cdots + \tilde{D}_p z^{-p}, \tag{8}$$

namely, a VARMA (q,p) model, and adjust their coefficients; see the right-hand side of Fig. 2.[2] The number of parameters is equal to that in the mixer in Eq. (1), so that we can expect more efficient learning. In fact, the parameters converge with much fewer samples than the conventional one in Sect. 2.4. This is advantageous especially in control engineering since the frequency band we treat is often lower (hence we can obtain sample data more slowly) than other applications such as acoustics.

BSD in this configuration is achieved as follows. First, we rewrite the demixer Eq. (8) as $\hat{U}(t) = WY(t)$, where

$$\hat{U}(t) = \begin{pmatrix} \tilde{u}(t) \\ \tilde{y}(t) \\ \hat{u}(t) \end{pmatrix}, \; Y(t) = \begin{pmatrix} \tilde{u}(t) \\ \tilde{y}(t) \\ y(t) \end{pmatrix}, \; \tilde{u}(t) = \begin{pmatrix} \hat{u}(t-q) \\ \vdots \\ \hat{u}(t-1) \end{pmatrix}, \; \tilde{y}(t) = \begin{pmatrix} y(t-p) \\ \vdots \\ y(t-1) \end{pmatrix},$$

$$W = \begin{pmatrix} I & O & O \\ O & I & O \\ -\left(\tilde{N}_q \cdots \tilde{N}_1\right) & \left(\tilde{D}_p \cdots \tilde{D}_1\right) & \tilde{D}_0 \end{pmatrix}. \tag{9}$$

[2] This does not give a minimal state-space realization, because it has redundant delay elements.

Thus our problem is reduced to the instantaneous mixture case superficially. The difference from Sect. 2.3 is that the entries in W are fixed except for the last block-row, and that some of the components of the vector $\hat{U}(t)$ can not be independent. These violate the assumption in Sect. 2.

To solve this problem, we adopt the following constrained optimization:

$$W \leftarrow W + \alpha \Delta W, \quad \Delta W = \Pi \left(I - \mathscr{E}[\psi(\hat{U})\hat{U}\,']\right) W \tag{10}$$

where $\Pi = \text{block-diag}(O, \cdots, O, I)$ and $\alpha > 0$ is a small number.

As in Sect. 2.3, the solution has the indeterminacy shown by Eq. (6). To avoid this, it has been proposed in [4] to add a further constraint on $N_0 = (n_{ij}^0)$:

$$n_{ii}^0 = 1, \quad |n_{ij}^0| < 1 \quad \text{for } i, j = 1, \cdots, m, \; i \neq j. \tag{11}$$

If the above learning law converges to the global optimum under this constraint, then the obtained demixer $H(z)$ gives the inverse system:

$$H(z)G(z) = I, \quad N(z) = \tilde{D}_0^{-1}\tilde{N}(z), \quad \text{and} \quad D(z) = \tilde{D}_0^{-1}\tilde{D}(z).$$

The parameter in (1) is thus obtained.

The effectiveness of the above method has been verified through numerical simulations; see [3, 4] for a detail.

Remark 2. The constraint (11) needs to be further investigated since it may restrict systems we can identify. In most of the applications in Sect. 4, however, we do not have to reduce the indeterminacy in this way. This is because there is no need to distinguish the pair $(N(z), u)$ from $(N(z)P\Lambda, (P\Lambda)^{-1}u)$, for any permutation and diagonal matrices P and Λ.

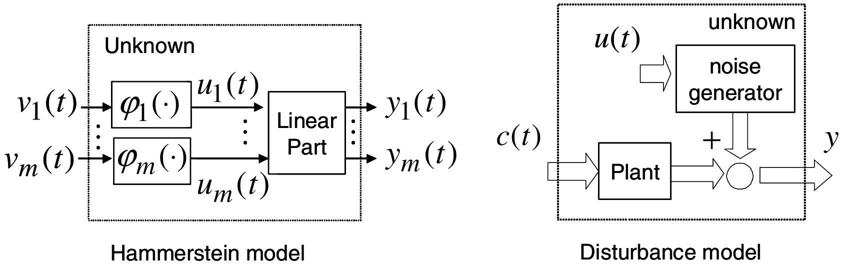
4 Applications

In this section we illustrate a couple of applications of blind identification. Sects. 4.1, 4.2, and 4.3 treat issues on control systems. Sect. 4.4 studies time-series prediction with focus on financial engineering.

