Lecture Notes in Control and Information Sciences

371

Vincent D. Blondel
Stephen P. Boyd
Hidenori Kimura (Eds.)

Recent Advances in Learning and Control

Springer

Lecture Notes
in Control and Information Sciences 371

Editors: M. Thoma, M. Morari

Vincent D. Blondel, Stephen P. Boyd,
Hidenori Kimura (Eds.)

Recent Advances in Learning and Control

Springer

Editors

Vincent D. Blondel, PhD

Département d'Ingénierie Mathématique
Université catholique de Louvain
avenue Georges Lemaître, 4
B-1348 Louvain-la-Neuve, Belgium
Email: vincent.blondel@uclouvain.be

Stephen P. Boyd, PhD

Information Systems Laboratory
Department of Electrical Engineering
Stanford University
Stanford, CA 94305-9510 ,USA
Email: boyd@stanford.edu

Hidenori Kimura, PhD

Bio-Mimetic Control Research Center
RIKEN (The Institute of Physical and
Chemical Research)
Anagahora, Shimoshidami, Moriyama-ku
Nagoya, 463-0003, Japan
Email: kimura@bmc.riken.jp

ISBN 978-1-84800-154-1 e-ISBN 978-1-84800-155-8
DOI 10.1007/978-1-84800-155-8

Lecture Notes in Control and Information Sciences ISSN 0170-8643

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2007941827

Typesetting by the authors and Scientific Publishing Services Pvt. Ltd.

Printed on acid-free paper

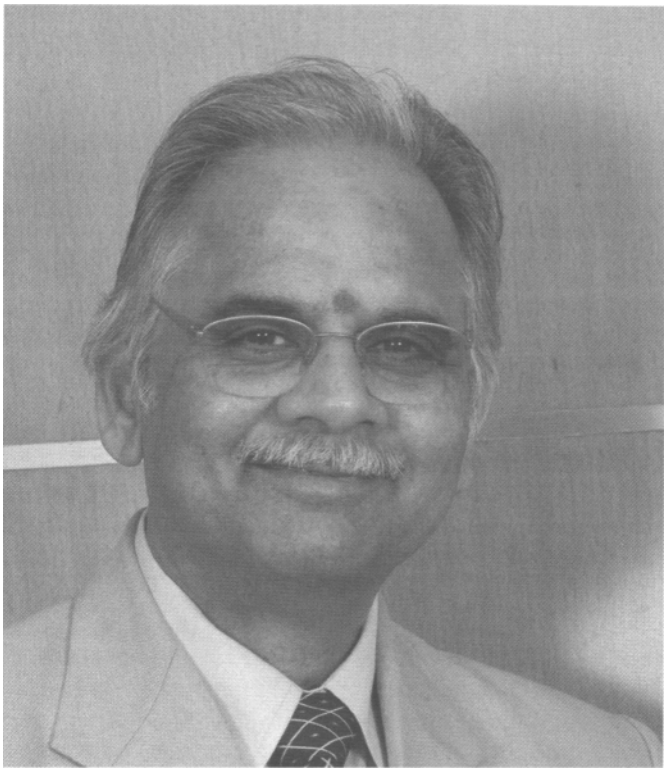
9 8 7 6 5 4 3 2 1

springer.com

This Festschrift contains a collection of articles by friends, co-authors, colleagues, and former Ph.D. students of Mathukumalli Vidyasagar. The articles have been collected at the occasion of his sixtieth birthday and are a tribute to his seminal work, which has been a continuous source of inspiration.

Vincent D. Blondel (Louvain-la-Neuve, Belgium)
Stephen P. Boyd (Stanford, USA)
Hidenori Kimura (Tokyo, Japan)

Mathukumalli Vidyasagar



Mathukumalli Vidyasagar was born on September 29, 1947 in Guntur, Andhra Pradesh, India. He moved to the USA along with his parents in 1960, and after completing his high school education in the USA, received his Bachelor's, Master's and Ph.D. degrees in Electrical Engineering from the University of Wisconsin, Madison, in 1965, 1967 and 1969 respectively. His Master's thesis was on adjoint systems for linear discrete-time systems, and his Ph.D. thesis was on the control of distributed parameter systems.

After completing his Ph.D. in 1969, he spent one year teaching at Marquette University in Milwaukee before moving to Canada where he spent the next nineteen years: Ten years at Concordia University in Montreal, and nine years at the University of Waterloo. He returned to his native India in 1989 to head the newly created Centre for Artificial Intelligence and Robotics (CAIR) under the Ministry of Defence, Government of India. In 2000 he joined Tata Consultancy Services, India's largest IT and Services Company, as an Executive Vice President, and set up the Advanced Technology Center (ATC).

During his twenty year long teaching career, Vidyasagar was a conventional professor, teaching courses and conducting research in control theory and robotics. In his first job in India as the Director of CAIR, he took the first tentative steps towards becoming a scientific administrator. During his tenure, he built CAIR into one of the leading research laboratories within DRDO with about 45 scientists and about 85 persons overall. In his current position, his charter is to identify areas of research that would lead TCS into non-traditional areas. At present the ATC consists of five groups, namely: E-Security (now comprising also Media & Entertainment), Life Sciences R&D, Open Source/Linux, Smart Card Technology, and most recently, Quantitative Finance. The current strength of ATC is around 90 persons. His duties as an EVP of India's largest IT company keep him occupied most of the time, and he finds it increasingly difficult to find time for research.

His scientific work spans several areas, beginning with circuit theory (his first love but abandoned very quickly during his Ph.D.); linear, nonlinear and robust control systems; robotics; neural networks; statistical learning theory; and most recently, statistical methods in computational biology. He is perhaps best known for having introduced the "stable factorization method," which expresses an unstable system as the ratio of stable systems, and significantly simplifies both the analysis and synthesis of feedback stability. Other contributions that seem to have made an impact are the introduction of ℓ_1-optimal control theory as an alternative to the better-known H_∞ control theory, and a general approach to the use of randomized algorithms in analyzing intractable problems in robust control.

During his career, Vidyasagar has written books at regular intervals, as he finds writing a book the best way to learn a subject. His book count will shortly enter double digits with the publication of *Hidden Markov Processes: Theory and Applications to Biology.* He has also published about 130 journal papers, and has supervised 14 Ph.D. students.

Vidyasagar has received several honors in recognition of his research, including the 2008 IEEE Control Systems Award. He was awarded the "Distinguished Service Citation" by his alma mater, the University of Wisconsin in 1995. In 2004, *IEEE Spectrum* named him as one of "Forty Tech Gurus" in the world, along with luminaries such as Craig Barret, Vinton Cerf *et al.* He is a Fellow of IEEE, the Indian Academy of Sciences, the Indian National Science Academy, the Indian National Academy of Engineering, and the Third World Academy of Sciences.

Vidyasagar has been married to Shakunthala for 35 years, and has one daughter Aparna. She is partially following in her father's footsteps by completing her Bachelor's degree and now pursuing her Ph.D. at the University of Wisconsin, Madison, but her specialization is in the life sciences.

Current Position and Responsibilities

Since 2000, Dr. Vidyasagar has been with Tata Consultancy Services (TCS), which is Asia's largest IT and consultancy firm, with more than 55,000 consultants and a turnover in 2004-06 of about U.S. $ 2.24 billion. Dr. Vidyasagar is one of seven Executive Vice Presidents, and directs the activities of the Advanced Technology Centre (ATC) whose charter is to develop new technologies in non-traditional (i.e., non-software) areas. In this capacity he oversees a team of about 70 persons working in three areas: E-Security including advanced encryption methods, Life Sciences R&D, and Open Source & Indian Language Computing.

In addition to overseeing the R&D activities of the ATC, Dr. Vidyasagar also continues with his own personal research. At present, his research interests are in the area of statistical learning theory and its applications to problems in computational biology.

Contact Information

Executive Vice President (Advanced Technology)
Tata Consultancy Services
No. 1, Software Units Layout, Madhapur
Hyderabad 500081, INDIA Tel: +91 40 6667 3001, Fax: +91 40 6667 2222
E-Mail: sagar@atc.tcs.co.in, m.vidyasagar@tcs.com
URL: www.atc.tcs.co.in/~sagar

Contents

Statistical Learning Theory: A Pack-based Strategy for Uncertain Feasibility and Optimization Problems

Teodoro Alamo[1], Roberto Tempo[2], and Eduardo F. Camacho[1]

[1] Departamento de Ingeniería de Sistemas y Automática, Universidad de Sevilla, Escuela Superior de Ingenieros, Camino de los Descubrimientos s/n. 41092 Spain
{alamo,eduardo}@cartuja.us.es
[2] IEIIT-CNR, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy
roberto.tempo@polito.it

Summary. In this paper, a new powerful technique, denoted as *pack-based strategy* is introduced in the context of statistical learning theory. This strategy allows us to derive bounds on the number of required samples that are manageable for "reasonable" values of probabilistic confidence and accuracy. Using this technique for feasibility and optimization problems involving Boolean expressions consisting of polynomials, we prove that the number of required samples grows with the accuracy parameter ϵ as $\frac{1}{\epsilon} \ln \frac{1}{\epsilon}$. This is a significant improvement when compared to the existing bounds which depend on $\frac{1}{\epsilon^2} \ln \frac{1}{\epsilon^2}$. We also apply this strategy to convex optimization problems. In this case, we show that the required sample size is inversely proportional to the accuracy for fixed confidence.

Keywords: Statistical learning theory, randomized algorithms, probabilistic robustness, uncertain systems.

1 Introduction

Uncertainty randomization is now widely accepted as an effective tool in dealing with uncertain feasibility and optimization problems which are computationally difficult, see e.g. [16]. In particular, regarding *synthesis* of a controller to achieve a given performance, two complementary approaches, sequential and non-sequential, have been proposed in recent years.

For sequential methods, the resulting algorithms are based on stochastic gradient [7, 9, 14] or ellipsoid iterations [12]; see also [3], [6] for other classes of sequential algorithms. Convergence properties in finite-time are in fact one of the focal points of these papers. Various control problems have been solved using these sequential randomized algorithms, including robust LQ regulators and uncertain Linear Matrix Inequalities.

A classic approach for non-sequential methods is based upon statistical learning theory, see [17] and [20] for further details. In particular, the use of this theory for feedback design of uncertain systems has been initiated in [21]; subsequent work along this direction include [22], [23] and [13]. However, the sample

V.D. Blondel et al. (Eds.) Recent Advances in Learning and Control, LNCIS 371, pp. 1–14, 2008.
springerlink.com

size bounds derived in these papers, which guarantee that the obtained solution meets a given probabilistic specification, may be too conservative for being practically useful in a systems and control context.

For convex optimization problems, a successful paradigm, denoted as the *scenario approach*, has been introduced in [4,5]. In this approach, the original robust control problem is reformulated in terms of a single convex optimization problem with sampled constraints which are randomly generated. The main result of this line of research is to derive explicit sample size bounds without resorting to statistical learning methods.

In this paper, we address two general (not necessarily convex) problems which encompass control of systems affected by uncertainty. In particular, we consider a performance function $g : \Theta \times \mathcal{W} \to \{0, 1\}$, which is a binary measurable function that serves to formulate the particular design problem under attention.

Semi-infinite feasibility problem: Find θ, if it exists, in the feasible set

$$\{ \theta \in \Theta \ : \ g(\theta, w) = 0, \text{ for all } w \in \mathcal{W} \}. \tag{1}$$

Semi-infinite optimization problem: If the feasible set is nonempty, find the optimal solution of the problem

$$\min_{\theta \in \Theta} J(\theta) \text{ subject to } g(\theta, w) = 0, \text{ for all } w \in \mathcal{W}, \tag{2}$$

where $J : \Theta \to (-\infty, \infty)$ is a measurable function. We notice that robust stabilization, robust model predictive control and control of uncertain nonlinear systems, to number only a few, may be reformulated as (1) or (2). We also recall that various specific applications which require the solution of these general semi-infinite optimization problems include congestion and cooperative control, robust localization and trajectory planning.

The robust design problems formulated in this paper are generally very difficult to solve deterministically when the set $\mathcal{W}$ has infinite cardinality. This is specially the case when the equality $g(\theta, w) = 0$ is *not* a convex constraint on the decision variable θ. One way of circumventing the possible infinity cardinality of set $\mathcal{W}$ consists in the use of randomization, see [16, 20].

The main contribution of this paper is to present the so-called *pack-based strategy*, and to discuss its use for uncertain feasibility and optimization problems. Using this strategy, we show that the sample size bounds which appear in the control literature can be significantly improved.

2 A Randomized Strategy

In this paper we assume that a probability measure $\Pr_{\mathcal{W}}$ over the sample space $\mathcal{W}$ is given; L, M, N and n_θ represent positive integers. Then, given $\mathcal{W}$, a collection of N independent identically distributed (i.i.d.) samples $\mathrm{w} = \{w^{(1)}, \ldots, w^{(N)}\}$ drawn from $\mathcal{W}$ is said to belong to the Cartesian product $\mathcal{W}^N = \mathcal{W} \times \cdots \times \mathcal{W}$ (N times). Moreover, if the collection w of N i.i.d. samples $\{w^{(1)}, \ldots, w^{(N)}\}$ is

generated from $\mathcal{W}$ according to the probability measure $\Pr_{\mathcal{W}}$, then the *multi-sample* w is drawn according to the probability measure $\Pr_{\mathcal{W}^N}$. The *accuracy*, *confidence* and constraint level (or *level*) are denoted by $\epsilon \in (0,1)$, $\delta \in (0,1)$ and $\rho \in [0,1)$, respectively. For $x \in \mathbb{R}$, $x > 0$, $\lceil x \rceil$ denotes the minimum integer greater than or equal to x, $\ln(\cdot)$ is the natural logarithm and e is the Euler number.

2.1 Probability of Violation and Empirical Mean

Given $\theta \in \Theta$, there might be a fraction of the elements of $\mathcal{W}$ 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* $\Pr_{\mathcal{W}}$ *over* $\mathcal{W}$ *and let* $\theta \in \Theta$ *be given. The probability of violation of* θ *for the function* $g : \Theta \times \mathcal{W} \to \{0,1\}$ *is defined as*

$$E_g(\theta) := \Pr_{\mathcal{W}} \{\, w \in \mathcal{W} \;:\; g(\theta, w) = 1 \,\}.$$

Given $\theta \in \Theta$, it is generally difficult to obtain the exact value of the probability of violation $E_g(\theta)$ 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 $\mathrm{w} = \{w^{(1)}, \ldots, w^{(N)}\}$ is defined as

$$\hat{E}_g(\theta, \mathrm{w}) := \frac{1}{N} \sum_{i=1}^{N} g(\theta, w^{(i)}). \tag{3}$$

Clearly, the empirical mean $\hat{E}_g(\theta, \mathrm{w})$ is a random variable. Since $g(\cdot, \cdot)$ is a binary function, $\hat{E}_g(\theta, \mathrm{w})$ is always within the closed interval $[0,1]$. As discussed below, the empirical mean is a key tool for solving the general semi-infinite problems considered in this paper by means of a randomized approach.

2.2 Randomized Feasibility and Optimization Problems

Suppose that a probability measure $\Pr_{\mathcal{W}}$ over the set $\mathcal{W}$ and the level $\rho \in [0,1)$ are given. Consider the following *randomized strategy*:

1. Draw N i.i.d. samples $\mathrm{w} = \{w^{(1)}, \ldots, w^{(N)}\}$ according to the probability $\Pr_{\mathcal{W}}$.
2. Find (if possible) a feasible solution $\theta \in \Theta$ of the constraint

$$\hat{E}_g(\theta, \mathrm{w}) \leq \rho. \tag{4}$$

3. If a feasible solution exists, solve the optimization problem

$$\min_{\theta \in \Theta} J(\theta) \;\text{ subject to } \hat{E}_g(\theta, \mathrm{w}) \leq \rho. \tag{5}$$

In this paper, the problems (4) and (5) are denoted as *randomized feasibility* and *randomized optimization*, respectively. We note that considering a level ρ larger than zero broadens the class of problems that can be addressed by the proposed methodology. To motivate the study of $\rho > 0$, we notice, for example, that in many industrial and manufacturing processes, allowing a (small) probability of violation has the effect of considerably reducing the production costs.

We remark that related randomized strategies can be found in the literature. For example, empirical mean minimization techniques are proposed in [22], min-max problems are presented in [10] and bootstrap learning methods are introduced in [13]. In these references, the main emphasis is on deriving explicit bounds on the required number of samples randomly drawn from $\mathcal{W}$ to guarantee that the obtained solution satisfies some probabilistic accuracy and confidence specifications. All these methods, however, are also based on the randomization of the design parameter set Θ, which implies that a finite family approach is followed.

3 Estimating the Probability of Failure

We address the problem of obtaining an explicit bound on the sample size to guarantee that the empirical mean is within a pre-specified accuracy $\epsilon \in (0,1)$ from the probability of violation with high confidence $1-\delta$, $\delta \in (0,1)$. The Hoeffding inequalities (see for example [11,16]) characterize how the empirical mean approximates, from a probabilistic point of view, the exact value of $E_g(\theta)$. Suppose that $\epsilon \in (0,1)$ is given. Then, from the Hoeffding inequalities we conclude that

$$\Pr_{\mathcal{W}^N}\{\, \mathrm{w} \in \mathcal{W}^N \;:\; |E_g(\theta) - \hat{E}_g(\theta, \mathrm{w})| \geq \epsilon \,\} \leq 2e^{-2N\epsilon^2}. \tag{6}$$

Thus, in order to guarantee confidence $1-\delta$ it suffices to take N such that $2e^{-2N\epsilon^2} \leq \delta$ is satisfied. That is, we obtain the (additive) Chernoff bound [8]

$$N \geq \left\lceil \frac{1}{2\epsilon^2} \ln \frac{2}{\delta} \right\rceil.$$

Unfortunately, this inequality is not applicable to the randomized framework stated in (4) and (5). In fact, the Chernoff bound is valid only for *fixed* θ. In the randomized strategy previously described, the parameter θ is a design variable varying in the set Θ. Motivated by this discussion, we now introduce the definition of probability of two-sided failure.

Definition 2. *[probability of two-sided failure] Given N, $\epsilon \in (0,1)$ and $g : \Theta \times \mathcal{W} \to \{0,1\}$, the probability of two-sided failure, denoted by $q_g(N,\epsilon)$ is defined as*

$$q_g(N,\epsilon) := \Pr_{\mathcal{W}^N}\{\, \mathrm{w} \in \mathcal{W}^N \;:\; \sup_{\theta \in \Theta} |E_g(\theta) - \hat{E}_g(\theta, \mathrm{w})| > \epsilon \,\}.$$

This definition can be easily explained in words. If N i.i.d. samples are drawn from $\mathcal{W}$ according to the probability $\Pr_{\mathcal{W}}$, the set of elements of Θ having an empirical mean not within ϵ of the exact value $E_g(\theta)$ is empty with probability

no smaller than $1-q_g(N,\epsilon)$. The function $g:\Theta\times\mathcal{W}\to\{0,1\}$ enjoys the Uniform Convergence of Empirical Means property (UCEM property) if $q_g(N,\epsilon)\to 0$ as $N\to 0$ for each $\epsilon>0$, see [19,20]. The probability of two-sided failure is based on the two inequalities $E_g(\theta)-\hat{E}_g(\theta,\mathrm{w})\leq\epsilon$ and $E_g(\theta)-\hat{E}_g(\theta,\mathrm{w})\geq-\epsilon$.

In the randomized strategy presented in this paper, we are interested in addressing the following question: after drawing N i.i.d. samples from $\mathcal{W}$, suppose that we compute $\hat{\theta}\in\Theta$ such that the corresponding empirical mean is smaller than ρ; then, what is the probability that the difference between $E_g(\hat{\theta})$ and the obtained empirical mean is larger than ϵ? To answer this question, we now introduce the formal definition of probability of one-sided constrained failure.

Definition 3. *[probability of one-sided constrained failure] Given N, $\epsilon\in(0,1)$, $\rho\in[0,1)$ and $g:\Theta\times\mathcal{W}\to\{0,1\}$, the probability of one-sided constrained failure, denoted by $p_g(N,\epsilon,\rho)$ is defined as*

$$p_g(N,\epsilon,\rho):=\mathrm{Pr}_{\mathcal{W}^N}\{\mathrm{w}\in\mathcal{W}^N:\ \textit{There exists } \theta\in\Theta \textit{ such that } \hat{E}_g(\theta,\mathrm{w})\leq\rho \textit{ and } E_g(\theta)>\hat{E}_g(\theta,\mathrm{w})+\epsilon\}.$$

4 Some One-sided and Two-sided Results from Statistical Learning Theory

Formally, let $\mathcal{G}$ denote the family of functions $\{\ g(\theta,\cdot)\ :\ \theta\in\Theta\ \}$, where $g:\Theta\times\mathcal{W}\to\{0,1\}$. Then, given the multisample $\mathrm{w}=\{w^{(1)},\ldots,w^{(N)}\}\in\mathcal{W}^N$, the binary vector $(g(\theta,w^{(1)}),\ldots,g(\theta,w^{(N)}))\in\{0,1\}^N$ can attain at most 2^N distinct values when θ varies in Θ. The maximum number of distinct binary vectors (denoted in the following by $\phi_g(\mathrm{w})$) that can be obtained grows with the number of samples N. The next definition introduces in a formal way the notion of growth function (also known as shatter coefficient), see [17], [18], [20].

Definition 4. *[growth function] Given the function $g:\Theta\times\mathcal{W}\to\{0,1\}$ and the multisample $\mathrm{w}=\{w^{(1)},\ldots,w^{(N)}\}\in\mathcal{W}^N$, $\phi_g(\mathrm{w})$ denotes the number of distinct binary vectors*

$$\{g(\theta,w^{(1)}),\ldots,g(\theta,w^{(N)})\}\in\{0,1\}^N$$

that can be obtained with the different elements of Θ. Then, the growth function $\pi_g(N)$ is defined as

$$\pi_g(N):=\sup_{\mathrm{w}\in\mathcal{W}^N}\phi_g(\mathrm{w}).$$

In words, $\phi_g(\mathrm{w})$ is the cardinality of the set $\{\ \{g(\theta,w^{(1)}),\ldots,g(\theta,w^{(N)})\}\ :\ \theta\in\Theta\ \}$ and the growth function is the supremum of $\phi_g(\mathrm{w})$ taken with respect to $\mathrm{w}\in\mathcal{W}^N$.

We are now ready to introduce the notion of VC-dimension, also known as the Vapnik-Chervonenkis dimension, see [17], [20].

Definition 5. *Given the function $g:\Theta\times\mathcal{W}\to\{0,1\}$, the VC-dimension, denoted as VC_g, is the largest integer N for which the equality $\pi_g(N)=2^N$ is satisfied.*

We now recall a bound on the VC-dimension stated in [22], for an important class of boolean functions. This bound has been used extensively to compute the VC-dimension of various classical control problems [22], [23], including the well-known static output feedback.

Lemma 1. *Suppose that the function* $g : \Theta \times \mathcal{W} \to \{0,1\}$ *can be written as a Boolean expression consisting of polynomials* $\beta_1(\theta, w), \ldots, \beta_k(\theta, w)$, *and the degree with respect to* $\theta \in \mathbb{R}^{n_\theta}$ *of all these polynomials is no larger than* α. *Then,*

$$\mathrm{VC}_g \leq 2n_\theta \log_2(4e\alpha k).$$

The VC-dimension establishes the "richness" of a given family $\mathcal{G}$. If the VC-dimension is bounded, the probability of two-sided failure converges to zero when N tends to infinity. This is stated in a precise way in the following result (see Theorems 7.2 and 10.2. in [20]).

Theorem 1. *Let* $g : \Theta \times \mathcal{W} \to \{0,1\}$ *be such that* $\mathrm{VC}_g \leq d < \infty$. *Suppose that* $\epsilon \in (0,1)$ *and* $\delta \in (0,1)$ *are given and* $N \geq d$. *Then,*

$$q_g(N,\epsilon) \leq 4\left(\frac{2eN}{d}\right)^d e^{\frac{-N\epsilon^2}{8}}.$$

Moreover, $q_g(N,\epsilon) \leq \delta$ *provided that*

$$N \geq \max\left\{\frac{16}{\epsilon^2}\ln\frac{4}{\delta}, \frac{32d}{\epsilon^2}\ln\frac{32e}{\epsilon^2}\right\}.$$

Theorem 1 is well-known in the control community. However, as discussed in [20], there exist other results in the literature that allow one to reduce by a factor close to 8 the explicit bound on the number of samples. Basically, these results differ from Theorem 1 in the exponent $-N\epsilon^2/8$ which is replaced by less conservative ones. For example, the exponent $-N\epsilon^2$ that can be found in [17] is used in [1] to obtain the following theorem.

Theorem 2. *Let* $g : \Theta \times \mathcal{W} \to \{0,1\}$ *be such that* $\mathrm{VC}_g \leq d < \infty$. *Suppose that* $\epsilon \in (0,1)$ *and* $\delta \in (0,1)$ *are given. Then, the probability of two-sided failure* $q_g(N,\epsilon)$ *is smaller than* δ *if*

$$N \geq \frac{1.2}{\epsilon^2}\left(\ln\frac{4e^{2\epsilon}}{\delta} + d\ln\frac{12}{\epsilon^2}\right).$$

We notice that the explicit bound obtained in this theorem is less conservative than Theorem 1. Theorems 1 and 2 refer to the probability of two-sided failure. The following explicit bound on the number of samples required to guarantee that the probability of one-sided failure is smaller than a pre-specified value δ has been recently obtained in [1].

Theorem 3. *Let $g : \Theta \times \mathcal{W} \to \{0,1\}$ be such that $\mathrm{VC}_g \leq d < \infty$. Suppose that $\epsilon \in (0,1)$, $\delta \in (0,1)$ and $\rho \in [0,1)$ are given. Then, the probability of one-sided constrained failure $p_g(N,\epsilon,\rho)$ is smaller than δ if*

$$N \geq \frac{5(\rho+\epsilon)}{\epsilon^2}\left(\ln\frac{4}{\delta} + d\ln\frac{40(\rho+\epsilon)}{\epsilon^2}\right).$$

The advantage of this one-sided result is that, for small values of the level parameter ρ, the number of required samples are several orders of magnitude smaller than those corresponding to the two-sided results.

5 A Pack-based Strategy in the Randomized Approach

In this section, a new technique denoted as *pack-based strategy* is used for obtaining improved sample size bounds for the case when the level parameter ρ is equal to zero. In particular, this strategy, presented in [2], is a key tool for proving the results stated in Section 6. Suppose that a probability measure $\Pr_{\mathcal{W}}$ over the set $\mathcal{W}$ is given. Consider the following *zero-level randomized strategy*:

1. Draw N i.i.d. samples $\mathrm{w} = \{w^{(1)}, \ldots, w^{(N)}\}$ according to the probability $\Pr_{\mathcal{W}}$.
2. Find (if possible) a feasible solution $\theta \in \Theta$ of the constraint

$$\hat{E}_g(\theta, \mathrm{w}) = 0. \tag{7}$$

3. If a feasible solution exists, solve the optimization problem

$$\min_{\theta\in\Theta} J(\theta) \text{ subject to } \hat{E}_g(\theta, \mathrm{w}) = 0. \tag{8}$$

In this *zero-level randomized strategy*, we are interested in addressing the following question: after drawing N i.i.d. samples from $\mathcal{W}$, suppose that we find $\hat{\theta} \in \Theta$ such that the corresponding empirical mean is equal to zero; then, what is the probability that $E_g(\hat{\theta})$ is larger than ϵ? This question is addressed in Section 6 by means of the *pack-based strategy*.

5.1 A Pack-based Formulation of the Zero-level Randomized Strategy

Given an integer L, we say that z belongs to the sample space $\mathcal{S}_L$ if z belongs to the Cartesian product $\mathcal{W} \times \cdots \times \mathcal{W}$ (L times). Then, given $g : \Theta \times \mathcal{W} \to \{0,1\}$, the function $g_L : \Theta \times \mathcal{S}^L \to \{0,1\}$ is defined as follows. For a given $\theta \in \Theta$ and a collection of L samples $z = \{w^{(1)}, \ldots, w^{(L)}\} \in \mathcal{S}_L$,

$$g_L(\theta, z) := \max_{i=1,\ldots,L} g(\theta, w^{(i)}). \tag{9}$$

The following notation is introduced to emphasize that a given collection of $N = LM$ samples may be considered as a collection of M packs of L samples

each. Given positive integers L and M, the collection $\mathrm{z} = \{z^{(1)}, \ldots, z^{(M)}\}$ is said to belong to the set $\mathcal{S}_L^M$ if $z^{(i)} \in \mathcal{S}_L$, $i = 1, \ldots, M$. Note that there is an equivalence one-to-one between the elements of $\mathcal{W}^{LM} = \mathcal{W}^N$ and the elements of $\mathcal{S}_L^M$.

The notion of probability of violation introduced in Definition 1, is now generalized to the function $g_L(\cdot, \cdot)$

$$E_{g_L}(\theta) := \mathrm{Pr}_{\mathcal{S}_L} \{ z \in \mathcal{S}_L \ : \ g_L(\theta, z) = 1 \}.$$

Clearly, the probability of L-violation can be estimated by its empirical mean. Given $\mathrm{z} = \{z^{(1)}, \ldots, z^{(M)}\} \in \mathcal{S}_L^M$, and $\theta \in \Theta$, the empirical mean of $g_L(\theta, z)$ with respect to the multisample z is defined as

$$\hat{E}_{g_L}(\theta, \mathrm{z}) := \frac{1}{M} \sum_{i=1}^{M} g_L(\theta, z^{(i)}). \tag{10}$$

The definition of probability of two-sided failure applies to function $g_L(\cdot, \cdot)$ in a natural way. Given M, $\tau \in (0, 1)$ and the function $g_L : \Theta \times \mathcal{S}_L \to \{0, 1\}$, the probability of two-sided failure $q_{g_L}(M, \tau)$ is defined by

$$q_{g_L}(M, \tau) := \mathrm{Pr}_{\mathcal{S}_L^M} \{ \mathrm{z} \in \mathcal{S}_L^M \ : \ \sup_{\theta \in \Theta} |E_{g_L}(\theta) - \hat{E}_{g_L}(\theta, \mathrm{z})| > \tau \}.$$

We are now in a position to introduce the following *pack-based zero-level randomized strategy*:

1. Given positive integers L and M, draw $N = LM$ independent identically distributed samples $\{w^{(1)}, \ldots, w^{(N)}\}$ according to the probability $\mathrm{Pr}_{\mathcal{W}}$.
2. Pack the N samples as follows:

$$\begin{aligned}
z^{(1)} &= (w^{(1)}, w^{(2)}, \ldots, w^{(L)}) \\
z^{(2)} &= (w^{(1+L)}, w^{(2+L)}, \ldots, w^{(2L)}) \\
&\vdots \quad \vdots \\
z^{(i)} &= (w^{(1+(i-1)L)}, w^{(2+(i-1)L)}, \ldots, w^{(iL)}) \\
&\vdots \quad \vdots \\
z^{(M)} &= (w^{(1+(M-1)L)}, w^{(2+(M-1)L)}, \ldots, w^{(ML)}).
\end{aligned}$$

3. Find (if possible) a feasible solution $\theta \in \Theta$ to the pack-based feasibility problem

$$\hat{E}_{g_L}(\theta, \mathrm{z}) = 0. \tag{11}$$

4. If this problem is feasible, find a solution to the pack-based optimization problem

$$\min_{\theta \in \Theta} J(\theta) \ \text{ subject to } \hat{E}_{g_L}(\theta, \mathrm{z}) = 0. \tag{12}$$

The following lemma, which is proved in Appendix 1, shows the relationships between the probabilities $E_{g_L}(\theta)$ and $E_g(\theta)$. We also remark that the "log-over-log" bound appearing here has been stated in [15], see also [16].

Lemma 2. *Suppose that $\theta \in \Theta$ satisfies $E_{g_L}(\theta) \leq \tau < 1$. Assume that $\epsilon \in (0,1)$ is given and*

$$L \geq \left\lceil \frac{\ln(1-\tau)}{\ln(1-\epsilon)} \right\rceil .$$

Then, we have that $E_g(\theta) \leq \epsilon$.

6 Explicit Sample Size Bounds for the Zero-level Randomized Problems

In this section, for the zero-level problem and when the function $g : \Theta \times \mathcal{W} \to \{0,1\}$ can be written as a Boolean polynomial expression, we derive a bound on the number of required samples that is manageable for "reasonable" values of probabilistic confidence and accuracy. In Subsection 6.1 we address the particular case when the constraints of the optimization problem are convex.

Before presenting the main result of this section, an intermediate result is stated.

Lemma 3. *Given $\epsilon \in (0,1)$, $\delta \in (0,1)$ and $N = LM$, suppose that M and L satisfy $q_{g_L}(M,\tau) \leq \delta$ and $L \geq \left\lceil \frac{\ln(1-\tau)}{\ln(1-\epsilon)} \right\rceil$. Then, with probability no smaller than $1-\delta$, either the zero-level randomized optimization problem (8) is unfeasible and, hence, also the general optimization problem (2) is unfeasible; or, (8) is feasible, and then any feasible solution θ satisfies the inequality $E_g(\theta) \leq \epsilon$.*

Proof. Notice that if the zero-level randomized optimization problem (8) is unfeasible then the general optimization problem (2) is also unfeasible. Consider now the case where problem (8) is feasible.

Clearly, if θ is a feasible solution to problem (8), then it is a feasible solution to the optimization problem (12). This implies that $\hat{E}_{g_L}(\theta, \mathrm{z}) = 0$. Denote by γ the probability of $E_g(\theta) > \epsilon$. Lemma 2 allows one to affirm that this probability γ is no larger than the probability of $E_{g_L}(\theta) > \tau$. We now prove that γ is no larger than $1-\delta$. That is,

$$\gamma \leq 1 - \Pr_{\mathcal{S}_L^M}\{\, \mathrm{z} \in \mathcal{S}_L^M \;:\; \sup_{\hat{E}_{g_L}(\theta,\mathrm{z})=0} E_{g_L}(\theta) > \tau \,\}$$

$$\leq 1 - \Pr_{\mathcal{S}_L^M}\{\, \mathrm{z} \in \mathcal{S}_L^M \;:\; \sup_{\theta\in\Theta} |E_{g_L}(\theta) - \hat{E}_{g_L}(\theta,\mathrm{z})| > \tau \,\} = 1 - q_{g_L}(M,\tau) \leq 1-\delta.$$

■

We are now in a position to introduce one of the main contributions of the paper. The following result states that to obtain a probabilistic feasible solution to problem (2), the notion of VC-dimension plays a secondary role. More precisely,

if ϵ is sufficiently small, an explicit bound that only depends on n_θ, δ and ϵ can be derived. On the other hand, for sufficiently large values of ϵ, the obtained explicit bound is inversely proportional to ϵ.

Theorem 4. *Suppose that the function $g : \Theta \times \mathcal{W} \to \{0,1\}$ can be written as a Boolean expression consisting of polynomials $\beta_1(\theta, w), \ldots, \beta_k(\theta, w)$, and the degree with respect to $\theta \in \mathbb{R}^{n_\theta}$ of all these polynomials is no larger than α. Letting $d = 2n_\theta \log_2(4e\alpha k)$, given $\epsilon \in (0, 0.25)$ and $\delta \in (0,1)$, if*

$$N \geq \bar{N}(\epsilon, \delta) := \begin{cases} \dfrac{3.54}{\epsilon}\left(\ln \dfrac{20.79}{\delta} + 11.5 n_\theta \log_2 \dfrac{2}{\epsilon}\right) & \text{if } \epsilon \leq \dfrac{1}{2e\alpha k}; \\ \dfrac{3.54}{\epsilon}\left(\ln \dfrac{20.79}{\delta} + 5.75 d\right) & \text{otherwise,} \end{cases}$$

then, with probability no smaller than $1-\delta$, either the problem (8) is unfeasible and, hence, also the general optimization problem (2) is unfeasible; or, (8) is feasible, and then any feasible solution $\hat{\theta}_N$ satisfies the inequality $E_g(\hat{\theta}_N) \leq \epsilon$.

Proof. Suppose that $\mathrm{VC}_{g_L} \leq d_L$ and that $\tau \in (0,1)$. As $\mathrm{VC}_{g_L} \leq d_L$, we infer from Theorem 2 that if

$$M \geq \frac{1.2}{\tau^2}\left(\ln \frac{4e^{2\tau}}{\delta} + d_L \ln \frac{12}{\tau^2}\right),$$

then $q_{g_L}(M, \tau) < \delta$. This means, by Lemma 3, that the number of samples N required to guarantee that the claim of Theorem 4 holds is bounded by

$$N \geq \left\lceil \frac{\ln(1-\tau)}{\ln(1-\epsilon)} \right\rceil \left(\frac{1.2}{\tau^2}\right)\left(\ln \frac{4e^{2\tau}}{\delta} + d_L \ln \frac{12}{\tau^2}\right). \tag{13}$$

Note that, from the assumptions of the theorem, it is easy to see that $g_L(\cdot,\cdot)$ can be written as a Boolean expression consisting of Lk polynomials $\tilde{\beta}_1(\theta, w), \ldots, \tilde{\beta}_{Lk}(\theta, w)$, and the degree with respect to $\theta \in \mathbb{R}^{n_\theta}$ of all these polynomials is no larger than α. Therefore, using Lemma 1 we obtain that $\mathrm{VC}_{g_L} \leq 2n_\theta \log_2(4e\alpha Lk)$. Thus, $d_L = 2n_\theta \log_2(4e\alpha Lk)$ can be adopted as a upper bound of the VC-dimension of $g_L(\cdot,\cdot)$.

To minimize the number of samples we fix τ to 0.824. This value corresponds to the numerical minimization of the term $-\left(\frac{\ln(1-\tau)}{\tau^2}\right)\ln\frac{12}{\tau^2}$. This choice of τ and the inequality $-\ln(1-\epsilon) > \epsilon$ leads to the bound

$$N \geq 1.768 \left\lceil \frac{1.738}{\epsilon} \right\rceil \left(\ln \frac{20.79}{\delta} + 5.75 n_\theta \log_2\left(4e\alpha k \left\lceil \frac{1.738}{\epsilon} \right\rceil\right)\right).$$

Since $\epsilon \in (0, 0.25)$, we have

$$\left\lceil \frac{1.738}{\epsilon} \right\rceil \leq \frac{1.738 + \epsilon}{\epsilon} \leq \frac{1.738 + 0.25}{\epsilon} < \frac{2}{\epsilon}.$$

Substituting this inequality in the previous bound we obtain

$$N \geq \frac{3.54}{\epsilon}\left(\ln \frac{20.79}{\delta} + 5.75 n_\theta \log_2 (\frac{8e\alpha k}{\epsilon})\right). \tag{14}$$

We now consider separately the two cases $\epsilon \leq \frac{1}{2e\alpha k}$ and $\epsilon > \frac{1}{2e\alpha k}$. First, if $\epsilon \leq \frac{1}{2e\alpha k}$, or equivalently $2e\alpha k \leq \frac{1}{\epsilon}$ we get

$$\log_2 \frac{8e\alpha k}{\epsilon} \leq \log_2 \frac{4}{\epsilon^2} = 2 \log_2 \frac{2}{\epsilon}.$$

Therefore, in this case the bound is

$$N \geq \frac{3.54}{\epsilon}\left(\ln \frac{20.79}{\delta} + 11.5 n_\theta \log_2 \frac{2}{\epsilon}\right).$$

Consider now the case $\epsilon > \frac{1}{2e\alpha k}$. Hence, $\frac{1}{\epsilon} < 2e\alpha k$ and

$$\log_2 \frac{8e\alpha k}{\epsilon} \leq \log_2 16(e\alpha k)^2 = 2 \log_2 (4e\alpha k).$$

Taking into account that $d = 2n_\theta \log_2 (4e\alpha k)$, the following bound is inferred from equation (14)

$$N \geq \frac{3.54}{\epsilon}\left(\ln \frac{20.79}{\delta} + 5.75 d\right). \qquad \blacksquare$$

Note that using the equation (13) in the proof of Theorem 4, more general results in which ϵ is not constrained to belong to the open interval $(0, 0.25)$ can be derived. In this sense, the restriction $\epsilon \in (0, 0.25)$ can be relaxed at the expense of obtaining larger constants appearing in the bound $\bar{N}(\epsilon, \delta)$.

6.1 The Scenario Approach

In this subsection, we study the so-called scenario approach for robust control introduced in [4, 5]. We address the semi-infinite optimization problem

$$\min_{\theta \in \Theta} J(\theta) \text{ subject to } g(\theta, w) = 0, \text{ for all } w \in \mathcal{W} \tag{15}$$

for 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 (15) is unique[1] and the level parameter ρ is equal to zero. The first two assumptions are now stated precisely.

Assumption 1. [convexity] *Let $\Theta \subset \mathbb{R}^{n_\theta}$ be a convex and closed set, and let $\mathcal{W} \subseteq \mathbb{R}^{n_w}$. We assume that*

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

where $f(\theta, w) : \Theta \times \mathcal{W} \to [-\infty, \infty]$ is convex in θ for any fixed value of $w \in \mathcal{W}$.

[1] We remark that this uniqueness assumption can be relaxed in most cases, as shown in Appendix A of [5].

Assumption 2. [uniqueness] *The optimization problem (8) is either unfeasible, or, if feasible, it attains a unique optimal solution for all possible multisample extractions* $\{w^{(1)}, \ldots, w^{(N)}\}$.

We now state a result which is based on the pack-based strategy previously introduced and on the proof of Theorem 1 of [5].

Corollary 1. *Let Assumptions 1 and 2 be satisfied. Given* $\epsilon \in (0,1)$ *and* $\delta \in (0,1)$, *if*

$$N \geq \frac{2}{\epsilon} \ln \frac{1}{2\delta} + 2n_\theta + \frac{2n_\theta}{\epsilon} \ln 4 \tag{16}$$

then, with probability no smaller than $1-\delta$, *either the problem (8) is unfeasible and, hence, also the semi-infinite convex optimization problem (15) is unfeasible; or, (8) is feasible, and then its optimal solution* $\hat{\theta}_N$ *satisfies the inequality* $E_g(\hat{\theta}_N) \leq \epsilon$.

Proof. This corollary is proved in [2].

In [5] it is shown that under the same assumptions of Corollary 1, it suffices to make N greater or equal than the bound

$$N \geq \left\lceil \frac{2}{\epsilon} \ln \frac{1}{\delta} + 2n_\theta + \frac{2n_\theta}{\epsilon} \ln \frac{2}{\epsilon} \right\rceil. \tag{17}$$

Corollary 1 constitutes another example of how the proposed *pack-based strategy* can be used to improve some of the sample size bounds available in the literature. In this particular case, the factor $\frac{2n_\theta}{\epsilon} \ln \frac{2}{\epsilon}$ in (17) is replaced with $\frac{2n_\theta}{\epsilon} \ln 4$.

7 Conclusion

In this paper, we proposed a new technique, denoted as *pack-based strategy*. This technique allows one to derive improved sample sizes bounds in the context of randomized algorithms. Using as starting point two-sided results from statistical learning theory, we obtain a new bound, that can be applied to optimization problems in which the robust constraints can be rewritten as Boolean expressions consisting of polynomials. This bound significantly improves upon those currently utilized in the control community. In particular, we have shown that the number of required samples grows with the accuracy parameter ϵ as $\frac{1}{\epsilon} \ln \frac{1}{\epsilon}$ instead of the usual $\frac{1}{\epsilon^2} \ln \frac{1}{\epsilon^2}$ dependence. Moreover, we have shown that the VC-dimension plays a secondary role. More precisely, when the accuracy parameter is sufficiently small, an explicit bound that only depends on the number of decision variables, and on the confidence and accuracy parameters is presented.

Acknowledgements

The authors acknowledge MCYT-Spain for funding this work (contracts DPI2007-66718-C04-01 and DPI2005-04568).

References

1. Alamo, T., Tempo, R., Camacho, E.F.: Revisiting statistical learning theory for uncertain feasibility and optimization problems. In: Proceedings of the 46th IEEE Conference on Decision and Control, New Orleans, USA (December 2007)
2. Alamo, T., Tempo, R., Camacho, E.F.: Improved sample size bounds for probabilistic robust control design: a pack-based strategy. In: Proceedings of the 46th IEEE Conference on Decision and Control, New Orleans, USA (December 2007)
3. Alamo, T., Tempo, R., Rodríguez, D., Camacho, E.F.: A sequentially optimal randomized algorithm for robust LMI feasibility problems. In: Proceedings of the European Control Conference, 2007, Kos, Greece (July 2007)
4. Calafiore, G., Campi, M.C.: Uncertain convex programs: randomized solutions and confidence levels. Math. Program. 102, 25–46 (2005)
5. Calafiore, G., Campi, M.C.: The scenario approach to robust control design. IEEE Transactions on Automatic Control 51(5), 742–753 (2006)
6. Calafiore, G., Dabbene, F.: An iterative localization method for probabilistic feasibility of uncertain LMIs. In: Proceedings of the 45th IEEE Conference on Decision and Control, San Diego, USA (December 2006)
7. Calafiore, G., Polyak, B.T.: Stochastic algorithms for exact and approximate feasibility of robust LMIs. IEEE Transactions on Automatic Control 46(11), 1755–1759 (2001)
8. 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)
9. Fujisaki, Y., Dabbene, F., Tempo, R.: Probabilistic design of LPV control systems. Automatica 39, 1323–1337 (2003)
10. 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)
11. Hoeffding, W.: Probability inequalities for sums of bounded random variables. Journal of American Statistics Association 58, 13–30 (1963)
12. Kanev, S., De Schutter, B., Verhaegen, M.: An ellipsoid algorithm for probabilistic robust controller design. Systems and Control Letters 49, 365–375 (2003)
13. 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)
14. Polyak, B.T., Tempo, R.: Probabilistic robust design with linear quadratic regulators. Systems & Control Letters 43, 343–353 (2001)
15. 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)
16. Tempo, R., Calafiore, G., Dabbene, F.: Randomized Algorithms for Analysis and Control of Uncertain Systems. Communications and Control Engineering Series. Springer, London (2005)
17. Vapnik, V.N.: Statistical Learning Theory. John Wiley & Sons, New York (1998)
18. Vapnik, V.N., Chervonenkis, A.Ya.: On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications 16, 264–280 (1971)
19. Vapnik, V.N., Chervonenkis, A.Ya.: Necessary and sufficient conditions for the uniform convergence of means to their expectations. Theory of Probability and its Applications 26(3), 532–553 (1981)

20. Vidyasagar, M.: A Theory of Learning and Generalization: with Applications to Neural Networks and Control Systems. Springer, London (1997)
21. Vidyasagar, M.: Statistical learning theory and randomized algorithms for control. Control Systems Magazine 18(6), 69–85 (1998)
22. Vidyasagar, M.: Randomized algorithms for robust controller synthesis using statistical learning theory. Automatica 37, 1515–1528 (2001)
23. Vidyasagar, M., Blondel, V.D.: Probabilistic solutions to some NP-hard matrix problems. Automatica 37(9), 1397–1405 (2001)

A Appendix

Appendix 1. *Proof of Lemma 2*

Two cases should be considered $\tau \leq \epsilon$ and $\tau > \epsilon$. The proof for the first case is trivial because $E_g(\theta) \leq E_{g_L}(\theta)$, for all $L \geq 1$. This implies that $E_g(\theta) \leq E_{g_L}(\theta) \leq \tau \leq \epsilon$. In what follows, the second case will be analyzed.

By definition, we obtain

$$E_g(\theta) = \Pr_{\mathcal{W}} \{ w \in \mathcal{W} : g(\theta, w) = 1 \}.$$

This means that the probability of drawing a collection of L samples $\mathrm{w} = \{w^{(1)}, \ldots, w^{(L)}\}$ and obtaining

$$g_L(\theta, \mathrm{w}) = \max_{i=1,\ldots,L} g(\theta, w^{(i)}) = 0$$

is equal to $(1 - E_g(\theta))^L$. From this we have that the pack probability of violation satisfies $E_{g_L}(\theta) = 1 - (1 - E_g(\theta))^L$. As $E_{g_L}(\theta) \leq \tau$ it follows that

$$1 - \tau \leq (1 - E_g(\theta))^L. \tag{18}$$

Since $(1 - E_g(\theta)) \in [0, 1]$ and L is supposed to satisfy the inequality

$$L \geq \left\lceil \frac{\ln(1-\tau)}{\ln(1-\epsilon)} \right\rceil \geq \frac{\ln(1-\tau)}{\ln(1-\epsilon)},$$

it is inferred from equation (18) that

$$1 - \tau \leq (1 - E_g(\theta))^{\frac{\ln(1-\tau)}{\ln(1-\epsilon)}}.$$

Applying the natural logarithm to both sides of the inequality, we obtain

$$\ln(1-\tau) \leq \left(\frac{\ln(1-\tau)}{\ln(1-\epsilon)} \right) \ln(1 - E_g(\theta)).$$

Multiplying both sides of previous inequality by $\dfrac{\ln(1-\epsilon)}{\ln(1-\tau)} > 0$, the following relations are inferred

$$\begin{aligned} \ln(1-\epsilon) &\leq \ln(1 - E_g(\theta)) \\ 1 - \epsilon &\leq 1 - E_g(\theta) \\ E_g(\theta) &\leq \epsilon. \end{aligned}$$

■

UAV Formation Control: Theory and Application*

Brian D.O. Anderson[1], Barış Fidan[1], Changbin Yu[1], and Dirk van der Walle[2]

[1] Research School of Information Sciences and Engineering, The Australian National University and National ICT Australia, Canberra, Australia
{Brian.Anderson,Baris.Fidan,Brad.Yu}@anu.edu.au

[2] Delft Center for Systems and Control, Delft University of Technology, Delft, The Netherlands
D.vanderWalle@student.tudelft.nl

Summary. Unmanned airborne vehicles (UAVs) are finding use in military operations and starting to find use in civilian operations. UAVs often fly in formation, meaning that the distances between individual pairs of UAVs stay fixed, and the formation of UAVs in a sense moves as a rigid entity. In order to maintain the shape of a formation, it is enough to maintain the distance between a certain number of the agent pairs; this will result in the distance between all pairs being constant. We describe how to characterize the choice of agent pairs to secure this shape-preserving property for a planar formation, and we describe decentralized control laws which will stably restore the shape of a formation when the distances between nominated agent pairs become unequal to their prescribed values. A mixture of graph theory, nonlinear systems theory and linear algebra is relevant. We also consider a particular practical problem of flying a group of three UAVs in an equilateral triangle, with the centre of mass following a nominated trajectory reflecting constraints on turning radius, and with a requirement that the speeds of the UAVs are constant, and nearly (but not necessarily exactly) equal.

Keywords: Formation control, surveillance, UAV, rigid formation, persistent formation.

1 Introduction

Today, technology allows us to mimic the behavior of insects, animals, birds, etc. [1, 2] using robots, unmanned airborne vehicles (or indeed regular aircraft, underwater autonomous vehicles, and the like). The range of applications is steadily growing, and it includes military and civilian applications, very often involving surveillance or exploration of some region.

Why use a formation? When the agents in a formation, unmanned airborne vehicles (UAVs) for example, are engaged in surveillance or exploration activity, they are typically able to synthesize an antenna of dimension far larger than an individual agent. The benefit is the improved sensitivity. If source localization

* This work is supported by National ICT Australia, which is funded by the Australian Government's Department of Communications, Information Technology and the Arts and the Australian Research Council through the Backing Australia's Ability Initiative.

V.D. Blondel et al. (Eds.) Recent Advances in Learning and Control, LNCIS 371, pp. 15–33, 2008.
springerlink.com

is of concern, not only is the improved sensitivity beneficial, but additionally, some localization tasks inherently require multiple sensors with known relative positions. A different reason is that multiple sensors may have different functionalities, and the aggregate functionality may give a new functionality for the formation of sensors. Again, many sensors, such as UAVs or sensors in a sensor network, are weight-constrained, and small mobile sensors can be cheaper to deploy.

Sometimes several of these factors can be simultaneously operative. We consider later in this chapter by way of a case study an application involving a cooperative UAV surveillance task. The agents determine angle information associated with the object at an unknown position. There is sensor noise and a limited cone of visibility. So more agents need to be used in the formation than might at first be imagined.

Control of a formation requires the mixing of several tasks. One is the whole formation task of moving from point A to point B (or moving the centre of mass of the formation, and adopting a certain orientation). Another is to maintain the relative positions of the agents during formation motion, so that the shape is preserved. A third might be to avoid obstacles. A fourth might be to split the formation, etc. Generally speaking, in nature and a number of man-made systems, there is no single master agent which oversights or controls every other agent. Some kind of decentralized handling of the control task is needed, and therein lies much of the scientific challenge associated with artificial formations.

Indeed, the highest level problem seen from the viewpoint of a control engineer is probably how to define practical architectures for the formation, or the dependencies and signal flows associated with communications, sensing and control. A key requirement is that any architecture be scalable. The number of communication links required for a single agent should *not* grow linearly with the number of agents in the formation for example. This scalability requirement is indeed a driver of the decentralized approach. It is clear that nature respects this too. No one bird in a formation of birds would be expected to watch every other bird, and determine where it will fly using its perception of where every other bird in the formation is flying. In this chapter, we mainly focus on the question of what sort of sensing and control architectures are needed to maintain the shape of a formation, while the formation moves as a cohesive whole. We shall present some details on types of control law that can ensure such motion occurs.

Furthermore, we will impose yet one more level of limitation. Clearly, if a group of agents is maintaining formation shape, each must know the desired value of certain geometric variables and must sense the actual values of those geometric variables, in order to generate any corrective control. Many variables can be sensed, generally involving some combination of distances and angles, and it is possible that one agent may sense more than one *class* of measurement. See, for example, [3]. For the most part, we shall contemplate two dimensional formations and require that a typical agent in two dimensions use at least two distance measurements to neighboring agents to maintain its position (and often

also sensing the relative position of these neighboring agents) assuming it has no degrees of freedom through being a leader or co-leader of the formation[1].

The chapter is organized as follows. In Section 2, we present and elaborate two different control structures that can be used in maintaining shapes of autonomous formations, the symmetric and asymmetric control structures, and characteristics of autonomous formations controlled by such control structures. Two particular graph theoretical notions are introduced and used in analysis in this section: graph (formation) rigidity and persistence. Using such ideas, in Section 3, we summarize some recent results on decentralized control design to ensure preservation of the shape of a formation, particularly focusing on autonomous formations with asymmetric control structure. Sections 4–6 are dedicated to a particular autonomous formation control application: cooperative surveillance with a team of three UAVs. Section 4 introduces the particular cooperative surveillance task considered and presents the specifications of the corresponding formation control problem. A distributed control scheme for this problem is proposed with details in Section 5. The proposed control scheme is analyzed for a number of scenarios via simulations in Section 6. As will become evident in these later sections, we draw on various aspects of the rather idealized studies of Section 3 in order to cope with the practical constraints imposed in the application. The chapter finishes with a brief summary and concluding remarks in Section 7.

2 Autonomous Formations and Shape Maintenance

When distance control between pairs of agents is used, there are two approaches to controlling each distance, symmetric and asymmetric. Formations with symmetric control structure can be represented with undirected graphs, while directed graphs are used to represent those with asymmetric control structure.

When a symmetric control structure is used, to keep the distance between each pair (A_1, A_2) of neighbor agents, there is a joint effort of both agents A_1 and A_2 to simultaneously and actively maintain their relative positions. The underlying graph of the formation will have an undirected edge between vertices 1 and 2 (which represent agents A_1 and A_2). Evidently, if enough agent pairs explicitly maintain distances, all inter-agent distances and hence the formation shape will be maintained, i.e. the formation will be rigid. Just what constitutes 'enough' will be explained in more detail subsequently.

In contrast, when an asymmetric control structure is used, only one of the agents in each neighbor agent pair, e.g., A_1 in the agent pair (A_1, A_2), actively maintains its position relative to its neighbor. This means for the neighbor agent pair (A_1, A_2) that only A_1 has to receive the position information broadcast by A_2, or sense the position of A_2 and make decisions on its own. Therefore, in

[1] When an agent is required to maintain distances from two other agents, it will generally need relative position information, or equivalently distances to the two agents and the angle between them. The angle information may be derivable by running a Kalman filter using distance measurements, or using an additional distance measurement, viz. that between its two neighbors.

the asymmetric control structure, both the overall control complexity and the communication complexity (in terms of message sent or information sensed) for the formation are reduced by half, compared to symmetric control. (One could argue however that redundancy is lost, and this may be dangerous). From a graph theory point of view, this asymmetric control structure is modelled in the associated directed graph by a directed edge from vertex 1 to vertex 2 (representing a directed constraint link from A_1 to A_2).

In this section we provide some detail on the characteristics of 2-dimensional formations with symmetric and asymmetric control structures, mainly using *rigid graph theory* [4, 5]. Some but not all of the ideas are valid for 3-dimensional formations. However, we shall omit discussion of the 3-dimensional case in the interest of brevity.

2.1 Formations with Symmetric Control Structure

To fit within the (rigid) graph theory framework, a multi-agent system model is used where agents are assumed to be points in $\Re^2$, and agent pairs for which the inter-agent distance is actively constrained to be constant are thought of as being joined by bars with lengths enforcing the inter-agent distance constraints [6,7,8,9]. Evidently, the multi-agent system can be modeled by a graph: vertices represent point-like agents and edges join vertex pairs corresponding to agent pairs with prescribed inter-agent distance constraints. (Naturally, one can contemplate other constraints than just distance, e.g. those involving angle, or angle and distance, although discussion is omitted in this chapter for the sake of simplicity.) Rigid graph theory is used to study properties of graphs which ensure that the formation being modelled by the graph will be rigid; formal definitions are available of course, but in rough terms, a *rigid formation* is one in which the only smooth motions are those corresponding to translation or rotation of the whole formation.

There exist two key tool sets for rigidity analysis. Linear algebra and matroid theory provide the first: given knowledge of the positions of the agents at any one time, one can construct a matrix, the so-called *rigidity matrix* [4], and the dimensions and rank of this matrix allow one to conclude that the formation is or is not rigid. The dimensions and rank are the same for almost all positions of the agents, i.e. for generic agent positions. This means that rigidity matrices formed from two formations differing from each other only in terms of the values for the constrained distances will have the same rank, except for very special sets of the agent positions. Examples of special sets include those where two agents are at the same position, three agents are collinear, four agents form a parallelogram, etc.

The second tool set is a combinatorial one, i.e. it involves a number of purely counting-type characterization conditions. The key result is Laman's Theorem [10]. This central contribution of the rigid graph theory literature implies that it is also possible in 2 dimensions to characterize rigidity of a generic formation corresponding to a given graph in purely combinatorial conditions related to the graph (discarding the agent coordinates). By generic formation, we mean one where the agent positions are generic.

A rigid formation is further called *minimally rigid* if the removal of any single inter-agent distance constraint causes loss of rigidity. Unsurprisingly, a graph is called minimally rigid if almost all formations to which the graph corresponds are minimally rigid. Minimal rigidity is easily characterizable in 2-dimensions both with the rigidity matrix and with Laman's Theorem. An easily checked necessary condition is that $|E| = 2|V| - 3$, where $|E|$ and $|V|$ are the numbers of edges and vertices of the graph. Thus in a minimally rigid graphs the number of edges and vertices have the same order.

The formation shape of a minimally rigid formation (assuming $N > 0$ agents) is maintained using the minimum possible number of information links (for the given number of agents, N), which makes use of minimally rigid information architectures an optimal (i.e. most efficient) choice. However, in many cases, there are practical reasons to use non-minimally rigid graphs rather than minimally rigid ones to underpin the shape of a formation. A minimally rigid formation offers no protection against loss of a sensor, or a communication link, or a control actuator, and in practice, it will often be necessary to obtain robustness through the use of some measure of redundancy. Measures of robustness are needed to reflect ability to sustain agent and link losses (whether from a sensing, communication or control failure) [11].

2.2 Formations with Asymmetric Control Structure

This subsection comments on extensions of the concepts and results presented in Section 2.1 to formations with asymmetric control structure. The basic task is again maintenance of formation shape during motion, i.e. ensuring that the smooth motions of the formation are restricted to translation or rotation. This task is examined in detail in [12] and [9].

We call a formation with asymmetric control structure *persistent* if it is rigid (where rigidity is as explained in Section 2.1) and satisfies another condition termed *constraint consistence* [12, 9]; constraint consistence is equivalent to the requirement that *it is possible* to maintain the nominated inter-agent distances. (Note that the rigidity property only says that *if* certain inter-agent distances are maintained, then all inter-agent distances are maintained when the formation moves smoothly. Moreover, it is possible to assign directions to the edges of a rigid graph that render it not constraint consistent.) A *minimally persistent formation* is one that is minimally rigid and constraint consistent. Formal definitions and properties of constraint consistence and persistence can be found in [12, 9], where directed graph notions are used.

3 Control Laws for Minimally Persistent Formations

In this section, we summarize some recent results on the construction of decentralized control laws to ensure that the shape of a formation is preserved. The key references are [13, 14, 15]. However, we must not discount other work in this area. Among the earliest key studies linking graph theory nontrivially to

problems of stabilization of the associated formations were those of Olfati-Saber and Murray [16, 17]. A very recent work with close links is [18]. This reference considers rigid and persistent formation control.

Unsurprisingly, it is generally possible to separate the navigation task for the formation (getting from A to B) from the shape stabilization task, and in this section we will focus *solely* on the latter task. The first of the Olfati-Saber and Murray papers provides laws based on potential functions applying to undirected graphs, and leaves open the directed case. The second of the cited papers of Olfati-Saber and Murray deals extensively with formations for which the underlying (directed) graph was acyclic. It turns out that formation stabilization in this acyclic case is more straightforward: decentralization of the control laws is easy because there is one-way-only or triangular coupling between the agents. References [13,14,15] are concerned with formations where the underlying graph includes one or more cycles.

More complicated agent models than the point models used here are dealt with in some literature. See e.g. [19,20]. We note also other problems of holding a formation of a specialized shape, e.g. equilateral polygon [21]. If all agents are equipped with a compass, stabilization is easier, [8]; that assumption will not be made here.

The key assumptions made are as follows: the formation exists in the plane, agents are point agents, with velocities controllable; agents can measure the distance to their neighbors, they know the angle between any two neighbors, and they know the desired distance to each of their neighbor(s). We restrict attention to *minimally* persistent formations, and we consider first the simplest formation containing a cycle–a triangular formation of three agents.

The potential for instability when a cycle is present should be reasonably clear. If agents i, j, k form a cycle, i tries to fix its position relative to j, j tries to fix its position relative to k, and k tries to fix its position relative to i, then there is a clear feedback mechanism around the cycle. Without analysis, one does not know if the feedback will have positive or negative effects, including destabilization.