In Sects. 4.1 and 4.2 the method is applied in batch processing, as in the original setting. After obtaining the parameter we can improve the control performance via re-designing the system. On the other hand, in Sects. 4.3 and 4.4 we change our viewpoint and try to apply the method in real-time, with a slight abuse.

4.1 Hammerstein Model Identification

In control systems, we sometimes encounter nonlinearity in the input channels such as saturation or dead zone; see the left-hand side of Fig. 3. If such nonlinear functions are static (i.e., memoryless), then we can identify both the linear part and the



Hammerstein model

Disturbance model

Fig. 3 Application to Hammerstein and Disturbance Models

nolinear functions by applying the blind identification technique in Sect. 3. This is called Hammerstein model identification.

If we obtain the distorting functions φ_i, then the compensation of the distortion is straightforward by using their inverse functions:

$$v_i(t) = \varphi_i^{-1}(\bar{u}_i(t)), \quad i = 1, \cdots, m,$$

where $\bar{u}_i$ is what we want to apply as u_i to the linear part [5].

4.2 Disturbance Model Identification

Control systems suffer from various disturbances whose mechanism is often unknown. If their statistic property is identical, we can estimate such unknown dynamics by observing command input $c = (c_1, \cdots, c_r)'$ and measured output $y = (y_1, \cdots, y_m)'$; see the right-hand side of Fig. 3.

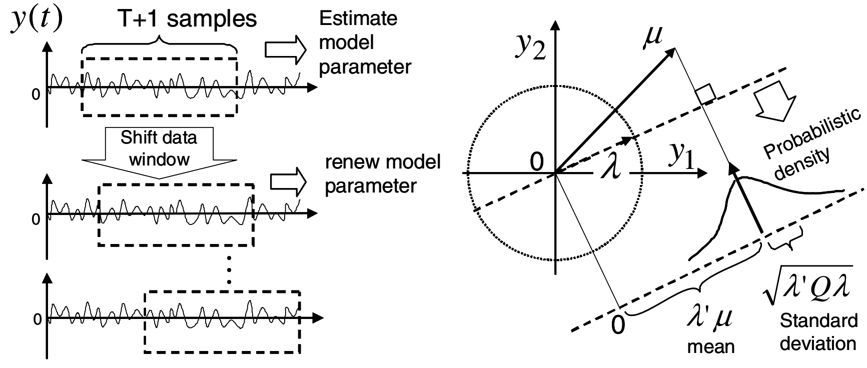
Note that $u = (u_1, \cdots, u_m)'$ is not considered to be a real world signal but a virtual input to drive the noise generator which generates a colored signal. In contrast to the previous issues, the command input c is assumed to be available for us, hence this scheme can be viewed as “semi-blind” identification.

After we identify both the plant and the noise generator, we can suppress the noise effect on y by applying feedback from y to c, say by H_∞ control methods: We can select weighting functions based on the identified parameters [6].

4.3 Change Detection

Since the method in Sect. 3 converges with a relatively small number of samples, we can apply it in a subinterval of the interval under consideration and renew the parameter estimation while shifting the subinterval (called window) with respect to time; see the left-hand side in Fig. 4.

One advantage of this window shifting is that we can trace the gradual change of the system parameter. Along this line, the authors [6] have studied a kind of fault detection technique of a mechanical system under vibration and verified its effectiveness via an experiment with a flexible structure. In spite that the frequency



Data window shifted for each identification

Probabilistic density for a brand combination

Fig. 4 Semi real-time applications

band of this structure is rather low, it can detect the change within around 10 (ten) seconds. The period is further shortened by combining this method with FastICA.

The semi real-time version of the proposed method in this sense is also adopted in the following subsection.

4.4 Time Series Prediction

The last issue is application to finance. Let $S_i(t)$ be the stock price of brand B_i at discrete time t, for $i = 1, \cdots, m$. Define the stock return vector by

$$y(t) = (y_1(t), \cdots, y_m(t))', \qquad y_i(t) = \log \frac{S_i(t)}{S_i(t-1)}.$$

Suppose that y satisfies the assumption in Sect. 2 for some transfer matrix $G(z)$ in Eq. (1), with input u representing various external factors. This model seems to be natural since stock prices are affected by the past values of their own as well as those of other stocks. Indeed, the model includes some existing ones.

Now suppose that we have estimated the parameter of Eq. (1) by observing $y(t)$ for $t = 0, \cdots, T$. Then we can predict the one-step-ahead return vector as the conditional expectation:

$$\mu = \mathscr{E}[y(T+1)] = -\sum_{i=1}^{p} D_i y(T+1-i) + \sum_{j=1}^{q} N_j \hat{u}(T+1-j). \tag{12}$$

Here, we have used $\mathscr{E}[N_0 u(T+1)] = N_0 \mathscr{E}[u(T+1)] = 0$ and substituted $u(t)$ by its estimation $\hat{u}(t)$ for $t = T+1-q, \cdots, T$.

If we take a ratio vector $\lambda = (\lambda_1, \cdots, \lambda_m)'$ such that $\|\lambda\| = 1$ and buy the combination of the stocks at this ratio (more precisely, "buy" B_i if $\lambda_i > 0$, and "sell" B_i if $\lambda_i < 0$ for $i = 1, \cdots, m$), then the return is expected to be $\mathscr{E}[\lambda' y(T+1)] = \lambda' \mu$.

We might expect a profit if this is positive. However, it is not sufficient to select λ such that $\lambda'\mu$ is large, because there is always an investment risk in it. In this context, the importance of variance (what they call volatility) is highly recognized in financial engineering.

In our model, the variance is evaluated as $\lambda' Q\lambda$ where

$$Q = \mathscr{E}\left[(y(T+1)-\mu)(y(T+1)-\mu)'\right].$$

By Eqs. (2) and (12), we have

$$Q = N_0 V N_0', \quad V = \mathscr{E}[u(T+1)u(T+1)'] = \mathrm{diag}(v_1,\cdots,v_m), \tag{13}$$

because the components of u are assumed to be independent. The diagonal entries $v_i = \mathscr{E}[u_i(T+1)^2]$ are unavailable for us so that we substitute them with

$$v_i = \frac{1}{T+1}\sum_{t=0}^{T}\hat{u}_i(t)^2, \quad i = 1,\cdots,m. \tag{14}$$

By Eqs. (13) and (14), we thus obtain Q to calculate the variance $\lambda' Q\lambda$.

In order to find the ratio λ such that the probability of loss $\lambda' y(T+1) < 0$ is as small as possible, we maximize

$$\gamma = \frac{\lambda'\mu}{\sqrt{\lambda' Q\lambda}} \tag{15}$$

with respect to λ; see the right-hand side of Fig. 4.[3] If γ is positive and large enough, then the ratio λ is recommendable. Otherwise it is too risky.

The effectiveness of this method has been verified in [7], by numerical simulation based on data of actual 8 brands. The method takes nontrivial computational time, but is feasible enough for financial application.

5 Conclusion

As in many other statistic learning methods, parametrization and learning law are two important issues in BSD. This article has focused on the parametrization issue from a system-theoretic point of view, and discussed its application to control and finance.

References

1. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley & Sons, Inc., Chichester (2001)
2. Cichocki, A., Amari, S.: Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications. John Wiley & Sons, Inc., Chichester (2002)

[3] As opposed to Sect. 2, we assume here that $\lambda' y$ is Gaussian (approximately), and hence its statistic property is determined only by μ and Q.

3. Nitta, M., Sugimoto, K.: Blind identification of polynomial matrix fraction via independent component analysis. Int. J. of Robust and Nonlinear Control, Special Issue on Polynomial Methods 17, 732–751 (2007)
4. Nitta, M., Sugimoto, K.: Blind identification of MIMO discrete-time systems based on independent component analysis. IEICE Trans. J90-A, 27–34 (2007) (in Japanese); English translation is available in Electronics and Communications in Japan, Part II, vol. 90, pp. 17–25
5. Even, J., Sugimoto, K.: Estimation of MIMO system with nonlinear distortion at the input: an adaptive approach. In: Proc. SICE-ICASE International Joint Conference (SICE-ICCAS 2006), Busan, Korea, October 2006, pp. 3984–3989 (2006)
6. Sugimoto, K., et al.: Polynomial matrix approach to independent component analysis (Part II) application. In: Proc. 16th IFAC World Congr., FR-A22-TO/1, Prague, Czech Republic (July 2005); Journal version is available in Japanese as: Nitta, M., Sugimoto, K.: Disturbance rejection system based on independent component analysis. Trans. of the Society of Instrument and Control Engineers 42, 1313–1319 (2006); Nitta et al.: Blind identification of mechanical systems via independent componet analysis and detection of structure change. Trans. Inst. of Systems, Control and Information Engineers 19, 177–184 (2006)
7. Sugimoto, K., Kondo, H.: Multivariate time series prediction by blind signal deconvolution. In: Proc. ICROS-SICE International Joint Conference 2009, Fukuoka, Japan (August 2009)
8. Amari, S., Cichocki, A., Yang, H.: A new learning algorithm for blind signal separation. In: Advances in Neural Information Processing Systems, vol. 8, pp. 757–763. MIT Press, Cambridge (1996)