3.1 Control of a Triangular Formation with Three Coleaders

This subsection summarises results of [14, 15]. Related results can be found in [18]. Denote the three agents by 1,2 and 3, and suppose their positions at any instant of time are denoted by $x_i, i = 1, 2, 3$. Suppose the nominal distances from 1 to 2, 2 to 3 and 3 to 1 are d_1, d_2, d_3, and suppose these distances satisfying the triangle inequality. Let z_1, z_2, z_3 denote the relative positions of 1 with respect to 2, etc, i.e. $z_1 = x_1 - x_2$ with two similar equations. The formation stabilization task is to ensure that $\lim_{t\to\infty} ||z_i(t)|| - d_i = 0, i = 1, 2, 3$. The question arises as to whether this behavior can reasonably be expected for all initial conditions, or only those for which $|||z_i(t)|| - d_i|$ is small; as it turns out, it is possible to secure this behavior for almost all initial conditions.

The control laws of [14,15], while different, have a common form. The law is

$$\dot{x}_i = -k_i z_i \tag{1}$$

In [14] and [15], the gains k_i are respectively:

$$k_i = [||z_i|| - d_i]/||z_i|| \tag{2}$$

$$k_i = [||z_i||^2 - d_i^2] \tag{3}$$

Clearly, movement of agent i is always directly towards or directly away from its neighbor, with the direction of movement such as to reduce the error between the actual and desired distance. This is intuitively reasonable. The gain differs between the two algorithms. Actually, they are both special cases of the algorithm

$$\dot{x}_i = -\phi_i(e_i, d_i)z_i \tag{4}$$

where ϕ_i for fixed d_i is a smooth first-third quadrant nonlinearity in $e_i = ||z_i||^2 - d_i^2$ (but not defined for $e_i < -d_i^2$).

Evidently, the whole system is nonlinear in a nontrivial way, and its analysis is by no means straightforward. Despite this, much can be shown. The following properties are probably true for the general family of laws just identified; however, at the moment, the properties have just been established for one at least of the two particular laws.

Property 1. Suppose that the initial positions of all agents are such that they are not collinear. If no two of the d_i are the same, and the second control law is used, the errors $e_i = ||z_i||^2 - d_i^2, i = 1, 2, 3$ will all converge exponentially fast to zero, or equivalently the correct triangle shape is assumed exponentially fast.

Property 2. If the initial positions result in the correct triangle shape being assumed exponentially fast, then the total translation of the triangle's centre of mass, and the total rotation of the triangle about the centre of mass over the interval $[0, \infty)$ are bounded, and all agents come to rest.

Property 3. If the initial positions of the agents are collinear, the subsequent trajectory will retain the collinearity property. However, the associated manifold is not stable; hence a random perturbation will result in the property 1 situation applying.

Property 4. If the initial positions of the agents are collinear but are otherwise generic, under the second law the agents will assume a common nonzero velocity exponentially fast.

3.2 Control of a Minimally Persistent Formation with Leader and First Follower

For this subsection, all graphs have at least $N \geq 3$ vertices. With the minimally persistent property, the edge count is precisely $2N - 3$. A further consequence of minimal persistence [9] is that vertices have an out-degree of at most 2, and apart from such vertices either there are exactly three vertices with an out-degree of 1, or one vertex with an out-degree of zero and one with an out-degree of 1. In the latter case, the zero out-degree vertex is termed a *leader*. The corresponding formation agent, having no neighbor from which its distance must

be maintained, is not constrained at all in its motion. When the out-degree 1 vertex has the leader as its neighbor, it is termed a *first follower*. The graphs considered here are leader-first-follower graphs. It turns out that some results for minimally persistent graphs without a leader-first-follower structure are, or appear to be, easily derivable from the result for a leader-first-follower structure. Hence leader-first-follower graphs are a suitable subclass to investigate initially. Also, in this subsection we restrict attention to graphs which contain one or more cycles, for reasons explained earlier.

Evidently before one could contemplate formation control for a non-minimally persistent graph, one would probably have to be able to treat minimally persistent graphs, and this might be the basis for treating non-minimally persistent graphs, though this has yet to be done.

By way of a final qualifying remark, in this subsection we are only able to present control laws for restoring a formation to its correct shape when it has undergone a *small* perturbation from its correct shape. The entire analysis is a linear one, and assumes that a linearized model is a valid approximation. Thus in comparison to the previous section, the vertex number is general but the convergence result is weaker.

We now explain the system model and outline the structure of the decentralized control law. We focus first on an agent j which has two neighbors, agents k and m, from which it must maintain its distance by amounts known to it. Suppose that the three agents concerned are all displaced from their nominal positions, call them x_{0j}, x_{0k}, x_{0m}, by small amounts $\delta x_j, \delta x_k, \delta x_m$. Assuming it can measure the relative positions of its neighbor agents, agent j determines to where it would need to move, in order to be at the correct distance from agents k and m, (assuming the latter do not move while it is moving). Identify this target position as $x_{0j} + \delta^* x_j$. Note that $\delta^* x_j$ is a function of δx_k and δx_m. Agent j actually moves to reduce the distance between where it currently is and this target position, by using the following law:

$$\dot{\delta x_j} = A_j(\delta^* x_j(\delta x_k, \delta x_m) - \delta x_j) \tag{5}$$

with the following condition guaranteeing the distance reduction property

$$A_j + A_j^T > 0 \tag{6}$$

By simple geometric arguments, it is possible to express δx_j^* in terms of $\delta x_k, \delta x_m$ and the nominal agent positions x_{0j}, x_{0k}, x_{0m}. There results

$$\dot{\delta x_j} = A_j B_j \begin{bmatrix} \delta x_j \\ \delta x_m \\ \delta x_k \end{bmatrix} \tag{7}$$

where B_j is a matrix depending on differences of the position coordinates x_{0j}, x_{0k}, x_{0m}, and is actually a submatrix of the rigidity matrix of the formation. Similar equations can be written for every agent, and for the leader and first follower. For details, the reader is referred to [13,14]. Putting all equations

together and with the first follower and leader labelled as vertices $N-1$ and N, there results:

$$\frac{d}{dt}\begin{bmatrix}\delta x_1\\ \delta x_2\\ \vdots\\ \delta x_N\end{bmatrix} = \Gamma\begin{bmatrix}R\\ 0_{3\times 2n}\end{bmatrix}\begin{bmatrix}\delta x_1\\ \delta x_2\\ \vdots\\ \delta x_N\end{bmatrix} \tag{8}$$

where Γ is obtained by stacking together into a block diagonal matrix the 'gain' matrices A_j each multiplied by a nonsingular 2×2 matrix , and R is actually the rigidity matrix of the formation with agents at their nominal locations.

It is convenient to drop the last three rows of this equation (and the last three columns of R), to form a smaller size equation. The associated modes are attributable to the *degrees of freedom (DOF)*[2] of the leader and the first follower, which allow the whole formation to translate or rotate. They are not used for shape stabilization however. The key question now is: how should the entries of Γ –a block diagonal matrix whose block entries are the A_j multiplied by known 2×2 nonsingular matrices–be chosen, in order that the linear equation set be stable. This will then ensure that when the agents of the formation are displaced a small distance from their nominal positions, the corrective action taken by each agent will drive the displacements to zero.

It is incidentally easy to find examples where the choice of $A_j = I$ is *destabilizing*. Thus the problem is nontrivial.

The key is the following result, [13,22], which actually comes with a constructive proof:

Theorem 1. *Consider the linear differential equation $\dot{x} = \Lambda A x$ in which a square real A is prescribed and Λ is diagonal, real and otherwise adjustable. Then a sufficient condition that there exists a choice of Λ such that the equation is asymptotically stable, i.e. the eigenvalues of ΛA all have negative real parts, is that the leading principal minors of A are nonzero.*

One can show further that full eigenvalue positionability is generically impossible via choice of Λ in a number of cases where stabilization is possible.

The matrix Λ of the theorem corresponds to Γ in (8), minus its last three rows and columns. Notice that Γ, being *block* diagonal, has more adjustable entries than Λ, which is strictly diagonal, and it is an open question as to how to exploit this fact. The matrix A of the theorem corresponds to the rigidity matrix R less its last three columns. A far from trivial argument appealing to various properties of minimally persistent graphs, see [13, 22], guarantees the leading principal minors are nonzero (perhaps after vertex reordering).

Evidently, the preceding results should only be the first of many. We still need to know which formations are easy to control, and to understand how to exploit the freedom in the choice of control laws for each agent to achieve some kind of tradeoffs. We need results which are not just local or linearized, and we would like results

[2] In $\Re^2$, a vertex has two, one or zero DOF(s) if it has no, one, or at least two outgoing edges; each outgoing edge uses up one DOF. A minor variation applies in 3 dimensions.

which can deal with agents with their own dynamics. One can also envisage control laws other than memoryless ones. Actually, a particular decentralized proportional integral control scheme is presented in Section 5 for the particular applications case study of the chapter, cooperative surveillance with UAV formations.

4 An Application: Surveillance with UAV Formations

In this section, as an application domain of UAV formation control, we consider the task of cooperative surveillance over a 2-dimensional region of interest using a non-hierarchical formation of three autonomous UAVs. We derive a distributed control scheme for the formation to maintain a set of desired inter-agent distances within the formation while moving on a pre-defined path, under the assumption of constant UAV speeds. The particular formation control structure to be adopted is the asymmetric one described in Section 2.2, with the three-coleader form described in Section 3.1. However, note again that the individual control schemes of the agents in this section are designed in the proportional-integral form as opposed to the designs presented in Section 3.1.

4.1 System Specification

The particular cooperative surveillance task to be considered is surveillance of a specified region of interest using a fleet of three UAVs with passive direction-finding sensor payloads, flying in an equilateral triangle formation.

This task is part of a research challenge problem posed by the Australian Defence Science and Technology Organisation (DSTO) with further specifications. One practical motivation for this research challenge is accurate cooperative localization of ground-based radar systems with small-size UAV fleets [23]. The cooperative surveillance task is planned to be experimented by DSTO using autonomous UAVs of the class *Aerosonde*, a small UAV developed by Aerosonde Pty Ltd [24]. An Aerosonde UAV typically has a wing span of 2.9 m and a maximum take-off mass of 15 kg. It can stay flying for 8–30 hours, depending on the payload it carries, without refuelling.

The speed of each Aerosonde UAV, in the prospective formation flight test, is preset to a certain constant value (between 20 m/s and 32 m/s) before take-off and is kept almost constant during the flight. However the constant speeds of different UAVs may be different, mainly due to different payloads. In our studies we assume a nominal constant speed of 32 m/s and, in some cases, consider mild variations from this nominal value for some of the UAVs. Complying with the preset constant nominal speed of 32 m/s and minimum turning radius of 400 m, the maximum turning rate is specified to be 0.08 rad/s.

The region of interest for surveillance is assumed to be a square with 30 km side length. For accuracy of localization of targets it is required to keep the inter-agent separation distances sufficiently large, and for coordination purposes the equilateral triangle formation shape and size is required to be maintained constant. Each of the three inter-agent distances is desired to be kept at 3 km.

4.2 Surveillance Task on a Spiral Path

For the surveillance task described in Section 4.1, we consider a particular path to be followed by the center of mass (CM) of the formation, an Archimedean spiral path originating from the center of the 30 km×30 km region of interest. This spiral path can be formulated in time-indexed form in 2-dimensional polar coordinates $(r, \bar{\theta})$ corresponding to the cartesian coordinates $(x, y) = (r\cos\bar{\theta}, r\sin\bar{\theta})$, with the center of the region of interest located at $(a, 0)$ (in both polar and cartesian coordinates), as

$$\begin{aligned} r(t) &= a + b\bar{\vartheta}(t) \\ \bar{\theta}(t) &= \bar{\vartheta}(t) \pmod{2\pi} \end{aligned} \tag{9}$$

where $\bar{\vartheta}(t)$ is a monotonically increasing function of t satisfying $\bar{\vartheta}(0) = 0$ and $\lim_{t\to\infty} \bar{\vartheta}(t) = \infty$, and the design constants $a, b \geq 0$ denote, respectively, the initial radial offset and the radial increase rate of the spiral. The main motivations for using a spiral path are as follows. The spiral path (9) originating from $(a, 0)$ (polar) scans the vicinity of $(a, 0)$ (polar) in a well-formulated polar form with constant increase rates of both the angle and radius. The periodicity rate of the scan angle $\bar{\theta}(t)$ (modulo 2π) and the increase rate of the scan radius $r(t)$ can be adjusted by selecting the spiral path parameters a, b accordingly. An agent, actual or fictitious, following this path will have a continuously increasing turning radius and hence continuously decreasing turning rate.

Now, consider a fictitious or virtual point agent A perfectly tracking the time-indexed spiral path (9) with constant speed $v(t) = v_c$ and agent kinematics

$$\begin{aligned} \dot{x}(t) &= v_c \cos(\theta(t)) \\ \dot{y}(t) &= v_c \sin(\theta(t)) \\ \dot{\theta}(t) &= \omega(t) \end{aligned} \tag{10}$$

where $p(t) = (x(t), y(t))$, $\theta(t)$, and $\omega(t)$ denote respectively the position, heading, and angular velocity of A at time instant $t \geq 0$. In this chapter, we define the *heading* of an agent moving in 2-dimensions as the counter-clock-wise angle from the cartesian x-axis to the motion direction vector of the agent.

Matching the virtual agent kinematics (10) with the spiral path equation (9) in order to satisfy the perfect tracking requirement, we have

$$\begin{aligned} x(t) &= r(t)\cos(\bar{\theta}(t)) \\ y(t) &= r(t)\sin(\bar{\theta}(t)) \\ \dot{x}(t) &= -r(t)\dot{\bar{\vartheta}}(t)\sin(\bar{\theta}(t)) + \dot{r}(t)\cos(\bar{\theta}(t)) \\ \dot{y}(t) &= r(t)\dot{\bar{\vartheta}}(t)\cos(\bar{\theta}(t)) + \dot{r}(t)\sin(\bar{\theta}(t)) \end{aligned}$$

Similarly, we have

$$v(t) = v_c \;=\; |\dot{p}(t)| = \sqrt{r^2(t)\dot{\bar{\vartheta}}^2(t) + \dot{r}(t))^2} = \dot{\bar{\vartheta}}(t)\sqrt{(a + b\bar{\vartheta}(t))^2 + b^2} \tag{11}$$

noting that $\dot{r}(t) = b\dot{\bar{\vartheta}}(t)$. In order to satisfy (11), $\dot{\bar{\vartheta}}(t)$ needs to be chosen as

$$\dot{\bar{\vartheta}}(t) = \frac{v_c}{\left((a + b\bar{\vartheta}(t))^2 + b^2\right)^{1/2}} \tag{12}$$

In the next section, we design a decentralized control scheme for the 3-UAV fleet mentioned in the beginning of this section to maintain the equilateral rigid formation described before while the CM of this formation is tracking the virtual agent A moving according to (9)–(12).

5 Non-hierarchical Formation Control for Surveillance

To design a decentralized control scheme meeting the spiral path tracking and formation control tasks specified in Section 4 we adopt the asymmetric control structure presented in Section 2.2. Furthermore, we assume the three-coleader formation structure presented in Section 3. Note here that a three-coleader (or non-hierarchical[3]) structure, as opposed to leader-follower (or hierarchical) structure, allows a balanced distribution of the tracking and formation maintenance tasks among the three agents (UAVs), and hence is expected to be more robust to speed variations of individual agents and atmospheric disturbances.

The formation maintenance and path tracking tasks are depicted in Fig. 1: The agents labelled as A_1, A_2,A_3, with respective positions $p_1(t), p_2(t), p_3(t)$ at each time instant t, are required to meet the inter-agent distance constraints $|p_1(t) - p_2(t)| = d_{12} = 3$ km, $|p_2(t) - p_3(t)| = d_{23} = 3$ km and $|p_3(t) - p_1(t)| = d_{31} = 3$ km, respectively, as well as keeping their distances to the virtual agent A (which is moving on the spiral path (9)) at $|p(t) - p_i(t)| = d_{ic} = \sqrt{3}$ km ($i \in \{1, 2, 3\}$) so that the formation CM traces (9).

5.1 Agent Models

Each agent A_i, $i \in \{1, 2, 3\}$, in compliance with the specifications of Section 4 is assumed to move with a constant speed $v_i(t) = v_{ci}$ and agent kinematics

$$\begin{aligned} \dot{x}_i(t) &= v_{ci}\cos(\theta_i(t)) \\ \dot{y}_i(t) &= v_{ci}\sin(\theta_i(t)) \\ \dot{\theta}_i(t) &= \omega_i(t) \end{aligned} \tag{13}$$

where $p_i(t) = (x_i(t), y_i(t))$, and $\theta_i(t)$ and $\omega_i(t)$ are respectively the heading and angular velocity of A_i at time instant $t \geq 0$. In the sequel we consider both of the cases (i) where $v_{c1} = v_{c2} = v_{c3}$ and (ii) where v_{ci} are different, and we

[3] Here, the term "hierarchy" is used in terms of leading/following (or path tracking/distance keeping) task distribution. For minimally persistent formations with three agents and 3-coleader structure, as opposed to the leader-follower case, note that this task distribution is uniform among the agents and hence there is no hierarchy in the above sense.

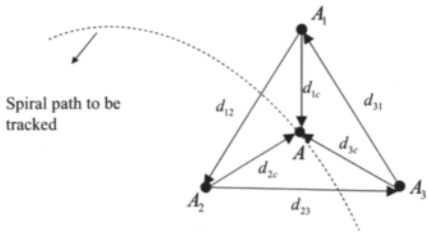


Fig. 1. Formation structure together with the formation control and path tracking tasks for the cooperative UAV surveillance task

demonstrate for case (ii) that there exist sensible upper bounds $\bar{\epsilon}_1, \bar{\epsilon}_2 > 0$ for which the formation maintenance is feasible, despite the speed difference between agents, if $|v_{c1} - v_{c2}| < \bar{\epsilon}_1$ and $|v_{c1} - v_{c3}| < \bar{\epsilon}_2$.

Each A_i is assumed to sense the location of itself as well as the agent it follows. Furthermore each A_i is assumed to know the spiral path to be tracked by the formation CM and hence the time trajectory $p(t)$ of the virtual agent A.

5.2 Control Design

We consider a non-hierarchical decentralized control design, i.e. we design three identical individual agent controllers, one for each of the agents A_1,A_2,A_3. For agent A_i $(i \in \{1,2,3\})$, the individual controller inputs at each time instant $t \geq 0$ are the locations $p_i(t)$,$p_j(t)$,$p(t)$ of, respectively, itself, the agent A_j it follows, and the virtual agent A introduced in Section 4.2. The controller output is the angular velocity $\omega_i(t)$ in (13). The corresponding feedback control law is designed in the proportional-integral (PI) form with proportional gain $k_P > 0$ and integrator gain $k_I > 0$ as

$$\omega_i(t) = \dot{\theta}_i(t) \quad = \quad k_P[\theta_{id}(t) - \theta_i(t)] + k_I \int_{t_0}^{t} [\theta_{id}(t) - \theta_i(t)]dt \qquad (14)$$

where generation of the desired heading signal $\theta_{id}(t)$ is explained in the sequel in detail. In short, $\theta_{id}(t)$ corresponds to the agent motion direction that would bring A_i to a location $p_{id}(t)$ with distance d_{ij} to $p_j(t)$ and distance d_{ic} to $p(t)$.

The desired location $p_{id}(t)$ is calculated, similarly to [25,26], using the following circle intersection rule:

$$p_{id}(t) = \arg\min \{\|p - p_i(t)\| \mid p \in C(p(t), d_{ic}) \cap C(p_j(t), d_{ij})\} \qquad (15)$$

where the notion $C(\cdot,\cdot)$ is used to formulate circles with the first argument denoting the center and the second denoting the radius, and it is assumed that $C(p(t), d_{ic}) \cap C(p_j(t), d_{ij})$ is non-empty. The case of an empty intersection is dealt with later.

The corresponding desired heading signal $\theta_{id}(t)$ is generated using

$$\theta_{id}(t) = \begin{cases} \angle(p_{id}(t) - p_i(t)) & \texttt{if} \quad |e_{ij}| > \varepsilon_i \quad \texttt{or} \quad |e_{ic}| > \varepsilon_i \\ \angle(p_{id}(t) - p_{id}(t - T_\Delta)) & \texttt{else} \end{cases} \tag{16}$$

where $T_\Delta > 0$ is a certain delay term used for interpolation,

$$e_{ij} = \|p_i(t) - p_j(t)\| - d_{ij},$$
$$e_{ic} = \|p_i(t) - p_c(t)\| - d_{ic},$$

and ε_i is a separation tolerance term. In our design and simulation studies, the values of the delay and tolerance terms are taken as $T_\Delta = 1$ sec. and $\varepsilon_i = 30$ m.

The switching law (16) determines the desired direction of motion to be from the current position of agent A_i towards the closest intersection point of the two circles $C(p(t), d_{ic})$ and $C(p_j(t), d_{ij})$ if either of the two separation errors $|e_{ij}|$,$|e_{ic}|$ is larger than the tolerated value ε_i, and in a direction parallel to that in which p_{id} is estimated to be moving otherwise.

6 Simulation-based Analysis of the Control Laws

The decentralized control scheme developed in Section 5 is numerically analyzed using a set of MATLAB®/Simulink® simulations. For all conducted simulations the parameters of the spiral path (9) are taken as $a = 0$ and $b = \frac{6000}{2\pi}$ and, as mentioned in Section 5, the desired separation distances are set to $d_{12} = d_{23} = d_{31} = 3$ km and $d_{ic} = \sqrt{3}$ km ($i \in \{1, 2, 3\}$). The area to be surveyed is assumed to lie in a 30 km×30 km square region.

Three cases are considered: (i) UAVs are all flying at the same constant speed and there exists no actuator or sensor noise affecting the system. (ii) UAVs are flying at different constant speeds and there exists no actuator or sensor noise affecting the system. (iii) UAVs are all flying at the same constant speed and there exist some actuator and sensor noises affecting the system. We present the simulation results and discussions for these three cases separately in the following subsections.

6.1 Ideal Motion Behavior

In this case we assume that all the UAVs are flying at the same constant speed of 32 m/s and there exists no actuator or sensor noise affecting the system. The decentralized control laws (14)–(16) are applied using the design parameters $k_P = 2$, $k_I = 0.0005$, $\varepsilon_i = 30$ m. The simulation results shown in Fig. 2 demonstrate that both the path tracking and formation maintenance tasks are successfully achieved.

6.2 Effects of Speed Variations

In this simulation case we assume that UAVs are flying at different constant speeds around a nominal value of 32 m/s and there exists no actuator or sensor noise affecting the system.

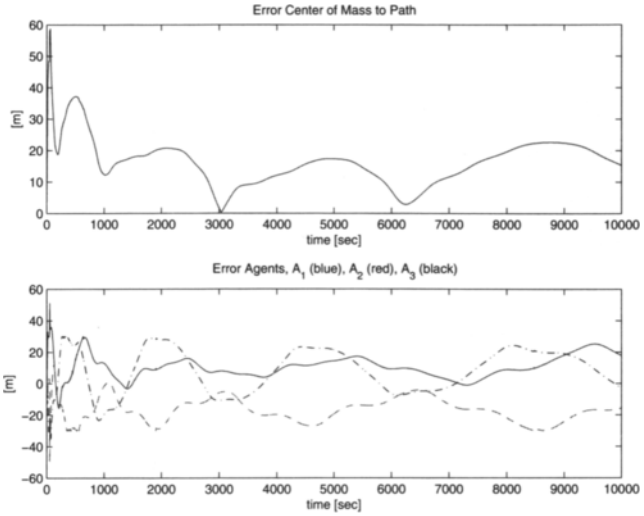


Fig. 2. Inter-agent distance keeping errors and the distance of CM to the path to be tracked

In the control design presented in Section 5, it is assumed that the two circles $C(p(t), d_{ic})$ and $C(p_j(t), d_{ij})$ always intersect, which is a valid assumption for the ideal case presented in Section 6.1. However, it is observed in simulation studies that this assumption may be violated occasionally if the differences between UAV speeds are sufficiently large. To circumvent such occasional cases, the control law (16) is modified as follows:

$$\theta_{id}(t) = \begin{cases} \angle(p_i(t) - p(t)) & \texttt{if} \quad \|p(t) - p_i(t)\| < d_{ij} - d_{ic} \\ \angle(p(t) - p_i(t)) & \texttt{else if} \quad \|p(t) - p_i(t)\| > d_{ij} + d_{ic} \\ \angle(p_{id}(t) - p_i(t)) & \texttt{else if} \quad |e_{ij}| > \varepsilon_i \quad \texttt{or} \quad |e_{ic}| > \varepsilon_i \\ \angle(p_{id}(t) - p_{id}(t - T_\Delta)) & \texttt{else} \end{cases} \tag{17}$$

The modified law (17) is the same as (16) when the circles $C(p(t), d_{ic})$ and $C(p_j(t), d_{ij})$ intersect. When they do not intersect, (17) defines the desired heading for agent A_i to be towards $p(t)$ if it is too distant to the virtual agent A, and away from $p(t)$ if it is too close to A.

The modified control scheme (14)–(17) is applied to various simulation settings with different UAV speeds, where A_1 is always assigned a constant speed of 32 m/sec and each of A_2 and A_3 is assigned a constant speed between 29 m/s and 32 m/s. Fig. 3 shows the root-mean-square of the error

$$\sqrt{\frac{1}{t_f}\int_0^{t_f} \left(\frac{p_1(t) + p_2(t) + p_3(t)}{3} - p(t)\right)^2 dt}$$

versus v_{c2} and v_{c3}, where t_f is the final time of surveillance. This figure demonstrates that agent speed differences within a certain limit are allowable in meeting the path tracking and formation maintenance tasks.

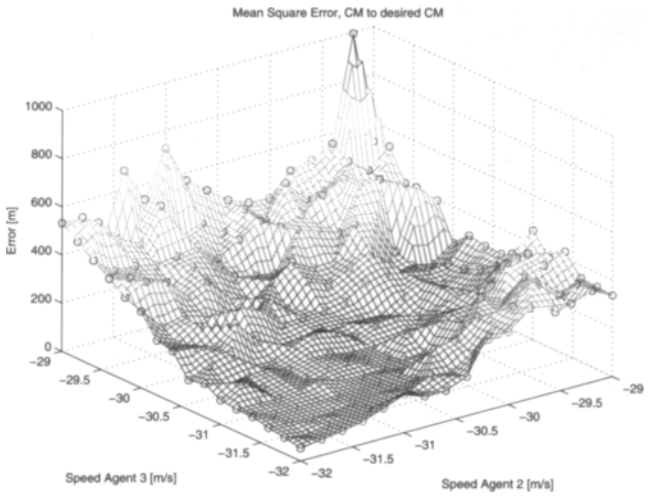


Fig. 3. Root-mean-square of the distance between the actual and desired locations of the formation CM versus agent speeds v_{c2} and v_{c3}

6.3 Effects of Sensor and Actuator Disturbances

UAVs same speed, sensor and actuator disturbance

In this final case, we assume that the three UAVs are all flying at the same constant speed and there exist some actuator and sensor noises affecting the system. The cumulative actuator and sensor noises are represented, respectively, by the parameters γ_a and γ_s represent in Fig. 4 depicting the diagram of the individual controllers of the agents. Various simulations have been performed with different standard deviation values for γ_a and γ_s, modelling the value of each of these two noises at any time instant as zero-mean Gaussian random variable. The results of these simulations are shown in Fig. 5, which demonstrates that actuator and sensor noises within a certain limit are allowable in meeting the path tracking and formation maintenance tasks.

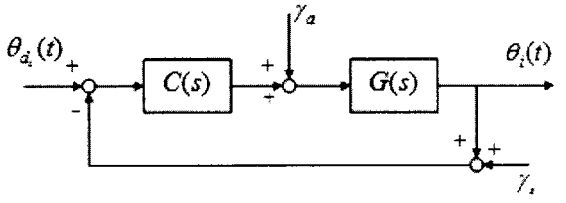


Fig. 4. Block diagram of the individual controller of agent A_i in the existence of actuator and sensor noises

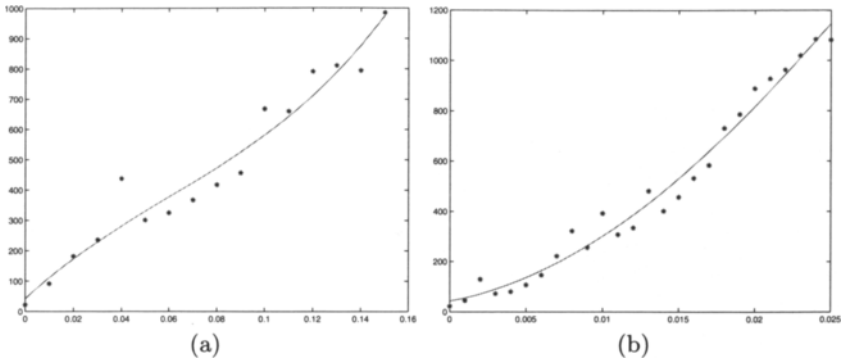


Fig. 5. Root-mean-square of the distance between the actual and desired locations of the formation CM versus standard deviations of (a) the actuator noise γ_a and (b) the sensor noise γ_s

7 Concluding Remarks

In the first part of this chapter we have described how to characterize the choice of agent pairs to secure this shape-preserving property for a planar formation based mainly on graph theory and linear algebra. We have described decentralized control laws which will stably restore the shape of a formation when the distances between nominated agent pairs become unequal to their prescribed values.

Some possible future research directions based on the first part of the chapter are as follows: In the ideal world of point agents, be it in two dimensions or three dimensions, it would be helpful to identify the class of decentralized control laws that could maintain formation shape. Of course, one must agree first on what variables are to be controlled, and here we have emphasized distance, and in the case of controlling distances, one must agree on what agent has the responsibility (or possibly shares the responsibility) for controlling what distance. Distance is naturally not the only variable, and other variables should be contemplated. It would be very helpful also to understand what formations were easy to control and what formations were hard to control: one that was easy to control would presumably require just small control signals to restore the formation shape when perturbed away from its ideal, and would be comparatively insensitive to noise perturbing the measurements. These questions also need to be posed when significant agent dynamics are introduced into the model; when there are speed, turning rate and sensing radius constraints (as in the second part of the chapter); and when there are constraints on sensing involving typically blind cones (an agent may not be allowed to have "eyes in the back of its head").

In the second part of the chapter, we have considered a particular application problem of flying a group of three UAVs in an equilateral triangle formation, with the center of mass following a nominated trajectory reflecting constraints on turning radius, and with a requirement that the speeds of the UAVs are constant. We have proposed a decentralized non-hierarchical proportional-integral control

scheme for this particular problem. The proposed control scheme is tested in various simulation scenarios to examine the basic characteristics of the proposed control scheme for the case where the UAV speeds are all equal as well as the affects of sensor and actuator disturbances and differences between the UAV speeds on the system performance. There is an ongoing study on analyzing the affects of the wind in details, even though this is partially covered within the context of sensor and actuator disturbances. The simulation results demonstrate the success and effectiveness of the proposed scheme within certain bounds of disturbances and UAV speed differences.

A particular future research topic related to the second part of the chapter is comprehensive mathematical analysis of the stability and robustness properties of the proposed scheme and mathematical explanation of the observed behavior of the path tracking error of the formation center of mass and the inter-agent distance keeping errors. A more practical research direction is revision of the control laws for further robustification against sensor and actuator disturbances and UAV speed differences and enhancement of the path tracking and formation maintenance performance.

References

1. Hubbard, S., Babak, B., Sigurdsson, S., Magnusson, K.: A model of the formation of fish schools and migrations of fish. Ecological Modelling 174, 359–374 (2004)
2. Janson, S., Middendorf, M., Beekman, M.: Honey bee swarms: How do scouts guide a swarm of uninformed bees? Animal Behaviour 70(1), 349–358 (2005)
3. Shao, J., Xie, G., Wang, L.: Leader-following formation control of multiple mobile vehicles. IET Control Theory and Applications 1, 545–552 (2007)
4. Tay, T., Whiteley, W.: Generating isostatic frameworks. Structural Topology 11, 21–69 (1985)
5. Jackson, B., Jordan, T.: Connected rigidity matroids and unique realizations of graphs. Journal of Combinatorial Theory B(94), 1–29 (2004)
6. Olfati-Saber, R., Murray, R.M.: Distributed cooperative control of multiple vehicle formations using structural potential functions. In: Proc. of the 15th IFAC World Congress, Barcelona, Spain, pp. 1–7 (2002)
7. Eren, T., Whiteley, W., Morse, A.S., Belhumeur, P.N., Anderson, B.D.: Sensor and network topologies of formations with direction, bearing and angle information between agents. In: Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, Hawaii, pp. 3064–3069 (December 2003)
8. Lin, Z., Francis, B., Maggiore, M.: Necessary and sufficient graphical conditions for formation control of unicycles. IEEE Trans. on Automatic Control 50, 121–127 (2005)
9. Yu, C., Hendrickx, J., Fidan, B., Anderson, B., Blondel, V.: Three and higher dimensional autonomous formations: Rigidity, persistence and structural persistence. Automatica, 387–402 (March 2007)
10. Laman, G.: On graphs and rigidity of plane skeletal structures. J. Engrg. Math. 4, 331–340 (1970)
11. Anderson, B., Yu, C., Fidan, B., Hendrickx, J.: Control and information architectures for formations. In: Proc. IEEE International Conference on Control Applications, Munich, Dermany, vol. 56, pp. 1127–1138 (October 2006)

12. Hendrickx, J., Anderson, B., Delvenne, J.-C., Blondel, V.: Directed graphs for the analysis of rigidity and persistence in autonomous agent systems. International Journal of Robust Nonlinear Control 17, 960–981 (2007)
13. Yu, C., Anderson, B., Dasgupta, S., Fidan, B.: Control of minimally persistent formations in the plane (submitted for publication, December 2006)
14. Anderson, B., Yu, C., Dasgupta, S., Morse, A.: Control of a three coleaders formation in the plane. Systems & Control Letters 56, 573–578 (2007)
15. Cao, M., Morse, A., Yu, C., Anderson, B., Dasgupta, S.: Controlling a triangular formation of mobile autonomous agents. In: IEEE Conference on Decision and Contol (to appear 2007)
16. Olfati-Saber, R., Murray, R.M.: Distributed cooperative control of multiple vehicle formations using structural potential functions. In: Proc. of the 15th IFAC World Congress, Barcelona, Spain, pp. 1–7 (2002)
17. Olfati-Saber, R., Murray, R.M.: Graph rigidity and distributed formation stabilization of multi- vehicle systems. In: Proc of the 41st IEEE Conf. on Decision and Control, Las Vegas, NV, pp. 2965–2971 (2002)
18. Krick, L.: Application of graph rigidity information control of multi-robot networks. Master's thesis, Department of Electrical and Computer Engineering, University of Toronto (2007)
19. Paley, D., Leonard, N.E., Sepulchre, R.: Collective motion: bistability and trajectory tracking. In: Proc. of the 43rd IEEE Conference on Decision and Control, vol. 2, pp. 1932–1937 (2004)
20. Justh, E.W., Krishnaprasad, P.S.: Equilibria and steering laws for planar formations. Systems and Control Letters 52(1), 25–38 (2004)
21. Smith, S.L., Broucke, M.E., Francis, B.A.: Stabilizing a multi-agent system to an equilibrium polygon formation. In: Proc. 17th International Symposium on Mathematical Theory of Networks and Systems, pp. 2415–2424 (2006)
22. Anderson, B., Dasgupta, S., Yu, C.: Control of directed formations with leader-first follower structure. In: IEEE Conference on Decision and Contol (to appear, 2007)
23. Drake, S., Brown, K., Fazackerley, J., Finn, A.: Autonomous control of multiple uavs for the passive location of radars. Tech. report, Defence Science and Technology Organisation, pp. 403–409 (2005)
24. Ledger, D.: Electronic warfare capabilities of mini UAVs. In: Proc. the Electronic Warfare Conference, Kuala Lumpur (2002)
25. Sandeep, S., Fidan, B., Yu, C.: Decentralized cohesive motion control of multi-agent formations. In: Proc. 14th Mediterranean Conference on Control and Automation (June 2006)
26. Fidan, B., Anderson, B., Yu, C., Hendrickx, J.: Modeling and Control of Complex Systems, ch. Persistent Autonomous Formations and Cohesive Motion Control, pp. 247–275. Taylor & Francis, London (2007)

Electrical and Mechanical Passive Network Synthesis

Michael Z.Q. Chen[1,2] and Malcolm C. Smith[1]

[1] Department of Engineering, University of Cambridge, U.K.
zc214@cam.ac.uk/mzqchen@gmail.com, mcs@eng.cam.ac.uk
[2] Department of Engineering, University of Leicester, U.K.

Summary. The context of this paper is the application of electrical circuit synthesis to problems of mechanical control. The use of the electrical-mechanical analogy and the inerter mechanical element is briefly reviewed. Classical results from passive network synthesis are surveyed including Brune's synthesis, Bott-Duffin's procedure, Darlington's synthesis, minimum reactance extraction and the synthesis of biquadratic functions. New results are presented on the synthesis of biquadratic functions which are realisable using two reactive elements and no transformers.

1 Introduction

Passive network synthesis is a classical subject in electrical circuit theory which experienced a "golden era" from the 1930s through to the 1960s. Renewed interest in this subject has recently arisen due to the introduction of a new two-terminal element called the *inerter* and the possibility to directly exploit electrical synthesis results for mechanical control [38]. Applications of this approach to vehicle suspension [39,30], control of motorcycle steering instabilities [19,20] and vibration absorption [38] have been identified.

Despite the relative maturity of the field, there are aspects of passive network synthesis which can be considered as incomplete. For example, the question of minimality of realisation in terms of the total number of elements used is far from solved. For mechanical networks, efficiency of realisation is much more important than for electrical networks. Also, for mechanical networks it is often desirable that no transformers are employed, due to the fact that levers with unrestricted ratios can be awkward to implement. However, the only general method for transformerless electrical synthesis—the method of Bott and Duffin [7] and its variants [29,31,40,21]—appears to be highly non-minimal.

The purpose of this paper is to review some of the background electrical circuit synthesis theory and present some new results on the transformerless synthesis of a sub-class of biquadratic functions.

2 The Electrical and Mechanical Analogy

The principal motivation for the introduction of the inerter in [38] was the synthesis of passive mechanical networks. It was pointed out that the standard

V.D. Blondel et al. (Eds.) Recent Advances in Learning and Control, LNCIS 371, pp. 35–50, 2008.
springerlink.com

form of the electrical-mechanical correspondences (where the spring, mass and damper are analogous to the inductor, capacitor and resistor) was restrictive for this purpose, because the mass element effectively has one terminal connected to ground. To allow the full power of electrical circuit synthesis theory to be translated over to mechanical networks, it is necessary to replace the mass element by a genuine two-terminal element with the property that the (equal and opposite) force applied at the terminals is proportional to the *relative* acceleration between them. In the notation of Fig. 1, the inerter obeys the force-velocity law $F = b(\dot{v}_1 - \dot{v}_2)$, where the constant of proportionality b is called the inertance and has the units of kilograms and v_1, v_2 are the velocities of the two terminals with $v = v_1 - v_2$. Fig. 2 shows the new table of element correspondences in the force-current analogy where force and current are the "through" variables and velocity and voltage are the "across" variables. The admittance $Y(s)$ is the ratio of through to across quantities, where s is the standard Laplace transform variable.

The mechanical realisation of an inerter can be achieved using a flywheel that is driven by a rack and pinion, and gears (see Fig. 3). The value of the inertance b is easy to compute in terms of the various gear ratios and the flywheel's moment

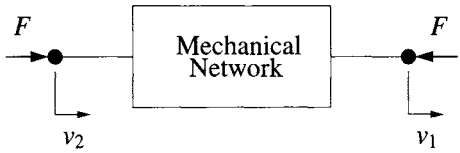


Fig. 1. Free-body diagram of a two-terminal mechanical element with force-velocity pair (F, v)

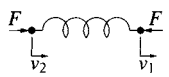
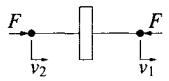
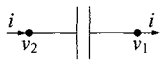
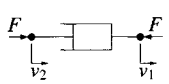
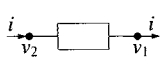
Mechanical		Electrical	
 $\frac{dF}{dt} = k(v_2 - v_1)$	$Y(s) = \frac{k}{s}$ spring	 $\frac{di}{dt} = \frac{1}{L}(v_2 - v_1)$	$Y(s) = \frac{1}{Ls}$ inductor
 $F = b\frac{d(v_2 - v_1)}{dt}$	$Y(s) = bs$ inerter	 $i = C\frac{d(v_2 - v_1)}{dt}$	$Y(s) = Cs$ capacitor
 $F = c(v_2 - v_1)$	$Y(s) = c$ damper	 $i = \frac{1}{R}(v_2 - v_1)$	$Y(s) = \frac{1}{R}$ resistor

Fig. 2. Circuit symbols and correspondences with defining equations and admittance $Y(s)$

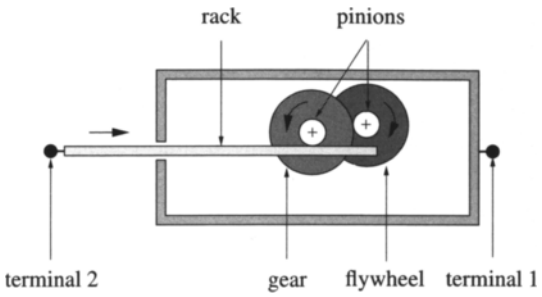


Fig. 3. Schematic of a mechanical model of an inerter

of inertia [38]. In general, if the device gives rise to a flywheel rotation of α radians per metre of relative displacement between the terminals, then the inertance of the device is given by $b = J\alpha^2$ where J is the flywheel's moment of inertia. Other methods of construction are described in [37].

With the new correspondence in Fig. 2, synthesis methods from electrical networks can be directly translated over to the mechanical case. In particular, Bott-Duffin's synthesis result [7] provides a means to realise an arbitrary positive-real mechanical admittance or impedance with networks comprising springs, dampers and inerters [38]. We will review some of the basic results from network synthesis in the next section.

3 Passive Network Synthesis

From the vast amount of literature which has been devoted to electrical circuit synthesis, we now highlight some of the fundamental results which are relevant for our application. Interesting overviews on the history of passive network synthesis can be found in [6] and [17]. Readers are referred to [1,4,5,27,28,43,47] for a more detailed treatment on the subject.

By the 1920s, researchers had started searching for a systematic way of realising passive networks. One of the early contributions is Foster's reactance theorem [22] and it is often described as the first devoted to synthesis of networks in the modern sense. The theorem states a necessary and sufficient condition for an impedance function to be realisable as the driving point impedance of a lossless one-port. The first paper dealing explicitly with the realisation of a one-port with the impedance being a prescribed function of frequency is Cauer's 1926 contribution, based on continuous fraction expansions [6,9]. With Cauer's and Foster's theorems, the synthesis problem for one-ports containing two kinds of elements only was solved.

In Brune's ground-breaking work [8], the class of positive-real functions was introduced. A rational function $Z(s)$ is defined to be *positive-real* if (i) $Z(s)$ is analytic in $\mathrm{Re}[s] > 0$ and (ii) $\mathrm{Re}[Z(s)] \geq 0$ for all s with $\mathrm{Re}[s] > 0$. He showed that there is a fundamental correspondence between positive-real functions and passive electrical circuits. In particular he showed that: (1) the

driving-point impedance or admittance of any linear two-terminal (one-port) network is positive-real, and conversely, (2) given *any* positive-real function, a two-terminal network comprising resistors, capacitors, inductors and transformers can be found which has the given function as its driving-point impedance or admittance. Brune's construction begins with the *Foster preamble* which reduces the positive-real function to a "minimum function", which is a positive-real function that has no poles or zeros on the imaginary axis or infinity and has a real part that vanishes for at least one finite real frequency. The next part of the construction is the "Brune cycle" which expresses the minimum function as a lossless coupling network connected to a positive-real function of strictly lower degree. The whole process is then repeated until a degree zero function (resistor) is reached.

For a number of years following Brune's paper, it was thought that the transformers appearing in the synthesis of general positive-real functions were unavoidable. It was therefore a surprise when a realisation procedure was published by Bott and Duffin which does not require transformers [7]. Similar to Brune's procedure, Bott-Duffin's approach also starts with the Foster preamble to reduce the positive-real function to a minimum function. It then makes use of the Richard's theorem [33], which is a generalisation of Schwarz's lemma [26], to express the minimum function as a lossless coupling network connected to *two* positive-real functions of strictly lower degree. Thus the procedure gives the appearance of being wasteful in terms of the number of components required. How wasteful it is remains an open question.

Since 1949 the only general simplifications of Bott-Duffin's method are just variants of the procedure, e.g. Pantell's procedure [29], Reza's procedure [31] and Storer's procedure [40]. All three variants work by unbalancing the bridge configuration within the lossless coupling network in Bott-Duffin's realisation to reduce the number of elements in the network from six to five in each cycle. Later, Fialkow and Gerst independently proved a similar result [21].

An important alternative proof of Brune's theorem was obtained in 1939 by Darlington [16]. The realisation method expressed the positive-real function as a lossless two-port terminated in a single resistor. The lossless two-port was realised using transformers as well as inductors and capacitors. The method was also called "minimum resistance synthesis". Connections of the method with classical interpolation were later identified [18] which have served to set the method in a general context.

A different set of techniques for passive network synthesis was based on a state-space formulation [1]. One of the central ideas is "reactance extraction" in which the impedance is represented as a multi-port with n of the ports terminated by inductors or capacitors, where n is the McMillan degree of the transfer-function. Central to the approach is the "positive-real lemma" which gives necessary and sufficient conditions for a rational transfer-function to be positive-real as a matrix condition in terms of the state-space realisation. The reactance extraction technique appears to have originated in a paper by Youla and Tissi [48], which deals with the rational bounded-real scattering matrix synthesis problem.

In the research work on electrical network synthesis, special attention has been paid to the biquadratic functions [24, 25, 23, 34, 44, 45, 42], where the impedance is given by

$$Z(s) = \frac{a_2 s^2 + a_1 s + a_0}{d_2 s^2 + d_1 s + d_0},$$

($a_i \geq 0$ and $d_i \geq 0$). For the biquadratic impedance function to be positive real, it is necessary and sufficient to have $(\sqrt{a_2 d_0} - \sqrt{a_0 d_2})^2 \leq a_1 d_1$ [23]. Biquadratic functions have been used as an important test case for the question of minimal realisation.

In [34], Seshu proved that at least two resistors are required for a transformerless realisation of a *biquadratic minimum* function, i.e. a biquadratic function that is minimum. (This result was also given by Reza in [32].) Seshu also proved that a transformerless realisation of any minimum function requires at least three reactive elements. The author went on to prove that, for a *biquadratic minimum* function, seven elements are generally required, except for the special cases $Z(0) = 4Z(\infty)$ and $Z(\infty) = 4Z(0)$, which are realisable with a five-element bridge structure. In fact, the seven-element realisations turned out to be the modified Bott-Duffin realisations [29, 40]. Following [34], it is sufficient to realise a general biquadratic function using eight elements (with one resistor to reduce a positive-real function to a minimum function). Whether it is necessary to use eight elements is still an open question.

At present, there exists no general procedure for realising biquadratic functions with the least number of elements without transformers. Given the lower order, it is very often the case that a census approach is used to cover all the possible combinations when the network structure or the number of elements is fixed (e.g. a five-element bridge network with 3 reactive elements). One attempt to generalise all biquadratic impedance functions realisable with one inductor and one capacitor (minimum reactive) without using a census approach was made by Auth [2, 3]. He formulated the problem as a three-port network synthesis problem and provided certain conditions on the physical realisability of the three-port resistive network that is terminated by one inductor and one capacitor. His approach combines elements from reactance extraction and transformerless synthesis. However, it seems that there is no general method to systematically check the conditions on the physical realisability that Auth derived. Also his direct use of Tellegen's form means that six resistors are needed [41] (see Section 4.2). In Section 4, we re-consider Auth's problem and derive a more explicit result. In particular, we show that only four dissipative elements (resistors or dampers) are needed.

4 Transformerless Second-order Minimum Reactance Synthesis

This section considers the sub-class of biquadratic functions realisable with one spring, one inerter, an arbitrary number of dampers with no levers (transformers),

which is exactly the problem considered by Auth [2, 3] under the force-current analogy. Here, we provide a more explicit characterisation of this class.

4.1 Problem Formulation

We consider a mechanical one-port network consisting of an arbitrary number of dampers, one spring and one inerter. We can arrange the network in the form of Fig. 4 where Q is a three-port network containing all the dampers. We bring in a mild assumption that the one-port has a well-defined admittance and the network Q has a well-defined impedance. As in the proof of [36, Theorem 8.1/2] we can derive an explicit form for the impedance matrix. This is defined by

$$\begin{bmatrix} \hat{v}_1 \\ \hat{v}_2 \\ \hat{v}_3 \end{bmatrix} = \begin{bmatrix} X_1 & X_4 & X_5 \\ X_4 & X_2 & X_6 \\ X_5 & X_6 & X_3 \end{bmatrix} \begin{bmatrix} \hat{F}_1 \\ \hat{F}_2 \\ \hat{F}_3 \end{bmatrix} =: X \begin{bmatrix} \hat{F}_1 \\ \hat{F}_2 \\ \hat{F}_3 \end{bmatrix} \tag{1}$$

where X is a non-negative definite matrix ($\hat{}$ denotes the Laplace transform). Setting $\hat{F}_3 = -bs\hat{v}_3$ and $\hat{F}_2 = -\frac{k}{s}\hat{v}_2$, and eliminating $\hat{v}_2$ and $\hat{v}_3$ gives the following expression for the admittance

$$Y(s) = \frac{\hat{F}_1}{\hat{v}_1} = \frac{bX_3 s^2 + \left[1 + kb(X_2X_3 - X_6^2)\right] s + kX_2}{b(X_1X_3 - X_5^2)s^2 + (X_1 + kb \det X)s + k(X_1X_2 - X_4^2)} \tag{2}$$

where $\det X = X_1X_2X_3 - X_1X_6^2 - X_2X_5^2 - X_3X_4^2 + 2X_4X_5X_6$. Note that $X_1 = 0$ requires that $X_4 = X_5 = 0$ for non-negative definiteness which means that the admittance does not exist. Thus the assumption of existence of the admittance requires that $X_1 > 0$.

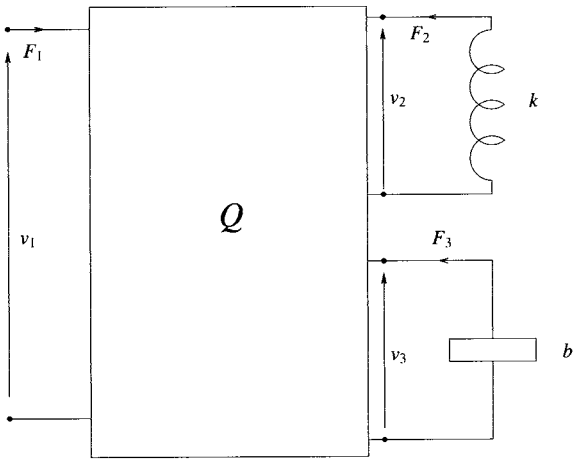


Fig. 4. Three-port damper network terminated with one inerter and one spring

The values of b and k can be set to 1 and the following scalings are carried out: $X_1 \to R_1$, $kX_2 \to R_2$, $bX_3 \to R_3$, $\sqrt{k}X_4 \to R_4$, $\sqrt{b}X_5 \to R_5$ and $\sqrt{kb}X_6 \to R_6$. The resulting admittance is

$$Y(s) = \frac{R_3 s^2 + \left[1 + (R_2 R_3 - R_6^2)\right] s + R_2}{(R_1 R_3 - R_5^2)s^2 + (R_1 + \det R)s + (R_1 R_2 - R_4^2)} \tag{3}$$

and

$$R := \begin{bmatrix} R_1 & R_4 & R_5 \\ R_4 & R_2 & R_6 \\ R_5 & R_6 & R_3 \end{bmatrix} = T \begin{bmatrix} X_1 & X_4 & X_5 \\ X_4 & X_2 & X_6 \\ X_5 & X_6 & X_3 \end{bmatrix} T,$$

where

$$T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sqrt{k} & 0 \\ 0 & 0 & \sqrt{b} \end{bmatrix} \tag{4}$$

and R is non-negative definite. From the expression $\det(R) = R_1 R_2 R_3 - R_1 R_6^2 - R_2 R_5^2 - R_3 R_4^2 + 2R_4 R_5 R_6$, we note that (3) depends on $\text{sign}(R_4 R_5 R_6)$ but not on the individual signs of R_4, R_5 and R_6.

The reactance extraction approach to network synthesis [1,48] allows the following conclusion to be drawn: any positive-real biquadratic (immitance) function should be realisable in the form of Fig. 4 for some non-negative definite X. It is also known that any non-negative definite matrix X can be realised as the driving-point impedance of a network consisting of dampers and levers (analogously, resistors and transformers) [10, Chapter 4, pages 173–179]. We now examine the question of the additional restrictions that are imposed when no transformers are allowed in Q.

4.2 Transformerless Realisation and Paramountcy

This section reviews the concept of paramountcy and its role in transformerless synthesis. We also state some relevant results from [13,14] which will be needed for our later results.

A matrix is defined to be *paramount* if its principal minors, of all orders, are greater than or equal to the absolute value of any minor built from the same rows [11,35]. It has been shown that paramountcy is a necessary condition for the realisability of an n-port resistive network without transformers [11, 35]. In general, paramountcy is not a sufficient condition for the realisability of a transformerless resistive network and a counter-example for $n = 4$ was given in [12,46]. However, in [41, pp.166–168], it was proven that paramountcy is necessary and sufficient for the realisability of a resistive network without transformers with order less than or equal to three ($n \leq 3$). The construction of [41] for the $n = 3$ case makes use of the network containing six resistors shown in Fig. 5. It is shown that this circuit is sufficient to realise any paramount matrix subject to judicious relabelling of terminals and changes of polarity. A reworking (in English) of Tellegen's proof is given in [13].

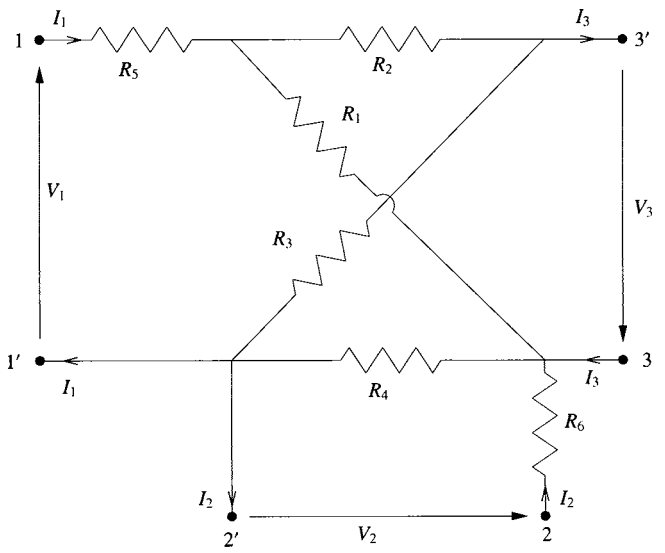


Fig. 5. Tellegen's circuit for the construction of resistive 3-ports without transformers

In the next two lemmas we establish a necessary and sufficient condition for a third-order non-negative definite matrix

$$R = \begin{bmatrix} R_1 & R_4 & R_5 \\ R_4 & R_2 & R_6 \\ R_5 & R_6 & R_3 \end{bmatrix} \tag{5}$$

to be reducible to a paramount matrix using a diagonal transformation. See [13,14] for the proofs.

Lemma 1. *Let R be non-negative definite. If any first- or second-order minor of R is zero, then there exists an invertible $D = diag\{1, x, y\}$ such that DRD is a paramount matrix.*

Lemma 2. *Let R be non-negative definite and suppose that all first- and second-order minors are non-zero. Then there exists an invertible $D = diag\{1, x, y\}$ such that DRD is a paramount matrix if and only if one of the following holds:*

(i) $R_4R_5R_6 < 0$;
(ii) $R_4R_5R_6 > 0$, $R_1 > \frac{R_4R_5}{R_6}$, $R_2 > \frac{R_4R_6}{R_5}$ *and* $R_3 > \frac{R_5R_6}{R_4}$;
(iii) $R_4R_5R_6 > 0$, $R_3 < \frac{R_5R_6}{R_4}$ *and* $R_1R_2R_3 + R_4R_5R_6 - R_1R_6^2 - R_2R_5^2 \geq 0$;
(iv) $R_4R_5R_6 > 0$, $R_2 < \frac{R_4R_6}{R_5}$ *and* $R_1R_2R_3 + R_4R_5R_6 - R_1R_6^2 - R_3R_4^2 \geq 0$;
(v) $R_4R_5R_6 > 0$, $R_1 < \frac{R_4R_5}{R_6}$ *and* $R_1R_2R_3 + R_4R_5R_6 - R_3R_4^2 - R_2R_5^2 \geq 0$.

4.3 Synthesis of Biquadratic Functions with Restricted Complexity

This section derives a necessary and sufficient condition for the realisability of an admittance function using one spring, one inerter, an arbitrary number

of dampers and no levers (transformers) (Theorem 1). The proof relies on the results of Section 4.2 and the construction of [41]. A stronger version of the sufficiency part of this result, which shows that at most four dampers are needed, is given in Theorem 2 with explicit circuit constructions. Singular cases are treated in Theorem 3.

Lemma 3. *A positive-real function $Y(s)$ can be realised as the driving-point admittance of a network in the form of Fig. 4, where Q has a well-defined impedance and is realisable with dampers only and $b, k \neq 0$, if and only if $Y(s)$ can be written in the form of*

$$Y(s) = \frac{R_3 s^2 + \left[1 + (R_2 R_3 - R_6^2)\right] s + R_2}{(R_1 R_3 - R_5^2)s^2 + (R_1 + \det R)s + (R_1 R_2 - R_4^2)}, \tag{6}$$

where

$$R = \begin{bmatrix} R_1 & R_4 & R_5 \\ R_4 & R_2 & R_6 \\ R_5 & R_6 & R_3 \end{bmatrix}$$

is non-negative definite, and there exists an invertible diagonal matrix $D = diag\{1, x, y\}$ such that DRD is paramount.

Proof: (Only if.) As in Section 4, we can write the impedance of Q in the form of (1). Since Q is realised using dampers only (no transformers), we claim that the matrix X in (3) is paramount. The transformation to (3), as in Section 4, now provides the required matrix R with the property that $X = DRD$ is paramount where $x = 1/\sqrt{k}$ and $y = 1/\sqrt{b}$.

(If.) If we define $k = 1/x^2$ and $b = 1/y^2$, then $X = DRD$ is paramount. Using the construction of Tellegen (see Section 4.2, Fig. 5), we can find a network consisting of 6 dampers and no transformers with impedance matrix equal to X. Using this network in place of Q in Fig. 4 provides a driving-point admittance given by (2) which is equal to (6) after the same transformation of Section 4. ■

We now combine Lemmas 1, 2 and 3 to obtain the following theorem.

Theorem 1. *A positive-real function $Y(s)$ can be realised as the driving-point admittance of a network in the form of Fig. 4, where Q has a well-defined impedance and is realisable with dampers only and $b, k \neq 0$, if and only if $Y(s)$ can be written in the form of (6) and R satisfies the conditions of either Lemma 1 or Lemma 2.*

In Theorem 2, we provide specific realisations for the $Y(s)$ in Theorem 1 for all cases where R satisfies the conditions of Lemma 2. The realisations are more efficient than the construction of Tellegen (see Section 4.2, Fig. 5) in that only four dampers are needed. The singular cases satisfying the conditions of Lemma 1 are also treated in Theorem 3.

Theorem 2. *Let*

$$Y(s) = \frac{R_3 s^2 + \left[1 + (R_2 R_3 - R_6^2)\right] s + R_2}{(R_1 R_3 - R_5^2)s^2 + (R_1 + \det R)s + (R_1 R_2 - R_4^2)} \tag{7}$$

where

$$R := \begin{bmatrix} R_1 & R_4 & R_5 \\ R_4 & R_2 & R_6 \\ R_5 & R_6 & R_3 \end{bmatrix}$$

is non-negative definite and satisfies the conditions in Lemma 2. Then $Y(s)$ can be realised with one spring, one inerter and four dampers in the form of Fig. 6(a)–6(e).

Proof: Fig. 6(a)–6(e) correspond to Cases (i)–(v) in Lemma 2, respectively. Explicit formulae can be given for the constants in each circuit arrangement. Here we consider only the case of Fig. 6(a).

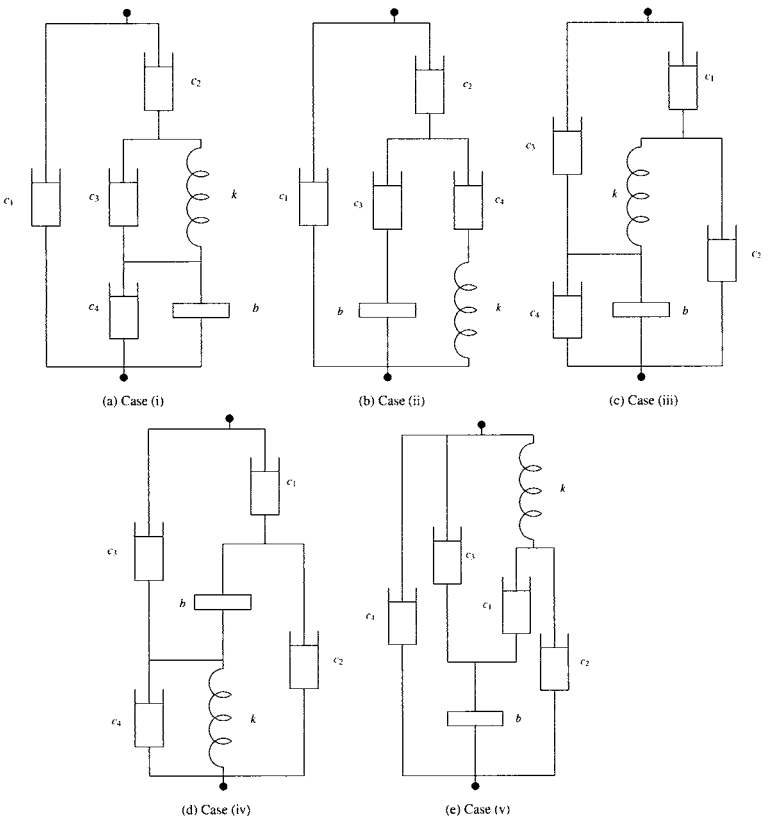


Fig. 6. Five circuit arrangements of Theorem 2

If $R_4R_5R_6 < 0$, $Y(s)$ can be realised in the form of Fig. 6(a) with

$$c_1 = \frac{1}{R_1 - \frac{R_4R_5}{R_6}}, \qquad c_2 = \frac{(R_3R_4 - R_5R_6)(R_4R_6 - R_2R_5)}{\det R(R_1R_6 - R_4R_5)},$$
$$c_3 = \frac{R_5^2(R_2 - \frac{R_4R_6}{R_5})}{(R_1R_6 - R_4R_5)^2}, \qquad c_4 = \frac{R_4^2(R_3 - \frac{R_5R_6}{R_4})}{(R_1R_6 - R_4R_5)^2},$$
$$b = \frac{(R_3R_4 - R_5R_6)^2}{(R_1R_6 - R_4R_5)^2}, \qquad k = \frac{(R_4R_6 - R_2R_5)^2}{(R_1R_6 - R_4R_5)^2}.$$

These formulae were derived directly in [15]. They can also be checked by direct substitution. See [15] for the procedures and the expressions of other cases. (A similar procedure has appeared in [13, 14].) ■

Theorem 3. *Let*

$$Y(s) = \frac{R_3s^2 + \left[1 + (R_2R_3 - R_6^2)\right]s + R_2}{(R_1R_3 - R_5^2)s^2 + (R_1 + \det R)s + (R_1R_2 - R_4^2)}$$

where R as defined in (5) is non-negative definite. If one or more of the first- or second-order minors of R is zero, then $Y(s)$ can be realised with at most one spring, one inerter and three dampers.

Proof: The proof is omitted for brevity. See [15] for details. ■

4.4 Example of Non-realisability

We now provide an explicit example of a biquadratic function which cannot be realised with two reactive elements and no transformers. First of all, we need to establish the following result.

Theorem 4. *The positive-real biquadratic function*

$$Y(s) = \frac{1}{h} \cdot \frac{a_0s^2 + a_1s + 1}{d_0s^2 + d_1s + 1} \tag{8}$$

can be realised in the form of (3), equivalently Fig. 4, for a given non-negative definite R if and only if R_2 satisfies

$$R_2 \geq \max\left\{a_1^{-1}, d_1^{-1}, d_0/(a_0d_1)\right\}, \tag{9}$$
$$0 \leq 1 - a_1R_2 + a_0R_2^2, \tag{10}$$

with R_4^2 determined by

$$(a_1^2 - 4a_0)R_4^4 - \tag{11}$$
$$2hR_2\left((a_1d_0 - 2a_0d_1 + a_0a_1)R_2 + 2(a_0 - d_0) + a_1(d_1 - a_1)\right)R_4^2 \tag{12}$$
$$+R_2^2h^2\left((a_0 - d_0)R_2 + d_1 - a_1\right)^2 = 0 \tag{13}$$

and satisfying

$$hR_2(\frac{d_0}{a_0} - 1) \leq R_4^2 \leq hR_2(d_1R_2 - 1), \tag{14}$$

and R_1, R_3, R_5 *and* R_6 *are determined by (15)–(18) as follows*

$$R_1 = h + \frac{R_4^2}{R_2}, \tag{15}$$

$$R_3 = a_0R_2, \tag{16}$$

$$R_5^2 = h(a_0 - d_0)R_2 + a_0R_4^2, \tag{17}$$

$$R_6^2 = 1 - a_1R_2 + a_0R_2^2, \tag{18}$$

with $sgn(R_4R_5R_6) = sgn(P)$ *where*

$$P := h(d_1 - a_1)R_2 + h(a_0 - d_0)R_2^2 + 2a_0R_2R_4^2 - a_1R_4^2. \tag{19}$$

Proof: Equating (3) and (8), we have

$$h = \frac{R_1R_2 - R_4^2}{R_2} = R_1 - \frac{R_4^2}{R_2}, \tag{20}$$

$$a_0 = \frac{R_3}{R_2}, \tag{21}$$

$$d_0 = \frac{R_1R_3 - R_5^2}{R_1R_2 - R_4^2}, \tag{22}$$

$$a_1 = \frac{1}{R_2} + R_3 - \frac{R_6^2}{R_2}, \tag{23}$$

$$d_1 = \frac{R_1 + \det R}{R_1R_2 - R_4^2}. \tag{24}$$

Equations (15)–(18) then follow from (20)–(23). Substituting (15)–(18) into (24) gives

$$2R_4R_5R_6 = h(d_1 - a_1)R_2 + h(a_0 - d_0)R_2^2 + 2a_0R_2R_4^2 - a_1R_4^2. \tag{25}$$

Thus the sign of $R_4R_5R_6$ is the same as the sign of P defined in (19). By squaring both sides of (25), substituting from (17,18) and rearranging the terms, we obtain (11).

From (17), we can see that the non-negativity of R_5^2 is equivalent to the lower inequality in (14). The non-negativity of R_6^2 in (18) is equivalent to (10).

To ensure the non-negative definiteness of R, it is necessary that the principal minors are non-negative. Given a non-negative R_2, the non-negativity of R_1, R_3, $R_1R_2 - R_4^2$ and $R_1R_3 - R_5^2$ is guaranteed by (15), (16), (20) and (22). Substituting from (16) and (18), we have $R_2R_3 - R_6^2 = a_0R_2^2 - (1 - a_1R_2 + a_0R_2^2) = a_1R_2 - 1$, which shows the necessity of the inequality

$$R_2 \geq a_1^{-1}. \tag{26}$$

Substituting (20) into (24) and rearranging the terms, we have $R_2 \det R = h(d_1R_2 - 1)R_2 - R_4^2$, and therefore

$$\det R \geq 0 \Leftrightarrow h(d_1R_2 - 1)R_2 \geq R_4^2 \tag{27}$$

which shows the necessity of the upper inequality in (14). For (27) to have a solution it is necessary that

$$R_2 \geq d_1^{-1}. \tag{28}$$

For the range defined in (14) to be non-empty, it is necessary that

$$R_2 \geq d_0/(a_0d_1). \tag{29}$$

Combining (26), (28) and (29) gives (9). ■

Now we will show in the example below that it is not always possible to realise a biquadratic in the form of Fig. 4 without transformers (levers).

Example (non-realisability). Consider the admittance function

$$Y(s) = \frac{2s^2 + s + 1}{s^2 + s + 1},$$

which takes the form (8) with $a_0 = 2$, $a_1 = d_0 = d_1 = h = 1$. Since $a_1d_1 > (\sqrt{a_0} - \sqrt{d_0})^2$, $Y(s)$ is positive-real. Now we apply the procedure in Theorem 4. By (9) and (14), it is necessary to have $R_2 \geq 1$ and

$$0 \leq R_4^2 \leq R_2(R_2 - 1). \tag{30}$$

Note that (10) is redundant in this case. For a particular R_2, R_4^2 is solved by (11). Then R_1, R_3, R_5 and R_6 are determined by (15)–(18). The solution of R_4^2 from (11) and the upper bound in (14) are plotted in Fig. 7. Therefore, we can see that any R_2 sufficiently large (in fact $R_2 \geq 1.5$) gives a non-negative definite R satisfying Theorem 4.

Now we would like to show that it is not possible to realise this admittance function in the form of Fig. 4 without transformers (levers). First, we note from (19) that $P = R_2^2 + (4R_2 - 1)R_4^2 \geq 0$ for all $R_2 \geq 0$. Therefore $R_4R_5R_6 > 0$ for any admissible R_2. By substituting from (17) and (18), it is easy to show that

$$\frac{R_5^2R_6^2}{R_4^2} - R_3^2 = \frac{1}{R_4^2}\left(R_2(R_2 + 1) + 2R_4^2 + 2R_2\left(R_2(R_2 - 1) - R_4^2\right)\right) \geq 0,$$

which implies that $R_3 < R_5R_6/R_4$ for any admissible R_2 . However,

$$\begin{aligned} &R_1R_2R_3 + R_4R_5R_6 - R_1R_6^2 - R_2R_5^2 \\ &= \frac{1}{2R_2}\left((R_2 - 2)R_4^2 - R_2(R_2^2 - 2R_2 + 2)\right) \\ &\leq \frac{1}{2R_2}\left((R_2 - 2)R_2(R_2 - 1) - R_2(R_2^2 - 2R_2 + 2)\right) \\ &= -\frac{R_2}{2} < 0, \end{aligned}$$

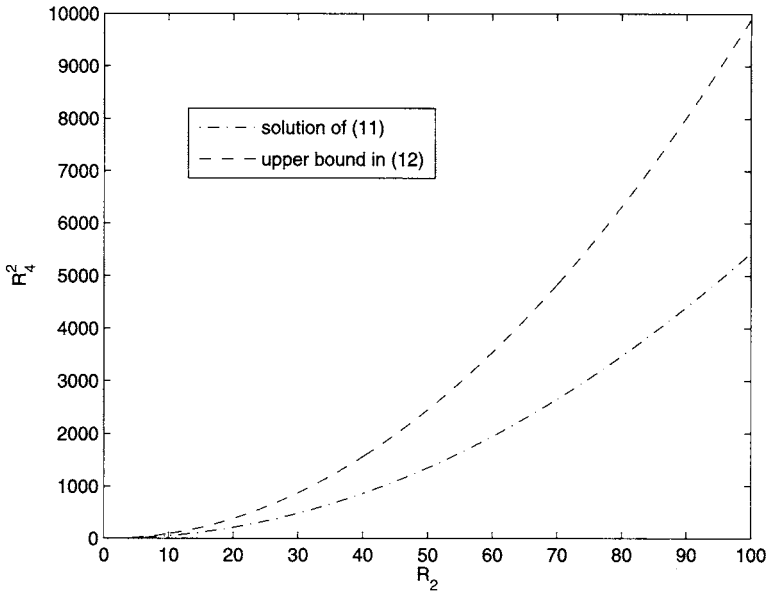


Fig. 7. Solution of R_4^2 and upper bound versus R_2

where the first inequality made use of (30). Therefore the second condition of Case (iii) in Lemma 2 fails for any admissible R_2. Therefore, $Y(s)$ cannot be realised in the form of Fig. 4 without transformers (levers).

5 Conclusions

The theme of this paper is the application of electrical circuit synthesis to mechanical networks. Relevant results from the field of passive networks have been surveyed. It was pointed out that the problem of minimal realisation (in terms of the number of elements used) is still unsolved, and that this is an important question for mechanical implementation. The class of biquadratic positive-real functions was highlighted as an interesting test case. For this class, an explicit procedure was provided to determine if a given function can be realised with two reactive elements and no transformers.

References

1. Anderson, B.D.O., Vongpanitlerd, S.: Network Analysis and Synthesis: A Modern Systems Theory Approach. Prentice Hall, Englewood Cliffs (1973)
2. Auth, L.V.: Synthesis of a Subclass of Biquadratic Immittance Functions, PhD thesis, University of Illinois, Urbana, Ill. (1962)
3. Auth, L.V.: RLC Biquadratic Driving-Point Synthesis using the Resistive Three-port. IEEE Trans. on Circuit Theory, 82–88 (1964)
4. Baher, H.: Synthesis of Electrical Networks. Wiley, Chichester (1984)

5. Balabanian, N.: Network Synthesis. Prentice-Hall, Englewood Cliffs (1958)
6. Belevitch, V.: Summary of the history of circuit theory. Proc. IRE 50(5), 848–855 (1962)
7. Bott, R., Duffin, R.J.: Impedance synthesis without use of transformers. J. Appl. Phys. 20, 816 (1949)
8. Brune, O.: Synthesis of a Finite Two-terminal Network Whose Driving-Point Impedance is a Prescribed Function of Frequency. J. Math. Phys. 10, 191–236 (1931)
9. Cauer, W.: Die Verwirklichung von Wechselstrom-Widerstanden Vorgescriebener Frequenzabhängigkeit. Arch. Elektrotech. 17, 355 (1926)
10. Cauer, W.: Synthesis of Linear Communication Networks. McGraw-Hill, New York (1958)
11. Cederbaum, I.: Conditions for the impedance and admittance matrices of n-ports without ideal transformers. Proc. IEE 105, 245–251 (1958)
12. Cederbaum, I.: Topological considerations in the realization of resistive n-port networks. IRE Trans. on Circuit Theory CT-8(3), 324–329 (1961)
13. Chen, M.Z.Q., Smith, M.C.: Mechanical networks comprising one damper and one inerter, Technical Report, CUED/F-INFENG/TR.569, Cambridge University Engineering Department, England (December 2006)
14. Chen, M.Z.Q., Smith, M.C.: Mechanical networks comprising one damper and one inerter. In: Proceedings of European Control Conference, Kos, Greece, pp. 4917–4924 (2007)
15. Chen, M.Z.Q.: Passive Network Synthesis of Restricted Complexity, PhD thesis, University of Cambridge, Cambridge, UK (2007)
16. Darlington, S.: Synthesis of reactance 4-poles which produce prescribed insertion loss characteristics. J. Math. Phys. 18, 257–353 (1939)
17. Darlington, S.: A History of Network Synthesis and Filter Theory for Circuits Composed of Resistors, Inductors, and Capacitors. IEEE Trans. on Circuits and Systems 46(1) (1999)
18. Dewilde, P., Viera, A.C., Kailath, T.: On a Generalized Szegö-Levinson Realization Algorithm for Optimal Linear Predictors based on a Network Synthesis Approach. IEEE Trans. on Circuits and Systems 25, 663–675 (1978)
19. Evangelou, S., Limebeer, D.J.N., Sharp, R.S., Smith, M.C.: Control of motorcycle steering instabilities—passive mechanical compensators incorporating inerters. IEEE Control Systems Magazine, 78–88 (October 2006)
20. Evangelou, S., Limebeer, D.J.N., Sharp, R.S., Smith, M.C.: Mechanical steering compensation for high-performance motorcycles. Transactions of ASME, J. of Applied Mechanics 74(2), 332–346 (2007)
21. Fialkow, A., Gerst, I.: Impedance synthesis without mutual coupling. Quart. Appl. Math. 12, 420–422 (1955)
22. Foster, R.M.: A reactance theorem. Bell System Tech. J. 3, 259–267 (1924)
23. Foster, R.M., Ladenheim, E.L.: A Class of Biquadratic Impedances. IEEE Trans. on Circuit Theory 10(2), 262–265 (1963)
24. Foster, R.M.: Biquadratic impedances realizable by a generalization of the five-element minimum-resistance bridges. IEEE Trans. on Circuit Theory, 363–367 (1963)
25. Foster, R.M.: Comment on Minimum Biquadratic Impedances. IEEE Trans. on Circuit Theory, 527 (1963)
26. Garnett, J.B.: Bounded Analytic Functions. Academic Press, London (1981)
27. Guillemin, E.A.: Synthesis of Passive Networks. John Wiley, Chichester (1957)

28. Newcomb, R.W.: Linear Multiport Synthesis. McGraw-Hill, New York (1966)
29. Pantell, R.H.: A new method of driving point impedance synthesis. Proc. IRE 42, 861 (1954)
30. Papageorgiou, C., Smith, M.C.: Positive real synthesis using matrix inequalities for mechanical networks: application to vehicle suspension. IEEE Trans. on Contr. Syst. Tech. 14, 423–435 (2006)
31. Reza, F.M.: A Bridge Equivalent for a Brune Cycle Terminated in a Resistor. Proc. IRE 42(8), 1321 (1954)
32. Reza, F.M.: A supplement to the Brune synthesis. AIEE Communication and Electronics 17, 85–90 (1955)
33. Richards, P.I.: A special class of functions with positive real parts in a half-plane. Duke J. of Math. 14, 777–786 (1947)
34. Seshu, S.: Minimal Realizations of the Biquadratic Minimum Functions. IRE Trans. on Circuit Theory, 345–350 (1959)
35. Slepian, P., Weinberg, L.: Synthesis applications of paramount and dominant matrices. In: Proc. Nat. Elec. Conf., vol. 14, pp. 611–630 (1958)
36. Smith, M.C., Walker, G.W.: A mechanical network approach to performance capabilities of passive suspensions. In: Proceedings of the Workshop on Modelling and Control of Mechanical Systems, pp. 103–117. Imperial College Press, Imperial College, London (1997)
37. Smith, M.C.: Force-controlling mechanical device, patent pending, Intl. App. No. PCT/GB02/03056 (July 4, 2001)
38. Smith, M.C.: Synthesis of mechanical networks: the inerter. IEEE Trans. Automatic Control 47(10), 1648–1662 (2002)
39. Smith, M.C., Wang, F-C.: Performance benefits in passive vehicle suspensions employing inerters. Vehicle System Dynamics 42, 235–257 (2004)
40. Storer, J.E.: Relationship between the Bott-Duffin and Pantell Impedance Synthesis. Proc. IRE 42(9), 1451 (1954)
41. Tellegen, B.D.H.: Théorie der Wisselstromen, P. Noordhoff (1952)
42. Tow, J.: Comments on On Biquadratic Impedances with two reactive elements. IEEE Trans. on Circuits and Systems 19 (1972)
43. Van Valkenburg, M.E.: Introduction to Modern Network Synthesis. Wiley, Chichester (1960)
44. Vasiliu, C.G.: Series-Parallel six-element synthesis of the biquadratic impedances. IEEE Trans. on Circuit Theory, 115–121 (1970)
45. Vasiliu, C.G.: Four-reactive six-element biquadratic structure. IEEE Trans. on Circuit Theory (1972)
46. Weinberg, L.: Report on Circuit Theory, Technical Report, XIII URSI Assembly, London, England (September 1960)
47. Yengst, W.C.: Procedures of Modern Network Synthesis. MacMillan, NYC (1964)
48. Youla, D.C., Tissi, P.: N-port synthesis via reactance extraction, part I. IEEE International Convention Record, 183–205 (1966)

Output Synchronization of Nonlinear Systems with Relative Degree One*

Nikhil Chopra[1] and Mark W. Spong[2]

[1] Department of Mechanical Engineering, and The Institute for Systems Research, University of Maryland, College Park, MD 20742, USA
nchopra@umd.edu

[2] Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, 1308 West Main Street, Urbana, IL 61801, USA
mspong@uiuc.edu

Summary. In this paper we extend our earlier results on output synchronization of nonlinear passive systems to the case of nonlinear systems with relative degree one. It is well known [5] that weakly minimum phase systems with relative degree one are feedback equivalent to a passive system with a positive definite storage function. We exploit this feedback equivalence to develop control laws for output synchronization of such systems, exchanging outputs on balanced graphs, and in the presence of communication delays, and switching interconnection topologies. We further show that the balanced graph assumption can be removed provided the internal dynamics in the normal form are Input-to-State-Stable (ISS) for each agent. Simulation results are presented to verify the obtained results.

Keywords: Output Synchronization, Nonlinear Systems, Time Delay, Normal Form, Passivity, Input-to-State-Stability, Hybrid Systems.

1 Introduction

The problem of coordination and control of multi-agent systems is important in numerous practical applications, such as sensor networks, unmanned air vehicles, and robot networks. Thus there has recently been considerable research devoted to the analysis and control of the coordinated behavior of such systems. The goal is to generate a desired collective behavior by local interaction among the agents. Consensus and agreement behavior has been studied in [2, 15, 21, 23, 27, 28, 30, 31, 34, 35]. Group coordination and formation stability problems have been recently addressed in [1, 11, 16, 17, 19, 20, 33] among others. We refer the readers to [22, 26] for surveys on these research efforts.

The concepts of passivity and dissipativity have been used for analyzing synchronization behavior in [7, 9, 24, 25, 32]. In this work we extend our earlier

* This research was partially supported by the Office of Naval Research under Grant N00014-02-1-0011, N00014-05-1-0186, and by the National Science Foundation under Grants ECS-0122412 and INT-0128656. The first author also acknowledges the support of the Faculty Startup Grant at the University of Maryland.

V.D. Blondel et al. (Eds.) Recent Advances in Learning and Control, LNCIS 371, pp. 51–64, 2008.
springerlink.com

work on output synchronization of passive nonlinear systems [7, 8, 9, 10] in two main directions. Instead of relying on the passivity assumption on the dynamics, we consider systems that have relative degree one and that can be globally transformed into the normal form [4]. It is well known [5] that systems that are weakly minimum phase and have relative degree one are feedback equivalent to a passive system with a positive definite storage function. Under the assumption that the zero dynamics are globally weakly minimum phase, using the results in [7, 8, 9, 10] and [5], we show output synchronization under time delays and switching interconnection topologies.

However, the above results are valid only for balanced interconnection graphs among the agents. It turns out that output synchronization can be guaranteed for weaker interconnection topologies, provided stronger assumptions are made on the system dynamics. We show that the continuous time consensus protocol developed in [27], which assumes the interconnection topology to be jointly rooted, can be used for output synchronization provided the remaining system dynamics (obtained by excluding the output dynamics in the normal form) are Input-to-State-Stable (ISS) [18].

The present paper is organized as follows. In Section 2 we give some background on nonlinear passive systems, the normal form and graph theory. This is followed by our main results in Section 3 where we demonstrate output synchronization for a system of N agents, with possibly time-varying coupling gains, switching network topologies and time delays in the network. In Section 4 we present numerical examples to validate the proposed results. This is followed up by a summary of obtained results and future work in Section 5.

2 Background

To set the background and notation for what follows, consider a control affine nonlinear system of the form

$$\Sigma \begin{cases} \dot{x} = f(x) + g(x)u \\ y = h(x) \end{cases} \tag{1}$$

where $x \in R^n$, $u \in R^m$, and $y \in R^m$. The functions $f(.) \in R^n$, $g(.) \in R^{n \times m}$, and $h(.) \in R^m$ are assumed to be sufficiently smooth. The admissible inputs are taken to be piecewise continuous and locally square integrable and we note that the dimensions of the input and output are assumed to be the same and we assume, for simplicity, that $f(0) = 0$ and $h(0) = 0$.

Definition 1. *The nonlinear system Σ is said to be passive if there exists a C^1 scalar function $V(x) \geq 0$, $V(0) = 0$, called a storage function, and a scalar function $S(x) \geq 0$ such that for all $t \geq 0$:*

$$V(x(t)) - V(x(0)) = \int_0^t u^T(s)y(s)ds - \int_0^t S(x(s))ds \tag{2}$$

The system Σ is strictly passive if $S(x) > 0$ and lossless if $S(x) = 0$.

If the following conditions hold [4]

H1 The matrix $L_g h(x)$ is nonsingular for each $x \in R^n$
H2 The vector fields $\tilde{g}_1(x), \ldots, \tilde{g}_m(x)$ are complete where $[\tilde{g}_1(x) \ldots \tilde{g}_m(x)] = g(x)[L_g h(x)]^{-1}$.
H3 The vector fields $\tilde{g}_1(x), \ldots, \tilde{g}_m(x)$ commute.

then there exists a globally defined diffeomorphism that transforms Σ into the celebrated normal form [14]

$$\begin{aligned} \dot{z} &= q(z, y) \\ \dot{y} &= b(z, y) + a(z, y)u \end{aligned}$$

where the matrix $a(z, y)$ is non-singular and $(n - m)$ real value functions $z_1(x), \ldots, z_{n-m}(x)$, along with the m-dimensional output $y = h(x)$, define the new set of coordinates.

Suppose now that we have N independent agents, each governed by dynamics of the form (1), i.e., for $i = 1, \ldots, N$ we have

$$\begin{aligned} \dot{x}_i &= f_i(x_i) + g_i(x_i)u_i \\ y_i &= h_i(x_i) \end{aligned} \tag{3}$$

We assume that all agents above satisfy assumptions $H1 - H3$ and therefore can be transformed into the globally defined normal form

$$\dot{z}_i = q_i(z_i, y_i) \tag{4}$$

$$\dot{y} = b_i(z_i, y_i) + a_i(z_i, y_i)u_i \tag{5}$$

As $a_i(z_i, y_i)$ is nonsingular, the following preliminary feedback law

$$u_i = a_i(z_i, y_i)^{-1}(-b_i(z_i, y_i) + v_i)$$

is well defined and the resultant dynamical system can be written as

$$\begin{aligned} \dot{z}_i &= q_i(z_i, y_i) \\ \dot{y}_i &= v_i \qquad\qquad i = 1, \ldots, N \end{aligned} \tag{6}$$

where the zero-dynamics are given as $\dot{z}_i = q_i(z_i, 0)$.

In the sequel, the control term v_i is used to couple the agents to achieve output synchronization based on the interconnection graph among the individual agents.

2.1 Graph Theory

Information exchange between agents can be represented as a graph. We give here some basic terminology and definitions from graph theory [12] sufficient to follow the subsequent development.

Definition 2. *By an (information) graph $\mathcal{G}$ we mean a finite set $\mathcal{V}(\mathcal{G}) = \{v_i, \ldots, v_N\}$, whose elements are called nodes or vertices, together with set $\mathcal{E}(\mathcal{G}) \subset \mathcal{V} \times \mathcal{V}$, whose elements are called edges. An edge is therefore an ordered pair of distinct vertices.*

If, for all $(v_i, v_j) \in \mathcal{E}(\mathcal{G})$, *the edge* $(v_j, v_i) \in \mathcal{E}(\mathcal{G})$ *then the graph is said to be undirected. Otherwise, it is called a directed graph.*

An edge (v_i, v_j) *is said to be incoming with respect to* v_j *and outgoing with respect to* v_i *and can be represented as an arrow with vertex* v_i *as its tail and vertex* v_j *as its head.*

The in-degree of a vertex $v \in \mathcal{G}$ *is the number of edges that have this vertex as a head. Similarly, the out-degree of a vertex* $v \in \mathcal{G}$ *is the number of edges that have this vertex as the tail.*

If the in-degree equals the out-degree for all vertices $v \in \mathcal{V}(\mathcal{G})$, *then the graph is said to be balanced.*

A path of length r *in a directed graph is a sequence* $v_0, \ldots, v_r$ *of* $r+1$ *distinct vertices such that for every* $i \in \{0, \ldots, r-1\}$, (v_i, v_{i+1}) *is an edge.*

A weak path is a sequence $v_0, \ldots, v_r$ *of* $r+1$ *distinct vertices such that for each* $i \in \{0, \ldots, r-1\}$ *either* (v_i, v_{i+1}) *or* (v_{i+1}, v_i) *is an edge.*

A directed graph is strongly connected if any two vertices can be joined by a path and is weakly connected if any two vertices can be joined by a weak path.

A directed graph is said to contain a directed spanning tree if there exists $v_i \in \mathcal{V}(\mathcal{G})$ *such that* $(v_i, v_j) \in \mathcal{E}(\mathcal{G})$, $\forall j \neq i$. *Such a graph is said to be rooted.*

3 Output Synchronization

Consider a network of N agents as above.

Definition 3. *The agents are said to output synchronize if*

$$\lim_{t \to \infty} ||y_i(t) - y_j(t)|| = 0 \quad \forall i, j = 1, \ldots, N$$

The agents are said to achieve output consensus, if additionally they converge to a common constant value.

We note that the first equation in the transformed agent dynamics (6) can be rewritten as

$$\dot{z}_i = q_i(z_i, 0) + p_i(z_i, y_i) y_i \tag{7}$$

We first assume that the individual system dynamics (6) are globally weakly minimum phase. This property guarantees the existence of a C^2 positive definite, radially unbounded function $W_i(z_i)$ such that $L_{q_i(z_i,0)} W_i \leq 0$.

The set $\mathcal{G}_N$ denotes the finite collection of possible directed graphs among the N agents, $\mathcal{P} = \{1, 2, \ldots, v\}$; $v \in \mathbb{N}$ is the finite index set associated with the elements of $\mathcal{G}_N = \{\mathcal{G}^1, \ldots, \mathcal{G}^v\}$. Let $z_s = [z_1 \ldots z_N]^T$ and $y_s = [y_1 \ldots y_N]^T$, then the coupling control law is given as

$$v_i(z_i, y_s) = -\left(L_{p_i(z_i,y_i)} W_i\right)^T + \sum_{j \in \mathcal{N}_i(\mathcal{G}^p)} K(y_j - y_i), \quad i = 1, \ldots, N \tag{8}$$

where $\mathcal{N}_i(\mathcal{G}^p)$ denotes the set of neighbors of the i^{th} agent in the interconnection graph $\mathcal{G}^p$.

Theorem 1. *Consider the dynamical system* (6) *along with the coupling control* (8). *If all multiagent systems are globally weakly minimum phase and the interconnection graph is balanced, strongly connected and time-invariant, then all signals in the dynamical system* (6) *and* (8) *are bounded and* (6) *output synchronizes.*

Proof. Consider a positive definite, radially unbounded Lyapunov function for the N-agent system as

$$V(z_s, y_s) = \sum_{i=1}^{N} (2W_i(z_i) + y_i^T y_i) \tag{9}$$

Let $c > 0$ be such that $V(z_s(t_0), y_s(t_0)) \leq c$. The derivative of $V(z_s, y_s)$ along trajectories generated by (6) and (8) is given as

$$\begin{aligned}
\dot{V}(z_s, y_s) &= \sum_{i=1}^{N} (2\dot{W}_i + 2y_i^T v_i) \\
&= \sum_{i=1}^{N} (2L_{q_i(z_i,0)} W_i + 2L_{p_i(z_i,y_i)} W_i y_i) - 2\sum_{i=1}^{N} y_i^T \left(L_{p_i(z_i,y_i)} W_i\right)^T \\
&\quad + 2K \sum_{i=1}^{N} y_i^T \sum_{j \in \mathcal{N}_i(\mathcal{G}^p)} (y_j - y_i)
\end{aligned} \tag{10}$$

As the graph is balanced, we can write

$$K \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i(\mathcal{G}^p)} y_j^T y_j = K \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i(\mathcal{G}^p)} y_i^T y_i \tag{11}$$

Using the above expression in (10) yields,

$$\begin{aligned}
\dot{V}(z_s, y_s) &= 2\sum_{i=1}^{N} L_{q_i(z_i,0)} W_i - K \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i(\mathcal{G}^p)} y_i^T y_i - K \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i(\mathcal{G}^p)} y_j^T y_j \\
&\quad + 2K \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i(\mathcal{G}^p)} y_j^T y_i \\
&= 2\sum_{i=1}^{N} L_{q_i(z_i,0)} W_i - K \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i(\mathcal{G}^p)} (y_j - y_i)^T (y_j - y_i) \leq 0
\end{aligned} \tag{12}$$

Therefore the compact set $V(z_s(t), y_s(t)) \leq c$ is positively invariant $\forall t \geq 0$ and hence all solutions of the dynamical system (6) and (8) are bounded. Consider the set $E = \{z_i \in R^{n-m}, y_i \in R^m \; i = 1, \ldots, N \mid \dot{V}(z_s, y_s) \equiv 0\}$. The set E is characterized by all trajectories such that $\{L_{q_i(z_i,0)} W_i \equiv 0, \; (y_i - y_j)^T (y_i - y_j) \equiv 0 \;\; \forall j \in \mathcal{N}_i(\mathcal{G}^p), \;\; \forall i = 1, \ldots, N\}$. Using Lasalle's Invariance Principle [18],

all trajectories of (6) and (8) converge to M as $t \to \infty$, where M is the largest invariant set contained in E. Strong connectivity of the interconnection graph then implies output synchronization of (6). ■

We next treat the practically important case where there are time delays in communication among the agents. The delays are assumed to be constant and bounded. As there can be multiple paths between two agents, we use T_{ij}^k to denote the delay along the k^{th} path from the i^{th} agent to the j^{th} agent, which we refer to as the *path delay*. We only impose the restriction that delays along all paths of length one are unique, i.e. one-hop transmission delays are uniquely defined.

Definition 4. *In the presence of delays the agents are said to output synchronize if*

$$\lim_{t\to\infty} ||y_i(t - T_{ij}^k) - y_j(t)|| = 0 \quad \forall i, j \quad \forall k \tag{13}$$

Let the coupling control law in this case be given as

$$v_i(z_i, y_s) = -\left(L_{p_i(z_i,y_i)}W_i\right)^T + \sum_{j\in\mathcal{N}_i(\mathcal{G}^p)} K(y_j(t - T_{ji}) - y_i), \quad p \in \mathcal{P} \tag{14}$$

where $\mathcal{N}_i(\mathcal{G}^p)$ denotes the set of neighbors of the i^{th} agent in the interconnection graph $\mathcal{G}^p$.

Theorem 2. *Consider the dynamical system* (6) *along with the coupling control* (14). *If all multiagent systems are globally weakly minimum phase and the interconnection graph is balanced, strongly connected and time-invariant, then all signals are bounded and* (6) *output synchronizes in the sense of* (13).

Proof. Consider a positive definite storage function for the N-agent system as

$$V_p(z_s, y_s) = K\sum_{i=1}^{N}\sum_{j\in\mathcal{N}_i(\mathcal{G}^p)}\int_{-T_{ji}}^{0} y_j^T(t+s)y_j(t+s)ds + \sum_{i=1}^{N}(2W_i(z_i) + y_i^T y_i) \tag{15}$$

A similar calculation as in Theorem 1, using the assumption that the interconnection graph is balanced, can be used to show that the derivative of V_p along solutions of (6) and (14) is given as

$$\begin{aligned}\dot{V}_p &= 2\sum_{i=1}^{N} L_{q_i(z_i,0)}W_i \\ &-K\sum_{i=1}^{N}\sum_{j\in\mathcal{N}_i(\mathcal{G}^p)}(y_j(t - T_{ji}) - y_i)^T(y_j(t - T_{ji}) - y_i) \le 0\end{aligned} \tag{16}$$

As V_p is positive definite and $\dot{V}_p \le 0$, $\lim_{t\to\infty} V_p(t)$ exists, is finite and all signals in the dynamical system (6) and (14) are bounded. Using Barbalat's Lemma [18], $\lim_{t\to\infty} \dot{V}_p(t) = 0$. Output synchronization in the sense of (13) follows from (16) and strong connectivity of the interconnection graph. ■

We next consider the case when the graph topology is not constant, such as in nearest neighbor scenarios, and there are time delays in communication. Consequently the information graph $\mathcal{G}(t)$ is not decided a priori, but is time varying. Thus we have a switched system with the continuous state $[z_s \quad y_s]^T$ and the discrete state $\mathcal{G}(t) \in \mathcal{G}_N$ where $\mathcal{G}_N$ is the finite collection of possible directed graphs among the N agents.

As the admissible inputs are restricted to be piecewise continuous, the agent dynamics are now given as

$$\begin{aligned} \dot{z}_i &= q_i(z_i, y_i) \\ \dot{y}_i &= v_{\sigma_i} \end{aligned} \tag{17}$$

where the vector fields have been previously defined, $\sigma : [0, \infty) \to \mathcal{P}$ is the right continuous switching signal and, as noted before, $\mathcal{P} = \{1, 2, \ldots, v\}$; $v \in \mathbb{N}$ is the finite index set associated with the elements of $\mathcal{G}_N = \{\mathcal{G}^1, \ldots, \mathcal{G}^v\}$. We assume that the switching signal is piecewise continuous and denote by $t_w, w = 1, 2, \ldots$ the consecutive discontinuities of the switching signal $\sigma(t)$.

Dwell Time Assumption. We impose the restriction that there exists $d > 0$ such that for every $T_d > 0$ we can find a positive integer w for which $t_{w+1} - d \geq t_w \geq T_d$.

To study output synchronization in switched multiagent systems that are weakly minimum phase with relative degree one we use Theorem 2 and the asymptotic convergence result developed in [13]. The results in [13] provide a Barbalat-like lemma for addressing switched systems and we adapt their result for demonstrating output synchronization when the interconnection topology is switching and there are time delays in communication. Let the coupling control for the agents be given as

$$v_{\sigma_i}(z_i, y_s, t) = -\left(L_{p_i(z_i, y_i)} W_i\right)^T + \textstyle\sum_{j \in \mathcal{N}_i(\mathcal{G}^p(t))} K(y_j(t - T_{ji}) - y_i) \tag{18}$$

The interconnection topology among the agents is not allowed to switch arbitrarily. The switching logic (see [3] and [13]) among the possible interconnection topologies is given by the following condition:

For every pair of switching times $t_z > t_w$ and every $p \in \mathcal{P}$ such that $\sigma(t_w) = \sigma(t_z) = p$

$$V_p(z_s(t_z), y_s(t_z)) \leq V_p(z_s(t_{w+1}), y_s(t_{w+1})) \tag{19}$$

Theorem 3. *Consider the dynamical system* (17) *along with the coupling control* (18). *If all systems in the network are globally weakly minimum phase, the interconnection graph is balanced, strongly connected, the switching law satisfies* (19) *and the dwell-time assumption, then all solution trajectories are bounded and the agents output synchronize in the sense of* (13).

Proof. Using Theorem 2, for each interconnection topology $p \in \mathcal{P}$, we have $\dot{V}_p \leq 0$. Using this fact in conjunction with the switching law (19) and observing that the index set $\mathcal{P}$ is finite, we conclude that all solutions are bounded

(see also [3] for more details). The Barbalat-like result (see the proof of asymptotic convergence in Theorem 7 of [13]) is now used to complete the proof. Using the dwell time assumption and finiteness of the index set $\mathcal{P}$, there exists an infinite subsequence of switching times $t_{w_1}, t_{w_2}, \ldots$ such that the time intervals $t_{w_{k^*+1}} - t_{w_{k^*}} \geq d$, $k^* = 1, 2, \ldots$and $\sigma(t) = h$ on these time intervals.

Denote the union of these time intervals by $\mathcal{H}$ and construct the auxiliary function

$$y_{\mathcal{H}}(t) = \begin{cases} -\dot{V}_h(t), & \text{if } t \in \mathcal{H} \\ 0 & \text{otherwise} \end{cases}$$

Using (16) and (19), $\forall t \geq 0$

$$\int_0^t y_{\mathcal{H}}(s)ds \leq V_h(z_s, y_s)_{(t=t_{w_1})} - V_h(z_s, y_s)_{(t=t)} \leq V_h(z_s, y_s)_{(t=t_{w_1})} \tag{20}$$

As $y_{\mathcal{H}}(t)$ is positive semi-definite, using (20) and letting $t \to \infty$, we have $y_{\mathcal{H}}(t) \in \mathcal{L}_1$. To show that $\lim_{t\to\infty} \dot{V}_h(t) = 0$ we need to prove that $\lim_{t\to\infty} y_{\mathcal{H}}(t) = 0$. Let us suppose that this is not true. Then $\exists \epsilon > 0$ and an infinite sequence of times s_k, $k = 1, 2, \ldots \in \mathcal{H}$ such that $y_{\mathcal{H}}(s_k) \geq \epsilon \ \forall k$. As the solution trajectories are ultimately bounded, $y_{\mathcal{H}}$ is uniformly continuous on $\mathcal{H}$. Consequently, $\exists \delta > 0$ such that s_k belongs to a time interval of length δ on which $y_{\mathcal{H}}(t) \geq \frac{\epsilon}{2}$. Therefore, a contradiction results as $y_{\mathcal{H}}(t) \in \mathcal{L}_1$. This implies that $\lim_{t\to\infty} y_{\mathcal{H}}(t) = 0$ and hence $\lim_{t\to\infty} \dot{V}_h(t) = 0$. As the communication graph is strongly connected at all times, the agents output synchronize in the sense of (13). ■

If there are no communication delays in the network, the following result follows from Theorem 3.

Corollary 1. *Consider the dynamical system* (17) *along with the coupling control* (18). *If all multiagent systems are globally weakly minimum phase, the interconnection graph is balanced, strongly connected, there are no communication delays among the agents and the switching law satisfies the dwell-time assumption, then all solution trajectories are bounded and the agents output synchronize.*

Proof. It is easy to see that in this case (15) reduces to

$$V_p(z_s, y_s) = V(z_s, y_s) = \sum_{i=1}^{N} (2W_i(z_i) + y_i^T y_i) \tag{21}$$

which is independent of the interconnection topology and hence is a common positive definite storage function for the switched system. As $V(z_s, y_s)$ is a continuous function and $\dot{V}_p(z_s, y_s) = \dot{V}(z_s, y_s) \leq 0$, the switching condition (19) is satisfied. Invoking Theorem 3 we conclude that the agents output synchronize. ■

The previous results were developed under the condition that the systems are globally weakly minimum phase and that the interconnection graph is balanced. We now seek to relax the balanced graph assumption, however it will be clear in the subsequent results that this comes at the cost of stronger assumptions

on the system dynamics. Observing (17), note that the agent's output dynamics are simple integrator dynamics and are decoupled from the zero dynamics. Therefore, we can treat (17) as an cascade system where the outputs y_i drive the z_i dynamics.

As the output dynamics are decoupled from the z_i dynamics, the continuous time consensus results in the literature [15, 23, 27] can be employed to ensure that the agents achieve output consensus. It is to be noted that additional consensus results have been developed in [6, 21, 34] among others, however the aforementioned results are in discrete time and hence are not directly applicable in the current setting.

Let the output coupling control be given as

$$v_{\sigma_i}(y_s, t) = \sum_{j \in \mathcal{N}(\mathcal{G}^p(t))} k_{ji}(t)(y_j(t) - y_i(t)); \quad p \in \mathcal{P} \tag{22}$$

where $0 < l_1 \leq k_{ji}(t) \leq l_2$. Therefore, the closed loop output dynamics of the various agents are given as

$$\dot{y}_i = \sum_{j \in \mathcal{N}(\mathcal{G}^p(t))} k_{ji}(t)(y_j(t) - y_i(t)) \quad \forall i = 1, \ldots, N \tag{23}$$

In various practical scenarios, the interconnection graph among the agents may not be connected at all times. To address this issue the notion of joint connectivity was introduced in [15] (see also [34]), and has been used by several authors including [21, 27, 31], among others. The agents are said to be jointly connected across the time interval $[t, t+T]$, $T > 0$ if the agents are weakly connected across the union $\cup_{\sigma \in [t, t+T]} \mathcal{E}(\mathcal{G}(\sigma))$. Furthermore, the agents are said to jointly rooted [6] across the time interval $[t, t+T]$, $T > 0$ if the union $\cup_{\sigma \in [t, t+T]} \mathcal{E}(\mathcal{G}(\sigma))$ contains a directed spanning tree. Note that the dwell time assumption implies that there are only finitely many distinct graphs in this union for each T. We recall the following result in [27]

Theorem 4. *Consider the dynamical system* (23). *If there exists infinitely many bounded, contiguous time intervals across which the agents are jointly rooted, then the outputs y_i are ultimately bounded and the agents achieve output consensus.*

We refer the reader to [27] for a proof of this result. The next result follows from the above result and the Input-to-State Stable (ISS) assumption on the z_i dynamics of the individual systems.

Theorem 5. *Consider the dynamical system* (17) *along with the coupling control* (22). *If there exists infinitely many bounded, contiguous time intervals across which the agents are jointly rooted and the z_i dynamics of every agent in* (17) *are ISS, then all signals are bounded and the agents achieve output consensus.*

Proof. Noting (17), the ISS assumption [18, 29] implies that there exists functions $\beta \in \mathcal{KL}$ and $\gamma \in \mathcal{K}$ such that the following estimate holds for all $t \geq 0$

$$||z(t)|| \leq \beta(||z(0), t||) + \gamma(\sup_{0 \leq \tau \leq t} ||y(\tau)||)$$

Using Theorem 4, the agents achieve output consensus and the outputs are ultimately bounded. Using this along with the ISS assumption we conclude that the z_i dynamics are ultimately bounded. ■

4 Examples

Example 1: For the sake of brevity and clarity, consider the following agent dynamics that are already in the normal form (6)

$$\begin{aligned}\dot{z}_i &= -z_i + z_i^2 y_i \\ \dot{y}_i &= v_i \quad i = 1, 2, 3\end{aligned}$$

where $z, y \in R$. The above system is globally minimum phase with $W(z) = \frac{1}{2}z^2$ as $\dot{W}(z)$ is negative definite along the zero dynamics. Therefore, the coupling control v_i for each agent is given as

$$\begin{aligned}v_i &= -\left(L_{p_i(z_i,y_i)} W_i\right)^T + \sum_{j \in \mathcal{N}_i} K y_j(t - T_{ji}) - y_i) \\ &= -z^3 + \sum_{j \in \mathcal{N}_i} K y_j(t - T_{ji}) - y_i)\end{aligned}$$

The agents are assumed to be interconnected using a time-invariant ring topology as shown in Figure 1. Therefore the closed loop dynamical system with $K = 1$ is given as

$$\begin{aligned}\dot{z}_i &= -z_i + z_i^2 y_i \quad i = 1, 2, 3 \\ \dot{y}_1 &= -z_1^3 + (y_3(t - T_{31}) - y_1) \\ \dot{y}_2 &= -z_2^3 + (y_1(t - T_{12}) - y_2) \\ \dot{y}_3 &= -z_3^3 + (y_2(t - T_{23}) - y_3)\end{aligned}$$

The time delays in the simulation were set as $T_{31} = .3, T_{12} = .5$ and $T_{23} = 1$. As expected from Theorem 2, the outputs synchronize and the z_i variables asymptotically converge to the origin in Figure 2.

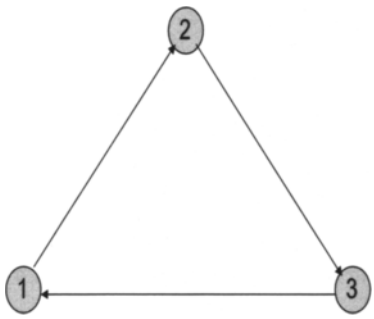


Fig. 1. The interconnection graph for the three agents shown as a ring

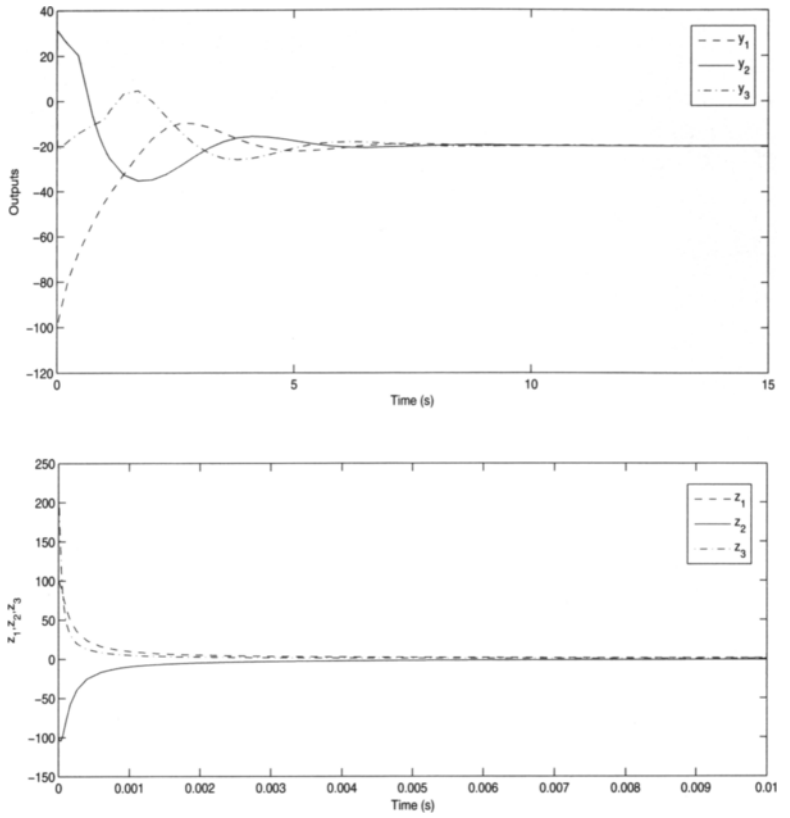


Fig. 2. The outputs of the three agents synchronize asymptotically (top) and the internal (z) dynamics converge to the origin (bottom)

Example 2: As a simple example, consider three identical point masses with the dynamics

$$\ddot{q}_i = \tau_i \ ; \ i = 1, 2, 3$$

where $q_i \in R$. Following [8, 9], let the preliminary feedback be given as

$$\tau_i = -\lambda \dot{q}_i + v_{\sigma_i}$$

where $\lambda > 0$. The closed loop system (in the normal form) reduces to

$$\dot{q}_i = -\lambda q + r_i \tag{24}$$

$$\dot{r}_i = v_{\sigma_i} \tag{25}$$

where $r_i = \dot{q}_i + \lambda q_i$. Consider a quadratic Lyapunov function for the q_i dynamics as $V(q_i) = \frac{1}{2} q_i^2$. It is evident that $\dot{V}(q_i) = -\lambda q_i^2 + r_i q_i$. Therefore, the q_i dynamics are ISS with r_i as the input.

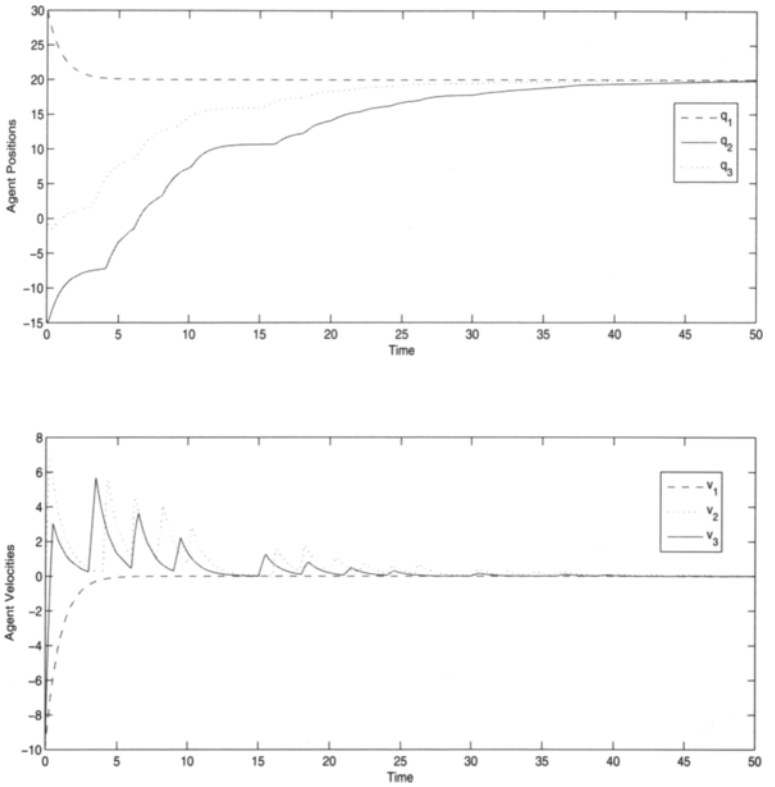


Fig. 3. The positions of the three agents synchronize asymptotically (top) and the agents' velocities synchronously converge to the origin (bottom)

Following (22), let the coupling control be given as

$$v_{\sigma_i}(r,t) = \sum_{j \in \mathcal{N}(\mathcal{G}^p(t))} (r_j(t) - r_i(t)); \quad p \in \mathcal{P} \tag{26}$$

where $r = [r_1 \quad r_2 \quad r_3]^T$, and the switching signal $\sigma(t)$ is such that there exists infinitely many bounded, contiguous time intervals across which the agents are jointly rooted. Then using Theorem 5, all signals in (24), (25) and (26) are bounded and $\lim_{t\to\infty} r_i = \mathrm{c} \quad \forall i$ where c is a constant. However, note that

$$\begin{aligned} r_j(t) - r_i(t) &= (\dot{q}_j(t) + \lambda q_j(t)) - (\dot{q}_i(t) + \lambda q_i(t)) \\ &= \dot{e}_{ij}(t) + \lambda e_{ij}(t) \end{aligned} \tag{27}$$

where $e_{ij}(t) = q_j(t) - q_i(t)$. As the agents achieve output consensus, from (27) $\lim_{t\to\infty} e_{ij}(t) = 0 \ \forall i, j$ and noting that (24) is a stable linear system, $\lim_{t\to\infty} \dot{q}_i = 0 \quad \forall i$. Therefore, the agents state synchronize asymptotically. The simulations were done with the first agent as the root and the interconnection graph jointly

rooted. As seen in Figure 3, the position of the various agents synchronize and the velocities synchronously approach the origin.

5 Conclusions

In this paper we extended our earlier results on output synchronization of nonlinear passive systems to the case of nonlinear systems with relative degree one. Exploiting the feedback equivalence of weakly minimum phase nonlinear systems with relative degree one [5] to passive systems with a positive definite storage function, control laws were developed for output synchronization of such systems in the presence of communication delays and switching topologies.

The aforementioned results are limited to the case when the interconnection graph among the agents is balanced. If the internal dynamics in the normal form are ISS, output synchronization was shown for the case when the interconnection graph is jointly rooted. Future work involves removing the relative degree assumption or the weakly minimum phase assumptions in the proposed results.

References

1. Arcak, M.: Passivity as a design tool for group coordination. In: Proc. Amer Control Conf., Minneapolis, MN (2006)
2. Blondel, V.D., Hendrickx, J.M., Olshevsky, A., Tsitsiklis, J.N.: Convergence in multiagent coordination, consensus, and flocking. In: Proc. Joint Conf. Decision Control & Euro Control Conf. Seville, Spain (2005)
3. Branicky, M.: IEEE Trans. Autom. Control 43, 475–482 (1998)
4. Byrnes, C., Isidori, A.: IEEE Trans. Autom. Control 36, 1122–1137 (1991)
5. Byrnes, C., Isidori, A., Willems, J.C.: IEEE Trans. Autom. Control 36, 1228–1240 (1991)
6. Cao, M., Morse, A.S., Anderson, B.D.O.: Reaching a consensus using delayed information. In: Proc. Conf. Decision Control, San Diego, CA (2006)
7. Chopra, N.: Output Synchronization of Networked Passive Systems. Ph.D. Thesis, University of Illinois at Urbana-Champaign, Urbana (2006)
8. Chopra, N., Spong, M.W.: Synchronization of networked passive systems with applications to bilateral teleoperation. In: Proc SICE Conf., Okayama, Japan (2005)
9. Chopra, N., Spong, M.W.: Passivity-based control of multi-agent systems. In: Kawamura, S., Svinin, M. (eds.) Advances in Robot Control: From Everyday Physics to Human-Like Movements, Springer, Heidelberg (2006)
10. Chopra, N., Spong, M.W.: Output synchronization of nonlinear systems with time delay in communication. In: Proc. Conf. Decision Control, San Diego, CA (2006)
11. Fax, J.A., Murray, R.M.: IEEE Trans. Autom. Control 49, 1465–1476 (2004)
12. Godsil, C., Royle, G.: Algebraic graph theory. Springer Graduate Texts in Mathematics, vol. 207. Springer, Heidelberg (2001)
13. Hespanha, J.P., Liberzon, D., Angeli, D., Sontag, E.D.: IEEE Trans. Autom. Control 50, 154–168 (2005)
14. Isidori, A.: Nonlinear Control Systems. Springer, Berlin (1995)
15. Jadbabaie, A., Lin, J., Morse, A.S.: IEEE Trans. Autom. Control. 48, 988–1001 (2003)

16. Justh, E.W., Krishnaprasad, P.S.: Syst. Control Lett. 52, 25–38 (2004)
17. Khalil, H.K.: Nonlinear systems. Prentice Hall, Upper Saddle River, New Jersey (2005)
18. Lee, D.J., Spong, M.W.: IEEE Trans. Autom. Control 52, 1469–1475 (2007)
19. Leonard, N.E., Fiorelli, E.: Virtual leaders, artificial potentials and coordinated control of groups. In: Proc. Conf. Decision Control, Orlando, FL (2001)
20. Marshall, J.A., Broucke, M.E., Francis, B.A.: IEEE Trans. Autom. Control 49, 1963–1974 (2004)
21. Moreau, L.: IEEE Trans. Autom. Control 50, 169–182 (2005)
22. Olfati-Saber, R., Fax, J.A., Murray, R.M.: Proc. IEEE 95, 215–233 (2007)
23. Olfati-Saber, R., Murray, R.M.: IEEE Trans. Autom. Control 49, 1520–1533 (2004)
24. Pogromsky, A.Y., Nijmeijer, H.: IEEE Trans. Circ. Syst-1 48, 152–162 (2001)
25. Pogromsky, A.Y.: Int. J. Bif Chaos 8, 295–319 (1998)
26. Ren, W., Beard, R.W., Atkins, E.: IEEE Control Syst. Mag. 27, 71–82 (2007)
27. Ren, W., Beard, R.W.: IEEE Trans. Autom. Control 50, 655–661 (2005)
28. Reynolds, C.W.: Comput. Graph 21, 25–34 (1987)
29. Sepulchre, R., Janković, M., Kokotović, P.V.: Constructive Nonlinear Control. Springer, London (1997)
30. Sepulchre, R., Paley, D., Leonard, N.: Collective motion and oscillator synchronization. In: Kumar, V., Leonard, N.E., Morse, A.S. (eds.) Cooperative Control. Lecture Notes in Control and Information Sciences, Springer, London (2004)
31. Slotine, J.J.E., Wang, W.: A study of synchronization and group cooperation using partial contraction theory. In: Kumar, V., Leonard, N.E., Morse, A.S. (eds.) Cooperative Control. Lecture Notes in Control and Information Sciences, Springer, London (2004)
32. Stan, G.B., Sepulchre, R.: IEEE Trans. Autom. Control 52, 256–270 (2007)
33. Tanner, H.G., Pappas, G.J., Kumar, V.: IEEE Trans. Robot. Autom. 20, 443–455 (2004)
34. Tsitsiklis, J.N., Bertsekas, D.P., Athans, M.: IEEE Trans. Autom. Control 31, 803–812 (1986)
35. Vicsek, T., Czirok, A., Ben-Jacob, E., Cohen, I., Schochet, O.: Phy. Rev. Lett. 75, 1226–1229 (1995)

On the Computation of Optimal Transport Maps Using Gradient Flows and Multiresolution Analysis

Ayelet Dominitz[1], Sigurd Angenent[2], and Allen Tannenbaum[1,3]

[1] Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa, Israel
ayeletdo@tx.technion.ac.il
[2] Department of Mathematics, University of Wisconsin, Madison, WI 53706, USA
angenent@math.wisc.edu
[3] Schools of Electrical & Computer and Biomedical Engineering, Georgia Institute of Technolgy, Atlanta, GA 30332-0250, USA
tannenba@ece.gatech.edu

Summary. The optimal mass transport methodology has numerous applications in econometrics, fluid dynamics, automatic control, statistical physics, shape optimization, expert systems, and meteorology. Further, it leads to some beautiful mathematical problems. Motivated by certain issues in image registration, visual tracking and medical image visualization, we outline in this note a straightforward gradient descent approach for computing the optimal L^2 optimal transport mapping which may be easily implemented using a multiresolution scheme. We discuss the well-posedness of our scheme, and indicate how the optimal transport map may be computed on the sphere.

Keywords: Optimal mass transport, multiresolution, visual tracking, wavelets.

1 Introduction

The problem of optimal mass transport problem was originally formulated by Gaspard Monge in 1781, who was concerned with finding the optimal way of moving a a given distribution of matter (a pile of soil or rubble for instance) into another (an excavation or fill) in the sense of minimal transportation cost. Monge's problem turned out to be the prototype for a class of questions arising in various areas of sciences including econometrics, fluid dynamics, automatic control, statistical physics, shape optimization, expert systems, and meteorology to name a few. A modern treatment of this problem was initiated by Kantorovich in 1942, leading to the Monge-Kantorovich formulation of the problem [10].

Our interest in the Monge-Kantorovich problem originally arose from our work in dynamic tracking and medical imaging applications. For the former application, optimal mass transport may be used as a key step in a geometric observer [13]. For the latter, optimal mass transport techniques may be employed for the registration of the proton density based imagery provided by MR. In this case, the intensity is a measure of density and thus the use of mass transport

V.D. Blondel et al. (Eds.) Recent Advances in Learning and Control, LNCIS 371, pp. 65–78, 2008.
springerlink.com

techniques is quite natural. Optimal transport techniques are also applicable to functional MR, where one wants to visualize the degree of activity in various features over time. In fact, these techniques can be employed in any application where volume or area preserving mappings are considered. In this paper, we will show that our method provides a means to obtain a regular area-preserving surface diffeomorphism starting with any diffeomorphism between surfaces with the same total surface area. We have found this technique useful for applications such as brain surface flattening [1].

Many algorithms have been considered for computing an optimal transport map. For example, one can find methods based on linear programming [14], Lagrangian mechanics closely related to ideas from the study of fluid dynamics [4]. While powerful, most of the the methods we have encountered are quite slow and computationally expensive in the L^2 case, which led us to consider a class of gradient descent algorithms more suited for image processing applications in [3].

Based upon this previous work, in the present paper we propose a solution for the optimal transport problem, which is obtained through the use of a multiresolution scheme. As in [3], we will employ a natural solution to the optimal transport problem based on the equivalent problem of polar factorization [5]. We then incorporate a multiresolution scheme into the solution which results in obtaining the optimal mapping via a gradient descent technique. The utility of our method is illustrated with a numerical example.

2 Problem Formulation

In this section, we give a precise mathematical formulation of the Monge-Kantorovich problem. Accordingly, consider two oriented Riemannian manifolds Ω_0 and Ω_1, each with a corresponding positive density function defined over it, denoted by μ_0 and μ_1, respectively. We assume that the total mass associated with each of the manifolds is equal,i.e.,

$$\int_{\Omega_0} \mu_0(x)\, dx = \int_{\Omega_1} \mu_1(x)\, dx \tag{1}$$

where dx is the standard Lebesgue measure. If this assumption is not satisfied, we can always scale one of the density distributions to make the total amount of mass equal.

We wish to find a smooth mass preserving transport map that takes the first distribution into the second one, i.e., $u : (\Omega_0, \mu_0) \to (\Omega_1, \mu_1)$. From change of variables such a diffeomorphism satisfies:

$$\mu_0 = |\nabla u| \mu_1 \circ u. \tag{2}$$

Here $|\nabla u|$ denotes the determinant of the Jacobian map ∇u and $\circ$ represents composition of functions. This equation is often referred to as the *Jacobian equation*, which constrains the mapping to be mass preserving (MP) with respect to the given density functions. A mapping u that satisfies this property may

thus be thought of as defining a redistribution of a mass of material from one distribution (Ω_0, μ_0) to another distribution (Ω_1, μ_1).

There may be many such mappings, and we want to choose an optimal one in some sense. This can be obtained by incorporating into the problem a penalty function of the following form:

$$C(u) = \int_{\Omega} \Phi(x, u(x)) \mu_0(x) dx \tag{3}$$

where $\Phi(x, u(x))$ is a positive twice-differentiable convex function, typically taken to be the geodesic distance (or its square) on the given manifold between x and $u(x)$. The contribution of this functional to the problem is to place a penalty on the distance the map u moves each bit of material, weighted by the material's mass. An optimal MP mapping (when it exists), is one that minimizes this functional over all while satisfying the mass preserving mappings.

A fundamental theoretical result [8,11] shows that a unique optimal MP map $u : (\Omega_0, \mu_0) \to (\Omega_1, \mu_1)$ exists for the L^2 case. In this paper, we will present a simple algorithm for the construction of the optimal MP map.