Author Index

Lecture Notes in Control and Information Sciences

Edited by M. Thoma, F. Allgöwer, M. Morari

Further volumes of this series can be found on our homepage:
springer.com

Vol. 398: Willems, J.C.; Hara, S.;
Ohta, Y.; Fujioka, H. (Eds.):
Perspectives in Mathematical System
Theory, Control, and Signal Processing
388 p. 2010 [978-3-540-93917-7]

Vol. 397: Yang, H.; Jiang, B.; Cocquempot, V.:
Fault Tolerant Control Design for
Hybrid Systems
191 p. 2010 [978-3-642-10680-4]

Vol. 396: Kozlowski, K. (Ed.):
Robot Motion and Control 2009
475 p. 2009 [978-1-84882-984-8]

Vol. 395: Talebi, H.A.:
Neural Network-Based State
Estimation of Nonlinear Systems
appro. 200 p. 2010 [978-1-4419-1437-8]

Vol. 394: Pipeleers, G.; Demeulenaere, B.;
Swevers, J.:
Optimal Linear Controller Design for
Periodic Inputs
177 p. 2009 [978-1-84882-974-9]

Vol. 393: Ghosh, B.K.; Martin, C.F.;
Zhou, Y.:
Emergent Problems in Nonlinear
Systems and Control
285 p. 2009 [978-3-642-03626-2]

Vol. 392: Bandyopadhyay, B.; Deepak, F.;
Kim, K.-S.:
Sliding Mode Control Using Novel
Sliding Surfaces
137 p. 2009 [978-3-642-03447-3]

Vol. 391: Khaki-Sedigh, A.; Moaveni, B.:
Control Configuration Selection for
Multivariable Plants
232 p. 2009 [978-3-642-03192-2]

Vol. 390: Chesi, G.; Garulli, A.;
Tesi, A.; Vicino, A.:
Homogeneous Polynomial Forms for
Robustness Analysis of Uncertain Systems
197 p. 2009 [978-1-84882-780-6]

Vol. 389: Bru, R.; Romero-Vivó, S. (Eds.):
Positive Systems
398 p. 2009 [978-3-642-02893-9]

Vol. 388: Jacques Loiseau, J.; Michiels, W.;
Niculescu, S-I.; Sipahi, R. (Eds.):
Topics in Time Delay Systems
418 p. 2009 [978-3-642-02896-0]

Vol. 387: Xia, Y.;
Fu, M.; Shi, P.:
Analysis and Synthesis of
Dynamical Systems with Time-Delays
283 p. 2009 [978-3-642-02695-9]

Vol. 386: Huang, D.;
Nguang, S.K.:
Robust Control for Uncertain Networked
Control Systems with Random Delays
159 p. 2009 [978-1-84882-677-9]

Vol. 385: Jungers, R.:
The Joint Spectral Radius
144 p. 2009 [978-3-540-95979-3]

Vol. 384: Magni, L.; Raimondo, D.M.;
Allgöwer, F. (Eds.):
Nonlinear Model Predictive Control
572 p. 2009 [978-3-642-01093-4]

Vol. 383: Sobhani-Tehrani E.:
Khorasani K.;
Fault Diagnosis of Nonlinear Systems
Using a Hybrid Approach
360 p. 2009 [978-0-387-92906-4]

Vol. 382: Bartoszewicz A.;
Nowacka-Leverton A.:
Time-Varying Sliding Modes for Second
and Third Order Systems
192 p. 2009 [978-3-540-92216-2]

Vol. 381: Hirsch M.J.; Commander C.W.;
Pardalos P.M.; Murphey R. (Eds.):
Optimization and Cooperative
Control Strategies: Proceedings of the 8th
International Conference on Cooperative
Control and Optimization
459 p. 2009 [978-3-540-88062-2]