3 Evolution of Monge-Kantorovich Functional

Our algorithm for finding an optimal mapping assumes we have an initial mapping $u^0 : (\Omega_0, \mu_0) \to (\Omega_1, \mu_1)$ with the MP property. For general domains the initial mapping can be obtained using a method described in [6]. The details for constructing such an initial mapping will be described below. Given an initial mapping, we will rearrange the distribution in the domain of the map in order to minimize the cost functional $C(u)$, while constraining u so that it continues to satisfy equation (2). We think of u as a function of time $t \geq 0$, and introduce the notation u^t with u^0 being the initial mapping. These maps evolve for t so that

$$u^0(x) = u^t(s^t(x)),$$

where $s^t : \Omega_0 \to \Omega_0$ preserves the measure $\mu_0(x)dx$. Notice, that for $t = 0$ we have that $s^{t=0}$ is the identity map.

The diffeomorphism s^t can be defined by a velocity field on Ω_0 which satisfies:

$$\frac{\partial s^t(x)}{\partial t} = v^t\left(s^t(x)\right)$$

The maps u^t then satisfy the *transport equation*:

$$\partial_t u + v \cdot \nabla u = 0, \tag{4}$$

while the cost functional evolves by

$$\frac{dM(u^t)}{dt} = \int_{\Omega} \mu_0(x) v^t(x) \cdot \Phi_x(x, u^t(x))\, dx. \tag{5}$$

The maps s^t will preserve the measure $\mu_0(x)dx$ if and only if the velocity field v^t satisfies

$$\nabla \cdot (\mu_0 v) = 0. \tag{6}$$

If Ω_j, $j = 0, 1$ is two dimensional, then the general divergence free vector field is of the form

$$\mu_0 v = -J\nabla H(x)$$

where J is rotation by $+90°$ and H is a "Hamiltonian function."

To obtain a gradient flow, one could set

$$H = \nabla \cdot (J\Phi_x(x, u(x))) \tag{7}$$

as in [9]. In the special case that $\Omega \subset \mathbb{R}^2$ and $\Phi(x, y) = \frac{1}{2}|x - y|^2$ this leads to $\Phi_x(x, u) = x - u$, and, since $\nabla \cdot Jx = 0$,

$$H = -\nabla \cdot Ju(x).$$

In this setting, the maps should evolve by the following initial value problem:

$$\partial_t u + v \cdot \nabla u = 0, \; v = \frac{1}{\mu_0(x)} \nabla \{\nabla \cdot Ju\}. \tag{8}$$

In general, using (7), one can compute that

$$\begin{aligned} \frac{d}{dt} M(u) &= -\int_\Omega (J\nabla H) \cdot \Phi_x(x, u)\, dx \\ &= \int_\Omega (\nabla H) \cdot J\Phi_x(x, u)\, dx \\ &= -\int_\Omega H \nabla \cdot (J\Phi_x(x, u))\, dx \end{aligned}$$

In the integration by parts calculation, we assume $H = 0$ on $\partial\Omega$, if this boundary is nonempty. This is natural, since then $J\nabla H$ is tangential to the boundary.

4 Well-posedness of the Solution

Unfortunately, the initial value problem (8) may not be well-posed, as we will see in this section. We moreover discuss the existence of solutions. Consider the case where $\Omega \subset \mathbb{R}^2$, and where the cost is quadratic in the distance. Then the maps u^t evolve by (8). We now compute the evolution equation for the Hamiltonian H. In index notation we have

$$H = -\nabla_k (J_{kl} u_l) = -J_{kl} \nabla_k u_l, \tag{9}$$

$$\partial_t u_k + v_l \nabla_l u_k = 0, \tag{10}$$

$$v_k = \frac{1}{\mu_0(x)} \nabla_k \nabla_l (J_{lm} u_m) = \frac{1}{\mu_0(x)} J_{lm} \nabla_k \nabla_l u_m, \tag{11}$$

where

$$(J_{kl}) = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

We find that

$$\begin{aligned}\partial_t H &= -J_{kl}\nabla_k(\partial_t u_l)\\ &= J_{kl}\nabla_k\left(v_m\nabla_m u_l\right)\\ &= v_m J_{kl}\nabla_k\nabla_m u_l + J_{kl}\left(\nabla_k v_m\right)\left(\nabla_m u_l\right)\\ &= -v_m\nabla_m H + J_{kl}\left(\nabla_k v_m\right)\left(\nabla_m u_l\right).\end{aligned}$$

Note that we have chosen the velocity v to be perpendicular to ∇H, since $\mu_0 v = -J\nabla H$. Hence the term $v\cdot\nabla H$ vanishes, and we get

$$\partial_t H = J_{kl}\left(\nabla_k v_m\right)\left(\nabla_m u_l\right).$$

Now use the formula for v_m to get

$$\begin{aligned}\partial_t H &= J_{kl}\nabla_k\left(\frac{-1}{\mu_0}J_{mn}\nabla_n H\right)\nabla_m u_l\\ &= J_{kl}J_{mn}\nabla_m u_l\nabla_k\left(\frac{-1}{\mu_0}\right)\nabla_n H - \frac{1}{\mu_0}J_{kl}(\nabla_m u_l)J_{mn}\nabla_k\nabla_n H.\end{aligned}$$

In other words, H satisfies a second order PDE of the form

$$\partial_t H = a_{kn}\nabla^2_{kn}H + b_m\nabla_m H \tag{12}$$

where

$$b_m = J_{kl}J_{nm}\nabla_n u_l\frac{\nabla_k m_0}{\mu_0^2}$$

and

$$a_{kl} = -\frac{1}{\mu_0}J_{kl}\nabla_m u_l J_{mn} = \frac{1}{\mu_0}\left(J^t\cdot\nabla u\cdot J\right)_{kl}$$

The antisymmetric part of a_{ij} is irrelevant since $\nabla^2 u$ is symmetric, so we may replace a_{kl} by

$$\tilde{a}_{kl} := \frac{1}{2\mu_0}\left(J^t\cdot\left[\nabla u + (\nabla u)^t\right]\cdot J\right)_{kl}.$$

The evolution equation for H is therefore forward (backward) parabolic if $\tilde{a}_{kl}$ is positive (negative) definite, i.e., if $(\nabla u)+(\nabla u)^t$ is positive (negative) definite. For instance, if u is close to the identity then the equation will be forward parabolic; if u is close to a 180° rotation, then the H-equation is backward parabolic.

4.1 Weak Solutions

If μ_0 is smooth, then one can rewrite the evolution equation for u in divergence form. Namely,

$$\begin{aligned}\mu_0\partial_t u_m &= -\mu_0 v_k\nabla_k u_m\\ &= J_{kl}\nabla_l H\ \nabla_k u_m\\ &= J_{kl}\nabla_l\left(H\nabla_k u_m\right) - J_{kl}H\nabla_k\nabla_l u_m\\ &= J_{kl}\nabla_l\left(H\nabla_k u_m\right)\\ &= -\nabla_l\left(HJ_{lk}\nabla_k u_m\right)\end{aligned}$$

where we have used that J_{kl} is antisymmetric and $\nabla_k \nabla_l u_m$ is symmetric in k, l to conclude that $H J_{kl} \nabla_k \nabla_l u_m = 0$. So we find that u and H satisfy

$$\partial_t (\mu_0 u) + \nabla \cdot (H(J \cdot \nabla)u) = 0, \; H = -\nabla \cdot (Ju) \tag{13}$$

To define a weak solution one only needs ∇u to belong to $L^2(\Omega)$, for if ∇u is square integrable, then both $H = \nabla \cdot (Ju)$ and $(J \cdot \nabla)u$ are in L^2, so that their product is a well defined integrable function. The first equation in (13) can then be interpreted in the sense of distributions.

Theorem 1. *Let the density μ_0 be smooth and strictly positive on $\bar{\Omega}$. Consider a C^1 weak solution u of (13) on $\Omega \times [0,T]$.*

If $(\nabla u) + (\nabla u)^t$ is everywhere positive definite then for all $t \in (0,T]$ the function $H(\cdot,t)$ belongs to $C^{1,\alpha}(\bar{\Omega})$.

If $(\nabla u) + (\nabla u)^t$ is everywhere negative definite, then $H(\cdot,t) \in C^{1,\alpha}(\bar{\Omega})$ for all $t \in [0,T)$. In particular the initial value $u(x,0)$ must have been such that $H = \nabla \cdot (Ju)$ is $C^{1,\alpha}$.

This follows immediately from the fact that the coefficients of the equation (12) which H satisfies are continuous.

The second part of the theorem shows that the initial value problem will have no C^1 solution on any short time interval $[0,T)$ if the initial map $u(x,0)$ is

— close to the map $u(x,0) \approx -x$ so that $\nabla u \approx -I$, and:
— C^1, while $H = \operatorname{curl} u = \nabla \cdot (Ju)$ is continuous but not differentiable.

Finally we note that the optimal map, when it exists, is often the gradient of a **convex** function, i.e. $u = \nabla\varphi$. In this case $\nabla u = \nabla\nabla\varphi$ is the second derivative of φ, and therefore is certainly positive definite. In other words, the flow defined by 12 is bound to be well-posed near the optimal map.

If at some point $(\nabla u) + (\nabla u)^t$ is not positive definite, then there are infinitely many directions in function space in which one can reduce the MK cost functional.

5 Multiresolution Based Gradient Flow

Since the initial value problem (8) may not be well-posed for all choices of initial data, one needs a different flow scheme. One option would be to take the Hamiltonian defined by (7) and smooth it. Here we take a different route which leads to an efficient multiscale scheme for the computation of the optimal transport map.

Here we choose a linearly independent set $\{\varphi_1, \varphi_2, \cdots\} \subset C^1(\bar{\Omega})$, and assume that the Hamiltonian has the form

$$H = \sum_{i=1}^{N} a_i(t)\varphi_i(x). \tag{14}$$

Such a choice of Hamiltonian leads to the following cost evolution

$$\frac{d}{dt}M(u^t) = -\sum_{i=1}^{N} a_i(t) \int_{\Omega} J\nabla\varphi_i(x) \cdot \Phi_x\left(x, u^t\right)\, dx. \tag{15}$$

This in turn leads one to choose H as in (14) with

$$a_i(t) = \int_{\Omega} J\nabla\varphi_i(x) \cdot \left(\Phi_x\left(x, u^t\right)\right)\, dx \tag{16}$$

Then

$$\frac{d}{dt}M(u^t) = -\sum_{i=1}^{N} a_i(t)^2.$$

Thus in this case, the maps should evolve by the following initial value problem

$$\begin{cases} \partial_t u + v \cdot \nabla u = 0 \\ v = -\frac{1}{\mu_0(x)} \sum_{i=1}^{N} a_i\left(t\right) J\nabla\varphi_i\left(x\right) \end{cases} \tag{17}$$

where $a_i\left(t\right)$ is as defined in 16. The flow will stop when $\nabla \cdot \left(J\Phi_x\left(x, u^t\right)\right)$ is orthogonal to $\{\varphi_1, \cdots, \varphi_N\}$.

A natural choice for an orthonormal basis $\{\varphi_i\}_{i=1}^{N}$ are the Laplace-Beltrami eigenfunctions. The eigenfunctions of the Laplace-Beltrami operator solve the stationary Helmholtz wave propagation equation $\Delta f = \lambda f$. The solutions of this differential equation on the surface are often referred to as the solutions of the general vibration problem. These can be calculated for any manifold and have certain multiresolution properties. Therefore by fixing the number of eigenfunctions we use at a given time, we can determine the various resolutions for which convergence to the optimal transport map will be obtained.

One could also choose the φ_i to be a wavelet basis $\psi_{j,m}$, such as the second generation wavelets introduced in [17,18]. These wavelet functions are L_2 basis functions which can be defined intrinsically on the manifolds and do not depend on any parameterization. Wavelet functions $\psi_{j,m}$ represent functions on a regular grid at multiple levels. of details. Here j indexes the scale (resolution of analysis) where higher j correspond to higher resolution analysis, and m indexes the spatial location of analysis. An important example of these wavelets is the construction of wavelets on the sphere, which was introduced in [15,16]. In this construction the lifting scheme is used to obtain locally supported, biorthogonal spherical wavelets and their associated fast transforms. The construction starts from a recursive triangulation of the sphere and is independent of parameterization. Since the construction does not rely on any specific properties of the sphere it can be generalized to other surfaces.

In our gradient descent formulation, we can group the wavelet functions $\psi_{j,m}$ according to their resolution level (j). Each of these groups can span any function defined over the manifold at a different resolution. We can then run our gradient flow at each resolution level separately, going from coarser to higher resolution,

until it converges. This gradient flow will be obtained by iterating the following steps (starting with $J = 1$) until the MK cost functional ceases to decrease (or until $\int H^2 \, dx < \varepsilon$ for some tolerance $\varepsilon > 0$):

1. Using the basis vectors $\{\psi_{j,m} | \, 1 \leq j \leq J \, , \, \forall m\}$ follow steepest descent.
2. Increase J by 1, so as to include basis functions corresponding to the next level of resolution.

6 The Case of the Sphere

The special case in which Ω_1 is the unit sphere is of particular interest for medical imaging applications since a significant portion of the surfaces of 3D anatomical structures are topological spheres (simply-connected compact surfaces). Therefore, a reasonable 1-1 mapping of minimal distortion of such surfaces to a sphere can be used for registration and visualization. After registration, the spherical representation can be analyzed using such spherical multiresolution functions as spherical harmonics or spherical wavelets.

Our method of finding the optimal area preserving mapping of some simply connected surface to the sphere is as follows [9]. We assume that we have an initial diffeomorphism from a given simply connected anatomical structure to the sphere. A method to find this mapping will be elaborated in Section 7. We define Ω_0 to be the flattened surface, i.e. the sphere, and μ_0 on Ω_0 will account for the area distortion by setting it to be the local ratio between the area on the original surface and the area of the surface once flattened. Now, we only need to find the optimal mapping u that will redistribute the mass of material μ_0 over the sphere to make it a uniform distribution. We set Ω_1 to be the sphere as well, and we set μ_1 to be uniformly 1 throughout Ω_1 and solve the optimal transport problem as described in the previous section. Since $\mu_1 = 1$, the constraint given in equation (2) reduces to $\mu_0 = |\nabla u|$, and so by the construction of μ_0, a mapping u which satisfies this constraint compensates exactly for the distortion in area that occurred during the flattening process.

In the case of the sphere, the above derivation for finding the optimal transport map can be rewritten in a very simple analytical form. For $\Omega_j = S^2 \subset \mathbb{R}^3$, $j = 0, 1$ and with $\Phi(x, y)$ as the geodesic distance from x to y we have

$$\Phi(x, y) = \arccos(x \cdot y).$$

Since Φ_x is the projection of the derivative of $\Phi(x, y)$ as function from $\mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}$ onto the tangent plane to S^2 at x, this leads to

$$\Phi_x(x, y) = -\frac{y - (x \cdot y)x}{\sqrt{1 - (x \cdot y)^2}}.$$

Since $Jy = x \times y$ and

$$(x \cdot u)^2 + |x \times u|^2 = |x^2| \, |u|^2 = 1,$$

we have

$$J\Phi_x(x,u) = -\frac{x \times u}{|x \times u|}. \tag{18}$$

Therefore, in this setting the maps should evolve by the initial value problem described in equation 17. Where for $a_i(t)$ defined as in (16) we have:

$$a_i(t) = \int_\Omega J\nabla\varphi_i(x) \cdot \left(\Phi_x\left(x, u^t\right)\right)\, dx = \int_{S^2} \nabla\varphi_i(x) \cdot \frac{x \times u}{|x \times u|}\, dx. \tag{19}$$

In the case of the sphere the intrinsic multiresolution bases we use also take a simple form:

Laplace-Beltrami Eigenfunctions: The stationary Helmholtz wave propagation equation $\Delta f = \lambda f$ has an analytical solution over the sphere. The resulting eigenfunctions coincide with the spherical harmonic functions. The spherical harmonics form a complete set of orthonormal functions. Thus, on the unit sphere, any square-integrable function can be expanded as a linear combination of these. The coefficients of the spherical harmonic functions of different degrees provide a measure of the spatial frequency constituents that comprise the function.

Spherical Wavelets: Spherical wavelets are second-generation wavelets, which are adapted to manifolds with non-regular grids [15, 16]. The main difference with traditional wavelet constructions is that it does not rely on the Fourier transform. This way one can entirely abandon the idea that wavelets are necessarily translates and dilates of one function. All the algorithmic details may be found in [15, 16]

7 Numerical Experiments

In this section, we outline the steps we perform to obtain the optimal MP mapping from a general domain to a sphere, using the techniques described above.

In general, there are different quality measures for the goodness of the mapping. From one point of view, we wish to obtain a mapping that preserves the local geometry of the original surface, and this can be obtained using a conformal (angle preserving) mapping [2]. On the other hand, it is reasonable to require the mapping to be area preserving. However, In general it is not possible to map a surface with non-constant Gaussian curvature to the sphere in a way which preserves both angles and areas. Such a mapping would be an isometry and thus curvature-preserving. Therefore, as a compromise, we wish to find the area preserving mapping that will minimize the geometrical distortion. In other words we are seeking an area preserving mapping of minimal distortion in some precise sense [1].

In order to do this, we first flatten the given structure to the sphere using conformal flattening. Then we minimally redistribute this mapping, so that it will be area preserving, using the method described in Section 6.

The algorithm is defined via a three step approach. The first step is to find the conformal mapping from some general manifold Ω_0 to the sphere. The area

distortion of this mapping will be interpreted as a density function μ_0 on S^2. The second step is to find an initial MP mapping $u^0 : (S^2, \mu_0) \to (S^2, \mu_1 = 1)$. The third step is to minimize the cost functional $C(u)$ of the map u^0, by rearranging the distribution in the domain of the map. We now give the details.

7.1 Construction Conformal Mapping

We employ the approach for conformal mapping described in [2]. This method is based on finding the solution of a second-order partial differential equation. Let $\Omega_0 \in \mathbb{R}^3$ represent the initial simply connected surface that we wish to map to the unit sphere S^2. Let δ_p denote the Dirac delta function at an arbitrary point p of the surface Ω_0, Δ is the Laplace-Beltrami operator on $\Omega_0 \backslash \{p\}$. In this framework, we obtain a conformal equivalence $z : \Omega_0 \backslash \{p\} \to S^2 \backslash \{\text{north pole}\}$as the solution of the following partial differential equation:

$$\Delta z = \left(\frac{\partial}{\partial u} - i \frac{\partial}{\partial v} \right) \delta_p, \tag{20}$$

where u and v are conformal coordinates defined in the neighborhood of p [7]. In [2], it is shown that the conformal mapping procedure may be implemented for a triangulated representation of the surface using a finite element method.

As noted above, the conformal mapping distorts the area of the surface. However one can quantify the change of area as the density function μ_0 at each point, so that the integral on the flattened surface (the unit sphere) $\int_{\Omega_0} \mu_0 dx$ will give us the area measure of the original surface. The density function is the determinant of the Jacobian of f^{-1}:

$$\mu_0 = |\nabla f^{-1}|$$

7.2 Construction of Initial Mapping

For a general domain, the initial mapping can be obtained using a technique proposed in [12,6]. We wish to find a diffeomorphism $\phi : S^2 \to S^2$ taking $\mu_0(x)\,dx$ into $\mu_1(x)\,dx = 1dx$. This mapping satisfies $\int_{S^2} \mu_0(x)\,dx = \int_{S^2} \mu_1(x)\,dx$, or equivalently $|\nabla \phi(x)|\mu_1(x) = \mu_0(x)$. The equations defined by these demands are underdetermined and have more than one possible solution. In his proof for the existence of such a diffeomorphism, Moser [12] derives a construction for obtaining such a mapping.

The basic idea is as follows. Let $\mu_t(x)\,dx$ be a family of closed 2-forms which are nondegenerate for $0 \leq t \leq 1$, i.e.

$$\int_{S^2} \mu_t(x)\,dx = \int_{S^2} \mu_0(x)\,dx.$$

For every $\mu_t(x)\,dx$ exists an automorphism ϕ_t taking $\mu_0(x)\,dx$ into $\mu_t(x)\,dx$, and therefore satisfying:

$$|\nabla \phi_t(x)|\mu_t(x) = \mu_0(x).$$

ϕ_0 is the identity map.

Instead of determining the mapping ϕ_t directly we first find a one parameter vector-field u_t for $t \in [0,1]$ on S^2 from which we shall recover by solving the ordinary differential equation:

$$\frac{\partial}{\partial t}\phi_t = v_t(\phi_t).$$

This has the advantage that v_t, is obtained from linear equations while ϕ_t, satisfies some nonlinear equations. We can then write

$$|\nabla \phi_t(x)|\left((1-t)\mu_0(x)+t\right) = \mu_0(x)$$

and through some manipulation of equations ,we find that in order to obtain v_t, it is sufficient to solve

$$\Delta\Theta = 1 - \mu_0(x)$$

and then

$$v_t = \frac{-\nabla\Theta}{(1-t)\mu_0 + t}.$$

Notice that the above formulation ignores the boundary values, because in the current settings we solve for surfaces with no boundary.

After performing this step, we have an area preserving mapping with respect to the original domain, which is obtained as the composition of the mapping obtained by the method elaborated above and the conformal mapping.

7.3 Construction of Optimal MP Map

We now use the MP mapping obtained in Section 7.2, denoted by u_0, as the initial mapping for the optimal transport algorithm. We obtain a gradient descent flow defined by the initial value problem in Section 6, which will lead us to the optimal area preserving mapping. In our case $\Omega_1 = S^2 \subset \mathbb{R}^3$ and our cost function is the geodesic distance on the sphere between our mapping and the conformal mapping. As shown in Section 6, in this setting the maps should evolve by the initial value problem (17).

We tested our algorithm using both multiresolution bases previously described, namely, spherical harmonics and spherical wavelets. Using either one of these two bases the gradient descent evolution can be calculated numerically on a discrete mesh by performing the following two steps at each iteration:

- Calculate the coefficients for the Hamiltonian $H = \sum_{i=1}^{N} a_i(t)\varphi_i(x)$ as

$$a_i(t) = \sum_j A_j \nabla\varphi_i[j] \cdot \frac{x[j] \times u_t[j]}{|x[j] \times u_t[j]|}$$

  where i indexes the wavelets, j indexes the mesh points and A_j is the area associated with each mesh point.

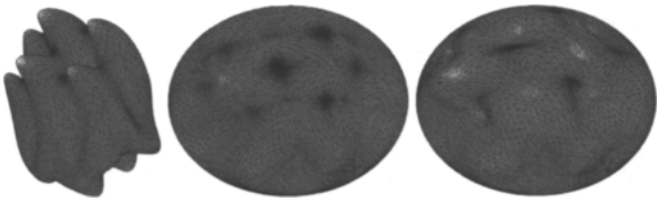


Fig. 1. On the left is initial "tooth" surface. In the middle is the tooth after conformal mapping to the sphere. On the right is the image of the area preserving mapping obtained using our gradient descent method. This mapping causes minimal local geometry distortion in the sense described in the text.

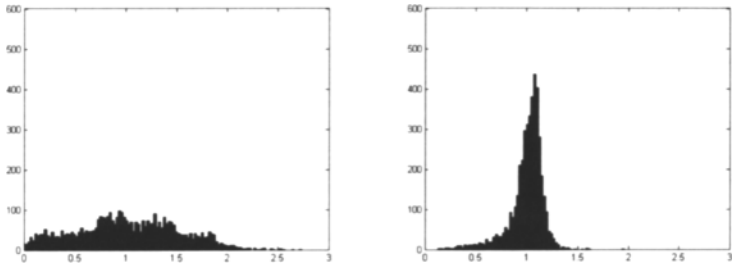


Fig. 2. Area Ratio Histograms: (a) On the left after conformal mapping; (b) On the right after Monge-Kantorovich

- Calculate the new vector field:

$$u_{t+\Delta t}[j] = u_t[j] + \Delta t \cdot \frac{1}{\mu_0[j]} \cdot \sum_i a_i(t) J\varphi_i[j] \cdot \nabla u[j].$$

We tested our method on a simple tooth image. Results are shown in Figures 1 and 2. Results were similar for the spherical wavelets and the spherical harmonics. However, it should be noted, that when we used the same number of basis functions, the spherical wavelet basis converged to smaller error bounds with fewer iterations (at a ratio of about 7).

8 Conclusions

In this paper, we derived a solution for the optimal transport problem based on a gradient flow, and implemented via a multiresolution scheme which guarantees fast convergence. We also showed how our work may be extended in a straightforward manner to the sphere which makes it useful for a number of medical imaging applications.

Acknowledgements

This work was supported in part by grants from NSF, AFOSR, ARO, MURI, MRI-HEL, NIH (NAC P41 RR-13218) through Brigham and Women's Hospital, and a Marie Curie Grant through the Technion. 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. Angenent, S., Haker, S., Tannenbaum, A., Kikinis, R.: On area preserving maps of minimal distortion. In: Djaferis, T., Schick, I. (eds.) System Theory: Modeling, Analysis, and Control, pp. 275–287. Kluwer, Holland (1999)
2. Angenent, S., Haker, S., Tannenbaum, A., Kikinis, R.: Laplace-Beltrami operator and brain surface flattening. IEEE Trans. on Medical Imaging 18, 700–711 (1999)
3. Angenent, S., Haker, S., Tannenbaum, A.: Minimizing flows for the Monge–Kantorovich problem. SIAM J. Math. Analysis 35, 61–97 (2003)
4. Benamou, J.-D., Brenier, Y.: A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. Numerische Mathematik 84, 375–393 (2000)
5. Brenier, Y.: Polar factorization and monotone rearrangement of vector-valued functions. Comm. Pure Appl. Math. 64, 375–417 (1991)
6. Dacorogna, B., Moser, J.: On a partial differential equation involving the Jacobian determinant. Ann. Inst. H. Poincaré Anal. Non Linéaire 7, 1–26 (1990)
7. Do Carmo, M.P.: Differential Geometry of Curves and Surfaces. Prentice-Hall, Inc., Englewood Cliffs (1976)
8. Feldman, M., McCann, R.J.: Monge's transport problem on a Riemannian manifold. Trans. Amer. Math. Soc. 354, 1667–1697 (2002)
9. Haker, S., Zhu, L., Tannenbaum, A., Angenent, S.: Optimal mass transport for registration and warping. Int. Journal Computer Vision 60, 225–240 (2004)
10. Kantorovich, L.V.: On a problem of Monge. Uspekhi Mat. Nauk. 3, 225–226 (1948)
11. McCann, R.: Polar factorization of maps on Riemannian manifolds. Geom. Funct. Anal. 11, 589–608 (2001)
12. Moser, J.: On the volume elements on a manifold. Trans. Amer. Math. Soc. 120, 286–294 (1965)
13. Niethammer, M., Vela, P., Tannenbaum, A.: Geometric observers for dynamically evolving curves. IEEE PAMI (submitted, 2007)
14. Rachev, S., Rüschendorf, L.: Mass Transportation Problems and Probability and Its Applications, vol. I, II. Springer, New York (1998)
15. Schröder, P., Sweldens, W.: Spherical wavelets: Efficiently representing functions on the sphere. In: Proceedings SIGGRAPH 1995 Computer Graphics, ACM Siggraph, pp. 161–172 (1995)

16. Schröder, P., Sweldens, W.: Spherical wavelets: Texture processing. In: Hanrahan, P., Purgathofer, W. (eds.) Rendering Techniques 1995, pp. 252–263. Springer, New York (1995)
17. Sweldens, W.: The Lifting Scheme: A New Philosophy in Biorthogonal Wavelet Constructions. In: Laine, A.F., Unser, M. (eds.) Proc. SPIE, Wavelet Applications in Signal and Image Processing III, pp. 68–79 (1995)
18. Sweldens, W.: The lifting scheme: A construction of second generation wavelets. SIAM J. Math. Anal. 29, 511–546 (1997)

Realistic Anchor Positioning for Sensor Localization*

Barış Fidan[1], Soura Dasgupta[2], and Brian D.O. Anderson[1]

[1] Research School of Information Sciences and Engineering, The Australian National University and National ICT Australia, Canberra, Australia
{Baris.Fidan,Brian.Anderson}@anu.edu.au
[2] Dept. of ECE, University of Iowa, Iowa City, IA 52242, USA
dasgupta@engineering.uiowa.edu

Summary. This paper considers localization of a source or a sensor from distance measurements. We argue that linear algorithms proposed for this purpose are susceptible to poor noise performance. Instead given a set of sensors/anchors of known positions and measured distances of the source/sensor to be localized from them we propose a potentially non-convex weighted cost function whose global minimum estimates the location of the source/sensor one seeks. The contribution of this paper is to provide nontrivial ellipsoidal and polytopic regions surrounding these sensors/anchors of known positions, such that if the object to be localized is in this region, localization occurs by globally exponentially convergent gradient descent in the noise free case. Exponential convergence in the noise free case represents practical convergence as it ensures graceful performance degradation in the presence of noise. These results guide the deployment of sensors/anchors so that small subsets can be made responsible for practical localization in geographical areas determined by our approach.

Keywords: Localization, Sensors, Global Convergence, Optimization, Gradient Descent.

1 Introduction

Over the last few years the problem of source/sensor localization has assumed increasing significance. Specifically *source localization* refers to a set of sensors estimating the precise location of a source using information related to their relative position to the source. In *sensor localization* a sensor estimates its own position using similar information relative to several nodes of known positions called *anchors or beacons.* This information can be distance, bearing, power level (indirectly related to distance) and time difference of arrival (TDOA).

We observe that localization is a fundamental component of a number of emerging applications, [1]. For example a network of sensors deployed to combat bioterrorism, must not only detect the presence of a potential threat, but must

* This work is supported by National ICT Australia, which is funded by the Australian Government's Department of Communications, Information Technology and the Arts and the Australian Research Council through the Backing Australia's Ability Initiative and NSF grants CCF-0729025 and ECS-0622017.

V.D. Blondel et al. (Eds.) Recent Advances in Learning and Control, LNCIS 371, pp. 79–94, 2008.
springerlink.com

also locate the source of the threat. Similarly, in pervasive computing, [2], [5], [6], locating printers in a building permits a computer aware of its own position to send its print job to the nearest printer. Likewise, [7], in sensor networks individual sensors must know their own positions, to route packets, detect faults, and detect and record events.

As compellingly, [8] catalogs a burgeoning multibillion dollar market surrounding wireless location technology. Examples include, E911 services that must respond to a growing number of emergency calls placed by mobile phones, whose time critical localization is crucial to timely response; mobile advertising that provides location specific advertisement to mobile phones or personal digital assistants (PDAs); indoor and outdoor asset tracking for such advanced public safety applications as retrieving lost children, patients and the like; fleet management to permit taxi operators, police forces and emergency reponders to minimize response times; location based wireless access security that heightens wireless network security and avoids the malicious interception of digital information, by permiting only location specific users to access information through Wireless Local Area Networks (WLANs); and location specific billing.

As has been noted in [2] in large scale sensor networks where nodes may move, manual configuration is not an option. Further GPS, [9], is too expensive in terms of both hardware and power, and ineffective in the absence of line-of-sight between satellites and receivers. In many applications the line-of-sight between satellites and receivers is absent due to the presence of foliage and other obstacles, or because of indoor settings.

To support the full scope of localization based applications, it is important to ensure that localization occurs in an efficient and time critical manner. As we will argue in the sequel the efficiency and accuracy of localization largely depends on how the sensors/anchors are delpoyed. Thus, in the sensor localization problem one would like to deploy anchors in a manner that achieves an efficient sensor localization over as large a geographical area as permitted by anchor and sensor capabilities. Likewise, in the source localization problem one would like to deploy sensors to achieve a similar coverage that is consistent with source strength and sensing power. Indeed, the principal focus of this paper is to study the issue of network deployment to achieve localization in an efficient and time critical fashion.

We do so by focusing on the case where distance measurements are available. Localization in this context means that given known 2 or 3- dimensional vectors $x_1, \cdots, x_N$ ($N > 2$ and $N > 3$ in 2 and 3 dimensions respectively) and an unknown vector y^*, one must estimate the value of y^*, from the measured distances $d_i = \|y^* - x_i\|$. Here, as elsewhere in this paper, unless otherwise noted, all vector norms are 2-norms. In the source localization problem, y^* represents the position of the unknown source, and the x_i the positions of the sensors seeking to estimate its location. In the sensor localization problem, the x_i are the positions of the anchors, and y^* the position of the sensor estimating its own position.

We note that distances can be estimated through various means. For example if a source emits a signal, the signal intensity and the characteristics of the medium provides a distance estimate. In this case with A the source signal strength, and η the power loss coefficient, the received signal strength (RSS) at a distance d from the source is given by

$$s = A/d^{\eta}. \tag{1}$$

Thus, A, s and η provide d. Alternatively, a sensor may transmit signals of its own, and estimate the distance by measuring the time it takes for this signal reflected off the target to return. Another means of estimating distances is when a group of sensors collaboratively use TDOA information.

Research in sensor/anchor deployment has been largely conducted at the network level, [2]- [4]. To place these papers in context consider the issue of anchor deployment for sensor network localization, where sensors are able to measure their distances from designated anchors within their geographical area. Indeed suppose a sensor located at a position y^* seeks to estimate its location by measuring distances from N anchors with known locations $\{x_1, \cdots, x_N\}$. In 2-dimensions, localization from distance measurements generically requires that distances of y^* from at least three non-collinearly situated x_i be available. To be precise, with just two distances, the position can be determined to within a binary ambiguity. Occasionally, *a priori* information may be available which will resolve that ambiguity. Otherwise, a third distance is needed. In three dimensions, barring additional information one generically needs at least four non-coplanar x_i. Papers on network localization are concerned with deploying enough anchors so that a sufficient number of distance measurements are available to localize all the sensors. In doing so [3], for example, treats the anchors and sensors as nodes in a graph, and assumes the existence of an edge between two nodes should a distance measurement between these two modes be available, and shows that localizability is equivalent to this graph being *globally rigid*, [3]. In [2] the authors show that in general the problem of network localization is NP-hard. This and other references such as [2] - [4] identify specific graph topologies under which localization can be performed in polynomial time and in a robust fashion.

Implicit in this theory is the assumption that all it takes for a sensor to localize itself is that it acquire a sufficient number of distance measurements from anchors at known positions. Thus in 2-dimensions a sensor is assumed to be robustly localizable if it can measure with reasonable accuracy its distance from three anchors that are sufficiently noncollinear. In three dimensions one must have distances from at least four sufficiently non-coplanar anchors. Indeed as explained in section 2 if the distance $d_i = \|x_i - y^*\|$ uniquely specify y^*, then y^* can be estimated using linear algorithms, [8], [10]. In practice, as also explained in section 2, *such a linear algorithm with certain geometries may deliver highly inaccurate estimates with noisy measurements of the distances, even when the noise is small, and the anchors well positioned, e.g. in two dimensions they are non-collinear to a sufficient extent.* Thus in practice several papers adopt a nonlinear estimation approach, e.g. [11]- [15], [20].

On the other hand, several papers adopt a nonlinear estimation approach, [11]- [15]. Specifically, these papers formulate cost functions whose minimization provides y^*. In all these cases the underlying cost functions are non-convex and there minimization complicated by locally attractive false optima, which in practical terms preclude the possibility of fast accurate localization.

Consequently a central premise of this paper is that it is not just enough to place anchors and/or sensors to ensure that the objects to be localized have a sufficient number of distance measurements. Rather in making these placements it is equally important to take into consideration the processing problems posed by the underlying estimation procedure.

To this end we review an approach presented and analyzed in [20]. This proposes a new weighted cost function whose minimization achieves localization. Though potentially manifested with local minima, this cost function has the following attractive property. Given a small number of anchors, one can characterize nontrivial, easily determined convex regions surrounding them, such that whenever an object to be localized lies in this region, the gradient descent minimization of this cost function is globally exponentially convergent. As explained in Section 3 this guides what we deem to be anchor deployment for *practical localization.*

Section 2 explains why linear localization can be problematic and reviews some standard nonlinear approaches. Section 3 describes our approach. Sections 4 and 5 describe two types of regions of the type described above. Section 6 is the conclusion. All proofs are in [20].

2 Past Linear and Nonlinear Approaches

In this section we discuss the basic deficiency of linear algorithms and describe some conventional nonlinear algorithms.

2.1 Linear Algorithms

We discuss now the practical ramifications of linear localization algorithms. Consider three non-collinear x_i in 2-dimensions and equations

$$\|x_i - y^*\|^2 = d_i^2, \qquad \text{for } i \in \{1,2,3\}. \tag{2}$$

Subtracting the first equation from the remaining two one obtains,

$$2\begin{bmatrix}(x_1-x_2)^T\\(x_1-x_3)^T\end{bmatrix} y^* = \begin{bmatrix}\|x_1\|^2 - \|x_2\|^2 + d_2^2 - d_1^2\\ \|x_1\|^2 - \|x_3\|^2 + d_3^2 - d_1^2\end{bmatrix}. \tag{3}$$

For non-collinear x_i, $det([(x_1-x_2),(x_1-x_3)]^T) \neq 0$, i.e. y^* can be solved uniquely. But, the solution is invariant if for any α the d_i^2 are replaced by $d_i^2+\alpha$, suggesting and verified by example in [16], that such linear algorithms may have poor noise performance.

Indeed consider in two dimensions $x_1 = 0$, $x_2 = [43,7]^T$, $x_3 = [47,0]^T$ and $y^* = [17.9719, -29.3227]^T$. Its distances from the x_i are: $d_1 = 34.392$, $d_2 = 44.1106$

and $d_3 = 41.2608$. Now suppose the measured distances from x_1, x_2 and x_3 are 35, 42 and 43 respectively. These estimated distances are inconsistent in that no single y can simultaneously meet these distance constraints. The linear algorithm provides an estimate of y^* that is $[16.8617, -6.5076]^T$, i.e. relatively small errors in distance measurements translate to very substantial localization error. *On the other hand, the gradient descent optimization of the cost function we provide here, with unity weights, initialized with the estimate provided by the linear algorithm, converges to* $[18.2150, -29.2443]^T$ *a point that is much closer to* y^*.

2.2 Nonlinear Approaches

Many papers directly work with (1), with known A and η. Thus rather than assuming that the d_i are directly available, they work with the RSS at several sensors and choose y^*, to be the y that minimizes

$$\sum_{i=1}^{N} \left(s_i - \frac{A}{\|x_i - y\|^{\eta}} \right)^2 . \tag{4}$$

In (4) there are N-sensors, the i-th located at x_i, and s_i is the RSS at the i-th sensor. It should be noted that unless one makes the unrealistic assumption that the noise perturbing the RSS is Gaussian, (4) *does not provide the Maximum Liklihood (ML) esitmate of* y^*. A more realistic assumption on the noise perturbing s_i is that it is lognormal.

Cost functions such as (4) are inevitably non-convex and their minimization manifested with locally attractive false optima. This is true for example for [12]-[13] who conduct searches *in two dimensions*, that as noted in [14] are sensitive to spurious stationary points. While it is easy to detect convergence of search procedures to false minima, one would have to reinitialize the search process, potentially multiple times, wasting precious power, and impairing time critical localization.

As partial amelioration, [14] provides search alternatives involving the so called Projection on Convex Sets (POCS) approach also in two dimensions, with $\eta = 2$. It has however, the unique solution of y^* in the noise free case if and only if y^* is in the convex hull of the x_i. Convergence fails if y^* lies outside the convex hull.

As an extension [15] proposes the so called hyperbolic POCS, that does sometimes converge even if y^* is outside the convex hull of the x_i. However, no characterization of conditions for convergence is given, and false stationary points abound.

Yet another approach is to use the lognormal model, [21], that assumes that RSS of (1) is replaced by,

$$10 \log_{10}(s_i) = 10 \log_{10}(A) - \eta 10 \log_{10}(\|x_i - y\|) + n_i \tag{5}$$

where n_i are mutually uncorrelated zero mean white Gaussian noise, with variance σ^2. The lognormal model reflects the modeling error in the physical estimation of η.

In this case straightforward calculations show that the ML estimate of y^* is the global minimum of

$$\sum_{i=1}^{N} \left[\log_{10} \left(\frac{s_i \|x_i - y\|}{A} \right) \right]^2 .$$

Again this cost function has mutiple spurious minima whose ready characterization is unavailable.

3 The Approach

The foregoing indicates that sensor/anchor placement that guarantee availability of sufficient numbers of distance measurements will not suffice for practical convergence. One must also take into account the convergence behavior of any nonlinear algorithm one employs. In particular, to avoid repeated re-initializations that sap energy and induce delays, it is important to characterize the sorts of regions we have described above. In particular we wish to characterize regions surrounding *small numbers of sensors/anchors* such that if a source/sensor lies in them, the resulting cost function can be minimized using a gradient descent law that is *globally exponentially convergent* in the noise free case. We assert that such a characterization facilitates *practical localization in three respects.*

First, there are several results in the nonlinear stability literature (e.g. [18], chapter 5) that show that global exponential convergence in the noise free case guarantees graceful performance degradation in the presence of noise. These include results where convergence in distribution, [17], with variance increasing with increasing noise variance is proven. Second, assigning just a small number of sensors/anchors the responsibility of localizing a given region also has practical benefits: Acquisition of large numbers of distances strains the systems resources both in terms of computational and in some cases communication costs. Third, such regions remove the need for repeated initializations which deplete resources and are inimical to time critical localization. Characterization of such regions for cost functions such as (4), which in any case does not lead to a ML estimate, is difficult as the corresponding gradient descent update kernel is rational for integer η and irrational in general. The corresponding cost function for ML estimation under a more realistic log normal noise assumption is even more problematic, due to the presence of logarithms in the update kernel. Consequently any benefits accruing from obtaining an ML estimate are outweighed by the lack of guaranteed convergence.

Instead our starting point is that some how distance measurements have been obtained without reference to the underlying signal model except possibly, as in [15], in its use in obtaining distance estimates. Localization is accomplished by obtaining y^* as the y minimizing the cost function defined below for certain designer selected weights $\lambda_i > 0$,

$$J(y) = 0.5 \sum_{i=1}^{N} \lambda_i \left(\|x_i - y\|^2 - d_i^2 \right)^2 , \tag{6}$$

In the cost function (6), each addend term $\lambda_i \left(\|x_i - y\|^2 - d_i^2\right)^2$, for $i = 1, \ldots, N$, penalizes the difference between the calculated distance of x_i to the source/sensor position estimate y and the measured distance of x_i to y^*. The coefficients λ_i ($i = 1, \cdots, N$) are weighting terms which can be chosen based on the any additional *a priori* information that maybe available. For example, if it is known that certain d_i estimates are more reliable than others, then one may choose the corresponding λ_i to be larger. On the other hand, see Section 5, these weight selections may also facilitate practical convergence in the sense adopted in this paper.

Of course, other indices can be contemplated. For example one could work with $\|x_i - y\| - d_i$, rather than the difference of the squares, and one could work with the $2m$-th power of the difference, rather than the second power. Working with the difference $\|x_i - y\| - d_i$, rather than the difference of the squares would be much harder to treat by the methods of this paper, as the derivative of the index is more awkward analytically. On the other hand, working with a power $2m$ rather than 2 may be a more tractable extension.

3.1 Preliminaries of (6)

Our standing assumption below ensures that $J(y) = 0$ if and only if $y = y^*$.

Assumption 1. *In two dimensions $N > 2$ and the x_i, $i \in \{1, \cdots, N\}$, are non-collinear. In three dimensions $N > 3$ and they are non-coplanar.*

As an initial point we will seek to find conditions under which

$$\partial J(y)/\partial y = \sum_{i=1}^{N} \lambda_i (\|y - x_i\|^2 - d_i^2)(y - x_i) = 0 \text{ iff } y = y^*. \tag{7}$$

The fact that $y = y^*$ guarantees that $\partial J(y)/\partial y = 0$ is trivially seen by noting that by definition $\|y^* - x_i\|^2 = d_i^2$ for all i. Examples presented in this section show that in fact the reverse implication does not always hold.

Consider the iterative gradient descent algorithm

$$y[k+1] = y[k] - \mu \sum_{i=1}^{N} \lambda_i \left(\|y[k] - x_i\|^2 - d_i^2\right) (y[k] - x_i) \tag{8}$$

where $y[k]$, for $k = 1, 2, \ldots$, denotes the estimate of y^* at the kth iteration, and $y[0]$ is the initial estimate to be chosen. Under (7), it is well known that given an arbitrary constant $C > 0$, there exists μ^* dependent on C such that for all $\mu \leq \mu^*$, the algorithm (8) is globally uniformly asymptotically convergent to $y = y^*$ for every $\|y[0]\| \leq C$. As a matter of fact in keeping with our goal of ensuring practical localization, the conditions we obtain are in fact stronger, in that they guarantee exponential rather than just the uniform asymptotic convergence guaranteed by (7).

We provide two examples where (7) fails. First leads to a situation where the resulting false stationary point is locally unstable.

Example 1. Thus consider the 2-dimensional case where $\lambda_i = 1$, $x_1 = [-1, 0]^T$, $x_2 = [0, -1]^T$, $x_3 = 0$, and $y^* = [-1, -1]^T$ depicted in fig. 1(a). In this case $d_1^2 = d_2^2 = 1$ and $d_3^2 = 2$. Observe in this case, with $y = [y_1, y_2]^T$ the first element of $\partial J/\partial y$ is given by

$$((y_1+1)^2 + y_2^2 - 1)(y_1+1) + ((y_2+1)^2 + y_1^2 - 1)y_1 + (y_1^2 + y_2^2 - 2)y_1$$

just as the second element is provided by

$$((y_1+1)^2 + y_2^2 - 1)y_2 + ((y_2+1)^2 + y_1^2 - 1)(y_2+1) + (y_1^2 + y_2^2 - 2)y_2.$$

The underlying symmetry of the two expressions ensures that they are simultaneously zero only if

$$y_1 = y_2 = \delta$$

and δ must obey:

$$\begin{aligned} ((\delta+1)^2 + \delta^2 - 1)(2\delta+1) + (2\delta^2 - 2)\delta = 0 \\ \Leftrightarrow 2\delta\left((\delta+1)(2\delta+1) + \delta^2 - 1\right) = 0 \\ \Leftrightarrow 6\delta^2(\delta+1) = 0. \end{aligned}$$

Thus the only two stationary points are $y = y^*$ and $y = 0$.

In (8) $y_1[k+1] - y_1[k]$ equals

$$-\mu\left[2y_1^2[k] + (y_1[k] + y_2[k])^2 + 3y_1[k](y_1^2[k] + y_2^2[k])\right].$$

For sufficiently small $\|y[k]\|$, the first two terms dominate. Thus, if $y_1[k] < 0$, and $y[k]$ is in the vicinity of the origin, $y_1[k+1] < y_1[k]$, exhibiting the local instability of $y = 0$. In practical terms, as is well known (see e.g. [18] chapter 2), the local instability of this stationary point will make it unattainable in that trajectories will rarely stick to it. Nonetheless this stationary point is inconsistent with the requirements of global exponential convergence.

On the other hand, there are examples, e.g. Example 2 below where the spurious stationary points are locally stable. This occurs, [19] if the Hessian

$$\begin{aligned} \mathcal{H} &= \frac{\partial}{\partial y}\left[\sum_{i=1}^{N} \lambda_i(\|y - x_i\|^2 - d_i^2)(y - x_i)\right] \\ &= 2\sum_{i=1}^{N} \lambda_i(y - x_i)(y - x_i)^T - \sum_{i=1}^{N} \lambda_i(\|y - x_i\|^2 - d_i^2)I \end{aligned}$$

is positive definite at such a stationary point.

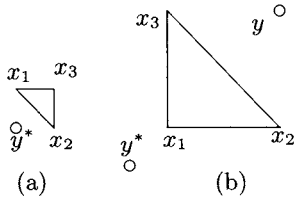


Fig. 1. (a) False unstable stationary point at x_3. (b) False stable stationary point at y.

Example 2. Choose $\lambda_i = 1$, $x_1 = [1,1]^T$, $x_2 = [1,3]^T$, $x_3 = [3,1]^T$, and the true $y^* = 0$ depicted in fig. 1(b). In this case $d_3^2 = d_2^2 = 10$ and $d_1^2 = 2$. By direct verification it can be seen that $y = [3,3]^T$ is a spurious stationary point. At this point

$$\mathcal{H} = 2\begin{bmatrix} 8 & 4 \\ 4 & 8 \end{bmatrix} - 6I$$
$$= \begin{bmatrix} 10 & 8 \\ 8 & 10 \end{bmatrix}$$

which is positive definite. Thus $y = [3,3]^T$ is in fact a local minimum.

With respect to example 2 note also that the basin of attraction of the false minimum at $[3,3]^T$ is quite large, in fact larger than that of $y^* = 0$. Thus one cannot hope that a sufficiently large reinitialization will move the trajectory away from this false minimum.

A subtle point to note is the following. One cannot just say that if the object to be localized lies in the basin of attraction of y^*, then localization is guaranteed. This is so as different y^* lead to different sets of d_i and consequently *different cost functions.* All one can say is that this basin of attraction provides the intializations of (8) under which a source located at this y^* can be localized by the given x_i.

One could of course seek to characterize these basins of attraction. However the argument above also shows that these must be computed anew for each different source location; a task that is computationally infeasible. This motivates our approach that instead, given x_i provides regions where the underlying cost function has *no false minima*, so that globally convergent gradient descent minimization is guaranteed.

4 Guaranteed Convergence for Given Weights

The last section demonstrates that (6) may well have local minima. In this section we provide a sufficient condition under which (7) holds for *fixed* $\lambda_i > 0$. There are various equivalent statements of the sufficient condition. One (see Theorem 3) is that y^* lies in a certain ellipsoid (or ellipse in two dimensions).

The sufficient condition provided here in fact has broader implications. Not only does it guarantee convergence of gradient descent minimization, but in fact induces exponentially fast convergence.

First some notation: We define the N-vector

$$u_N = [1, \cdots, 1]^T, \tag{9}$$

the $3 \times N$ or $4 \times N$ matrix in 2 and 3-dimensions respectively,

$$\mathcal{X} = [[x_1, \cdots, x_N]^T, u_N]^T, \tag{10}$$

the 3×1 or 4×1 vector in 2 and 3-dimensions respectively,

$$\bar{y} = [y^{*T}, 1]^T \tag{11}$$

and $\text{Co}(x_1, \cdots, x_N)$ as the convex hull of $\{x_1, \cdots, x_N\}$. We provide an initial result.

Lemma 1. *Consider x_i in 2 or 3-dimensions with assumption 1 holding. Then for every y^* there exist scalar β_i obeying*

$$\sum_{i=1}^{N} \beta_i = 1 \tag{12}$$

for which

$$\sum_{i=1}^{N} \beta_i x_i = y^*. \tag{13}$$

If $\beta_i \geq 0$ for all i, then $y^* \in \text{Co}\{x_1, \cdots, x_N\}$. Further, for $N > 3$ in 2-dimensions and $N > 4$ in 3-dimensions, the β_i that obey (12), (13) are in general non-unique. We now develop a condition (contained in Lemma 2 below) involving β_i and λ_i to ensure (7).

Define

$$\beta = [\beta_1, \cdots, \beta_N]^T, \tag{14}$$

and

$$\tilde{x}_i = x_i - y^* \text{ and } \tilde{y} = y - y^*. \tag{15}$$

Then (7) holds if and only if

$$\sum_{i=1}^{N} \lambda_i (\|\tilde{y} - \tilde{x}_i\|^2 - \|\tilde{x}_i\|^2)(\tilde{y} - \tilde{x}_i) = 0 \Leftrightarrow \tilde{y} = 0. \tag{16}$$

Further because of (12), (13),

$$\sum_{i=1}^{N} \beta_i \tilde{x}_i = 0. \tag{17}$$

With

$$\Lambda = \text{ diag } \{\lambda_1, \cdots, \lambda_N\}, \tag{18}$$

$$e_i(y) = \tilde{y} - \tilde{x}_i \text{ and } E(y) = [e_1(y), \cdots, e_N(y)], \tag{19}$$

as distances are invariant under a coordinate translation,

$$\begin{aligned} \partial J/\partial y &= \sum_{i=1}^{N} \lambda_i (\|\tilde{y} - \tilde{x}_i\|^2 - \|\tilde{x}_i\|^2)(\tilde{y} - \tilde{x}_i) \\ &= \sum_{i=1}^{N} \lambda_i \tilde{y}^T (\tilde{y} - 2\tilde{x}_i)(\tilde{y} - \tilde{x}_i) \\ &= \sum_{i=1}^{N} \lambda_i \left(2(\tilde{y} - \tilde{x}_i)(\tilde{y} - \tilde{x}_i)^T - (\tilde{y} - \tilde{x}_i)\tilde{y}^T\right) \tilde{y} \\ &= 2E(y)\Lambda E^T(y)\tilde{y} - E(y)\Lambda u_N \tilde{y}^T \tilde{y}. \end{aligned} \tag{20}$$

Because of (17), and (12)

$$E(y)\beta = \tilde{y}. \tag{21}$$

Define

$$P = \Lambda \left(2I - u_N \beta^T\right), \tag{22}$$

and observe from (20) and (21), that

$$\begin{aligned} \partial J/\partial y &= 2E(y)\Lambda E(y)^T \tilde{y} - E(y)\Lambda u_N \beta^T E(y)^T \tilde{y} \\ &= E(y)PE^T(y)\tilde{y}. \end{aligned} \tag{23}$$

Then we provide an intermediate result that concerns a sufficient condition on P for assuring (7).

Lemma 2. *Consider y^*, $x_1, \cdots x_N$ with assumption 1 in place. Suppose $\beta = [\beta_1, \cdots, \beta_N]^T$, obeys (12) and (13), $\lambda_i > 0$, E is as in (15), (19), and P is defined through (9), (18), (22). Suppose also that*

$$P + P^T > 0. \tag{24}$$

Then we have:

(A)There exists $\alpha > 0$, such that for all y,

$$E(y)(P + P^T)E^T(y) \geq \alpha I \tag{25}$$

(B)Condition (7) holds.

The Lemma provides only a sufficient condition for (7). Even if (24) is violated, (7) will hold unless some $\tilde{y} \neq 0$ is in the null space of $E(y)PE^T(y)$ which itself depends on $\tilde{y}$. Yet we show below that the condition of the Lemma does quantify

a nontrivial domain where (7) holds. Secondly, as we show in Theorem 1 under this condition not only does convergence occur but does so at an exponential rate.

Theorem 1. *Consider the algorithm update equation (8) and the various quantities defined in Lemma 2. Suppose (24) holds. Then for every $M > 0$, there exists a $\mu^*(M)$ such that $y[k] - y^*$ converges exponentially to zero whenever*

$$\|y[0] - y^*\| \leq M \tag{26}$$

and

$$0 < \mu \leq \mu^*(M). \tag{27}$$

Observe from (22) that P depends only on the λ_i and β_i. Thus our next goal is to characterize conditions on λ_i and β_i, for which (24) holds.

Theorem 2. *Consider in 2 or 3-dimensions, x_i, obeying assumption 1, $\lambda_i > 0$, Λ as in (18), P defined through (9), (18), (22) and any y^* and β obeying (14), (12) and (13). Then (24) holds if and only if:*

$$(\beta^T \Lambda^{-1} \beta)(u_N^T \Lambda u_N) = \left(\sum_{i=1}^{N} \beta_i^2/\lambda_i\right)\left(\sum_{i=1}^{N} \lambda_i\right) < 9. \tag{28}$$

Further (28) holds if (7) is true.

Consider now the special case of unity weights, i.e. $\lambda_i = 1$, and $N < 9$. Then we argue that $\mathrm{Co}(x_1, \cdots, x_N)$ is a *proper subset* of the region characterized by Theorem 2. Indeed if $y^* \in \mathrm{Co}(x_1, \cdots, x_N)$ then in (12) and (13) $\beta_i \geq 0$, and

$$\sqrt{\beta^T \beta} \leq \beta^T u_N = 1.$$

Thus as $\Lambda = I$,

$$(\beta^T \Lambda^{-1} \beta)(u_N^T \Lambda u_N) \leq 8.$$

Recalling that in 2 and 3-dimensions it suffices to have $N = 3$ and $N = 4$, respectively, for given x_i satisfying assumption 1, the set characterized by Theorem 2 can be chosen to be significantly larger than their convex hull. This means that in the sensor localization problem, just a few anchors will achieve substantial geographical coverage, just as in source localization just a few sensors will achieve a large coverage.

Theorem 2 characterizes the set for which (24) holds in terms of β, but not directly in terms of y^*. Of course the β_i themselves have a relationship to y^*. The next theorem exploits this relationship to characterize this set directly in terms of y^*, and in fact shows that the set of y^* satisfying (12), (13) and (28) is an *ellipsoid* (an ellipse in two dimension).

Theorem 3. *For every $\lambda_i > 0$, and x_i, obeying assumption 1, the set of y^* for which scalar β_i satisfying (12), (13) and (28) exist, is a nonempty ellipsoid, determined entirely by x_i and λ_i.*

5 Choosing the Weights

In the previous section, we took the λ_i as given, and characterized the y^* and β_i for which the condition for exponential convergence holds. Now the λ_i are effectively user-chosen parameters. One way of choosing them is to use *a priori* knowledge of the type described in the introduction. However, more in keeping with our objectives of sensor/anchor placement for practical convergence, in this section we quantify the set of y^* for which there exists a set of λ_i so that (28) holds. The first question we address is the following. By changing the λ_i can one alter false stationary points should they exist? An affirmative answer to this question permits one to detect convergence to a false stationary point by changing the λ_i by a nontrivial amount. The Theorem below shows that barring a highly non-generic situation, the answer is indeed yes. This non-generic situation arises when one of the x_i has exactly the same distance from each of the other x_j as y^*. In this case x_i is a false stationary point. Example 1 exemplifies this situation.

Theorem 4. *Consider distinct x_i, obeying assumption 1. Suppose for some $y \neq y^*$ and all nonnegative λ_i,*

$$\sum_{i=1}^{N} \lambda_i (\|y - x_i\|^2 - d_i^2)(y - x_i) = 0. \tag{29}$$

Then there exists $K \subset \{1, \cdots, N\}$, $|K| = N - 1$, such that for all $i \in K$ and $j \notin K$, $\|x_i - x_j\| = d_i$. Further, $y = x_j$.

Since generically the choice of λ_i affects the location and existence of false stationary points, we now characterize conditions under which λ_i exist for the sufficiency condition (24) to hold thereby guaranteeing practical convergence for the algorithm. A related question is: Given the existence of such a set of λ_i, how is one to select them? To understand the underlying intuition on this last question, suppose y^* is much closer to x_1 than the other x_i. Intuition suggests that one should emphasize d_1 more than the other d_i, by choosing λ_1 to be relatively larger. This is not just because the distance estimates at a closer location will be more reliable. Rather it is inherent in the underlying geometry of the problem. For example in an extreme case if a $d_i = 0$, then this measurement alone suffices for localization, i.e. the other weights can in the limit be selected as zero. The results of this section should be viewed in this context. We first present the following Theorem that characterizes regions in question in terms of the β_i.

Theorem 5. *Under the hypotheses of Theorem 2 there exist $\lambda_i > 0$ for which (24) holds if and only if*

$$\sum_{i=1}^{N} |\beta_i| < 3. \tag{30}$$

Further under (30)

$$\lambda_i = |\beta_i| \tag{31}$$

always guarantee (24).

Observe (31) in particular is in accord with the intuition we have foreshadowed. In particular, if y^* is very close to x_1 then there exists at least one choice of the β_i, such that β_1 has a much larger magnitude than the remaining β_i. For such a choice, (31) forces λ_1 to be much larger than the remaining λ_i.

Nonetheless, this result is in terms of β_i and thus only indirectly characterizes the spawn of the geographical area that guarantees (24). Theorem 6 below provides a direct characterization.

Theorem 6. *With x_i, obeying assumption 1, the largest set of y^* for each member of which one can find a choice of nonnegative λ_i that guarantee (24), has the following properties:*

(a) It is a convex polytope.
(b) It has Co$\{x_1, \cdots, x_N\}$ as a proper subset.
(c) It can be quantified by a linear program that is entirely determined by the x_i.

The fact that a linear program can determine this polytope is of course attractive from a computational view point. While this polytope contains Co$\{x_1, \cdots, x_N\}$ it is in fact much larger than just Co$\{x_1, \cdots, x_N\}$. Indeed consider the 2-dimensional example depicted in figure 2 where 1,2,3 represent the x_i locations. Choose $\beta_1 = 0$. Then y^* satisfying (12) and (30) are in the interval (4,5). Here [2,3] is a closed subinterval of (4,5), and the lengths of the segments joining 4 and 2, 3 and 5 and 2 and 3 are all equal. By similarly extending [1,3] and [1,2], one could come up with a hexagon 6,7,4,8,9,5 that defines the desired polytope.

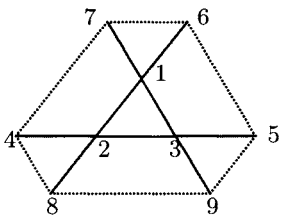


Fig. 2. Illustration of the polytope suggested by Theorem 5

Note also that though in this set (31) provides a choice of the λ_i, these are not the only choice one can make. The more one enters the interior of this polygon, the more the choices of λ_i, and indeed the larger the region where a common set of λ_i guarantees (7). This in particular has implications to the positioning of the x_i. Thus, with a potentially rough estimate of the position of a source, groups of sensors can collaborate to determine whether they can settle upon a λ_i which ensures (7). This provides guidance on how to deploy fewer sensors to achieve greater coverage.

Similar conclusions will follow for two-dimensional examples with $N > 3$, and for three-dimensional examples.

6 Conclusion

We have studied conditions under which a localization algorithm involves a globally exponentially convergent gradient descent minimization problem. In particular given a set of nodes with known positions (e.g. sensors), this algorithm seeks to localize an object (e.g. a source) whose distances from these nodes are available. Given a set of such sensors and a set of weights we characterize a nontrivial ellipsoidal geographical region surrounding these sensors such that a source lying in this region can be localized through a minimization described above. We also characterize a polytopic region for which there exist weights that permit similar localization. These characterizations provide guidance for placing sensors/anchors to achieve a desired level of geographical coverage.