Vol. 380: Basin M.:
New Trends in Optimal Filtering and Control for Polynomial and Time-Delay Systems
206 p. 2008 [978-3-540-70802-5]

Vol. 379: Mellodge P.; Kachroo P.:
Model Abstraction in Dynamical Systems:
Application to Mobile Robot Control
116 p. 2008 [978-3-540-70792-9]

Vol. 378: Femat R.; Solis-Perales G.:
Robust Synchronization of Chaotic Systems
Via Feedback
199 p. 2008 [978-3-540-69306-2]

Vol. 377: Patan K.:
Artificial Neural Networks for
the Modelling and Fault
Diagnosis of Technical Processes
206 p. 2008 [978-3-540-79871-2]

Vol. 376: Hasegawa Y.:
Approximate and Noisy Realization of
Discrete-Time Dynamical Systems
245 p. 2008 [978-3-540-79433-2]

Vol. 375: Bartolini G.; Fridman L.;
Pisano A.; Usai E. (Eds.):
Modern Sliding Mode Control Theory
465 p. 2008 [978-3-540-79015-0]

Vol. 374: Huang B.; Kadali R.:
Dynamic Modeling, Predictive Control
and Performance Monitoring
240 p. 2008 [978-1-84800-232-6]

Vol. 373: Wang Q.-G.; Ye Z.; Cai W.-J.;
Hang C.-C.:
PID Control for Multivariable Processes
264 p. 2008 [978-3-540-78481-4]

Vol. 372: Zhou J.; Wen C.:
Adaptive Backstepping Control of Uncertain
Systems
241 p. 2008 [978-3-540-77806-6]

Vol. 371: Blondel V.D.; Boyd S.P.;
Kimura H. (Eds.):
Recent Advances in Learning and Control
279 p. 2008 [978-1-84800-154-1]

Vol. 370: Lee S.; Suh I.H.;
Kim M.S. (Eds.):
Recent Progress in Robotics:
Viable Robotic Service to Human
410 p. 2008 [978-3-540-76728-2]

Vol. 369: Hirsch M.J.; Pardalos P.M.;
Murphey R.; Grundel D.:
Advances in Cooperative Control and
Optimization
423 p. 2007 [978-3-540-74354-5]

Vol. 368: Chee F.; Fernando T.
Closed-Loop Control of Blood Glucose
157 p. 2007 [978-3-540-74030-8]

Vol. 367: Turner M.C.; Bates D.G. (Eds.):
Mathematical Methods for Robust and
Nonlinear Control
444 p. 2007 [978-1-84800-024-7]

Vol. 366: Bullo F.; Fujimoto K. (Eds.):
Lagrangian and Hamiltonian Methods for
Nonlinear Control 2006
398 p. 2007 [978-3-540-73889-3]

Vol. 365: Bates D.; Hagström M. (Eds.):
Nonlinear Analysis and Synthesis
Techniques for Aircraft Control
360 p. 2007 [978-3-540-73718-6]

Vol. 364: Chiuso A.; Ferrante A.;
Pinzoni S. (Eds.):
Modeling, Estimation and Control
356 p. 2007 [978-3-540-73569-4]

Vol. 363: Besançon G. (Ed.):
Nonlinear Observers and Applications
224 p. 2007 [978-3-540-73502-1]

Vol. 362: Tarn T.-J.; Chen S.-B.;
Zhou C. (Eds.):
Robotic Welding, Intelligence and
Automation
562 p. 2007 [978-3-540-73373-7]

Vol. 361: Méndez-Acosta H.O.; Femat R.;
González-Álvarez V. (Eds.):
Selected Topics in Dynamics and Control of
Chemical and Biological Processes
320 p. 2007 [978-3-540-73187-0]

Vol. 360: Kozlowski K. (Ed.):
Robot Motion and Control 2007
452 p. 2007 [978-1-84628-973-6]

Vol. 359: Christophersen F.J.:
Optimal Control of Constrained
Piecewise Affine Systems
190 p. 2007 [978-3-540-72700-2]

Vol. 358: Findeisen R.; Allgöwer
F.; Biegler L.T. (Eds.):
Assessment and Future
Directions of Nonlinear
Model Predictive Control
642 p. 2007 [978-3-540-72698-2]

Vol. 357: Queinnec I.; Tarbouriech
S.; Garcia G.; Niculescu S.-I. (Eds.):
Biology and Control Theory:
Current Challenges
589 p. 2007 [978-3-540-71987-8]