References

1. IEEE Signal Processing Magazine (July 2005)
2. Aspnes, J., Eren, T., Goldenberg, D.K., Morse, A.S., Whiteley, W., Yang, Y.R., Anderson, B.D.O., Belheumer, P.N.: A theory of network localization. Transactions on Mobile Computing, 1663–1678 (December 2006)
3. Anderson, B.D.O., Belheumer, P., Eren, T., Goldenberg, D.K., Morse, A.S., Whiteley, W., Yang, Y.R.: Graph properties of easily localizable networks. Preprint Australian National University (2005)
4. He, T., Huang, C., Blum, B., Stankovic, J., Abedelhazer, T.: Range-free localization schemes in large scale sensor networks. In: Proceedings Ninth International Conference on Mobile Computing and Networking (September 2003)
5. Forman, G.H., Zahorzan, J.: The challenges of mobile computing. Computer, 38–47 (April 1994)
6. Weiser, M.: Some computer science problems in ubiquitous computing. Comm. ACM (July 1993)
7. Karp, B., Kung, H.T.: GPSR: Greedy perimeter stateless routing for wireless networks. In: Proceedings of Sixth International Conference on Mobile Computing and Networking (August 2000)
8. Sayed, A.H., Tarighat, A., Khajehnouri, N.: Network based wireless location. IEEE Signal Processing Magazine, 24–40 (July 2005)
9. Hoffman-Wellenhof, B., Lichtenhegger, H., Collins, J.: Global Positioning Systems: Theory and Practice, 4th edn. Springer, Heidelberg (1997)
10. Dandach, S., Fidan, B., Dasgupta, S., Anderson, B.D.O.: Adaptive source localization with mobile agents. In: Proceedings of CDC, SanDiego, CA, pp. 2045–2050 (December 2006)
11. Patwari, N., Ash, J.N., Kyperountas, S., Hero, A.O., Moses, R.L., Correal, N.S.: Locating the nodes: cooperative localization in wireless sensor networks. IEEE Signal Processing Magazine, 54–69 (July 2005)
12. Rabbat, M.G., Nowak, R.D.: Decentralized source localization and tracking. In: Proceedings of ICASSP, Montreal, Canada, vol. III, pp. III-921–III-924 (May 2004)
13. Rabbat, M.G., Nowak, R.D.: Distributed optimization in sensor networks. In: Proceedings of 3rd International Symposium on Information Processing in Sensor Networks, Berkeley, CA, pp. 20–27 (April 2004)

14. Blatt, D., Hero, A.O.: Energy-based sensor network source localization via projection onto convex sets. IEEE Transactions on Signal Processing, 3614–3619 (September 2006)
15. Rydstrom, M., Strom, E.G., Svensson, A.: Robust sensor network positioning based on projection onto circular and hyperbolic convex sets (POCS). In: Proceedings of SPAWC, Cannes, France (July 2006)
16. Cao, M., Anderson, B.D.O., Morse, A.S.: Sensor network localization with imprecise distance measurements. Systems and Control Letters 55, 87–93 (2006)
17. Bitmead, R.R.: Convergence in distribution of LMS-type adaptive parameter estimates. IEEE Transactions on Automatic Control, 54–60 (January 1983)
18. Khalil, H.K.: Nonlinear Systems. Prentice Hall, Englewood Cliffs (2002)
19. Vidyasagar, M.: Nonlinear Systems Analysis, ch. 5, Lyapunov Stability. Prentice Hall, Englewood Cliffs (1993)
20. Fidan, B., Dasgupta, S., Anderson, B.D.O.: Conditions for Guaranteed Convergence in Sensor and Source Localization. In: Proceedings of ICASSP, Honolulu, HI (April 2007) (also submitted to IEEE Transactions on Signal Processing)
21. Rappaport, T.S.: Wireless Communications: Principles & Practice. Prentice Hall, Englewood Cliffs (1996)

Graph Implementations for Nonsmooth Convex Programs

Michael C. Grant[1] and Stephen P. Boyd[2]

[1] Stanford University
mcgrant@stanford.edu
[2] Stanford University
boyd@stanford.edu

Summary. We describe *graph implementations*, a generic method for representing a convex function via its epigraph, described in a disciplined convex programming framework. This simple and natural idea allows a very wide variety of smooth and nonsmooth convex programs to be easily specified and efficiently solved, using interior-point methods for smooth or cone convex programs.

Keywords: Convex optimization, nonsmooth optimization, disciplined convex programming, optimization modeling languages, semidefinite programming, second-order cone programming, conic optimization, nondifferentiable functions.

1 Introduction

It is well known that convex programs have many attractive properties, including the proven existence of efficient methods to solve them. What is not as widely appreciated, however, is that nonsmooth convex programs—*i.e.*, models with nondifferentiable constraints or objectives—can, in theory, be solved just as efficiently as their smooth counterparts. But here, as is often the case, theory and practice do not coincide. Methods that are theoretically efficient for general nondifferentiable problems, such as the ellipsoid method [3], are notoriously slow in practice.

In contrast, there are many solvers available for smooth convex programs, as well as for certain standard forms such as semidefinite programs (SDPs), that are efficient in both theory and practice; *e.g.*, [13, 16, 17, 18]. These solvers can often be used to solve a general nonsmooth problem as well—not directly of course, but by first transforming it into an equivalent form supported by the solver. The equivalent problem is usually larger than the original, but the superior efficiency of the solver more than compensates for the increase in size, especially if problem structure is taken into account.

The transformation approach dates back to the very first days of linear programing [7]. It is usually taught as a collection of tricks that a modeler can use to (hopefully) reformulate problems by hand. The versatility of the approach, of course, depends upon the variety of transformations known by the modeler. But while some transformations are fairly obvious and widely known, others are

V.D. Blondel et al. (Eds.) Recent Advances in Learning and Control, LNCIS 371, pp. 95–110, 2008.
springerlink.com

neither obvious nor well known, even to some experts in convex optimization. Furthermore, even if a transformation is identified, the reformulation process is often time consuming and error prone, for both experts and applications-oriented modelers alike.

We propose to enable modeling frameworks to largely *automate* the process of identifying and applying these transformations, so that a much wider variety of models—smooth and nonsmooth alike—can be both easily specified and efficiently solved. Our approach depends upon two distinct but interrelated developments. The first is a methodology for constructing convex optimization models called *disciplined convex programming*. The methodology imposes a set of rules or conventions that must be followed when constructing convex programs. The rules are simple and teachable, drawn from basic principles of convex analysis, and follow the practices of those who regularly use convex optimization. Conforming problems are called, appropriately, *disciplined convex programs*, or *DCPs*. The DCP ruleset does not limit generality, but it does require that the modeler explicitly provide just enough structure to allow further analysis and solution of a problem to be automated.

The second development is a new method for defining or implementing a function in a modeling framework, as as the optimal value of a parameterized convex program (specifically, a DCP). We call such a function definition a *graph implementation* because it exploits the relationship between convex and concave functions and their epigraphs and hypographs, respectively. A graph implementation encapsulates a method for transforming instances of a specific function in a constraint or objective into a form compatible with the underlying solver's standard form. The conditions imposed by the DCP ruleset ensure that these transformations always preserve equivalence and convexity. The most significant benefit of graph implementations is their ability to efficiently implement nondifferentiable functions. But in fact, graph implementations can also be used to implement many *smooth* functions as well when the target standard form is nonsmooth (*e.g.*, an SDP).

We have created a modeling framework called `cvx` [8] that supports disciplined convex programming and graph implementations. `cvx` uses the object-oriented features of MATLAB® to turn it into an optimization modeling language: optimization variables can be declared and constraints and objectives specified using natural MATLAB® syntax. `cvx` verifies compliance with the DCP ruleset, transforms conforming models to solvable form, calls an appropriate numerical solver, and translates the numerical results back to the original problem—all without user intervention. The framework includes a large library of common convex and concave functions, both smooth and nonsmooth, and more can be added.

To an applications-oriented user, the conceptual model presented by `cvx` is very simple: `cvx` solves any problem (up to some practical size limits, of course) constructed according to the DCP ruleset from functions found in the `cvx` library. The modeler need not know what transformations are taking place, or even that a transformation is necessary. That is, graph implementations are entirely *opaque* or hidden from a standard `cvx` user. On the other hand, expert users

can use graph implementations to add new transformations to the system that general users can exploit simply by calling the new functions in their models. This division of labor allows modelers to focus on building convex programs, and experts to focus on solving them.

In what follows, we will describe disciplined convex programming and graph implementations in detail. Both are abstract, language-independent concepts; nevertheless, it will be easier to explain both using examples from an actual modeling framework such as `cvx`. So we begin by introducing `cvx` with a few simple examples. A basic familiarity with MATLAB® is assumed throughout.

2 A Brief Introduction to `cvx`

To begin, consider the simple linear program

$$\begin{array}{l} \text{minimize} \quad c^T x \\ \text{subject to } Ax \leq b, \end{array} \tag{1}$$

with variable $x \in \mathbf{R}^n$ and data $A \in \mathbf{R}^{m \times n}$, $b \in \mathbf{R}^m$, and $c \in \mathbf{R}^n$. The following (MATLAB®/`cvx`) code generates and solves a random instance of (1):

```
m = 16; n = 8;
A = randn(m,n); b = randn(m,1); c = randn(n,1);
cvx_begin
    variable x(n)
    minimize( c' * x )
    subject to
        A * x <= b;
cvx_end
```

The indentation is purely for stylistic reasons and is optional. The code is relatively self-explanatory, but a few notes are in order:

- The `cvx_begin` and `cvx_end` commands mark the beginning and end, respectively, of any `cvx` model.
- Variables must be declared before their first use. For example, the `variable` statement above declares `x` to be a vector of length `n`.
- The `subject to` statement is optional—it is provided by `cvx` only to make models more readable and has no mathematical effect.
- Objectives and constraints may be placed in any order.

When `cvx_end` is reached, `cvx` will complete the conversion of the model to solvable form and call the underlying numerical solver. It will replace the MATLAB® variable `x`, which up to that point was a special `cvx` variable object, with a numeric vector representing an optimal value.

Now consider a norm minimization problem with box constraints:

$$\begin{array}{l} \text{minimize} \quad \|Ay - b\|_2 \\ \text{subject to } \ell \leq y \leq u \end{array} \tag{2}$$

The following cvx/MATLAB® code constructs and solves a version of (2), reusing A and b from above and generating random bounds:

```
l = -abs(randn(n,2)); u = +abs(randn(n,2));
cvx_begin
    variable y(n)
    minimize( norm(A*y-b,2) )
    subject to
        y <= u;
        y >= l;
cvx_end
```

It is well known that (2) can be reformulated as a (convex) quadratic program (QP) or a second-order cone program (SOCP). (cvx, in fact, converts this problem to an SOCP.) The transformation in this case is not particularly complicated; still, it is nice to have it completely automated.

cvx supports a variety of other norms and penalties for this model simply by replacing the objective function; for example:

```
minimize( norm(A*y-b,2) + 0.1*norm(y,1) )
minimize( norm(A*y-b,3.5) )
minimize( sum(huber(A*y-b)) )
```

All of these examples can be reformulated as SOCPs as well. Here, however, the transformations are not at all obvious, even to experts; and in all three cases working out the transformation by hand would be tedious and error prone. cvx, however, can solve all three problems automatically.

As a final example, consider the task of determining the minimum volume ellipsoid (also known as the Löwner-John ellipsoid) $\mathcal{E}$ containing a finite set of points $z_1, z_2, \ldots, z_n \in \mathbf{R}^d$:

$$\begin{array}{ll} \text{minimize} & \text{vol}(\mathcal{E}) \\ \text{subject to} & z_i \in \mathcal{E}, \quad i = 1, \ldots, n. \end{array} \tag{3}$$

The parameterization we will use for $\mathcal{E}$ is

$$\mathcal{E} \triangleq \{\, u \mid \|Pu + q\|_2 \leq 1 \,\}, \tag{4}$$

where $(P, q) \in \mathbf{R}^{d \times d} \times \mathbf{R}^d$, and P is symmetric positive semidefinite. In this case, $\text{vol}(\mathcal{E})$ is proportional to $\det(P^{-1})$ (see [6, §8.4]). With this parametrization we can cast the problem above as

$$\begin{array}{ll} \text{minimize} & \det P^{-1} \\ \text{subject to} & \|Pz_i + q\|_2 \leq 1, \quad i = 1, 2, \ldots, n, \end{array} \tag{5}$$

with variables $P = P^T \in \mathbf{R}^{d \times d}$ and $q \in \mathbf{R}^d$. We have written the objective informally as $\det P^{-1}$; a more precise description is $f_{\texttt{det\_inv}}(P)$, where

$$f_{\texttt{det\_inv}}(P) \triangleq \begin{cases} \det(P)^{-1} & P = P^T \succ 0 \\ +\infty & \text{otherwise.} \end{cases} \tag{6}$$

This function implicitly constrains P to be symmetric and positive definite. The function $f_{\texttt{det\_inv}}$ is convex, so the problem above is a convex problem.

The following cvx/MATLAB® code generates a random set of points and computes the optimal ellipsoid by solving (5):

```
d = 2;
z = randn(d,n);
cvx_begin
    variables P(d,d) q(d)
    minimize( det_inv(P) )
    subject to
        for i = 1 : n,
            norm( P*z(:,i)+q,2 ) <= 1;
        end
cvx_end
```

The function `det_inv` represents $f_{\texttt{det\_inv}}(\cdot)$, including the implicit constraint that its argument be symmetric and positive definite. It is known that this problem can be cast as a semidefinite program (SDP), but the required conversion is quite complicated. Fortunately, that conversion is buried inside `cvx`'s definition of `det_inv` and performed automatically.

This is, of course, a considerably abbreviated introduction to `cvx`, intended only to give the reader an idea of the basic syntax and structure of `cvx` models. The reader is encouraged to read the user's guide [8] for a more thorough treatment, or to download the software and try it. The examples presented here can be entered exactly as listed.

3 Disciplined Convex Programming

Disciplined convex programming was first named and described by Grant, Boyd, and Ye in [9] and Grant in [10]. It was modeled on the methods used by those who regularly construct convex optimization models. Such modelers do not simply construct arbitrary nonlinear programs and attempt to verify convexity after the fact; rather, they begin with a mental library of functions and sets with known geometries, and combine them in ways which convex analysis guarantees will preserve convexity.

Disciplined convex programming is an attempt to formalize and this practice and codify its techniques. It consists of two key components:

- an *atom library*—a collection of functions or sets with known properties of curvature (convexity and concavity) and monotonicity; and
- the *DCP ruleset*—a finite enumeration of ways in which atoms may be combined in objectives and constraints while preserving convexity.

The rules are drawn from basic principles of convex analysis, and are easy to learn, once you have had an exposure to convex analysis and convex optimization. They constitute a set of sufficient but not necessary conditions for convexity,

which means that it is possible to build models that violate the rules but are still convex. We will provide examples of such violations and their resolution later in this section.

3.1 Preliminaries

The rules of disciplined convex programming depend primarily upon the *curvature* of numeric expressions. The four categories of curvature considered are *constant*, *affine*, *convex*, and *concave*. The usual definitions apply here; for example, a function $f : \mathbf{R}^n \to \mathbf{R}$ is convex if its domain is a convex set, and

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) \quad \forall x, y \in \mathbf{R}^n,\ \alpha \in [0,1]. \tag{7}$$

Of course, there is significant overlap in these categories: constant expressions are affine, and real affine expressions are both convex and concave. Complex constant and affine expressions are considered as well, but of course convex and concave expressions are real by definition.

Functions in the atom library are chracterized not just by curvature but by *monotonicity* as well. Three categories of monotonicity are considered: *nondecreasing*, *nonincreasing*, and *nonmonotonic*. Again, the usual mathematical definitions apply; for example, a function $f : \mathbf{R} \to \mathbf{R}$ is nondecreasing if

$$x \geq y \quad \Longrightarrow \quad f(x) \geq f(y). \tag{8}$$

Two technical clarifications are worth making here. First, monotonicity is considered in a global, extended-valued sense. For example, the MATLAB® square root function `sqrt` is interpreted in `cvx` as follows:

$$f_{\texttt{sqrt}} : \mathbf{R} \to (\mathbf{R} \cup -\infty), \quad f_{\texttt{sqrt}}(x) \triangleq \begin{cases} \sqrt{x} & x \geq 0 \\ -\infty & x < 0 \end{cases} \tag{9}$$

Under this interpretation, it is concave and nondecreasing. Secondly, for functions with multiple arguments, curvature is considered *jointly*, while montonicity is considered *separately* for each argument. For example, the function `quad_over_lin` in `cvx`, given by

$$f_{\texttt{qol}}(x,y) : (\mathbf{R}^n \times \mathbf{R}) \to (\mathbf{R} \cup +\infty), \quad f_{\texttt{qol}}(x,y) \triangleq \begin{cases} x^T x/y & y > 0 \\ +\infty & y \leq 0 \end{cases} \tag{10}$$

is jointly convex in x and y, but nonincreasing in y alone.

With terminology defined, we now proceed to the ruleset itself.

3.2 Constraints and Objectives

A disciplined convex program may either be an *optimization* problem consisting of a single objective and zero or more constraints, or a *feasibility* problem

consisting of one or more constraints and no objective. The rules for objectives and constraints are as follows:

- A valid objective is
  - the minimization of a convex expression;
  - the maximization of a concave expression.
- A valid constraint is
  - a set membership relation ($\in$) in which the left-hand side (LHS) is affine and the right-hand side (RHS) is a convex set.
  - an equality ($=$) with an affine LHS and an affine RHS.
  - a less-than inequality ($\leq$) with a convex LHS and a concave RHS.
  - a greater-than inequality ($\geq$) with a concave LHS and a convex RHS.

For any problem that conforms to these rules, the constraint set is convex. These rules, however, require more than just convexity of the constraint set: They constrain *how* the constraint set is described. For example, the constraint `square(x)==0`, where `x` is a scalar variable, defines the convex set $\{0\}$. But it is rejected by the rules above, since the LHS of this equality constraint is not affine. When the constraint is written in the equivalent form `x==0`, however, which is accepted by the rules above, since both sides are affine.

3.3 Simple Expressions

Disciplined convex programming determines the curvature of numeric and set expressions by recursively applying the following rules. This list may seem long, but it is for the most part an enumeration of basic rules of convex analysis for combining convex, concave, and affine forms: sums, multiplication by scalars, and so forth. For the basics of convex analysis, see, *e.g.*, [2, 4, 6, 15, 20].

- A valid affine expression is
  - a valid constant expression;
  - a declared variable;
  - a valid call to a function with an affine result;
  - the sum or difference of affine expressions;
  - the product of an affine expression and a constant.
- A valid convex expression is
  - a valid constant or affine expression;
  - a valid call to a function with a convex result;
  - the sum of two or more convex expressions;
  - the difference between a convex expression and a concave expression;
  - the product of a convex expression and a nonnegative constant;
  - the product of a concave expression and a nonpositive constant;
  - the negation of a concave expression.
- A valid concave expression is
  - a valid constant or affine expression;
  - a valid call to a function in the atom library with a concave result;
  - the sum of two or more concave expressions;

  - the difference between a concave expression and a convex expression;
  - the product of a concave expression and a nonnegative constant;
  - the product of a convex expression and a nonpositive constant;
  - the negation of a convex expression.
- A valid convex set expression is
  - a valid reference to a convex set in the atom library;
  - the intersection of two or more convex set expressions;
  - the sum or difference of convex set expressions;
  - the sum or difference of a convex set expression and a constant;
  - the product of a convex set expression and constant.

If an expression cannot be categorized by this ruleset, then it is rejected by `cvx`. For matrix and array expressions, these rules are applied on an elementwise basis. We note that the set of rules listed above is redundant; there are much smaller, equivalent sets of rules.

Of particular note is that these expression rules forbid *products* between non-constant expressions. We call this the *no-product rule* for obvious reasons. For example, the expression $x\sqrt{x}$, written in `cvx` as `x*sqrt(x)`, is convex (at least when x is positive) but is rejected by `cvx` as violating the above rules. Fortunately, `cvx` provides a function called `pow_pos(x,p)` that implements the convex and concave branches of `x^p`, so this expression can be written as `pow(x,3/2)`.

3.4 Compositions

A basic rule of convex analysis is that convexity is closed under composition with an affine mapping. This is part of the DCP ruleset as well:

- A convex, concave, or affine function may accept as an argument an affine expression (assuming it is of compatible size).

For example, consider the function `square`, which is provided in the `cvx` atom library. This function squares its argument; *i.e.*, it computes `x.*x`. (For array arguments, it squares each element independently.) It is known to be convex, provided its argument is real. So if `x` is a real variable, then

```
square( x )
```

is accepted by `cvx`; and, thanks to the above rule, so is

```
square( A * x + b )
```

if `A` and `b` are constant matrices of compatible size.

The DCP ruleset also provides for certain nonlinear compositions as well. The four composition rules are:

- If a convex function is nondecreasing in a given argument, then that argument may be convex.
- If a convex function is nonincreasing in a given argument, then that argument may be concave.

- If a concave function is nondecreasing in a given argument, then that argument may be concave.
- If a concave function is nonincreasing in a given argument, then that argument may be convex.

(In each case, we assume that the argument is of compatible size.) In fact, nearly every rule in the DCP ruleset can be derived from these composition rules.

For example, the pointwise maximum of convex functions is convex, because the maximum function is convex and nondecreasing. Thus if `x` is a vector variable then

```
max( abs( x ) )
```

obeys the first of the four composition rules and is therefore accepted by `cvx`. In fact, the infinity-norm function `norm( x, Inf )` is defined in exactly this manner. Affine functions must obey these composition rules as well; but because they are both convex and concave, they prove a bit more flexible. So, for example, the expressions

```
sum( square( x ) )
sum( sqrt( x ) )
```

are both valid nonlinear compositions in `cvx` since the rules for both the convex-nondecreasing and convex-nonincreasing cases apply to `sum`.

3.5 The Ruleset in Practice

As we stated in the introduction to this section, the DCP rules are sufficient but not necessary conditions for the convexity (or concavity) of an expression, constraint, or objective. Some expressions which are obviously convex or concave will fail to satisfy them. For example, if `x` is a `cvx` vector variable, then the expression

```
sqrt( sum( square( x ) ) )
```

is rejected by `cvx`, because there is no rule governing the composition of a concave nondecreasing function with a convex function. Fortunately, there is a simple workaround in this case: use `norm( x )` instead, since `norm` is in the atom library and is known by `cvx` to be convex.

This is an example of what is meant by our statement in the introduction that disciplined convex programming requires the modeler to supply "just enough" structure to enable the automation of the solution process. Obviously, both `norm` and the longer, non-compliant version are equivalent numerically, but the latter form enables `cvx` to complete the verification and conversion process. Of course, because the library is finite, there will inevitably be instances where a simple substitution is not possible. Thus to insure generality, the atom library must be expandable.

4 Graph Implementations

Any modeling framework for optimization must provide a computational description of the functions its supports to the underlying solver. For a smooth function, this traditionally consists of code to compute the value an derivatives of the function at requested points. In cvx, it is possible to define a convex or a concave function as the solution of a parameterized DCP. We call such a definition a *graph implementation*, a term first coined in [10] and inspired by the properties of epigraphs and hypographs of convex and concave functions, respectively.

4.1 The Basics

Recall the definition of the *epigraph* of a function $f : \mathbf{R}^n \to (\mathbf{R} \cup +\infty)$:

$$\mathbf{epi}\, f \triangleq \{\, (x, y) \in \mathbf{R}^n \times \mathbf{R} \mid f(x) \leq y \,\}. \tag{11}$$

A fundamental principle of convex analysis states that f is a convex function if and only if $\mathbf{epi}\, f$ is a convex set. The relationship between the two can be expresed in a reverse fashion as well:

$$f(x) \equiv \inf \{\, y \mid (x, y) \in \mathbf{epi}\, f \,\}. \tag{12}$$

(We adopt the convention that the infimum of an empty set is $+\infty$.) Equation (12) expresses f as the solution to a convex optimization problem—or, more accurately, a family of such problems, parameterized by the argument x.

A *graph implementation* of f takes the relationship in (12) and makes it concrete, by expressing $\mathbf{epi}\, f$ in a solvable manner—that is, with an equivalent collection of constraints in x and y that are compatible with the target solver. For example, consider the real absolute value function $f_{\mathtt{abs}}(x) = |x|$. Its epigraph can be represented as an intersection of two linear inequalities:

$$\mathbf{epi}\, f_{\mathtt{abs}} = \{\, (x, y) \mid |x| \leq y \,\} = \{\, (x, y) \mid x \leq y, -x \leq y \,\} \tag{13}$$

A graph implementation is just a description or *encapsulation* of that transformation, justified mathematically through a simple equivalency of sets.

In cvx, graph implementations can be specified using the same syntax as other cvx models, and are subject to the same DCP ruleset as well. The following cvx/MATLAB® code is a representation of $f_{\mathtt{abs}}$:

```
function y = f_abs(x)
cvx_begin
    variable y
    minimize( y )
    subject to
         x <= y;
        -x <= y;
cvx_end
```

(The absolute value function `abs` in `cvx` is actually implemented a bit differently; for example, it supports complex values and vector-valued arguments, in an elementwise fashion.)

If `f_abs` is called with a numeric value of `x`, then the `cvx` specification it contains will construct a linear program with a single variable and two inequalities. Upon reaching `cvx_end`, `cvx` will call the underlying solver and compute the correct result—at least to within the tolerances of the solver. This is, of course, a rather impractical way to compute the absolute value; in the real implementation of `abs` in `cvx` we avoid this inefficiency. But it is, at least, technically correct, and it is also a useful way to debug a graph implementation.

The more interesting case is when `f_abs` is used within a `cvx` model, with an affine `cvx` expression for an argument. In this case, the `cvx` specification will be incomplete, because the value of `x` is not yet known. What `cvx` does in this circumstance is to incorporate the specification *into the surrounding model* itself, in a manner not unlike the expansion of a `inline` function in C++. For example, if `z` is a scalar `cvx` variable, then the constraint

```
f_abs(z-3) <= 1;
```

will be translated internally by `cvx` as follows:

```
y <= 1;
x == z-3;
x <= y;
-x <= y;
```

(Steps are taken as needed to avoid name conflicts with existing variables.) The constraint is now in a form compatible with an efficient solver. Of course, two new variables and several new constraints have been added, but in the long run the added costs of expansions like this are far outweighed by the fact that a much more efficient solver can now be used, because the nondifferentiability has been eliminated.

Of course, the transformation of the absolute value function into an efficiently solvable form is relatively well known. But while it may be obvious to some, it is certainly not to everyone; and it is certainly convenient to have the transformation automated. For more advanced functions, the benefits should be more clear.

4.2 Advanced Usage

Graph implementations of convex functions are not, in fact, limited to strict epigraph representations. Suppose that $S \subset \mathbf{R}^n \times \mathbf{R}^m$ is a convex set and $\bar{f} : (\mathbf{R}^n \times \mathbf{R}^m) \to (\mathbf{R} \cup +\infty)$ is jointly convex in x and y; then

$$f : \mathbf{R}^n \to (\mathbf{R} \cup +\infty), \quad f(x) \triangleq \inf \left\{ \bar{f}(x,y) \;\middle|\; (x,y) \in S \right\} \tag{14}$$

is a convex function of x. If $m = 1$ and $\bar{f}(x,y) \triangleq y$, then the epigraph form (12) is recovered; but `cvx` fully supports this more general form.

For example, consider the unit-halfwidth Huber penalty function $h(x)$:

$$h : \mathbf{R} \to \mathbf{R}, \quad h(x) \triangleq \begin{cases} x^2 & |x| \leq 1 \\ 2|x| - 1 & |x| \geq 1 \end{cases} \tag{15}$$

This function cannot be used in an optimization algorithm utilizing Newton's method, because its Hessian is discontiuous at $x = \pm 1$, and zero for $|x| \geq 1$. However, it can be expressed in the form (14) in this manner:

$$h(x) \triangleq \inf \left\{ 2v + w^2 \mid |x| \leq v + w, \ w \leq 1 \right\} \tag{16}$$

We can implement the Huber penalty function in `cvx` as follows:

```
function cvx_optval = huber( x )
cvx_begin
    variables w v;
    minimize( 2 * v + square( w ) );
    subject to
        abs( x ) <= w + v;
        w <= 1;
cvx_end
```

If `huber` is called with a numeric value of `x`, then `cvx` will solve the resulting QP and return the numeric result. (As with `f_abs`, there is a simpler way to compute the Huber penalty when its argument is a numeric constant.) But if `huber` is called from within a larger `cvx` specification, then `cvx` will use this implementation to transform the call into a form compatible with the underlying solver. Note that the precise transformation depends on how `square` and `abs` are themselves implemented; multilevel transformations like this are quite typical.

There is a corresponding development for concave functions as well. Given the set S above and a concave function $\bar{g} : (\mathbf{R}^n \times \mathbf{R}^m) \to (\mathbf{R} \cup +\infty)$ is concave, the function

$$\bar{f} : \mathbf{R} \to (\mathbf{R} \cup +\infty), \quad \bar{f}(x) \triangleq \sup \{ g(x, y) \mid (x, y) \in S \} \tag{17}$$

is also a concave function. If $\bar{g}(x, y) \triangleq y$, then

$$\bar{f}(x) \triangleq \sup \{ y \mid (x, y) \in S \} \tag{18}$$

gives the *hypograph* representation of $\bar{f}$; that is, $S = \mathbf{hypo}\, f$. In `cvx`, a concave incomplete specification is simply one that uses a `maximize` objective instead of a `minimize` objective.

Some functions are not thought of as nondifferentiable in a casual setting but are technically so, and must be dealt with as such in an optimization algorithm. Consider, for example, the real square root function (9) above. This function is concave, and is smooth for positive x, but not at $x = 0$. Its hypograph, however, is

$$\mathbf{hypo}\, f_{\mathtt{sqrt}} \triangleq \left\{ (x, y) \mid x \geq 0, \ \sqrt{x} \geq y \right\} = \left\{ (x, y) \mid \max\{y, 0\}^2 \leq x \right\} \tag{19}$$

Thus a graph implementation can solve the nondifferentiability problem. In cvx, this function can be implemented as follows:

```
function y = f_sqrt(x)
cvx_begin
    variable y
    maximize( y )
    subject to
        square( y ) <= x
cvx_end
```

This particular type of nondifferentiability also occurs in the concave entropy function; it can be eliminated with a similar transformation.

4.3 Conic Solver Support

The most obvious benefit of graph implementations is their ability to describe nonsmooth functions in a computationally efficient manner. But the solvers used in the first publicly released versions of cvx posed a different challenge: they did not support smooth functions either. Rather, these solvers solved *semidefinite-quadratic-linear programs* (SQLPs)—problems of the form

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & \mathcal{A}x = b \\ & x \in \mathcal{K}_1 \times \mathcal{K}_2 \times \ldots \mathcal{K}_L \end{array} \tag{20}$$

where x is the optimization variable, $\mathcal{A}$ is a linear operator, b and c are vectors, and the sets $\mathcal{K}_i$ are convex cones from a select list: the nonnegative orthant $\mathbf{R}^n_+$, the second-order cone $\mathcal{Q}^n$, and the semidefinite cone $\mathcal{S}^n_+$:

$$\mathcal{Q}^n \triangleq \{ (x, y) \in \mathbf{R}^n \times \mathbf{R} \mid \|x\|_2 \leq y \} \tag{21}$$

$$\mathcal{S}^n_+ \triangleq \left\{ X \in \mathbf{R}^{n \times n} \;\middle|\; X = X^T,\ \lambda_{\min}(X) \geq 0 \right\} \tag{22}$$

Clearly, SQLPs are very closely related to SDPs; in fact, all SQLPs can be solved as SDPs. For more information about these problems, consult [12, 19], or the documentation on the solvers themselves [16, 17].

In practice, few application-driven models naturally present themselves as SQLPs; rather, modelers have simply recognized that their problems can be transformed into that form. In fact, as is known to readers of certain well-known texts on the subject [1, 5, 12, 14], SQLPs are *very* versatile, and can represent a wide variety of smooth and nonsmooth convex programs. The challenge, then, lies in finding an SQLP representation of of a given convex program—assuming one exists.

Using graph implementations, a variety of both smooth and nonsmooth functions were added to the cvx atom library for SQLP solvers, including minimums and maximums, absolute values, quadratic forms, convex and concave branches of the power function x^p, ℓ_p norms, convex and concave polynomials, geometric

means, eigenvalue and singular value functions, and determinants. Key omissions include logarithms, exponentials, and entropy; such functions simply cannot be exactly represented in an SQLP solver. (On the other hand, smooth solvers cannot solve many of the eigenvalue and determinant problems for which SQLP solvers excel.)

For a simple example, consider the function $f_{\mathtt{sq}}(x) \triangleq x^2$; its epigraph form (12) can be expressed using a single semidefinite cone:

$$f_{\mathtt{sq}}(x) \triangleq \inf \left\{ y \;\middle|\; \left[\begin{smallmatrix} y & x \\ x & 1 \end{smallmatrix}\right] \in \mathcal{S}^2_+ \right\}. \tag{23}$$

The `cvx` version of this function is

```
function y = f_sq(x)
cvx_begin
    variable y
    minimize( y )
    [ y, x ; x, 1 ] == semidefinite(2);
cvx_end
```

(Since MATLAB® does not have a set membership $\in$ operator, `cvx` uses equality constraints and functions like `semidefinite` to accomplish the same result.)

For a somewhat more complex example, consider the matrix fractional function $f_{\mathtt{mf}} : (\mathbf{R}^n \times \mathbf{R}^{n\times n}) \to (\mathbf{R} \cup +\infty)$, where

$$f_{\mathtt{mf}}(x,Y) = \begin{cases} \frac{1}{2} x^T Y^{-1} x & Y = Y^T \succ 0 \\ +\infty & \text{otherwise} \end{cases} \tag{24}$$

This function is convex in both arguments, and implicitly constrains `Y` to be both symmetric and positive definite. Its epigraph representation is

$$f_{\mathtt{mf}}(x,Y) \triangleq \sup \left\{ z \;\middle|\; \left[\begin{smallmatrix} Y & x \\ x^T & z \end{smallmatrix}\right] \in \mathcal{S}^{n+1}_+ \right\} \tag{25}$$

so it may be implemented in `cvx` as follows:

```
function cvx_optval = f_mf( x, Y )
n = length( x );
cvx_begin
    variable z;
    minimize( z );
    subject to
        [ Y, x ; x', z ] == semidefinite(n+1);
cvx_end
```

Both `f_sq` and `f_mf` are relatively simple examples in comparison to other functions in the `cvx` library. The complexity of some SQLP implementations is in some cases quite striking. For example, the ℓ_p norm can be represented exactly in an SQLP whenever $p = n/d$ is rational. The number of cone constraints required to represent it, however, depends not only on the size of the vector involved, but also in the pattern of bits in a binary representation of n and d! Needless to say, performing such transformations by hand is quite impractical—but once implemented, quite reasonable for a computer.

5 Final Words

We believe that disciplined convex programming closes a significant gap between the theory and practice of convex optimization. A large fraction of useful convex programs are nonsmooth; and until now, those who wished to solve them were faced with unattractive options: transform them by hand to a different, more easily solved form; develop a custom solver; utilize a poorly-performing subgradient-based method; or approximate. A modeling framework that supports disciplined convex programming provides a truly attractive alternative in most of these cases.

References

1. Alizadeh, F., Goldfarb, D.: Second-order cone programming (January 2004)
2. Bertsekas, D.P.: Convex Analysis and Optimization. In: Nedić, A., Ozdaglar, A.E. (eds.) Athena Scientific (2003)
3. Bland, R., Goldfarb, D., Todd, M.: The ellipsoid method: A survey. Operations Research 29(6), 1039–1091 (1981)
4. Borwein, J., Lewis, A.: Convex Analysis and Nonlinear Optimization: Theory and Examples. Springer, Heidelberg (2000)
5. Ben-Tal, A., Nemirovski, A.: Lectures on Modern Convex Optimization: Analysis, Algorithms and Engineering Applications. MPS/SIAM Series on Optimization. SIAM, Philadelphia (2001)
6. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge Univ. Press, Cambridge (2004), `http://www.stanford.edu/~boyd/cvxbook.html`
7. Dantzig, G.: Linear Programming and Extensions. Princeton University Press, Princeton (1963)
8. Grant, M., Boyd, S.: CVX: MATLAB® software for disciplined convex programming, version 1.1 (September 2007), `http://www.stanford.edu/~boyd/cvx/`
9. Grant, M., Boyd, S., Ye, Y.: Disciplined convex programming. In: Liberti, L., Maculan, N. (eds.) Global Optimization: from Theory to Implementation, Nonconvex Optimization and Its Applications, pp. 155–210. Springer Science+Business Media, Inc., New York (2006)
10. Grant, M.: Disciplined Convex Programming. PhD thesis, Department of Electrical Engineering, Stanford University (December 2004)
11. Löfberg, J.: YALMIP: A toolbox for modeling and optimization in MATLAB®. In: Proceedings of the CACSD Conference, Taipei, Taiwan (2004), `http://control.ee.ethz.ch/~joloef/yalmip.php`
12. Lobo, M., Vandenberghe, L., Boyd, S., Lebret, H.: Applications of second-order cone programming. Linear Algebra and its Applications 284, 193–228 (1998) (special issue on Signals and Image Processing)
13. MOSEK ApS. Mosek (software package) (September 2007), `http://www.mosek.com`
14. Nesterov, Yu., Nemirovsky, A.: Interior-Point Polynomial Algorithms in Convex Programming : Theory and Algorithms of Studies in Applied Mathematics, vol. 13. SIAM Publications, Philadelphia, PA (1993)
15. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)

16. Sturm, J.: Using SeDuMi 1.02, a MATLAB® toolbox for optimization over symmetric cones. Optimization Methods and Software 11, 625–653 (1999), Updated version available at `http://sedumi.mcmaster.ca`
17. Toh, K., Todd, M., Tutuncu, R.: SDPT3 — a MATLAB® software package for semidefinite programming. Optimization Methods and Software 11, 545–581 (1999)
18. Vanderbei, R.: LOQO: An interior point code for quadratic programming. Optimization Methods and Software 11, 451–484 (1999)
19. Vandenberghe, L., Boyd, S.: Semidefinite programming. SIAM Review 38(1), 49–95 (1996)
20. van Tiel, J.: Convex Analysis. An Introductory Text. John Wiley & Sons, Chichester (1984)

When Is a Linear Continuous-time System Easy or Hard to Control in Practice?

Shinji Hara[1] and Masaaki Kanno[2]

[1] Department of Information Physics and Computing, The University of Tokyo, Japan
Shinji_Hara@ipc.i.u-tokyo.ac.jp
[2] Japan Science and Technology Agency, Japan
M.Kanno.99@cantab.net

Summary. This paper is focused on characterization of easily controllable plants in practical control applications rather than to design an optimal or a robust controller for a give plant. After explaining the background and the motivation of the research topic, we first provide two notions, namely finite frequency positive realness (FFPR) and Condition (π), which represent desirable phase/gain properties for easily controllable plants. We then show closed-form analytical expressions of best achievable $\mathcal{H}_2$ tracking and regulation performances, and we provide the connection between Condition (π) and the achievable robust performance based on $\mathcal{H}_\infty$ loop shaping design procedure.

Keywords: Feedback control, Control performance limitation, Phase/gain property, $\mathcal{H}_2$ control, $\mathcal{H}_\infty$ loop shaping design, Reciprocal transform.

1 Introduction

More robust and high performance is required for feedback control in a lot of applications of advanced technology. It is necessary to design a good plant as well as an optimal controller, since there exist control performance limitations caused by plant properties such as unstable poles, non-minimum phase zeros, lightly damped modes, and time delays.

This paper focuses on characterization of easily controllable plants in practical control applications rather than to design an optimal or a robust controller for a give plant. In other words, we consider a plant design to guarantee existence of a controller that achieves desirable closed-loop performance. Once we design a plant with such a property, standard optimal and/or robust control methods can be applied to complete the whole design process. The key point is that the control performance is explicitly taken into account in the process of plant design.

The well-known Bode integral type relations tell us that unstable or non-minimum phase plants are not easy to control. However, it is not clear what kind of properties are really required for plants to be controlled in order to achieve the desirable feedback performance under physical constraints such as control effort limit and measurement accuracy. In this paper, we will make a partial answer to the question by summarizing several recent results by the authors, which relate researches on control performance limitations [13, 14].

V.D. Blondel et al. (Eds.) Recent Advances in Learning and Control, LNCIS 371, pp. 111–124, 2008.
springerlink.com

The paper is organized as follows. Section 2 is devoted to motivation and background of the research, which indicates that the notion of minimum phase is not enough for our purpose. Two notions based on the phase/gain property, finite frequency positive realness (FFPR) [7] and Condition (π) [6,8] are introduced in Section 3 for characterizing a set of easily controllable plants in practice. Sections 4 and 5 are devoted to $\mathcal{H}_2$ and $\mathcal{H}_\infty$ control performance limitations. Two types of $\mathcal{H}_2$ tracking and regulation performance limits are provided in Section 4. Section 5 investigates the connection between Condition (π) and the achievable robust performance. Some concluding remarks are made in Section 6.

Notation: $\mathbf{R}$ and $\mathbf{C}$ respectively denote the sets of real and complex numbers, and we define $\mathbf{C}_+ := \{s \in \mathbf{C} \mid \mathrm{Re}(s) > 0\}$ and $\mathbf{C}_- := \{s \in \mathbf{C} \mid \mathrm{Re}(s) < 0\}$.

2 Intrinsic Control Performance Limits

2.1 Bode Integral Relations

Consider the typical SISO (single-input and single-output) unity feedback control system depicted in Fig. 1, where $P(s)$ denotes the continuous-time plant to be controlled and $K(s)$ is the continuous-time controller to be designed. $r(t)$, $d(t)$, $u(t)$, $y(t)$, and $e(t) := r(t) - y(t)$ are the reference command, disturbance input, control input, plant output, and error signal, respectively.

The sensitivity function $S(s)$ and the complementary sensitivity function $T(s)$ respectively defined by

$$S(s) := \frac{1}{1+L(s)}\ , \quad T(s) := \frac{L(s)}{1+L(s)}$$

play an important role for evaluating the control performance, where $L(s) := P(s)K(s)$ is the loop transfer function. The most well-known result on control performance limitation is the so called Bode integral relation on sensitivity gain which is given as follows [3,13] :

Suppose that $L(s)$ is strictly proper, i.e., $S(\infty) = 1$, and that the closed-loop system is stable. Then, we have

$$\frac{1}{\pi}\int_0^\infty \log|S(j\omega)|\,d\omega = \sum_{k=1}^{n_p} p_k^{\mathrm{a}} - \frac{1}{2}\nu_\infty \ ; \ \nu_\infty := \lim_{s\to\infty} s\cdot L(s), \tag{1}$$

where $p_k^{\mathrm{a}} \in \mathbf{C}_+$ $(k = 1, \cdots, n_p)$ are unstable poles of $L(s)$.

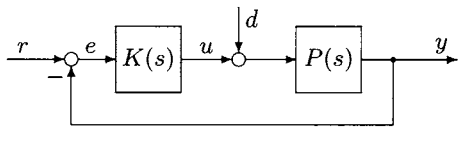


Fig. 1. Unity feedback control system

The counterpart for the complementary sensitivity function $T(s)$ is given as follows [12,13]:

Suppose that $L(s)$ has at least one integrator, i.e., $T(0) = 1$ and that the closed-loop system is stable. Then,

$$\frac{1}{\pi}\int_0^\infty \log|T(j\omega)|\,\frac{d\omega}{\omega^2} = \sum_{k=1}^{n_z}\frac{1}{z_k^{\mathrm{a}}} + \frac{1}{2}\cdot T'(0) \tag{2}$$

holds, where $z_k^{\mathrm{a}} \in \mathbf{C}_+$ $(k = 1, \cdots, n_z)$ are non-minimum phase zeros of $L(s)$.

The two Bode integral type relations above clearly show limitations of achievable control performances caused by unstable poles and/or non-minimum phase (NMP) zeros of the loop transfer function $L(s)$. Since any unstable poles and NMP zeros of the plant can not be cancelled out by the stabilizing controller, they remain in $L(s)$ no mater how we choose a controller which stabilizes the feedback system. Hence, we can conclude that any plant which is either unstable or NMP is not so easy to control.

2.2 Motivating Example: Three-disk Torsion System

One may raise a very simple question related to the conclusion in the previous subsection: *Is any (marginally) stable and minimum phase (MP) plant always easy to control in practice, where we have several physical constraints to be considered in the feedback control implementation?*

In order to answer the question, let us consider a three-disk torsion system, where three disks are connected by wires in vertical and they are actuated by a DC motor directly connected to the bottom disk. That is, the control input u is the voltage applied to the motor, and an encoder is attached at each disk to measure the angular displacements of the bottom, middle and top disks, which are respectively denoted by θ_b, θ_m, and θ_t.

We denote the transfer functions from applied voltage u to the regulated outputs θ_b, θ_m and θ_t by $P_b(s)$, $P_m(s)$ and $P_t(s)$, respectively. They are represented by 6th order models of the forms

$$P_b(s) = \frac{N_b(s)}{D(s)},\quad P_m(s) = \frac{N_m(s)}{D(s)},\quad P_t(s) = \frac{N_t(s)}{D(s)},$$

where

$$\begin{aligned} N_b(s) &= K_b(s^2 + 2\zeta_{b1}\omega_{b1}s + \omega_{b1}^2)(s^2 + 2\zeta_{b2}\omega_{b2}s + \omega_{b2}^2),\\ N_m(s) &= K_m(s^2 + 2\zeta_m\omega_m s + \omega_z^2),\quad N_t(s) = K_t,\\ D(s) &= s(s+c)(s^2 + 2\zeta_1\omega_1 s + \omega_1^2)(s^2 + 2\zeta_2\omega_2 s + \omega_2^2). \end{aligned}$$

Note that all the parameters in the above three transfer functions are positive due to the mechanical structure. For example, the values in our experimental equipment are given by

$$K_b = 368.5957,\ (\omega_{b1}, \zeta_{b1}) = (24.9324, 0.0204),\ (\omega_{b2}, \zeta_{b2}) = (63.4323, 0.0080),$$

$$K_m = 6.2526 \times 10^5,\ (\omega_m, \zeta_m) = (38.3990, 0.01325),\ K_t = 9.219326849 \times 10^8,$$
$$c = 1.9718,\ (\omega_1, \zeta_1) = (36.0887, 0.0279),\ (\omega_2, \zeta_2) = (66.8441, 0.0096).$$

Hence, we can see the following properties:

- All the transfer functions share the common denominator $D(s)$, or they have the same set of poles, which are all in $\mathbf{C}_- \cup \{0\}$.
- All the zeros of $P_b(s)$ and $P_m(s)$ are in $\mathbf{C}_-$ and $P_t(s)$ has no zeros.

These imply that all the transfer functions $P_b(s)$, $P_m(s)$ and $P_t(s)$ are marginally stable and minimum phase, and hence the notion of MP/NMP does not give any help to distinguish the easiness or hardness of control among those three systems.

However, we can naturally guess from physical intuition that the bottom disk is the easiest and the top one is the hardest for control. The guess is actually correct as seen in Fig. 2. It compares the $\mathcal{H}_2$ tracking performance achieved by feedback under control input penalty, where the objective function to be minimized is given by

$$J^* := \inf_{K \in \mathcal{K}_s} \int_0^\infty \left(|e(t)|^2 + W_u^2 |u(t)|^2 \right) dt \tag{3}$$

for unit step reference command $r(t)$ with no disturbance input or $d(t) \equiv 0$ in the feedback control system depicted in Fig. 1 (See Section 4 for the problem setting and the closed-form expression of performance limitation.). The figure clearly shows that the bottom disk provides the best tracking performance and the top disk gives the worst for all weights W_u.

3 Desirable Phase/Gain Properties

The investigation in the previous section concludes the following: While the minimum phase (MP) property is well known to be important for achieving

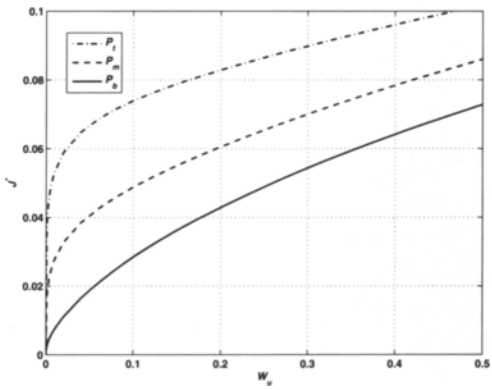


Fig. 2. Optimal tracking performance of three-disk torsion system

good regulation/tracking performance in control community, it is effective for a rather ideal case where the sensor noise is arbitrarily small and there is no limitation on the control effort. This section will provide desirable phase/gain properties of plants to be controlled, which is different from the notion of MP and have a potential of distinguishing the easiness and hardness for control in practice under several physical constraints.

3.1 Finite Frequency Positive Realness (FFPR) and Condition (π)

Two notions based on the phase/gain property, finite frequency positive realness (FFPR) [7] and Condition (π) [6,8], have been proposed for characterizing a set of easily controllable plants in practice. Before introducing them, let us investigate the Bode plots of three transfer functions of the torsion system.

The three phase plots are completely different as seen in Fig. 3. The phases for $P_b(s)$ are greater than -180[degree] for all $\omega > 0$. This is a typical phenomenon for the collocated case. The phases for $P_m(s)$ and $P_t(s)$ are below -180[degree] for all $\omega > \omega_1$, where ω_1 is the frequency of first flexible mode. This implies that the directions of the first modes for $P_m(s)$ and $P_t(s)$ are opposite to that for the rigid mode, while the direction of the first mode for $P_b(s)$ is coincident with that for the rigid mode. The flexible mode for the latter case is referred to as an "in-phase mode," and it is known in the mechanical engineering community that in-phase modes are easy to control [7]. The difference between $P_m(s)$ and $P_t(s)$ is not in the first mode but in the second mode. The second mode for $P_m(s)$ is in-phase, and hence the phases are greater than -360[degree] for all $\omega > 0$. However, the phases for $P_t(s)$ are below -360[degree] for all $\omega > \omega_2$, and hence the second mode for $P_t(s)$ is not in-phase or called "reverse phase."

The notion of in-phase (IP) captures the mode directions, and it has a potential to distinguish the easiness and difficulty of control for flexible mechanical

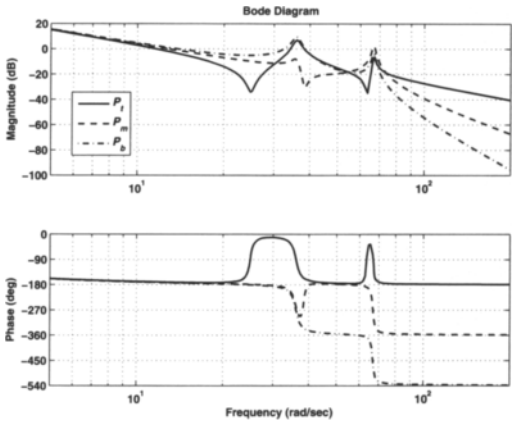


Fig. 3. Bode plots for torsion system

structures. However, the systematic treatment of the IP property is not so easy, because the mode decomposition is required.

The notion of "finite frequency positive realness (FFPR)" was introduced in [7] to overcome the drawbacks and the following knowledge has been obtained for the class of plants

$$\mathcal{P}_{ms} := \{P(s) \mid P(s) \text{ is a strictly proper transfer function with all the poles in the closed complex left half plane}\}.$$

FFPR [7]: Let $P(s) \in \mathcal{P}_{ms}$ be the transfer function of the plant to be controlled. Then the maximum control bandwidth ω_{b*} achievable by a dynamic feedback controller is approximately the same as the maximum frequency ϖ that satisfies

$$G(j\omega) + G^*(j\omega) \geq 0 \quad (\forall\ |\omega| \leq \varpi)\ ;\ G(s) := sP(s).$$

The knowledge quoted above implies that there is some connection between the phase delay of -180[degree] and the achievable control bandwidth. However, the further theoretical investigation is not possible, because the higher oscillated modes may affect the achievable control performance. The following condition defining a phase/gain property was introduced to improve the FFPR property, where a small gain requirement is added to it [6].

Condition (π) [6]: Either one of the following conditions holds for $P(s) \in \mathcal{P}_{ms}$:
(1) $\forall \omega \in \mathbb{R}$, $\angle P(j\omega) \geq -\pi$
(2) $\exists\ \omega_\pi > 0$ such that

$$\begin{aligned} 0 \leq \omega < \omega_\pi &: \angle P(j\omega) \geq -\pi, \\ \omega = \omega_\pi &: \angle P(j\omega) = -\pi,\ |P(j\omega)| \leq 1, \\ \omega > \omega_\pi &: |P(j\omega)| < 1. \end{aligned}$$

That is, ω_π is the frequency at which the phase of the plant reaches -180[degree] and from which the gain of the plant is always less than 0[dB]. In other words, a good phase property in the low frequency range and a small gain in the high frequency range are both required.

3.2 Numerical Example: A Flexible System

In order to confirm the effectiveness of two properties, namely FFPR and Condition (π), let us consider the simplest flexible plant expressed as

$$P(s) = \frac{1}{s^2} + \frac{k}{s^2 + 2\zeta s + 1}\ ;\ 0 < \zeta < 1. \tag{4}$$

Since the transfer function can be rewritten as

$$P(s) = \frac{(1+k)s^2 + 2\zeta s + 1}{s^2 + 2\zeta s + 1}, \tag{5}$$

we can see the following facts: MP for $k > -1$ and In-Phase for $k > 0$. We can also see that the phases for the reverse phase plant, i.e., $k < 0$, are below -180[degree] for all ω, while the phases for the in-phase plant, i.e., $k > 0$, are greater than -180[degree] for all ω. In other words, Condition (π) holds if $k \geq 0$ and it does not hold if $k < 0$, and the critical plant is the double integrator, i.e., the plant with $k = 0$.

We now compare the optimal tracking performances J^* with $W_u = 1$ in (3) for $k \in (-1.5, 1.5)$. The solid and dashed plots in Fig. 4 indicate the values of J^* for $\zeta = 0.01$ and $\zeta = 0.1$, respectively. We can see from the plots (especially from the solid plot) that a significant difference appears at $k = 0$ rather than at $k = -1$.

This implies that the notion of IM is much more suited than that of MP for characterizing a set of easily controllable plants in practice and that Condition (π) is one of possible candidates for the measure of easiness of control. The theoretical justifications will be done in Section 5, where investigating the connection between Condition (π) and the best achievable robust control performance based on $\mathcal{H}_\infty$ loop shaping design proposed by McFarlane and Glover [11].

4 $\mathcal{H}_2$ Control Performance Limitations

4.1 Problem Setting

This section investigates the $\mathcal{H}_2$ control performance limitations in a more general setting than that used in Section 2. In other words, we here consider the optimal tracking problem with frequency weighted control input penalty and the optimal regulation problem with frequency weighted output penalty for the SISO unity feedback control system depicted in Fig. 1.

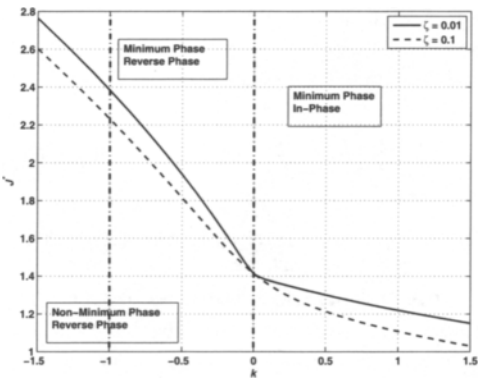


Fig. 4. J^* for flexible plant

We assume for the tracking problem that the reference command $r(t)$ is the unit step function and there is no disturbance input, i.e., $d(t) \equiv 0$. The performance index to be minimized is given by

$$J^*(P, W_u) := \inf_{K \in \mathcal{K}_s} \int_0^\infty \left(|e(t)|^2 + |u_w(t)|^2 \right) \, dt, \tag{6}$$

where $u_w(t)$ is the weighted control input with weight $W_u(s)$, and $\mathcal{K}_s$ denotes the set of stabilizing controllers. On the other hand, we assume that the disturbance input is impulse input, i.e., $d(t) = \delta(t)$, and there is no reference command, i.e., $r(t) \equiv 0$, in the optimal regulation problem. The performance index to be minimized is given by

$$E^*(P, W_y) := \inf_{K \in \mathcal{K}_s} \int_0^\infty \left(|u(t)|^2 + |y_w(t)|^2 \right) \, dt, \tag{7}$$

where $y_w(t)$ denotes the weighted plant output with weight $W_y(s)$.

4.2 Analytical Closed-form Expressions

The $\mathcal{H}_2$ optimal control problems defined in the previous subsection are quite standard, and hence we can get the optimal costs as well as the optimal controllers by solving either the corresponding Riccati equations or the LMI problems. However, those numerical methods do not give any insights on easiness of the plant for control. The following analytical expressions of J^* for marginally stable plants and of E^* for minimum phase plants derived in [2] are very helpful to understand what kind of properties easily controllable plants should have, because the forms are represented in terms of only plant properties such as unstable poles, non-minimum phase zeros, and plant gain.

Theorem 1

$$J^*(P, W_u) = 2 \sum_{k=1}^{n_z} \frac{1}{z_k^{\mathrm{a}}} + \frac{1}{\pi} \int_0^\infty \log \left(1 + \frac{|W_u(j\omega)|^2}{|P(j\omega)|^2} \right) \frac{d\omega}{\omega^2} \, , \tag{8}$$

where $z_k^{\mathrm{a}} \in \mathbf{C}_+ \ (k = 1, \cdots, n_z)$ *are non-minimum phase zeros of* $P(s)$.

$$E^*(P, W_y) = 2 \sum_{k=1}^{n_p} p_k^{\mathrm{a}} + \frac{1}{\pi} \int_0^\infty \log \left(1 + |W_y(j\omega) P(j\omega)|^2 \right) d\omega \, , \tag{9}$$

where $p_k^{\mathrm{a}} \in \mathbf{C}_+ \ (k = 1, \cdots, n_p)$ *are unstable poles of* $P(s)$.

The expression for J^* implies that the NMP zeros of the plant make the tracking performance worse. Especially, the closer the NMP zero is to the origin, the worse the achievable performance becomes. Moreover, we can see from the second term that the smaller plant gain in the lower frequency range gives the better performance. Hence, the performance limit is affected by MP zeros of the plant as

well as the NMP zeros as seen in the example of torsion system in Subsection 2.2. The expression for E^* is similar to that for J^* but different, and we note the following relations:

$$1/z_k^{\mathrm{a}},\ 1/|P(j\omega)|\ \text{ in }\ J^* \iff p_k^{\mathrm{a}},\ |P(j\omega)|\ \text{ in }\ E^*\ .$$

This motivates us to introduce a new transform for continuous-time systems named "*Reciprocal Transform*" proposed in [4].

The class of continuous-time systems considered is given by

$$\mathbf{\Sigma} := \{\ G(s) \in \mathbf{R}_p^{m\times m} \mid |G(0)| \neq 0\ \}, \tag{10}$$

and the definition of *Reciprocal Transform* $\mathcal{R}(\cdot)$ is as follows:

Definition 1. *For $G(s) \in \mathbf{\Sigma}$, we define the* Reciprocal Transform $\mathcal{R}(\cdot)$ *by*

$$\mathcal{R}(G(s)) := G^{-1}(1/s). \tag{11}$$

The reciprocal transform consists of the following two actions.

- We first consider the inverse system (change of the role of input and output of the system), and
- replace the frequency variable s by s^{-1} (change of the role of low-frequency range and high-frequency range).

We have the following theorem on $\mathcal{H}_2$ control performance limits.

Theorem 2. *Suppose $W(s) \in \mathbf{\Sigma}$ is an outer function. Then, we have*

(i) For any $P(s) \in \mathbf{\Sigma}$ which has at least one integrator,

$$J^*(P,W) = E^*(\mathcal{R}(P), \{\mathcal{R}(W)\}^{-1}). \tag{12}$$

(ii) For any $P(s) \in \mathbf{\Sigma}$ which is strictly proper,

$$E^*(P,W) = J^*(\mathcal{R}(P), \{\mathcal{R}(W)\}^{-1}). \tag{13}$$

The theorem is quite useful for deriving J^* from E^* and vice versa.

4.3 Sum of Roots Characterization

This subsection provides different expressions for J^* and E^* which are given in terms of *two sums of roots* (SoRs) [10], where we respectively assume $W_u(s) = 1$ and $W_y(s) = 1$ for simplicity.

Write the SISO strictly proper plant $P(s)$ with order n as $P(s) = \frac{P_N(s)}{P_D(s)}$, where P_N and P_D are coprime polynomials and P_D is assumed to be monic without loss of generality. Let $M_D(s)$ be the spectral factor of

$$P_N(s)P_N(-s) + P_D(s)P_D(-s)\ \ \Big(= M_D(s)M_D(-s)\Big)\ . \tag{14}$$

It should be noticed that M_D is the characteristic polynomial of the closed-loop system constructed with the optimal controller for both problems.

We now define the roots of $P_N(s)$, $P_D(s)$, and $M_D(s)$ as follows.

$$\begin{aligned} z_\ell \ (\ell = 1 \sim m) &: \text{ roots of } P_N(s) \ \text{ (zeros of } P(s)), \\ p_j \ (j = 1 \sim n) &: \text{ roots of } P_D(s) \ \text{ (poles of } P(s)), \\ \alpha_i \ (i = 1 \sim n) &: \text{ roots of } M_D(s) \ \text{ (closed-loop poles).} \end{aligned}$$

Then, we have the following theorem, which provides SoR characterizations for best achievable $\mathcal{H}_2$ tracking and regulation performances.

Theorem 3

$$J^*(P,1) = \sum_{i=1}^{n}\left(-\frac{1}{\alpha_i}\right) - \sum_{\ell=1}^{m}\left(-\frac{1}{z_\ell}\right), \tag{15}$$

$$E^*(P,1) = \sum_{i=1}^{n}\left(-\alpha_i\right) - \sum_{j=1}^{n}\left(-p_j\right). \tag{16}$$

Note that the summations in the second terms in Theorem 3 count all the poles or zeros of the plant rather than the unstable poles or NMP zeros as in the expressions in the previous subsection.

The above theorem indicates an intriguing fact:

The best achievable regulation performance level is the difference between the degrees of "*average stability*" of the plant to be controlled and of the closed-loop system constructed with the optimal controller. Roughly speaking, the more unstable the plant is, the worse performance one can get.

Contrary to the $\mathcal{H}_2$ regulation case the best $\mathcal{H}_2$ tracking performance level is expressed in terms of *time constants*, or more specifically, as the difference between the "average response speed" of the closed-loop system and that of the plant *zeros*. Under the assumption that the achieved closed-loop poles are the same, it is observed that having zeros of the plant far away from the origin does not affect the performance level whether they are minimum phase or non-minimum phase. As the time constant of a minimum phase zero increases, the plant bandwidth widens and the controller can 'see' the plant more. As a result a better performance can be accomplished. Conversely a non-minimum phase zero near the origin brings about an inverse response and degrades the achievable performance.

5 $\mathcal{H}_\infty$ Performance Limitations

This section is devoted to one of typical $\mathcal{H}_\infty$ control problems, namely the $\mathcal{H}_\infty$ loop shaping design procedure for connecting Condition (π) and the best achievable robust control performance.

5.1 $\mathcal{H}_\infty$ Loop Shaping Design

The $\mathcal{H}_\infty$ loop shaping design procedure is known to be one of nice ways to link robust $\mathcal{H}_\infty$design to the classical design, and it is defined as follows ([11]).

Consider the feedback control system depicted in Fig. 5, where P_0 denotes the plant to be controlled and W is an appropriate weighting function. The problem is to find a controller K which stabilizes the modified plant $P := P_0 W$ and minimizes the $\mathcal{H}_\infty$ norm of

$$\Phi := \begin{bmatrix} I \\ K \end{bmatrix} (I + PK)^{-1} \begin{bmatrix} I & P \end{bmatrix}. \tag{17}$$

In other words, the problem to be solved is defined as

$$\gamma_{\text{opt}}(P) := \inf_{K \in \mathcal{K}_s} \|\Phi\|_\infty, \tag{18}$$

and the implemented controller is WK instead of K. The performance index to be minimized is the $\mathcal{H}_\infty$ norm of a 2×2 transfer matrix which includes both the sensitivity and complementary sensitivity functions, and hence we can get a stabilizing controller which makes a good balance between sensitivity reduction and robustness improvement. It is known that, no iteration procedure is required to derive the best achievable cost γ_{opt}, and hence we may obtain an algebraic expression of γ_{opt}.

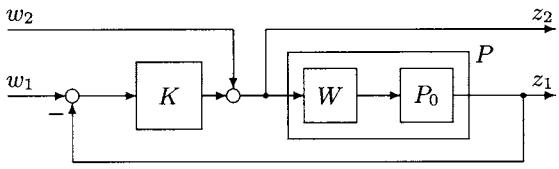


Fig. 5. $\mathcal{H}_\infty$ loop shaping design

The index γ_{opt} indicates in a certain sense the robustness of the obtained closed-loop system (the smaller, the better), because the performance index includes both the sensitivity and complementary sensitivity functions. In addition, for SISO systems, lower bounds for the phase margin of the designed closed-loop system can be expressed by PM $\geq 2 \arcsin(1/\gamma_{\text{opt}})$ [15].

Furthermore, it is known that the controller synthesized by $\mathcal{H}_\infty$ loop shaping design improves the robustness without changing the crossover frequency of P drastically if γ_{opt} is approximately less than $2\sqrt{2}$ [11]. Therefore, γ_{opt} indicates how 'good' the closed-loop system can be if the crossover frequency of P is chosen to be the control bandwidth. A simple example is the double integrator $P(s) = 1/s^2$, and the achievable performance γ_{opt} is given by

$$\gamma_{\text{opt}} = \gamma_{\text{DI}} :=\leq \sqrt{4 + 2\sqrt{2}} < 2\sqrt{2} \tag{19}$$

and the resulting PM is 45[degree].

We can also readily see that the $\mathcal{H}_\infty$ loop shaping design problem is self reciprocal, or we have the following theorem.

Theorem 4. *For any* $P(s) \in \mathbf{\Sigma}$, *we have*

$$\gamma_{\rm opt}(P(s)) = \gamma_{\rm opt}(\mathcal{R}(P(s))). \tag{20}$$

5.2 2nd Order Systems

We here focus on the 2nd order case. All the 2nd order continuous-time systems belonging to $\mathcal{P}_{ms}$ can be parametrized as follows:

$$\mathcal{P}_2 := \left\{ P(s) = k\frac{\beta s + 1}{s^2 + \alpha_1 s + \alpha_0} \;\middle|\; \alpha_0 \geq 0,\ \alpha_1 \geq 0,\ k > 0 \right\} \subset \mathcal{P}_{ms}\ . \tag{21}$$

Depending on the sign of β, $P(s) \in \mathcal{P}_2$ exhibits a distinct property; P is minimum phase if $\beta \geq 0$, while P is non-minimum phase if $\beta < 0$. Hence, we can define two sets:

$$\text{Minimum Phase (MP):} \quad \mathcal{P}_{2+} := \left\{ P(s) \in \mathcal{P}_2 \;\middle|\; \beta \geq 0 \right\},$$
$$\text{Non-Minimum Phase (NMP):} \quad \mathcal{P}_{2-} := \left\{ P(s) \in \mathcal{P}_2 \;\middle|\; \beta < 0 \right\}.$$

It is trivial that all systems in $\mathcal{P}_{2+}$ belong to $\mathcal{P}^{\rm g}_{ms}$ (the set of $P(s) \in \mathcal{P}_{ms}$ satisfying Condition (π)). On the other hand, the phase delay of any system $P(s)$ in $\mathcal{P}_{2-}$ exceeds 180[degree], and we have

$$\mathcal{P}^{\rm g}_{2-} := \mathcal{P}_{2-} \cap \mathcal{P}^{\rm g}_{ms} = \left\{ P(s) \in \mathcal{P}_{2-} \;\middle|\; 0 < k \leq \frac{\alpha_1}{-\beta} \right\}. \tag{22}$$

Then the following theorem holds [6].

Theorem 5 ([6]). *For any system* $P(s)$ *in*

$$\mathcal{P}^{\rm g}_2 := \mathcal{P}_2 \cap \mathcal{P}^{\rm g}_{ms} = \mathcal{P}_{2+} \cup \mathcal{P}^{\rm g}_{2-}\ , \tag{23}$$

it holds that $\gamma_{\rm opt} \leq \gamma_{\rm DI}$. *The equality holds when* $\alpha_0 = \alpha_1 = \beta = 0$ *for* $P(s) \in \mathcal{P}_{2+}$ *and* $\alpha_0 = 0$, $k = \alpha_1/(-\beta)$ *for* $P(s) \in \mathcal{P}^{\rm g}_{2-}$.

For $P(s) \in \mathcal{P}_{2-}$, as k increases, the control bandwidth gets wider, but the phase delay exceeds 180[degree]. Under this situation the control performance measured by $\gamma_{\rm opt}$ becomes worse without bound. In other words the bandwidth that can be achievable in practice has some limitation, which also agrees with the knowledge implied by FFPR.

5.3 Higher Order Systems

It is not so easy to derive theoretical results for general higher order systems, although it is shown that the relation between Condition (π) and $\gamma_{\rm DI}$ still holds for some classes of 3rd order systems in [6,8]. Here, we will show it is also valid for a 4th order system investigated in Subsection 3.2. The optimal costs $\gamma_{\rm opt}$ for $k = -0.5$, $k = 0$, and $k = 0.5$ are plotted in Fig. 6. The two cases with $\zeta = 0.01$ and $\zeta = 0.1$ are plotted for $k = -0.5$ and $k = 0.5$, where the upper plots correspond to the case of $\zeta = 0.01$. We can see from this figure with further computations that the following facts are true for all $\zeta > 0$: $\gamma_{\rm opt} > \gamma_{\rm DI}$ $(^\forall k < 0)$, and $\gamma_{\rm opt} < \gamma_{\rm DI}$ $(^\forall k > 0)$. This clearly shows the effectiveness of Condition (π) for characterizing a set of easily controllable plants.

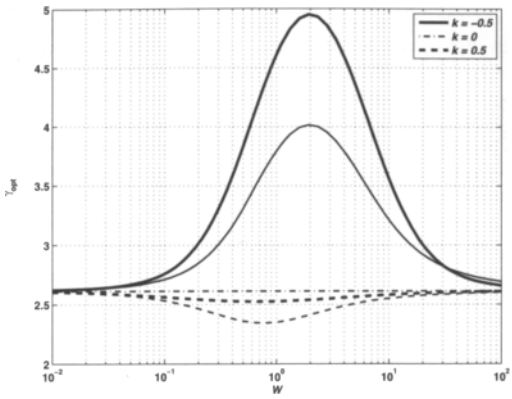


Fig. 6. Achievable robust performance γ_{opt}

6 Conclusion

This paper has been focused on characterization of easily controllable plants in practical control applications rather than to design an optimal or a robust controller for a give plant. After explaining the background and the motivation of the research topic, we first provided two notions, namely finite frequency positive realness (FFPR) and Condition (π), which represent desirable phase/gain properties for easily controllable plants.

We then showed closed-form analytical expressions for best achievable $\mathcal{H}_2$ tracking and regulation performances. They capture the performance limits caused by several plant properties. The more general results including discrete-time case are found in [5, 1]. Another type of expressions in terms of sum of roots have been also proposed. They are not directly linked to the plant property, but a kind of parametric design is possible based on the formula using recently developed tools in computer algebra [9].

Finally, we have investigated the connection between Condition (π) and the achievable robust performance based on $\mathcal{H}_\infty$ loop shaping design. The results confirm the effectiveness of Condition (π) for distinguishing good and bad plants and bring us a new robust control design, which is a nice combination of the classical design concept and modern systematic procedure [8].

References

1. Bakhtiar, T., Hara, S.: H_2 Regulation Performance Limitations for SIMO Linear Time-invariant Feedback Control Systems. Automatica (to appear, 2007)
2. Chen, J., Hara, S., Chen, G.: Best tracking and regulation performance under control energy constraint. IEEE Trans. on Automatic Control 48(8), 1320–1336 (2003)
3. Freudenberg, J.S., Looze, D.P.: Right half plane zeros and poles and design tradeoffs in feedback systems. IEEE Trans. on Automatic Control 30(6), 555–565 (1985)

4. Hara, S.: A unification of analytical expressions for control performance limitations via reciprocal transform. In: SSSC 2007. 3rd IFAC Symp. on System, Structure and Control (submitted, 2007)
5. Hara, S., et al.: The Best Achievable H_2 Tracking Performances for SIMO Feedback Control Systems. J. of Control Science and Engineering (2007)
6. Hara, S., Kanno, M., Onishi, M.: Finite frequency phase property versus achievable control performance in $\mathcal{H}_\infty$ loop shaping design. In: Proc. of SICE-ICASE Int. Joint Conf. 2006, Busan, Korea, pp. 3196–3199 (October 2006)
7. Iwasaki, T., Hara, S., Yamauchi, K.: Dynamical system design from a control perspective: Finite frequency positive-realness approach. IEEE Trans. on Automatic Control 48(8), 1337–1354 (2003)
8. Kanno, M., Hara, S., Onishi, M.: Characterization of Easily Controllable Plants Based on the Finite Frequency Phase/Gain Property. In: Proc. of American Control Conference 2007, New York, pp. 5816–5821 (July 2007)
9. Kanno, M., et al.: Parametric optimization in control using the sum of roots for parametric polynomial spectral factorization. In: Proc. Int. Symposium on Symbolic and Algebraic Computation, Waterloo, Ontario, Canada, pp. 211–218 (July-August 2007)
10. Kanno, M., Hara, S., Anai, H., Yokoyama, K.: Sum of Roots, Polynomial Spectral Factorization, and Control Performance Limitations. In: IEEE Conf. on Decision and Control, New Orleans (to be presented, December 2007)
11. McFarlane, D.C., Glover, K.: Robust Controller Design Using Normalized Coprime Factor Plant Descriptions. Lecture Notes in Control and Inf. Sciences, vol. 138. Springer, Heidelberg (1990)
12. Middleton, R.H.: Trade-offs in linear control systems design. Automatica 27(2), 281–292 (1991)
13. Seron, M.M., Braslavsky, J.H., Goodwin, G.C.: Fundamental Limitations in Filtering and Control. Springer, London (1997)
14. Skogestad, S., Postlethwaite, I.: Multivariable Feedback Control: Analysis and Design, 2nd edn. Wiley, Chichester (2005)
15. Vinnicombe, G.: Uncertainty and Feedback — $\mathcal{H}_\infty$ Loop-shaping and the ν-gap Metric. Imperial College Press, London (2001)

Metrics and Morphing of Power Spectra

Xianhua Jiang, Shahrouz Takyar, and Tryphon T. Georgiou

Dept. of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455
{jiang082,shahrouz,tryphon}@umn.edu

Summary. Spectral analysis of time-series has been an important tool of science for a very long time. Indeed, periodicities of celestial events and of weather phenomena sparked the curiosity and imagination of early thinkers in the history of science. More recently, the refined mathematical techniques for spectral analysis of the past fifty years form the basis for a wide range of technological developments from medical imaging to communications. However, in spite of the centrality of a spectral representation of time-series, no universal agreement exists on what is a suitable metric between such representations. In this paper we discuss three alternative metrics along with their application in morphing speech signals. Morphing can be naturally effected via a deformation of power spectra along geodesics of the corresponding geometry. The acoustic effect of morphing between two speakers is documented at a website.

1 Introduction

The latter half of the 20th century witnessed the fast development of the mathematics of signals and systems, driven by technological needs for modeling, identification, and control. Early thinkers attempted to quantify in various ways the amount of information supplied by data and how to quantify uncertainty. In particular, in statistics, we encounter the Fisher information metric and Kullback-Leibler distance on probability density functions with natural interpretations based on a Bayesian viewpoint. Then in control theory, we encounter the graph topology and the gap/graph metrics which quantify distance between dynamical systems in a way suitable for ensuring robustness of stability. The fundamental contributions of M. Vidyasagar in this latter field, in the 80's, underscored the significance of searching for the "correct" topology when studying properties of signals and systems. These contributions have educated a generation of control theorists and have influenced our current program.

Leaping across to signal processing, in spite of great strides over the past half a century, there is no universal agreement on how to naturally measure distance between time-series. Yet, the importance of such natural metrics in quantifying resolution and in developing techniques for robust identification is self-evident. Thus, the purpose of this paper is to present three alternative metrics between power spectral densities which indeed have a natural interpretation as distances

V.D. Blondel et al. (Eds.) Recent Advances in Learning and Control, LNCIS 371, pp. 125–135, 2008.
springerlink.com

between time-series, and then, to discuss their experimental relevance in morphing of speech. The metrics induce natural geodesics in the space of power spectral densities, and in our example of speech-morphing, it is reasonable to follow such geodesic paths linking the spectral content of the voices of two speakers. In this printed document, experimental results are illustrated using figures while their acoustic counterparts are documented at a website.

2 The $\mathcal{L}_1$-distance between Power Spectra

Consider that $\{u_{f_i}(k)\}$ with $i \in \{0,1\}$ and $k \in \mathbb{Z}$, represent two (i.e, for $i \in \{0,1\}$) discrete-time, stationary, zero-mean, scalar stochastic processes whose spectral content is characterized by the two power spectral density functions f_0 and f_1. We postulate that there exist processes ψ_0 and ψ_1 so that

$$u_{f_0}(k) + \psi_0(k) = u_{f_1}(k) + \psi_1(k),$$

and seek such a perturbation of minimal total combined variance

$$E\{\psi_0(k)^2\} + E\{\psi_1(k)^2\}$$

which is sufficient to "reconcile" the readings of our original two processes $\{u_{f_i}(k)\}$. The combined variance represents the minimal amount of "energy" needed to render the two indistinguishable, and qualifies as a reasonable distance between the two.

Given f_0, f_1, the optimal choice consists of processes such that ψ_0 and ψ_1 are independent, $u_{f_i}(k)$ and ψ_i are also independent, and the spectral densities of the perturbations are

$$f_{\psi_0}(\theta) = \begin{cases} f_1(\theta) - f_0(\theta) & \text{if } f_1(\theta) - f_0(\theta) \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

$$f_{\psi_1}(\theta) = \begin{cases} f_0(\theta) - f_1(\theta) & \text{if } f_1(\theta) - f_0(\theta) \leq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Then, the distance between the two is

$$\begin{aligned} d(f_0, f_1) &:= E\{\psi_0(k)^2\} + E\{\psi_1(k)^2\} \\ &= \int_{-\pi}^{\pi} (f_{\psi_0}(\theta) + f_{\psi_1}(\theta)) d\theta \\ &= \int_{-\pi}^{\pi} |f_0(\theta) - f_1(\theta)| \, d\theta \\ &= \|f_0 - f_1\|_1. \end{aligned}$$

It is easy to see that a geodesic path between f_0 and f_1 is simply the linear family of spectral densities

$$f_\tau(\theta) = (1-\tau) f_0(\theta) + \tau f_1(\theta) \tag{1}$$

for $\tau \in [0,1]$. Also, the total geodesic distance between the two is equal to $d(f_0, f_1)$, hence this metric is intrinsic.

3 The *Prediction Metric*

Following [4], we motivate and present a metric between power spectral density functions which, in a certain precise sense, quantifies consistency in prediction. We will refer to this as the *prediction metric* (for lack of a better name).

Our starting point is a minimum-variance prediction problem. We first quantify the degradation in the variance of the prediction error when we design a predictor using the wrong choice among two alternatives. More precisely, assume that $\{u_{f_i}(k)\}$ with $i \in \{0,1\}$ and $k \in \mathbb{Z}$, are discrete-time stochastic processes as before, whose spectral content is characterized by the two power spectral density functions f_0 and f_1. We denote by $\mathcal{E}\{\ \}$ the expectation operator, and by

$$\mathbf{p}_i := (p_{f_i}(1), p_{f_i}(2), p_{f_i}(3), \dots)$$

the values for the coefficients that minimize the linear one-step-ahead prediction error variance

$$J_i(\mathbf{p}) := \mathcal{E}\{|u_{f_i}(0) - \sum_{\ell=1}^{\infty} p(\ell) u_{f_i}(-\ell)|^2\},$$

i.e., $\mathbf{p}_i = \text{argmin}(J_i(\mathbf{p}))$. A quantitative assessment of the difference between f_0 and f_1 can be based on the degradation of the error variance when the predictor is designed based on one of the two, say f_1, and then used to predict a process having power spectral density f_0. In particular, we compute the ratio of the "degraded" predictive error variance over the optimal error variance

$$\rho(f_0, f_1) := \frac{J_0(\mathbf{p}_1)}{J_0(\mathbf{p}_0)} = \frac{\mathcal{E}\{u_{f_0}(0) - \sum_{\ell=1}^{\infty} p_{f_1}(\ell) u_{f_0}(-\ell)|^2\}}{\mathcal{E}\{u_{f_0}(0) - \sum_{\ell=1}^{\infty} p_{f_0}(\ell) u_{f_0}(-\ell)|^2\}}$$

as a function of the two densities f_0 and f_1. This turns out to be equal to the ratio of the arithmetic over the geometric means of the fraction f_0/f_1, i.e.,

$$\rho(f_0, f_1) = \left(\int_{-\pi}^{\pi} \left(\frac{f_0(\theta)}{f_1(\theta)}\right) \frac{d\theta}{2\pi}\right) / \exp\left(\int_{-\pi}^{\pi} \log\left(\frac{f_0(\theta)}{f_1(\theta)}\right) \frac{d\theta}{2\pi}\right).$$

Since the degradation

$$\rho(f_0, f_1) - 1 \geq 0,$$

with equality when the fraction f_0/f_1 is constant, it can be used as a measure of dissimilarity between the shapes of the two spectral density functions. Then again, the quadratic term in Δ in the expansion of

$$\rho(f, f + \Delta) - 1,$$

defines a quadratic form on positive density functions; this can be readily computed as (see [4] for technical details)

$$g_f(\Delta) := \int_{-\pi}^{\pi} \left(\frac{\Delta(\theta)}{f(\theta)}\right)^2 \frac{d\theta}{2\pi} - \left(\int_{-\pi}^{\pi} \frac{\Delta(\theta)}{f(\theta)} \frac{d\theta}{2\pi}\right)^2.$$

A geodesic path f_τ ($\tau \in [0,1]$) connecting two spectral densities f_0, f_1, i.e., a path having minimal length

$$\int_0^1 \sqrt{g_{f_\tau}(\frac{\partial f_\tau}{\partial \tau})}d\tau,$$

turns out to be the exponential familiy

$$f_\tau(\theta) = f_0^{1-\tau}(\theta) f_1^\tau(\theta), \tag{2}$$

for $\tau \in [0,1]$. The geodesic distance is once more equal to the distance

$$d_{\text{prediction}}(f_0, f_1) := \sqrt{\int_{-\pi}^{\pi} \left(\log \frac{f_0(\theta)}{f_1(\theta)}\right)^2 \frac{d\theta}{2\pi} - \left(\int_{-\pi}^{\pi} \log \frac{f_0(\theta)}{f_1(\theta)} \frac{d\theta}{2\pi}\right)^2}$$

between the two densities, i.e., the metric $d_{\text{prediction}}(\cdot,\cdot)$ is intrinsic just as the $\mathcal{L}_1$-distance $d(\cdot,\cdot)$ that we discussed earlier.

4 Transportation-related Metrics

Morphing and interpolation of distributions can be quite naturally dealt with as a mass transport problem. The original optimal transport problem was introduced by Gaspard Monge two centuries ago. The so-called Monge-Kantorovich theory that has since emerged is a very active area of research with a wide range of applications ranging from economics to theoretical physics and image processing [11, 2, 6]. In the present section we outline only certain basic facts.

Consider the Wasserstein distance of order 2

$$W_2(f_0, f_1) := \inf\{\sqrt{\int_{-\pi}^{\pi} |\theta - s(\theta)|^2 f_0(\theta) d\theta} \\ \text{for } s \,:\, |s'(\theta)| \cdot f_1(s(\theta)) = f_0(\theta)\}$$

between density functions f_0 and f_1. For our purposes the density functions are normalized to have equal mass, e.g., $\int_{-\pi}^{\pi} f_0(\theta)d\theta = \int_{-\pi}^{\pi} f_1(\theta)d\theta = 1$. Alternatively, one may consider a penalty to weigh in the difference in mass between the two (see e.g., [6]). The integrand $|\theta - s(\theta)|^2$ represents the cost of transferring the mass element $f_0(\theta)d\theta$ to a new location $s(\theta)$. The (nonlinear) constraint that the first distribution is carried onto the second is expressed by the requirement that the transport map ψ satisfies

$$|s'(\theta)| \cdot f_1(s(\theta)) = f_0(\theta)$$

or, equivalently,

$$\int_{-\pi}^{\theta} f_0(\sigma)d\sigma = \int_{-\pi}^{s(\theta)} f_1(\sigma)d\sigma. \tag{3a}$$

The latter allows determining $s(\theta)$ by integrating f_0, f_1 and comparing the respective values of their integrals (cf. [6, section 3.2]). The scaling function $s(\theta)$ is monotonically increasing (when f_0, f_1 are positive), i.e., there are no "folds" in transfering f_0 to f_1.

The relevant theory extends to transport of probability measures on $\mathbb{R}^n$ and even on Riemannian manifolds [2, 11]. The transport map s is characterized by Brenier and McCann's theorems [11, pages 66 and 96]. In general the Wasserstein distance defines a metric on the space of (normalized) density functions where the geodesic distance between two densities in fact equals the actual distance between the two. Geodesics f_τ ($\tau \in [0,1]$) between the extreme "points" f_0 and f_1 are determined by a gradient flow [11, page 252], which, in this very special one-dimensional case, specifies that the geodesic via

$$((1-\tau) + \tau s'(\theta))\, f_\tau((1-\tau)\theta + \tau s(\theta)) = f_0(\theta), \tag{3b}$$

for $\tau \in [0,1]$, and where $s(\theta)$ was computed to satisfy (3a) for the two "end points" f_0 and f_1. Interestingly, for the special case of density functions in one dimension, as in the present paper, the transport map s is not only the unique optimal transport map for the Wasserstein distance of order 2, but also for all other Wasserstein distances of order $p > 1$ (see [2, Section 3]) and hence, it is optimal for $p = 1$ as well (though not unique).

5 Morphing of Speech

Modeling and synthesizing speech is a well studied subject driven by the availability of cheap, high speed processors and by a wide range of applications; these include wireless communications, voice mail, data compression, speech recognition, speaker identification, text-to-speech translation, altering or synthesizing voice with particular characteristics as well as a multitude of applications in the entertainment industry [3]. The purpose of the present work is to consider and study geodesic paths between power spectral density functions as a means to morph the voice of one individual to the voice of another. Despite great advances in the theory and technology of speech processing, speech morphing is still in an early phase and far from being a standard application [8]. We first provide a rudimentary exposition of certain standard techniques and practices in modeling speech and then we discuss our experimental results.

Speech sounds are produced by acoustic excitation of cavities—the vocal and nasal tracts, manipulated by the position and stiffness of the tongue, lips, jaw, etc. The excitation is produced by the vocal chords in the larynx, or by turbulence at other parts of the vocal system (as with fricatives **f**, **sh** or affricates **j**, etc., where the tongue and lips create suitable constrictions that generate turbulence). Typically, the shape of the vocal tract results in certain resonances called formants. These dominate the character of voiced sounds. However, coupling of the oral and nasal tracts, as in nasal sounds, causes the power spectra to have deep "valleys" as well. This is also the case when the excitation originates in the interior of the vocal tract, instead of the glottis. Typical models for speech

suggest $\sim$ 25 [ms] intervals of "quasi-stationarity". Each $\sim$ 25 [ms] fragment is represented by a linear filter driven by an excitation signal. Typically, the latter is either a periodic excitation, e.g., a train of pulses, which creates a pitch, or white noise. Voiced sounds typically require the filter to be autoregressive (AR) since the formant-structure is mainly responsible for their character. However, in general, pole-zero models are appropriate. Thus, common practice for speech analysis and synthesis is mainly based on either construction of detailed models or on linear time-frequency techniques which identify/generate the spectral content over sufficiently small time-intervals.

Speech morphing is the process of transforming one person's speech pattern (e.g., Alice's) into another's (Bob's), gradually, creating a new pattern with a distinct identity, while preserving the speech-like quality and content of the spoken sentence. In practice, there are several technical issues that need to be dealt with depending on the application. A typical experiment requires first that A (i.e., Alice) and B (i.e., Bob) recite the same sentence. These sentences need to be segmented appropriately and an exact correspondence be drawn between the beginning and end of the various quasi-stationary intervals that correspond to similar sounds. Time-frequency analysis can be used to automate marking of such intervals. Then, a suitable algorithm is required to interpolate the spectral qualities of the sounds produced by A and B.

Earlier attempts can be roughly classified in two groups, those who use direct time-frequency models and those who use linear predictive and other nonlinear methods for identifying suitable models. For instance, Abe [1] modified spectra by stitching together the low frequency part of Alice's spectrum, below a pre-specified frequency, with the part of Bob's power spectrum above the given frequency. This pre-specified frequency is then used as the control parameter that regulates the mixing. Ye and Young [14] proposed a "perceptually weighted" linear transformation method based on certain a Gaussian mixture model for the power spectra of the two subjects. Kawahara and Matsui [7] manually selected anchor points to define correspondence between two time-frequency representations for Alice and Bob. Based on these anchor points, a piecewise bilinear transformation was used to map the target time-frequency coordinate onto the reference coordinate. Others focus on the modification of the excitation signal and AR spectral parameter; Goncharoff and Kaine-Krolak [5] interpolated the AR-spectra by pole shifting. He first applied a perceptually-based pole pairing algorithm, then generated a path between the pole pairs to achieve linear changes in pole's frequency and bandwidth. Pfitzinger [8] on the other hand applied a dynamic programming technique to align the residuals and power spectra of AR models of the respective speech signals.

The approach taken in our work shares a number of basic steps with those earlier attempts, such as marking and processing separately quasi-stationary segments over 25 [ms] time intervals, etc. The key difference is in the algorithm for interpolating the resulting power spectra—in our work we suggest that this can be most conveniently, and perhaps naturally, effected by following geodesics in suitable metrics (such as those discussed in Sections 2-4). Below, we briefly

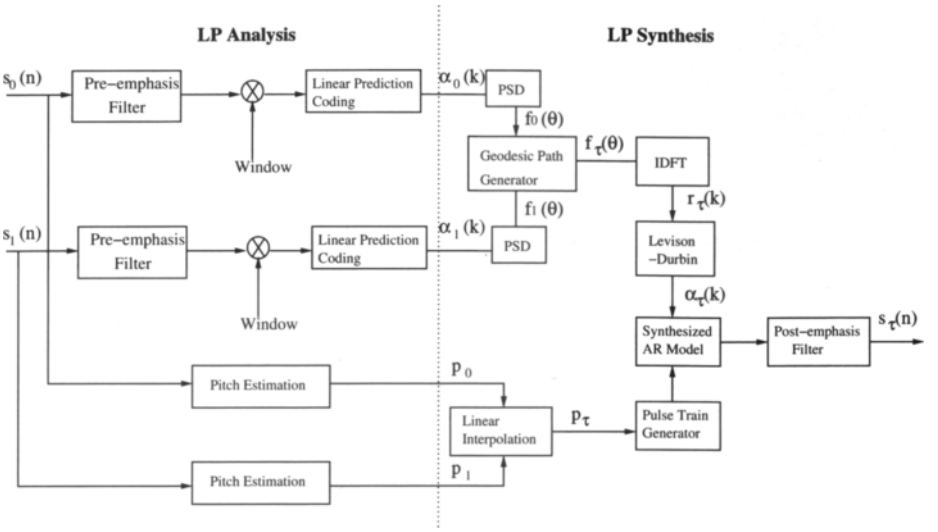


Fig. 1. Linear Prediction Analysis and Synthesis Framework

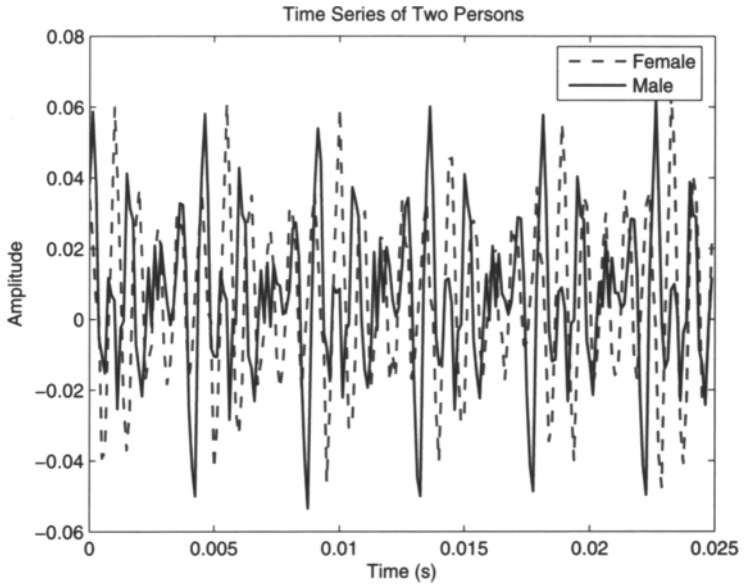


Fig. 2. Time signals corresponding to the phoneme "a" for A (female) and B (male)

discuss the steps taken for analyzing a voiced phoneme for A and B, generating the respective spectra, and then generating the morphed sound. Examples of geodesics of complete words and sentences are posted at [12] in an audible format.

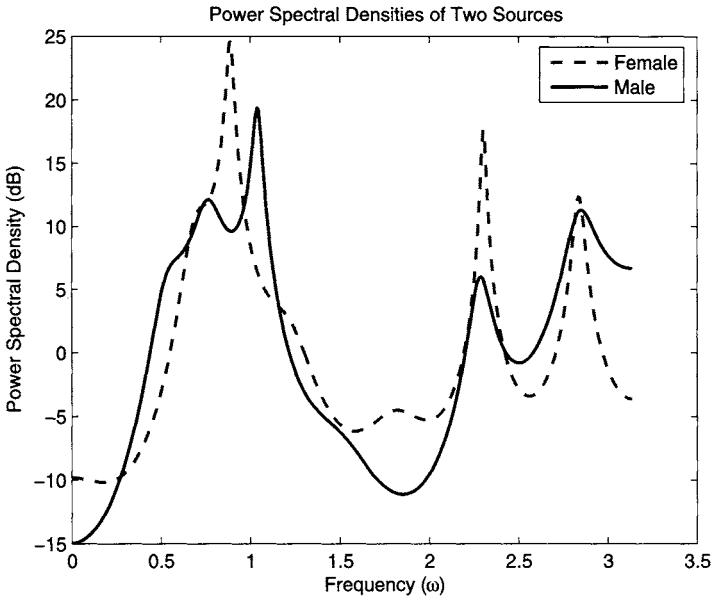


Fig. 3. Power spectra for subjects A (female) and B (male)

For analyzing a voiced phoneme we use standard linear prediction techniques [10]. The basic framework is shown in Figure 1. Linear prediction analysis is followed by generation of a power spectrum at $\tau \in [0,1]$ on the geodesic path, and then followed by a synthesis stage. Linear prediction is used to obtain the coefficients of a modeling filter as well as to estimate the pitch. The voiced sound that is being analyzed (an "a" in the present case) is shown as two superimposed time-waveforms in Figure 2 (one for speaker A and one for B); it was sampled at 8KHz. For modeling, we choose an AR model of order 14. The frame size is 25 [ms], and the frame interval is 12 [ms],. A standard "pre-emphasis" filter is used to reduce the low-frequency content in the signal. The filtered data is weighted using a Hamming window. For modeling, we chose to use standard methods; we use the autocorrelation method for estimating the covariance lags and then the Levinson-Durbin method for obtaining the coefficients of the AR model. The AR model of the phoneme, for each of the two speaker, provides a corresponding power spectral density. A power spectrum at a point $\tau \in [0,1]$ on the geodesic path is then determined for each of the three metrics of Sections 2-4. We estimate the pitch period for speakers A and B, using either residual-based estimation or Harmonic sinusoidal wave-based estimation. These periods are linearly interpolated along the path. The synthesized power spectral densities are approximated as AR-spectra and, finally, a pulse train with the interpolated pitch period is used to drive the synthesized AR-filter (at $\tau \in [0,1]$ on the geodesic) in order to produce the synthesized sound. A post-emphasis filter is used to compensate the effect of the pre-emphasis filter–a standard practice.

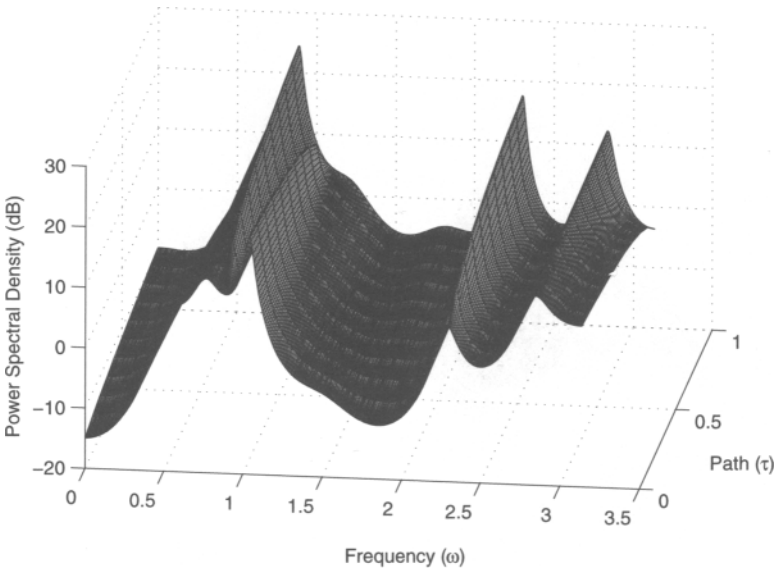


Fig. 4. Geodesic path between the two spectra following (1)

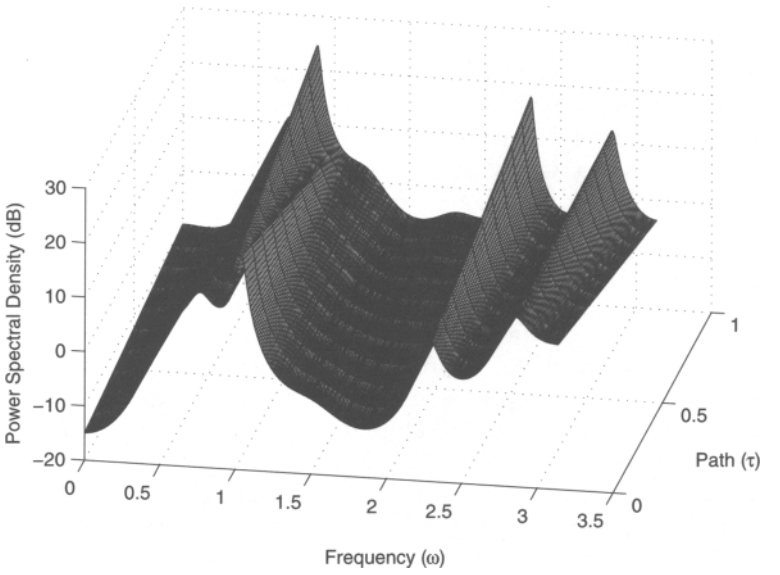


Fig. 5. Geodesic path between the two spectra following (2)

In synthesizing complete sentences, a number of issues are critical but relatively standard. For instance, dynamic time warping is needed for an automatic alignment between two persons' speech, which may have different duration for

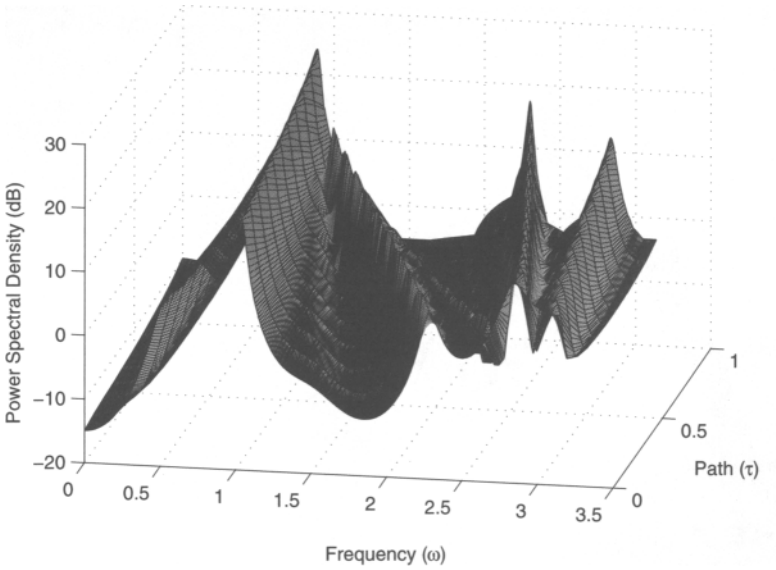


Fig. 6. Geodesic path between the two spectra following (3)

the same phoneme and detecting voiced and unvoiced part within a sentence. More complicated methods have been suggested to further improve the voice quality and some of these could be incorporated in the above framework. For example, Wong *etal.* [13] suggest that the AR coefficients be estimated when the glottis is closed. This requires further processing and, for the high-pitch speakers, it becomes challenging. Others suggest a glottal-flow-like excitation [9], such as in the Liljencrants-Fant (LF) model, instead of simple pulse trains.

6 Concluding Remarks

Examples of geodesics for complete words and sentences are posted at [12] in an audible format. The quality of sound produced by the three alternative geodesics (1), (2), and (3) appears to be ordered in reverse, with geodesics (3) and (2) performing best. Geodesic (3), shown in Figure 6, retains the formant structure of the underlying spectra, though the effect on the acoustic quality is very subjective. Interestingly, geodesic (2) has surprisingly good acoustic qualities, in spite of the fact that visually, in Figures 4 and 5, there is an apparent "fade in" and "fade out" of the respective formants in the two power spectra.

Acknowledgements

The authors are grateful to Professor Allen Tannenbaum for his input.

The research has been supported in part by the NSF, AFOSR, and the Vincentinne Hermes-Luh endowment.

References

1. Abe, M.: Speech morphing by gradually changing spectrum parameter and fundamental frequency. In: Proc. ICSLP 1996, vol. 4, pp. 2235–2238 (1996)
2. Ambrosio, L.: Lectures Notes on Optimal Transport Problems, CVGMT (July 2000) (preprint)
3. Childers, D.G.: Speech Processing and Synthesis Toolboxes. Wiley, Chichester (2000)
4. Georgiou, T.T.: Distances and Riemannian metrics for spectral density functions. IEEE Trans. on Signal Processing 55(8), 3995–4003 (2007)
5. Goncharoff, V., Kaine-Krolak, M.: Interpolation of LPC spectra via pole shifting. In: Proc. ICASSP 1995, vol. 1, pp. 780–783 (1995)
6. Haker, S., Zhu, L., Tannenbaum, A., Angenent, S.: Optimal mass transport for registration and warping. International Journal on Computer Vision 60(3), 225–240 (2004)
7. Kawahara, H., Matsui, H.: Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. In: Proc. ICASSP 2003, vol. 1, pp. 256–259 (2003)
8. Pfitzinger, H.R.: Unsupervised Speech Morphing between Utterances of any Speakers. In: Proceedings of the 10th Australian Intern. Conf. on Speech Science and Technology, Sydney, pp. 545–550 (2004)
9. Sambur, M.R., Rosenberg, A.E., Rabiner, L.R., McGonegal, C.A.: On reducing the buzz in LPC synthesis. J. Acoust. Soc. Am. 63(3), 918–924 (1978)
10. Stoica, P., Moses, R.: Introduction to Spectral Analysis. Prentice Hall, Englewood Cliffs (2005)
11. Villani, C.: Topics in Optimal Transportation. In: GSM, vol. 58, AMS (2003)
12. `http://www.ece.umn.edu/~georgiou`
13. Wong, D.Y., Markel, J.D., Gray, A.H.: Least Squares Glottal Inverse Filtering from the Acoustic Speech Waveform. IEEE Trans. on Acoustics, Speech, and Signal Processing ASSP-27(4), 350–355 (1979)
14. Ye, H., Young, S.: Perceptually Weighted Linear Transformations for Voice Conversion. Eurospeech (2003)

A New Type of Neural Computation

Hidenori Kimura and Shingo Shimoda

Bio-Mimetic Control Research Center, RIKEN
2271-130, Anagahora, Shimoshidami, Moriyama-ku, Nagoya 463-0003, Japan
kimura@bmc.riken.jp, shimoda@bmc.riken.jp

Summary. A new type of neural computation scheme is proposed, in which the numerical values are represented as the number of firing neurons of neural clusters. The scheme is quite classical in the sense that it is essentially based on McCulloch-Pitts model and Hebbian rule, but it is drastically new in the sense that it is free from any symbolism. The computational capability of this neural shceme is extensively discussed.

Keywords: Neural computation, McCulloch-Pitts model, Hebbian rule, non-symbolic computation, input replacement.

1 Introduction

Computational neuroscience dates back to the celebrated paper by McCulloch and Pitts [1] who modeled the neuron mathematically as a hybrid elements with analog inputs and digital output. They showed that the weighted connection of these neurons was capable of generating any function that can be generated by a universal Turing Machine. In some sense, McCulloch-Pitts model of brain is a connectionism equivalent of Turing Machine. There have been proposed many kinds of models of neurons and its network, but the McCulloch-Pitts model continued to provide them with the basic framework [2].

In addition to logical capability, plasticity is another important characteristic of brain that represents learning, memory and adaptation which are crucial factors of the life. The issue was first formulated in terms of neuron behaviors by Hebb who reduced the plasticity of brain functions to the change of synaptic connectivities [3]. This idea which is referred to as *Hebbian rule*, and has been the leading paradigm of the neural basis of learning, as well as for various schemes of learning machines [2] [4]. Artificial neural network, based on these two biologically inspired ideas of McCulloch-Pitts and Hebb, opened up a new field of computational neuroscience and gave a great impact to modern technology, as well as to cognitive science.

The main target of artificial neural network is to construct a functional map between inputs and outputs through its samples of input/output pairs. In its most abstract form, artificial neural network no longer has any direct relevance to *real brain*, as its name "artificial" demonstrates. However, computational neuroscience which tries to capture the algorithmic features of brain mathematically

V.D. Blondel et al. (Eds.) Recent Advances in Learning and Control, LNCIS 371, pp. 137–147, 2008.
springerlink.com

relies essentially on the paradigm of learning theory based on artificial neural network. It is not unusual that experimental data obtained through various imaging devices are called upon to validate a learning paradigm of artificial neural network [5].

Observed data of brain activities through imaging devices reflect the number of firing neurons in a certain cortical area of brain. On the other hand, artificial neural network usually represents numerical values as a weighted sum of firing signals.

In other words, imaging devices assume the number of firing elemnts as indicators of neural activities, while artificial neural network represents its output as synaptic modifications of neuron groups. Although the brain way of representing neural information is still in the mist, it is worth considering the possibility of neural computation scheme, in which each numerical value is represented as a number of firing neurons, rather than weighted sum of each neuron.

Representing numerical values as a number of firing neurons is similar to counting numbers by fingers, and its the basis of fundamental arithmetic. It is free from any type of symbolisms which only human being has acquired at the last stage of evolution. We notice that most *natural* computations carried out by living organisms are essentially symbol-free. In metabolic control in the cellular level, all the numerical aspects of control are ascribed to the concentrations of chemical agents, which share many common features with the number of firing neurons in neural computation network.

This paper reports our attempt to construct a basic sheme of *non-symbolic* neural computaitons. Non-symbolic neural computation is strongly connected to the notion of *compound control* which we have proposed as a basic mode of biological control [6] [7]. We propose this neural computation scheme as a common link connecting each level of biological control raging from cellular to behavioral control.

This control scheme was born as a tool of tacit learning and has been applied to various robot controls including biped walking [8].

2 Extended McCulloch-Pitts Model and Hebbian Rule

Our model of neuron is the same as the one McCulloch and Pitts introduced sixty years ago [1] which is mathematically described as

$$y(t) = \mathbf{1}\Big(\sum_{i=1}^{N} w_i y_i(t) - \theta(t)\Big), \tag{1}$$

where $\mathbf{1}(x)$ is a discrete sigmoid function characterized as

$$\mathbf{1}(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{2}$$

and $y_i(t)$ is the output of a presynaptic neuron connected to the neuron with connecting weight w_i.

In our model, the threshold, denoted by θ, is assumed to be adjusted in a way that it reaches to a certain steady level. It is assumed to be increased after the neuron fires in order to prevent excessive firing, while it decreases after some rest (non-firing) state to encourage next firing. In the discrete time frame, it is mathematically described as

$$\theta(t+1) - \theta(t) = \overline{\Delta\theta} y(t) + \underline{\Delta\theta}(y(t) - 1), \tag{3}$$

where $\overline{\Delta\theta}$ and $\underline{\Delta\theta}$ denote the incremental and decremental magnitudes, respectively. We assume both $\overline{\Delta\theta}$ and $\underline{\Delta\theta}$ are positive.

In order to represent the plasticity of neural network, the celebrated Hebbian rule is extended to include the *mediator* neuron. The behavior of connecting weight w_{ij} which connects the neuron with output $y_i(t)$ and the other with output $y_j(t)$ is described as

$$w_{ij}(t+1) - w_{ij}(t) = \overline{\Delta w} y_i(t) y_j(t) y_k(t) + \underline{\Delta w}(y_i(t) y_j(t) y_k(t) - 1), \tag{4}$$

where $\overline{\Delta w}$ and $\underline{\Delta w}$ denote the incremental and decremental magnitudes, respectively. Here, we introduced the mediator neuron with output $y_k(t)$. If the mediator always fires, i.e., $y_k(t) = 1$ for each t, it can be excluded from (4) and the equation represents the classical Hebbian rule. If both $\overline{\Delta w}$ and $\underline{\Delta w}$ are positive, the rule (4) represents the *facillitation*, while if they are negative, it represents the *inhibition*. We graphically represent the above rule as illustrated in Figure 1.

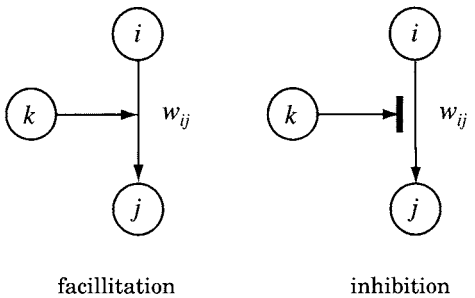


Fig. 1. Extended Hebbian Rule

3 Neuron Cluster and Representation of Numerical Values

The most salient feature of our neural computation scheme lies in its way of representing numerical values associated with neural activities. We form a group of neurons, called a *cluster*, to represent a component of the neural state. The number of firing neurons in a cluster is taken as the quantitative indicator of the state. Since the number of firing elements is always a positive integer, we must

approximate a positive number x to be the largest integer which is not beyond x and is normally designated by $[x]$. The negative number x is treated similarly as $-[-x]$.

An obvious identity

$$[x] = \sum_{j=1}^{N} \mathbf{1}(x - j) \tag{5}$$

for positive number x, where N is a sufficiently large number beyond x, we can devise a way of encoding x to $[x]$ using our neurons as shown in Figure 2, where the threshold of the i-th neuron is taken to be i. The number of firing neurons is equal to $[x]$. We call the cluster with thresholds being arranged in a natural order as

$$\theta_j = j \tag{6}$$

a *natural cluster*. To deal with negative values, we use the same network with weights -1 instead of $+1$ and join the two clusters.

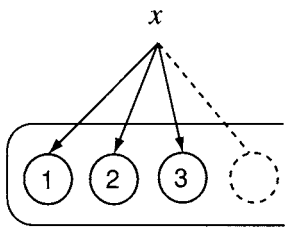


Fig. 2. Neumerical Coder using natural cluster

4 Connections of Clusters for Elementary Arithmetic Operations

Figures 3-7 show the cluster connections to perform addition, subtraction, multiplication and division of two given numbers. Addition is performed by perfectly connecting the two input clusters to a coder with unity weight, as is shown in Figure 3. Figure 4 illustrates a simplified version of Figure 3, in which the number of connections are reduced from $2N^2$ to $3N$, with N being the number of neurons in one cluster. The subtraction is simply performed by changing the weights connecting y cluster to the output cluster from 1 to -1, shown in Figure 5. This scheme actually computes $\max(x - y, 0)$. To implement the real subtraction $x - y$, we must prepare a similar connection of clusters with signs of weights reversed to compute $\max(y - x, 0)$ and subtract from the former to the latter to produce $x - y = \max(x - y, 0) - \max(y - x, 0)$. Multiplication shown in Figure 6 is a bit complicated. We must prepare N^2 neurons for the output cluster z, which are suffixed in the way $z_{11}, z_{12}, \cdots, z_{1N}, z_{21}, z_{22}, \cdots, z_{2N}, \cdots, z_{NN}$. The neuron designated by z_{ij} is connected to x_i, the i-th neuron of x cluster, and y_j, the j-th neuron of y cluster with unity weight. All the thresholds of z cluster

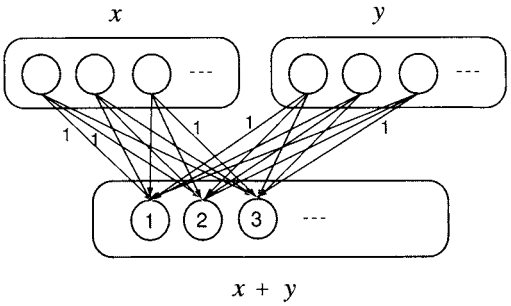


Fig. 3. Addition

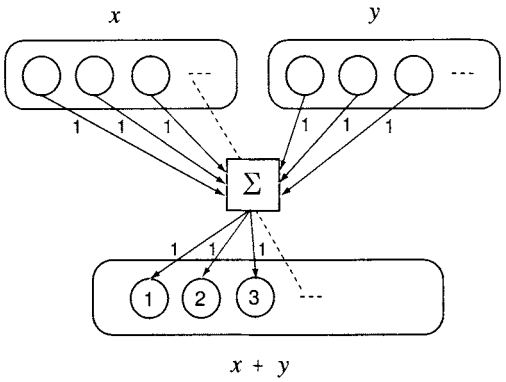


Fig. 4. Simplification of Fig.3

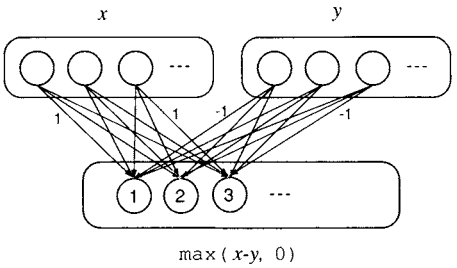


Fig. 5. Subtraction

are chosen to be equal to 2. To see the scheme of Figure 6 can perform multiplication, consider the case where $x = 2$ and $y = 3$, i.e., the first two neurons of the x-cluster and the first three neurons of the y-cluster fire. All the other neurons are in the rest. Then, the inputs to $z_{11}, z_{12}, z_{13}, z_{21}, z_{22}, z_{23}$ are equal to 2 and the inputs to all the other neurons in z cluster are less than 2. Hence, $6 = 2 \times 3$ neurons in z cluster fire. The division shown in Figure 7 is relatively

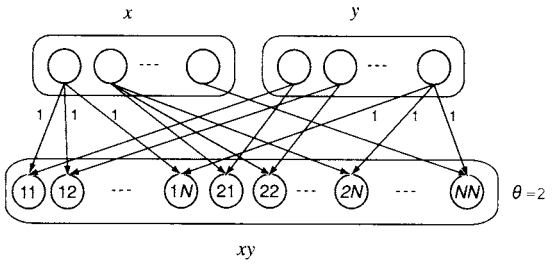


Fig. 6. Multiplication

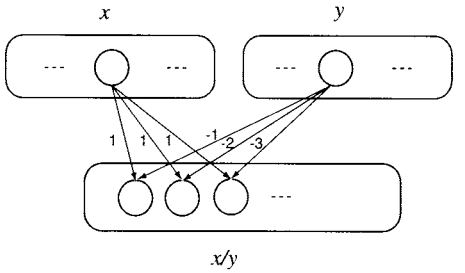


Fig. 7. Subtraction

simple. Both of the input clusters are connected perfectly to the output cluster. The connecting weights between x cluster and the output cluster are all unity, while the connecting weights between a neuron of y cluster and the j-th neuron of output cluster is chosen to be $-j$. Therefore, the number of firing neurons in the output cluster is equal to $\max(j : x - jy \geq 0)$, which is equal to $[x/y]$. Obvious modifications are required to deal with negative values.

5 Generation of Arbitrary Functions

An arbitrary function $y = f(x)$ with $f(0) = 0$ can be generated by connecting two clusters x and y. Actually, we can construct its approximation $y = [f([x])]$ by choosing the weight w_{ij} connecting the i-th neuron of the x cluster and j-th neuron of the y cluster as

$$w_{ij} = f(i) - f(i-1), \quad j = 1, 2, \cdots \tag{7}$$

Then, the input to each neuron of y cluster becomes equal to

$$y = \sum_{i=1}^{[x]} \big(f(i) - f(i-1)\big) = f([x])$$

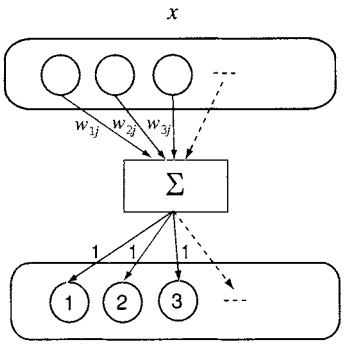


Fig. 8. Function Generator

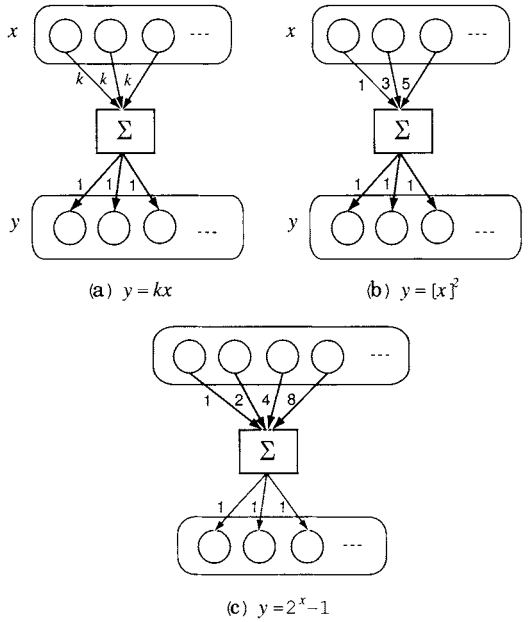


Fig. 9. Computational Configuration of Various Functions

By choosing the thresholds of each neuron to be in a natural order (6) the number of firing elements is equal to $[f([x])]$. Figure 8 illustrates a simplified implementation of the network (7). In Figure 9, some typical functions are illustrated.

6 Dynamic Characteristics

This section shows the capability of our neural computing scheme concerning dynamical characteristics which the extended Hebbian rule is responsible for.

6.1 Lateral Inhibition and Its Release

A frequently observed neural module is the so-called lateral inhibition. This is a characteristics of a parallel neural pathway in which the firing of one string of neurons inhibits the firing of other neural path. This module is frequently found in visual cortex possibly for the purpose of enhancing object edges [9]. It is also supposed to be used for rhythm generation [10].

Figure 10 shows a simplest assembly of lateral inhibition. In Figure 10(a), we take $w_e < -w_d$ and the threshold θ_E of E is taken to be zero. Then, as I' is applied and the neuron A fires, the weight w_a increases and finally the neuron C fires. This prevents E to fire. Thus, the path $B \rightarrow D \rightarrow E$ connecting input I to output O is inhibited by firing of neuron A.

If the neuron A inhibits the weight w_a, as in Figure 10(b), then the firing of C is prevented, which results in the release of inhibition.

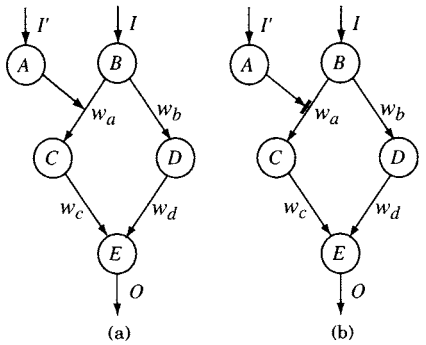


Fig. 10. Lateral Inhibition and Its Release

6.2 Input Replacement

If an input to a neural pathway is associated with another input for some time interval, the associated input can generate the output which was generated by the original input. This is frequently seen in neural network and is regarded as a fundamental building block of learning, or brain plasticity. The conditional reflex which was initially investigated by Pavlov [11] is a typical macroscopic representation of imput replacement. Actually, Hebb proposed his plasticity rule motivated greately by Pavlov's work. The adaptation and learning are believed to be reduced to this neural process.

Figure 11 illustrates a simple neural configuration of input replacement. It is the same as the release of lateral inhibition except that a new edge is added connecting neurons A and E. If the neurons A and B fire simultaneously for sometime, only the firing A become capable of the firing E. This is the simplest configuration composed of simple neurons. We can construct a more complex input replacement with complex output pattern.

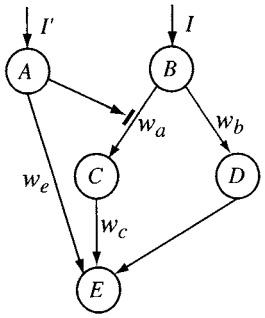


Fig. 11. Input Replacement

6.3 Integration

The integration of the input can be implemented by the scheme of Figure 12. The input I_1 can be integrated in the output 0. Here, the thresholds of the neurons in cluster E is taken to be zero.

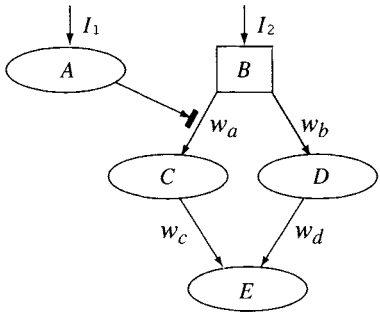


Fig. 12. Integration

The idea is as follows. The weights w_a and w_b are chosen to be equal and the weight w_b is chosen to be $w_b < -w_c$ so that no neuron in the cluster E fires even if I_2 is added. This is a sort of lateral inhibition. If, at time $t = 1$, I_1 is applied to the A cluster which fires $I_1(1)$ neurons, then the number of firing neurons of C cluster is decreased by $I_1(1)$. Hence, $I_1(1)$ neurons are released from being inhibited and $I_1(1)$ neurons of the E cluster fire. At time $t = 2$, $I_1(2)$ neurons of A cluster fire. Then, due to the extended Hebbian rule, the weight w_a is decreased by $I_1(2)$ that connects the A cluster to the previously fired neurons. Then, at $t = 2$, *additional* $I_1(2)$ neurons are inhibited so that the total $I_1(1) + I_1(2)$ neurons in the C cluster inhibited. Thus, the number of firing neurons of E cluster becomes equal to $I_1(1) + I_1(2)$. In this way, the number of firing neurons of the E-cluster at time t is equal to the sum of the input $\sum_{j=1}^{t} I_1(j)$.

The above reasoning applies only to positive inputs. To deal with negative inputs, we must duplicate the scheme and operate the two simultaneously.

6.4 Use of the Scheme as a Controller

Our scheme can be connected to the outer world to carry out tasks as a controller. Figure 13 illustrates a use of our scheme as a proportional feedback controller. Figure 14 illustrates an example of our scheme to learning control of simple robot arm. Since our scheme can perform all the elementary arithmetic, we can devise any type of controller using our scheme.

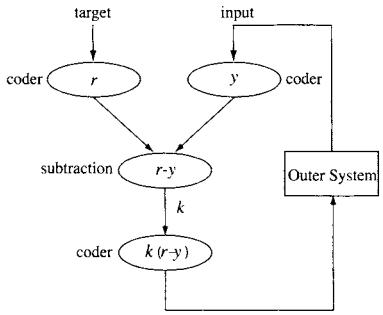


Fig. 13. Proportional Controller

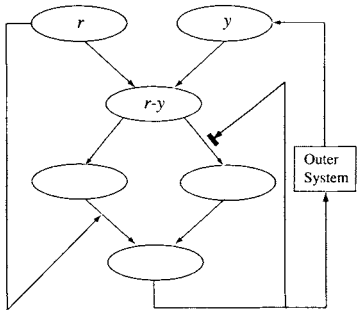


Fig. 14. An Adaptive Controller

7 Conclusion

The neural computational scheme presented in this paper certainly needs a huge number of neurons to perform even an elementary computation. It is also able to compute only an approximation of the value to be computed, because it deals only with integers. We need to increase the number of neurons in each cluster in order to perform accurate computations.

In spite of these uneconomical features of our computational scheme, we believe it important to propose a non-symbolic neural computation scheme. First, biological way of computation has evolved from non-symbolic way of simple organism. The brain, which is a final product of the long history of the evolution of

life, must reflect some non-symbolic way of computation. Second, since computational behaviors of neurons are locally controlled and driven by their rules for surviving, it is hard to imagine that the numerical values are symbolically coded in their local activities. Third, various imaging devices to observe and record the brain activities through the number of firing neurons. In order to make theory consistent with experimental data, we need to develop a framework of computation in which activities are represented in terms of numbers of firing neurons. These rationale seem to be enough to justify the importance of non-symbolic neural computation scheme.

This scheme is now being implemented in a walking robot which exhibits surprizing robustness against various changes of walking surfaces.

References

1. McCulloch, W.S., Pitts, W.H.: A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biol. 5, 115–133 (1943)
2. Rosenblatt, F.: Principles of Neurodynamics; Perceptron and the theory of Brain Mechanisms, Spartan Books (1962)
3. Hebb, D.O.: The Organization of Behavior: A Neuropsychological Theory (1949)
4. Amari, S.: Theory of adaptive pattern classifiers. IEEE Trans. EC-16, 299–307 (1967)
5. Doya, K., et al. (eds.): Bayesian Brain: Probabilistic Aproaches to Neural Coding (2007)
6. Tanaka, R.J., et al.: Mathematical description of gene regulatory units. Biophysical J. 91, 1235–1247 (2006)
7. Kimura, H., et al.: Stochastic approach to molecular interactions and computational theory of metabolic and genetic regulations. J. Theor. Biol. (in press, 2007)
8. Shimoda, S., et al.: A neural computation scheme of compound control with application to biped locomotion (submitted for publication)
9. Hubel, D.H., Wiesel, T.N.: Brain mechanism of vision, Scientific America (September 1979)
10. Matsuoka, K., et al.: Mechanisms of frequency and pattern control in the neural rhythm generators. Biolg. Cybern. 56, 345–353 (1987)
11. Pavlov, I.P.: Conditioned Reflex: An Investigation of Physiological Activities of the Cerebral Cortex, Translated by Arep, G.V., Oxford Univ. Press, Oxford (1927)

Getting Mobile Autonomous Robots to Form a Prescribed Geometric Arrangement

Laura Krick, Mireille Broucke, and Bruce Francis

Department of Electrical and Computer Engineering, University of Toronto, 10 King's College Road, Toronto, Ontario, Canada M5S 3G4
lkrick@control.utoronto.ca, broucke@control.utoronto.ca, bruce.francis@utoronto.ca

Summary. Mobile autonomous robots present an interesting example within the subject of distributed control systems. There are several motor-driven wheeled robots that are autonomous in that their actions are subject only to sensor inputs and pre-loaded programs; there is no leader and no supervisor. The problem is to design the onboard controllers so that the robots perform a useful cooperative task. For example, suppose the robots all have antennas, forming an antenna array, and the collective task is to shape and point the radiated beam of the array. This requires the robots to form a certain geometric pattern. Previous work in this area has focussed mainly on the rendezvous problem, where the desired task is to meet at a common location (without navigational instruments). In this paper the task is to form a prescribed geometric arrangement, such as a regular polygon. The arrangement is defined by a rigid graph with link lengths. Nonlinear control laws are developed and the closed-loop equilibria are studied in detail.

Keywords: Autonomous robots, formation control.

1 Introduction

Distributed control refers to systems of modules or agents that are interconnected dynamically or have a common objective and where control is local, with the possible exception of high-level intermittent centralized supervision. Undoubtedly these kinds of systems will become more and more prevalent as embedded hardware evolves. An interesting example and area of ongoing research is the control of a group of autonomous mobile robots, ideally without centralized control or a global coordinate system, so that they work cooperatively to accomplish a common goal. The aims of such research are to achieve systems that are scalable, modular, and robust. These goals are similar to those of sensor networks—networks of inexpensive devices with computing, communications, and sensing capabilities. Such devices are currently commercially available and include products like the Intel Mote. A natural extension of sensor networks would be to add simple actuators to the sensors to make them mobile, and then to adapt the network configuration to optimize network coverage.

If global coordinates are known and there is an omniscient supervisor, these problems are routine: Each robot could be stabilized to its assigned location.

V.D. Blondel et al. (Eds.) Recent Advances in Learning and Control, LNCIS 371, pp. 149–159, 2008.
springerlink.com

The current technology to provide global coordinates is the Global Positioning System (GPS). However, the use of GPS for position information in multi-agent applications has several problems: The precision of GPS depends on the number of satellites currently visible; during periods of dense cloud cover, in urban areas, and under dense vegetation, there may be no line of sight between the receiver and the GPS satellite. These problems in obtaining global coordinates make it natural to study decentralized control.

The simplest problem is stabilizing the robots to a common location, frequently called the *rendezvous problem*. Many different techniques have been used to solve this problem, for example, cyclic pursuit [6] and the circumcentre law [1]. The solution in [5] involves asynchronous stop-and-go cycles. Reference [3] considers a problem of decentralized, self organizing communication nodes with mobility. In this case, the mobility is used to improve communication performance. Another possible goal for mobile sensor networks is to optimize sensor placement by minimizing a cost function. This type of problem is studied in [2].

An interesting approach to formation control is that of Olfati-Saber [8]. The robots are point masses (double integrators) with limited vision, and he proposes using rigid graph theory to define the formation; he also proposes a gradient control law involving prescribed distances. The limitation is that the network is not homogeneous—special so-called γ-agents are required to achieve flocking.

Finally, reference [9] considers the problem of achieving polygon formations without global coordinates, but the solution is complete only for three robots.

The starting point for our paper is [8] but with the following differences: Our robots are simpler, kinematic points (integrators), and the visibility graph is fixed; on the other hand, there is no need for special agents.

Proofs are largely omitted due to page limitations; complete proofs are available in the MASc thesis of the first author [4] (available by email request). Moreover, the main goal of the paper is to be of tutorial value, so much of the discussion relates to the example of six robots in a regular polygon formation.

2 The Problem

Consider n robots moving in the plane, $\mathbb{R}^2$. They are modeled by the simplest possible kinematic equation, namely $\dot{z}_i = u_i$, where $z_i(t)$ is the position of robot i and $u_i(t)$ is its velocity input. The goal is to have the points $z_1(t), \ldots, z_n(t)$ converge to form a stable equilibrium formation. This requirement could be described in general terms, but we prefer a specific example in order to derive explicit formulas. Thus, the desired formation is taken to be an ordered regular polygon with vertices in the order $z_1, z_2, \ldots, z_n$ and with sides a prescribed length d. Figure 1 shows six robots in a regular polygon. We call such an arrangement a d-polygon.

Additionally, we assume the robots have only onboard sensors. They can sense relative positions of other robots but not absolute positions. With this setup, we have the following problem statement: Given a distance d, design control laws so that d-polygons are stable equilibrium formations.

Our solution begins with rigid graph theory, which is now briefly reviewed.

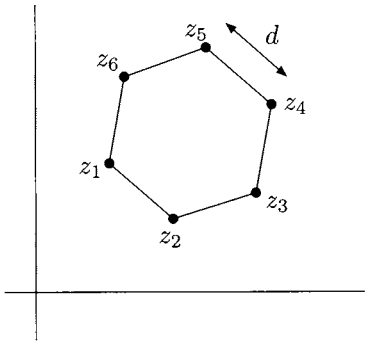


Fig. 1. Six robots in a regular polygon formation

3 Frameworks

A framework is an embedding of an undirected graph into the plane, as follows. Consider n points $p_1, \ldots, p_n$ in $\mathbb{R}^2$. If we think of these points as the nodes of a graph, we can introduce some edges, which are specified by a set E of ordered pairs of integers (i, j), $i, j \in \{1, \ldots, n\}$. Since the graph is undirected, we may as well consider only those edges where $i < j$. We associate with an edge (i, j) a *link* $p_j - p_i$. Geometrically, this is the vector from p_i to p_j. We can order the links in this way: increasing j then increasing i where there's a link from p_i to p_j and $i < j$. Figure 2 shows an example with five nodes and six links, denoted e_k. Such a planar figure is called a *framework*. It is specified by an ordered list of its nodes, $p = (p_1, \ldots, p_n)$, together with the edge set, E. The vector p lives in $\mathbb{R}^{2n}$.

Two frameworks (p, E), (q, E), with the same number of nodes and the same edge sets, are *congruent* if one can be obtained from the other by a rigid body motion—translation followed by rotation.

A framework is rigid if it is locally uniquely determined up to conguence by its link lengths. This can be defined more precisely as follows. Suppose the

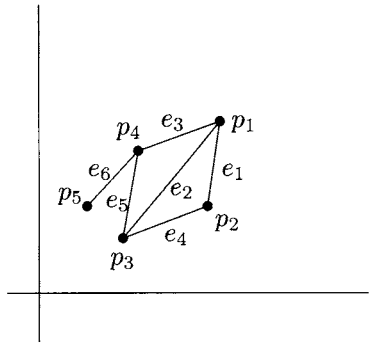


Fig. 2. Example of a framework

framework has n nodes and m links. The links have an ordering; enumerate them as $e_1, \ldots, e_m$. Define the *rigidity function*

$$r(p) = (\|e_1\|^2, \ldots, \|e_m\|^2), \quad r : \mathbb{R}^{2n} \longrightarrow \mathbb{R}^m.$$

Notice that r implicitly depends on the edge set E. The framework (p, E) is *rigid* if there exists an open neighbourhood $\mathcal{U}$ of p such that if $q \in \mathcal{U}$ and $r(q) = r(p)$, then (q, E) and (p, E) are congruent.

From a computational point of view, a more useful but slightly weaker, linear notion is infinitesimal rigidity. This means that, up to first order, $r(p)$ will not change if p is perturbed. The Taylor series of r at p is

$$r(p + \delta p) = r(p) + Dr(p)\delta p + \cdots .$$

Here $Dr(p)$ is the derivative, i.e., Jacobian matrix, of r at p. Thus $r(p+\delta p) = r(p)$ up to first order iff $Dr(p)\delta p = 0$, that is, $\delta p \in \text{Ker } Dr(p)$. The matrix $Dr(p)$ is called the *rigidity matrix*. Its kernel necessarily has dimension at least 3, because $r(p+\delta p)$ will be equal to $r(p)$ if (p, E) is perturbed to $(p+\delta p, E)$ by a rigid body transformation. The framework is said to be *infinitesimally rigid* if the kernel has exactly dimension equal to 3:

$$\dim(\text{Ker } Dr(p)) = 3.$$

Example. For the triangle in Figure 3 we have

$$r(p) = (\|e_1\|^2, \|e_2\|^2, \|e_3\|^2), \quad Dr(p) = 2 \begin{bmatrix} -e_1^T & e_1^T & 0 \\ -e_2^T & 0 & e_2^T \\ 0 & -e_3^T & e_3^T \end{bmatrix}.$$

The latter matrix, via elementary column operations, has the same rank as

$$\begin{bmatrix} e_1^T & 0 & 0 \\ e_2^T & e_2^T & 0 \\ 0 & e_3^T & 0 \end{bmatrix}.$$

This 3×6 matrix has linearly independent rows unless two of the e_i's are collinear. Thus the triangle is infinitesimally rigid iff it has positive area. □

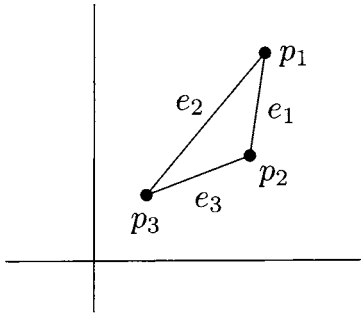


Fig. 3. Triangle example

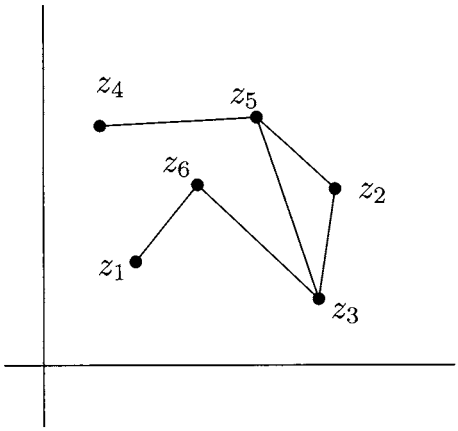


Fig. 4. Rendezvous example

4 Review of Rendezvous

Consider six robots, modeled by $\dot{z}_i = u_i$. Each $z_i(t)$ and $u_i(t)$ belongs to $\mathbb{R}^2$. The state of this system of six robots is $z = (z_1, \ldots, z_6)$ and state space is $\mathcal{Z} = \mathbb{R}^{12}$. Suppose the goal is for them to *rendezvous*—converge to a common location. One way to induce this to happen is to start with a connected, undirected graph with six nodes. Embed this into configuration space $\mathbb{R}^2$ as a framework, as in Figure 4. There are six links in this framework; enumerate them as $e_1, \ldots, e_6$. Next, define the potential function

$$V(z) = \frac{1}{2} \sum_{i=1}^{6} \|e_i\|^2.$$

Obviously, $V(z) = 0$ is the rendezvous condition. Then the proposed control law is gradient descent:

$$u = (u_1, \ldots, u_6), \quad u = -DV(z)^T.$$

For example, the control law for robot 2 is

$$\begin{aligned} u_2 &= -\left[\frac{\partial}{\partial z_2} \frac{1}{2}(\|z_5 - z_2\|^2 + \|z_3 - z_2\|^2)\right]^T \\ &= (z_5 - z_2) + (z_3 - z_2). \end{aligned}$$

Thus the sensor requirements are exactly as shown in the framework: Robot 2 must see robots 5 and 3. Such control laws can be implemented using only onboard sensors.

The closed-loop equations are

$$\dot{z} = -(L \otimes I_2)z,$$

where L is the Laplacian of the graph.[1] Classical algebraic graph theory says that, because the graph is connected, $-L$ has one zero eigenvalue and the others have negative real parts. This is enough to conclude that rendezvous will occur from all initial positions. The rendezvous formation is defined by the 2-dimensional subspace

$$\mathcal{R} = \{z : z_1 = \cdots = z_6\} \subset \mathcal{Z}.$$

For every $z(0)$, $z(t)$ converges to a point on $\mathcal{R}$.

5 Polygon Formation

Consider now the same six robots but where the goal is for them to form a d-polygon. Prescribing only the six lengths doesn't constrain the polygon to be regular, because the framework isn't rigid. To rectify this we have to add more links, say as in Figure 5. This is more links than necessary, but the symmetry is appealing. The inner links are somewhat longer, length d_1 (uniquely determined by d and the number of robots).

The goal is to design control laws u_i that stabilize the polygon formation. In contrast to rendezvous, such laws must be nonlinear. Our control law is a gradient law. Let's say the desired length of e_i is δ_i, which equals d or d_1, as appropriate. Define the potential function

$$V(z) = \frac{1}{2}\sum_{i=1}^{12} \left(\|e_i\|^2 - \delta_i^2\right)^2. \tag{1}$$

(This form is preferred over $\sum \left(\|e_i\| - \delta_i\right)^2$ because the latter leads to control terms with $1/\|e_i\|$.) Then the control law is

$$u = (u_1, \ldots, u_6), \quad u = -DV(z)^T.$$

For example, the control law for robot 1 is

$$\begin{aligned} u_1 = {} & \left(\|z_2 - z_1\|^2 - d^2\right)(z_2 - z_1) + \left(\|z_6 - z_1\|^2 - d^2\right)(z_6 - z_1) \\ & + \left(\|z_3 - z_1\|^2 - d_1^2\right)(z_3 - z_1) + \left(\|z_5 - z_1\|^2 - d_1^2\right)(z_5 - z_1). \end{aligned}$$

Thus the sensor requirements are exactly as shown in the rigid framework: Robot 1 must see robots 5,6,2,3.

Control laws derived in this way, from an infinitesimally rigid framework and a single potential function, are symmetrical in that if robot i sees robot j, so does robot j see robot i.

The closed-loop equation for the six robots is

$$\dot{z} = f(z) := -DV(z)^T.$$

[1] The inflation from L to $L \otimes I_2$ is necessary because each z_i is 2-dimensional. Had we modeled z_i as a point in the complex plane, we would have had $\dot{z} = -Lz$.

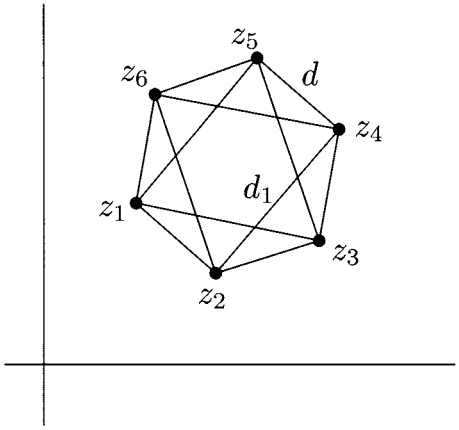


Fig. 5. Regular polygon made rigid by adding links

We turn to the concept of formation manifold, instead of formation subspace as for rendezvous. For our polygon formation, the formation manifold, denoted $\mathcal{P}$, is the set of all $z \in \mathcal{Z}$ such that $z_1, \ldots, z_6$ form a regular polygon with counterclockwise orientation and side lengths d. It is a 3-dimensional differentiable manifold. The three degrees-of-freedom come from translation and rotation in the plane. All points on $\mathcal{P}$ are equilibria, but there are other equilibria, including $\mathcal{R}$. So $\mathcal{P}$ cannot be globally attractive and we therefore take up the task of proving local asymptotic stability of $\mathcal{P}$, which is to say, if the six robots are in the polygon formation with the correct side lengths, and then their positions are perturbed, they will converge to a polygon formation with the correct side lengths. The initial polygon and the final polygon will not in general have the same locations in $\mathbb{R}^2$—the latter will be a translation/rotation of the former.

So the problem is local asymptotic stability of the manifold $\mathcal{P}$ of equilibria for the system $\dot{z} = f(z)$. It will turn out that the Jacobian $Df(p)$, for p in $\mathcal{P}$, has three zero eigenvalues and the remaining ones have negative real parts. Because of the three zero eigenvalues, local asymptotic stability can't be immediately concluded from $Df(p)$.

We remark that LaSalle's theorem is relevant and yields convergence of $z(t)$ to $\mathcal{P}$ but not to a point on $\mathcal{P}$. That is, one cannot conclude from LaSalle that the formation is stationary in the limit (even though it's easy to show that the centroid of the robots is stationary).

One might think of an alternative line of attack—to use a model more appropriate for the formation. Select a robot, say robot 1, to be the origin for a moving coordinate frame; select another, say robot 2, to provide a reference direction for the "x-axis" of the moving frame; for this to work, it must be that $z_2(t) \neq z_1(t)$ always. Then use position vectors $w_1, \ldots, w_5$ in the moving frame for robots 2 - 6: The vector $w = (w_1, \ldots, w_5)$, expressed in the moving frame, has dimension 9 (the y-component of w_1 is zero). So the Jacobian of the w-dynamics model

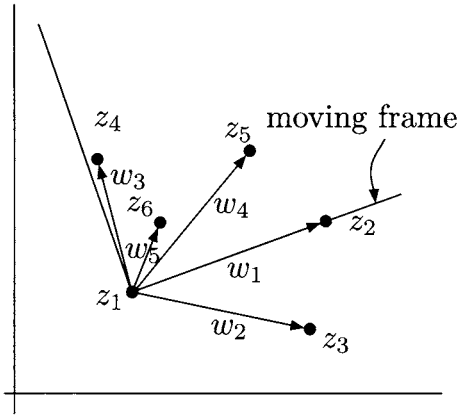


Fig. 6. Moving frame

should have only stable eigenvalues—no zero eigenvalues. The difficulty with this approach is loss of structure: The Jacobian of the w-dynamics is difficult to analyze for a general number of robots. Moreover one can't infer from w if the formation is, in global coordinates, stationary in the limit.

6 Local Asymptotic Stability of $\mathcal{P}$

Under the distributed control law designed above, the regular polygon formation is asymptotically stable. That is, if the robot positions are perturbed, the robots will return to nearby points that form the same regular polygon. The precise statement is this:

Theorem 1. *For every $p \in \mathcal{P}$ and every open neighbourhood $\mathcal{V}$ of p there exists an open neighbourhood $\mathcal{U}$ of p such that, if $z(0) \in \mathcal{U}$, then $z(t)$ converges to a point q in $\mathcal{V} \cap \mathcal{P}$.*

The proof follows from centre manifold theory.

7 Experiments

The preceding theory was tried out on real robot rovers. This section describes how that was done and what was learned.

The Robots

Our mobile robot is a two-wheeled rover—Figure 7. The robot does not have vision capability. Instead, there is an overhead camera connected to a PC. The PC

in turn communicates to the robot via bluetooth wireless. The overhead camera identifies each robot by its coloured "hat". A bright coloured disk is used to identify the robot and locate its position and a black stripe is used to mark its heading. The PC implements the control algorithmbecause the robot has limited processing and sensing capabilities. Each wheel on the robot is equipped with only an encoder that measures the distance the robot has traveled and so cannot measure any information about the other robots. The information from the wheel encoders is transmitted back to the PC via bluetooth wireless. So the central computer receives information about the robot positions from two sources: the overhead camera and the wheel encoders. This information is combined using a sensor fusion algorithm. Specific details of the sensor fusion can be found in [7].

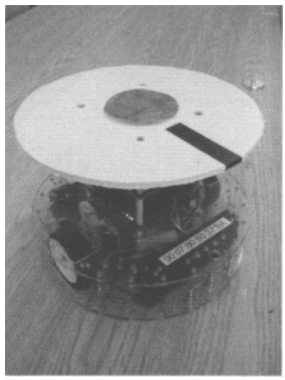


Fig. 7. Robot used in the experiments. The distinctive top is used to identify the robot and the black stripe is used to determine the robot's heading by the central computer's vision algorithm.

The robots have a large power/mass ratio, so a kinematic model is appropriate. However, there are several ways in which the real robots differ from the mathematical models. First, there is approximately a 400 ms delay in the feedback loop due to image processing, wireless communication, and computation. Secondly, because of limited camera resolution, the error in the position measurements from the camera is approximately 1 cm and the error in the heading measurement is $\pm 8°$.

Control Laws

Our theory is for a kinematic point model. Closer to our robot rover is the unicycle model

$$\begin{aligned} \dot{x} &= v \cos\theta \\ \dot{y} &= v \sin\theta \\ \dot{\theta} &= \omega. \end{aligned}$$

Here (x, y, θ) is the state, θ being the heading angle, and (v, ω) is the control input, v being the forward speed and ω the turning speed. The model cannot be linearized at a point (the linearization is uncontrollable) but it can be linearized about a point just in front. Define the new position vector

$$z = \begin{bmatrix} x + l\cos\theta \\ y + l\sin\theta \end{bmatrix},$$

where $l > 0$, and the new input

$$u = \begin{bmatrix} \cos\theta & -l\sin\theta \\ \sin\theta & l\cos\theta \end{bmatrix} \begin{bmatrix} v \\ \omega \end{bmatrix}$$

With this change of variables, the unicycle becomes our kinematic point in the plane, $\dot{z} = u$, and the derived control laws can be implemented.

Results

Our experiments were for four robots tasked with forming a square. We selected l to be approximately one tenth the length of the side of the square. Although using a relatively small l value results in more accurate positioning of the robots, there is a tradeoff: The control signals can be very large when the robot makes tight turns. (Imagine trying to park a car by linearizing about a point just ahead.) This phenomenon is especially common close to $t = 0$. If the robot's initial heading is not close to the direction of z's initial velocity, the unicycle will make a very tight turn if l is very small, resulting in a large initial angular velocity.

In the experiments there were some unexpected results. In particular, a limit cycle was observed as the robots reached the target formation. Simulations showed that such a limit cycle occurs in systems where the maximum velocity of the robots is limited and the control is calculated based on delayed information. Detailed experimental results are in [4].

8 Conclusion

In summary, the method for formation control is this: Represent the desired formation as a framework in the plane; add links if necessary until the framework is infinitesimally rigid; form a global potential function from all the links in the augmented framework; apply the negative gradient of the potential function as the control law. The resulting visibility graph—who has to see whom—is the same as the infinitesimally rigid framework. The formation is then locally asymptotically stable, though there are other equilibria.

The formation control problem can be expressed quite succinctly as one of output stabilization. As in Section 5, the setup involves

$$\dot{z} = u, \quad z = (z_1, \ldots, z_6), \quad u = (u_1, \ldots, u_6).$$

The polygon formation is defined by saying that $V(z) = 0$, where V is the potential function defined in (1). This suggests making $V(z)$ an output, say w, and driving w to 0. An added side condition is that $z(t)$ must converge (the robots themselves must asymptotically stop moving). But the challenge comes from requiring distributed control—u_i can depend only on the relative distances to specified neighbours, say, as in Figure 5. It would be interesting to attack the problem from this angle.

End Note by Bruce Francis

It is a great pleasure to contribute to this Festschrift to honour Sagar on the occasion of his 60th birthday. Sagar's breadth of contributions in systems and control is, I'm certain, unsurpassed. For me personally it has been a joy and an inspiration to know him since 1971 as a teacher, collaborator on research, mentor, and friend.

References

1. Cortés, J., Martínez, S., Bullo, F.: Robust rendezvous for mobile autonomous agents via proximity graphs in d dimensions. IEEE Transactions on Automatic Control 51(8), 1289–1298 (2004)
2. Cortés, J., Martínez, S., Karatas, T., Bullo, F.: Coverage control for mobile sensing networks. IEEE Transactions on Robotics and Automation 2(2), 243–255 (2004)
3. Goldenberg, D.K., Lin, J., Morse, A.S., Rosen, B.E., Yang, Y.R.: Towards mobility as a network control primitive. In: Proceedings of the 5th ACM International Symposium on Mobile ad hoc Networking and Computing, Tokyo, Japan, pp. 163–174 (2004)
4. Krick, L.: Application of Graph Rigidity in Formation Control of Multi-Robot Networks. MASc Thesis, University of Toronto, Electrical and Computer Engineering Dept. (2007)
5. Lin, J., Morse, A.S., Anderson, B.D.O.: The multi-agent rendezvous problem - the asynchronous case. In: Proceedings of the 43rd IEEE Conference on Decision and Control, Atlantis, Bahamas, vol. 2, pp. 1926–1931 (December 2004)
6. Marshall, J.A., Broucke, M.E., Francis, B.A.: Formations of vehicles in cyclic pursuit. IEEE Transactions on Automatic Control 9(11), 1963–1974 (2004)
7. Nawrot, J.: Time Optimal Control For Collision Avoidance Recovery of Two Unicycles. MASc Thesis, University of Toronto, Electrical and Computer Engineering Dept. (2005)
8. Olfati-Saber, R.: Flocking for multi-agent dynamic systems: Algorithms and theory. IEEE Transactions on Automatic Control 51, 401–420 (2006)
9. Smith, S.L., Broucke, M.E., Francis, B.A.: Stabilizing a multi-agent system to an equilateral polygon formation. In: MTNS 2006. Proceeds of the 17th International Symposium on Mathematical Theory of Networks and Systems, Kyoto, Japan (2006)

Convex Optimization in Infinite Dimensional Spaces*

Sanjoy K. Mitter

Department of Electrical Engineering and Computer Science, The Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, USA
mitter@mit.edu

Summary. The duality approach to solving convex optimization problems is studied in detail using tools in convex analysis and the theory of conjugate functions. Conditions for the duality formalism to hold are developed which require that the optimal value of the original problem vary continuously with respect to perturbations in the constraints only along feasible directions; this is sufficient to imply existence for the dual problem and no duality gap. These conditions are also posed as certain local compactness requirements on the dual feasibility set, based on a characterization of locally compact convex sets in locally convex spaces in terms of nonempty relative interiors of the corresponding polar sets. The duality theory and related convex analysis developed here have applications in the study of Bellman–Hamilton Jacobi equations and Optimal Transportation problems. See Fleming–Soner [8] and Villani [9].

Keywords: Convexity, Optimization, Convex Conjugate Functions? Quantum Detection.

Introduction

The duality approach to solving convex optimization problems is studied in detail using tools in convex analysis and the theory of conjugate functions. Conditions for the duality formalism to hold are developed which require that the optimal value of the original problem vary continuously with respect to perturbations in the constraints only along feasible directions; this is sufficient to imply existence for the dual problem and no duality gap. These conditions are also posed as certain local compactness requirements on the dual feasibility set, based on a characterization of locally compact convex sets in locally convex spaces in terms of nonempty relative interiors of the corresponding polar sets. The duality theory and related convex analysis developed here have applications in the study of Bellman–Hamilton Jacobi equations and Optimal Transportation problems. See Fleming–Soner [8] and Villani [9].

* Support for this research was provided by the Department of Defense MURI Grant: Complex Adaptive Networks for Cooperative Control Subaward #03-132, and the National Science Foundation Grant CCR-0325774.

V.D. Blondel et al. (Eds.) Recent Advances in Learning and Control, LNCIS 371, pp. 161–179, 2008.
springerlink.com

1 Notation and Basic Definitions

This section assumes a knowledge of topological vector spaces and only serves to recall some concepts in functional analysis which are relevant for optimization theory. The extended real line $[-\infty, +\infty]$ is denoted by $\overline{R}$. Operations in $\overline{R}$ have the usual meaning with the additional convention that

$$(+\infty) + (-\infty) = (-\infty) + (+\infty) = +\infty$$

Let X be a set, $f\colon X \to \overline{R}$ a map from X into $[-\infty, +\infty]$. The *epigraph* of f is

$$\mathrm{epi} f \triangleq \{(x, r) \in X \times R\colon r \geq f(x)\} \ .$$

The *effective domain* of f is the set

$$\mathrm{dom} f \triangleq \{x \in X\colon f(x) < +\infty\} \ .$$

The function f is *proper* iff $f \not\equiv +\infty$ and $f(x) > -\infty$ for every $x \in X$. The *indicator* function of a set $A \subset X$ is the map $\delta_A\colon X \to \overline{R}$ defined by

$$\delta_A(x) = \begin{cases} +\infty \text{ if } & x \notin A \\ 0 \quad \text{if } & x \in A \end{cases} \ .$$

Let X be a vector space. A map $f\colon X \to \overline{R}$ is *convex* iff $\mathrm{epi} f$ is a convex subset of $X \times R$, or equivalently iff

$$f(\epsilon x_1 + (1-\epsilon)x_2) \leq \epsilon f(x+1) + (1-\epsilon) f(x_2)$$

for every $x_1, x_2 \in X$ and $\epsilon \in [0,1]$. The *convex hull* of f is the largest convex function which is everywhere less than or equal to f; it is given by

$$\begin{aligned} \mathrm{co} f(x) &= \sup\{f'(x)\colon f' \text{ is convex } X \to \overline{R}\ ,\ f' \leq f\} \\ &= \sup\{f'(x)\colon f' \text{ is linear } X \to \overline{R}\ ,\ f' \leq f\} \end{aligned}$$

Equivalently, the epigraph of $\mathrm{co} f$ is given by

$$\mathrm{epi}(\mathrm{co} f) = \{(x, r) \in X \times R\colon (x, s) \in \mathrm{coepi} f \ \text{ for every } s > r\} \ ,$$

where $\mathrm{coepi} f$ denotes the convex hull of $\mathrm{epi} f$.

Let X be a topological space. A map $f\colon X \to \overline{R}$ is lower semicontinuous} (ℓsc) iff $\mathrm{epi} f$ is a closed subset of $X \times R$, or equivalently iff $\{x \in X\colon f(x) \leq r\}$ is a closed subset of X for every $r \in R$. The map $f\colon X \to \overline{R}$ is ℓsc at x_0 iff given any $r \in (-\infty, f(x_0))$ there is a neighborhood N of x_0 such that $r < f(x)$ for every $x \in N$. the *lower semicontinuous hull* of f is the largest lower semicontinuous functional on X which everywhere minorizes f, i.e.

$$\ell\mathrm{sc} f(x) = \sup\{f'(x)\colon f' \text{ is } \ell\mathrm{sc} X \to \overline{R}\ ,\ f' \leq f\} = \liminf_{x' \to x} f'(x)$$

Equivalently, $\mathrm{epi}(\ell\mathrm{sc} f) = \mathrm{cl}(\mathrm{epi} f)$ in $X \times R$.

A *duality* $\langle X, X^* \rangle$ is a pair of vector spaces X, X^* with a bilinear form $\langle \cdot, \cdot \rangle$ on $X \times X^*$ that is separating, i.e. $\langle x, y \rangle = 0 \forall y \in X^* \Rightarrow x = 0$ and $\langle x, y \rangle = 0 \forall x \in X \Rightarrow x = 0$. Every duality is equivalent to a Hausdorff locally convex space X paired with its topological dual space X^* under the natural bilinear form $\langle x, y \rangle \triangleq y(x)$ for $x \in X, y \in X^*$. We shall also write $xy \equiv \langle x, y \rangle \equiv y(x)$ when no confusion arises.

Let X be a (real) Hausdorff locally convex space (HLCS), which we shall always assume to be real. X^* denotes the topological dual space of X. The *polar* of a set $A \subset X$ and the *(pre-)polar* of a set $B \subset X^*$ are defined by[1]

$$A^0 \triangleq \{y \in X^* : \sup_{x \in A} \langle x, y \rangle \leq 1\}$$

$$^0B \triangleq \{x \in X^* : \sup_{y \in B} \langle x, y \rangle \leq 1\} \ .$$

The *conjugate* of a functional $f : X \to \overline{R}$ and the *(pre-)conjugate* of a functional $g : X^* \to \overline{R}$ are defined by

$$f^* : X^* \to \overline{R} : y \mapsto X^* \sup_{x \in X} [\langle x, y \rangle - f(x)]$$

$$g^* : X \to \overline{R} : x \mapsto X \sup_{y \in Y} [\langle x, y \rangle - g(y)] \ .$$

If X is a HLCS there are several topologies on X which are important. By τ we denote the original topology on X; by the definition of equicontinuity τ is precisely that topology which has a basis of 0-neighborhoods consisting of polars of equicontinuous subsets of X^*. The *weak topology* $w(X, X^*)$ is the weakest topology compatible with the duality $\langle X, X^* \rangle$, i.e. it is the weakest topology on X for which the linear functionals $x \mapsto \langle x, y \rangle, y \in X^*$ are continuous. Equivalently, $w(X, X^*)$ is the locally convex topology on X generated by the seminorms $x \mapsto |\langle x, y \rangle|$ for $y \in X^*$; it has a basis of 0-neighborhoods given by polars of finite subsets of X^*. The *Mackey topology* $m(X, X^*)$ on X is the strongest topology on X compatible with the duality $\langle X, X^* \rangle$[2]; it has a 0-neighborhood basis consisting of polars of all $w(X^*, X)$-compact convex[3] subsets of X^*. The *strong topology* $s(X, X^*)$ is the strongest locally convex topology on X that still has a basis consisting of $w(X, X^*)$-closed sets; it has as 0-neighborhood basis all $w(X, X^*)$-closed convex absorbing subsets of X, or equivalently all polars of $w(X^*, X)$-bounded subsets of X^*. We shall often write w, m, s for $w(X, X^*), m(X, X^*), s(X, X^*)$, and also w^* for $w(X^*, X)$. The strong topology need not be compatible with the duality $\langle X, X^* \rangle$. In general we have

[1] We use the convention $\sup \emptyset = -\infty, \inf \emptyset = +\infty$. Hence $\emptyset^0 = X^*$.

[2] A topology τ_0 on the vector space $X$is *compatible* with the duality $\langle X, X^* \rangle$ iff $(X, \tau_0)^* = X^*$, i.e. the space of all continuous linear functionals on X with the τ_0-topology may be identified with X^*.

[3] The word "convex" here may not be omitted unless X is a barrelled space. In general there may be $w(X^*, X)$-compact subsets of X^* whose closed convex hulls are not compact for the $w(X^*, X)$ topology.

$w(X, X^*) \subset \tau \subset m(X, X^*) \subset s(X, X^*)$. For a *convex* set A, however, it follows from the Hahn–Banach separation theorem that A is closed iff A is $w(X, X^*)$-closed iff A is $m(X, X^*)$-closed. More generally

$$w - \mathrm{cl}A = \mathrm{cl}A = m - \mathrm{cl}A \supset s - \mathrm{cl}A$$

when A is convex. Similarly, if a convex function $f\colon X \to \overline{R}$ is $m(X, X^*)$-ℓsc then it is ℓsc and even $w(X, X^*)$-lsc. It is also true that the bounded sets are the same for every compatible topology on X.

Let X be a HLCS and $f\colon X \to \overline{R}$. The *conjugate function* $f^*\colon X^* \to \overline{R}$ is convex and $w(X^*, X)$-ℓsc since it is the supremum of the $w(X^*, X)$-continuous affine functions $y \mapsto \langle x, y\rangle - f(x)$ over all $x \in \mathrm{dom}f$. Similarly, for ${}^*g\colon X \to R$ is convex and lsc. The conjugate functions f^*, ${}^*g$ never take on $-\infty$ values, unless they are identically $-\infty$ or equivalently $f \equiv +\infty$ or $g \equiv +\infty$. Finally, from the Hahn–Banach separation theorem it follows that

$$^*(f^*) = \ell\mathrm{scco}f \tag{1}$$

whenever f has an affine minorant, or equivalently whenever $f^* \not\equiv +\infty$; otherwise $\ell\mathrm{scco}f$ takes on $-\infty$ values and $f^* \equiv +\infty, {}^*(f^*) \equiv -\infty$.

The following lemma is very useful.

Lemma 1. *Let X be a HLCS, and let $f\colon X \to \overline{R}$. Then* $\mathrm{co}(\mathrm{dom}f) = \mathrm{dom}(\mathrm{co}f)$. *If $f^* \not\equiv +\infty$, then* $\mathrm{clcodom}f = \mathrm{cldom}^*(f^*)$.

A *barrelled space* is a HLCS X for which every closed convex absorbing set is a 0-neighborhood; equivalently, the $w(X^*, X)$-bounded sets in X^* are conditionally $w(X^*, X)$-compact. It is then clear that the $m(X, X^*)$ topology is the original topology, and the equicontinuous sets in X^* are the conditionally w^*-compact sets. Every Banach space or Frechet space is barrelled, by the Banach–Steinhaus theorem.

We use the following notation. If $A \supset X$ where X is HLCS, then $\mathrm{int}A$, $\mathrm{cor}A$, $\mathrm{ri}A$, $\mathrm{rcor}A$, $\mathrm{cl}A$, $\mathrm{span}A$, $\mathrm{aff}\,A$, $\mathrm{co}A$ denote the *interior* of A, the *algebraic interior* of A or *core*, the *relative interior* of A, the *relative core* or *algebraic interior* of A, the *closure* of A, the *span* of A, the *affine hull* of A, and the *convex hull* of A. By *relative interior* of A we mean the interior of A in the relative topology of X on $\mathrm{aff}\,A$; that is $x \in \mathrm{ri}A$ iff there is a 0-neighborhood N such that $(x + N) \cap \mathrm{aff}\,A \subset A$. Similarly, $x \in \mathrm{rcor}A$ iff $x \in A$ and $A - x$ absorbs $\mathrm{aff}\,A - x$, or equivalently iff $x + [0, \infty] \cdot A \supset A$ and $x \in A$. By *affine hull* of A we mean the smallest (not necessarily closed) affine subspace containing A; $\mathrm{aff}\,A = A + \mathrm{span}(A - A) = x_0 + \mathrm{span}(A - x_0)$ where x_0 is any element of A.

Let A be a subset of the HLCS X and B a subset of X^*. We have already defined $A^0, 0B$. In addition, we make the following useful definitions:

$$A^+ \triangleq \{y \in X^*\colon \langle x, y\rangle \geq 0 \forall x \in A\}$$
$$A^- \triangleq -A^+ = \{y \in X^*\colon \langle x, y\rangle \leq 0 \forall x \in A\}$$
$$A^\perp \triangleq A^+ \cap A^- = \{y \in X^*\colon \langle x, y\rangle = 0 \forall x \in A\} \quad .$$

Similarly, for $B \subset X^*$ the sets ${}^+B, {}^-B, {}^\perp B$, are defined in X in the same way. Using the Hahn–Banach separation theorem it can be shown that for $A \subset X, {}^0(A^0)$ is the smallest closed convex set containing $A \cup \{0\}$; ${}^+(A^+) = {}^-(A^-)$ is the smallest closed convex cone containing A; and ${}^\perp(A^\perp)$ is the smallest closed subspace containing A. Thus, if A is nonempty[4] then

$$\begin{aligned} {}^0(A^0) &= \operatorname{clco}(A \cup \{0\}) \\ {}^+(A^+) &= \operatorname{cl}[0, \infty) \cdot \operatorname{co} A \\ {}^\perp(A^\perp) &= \operatorname{clspan} A \\ A + {}^\perp((A - A)^\perp) &= \operatorname{cl} \operatorname{aff} A \ . \end{aligned}$$

2 Some Results from Convex Analysis

A detailed study of convex functions, their relative continuity properties, their sub-gradients and the relation between relative interiors of convex sets and local equicontinuity of polar sets is presented in the doctoral dissertation of S.K. Young [1977], written under the direction of the present author. In this section, we cite the relevant theorems needed in the sequel. The proofs may be found in the above-mentioned reference.

Theorem 1. *Let X be a HLCS, $f\colon X \to \overline{\mathbb{R}}$ convex and M an affine subset of X with the induced topology, $M \supset \operatorname{dom} f$.*

Let $f(\cdot)$ be bounded above on a subset C of X where $\operatorname{ri} C \neq \emptyset$ and $\operatorname{aff} C$ is closed with finite co-dimension in M. Then, $\operatorname{rcorcodom} f \neq \emptyset$, $\operatorname{co} f$ restricted to $\operatorname{rcorcodom} f$ is continuous and $\operatorname{aff} \operatorname{dom} f$ is closed with finite co-dimension in M. Moreover, $f^* \equiv +\infty$ of $\exists x_0 \in X$, $r_0 > -f(x_0)$, such that $\{y \in X^* | f^*(y) - \langle x_0, y\rangle \leq r_0\}$ is $w(X^*, X)/M^\perp$ locally bounded.

Proposition 1. *Let $f\colon X \to \overline{R}$ convex be a function on the HLCS X. The following are equivalent:*

(1) $y \in \partial f(x_0)$
(2) $f(x) \geq f(x_0) + \langle x - x_0, y\rangle \ \forall x \in X$
(3) x_0 solves $\inf_x [f(x) - xy]$, i.e. $f(x_0) - \langle x_0, y\rangle = \inf_x [f(x) - \langle x, y\rangle]$
(4) $f^*(y) = \langle x_0, y\rangle - f(x_0)$
(5) $x_0 \in \partial f^*(y)$ and $f(x_0) = {}^*(f^*)(x_0)$.

If $f(\cdot)$ is convex and $f(x_0) \in R$, then each of the above is equivalent to

(6) $f'(x_0\,;\, x) \geq \langle x, y\rangle \ \forall x \in X$.

Theorem 2. *Let $f\colon X \to \overline{R}$ convex be a function on the HLCS X, with $f(x_0)$ finite. Then the following are equivalent:*

(1) $\partial f(x_0) \neq \emptyset$
(2) $f'(x_0\,;\,\cdot)$ is bounded below on a 0-neighborhood in X, i.e. there is a 0-neighborhood N such that $\inf_{x \in N} f'(x_0\,;\, x) > -\infty$

[4] If $A = \emptyset$, then ${}^0(A^0) = {}^+(A^+) = {}^\perp(A^\perp) = \{0\}$.

(3) $\exists 0-\text{nbhd } N,\ \delta > 0 \text{ st } \inf_{\substack{x \in N \\ 0 < t < \delta}} \dfrac{f(x_0 + tx) - f(x_0)}{t} > -\infty$

(4) $\liminf_{x \to 0} f'(x_0\,;\, x) > -\infty$

(5) $\liminf_{\substack{x \to 0+ \\ t \to 0}} \dfrac{f(x_0 + tx) - f(x_0)}{t} > -\infty$

(6) $\exists\ y \in X^* \text{ st } f(x_0 + x) - f(x_0) \geq \langle xy \rangle\ \forall x \in X.$

If X is a normed space, then each of the above is equivalent to:

(7) $\exists M > 0 \text{ st } f(x_0 + x) - f(x_0) \geq -M|x|\ \forall x \in X$

(8) $\exists M > 0\ , \epsilon > 0 \text{ st whenever } |x| \leq \epsilon\ ,\ f(x_0 + x) - f(x_0) \geq -M|x|$

(9) $\liminf_{|x| \to 0} \dfrac{f(x_0 + x) - f(x_0)}{|x|} > -\infty$

Definition 1. *The* recession function f_∞ *of a function* $f\colon X \to \overline{\mathbb{R}}$ *is defined to be*

$$f_\infty = \sup_{y \in \mathrm{dom} f^*} \langle x, y \rangle\ .$$

Proposition 2. *Let* $f\colon X \to \overline{\mathbb{R}}$ *be a convex* ℓsc *proper function on the HLCS* X. *Then* $f_\infty(x)$ *is given by each of the following:*

(1) $\min\{r \in R\colon (x, r) \in (\mathrm{epi} f)_\infty\}$

(2) $\sup_{a \in \mathrm{dom} f} \sup_{t > 0} [f(a + tx) - f(a)]/t$

(3) $\sup_{t > 0} [f(a + tx) - f(a)]/t$ *for any fixed* $a \in \mathrm{dom} f$

(4) $\sup_{a \in \mathrm{dom} f} [f(a + x) - f(a)]$

(5) $\sup_{y \in \mathrm{dom} f^*} \langle x, y \rangle.$

In (1), the minimum is always attained (whenever it is not $+\infty$*), since* $(\mathrm{epi} f)_\infty$ *is a closed set.*

Theorem 3. *Let* X *be HLCS,* $f\colon X \to \overline{R}$ *convex. Assume* $\mathrm{riepi} f \neq \emptyset$. *Then* $f(\cdot)$ *is continuous relative to* $\mathrm{aff\,dom} f$ *on* $\mathrm{rcordom} f$, *and the following are equivalent for a point* $x_0 \in X$*:*

(1) $f(\cdot)$ *is relatively continuous at* $x_0 \in \mathrm{dom} f$

(2) $x_0 \in \mathrm{rcordom} f$

(3) $\mathrm{dom} f - x_0$ *absorbs* $x_0 - \mathrm{dom} f$

(4) $\forall x \in \mathrm{dom} f,\ \exists \epsilon > 0 \text{ st } (1 + \epsilon)x_0 - \epsilon x \in \mathrm{dom} f$

(5) $[\mathrm{dom} f - x_0]^- \subset [\mathrm{dom} f - x_0]^\perp \equiv \{y \in X^*\colon y \equiv$ *constant on* $\mathrm{dom} f$

(6) $[\mathrm{dom} f - x_0]^-$ *is a subspace*

(7) $\{y \in X^*\colon (f^*)_\infty(y) - x_0 y \leq 0\}$ *is a subspace*

(8) $x_0 \in \mathrm{dom} f$ *and* $\{y \in X^*\colon f^*(y) - x_0 y \leq r\}_\infty$ *for some* $r \geq -f(x_0)$

(9) $\partial f(x_0) \neq \emptyset$ *and* $(\partial f(x_0))_\infty$ *is a subspace*

(10) $\partial f(x_0)$ *is nonempty and* $w(X^*, \mathrm{aff\,dom} f - x_0)$*-compact.*

3 Duality Approach to Optimization

3.1 Introduction

The idea of duality theory for solving optimization problems is to transform the original problem into a "dual" problem which is easier to solve and which has the same value as the original problem. Constructing the dual solution corresponds to solving a "maximum principle" for the problem. This dual approach is especially useful for solving problems with difficult implicit constraints and costs (e.g. state constraints in optimal control problems), for which the constraints on the dual problem are much simpler (only explicit "control" constraints). Moreover the dual solutions have a valuable sensitivity interpretation: the dual solution set is precisely the subgradient of the change in minimum cost as a function of perturbations in the "implicit" constraints and costs.

Previous results for establishing the validity of the duality formalism, at least in the infinite-dimensional case, generally require the existence of a feasible interior point ("Kuhn–Tucker" point) for the implicit constraint set. This requirement is restrictive and difficult to verify. Rockafellar [5, Theorem 11] has relaxed this to require only continuity of the optimal value function. In this chapter we investigate the duality approach in detail and develop weaker conditions which require that the optimal value of the minimization problem varies continuously with respect to perturbations in the implicit constraints only along feasible directions (that is, we require relative continuity of the optimal value function); this is sufficient to imply existence for the dual problem and no duality gap. Moreover we pose the conditions in terms of certain local compactness requirements on the dual feasibility set, based on results characterizing the duality between relative continuity points and local compactness.

To indicate the scope of our results let us consider the Lagrangian formulation of nonlinear programming problems with generalized constraints. Let U, X be normed spaces and consider the problem

$$P_0 = \inf\{f(u)\colon u \in C, g(u) \leq 0\}$$

where C is a convex subset of U, $f\colon C \to R$ is convex, and $g\colon C \to X$ is convex in the sense that

$$g(tu_1 + (1-t)u_2) \leq tg(u_1) + (1-t)g(u_2)\ ,\ u_1, u_2 \in C\ ,\ t \in [0,1]\ .$$

We are assuming that X has been given the partial ordering induced by a nonempty closed convex cone Q of "positive vectors"; we write $x_1 > x_2$ to mean $x_1 - x_2 \in Q$. The dual problem corresponding to P_0 is well-known to be

$$D_0 = \sup_{y \in Q^+} \inf_{u \in C} [f(u) + \langle g(u), y\rangle]\ ;$$

this follows from (6) below by taking $L \equiv 0$, $x_0 = 0$, and

$$F(u,x) = \begin{cases} f(u) \text{ if } u \in C, g(u) \leq x \\ +\infty \text{ otherwise.} \end{cases} \tag{2}$$

We also remark that it is possible to write

$$P_0 = \inf_u \sup_y (u, y) \qquad D_0 = \sup_y \inf_u (u, y)$$

where we have defined the Lagrangian function by

$$\ell(u, y) = \begin{cases} +\infty & \text{if } u \in C \\ f(u) - \langle g(u), y \rangle & \text{if } u \in C, y \in Q^- \\ -\infty & \text{if } u \in C, y \in Q^- \end{cases} .$$

In analyzing the problem P_0, we imbed it in the family of perturbed problems

$$P(x) = \inf\{f(u) \colon u \in C, g(u) \leq x\} .$$

It then follows that the dual problem is precisely the second conjugate of P_0 evaluated at 0: $D_0 =^* (P^*)(0)$. Moreover if there is no duality gap ($P_0 = D_0$) then the dual solution set is the subgradient $\partial P(0)$ of $P(\cdot)$ at 0. The following theorem summarizes the duality results for this problem.

Theorem 4. *Assume P_0 is finite. The following are equivalent:*

(1) $P_0 = D_0$ and D_0 has solutions
(2) $\partial P(0) \neq \emptyset$
(3) $\exists \hat{y} \in Q^+$ st $P_0 = \inf_{u \in C}[f(u) + \langle g(u), \hat{y} \rangle]$
(4) $\exists \epsilon > 0, M > 0$ st $f(u) \geq P_0 - M|x|$ whenever $u \in C, |x| \leq \epsilon, g(u) \leq x$.

If (1) is true then $\hat{u}$ is a solution for P_0 iff $\hat{u} \in C, g(u) \leq 0$, and there is a $\hat{y} \in Q^+$ satisfying

$$f(u) + \langle g(u), \hat{y} \rangle \geq f(\hat{u}) \forall u \in C ,$$

in which case complementary slackness holds, i.e. $\langle g(\hat{u}), \hat{y} \rangle = 0$, and $\hat{y}$ solves D_0.

Proof. This follows directly from Theorem 6 with F defined by (2). ∎

We remark here that criterion (4) is necessary and sufficient for the duality result (1) to hold, and it is crucial in determining how strong a norm to use on the perturbation space X (equivalently, how large a dual space X^* is required in formulating a well-posed dual problem).

The most familiar assumption which is made to insure that the duality results of Theorem 4 hold is the existence of a Kuhn–Tucker point:

$$\forall \overline{u} \in C \text{ st } -g(\overline{u}) \in\in Q$$

This is a very strong requirement, and again is often critical in determining what topology to use on the perturbation space X. More generally, we need only require that $P(\cdot)$ is continuous as 0. Rockafellar has presented the following result [5]: if U is the normed dual of a Banach space V, if X is a Banach space, if g is lower semicontinuous in the sense that

$$\text{epi} g \triangleq \{(u, x) \colon g(u) \leq x\}$$

is closed in $U \times X$ (e.g. if g is continuous), then the duality results of Theorem 4 hold whenever

$$0 \in \text{core}[g(C) + Q] \ .$$

In fact, it then follows that $P(\cdot)$ is continuous at 0. The following theorem relaxes this result to relative continuity and also provides a dual characterization in terms of local compactness requirements which are generally easier to verify.

Theorem 5. *Assume P_0 is finite. The following are equivalent:*

(1) $\text{aff}[g(C) + Q]$ *is closed; and* $0 \in \text{rcor}[g(C) + Q]$, *or equivalently* $\forall u \in C, \forall x \leq g(u) \exists \epsilon > 0$ *and* $u_1 \in C$ st $g(u_1) + \epsilon x \leq 0$.

(2) $Q^+ \cap g(C)^+$ *is a subspace M; and there is an $\epsilon > 0$, and $x_1 \in X$, and $r_1 \in R$ such that $\{y \in Q^+ \colon \inf_{|v| \leq \epsilon} \sup_{u \in C}[f(u) + g(u)y - uv] > r_1\}$ is nonempty and $w(X^*, X)/M$-locally bounded.*

If either of the above holds, then $P(\cdot)$ is relatively continuous at 0 and hence Theorem 4 holds. Moreover the dual solutions have the sensitivity interpretation

$$P'(0; x) = \max\{\langle x, y\rangle \colon y \text{ solves } D_0\}$$

where the maximum is attained and $P'(0; \cdot)$ denotes the directional derivative of the optimal value function $P(\cdot)$ evaluated at 0.

Proof. This follows directly from Theorem 9. ■

3.2 Problem Formulation

In this section we summarize the duality formulation of optimization problems. Let U be a HLCS of controls; X a HLCS of states; $u \mapsto Lu + x_0$ an affine map representing the system equations, where $x_0 \in X$, and $L \colon U \to X$ is linear and continuous; $F \colon U \times X \to \bar{R}$ a cost function. We consider the *minimization problem*

$$P_0 = \inf_{u \in U} F(u, Lu + x_0) \quad , \tag{3}$$

for which feasibility constraints are represented by the requirement that $(u, Lu + x_0) \in \text{dom}F$. Of course, there are many ways of formulating a given optimization problem in the form (3) by choosing different spaces U, X and maps L, F; in general the idea is to put explicitly, easily characterized costs and constraints into the "control" costs on U and to put difficult implicit constraints and costs into the "state" part of the cost where a Lagrange multiplier representation can be very useful in transforming implicit constraints to explicit constraints. The dual variables, or multipliers will be in X^*, and the dual problem is an optimization in X^*.

In order to formulate the dual problem we consider a family of *perturbed problems*

$$P(x) = \inf_{u \in U} F(u, Lu + x) \tag{4}$$

where $x \in X$. Note that if $F: U \times X \to \overline{R}$ is convex then $P: X \to \overline{R}$ is convex; however F ℓsc does not imply that P is lsc. Of course $P_0 = P(x_0)$. We calculate the conjugate function of P

$$P^*(y) = \sup_x[\langle x, y\rangle - P(x)] = \sup_{ux,}[\langle x, y\rangle - F(u, Lu + x)] = F^*(-L^*y, y) \quad . \tag{5}$$

The dual problem of $P_0 = P(x_0)$ is given by the second conjugate of P evaluated at x_0, i.e.

$$D_0 =^* (P^*)(x_0) = \sup_{y \in X^*}[\langle x_0, y\rangle - F^*(-L^*y, y)] \tag{6}$$

The feasibility set for the dual problem is just $\mathrm{dom}P^* = \{y \in X^* : (-L^*y, y) \in \mathrm{dom}F^*\}$. We immediately have

$$P_0 \equiv P(x_0) \geq D_0 \equiv^* (P^*)(x_0) \quad . \tag{7}$$

Moreover, since the primal problem P_0 is an infimum, and the dual problem D_0 is a supremum, and $P_0 \geq D_0$, we see that if $\hat{u} \in U, \hat{y} \in X^*$ satisfy

$$F(\hat{u}, L\hat{u} + x_0) = \langle x_0, \hat{y}\rangle - F^*(-L^*\hat{y}, \hat{y}) \tag{8}$$

then $P_0 = D_0 = F(\hat{u}, L\hat{u} + x_0)$ and (assuming $P_0 \in R$) $\hat{u}$ is optimal for P, $\hat{y}$ is optimal for D. Thus, the existence of a $\hat{y} \in X^*$ satisfying (8) is a sufficient condition for optimality of a control $\hat{u} \in U$; we shall be interested in condition under which (8) is also necessary. It is also clear that any "dual control" $y \in X^*$ provides a lower bound for the original problem: $P_0 \geq \langle x_0, y\rangle - F^*(-L^*y, y)$ for every $y \in X^*$.

The duality approach to optimization problems P_0 is essentially to vary the constraints slightly as in the perturbed problem $P(x)$ and see how the minimum cost varies accordingly. In the case that F is convex, $P_0 = D_0$ or no "duality gap" means that the perturbed minimum costs function $P(\cdot)$ is ℓsc at x_0. The stronger requirement that the change in minimum cost does not drop off too sharply with respect to perturbations in the constraints, i.e. that the directional derivative $P'(x_0\,;\,\cdot)$ is bounded below on a neighborhood of x_0, corresponds to the situation that $P_0 = D_0$ and the dual problem D_0 has solutions, so that (8) becomes a necessary and sufficient condition for optimality of a control $\hat{u}$. It turns out that the solution of D_0 when $P_0 = D_0$ are precisely the element of $\partial P(x_0)$, so that the dual solutions have a sensitivity interpretation as the subgradients of the change in minimum cost with respect to the change in constraints.

Before stating the above remarks in a precise way, we define the Hamiltonian and Lagrangian functions associated with the problem P_0. We denote by $F_u(\cdot)$ the functional $F(u, \cdot): x \to F(u, x): X \to \overline{R}$, for $u \in U$. The *Hamiltonian* function $H: U \times X^* \to \overline{R}$ is defined by

$$H(u, y) = \sup_{x \in X}[\langle x, y\rangle - F(u, x)] = F_u^*(y) \quad . \tag{9}$$

Proposition 3. *The Hamiltonian H satisfies:*

(1) $({}^*H_u)(x) = {}^*(F_u^*)(x)$
(2) $({}^*H_u)^*(y) = H_u(y) = F_u^*(y)$
(3) $F^*(v,y) = \sup_u[\langle u,v\rangle + H(u,y)] = (-H(\cdot\,,y))^*(v)$.

Moreover $H(u,\cdot\,)$ is convex and w^* – ℓsc $X^* \to \overline{R}$; $H(\cdot\,,y)$ is concave $U \to \overline{R}$ if F is convex; if $F(u,\cdot\,)$ is convex, proper, and ℓsc then $H(\cdot\,,y)$ is concave for every y iff F is convex.

Proof. The equalities are straightforward calculations. $H(u,\cdot\,)$ is convex and ℓsc since $({}^*H_u)^* = H_u$. It is straightforward to show that $-H(\cdot\,,y)$ is convex if $F(\cdot\,)$ is convex. On the other hand if $({}^*F_u)^* = F_u$ and $H(\cdot\,,y)$ is concave for every $y \in X^*$, then

$$F(u,x) = {}^*(F_u^*)(x) = {}^*H_u(x) = \sup_y[xy - H(u,y)]$$

is the supremum of the convex functionals $(u,x) \mapsto \langle x,y\rangle - H(u,y)$ and hence F is convex. ∎

The *Lagrangian* function $\ell\colon U \times X^*$ to $\overline{R}$ is defined by

$$\ell(u,y) = \inf_x[F(u,Lu+x_0+x) - \langle x,y\rangle] = \langle Lu+x_0,y\rangle - F_u^*(y)$$
$$= \langle Lu+x_0,y\rangle - H(u,y) \quad . \quad (10)$$

Proposition 4. *The Lagrangian ℓ satisfies*

(1) $\inf_u \ell(u,y) = \langle x_0,y\rangle - F^*(-L^*y,y)$
(2) $D_0 \equiv {}^*(P^*)(x_0) = \sup_y \inf_u \ell(u,y)$
(3) ${}^*(-\ell_u)(x) = {}^*(F_u^*)(Lu+x_0+x)$
(4) $P_0 \equiv P(x_0) = \inf_u \sup_y \ell(u,y)$ *if* $F_u = {}^*(F_u^*)$ *for every* $u \in U$.

Moreover $\ell(u,\cdot\,)$ is convex and w^*-ℓsc $X^* \to \overline{R}$ for every $u \in U$; $\ell(\cdot\,)$ is convex $U \times X^* \to \overline{R}$ if F is convex; if $F_u = {}^*(F_u^*)$ for every $u \in U$ then ℓ is convex iff F is convex.

Proof. The first equality (1) is direct calculation; (2) then follows from (1) and (4). Equality (3) is immediate from (10); (4) then follows from (3) assuming that ${}^*(F_u^*) = F_u$. The final remarks follow from Proposition 3 and the fact that $\ell(u,y) = \langle Lu+x_0,y\rangle - H(u,y)$. ∎

Thus from Proposition 4 we see that the duality theory based on conjugate functions includes the Lagrangian formulation of duality for inf-sup problems. For, given a Lagrangian function $\ell\colon U \times X^* \to \overline{R}$, we can define $F\colon U \times X \to \overline{R}$ by $F(u,x) = {}^*(-\ell_u)(x) = \sup_y[\langle x,y\rangle + \ell(u,x)]$, so that

$$P_0 = \inf_u \sup_y \ell(u,y) = \inf_u F(u,0)$$
$$D_0 = \sup_y \inf_u \ell(u,y) = \sup_y -F^*(0,y) \quad ,$$

which fits into the conjugate duality framework.

For the following we assume as before that U, X are HLCS's; $L: U \to X$ is linear and continuous; $x_0 \in X$; $F: U \times X \to \overline{R}$. We define the family of optimization problems $P(x) = \inf_u F(u, Lu+x)$, $P_0 = P(x_0)$, $D_0 = \sup_y[\langle x, y\rangle - F^*(-L^*y, y)] =^* (P^*)(x_0)$. We shall be especially interested in the case that $F(\cdot)$ is convex, and hence $P(\cdot)$ is convex.

Proposition 5. (no duality gap). *It is always true that*

$$P_0 \equiv P(x_0) \geq \inf_u \sup_y \ell(u, y) \geq D_0 \equiv \inf_u \sup_y \ell(u, y) \equiv^* (P^*)(x_0) \ . \tag{11}$$

If $P(\cdot)$ is convex and D_0 is feasible, then the following are equivalent:

(1) $P_0 = D_0$

(2) $P(\cdot)$ *is* ℓsc *at* x_0, *i.e.* $\lim_{x \to x_0} \inf P(x) \geq P(x_0)$

(3) $\sup_{F \text{ finite} \subset X^*} \inf_{\substack{u \in U \\ x \in Lu + x_0 + {}^0F}} F(u, x) \geq P_0$

These imply, and are equivalent to, if $F_u =^* (F_u^*)$ *for every* $u \in U$,

(4) ℓ *has a saddle value, i.e.* $\inf_u \sup_y \ell(u, y) = \sup_y \inf_u \ell(u, y)$.

Proof. The proof is immediate since $P_0 = P(x_0)$ and $D_0 =^* (P^*)(x_0)$. Statement (4) follows from Proposition 4 and Eq. (11). ■

Theorem 6. (no duality gap and dual solutions). *Assume P_0 is finite. The following are equivalent:*

(1) $P_0 = D_0$ *and* D_0 *has solutions*
(2) $\partial P(x_0) \neq \emptyset$
(3) $\exists \hat{y} \in Y \text{ st } P_0 = \langle x_0, \hat{y}\rangle - F^*(-L^*\hat{y}, \hat{y})$
(4) $\exists \hat{y} \in Y \text{ st } P_0 = \inf_u \ell(u, \hat{y})$.

If $P(\cdot)$ is convex, then each of the above is equivalent to

(5) $\exists 0 -$ neighborhood $N \text{ st } \inf_{x \in N} P'(x_0; x) > -\infty$
(6) $\liminf_{x \to 0} P'(x_0; x) > -\infty$
(7)
$$\liminf_{x \to 0 + t \to 0} \frac{P(x_0 + tx) - P_0}{t}$$
$$\equiv \sup_{N = 0-\text{nbhd}} \inf_{t > 0} \inf_{x \in N} \inf_{u \in U} \frac{F(u, Lu + x_0 + tx) - P_0}{t} > -\infty \ .$$

If $P(\cdot)$ is convex and X is a normed space, then the above are equivalent to:

(8) $\exists \epsilon > 0, M > 0$ st $F(u, Lu + x_0 + x) - P_0 \geq -M|x| \forall u \in U, |x| \leq \epsilon$.
(9) $\exists \epsilon > 0, M > 0$ st $\forall u \in U, |x| \leq \epsilon, \delta > 0 \exists u' \in U$ st $F(u, Lu + x_0 + x) - F(u', Lu' + x_0) \geq -M|x| - \epsilon$.

Moreover, if (1) is true then $\hat{y}$ solves D_0 iff $\hat{y} \in \partial P(x_0)$, and $\hat{u}$ is a solution for P_0 iff there is a $\hat{y}$ satisfying any of the conditions (1′)–(3′) below. The following statements are equivalent:

(1′)$\hat{u}$ *solves* P_0, $\hat{y}$ *solves* D_0, *and* $P_0 = D_0$
(2′)$F(\hat{u}, L\hat{u} + x_0) = \langle x_0, \hat{y}\rangle - F^*(-L^*\hat{y}, \hat{y})$
(3′)$(-L^*\hat{y}, \hat{y}) \in \partial F(\hat{u}, L\hat{u} + x_0)$.

These imply, and are equivalent to, if $F(u, \cdot)$ *is proper convex lsc* $X \to \overline{R}$ *for every* $u \in U$, *the following equivalent statements:*

(4′)$0 \in \partial\ell(\cdot, \hat{y})(\hat{u})$ *and* $0 \in \partial(-\ell(\hat{u}, \cdot))(\hat{y})$, *i.e.* $(\hat{u}, \hat{y})$ *is a saddlepoint of* ℓ, *that is* $\ell(\hat{u}, y) \leq \ell(\hat{u}, \hat{y}) \leq \ell(u, \hat{y})$ *for every* $u \in U, y \in X^*$).
(5′)$L\hat{u} + x_0 \in \partial H(\hat{u}, \cdot)(\hat{y})$ *and* $L^*\hat{y} \in \partial(-H(\cdot, \hat{y}))(\hat{u})$, *i.e.* $\hat{y}$ *solves* $\inf_y[H(\hat{u}, y) - \langle L\hat{u} + x_0, y\rangle]$ *and* $\hat{u}$ *solves* $\inf_u[H(u, \hat{y}) + \langle u, L^*\hat{y}\rangle]$.

Proof. (1) => (2). Let $\hat{y}$ be a solution of $D_0 =^* (P^*)(x_0)$. Then $P_0 = \langle x_0, \hat{y}\rangle - P^*(\hat{y})$. Hence $P^*(\hat{y}) = \langle x_0, \hat{y}\rangle - P(x_0)$ and from Proposition 1, (4) => (1) we have $y \in \partial P(x_0)$.

(2) => (3). Immediate by definition of D_0.

(3) => (4) => (1). Immediate from (11).

If $P(\cdot)$ is convex and $P(x_0) \in R$, then (1) and (4)–(9) are all equivalent by Theorem 2. The equivalence of (1′)–(5′) follows from the definitions and Proposition 5. ■

Remark. *In the case that* X *is a normed space, condition (8) of Theorem 6 provides a necessary and sufficient characterization for when dual solutions exists (with no duality gap) that shows explicitly how their existence depends on what topology is used for the space of perturbations. In general the idea is to take a norm as weak as possible while still satisfying condition (8), so that the dual problem is formulated in as nice a space as possible. For example, in optimal control problems it is well known that when there are no state constraints, perturbations can be taken in e.g. an* L_2 *norm to get dual solutions* y *(and costate* $-L^*y$*) in* L_2, *whereas the presence of state constraints requires perturbations in a uniform norm, with dual solutions only existing in a space of measures.*

It is often useful to consider perturbations on the dual problem; the duality results for optimization can then be applied to the dual family of perturbed problems. Now the dual problem D_0 *is*

$$-D_0 = \inf_{y \in X^*}[F^*(-L^*y, y) - \langle x_0, y\rangle \quad .$$

In analogy with (4) we define perturbations on the dual problem by

$$D(v) = \inf_{y \in X^*}[F^*(v - L^*y, y) - \langle x_0, y\rangle] \ , \quad v \in U^* \tag{12}$$

Thus $D(\cdot)$ *is a convex map* $U^* \to \overline{R}$, *and* $-D_0 = D(0)$. *It is straightforward to calculate*

$$(^*D)(u) = \sup_v[\langle u, v\rangle - D(v)] =^* (F^*)(u, Lu + x_0) \quad .$$

Thus the "dual of the dual" is

$$-^*(D^*)(0) = \overset{*}{\inf_{u \in U}}(F^*)(u, Lu + x_0) \quad . \tag{13}$$

In particular, if $F =^* (F^*)$ *then the "dual of the dual" is again the primal, i.e.* $\mathrm{dom}^* D$ *is the feasibility set for* P_0 *and* $-^*(D^*)(0) = P_0$. *More generally, we have*

$$P_0 \equiv P(x_0) \geq -^*(D^*)(0) \geq D_0 \equiv -D(0) \equiv^* (P^*)(0) \quad . \tag{14}$$

3.3 Duality Theorems for Optimization Problems

Throughout this section it is assumed that U, X are HLCS's; $L\colon U \to X$ is linear and continuous; $x_0 \in X$ and $F\colon U \times X \to \overline{R}$. Again, $P(x) = \inf_u F(u, Lu + x_0 + x)$, $P_0 = P(x_0)$,

$$D_0 =^* (P^*)(x_0) = \sup_{y \in X^*} [\langle x_0, y\rangle - F^*(-L^*y, y)] \quad .$$

We shall be interested in conditions under which $\partial P(x_0) \neq \emptyset$; for then there is no duality gap and there are solutions for D_0. These conditions will be conditions which insure the $P(\cdot)$ is relatively continuous at x_0 with respect to $\mathrm{aff\,dom} P$, that is $P \uparrow \mathrm{aff\,dom} P$ is continuous at x_0 for the induced topology on $\mathrm{aff\,dom} P$. We then have

$$\begin{aligned} &\partial P(x_0) \neq \emptyset \\ &P_0 = D_0 \\ &\text{the solution set for } D_0 \text{ is precisely } \partial P(x_0) \\ &P'(x_0; x) = \max_{y \in \partial P(x_0)} \langle x, y\rangle \quad . \end{aligned} \tag{15}$$

This last result provides a very important sensitivity interpretation for the dual solutions, in terms of the rate of change in minimum cost with respect to perturbations in the "state" constraints and costs. Moreover if (15) holds, then Theorem 6, (1′)–(5′), gives necessary and sufficient conditions for $\hat{u} \in U$ to solve P_0.

Theorem 7. *Assume* $P(\cdot)$ *is convex (e.g.* F *is convex). If* $P(\cdot)$ *is bounded above on a subset* C *of* X, *where* $x_0 \in \mathrm{ri} C$ *and* $\mathrm{aff}\, C$ *is closed with finite codimension in an affine subspace* M *containing* $\mathrm{aff\,dom} P$, *then (15) holds.*

Proof. From Theorem 1, (1b) => (2b), we know that $P(\cdot)$ is relatively continuous at x_0. ∎

Corollary 1. (Kuhn–Tucker point). *Assume* $P(\cdot)$ *is convex (e.g.* F *is convex). If there exists a* $\overline{u} \in U$ *such that* $F(\overline{u}, \cdot)$ *is bounded above on a subset* C *of* X, *where* $L\overline{u} + x_0 \in \mathrm{ri}\ C$ *and* $\mathrm{aff}\, C$ *is closed with finite codimensions in an affine subspace* M *containing* $\mathrm{aff\,dom}\ P$, *then (15) holds. In particular, if there is a* $\overline{u} \in U$ *such that* $F(\overline{u}, \cdot)$ *is bounded above on a neighborhood of* $L\overline{u} + x_0$, *then (15) holds.*

Proof. Clearly

$$P(x) = \inf_u F(u, Lu + x) \leq F(\overline{u}, L\overline{u} + x) \ ,$$

so Theorem 1 applies. ∎

The Kuhn–Tucker condition of Corollary 1 is the most widely used assumption for duality [4]. The difficulty in applying the more general Theorem 7 is that, in cases where $P(\cdot)$ is not actually continuous but only relatively continuous, it is usually difficult to determine aff dom P. Of course,

$$\mathrm{dom}P = \bigcup_{u \in U} [\mathrm{dom}F(u, \cdot) - Lu] ,$$

but this may not be easy to calculate. We shall use Theorem 1 to provide dual compactness conditions which insure that $P(\cdot)$ is relatively continuous at x_0.

Let K be a convex balanced $w(U, U^*)$-compact subset of U; equivalently, we could take $K = {}^0N$ where N is a convex balanced $m(U^*, U)$-0-neighborhood in U^*. Define the function $g\colon X^* \to \overline{R}$ by

$$g(y) = \inf_{v \in K^0} F^*(v - L^*y, y) \quad . \tag{16}$$

Note that g is a kind of "smoothing" of $P^*(y) = F^*(-L^*y, y)$ which is everywhere majorized by P^*. The reason why we need such a g is that $P(\cdot)$ is not necessarily ℓsc, which property is important for applying compactness conditions on the levels sets of P^*; however ${}^*g$ is automatically ℓsc and ${}^*g$ dominates P, while at the same time ${}^*g$ approximates P.

Lemma 2. *Define $g(\cdot)$ as in (16). Then*

$$({}^*g)(x) \leq \inf_u [F(u, Lu + x] + \sup_v \in K^0 \langle u, v \rangle] \quad .$$

If $F = {}^*(F^*)$, then $P(x) \leq ({}^*g)(x)$ for every $x \in \mathrm{dom}P$. Moreover

$$\mathrm{dom}^*g \supset \bigcup_{u \in \mathrm{span}K} [\mathrm{dom}F(u, \cdot) - Lu] \quad .$$

Proof. By definition of ${}^*g$, we have

$$({}^*g)(x) = \sup_y \sup_{v \in K^0} [\langle x, y \rangle - F^*(v - L^*y, y)] \quad .$$

Now for every $u \in U$ and $y \in Y$, $F^*(v - L^*y, y) \geq \langle u, v - L^*y \rangle + \langle Lu + x, y \rangle - F(u, Lu + x) = \langle u, v \rangle + \langle x, y \rangle - F(u, Lu + x)$ by definition of F^*. Hence for every $u \in U$,

$$\begin{aligned} ({}^*g)(x) \leq \sup_{v \in K^0} [F(u, Lu + x) - \langle u, v \rangle] &= F(u, Lu + x) + \sup_{v \in -K^0} \langle u, v \rangle \\ &= F(u, Lu + x) + \sup_{v \in K^0} \langle u, v \rangle \end{aligned}$$

where the last equality follows since K^0 is balanced. Thus we have proved the first inequality of the lemma.

Now suppose $F =^* (F^*)$ and $x \in \mathrm{dom}P$. Since K^0 is a $m(U^*,U)$-0-neighborhood, we have

$$\begin{aligned}(^*g)(x) &= \sup_{v\in K^0}\sup_y[\langle x,y\rangle - F^*(v-L^*y,y)]\\ &\geq \limsup_{v\to 0}\sup_y[\langle x,y\rangle - F^*(v-L^*y,y)]\\ &= \liminf_{v\to 0}\inf_y[F^*(v-L^*y,y)-\langle x,y\rangle]\ ,\end{aligned}$$

where the $\liminf_{v\to 0}$ is taken in the $m(U^*,U)$-topology. Define

$$h(v) = \inf_y[F^*(v-L^*y,y)-\langle x,y\rangle]\ ,$$

so that

$$(^*g)(x) \geq -\liminf_{v\to 0} h(v)\ .$$

Now

$$\begin{aligned}(^*h)(u) &= \sup_v\sup_y[\langle u,v\rangle - F^*(v-L^*y,y)+\langle x,y\rangle]\\ &= {}^*(F^*)(u,Lu+x) = F(u,Lu+x)\ .\end{aligned}$$

Hence $P(x) < +\infty$ means that $\inf_u F(u,Lu+x) < +\infty$, i.e. $^*h \not\equiv +\infty$, so that we can replace the lim inf by the second conjugate:

$$(^*g)(x) \geq -\liminf_{v\to 0} h(v) = -(^*h)^*(0) = \inf_u F(u,Lu+x) = P(x)\ .$$

The last statement in the lemma follows from the first inequality in the lemma. For

$$x \in \bigcup_{u\in \mathrm{span}K}[\mathrm{dom}F(u,\cdot)-Lu] \text{ iff } \exists u \in [0,\infty)\cdot K \text{ st } F(u,Lu+x) < +\infty\ ,$$

$$\text{iff } \exists u \text{ st } \sup_{v\in K^0}\langle u,v\rangle < +\infty$$

and

$$F(u,Lu+x) < +\infty \quad (\text{since } K =^0 (K^0))$$

$$\text{iff } \exists u \text{ st } F(u,Lu+x) + \sup_{v\in K^0}\langle u,v\rangle < +\infty$$

and this implies that $x \in \mathrm{dom}^*g$. Hence $\mathrm{dom}^*g \subset \bigcup_{u\in \mathrm{span}K}[\mathrm{dom}F(u,\cdot)-Lu]$.

Note that $\mathrm{dom}P$ is given by $\bigcup_{u\in U}[\mathrm{dom}F(u,\cdot)-Lu]$. ■

Theorem 8. *Assume $F =^* (F^*)$, $P_0 < +\infty$, and there is a $w(U,U^*)$-compact convex subset K of U such that* $\mathrm{span}K \subset \bigcup_{x\in X}\mathrm{dom}F(\cdot,x)$. *Suppose*

(1) $\{y \in X^* : (F^*)_\infty(-L^*y,y)-\langle x_0,y\rangle \leq 0\}$ *is a subspace M;*

(2) $\exists m(U^*,U)$*-0-neighborhood N in U^*, an $x_1 \in X$, an $r_1 \in R$ such that* $\{y \in X^* : \inf_{v\in N} F^*(v-L^*y,y)-\langle x,y\rangle < r_i\}$ *is nonempty and locally $^\perp M$-equicontinuous for the $w(X^*,X)$-topology.*

Then aff dom$P$ *is closed,* $P(\cdot) \uparrow$ aff dom$P$ *is continuous at* x_0 *for the induced topology on* aff dom$P$, *and (15) holds.*

Proof. We may assume that K is balanced and contains N^0 by replacing K with co bal$(K \cup N^0) = {}^0 (K^0 \cap -K^0 \cap N \cap -N)$. Define $g(\cdot)$ as in (16). We first show that dom$P$ = dom$^* g$. Now

$$\text{dom}P = \bigcap_{u \in U} [\text{dom}F(u, \cdot) - Lu] = \bigcup_{u \in \text{span}K} [\text{dom}F(u, \cdot) - Lu] \subset \text{dom}^* g$$

by Lemma 2. But also by Lemma 2 we have $P(x) \leq ({}^*g)(x)$ for every $x \in X$ (since dom$P \subset$ dom$^* g$), so dom$P \supset$ dom$^* g$ and hence dom$P$ = dom$^* g$.

This also implies that cldom$^*(P^*)$ = cldom$^* g$, since cldom$^*(P^*)$ = cldom$P$ by Lemma 1 (note $P^* \not\equiv +\infty$ since P^* has a nonempty level set by hypothesis 2). Hence by Definition 1 of recession functions we have $(P^*)_\infty = g_\infty - (({}^*g)^*)_\infty$. A straightforward calculation using Proposition 2, and the fact that $P^*(y) = F^*(-L^*y, y)$ yields

$$g_\infty(y) = (P^*)_\infty y = (F^*)_\infty(-L^*y, y) \quad .$$

Now $M = \{y \in X^* \colon g_\infty(y) - \langle x_0, y\rangle \leq\} = [\text{dom}g - x_0]^-$ is a subspace, hence $M = [\text{dom}g - x_0]^\perp$ and $x_0 + {}^\perp M$ is a closed affine set containing dom$g$. But hypothesis (2) then implies that riepi$^* g \neq \emptyset$ and aff dom$g$ is closed with finite codimension in $x_0 + {}^\perp M$, by Theorem 1. Moreover, by Theorem 3, ${}^*g(\cdot)$ is actually relatively continuous at x_0. Now ${}^\perp M = {}^\perp([\text{dom}g - xc]^\perp) = \text{cl aff dom}^* g - x_0$; since aff dom$^* g$ is a closed subset of $x_0 + {}^\perp$ = cl aff dom$^* g$, we must have aff dom$^* g$ = cl aff dom$^* g$. Finally, since dom$P$ = dom$^* g$ and $P \leq^* g$, $P(\cdot)$ is bounded above on a relative neighborhood of x_0 and hence is relatively continuous at x_0. ■

We shall be interested in two very useful special cases. One is when U is the dual of a normed space V, and we put the $w^* = w(U, V)$ topology as the original topology on U; for then $U^* \cong V$ and the entire space U is the span of a $w(U, V)$-compact convex set (namely the unit ball in U). Hence, if $U = V^*$ where V is a normed space, and if $F(\cdot)$ is convex and $w(U \times X, V \times X^*) - \ell$sc, then conditions (1) and (2) of Theorem 6 are automatically sufficient for (1) to hold.

The other case is when X is a barrelled space, so that interior conditions reduce to core conditions for closed sets (equivalently, compactness conditions reduce to boundedness conditions in X^*). For simplicity, we consider only Frechet spaces for which it is immediate that all closed subspaces are barrelled.

Theorem 9. *Assume* $F =^* (F^*)$; $P_0 < +\infty$; X *is a Frechet space or Banach space; and there is a* $w(U, U^*$*-compact convex set* K *in* U *such that* span$K \supset \bigcup_{x \in X}$ dom$F(\cdot, x)$. *Then the following are equivalent:*

(1) aff dom$P$ *is closed; and* $x_0 \in$ rcordom$P$, *or equivalently* $F(u_0, Lu_0 + x_0 + x) < +\infty \Rightarrow \exists \epsilon > 0$ *and* $u_1 \in U$ st $F(u_1, Lu_1 + x_0 - \epsilon x) < +\infty$.

(2) $\{y \in X^*\colon (F^*)_\infty(-L^*y, y) - \langle x_0, y\rangle \leq 0\}$ *is a subspace* M*; and there exists a* $m(U^*, U)$*-0-neighborhood* N *in* U^**, an* $x_1 \in X$*, an* $r_1 \in R$ *such that* $\{y \in X^*\colon \inf_{v\in N} F^*(v - L^*y, y) - \langle x_0, y\rangle < r_1\}$ *is nonempty and* $w(X^*, X)/M$*-locally bounded.*

If either of the above holds, then $P(\cdot) \uparrow \operatorname{aff}\operatorname{dom}P$ *is continuous at* x_0 *for the induced metric topology on* $\operatorname{aff}\operatorname{dom}P$ *and (1) holds.*

Proof. We first note that since $\operatorname{span}K \supset \bigcup_{x\in X} \operatorname{dom}F(\cdot, x)$ we have as in Theorem 8 that $\operatorname{dom}P = \operatorname{dom}^* g$ and $g_\infty(y) = (P^*)_\infty(y) = (F^*)_\infty(-L^*y, y)$.

(1) => (2). We show that $g(\cdot)$ is relatively continuous at x_0, and then (2) will follow. Now $\operatorname{dom}P = \operatorname{dom}^* g$, so $x_0 \in \operatorname{rcordom}P$. Let $W = \operatorname{aff}\operatorname{dom}P - x_0$ be the closed subspace parallel to $\operatorname{dom}P$, and define $h\colon W \to \overline{R}\colon w \to {}^* g(x_0 + w)$. Since ${}^* g$ is ℓsc on X, h is ℓsc on the barrelled space W. But $0 \in \operatorname{coredom}h$ (in W), hence h is actually continuous at 0 (since W is barrelled), or equivalently ${}^* g$ is relatively continuous at x_0. Applying Theorem 3 we now see that M is the subspace $W^\perp$; the remainder of (2) then follows from Theorem 1, since $g(y) = \inf_{v\in N} F^*(v - L^*y, y) \geq ({}^* g)^*(y)$.

(2) => (1). Note that ${}^\perp M$ is a Frechet space in the induced topology, so $w(X^*, X)/M$-local boundedness is equivalent to $w(X^*, X)/M$-local compactness. But now we may simply apply Theorem 8 to get $P(\cdot)$ relatively continuous at x_0 and $\operatorname{aff}\operatorname{dom}P$ closed; of course, (1) follows. ∎

Corollary 2. *Assume* $P_0 < +\infty$*;* $U = V^*$ *where* V *is a normed space;* X *is a Frechet space or Banach space;* $F(\cdot)$ *is convex and* $w(U \times X, V \times X^*) - \ell$sc*. Then the following are equivalent:*

(1) $x_0 \in \operatorname{cordom}P \equiv \operatorname{cor}\bigcup_{u\in U}[\operatorname{dom}F(u, \cdot) - Lu]$

(2) $\{y \in X^*\colon (F^*)_\infty(-L^*y, y) - \langle x_0, y\rangle \leq 0\} = \{0\}$*; and there is an* $\epsilon < 0$*, an* $x_1 \in X$*, and* $r_1 \in R$ *such that*

$$\{y \in X^*\colon \inf_{|v|\leq\epsilon} F^*(v - L^*y, y) - \langle x_0, y\rangle < r_1\}$$

is nonempty and $w(X^*, X)$*-locally bounded.*

(3) There is an $\epsilon > 0$*, an* $r_0 \in R$ *such that*

$$\{y \in X^*\colon \inf_{|v|\leq\epsilon} F^*(v - L^*y, y) - \langle x_0, y\rangle < r_0\}$$

is nonempty and $w(X^*, X)$*-bounded.*

If any of the above holds, then $P(\cdot)$ *is continuous at* x_0 *and (1) holds.*

Proof. Immediate. ∎

We can also apply these theorems to perturbations on the dual problem to get existence of solutions to the original problem P_0 and no duality gap $P_0 = D_0$.

Corollary 3. *Assume $P_0 > -\infty$; $U = V^*$ where V is a Frechet space or Banach space; X is a normed space; $F(\cdot)$ is convex and $w(U \times X, V \times X^*) - \ell$sc. Suppose $\{u \in U\colon F_\infty(u, Lu + x_0) \leq 0\}$ is a subspace M, and there is an $\epsilon > 0$, an $x_1 \in X$, an $r_1 \in R$ such that*

$$\{u \in U\colon \inf_{|x| \leq \epsilon} F(u, Lu + x_0 + x) < r_1\}$$

is nonempty and $w(U, U^*)/M$-locally compact. Then $P_0 = D_0$ and P_0 has solutions.

Proof. Apply Theorem 9 to the dual problem (12). ■

References

1. Schäfer, H.H.: Topological Vector Space. Springer, Berlin-New York (1971)
2. Schatten, R.: Norm Ideals of Completely Continuous Operators. Springer, Berlin-New York (1960)
3. Treves, F.: Topological Vector Spaces, Distributions and Kernels. Academic Press, New York (1967)
4. Ekeland, I., Teman, R.: Convex Analysis and Variational Problems. North Holland, Amsterdam (1976)
5. Rockafellar, R.T.: Conjugate Duality and Optimization. In: Conference Board of Math. Sci. Series, vol. 16, SIAM Publications (1974), Much of the material presented in this paper represents joint work with S.K. Young as reported in his unpublished doctoral thesis: Convexity and Duality in Optimization Theory, Ph.D. dissertation, Mathematics Department, MIT (1977)
6. Mitter, S.K., Young, S.K.: Integration with respect to Operator Valued Measures with Applications to Quantum Estimation Theory. Annali di Matematica Pura ed Applicata, vol. CXXXVII, pp. 1–39 (1984); and the references cited there. The section on Quantum Detection has not been published before. There has been considerable recent activity in Infinite-Dimensional Mathematical Programming
7. Anderson, E.J., Nash, P.: Linear Programming in Infinite Dimensional Spaces. John Wiley & Sons, New York (1987)
8. Fleming, W.H., Soner, H.M.: Controlled Markov Processes and Viscosity Solutions. Springer, Berlin-New York (1993)
9. Villani, C.: Topics i Optimal Transportation of Graduate Studies in Mathematics, vol. 58. American Mathematical Society, Providence, RI (2003)

The Servomechanism Problem for SISO Positive LTI Systems

Bartek Roszak and Edward Davison

Systems Control Group, Dept of Electrical and Computer Engineering,
University of Toronto, Toronto, Ontario M5S 1A4, Canada
{bartek,ted}@control.utoronto.ca

Summary. In this paper, we study the servomechanism problem for positive LTI systems. In particular, we consider the robust servomechanism problem of nonnegative constant reference signals for stable SISO positive LTI systems with constant disturbances, for the case when the plant model of the system is not known.

A positive LTI system is an LTI system with the imposed constraint that the state, output and/or input variables be nonnegative for all time. This type of constraint is quite natural to apply, since the property of nonnegativity occurs quite frequently in practical applications and in nature, e.g. in pharmacodynamic applications, infusion of vasoactive drugs, chemical systems. A special class of positive systems that is of significant importance and appears quite frequently in the literature is the class of compartmental systems; these systems are composed of a finite number of storage devices or reservoirs and model various gas, fluid, and solid exchange networks.

Positive systems have been of great interest to numerous researchers over several decades. Interesting results obtained on positive systems span the topics of reachability and controllability, realization theory for positive systems, stability control, and control via pole-assignment. However, to date, there appears that no study has been carried out on the servomechanism problem for positive LTI systems; thus we focus our attention on this topic in the paper. Various applications of the results are applied to a number of positive system problems obtained from the literature.

Keywords: positive systems, robust servomechanism problem, tuning regulators, nonnegativity, unknown systems.

1 Introduction

In this paper we study the robust servomechanism problem for positive linear systems. In particular, we consider the tracking and regulation problem of nonnegative constant reference signals for unknown stable SISO positive linear systems with nonnegative constant disturbances.

Positive systems and their counterparts, compartmental systems, have been of great interest to numerous researchers, and some interesting results on positive systems have been recorded in the literature; we point the interested readers to [1], and the references therein, for numerous citations.

The paper is organized as follows. Preliminaries are given first, where the terminology, positive systems, *almost* positive systems, and tuning regulators

V.D. Blondel et al. (Eds.) Recent Advances in Learning and Control, LNCIS 371, pp. 181–193, 2008.
springerlink.com

are discussed. All assumptions on the system plant treated in the paper and the Problem Statement are described in Section 3. Section 4 provides the main results of the paper. Finally, several examples illustrating the theory are given in Section 5.

2 Background and Preliminaries

2.1 Terminology

Let the set $\mathbb{R}_+ := \{x \in \mathbb{R} \mid x \geq 0\}$, the set $\mathbb{R}_+^n := \{x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n \mid x_i \in \mathbb{R}_+, \ \forall i = 1, ..., n\}$. If exclusion of 0 from the sets will be necessary, then we'll denote the sets in the standard way $\mathbb{R}_+^n \setminus \{0\}$. The set of eigenvalues of a matrix $\mathcal{A}$ will be denoted as $\sigma(\mathcal{A})$. The ij^{th} entry of a matrix $\mathcal{A}$ will be denoted as a_{ij}. A *nonnegative* matrix $\mathcal{A}$ has all of its entries greater or equal to 0, $a_{ij} \in \mathbb{R}_+$. A *Metzler* matrix $\mathcal{A}$ is a matrix for which all off-diagonal elements of $\mathcal{A}$ are nonnegative, i.e. $a_{ij} \in \mathbb{R}_+$ for all $i \neq j$. A *compartmental* matrix $\mathcal{A}$ is a matrix that is Metzler, where the sum of the components within a column is less than or equal to zero, i.e. $\sum_{i=1}^{n} a_{ij} \leq 0$ for all $j = 1, 2, ..., n$.

In this section we give an overview of *positive linear systems* [2], [3], and point out a key subset of positive systems known as *compartmental systems* [3], [4].

We first define a positive linear system [3] in the traditional sense.

Definition 1. *A linear system*

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y &= Cx + Du \end{aligned} \tag{1}$$

where $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $C \in \mathbb{R}^{r \times n}$, *and* $D \in \mathbb{R}^{r \times m}$ *is considered to be a* positive linear system *if for every nonnegative initial state and for every nonnegative input the state of the system and the output remain nonnegative.*

The above definition states that any trajectory starting at an initial condition $x_0 \in \mathbb{R}_+^n$ will not leave the positive orthant, and moreover, that the output also remains nonnegative. For convenience, if for all time a state x satisfies $x \in \mathbb{R}_+^n$, the output y satisfies $y \in \mathbb{R}_+^r$, or input u satisfies $u \in \mathbb{R}_+^m$, then we'll say that the state, output, or input maintains *nonnegativity.* Notice that Definition 1 states that the input to the system must be nonnegative, a restriction that in applications is not always feasible.

It turns out that Definition 1 has a very nice interpretation in terms of the matrix quadruple (A, B, C, D).

Theorem 1 ([3]). *A linear system* (1) *is positive if and only if the matrix A is a Metzler matrix, and B, C, and D are nonnegative matrices.*

In real life systems, nonnegativity of states occurs quite often; however, the need for the input u to be also nonnegative is not a necessity, as was also pointed out in [5]. For this reason we introduce the following definitions.

Definition 2. *An arbitrary linear system is considered to be a* state positive linear system *if there exists a nonnegative initial state x_0 and an input u such that the state of the system remains nonnegative for all time.*

An equivalent definition can be given for *output positive linear systems.*

Definition 3. *An arbitrary linear system is considered to be an* output positive linear system *if there exists a nonnegative initial state x_0 and an input u such that the output of the system remains nonnegative for all time.*

An interesting subset of positive systems is that of compartmental systems. A compartmental system consists of n interconnected compartments [4] or reservoirs. The main mathematical distinction, for LTI systems, between a positive system and a compartmental system is that a positive system's A matrix is Metzler, while a compartmental system's A matrix is compartmental. The inclusion of compartmental systems is made because in general compartmental systems are stable, a property of great significance throughout the paper. For a more complete study and interesting results on compartmental systems see [4] and references therein.

2.2 "Almost" Positive Systems

In this subsection, we state several definitions that will lead us to the concept of *almost* positivity, which will be of great importance in Section 4.

Definition 4. *An arbitrary linear system is considered to be a* δ-state (δ-output) positive linear system *with respect to x_0 and u if for a given $\delta = (\delta_1, \delta_2, ..., \delta_{n(r)}) \in \mathbb{R}_+^{n(r)} \setminus \{0\}$, the state x (output y) of the system satisfies*

$$x_i(t)(y_i(t)) \geq -\delta_i, \quad \forall i = 1, 2, ..., n(r), \quad \forall t \in [0, \infty)$$

In the above definition there is no restriction on how small or large the components of δ should be. It will be of interest, however, to study the case when each component $\delta_i \to 0$, for all i.

Definition 5. *An arbitrary linear system is considered to be an* almost-state (output) positive linear system *(or simply* almost-state (output) positivity*) with respect to x_0 and u if for any given $\delta = (\delta_1, \delta_2, ..., \delta_{n(r)}) \in \mathbb{R}_+^{n(r)} \setminus \{0\}$, the state x (output y) of the system satisfies*

$$x_i(t)(y_i(t)) \geq -\delta_i, \quad \forall i = 1, 2, ..., n(r), \quad \forall t \in [0, \infty)$$

It is worth pointing out that in a practical setting, and even in the theoretical one, the notion of the number zero has been often questioned, i.e. when is a number small enough to consider it to be zero; with the above remark in mind it would appear that if δ can be made arbitrarily small, then one could argue that any *almost* state (output) positive system is basically a state (output) positive system.

2.3 Tuning Regulators

In this section we describe a particular compensator, known as the tuning controller or tuning regulator, which solves the servomechanism tracking and regulation problem for *unknown*[1] stable linear systems under unknown constant disturbances. Such unknown systems often occur in industrial application problems. The results of this section can be found in their entirety and in their general form in [6, 7].

Consider the plant

$$\begin{aligned} \dot{x} &= Ax + Bu + E\omega \\ y &= Cx + Du + F\omega \\ e &:= y - y_{ref} \end{aligned} \tag{2}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $y \in \mathbb{R}^r$, the unmeasurable disturbance vector $\omega \in \mathbb{R}^{\bar{\Omega}}$, and $y_{ref} \in \mathbb{R}^r$ is a desired tracking signal. Assume that the output y is measurable, that the matrix A is Hurwitz, $m = r$ ([6] does not assume $m = r$, but for the purpose of this paper this causes no loss of generality), and that the disturbance vector and tracking signal satisfy:

$$\begin{aligned} \dot{z}_1 &= \mathcal{A}_1 z_1 \\ \omega &= \mathcal{C}_1 z_1, \end{aligned} \quad \text{and} \quad \begin{aligned} \dot{z}_2 &= \mathcal{A}_2 z_2 \\ \sigma &= \mathcal{C}_2 z_2 \\ y_{ref} &= G\sigma, \end{aligned}$$

respectively, where $z_i \in \mathbb{R}^{n_i}$, $(\mathcal{C}_i, \mathcal{A}_i)$ is observable for $i = 1, 2$, $z_1(0)$ may or may not be known, and $z_2(0)$ is known. Then in the case of constant disturbances (ω) and constant tracking (y_{ref}) signals, the tuning regulator that solves the robust servomechanism problem[2], i.e. such that (i) the closed loop system is stable, (ii) for all tracking signals and disturbances $e \to 0$ as $t \to \infty$, and (iii), property (ii) occurs for all plant perturbations which maintain closed loop stability, is given by:

$$\begin{aligned} \dot{\eta} &= y - y_{ref} \\ u &= -\epsilon(D - CA^{-1}B)^{-1}\eta \end{aligned} \tag{3}$$

where $\epsilon \in (0, \epsilon^*]$, $\epsilon^* \in \mathbb{R}_+ \setminus \{0\}$.

We summarize the above discussion by a Theorem for the case of SISO linear systems.

Theorem 2 ([6]). *Consider the system* (2), *under the assumption that* $y_{ref} \in \mathbb{R}$ *and* $\omega \in \mathbb{R}$ *are constants. Then there exists an* ϵ^* *such that the tuning regulator* (3) *achieves robust tracking and regulation if and only if* $rank(D - CA^{-1}B) = r$.

We refer the interested reader to [6] for the procedure of experimentally obtaining the gain matrix $(D - CA^{-1}B)^{-1}$, for the case of unknown plant models (2). It is to be noted that if $rank(D - CA^{-1}B) = r$, then "on-line tuning" [6] is used to find an optimal value of ϵ in the controller (7).

[1] By unknown we mean that there is no knowledge of (A, B, C, D).

[2] Referred to in [6] as the "robust control of a general servomechanism problem".

3 Problem Statement and System Assumptions

In this section we provide the details of the plant, all accompanying assumptions made on the plant, and the problem of interest.

Throughout this paper we consider the following unknown plant:

$$\begin{aligned} \dot{x} &= Ax + bu + e_\omega \omega \\ y &= cx + du + f\omega \\ e &:= y - y_{ref} \end{aligned} \tag{4}$$

where A is an $n \times n$ Metzler stable matrix, $b \in \mathbb{R}_+^n$, $c \in \mathbb{R}_+^{1\times n}$, $d = 0$, $y_{ref} \in Y_{ref} \subset \mathbb{R}_+$, and $\omega \in \Omega \subset \mathbb{R}$ such that $e_\omega \omega \in \mathbb{R}_+^n$, $f\omega \in \mathbb{R}_+$.

Next, we provide an important assumption which will be commonly used in the sequel. This assumption is needed in order to ensure that the steady state values of the closed loop system will be nonnegative, under the choice of the reference signals and the disturbances of the plant. If this assumption was not true, then clearly we cannot attempt to satisfy any sort of nonnegativity of the states.

Assumption 1. *Given* (4) *assume that* $rank(d - cA^{-1}b) = 1$ *and that the sets* Ω *and* Y_{ref} *are chosen such that the steady state values of the plant's states are nonnegative, i.e. for all tracking and disturbance signals in question, it is assumed that the steady-state of the system* (4) *given by*

$$\begin{bmatrix} x_{ss} \\ u_{ss} \end{bmatrix} = -\begin{bmatrix} A & b \\ c & d \end{bmatrix}^{-1} \begin{bmatrix} e_\omega & 0 \\ f & -1 \end{bmatrix} \begin{bmatrix} \omega \\ y_{ref} \end{bmatrix} \tag{5}$$

has the property that $x_{ss} \in \mathbb{R}_+$. *It is to be noted that the identified inverse exists if and only if* $rank(d - cA^{-1}b) = 1$.

With the above plant and assumption given, we outline the main problem of interest.

Problem 1. Consider the unknown plant (4) under Assumption 1. Find a controller u that

(a) guarantees closed loop stability,
(b) ensures tracking and regulation of the output y, i.e. $e = y - y_{ref} \to 0$, as $t \to \infty$, $\forall y_{ref} \in Y_{ref}$ and $\forall \omega \in \Omega$, and
(c) ensures the plant (4) can be made state and an output positive system.
(d) For all perturbations of the nominal plant model, which do not de-stabilize the open loop system and do not change the positivity of the system (i.e. A is still Metzler and b, c, $d = 0$ are still nonnegative matrices), it is desired that the controller can still achieve asymptotic error regulation and state and output positivity, i.e. properties (a), (b) and (c) should still be attainable if the system model is allowed to perturb.

Notice that because the output matrix (c) is nonnegative, then the output will be nonnegative if state nonnegativity is ensured. Thus, the problem of interest boils down to state nonnegativity only.

Remark 1. In the sequel when almost-state and almost-output positivity will be considered, then in Problem 1 the words state and output should be replaced by almost-state and almost-output, respectively.

4 Main Results

In this section, we present the main results of the paper.

We first point out an interesting lemma that relates the tracking reference signal and the steady state output (with no input) to the steady state of the input.

Lemma 1. *Consider system* (4). *Let* $\overline{y}_{ss} = c\overline{x}_{ss}$, *where* $\overline{x}_{ss} = -A^{-1}e_{\omega}\omega$. *Then,*

$$\begin{aligned} y_{ref} - \overline{y}_{ss} > 0 &\iff u_{ss} > 0 \\ y_{ref} - \overline{y}_{ss} < 0 &\iff u_{ss} < 0 \\ y_{ref} - \overline{y}_{ss} = 0 &\iff u_{ss} = 0 \end{aligned} \tag{6}$$

Notice that from the above Lemma, if $u_{ss} = 0$ then in order to track y_{ref} it suffices to have no input, i.e. $y \to \overline{y}_{ss} = y_{ref}$ as $t \to \infty$, if $u = 0$ for (4). Next, we'll take the latter remark into account by concentrating first on the case when $u_{ss} > 0$ and secondly when $u_{ss} < 0$. We'll then finish this section by presenting results on *almost* positivity with and without an input bias.

Throughout the remainder of this section, we'll make great use of the tuning regulator [1] defined by

$$\dot{\eta} = y - y_{ref}, \quad \eta_0 = 0, \quad u = -\epsilon\eta \tag{7}$$

where $\epsilon > 0$. The reasons for using (7) instead of (3) has been justified in [1]. However, we point out that we still need Assumption 1 in order to use the above regulator.

We are now ready to introduce the first main result of this section.

Theorem 3. *Consider system* (4) *under Assumption 1 and with initial condition* $x_0 = 0$, *and assume the steady-state input* (u_{ss}) *value is positive; then there exists an* ϵ^* *such that for all* $\epsilon \in (0, \epsilon^*]$ *the tuning regulator* (7) *solves Problem 1, and moreover, the input to the system remains nonnegative for all time* $t \geq 0$.

The above result states that if the system is initially at rest, Assumption 1 is satisfied, and the steady-state value of the input is positive, then the tuning regulator will solve the servomechanism problem outlined by Problem 1 with a nonnegative control input.

Our next result shifts to the case when the steady-state value of the input is negative. In this case we clearly cannot obtain nonnegative control; however, we will point out that state and output nonnegativity can still be obtained for systems that are initially at rest.

Theorem 4. *Consider system* (4) *under Assumption 1 and with initial condition* $x_0 = 0$, *and assume the steady-state input* (u_{ss}) *value is negative, the steady-state of the system* (4) *with* $u = 0$ *is positive* ($\overline{x}_{ss} > 0$, *component-wise), and* $y_{ref} \in Y_{ref} \setminus \{0\}$; *then there exists an* ϵ^* *such that for all* $\epsilon \in (0, \epsilon^*]$ *the tuning regulator* (7) *solves Problem 1.*

In the latter Theorem we have added two extra assumptions that Theorem 3 did not have. Namely, the first extra assumption is that the steady state $\overline{x}_{ss}$ must be positive component-wise. The meaning of this assumption is that for the system:

$$\dot{x} = Ax + e_\omega \omega$$

$x \to x_{ss} = -A^{-1} e_\omega \omega > 0$ component-wise, as $t \to \infty$. The second assumption on the tracking reference signal is self-explanatory.

We are now ready to shift to the case where the initial conditions on the system are not necessarily at rest. For the rest of this section, we will consider Problem 1 under Remark 1, i.e. we will consider *almost* positivity only.

Theorem 5. *The tuning regulator* (7) *solves Problem 1 under Remark 1 and Assumption 1 regardless of the initial condition* $x_0 \in \mathbb{R}^n_+$.

The main difference between the latter Theorem and the previous two Theorems is the fact that under Assumption 1 Theorem 5 can satisfy *almost*-state and *almost*-output positivity regardless of the constraint on the input and the initial condition. We can further extend this result to the case where there is an input-bias, i.e. in the next result consider the plant to be defined by

$$\begin{aligned} \dot{x} &= Ax + b(u + u_b) + e_\omega \omega \\ y &= cx + f\omega \\ e &:= y - y_{ref}, \end{aligned} \tag{8}$$

where $u_b \in \mathbb{R}_+$ is an input bias into the system that causes a shift in the initial condition of the system.

Corollary 1. *Consider the plant* (8) *where* $e_\omega \omega \in \mathbb{R}^n$ *and* $f\omega \in \mathbb{R}$. *The tuning regulator* (7) *solves Problem 1 under Remark 1 and Assumption 1 (with* u *replaced by* $u + u_b$*) regardless of the initial condition* $x_0 \in \mathbb{R}^n_+$ *if the system* (8) *with* $u = 0$ *is state and output nonnegative.*

Intuitively the above Corollary states that initially the states and the output of the system are biased to some positive value by the constant inflow of u_b, allowing for arbitrary disturbances. We will illustrate this case in one of our examples in the next section.

5 Examples

Here we present several examples illustrating the results of the previous section.

Example 1 (Illustration of Theorem 4). The system considered in this example has been taken from [8].

The interior temperature of an electrically heated oven is to be controlled by varying the heat input u to the jacket. Let the heat capacities of the oven interior and of the jacket be c_2 and c_1, respectively, let the interior and exterior jacket surface areas be a_1 and a_2, and let the radiation coefficient of the interior and exterior jacket surfaces be r_1 and r_2. If the external temperature is T_0, the jacket temperature T_1 and the oven interior temperature is T_2, then the behaviour for the jacket is described by:

$$c_1\dot{T}_1 = -a_2r_2(T_1 - T_0) - a_1r_1(T_1 - T_2) + u + a_3\omega$$

where ω is a disturbance, and for the oven interior:

$$c_2\dot{T}_2 = a_1r_1(T_1 - T_2)$$

By setting the state variables to be the excess of temperature over the exterior

$$x_1 := T_1 - T_0 \quad \text{and} \quad x_2 := T_2 - T_0$$

results in the system:

$$\dot{x} = \begin{bmatrix} \dfrac{-(a_2r_2 + a_1r_1)}{c_1} & \left(\dfrac{a_1r_1}{c_1}\right) \\ \left(\dfrac{a_1r_1}{c_2}\right) & \left(\dfrac{-a_1r_1}{c_2}\right) \end{bmatrix} x + \begin{bmatrix} 1/c_1 \\ 0 \end{bmatrix} u + \begin{bmatrix} a_3/c_1 \\ 0 \end{bmatrix} \omega$$

Assume that the values of the constants above are chosen such that $c_1 = c_2 = 1$, $a_1 = a2 = a_3 = 1$ and $r_1 = r_2 = 1$, and that the disturbance vectors are $e_\omega\omega = [3\ 0]^T$ and $f\omega = 0$ then:

$$\dot{x} = \begin{bmatrix} -2 & 1 \\ 1 & -1 \end{bmatrix} x + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u + \begin{bmatrix} 3 \\ 0 \end{bmatrix} \omega$$

and the output equation is

$$y = [1\ \ 0]x.$$

We will now show that if the temperature excess in the jacket is initially zero, then we can maintain nonnegativity of the states and output, along with obtaining tracking control by using the tuning regulator. In this example we'll assume that we want to regulate y to $y_{ref} = 1$, i.e. the desired temperature excess should be 1°C higher in the jacket than the outside.

First, it is easy to show that the A matrix is stable, i.e. $\sigma(A) = \{-2.618, -0.382\}$, and that the assumptions of Theorem 4 hold; in particular, that $\bar{x}_{ss} > 0$ and $u_{ss} < 0$. In this case, the application of controller (7) with $\epsilon = 0.1$ will suffice. Figure 1 and Figure 2 illustrate the states x and the input u, for the case of $y_{ref} = 1$, with zero initial conditions.

The second example describes the monitoring and controlling the depth of anesthesia in surgery. The problem has been originally considered by Haddad et al and has been presented in its entirety in [9].

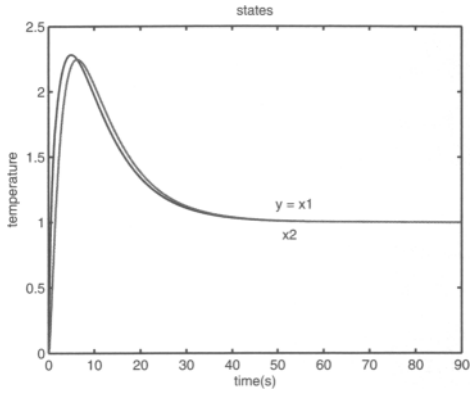


Fig. 1. State response for Example 1

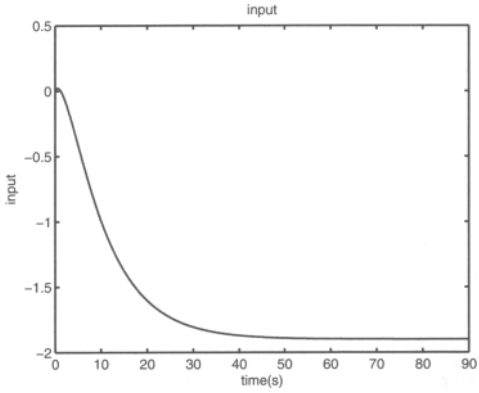


Fig. 2. Input response for Example 1

Example 2 (Illustration of Theorem 3 and Corollary 1)
The use of propofol as an intravenous anesthetic is common for both induction and maintenance of general anesthesia [10]. An effective patient model for the disposition of propofol is based on the three-compartmental mammillary model, see Figure 3 [11], [12]. The three-compartmental mammillary system provides a pharmacokinetic model for a patient describing the distribution of propofol into the central compartment and the other various tissue groups of the body. The mass balance equations for the compartmental system yields [9]:

$$\begin{aligned} \dot{x}_1 &= -(f_{01} + f_{21} + f_{31})x_1 + f_{12}x_2 + f_{13}x_3 + u \\ \dot{x}_2 &= f_{21}x_1 - f_{12}x_2 \\ \dot{x}_2 &= f_{21}x_1 - f_{12}x_2 \end{aligned} \tag{9}$$

where the states are masses in *grams* of propofol in the respective compartments. The input u is the infusion rate in $grams/min$ of the anesthetic propofol into the first

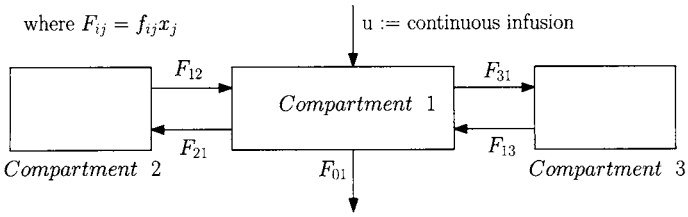


Fig. 3. Three compartmental mammillary model

compartment. The rate constant $f_{11} \geq 0$ is in min^{-1} and represents the elimination rate from the central compartment, while the rate constants $f_{ij} \geq 0$, which are also in min^{-1}, characterize drug transfer between compartments. It has been pointed out in [9] that the rate constants, although nonnegative, can be uncertain due to patient gender, weight, pre-existing disease, age, and concomitant medication. Note that it has been pointed out in [9], [13] that 2.5-6 $\mu g/ml$ blood concentration levels of propofol are required during the maintenance stage in general anesthesia.

In [9] the assumption made was that a 70 kg patient was treated with propofol concentration levels of $4\mu g/mol$, which led to the desired tracking value for $x_1 = 44.52mg$. It has also been pointed out that the values of f_{ij} may be uncertain; this however causes no problem since a mathematical model of the system is not required.

Our system matrices for (9) become:

$$A = \begin{bmatrix} -(f_{01} + f_{21} + f_{32}) & f_{12} & f_{31} \\ f_{21} & -f_{12} & 0 \\ f_{31} & 0 & -f_{13} \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

For our simulations, we will use the the parameters in Table 1, presented in [14]. The table presents two sets of data; in order to show that our controller works for uncertain systems, we will alternate between the two sets.

Table 1. Pharmacokinetic parameters [14]

Data	f_{01}	f_{21}	f_{12}	f_{31}	f_{13}
1	0.152	0.207	0.092	0.040	0.0048
2	0.119	0.114	0.055	0.041	0.0033

In addition to the model above, we assume that a disturbance exists, and that it affects the input to the first compartment, see Figure 4. With the disturbance in place, our system model becomes:

$$\dot{x} = \begin{bmatrix} -(f_{01} + f_{21} + f_{32}) & f_{12} & f_{31} \\ f_{21} & -f_{12} & 0 \\ f_{31} & 0 & -f_{13} \end{bmatrix} x + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} u + \begin{bmatrix} e_\omega \\ 0 \\ 0 \end{bmatrix} \omega, \quad y = [1\ 0\ 0]x.$$

The simulation, Figure 5 and Figure 6 show the response of $y = x_1$ and u with $\epsilon = 0.1$, and where at $t \in [0, 50min)$ data 1, from Table 1, is used with

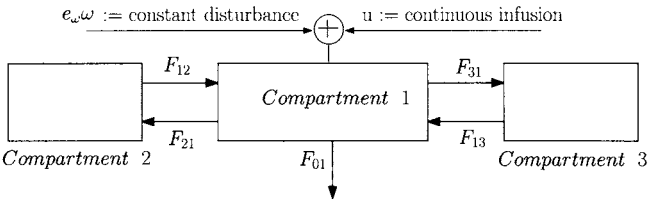


Fig. 4. Three compartmental mammillary model with disturbance

$e_\omega\omega = 0.5$ (by Theorem 3); at $t = 50min$ the system switches to data 2 with $e_\omega\omega = 0.5$ (by Corollary 1), and finally at $t = 100min$ the system undergoes further disturbance with $e_\omega\omega = 1.5$ (by Corollary 1).

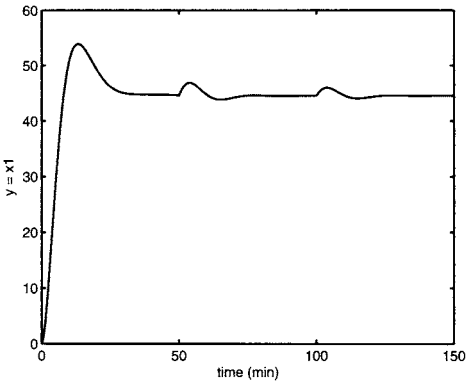


Fig. 5. Output response for Example 2

The last example illustrates a system of reservoirs network.

Example 3 (Illustration of Theorem 3). The following system, which is compartmental, has been taken from [3] pg.105. Consider the reservoirs network of Figure 7. The system is of dimension 6, as we can assume the pump dynamics can be neglected. As pointed out in [3], the dynamics of each reservoir can be captured by a single differential equation:

$$\dot{x}_i = -\alpha_i x_i + v, \qquad z = \alpha_i x_i$$

for all $i = 1, ..., 6$, with $\alpha > 0$, and where x_i is the water level (L) in the reservoir.

Consider now the case when $\gamma = 0.5$, $\phi = 0.7$, $\alpha_1 = 0.8$, $\alpha_2 = 0.7$, $\alpha_3 = 0.5$, $\alpha_4 = 1$, $\alpha_5 = 2$, and $\alpha_6 = 0.8$. Note that all the rates are measured in L/s. With the variables defined above we obtain the following system:

$$\dot{x} = \begin{bmatrix} -0.8 & 0 & 0 & 0 & 2 & 0 \\ 0 & -0.7 & 0 & 0 & 0 & 0 \\ 0.8 & 0.7 & -0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.15 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -2 & 0 \\ 0 & 0 & 0.35 & 0 & 0 & -0.8 \end{bmatrix} + \begin{bmatrix} 0.5 \\ 0.5 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} u, \quad y = [0\ 0\ 0\ 0\ 0\ 1]x.$$

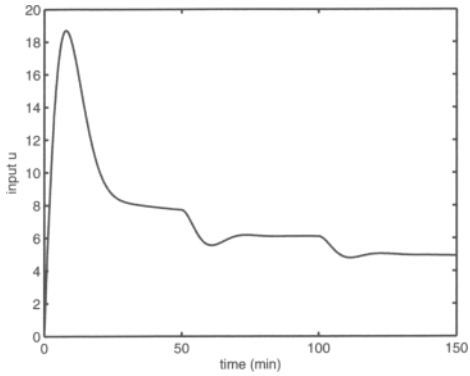


Fig. 6. Control input for Example 2

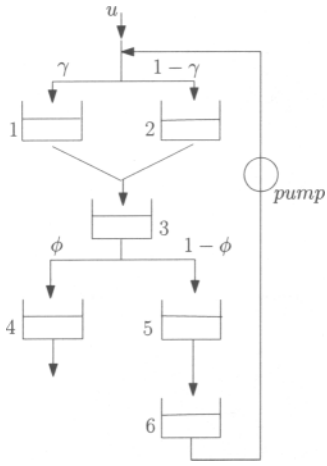


Fig. 7. System set up for Example 3

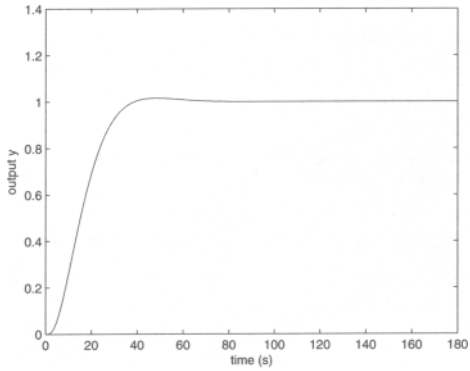


Fig. 8. Output response for Example 3

It is easy to check that the above compartmental system is stable, since it has $\sigma(A) = \{-0.8,\ -0.2112,\ -0.9924 \pm 0.5249i,\ -2.1039,\ -0.7000\}$. Assume now that the initial condition $x_0 = 0$, and that we would like to track the reference $y_{ref} = 1$. Now since $\omega = 0$, then by [1] $u_{ss} > 0$ and Theorem 3 can be used. The application of controller (7) with $\epsilon = 0.05$, solves the tracking problem, as in the previous example. Figure 8 illustrates the output y response. The plots of the states x and input are omitted.

References

1. Roszak, B., Davison, E.J.: Tuning regulators for tracking SISO positive linear systems. In: European Control Conference 2007, Kos, Greece (July 3-5, 2007)
2. Luenburger, D.: Introduction to Dynamic Systems: Theory, Models and Applications. Wiley, New York (1979)
3. Farina, L., Rinaldi, S.: Positive Linear Systems: Theory and Applications. John Wiley & Sons, Inc., New York (2000)
4. Jacquez, J.A., Simon, C.P.: Qualitative Theory of Compartmental Systems. Society for Industrial and Applied Mathematics 35(1), 43–79 (1993)
5. De Leenheer, P., Aeyels, D.: Stabilization of positive linear systems. Systems and Control Letters 44, 259–271 (2001)
6. Davison, E.J.: Multivariable Tuning Regulators: The Feedforward and Robust Control of a General Servomechanism Problem. IEEE Transactions on Automatic Control AC-21, 35–47 (1976)
7. Miller, D.E., Davison, E.J.: The Self-Tuning Robust Servomechanism Problem. IEEE Transactions on Automatic Control 34, 511–523 (1989)
8. Barnett, S., Cameron, R.G.: Introduction to Mathematical Control Theory, 2nd edn. Clarendon Press, Oxford (1985)
9. Haddad, W.M., Hayakawa, T., Bailey, J.M.: Adaptive control for non-negative and compartmental dynamical systems with applications to general anesthesia. Int. J. Adapt. Control Signal Process. 17, 209–235 (2003)
10. Doze, V.A., Westphal, L.M., White, P.F.: Comparison of propofol with methohexital for outpatient anesthesia. Aneth. Analg. 65, 1189–1195 (1986)
11. Linkens, D.A., Abbod, M.F., Peacock, J.E.: Clinical implementation of advanced control in anaesthesia. Trans. Inst. Meas. Control 22, 303–330 (2000)
12. Anderson, D.H.: Compartmental Modeling and Tracer Kinetics. Lecture Notes in Biomathematics 50. Springer, New York (1983)
13. White, M., Kenny, G.N.C.: Intravenous propofol anaesthesia using a computerised infusion system. Anaesthesia 45, 204–209 (1990)
14. Glass, P.S., Goodman, D.K., Ginsberg, B., Reves, J.G., Jacobs, J.R.: Accuracy of phamacokinetic model-driven infusion of propofol. Anesthesiology 71, A277 (1989)

Passivity-based Stability of Interconnection Structures

Eduardo D. Sontag[1] and Murat Arcak[2]

[1] Department of Mathematics, Rutgers University, USA
sontag@math.rutgers.edu
[2] Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, USA
arcakm@rpi.edu

In the 1970s, Vidyasagar developed an approach to the study of stability of interconnected systems. This paper revisits this approach and shows how it allows one to interpret, and considerably extend, a classical condition used in mathematical biology.

1 Introduction

In Chapter 7 of his influential book *Input-Output Analysis of Large Scale Interconnected Systems* [1], Vidyasagar describes an approach to studying the stability of networks made up of passive subsystems. This approach, which was pioneered in earlier work of Sundareshan and Vidyasagar [2] and Moylan and Hill [3], relies upon verifying the diagonal stability of an associated *dissipativity matrix* which incorporates information about the passivity properties of the subsystems, the interconnection structure of the network, and the signs of the interconnection terms.

As shown in the authors' work [4, 5], diagonal stability, combined with an excess of passivity property on components (the "secant gain" in the terminology in [6]) can be used to rederive, and extend considerably, the classical "secant condition" [7, 8] for the stability of cyclic feedback systems, well-known in mathematical biology. Cyclic feedback structures have classically been used to model autoregulatory feedback loops in gene networks [9, 10, 7, 11, 8], as well as metabolic pathways [12, 13] and cell signaling [14] In this expository paper, we provide a streamlined version of the key lemma on stability of interconnections due to Vidyasagar and Moylan and Hill, and then show how it its hypotheses may be verified for network structures of great interest in biology.

2 The Key Lemma

We denote by L^2_e the extended space of signals (thought of as time functions) $w : [0, \infty) \to \mathbb{R}$ which have the property that each restriction $w_T = w|_{[0,T]}$ is in $L^2(0, T)$, for every $T > 0$. Given an element $w \in L^2_e$ and any fixed $T > 0$, we

V.D. Blondel et al. (Eds.) Recent Advances in Learning and Control, LNCIS 371, pp. 195–204, 2008.
springerlink.com

write $\|w\|_T$ for the L^2 the norm of this restriction w_T, and given two functions $v, w \in L^2_e$ and any fixed $T > 0$, the inner product of v_T and w_T is denoted by $\langle v, w \rangle_T$. The same notation is used for vector functions.

We view a family of M subsystems to be interconnected as operators

$$\Sigma_i : L^2_e \to L^2_e : u_i \mapsto y_i,$$

and impose the following strict passivity property: there exist constants $\gamma_i > 0$ ("secant gains" in [6]) such that

$$\|y_i\|^2_T \leq \gamma_i \langle y_i, u_i \rangle_T \text{ for each } i = 1, \ldots, M \text{ and each } T > 0\,. \tag{1}$$

We then consider the interconnection where

$$u_i(t) = v_i(t) + A_i y(t)\,, \tag{2}$$

or just $u = v + Av$, where the v_i's are external inputs, $y = \text{col}(y_1, \ldots, y_M)$, $v = \text{col}(v_1, \ldots, v_M)$, and the A_i, $i = 1, \ldots, M$ are the rows of an interconnection matrix $A \in \mathbb{R}^{M \times M}$. In other words, the ith subsystem receives as inputs an external input plus an appropriate linear combination of outputs from the remaining systems (including possibly feedback from itself, if the corresponding diagonal entry of A is nonzero). We introduce:

$$E := A - \Gamma$$

where

$$\Gamma = \text{diag}\left(\frac{1}{\gamma_1}, \ldots, \frac{1}{\gamma_M}\right).$$

Lemma 1. Suppose that E is diagonally stable, that is, exists a diagonal positive definite matrix $D \in \mathbb{R}^{M \times M}$ such that

$$DE + E'D < 0\,.$$

Then, the system obtained from the systems Σ_i using the interconnection matrix A is L^2 stable as a system with input v and output y. More precisely, there is some constant $\rho > 0$ such that, for any $u, v, y \in (L^2_e)^M$ such that (1) and (2) hold, necessarily $\|y\|_T \leq \rho \|v\|_T$ for all $T > 0$ (and therefore also $\|y\| \leq \rho \|v\|$, if $v \in (L^2)^M$).

Proof: We pick an $\alpha > 0$ such that $DE + E'D < -2\alpha I$, and observe that, for any $T > 0$ and any function $z \in L^2(0, T)$, it holds that:

$$\begin{aligned}
\langle Dz, Ez \rangle &= \int_0^T z(s)' DEz(s)\, ds \\
&= \int_0^T \frac{1}{2} z'(s)(DE + E'D)z(s)\, ds \\
&\leq -\alpha \int_0^T z'(s)z(s)\, ds \\
&= -\alpha \|z\|^2\,.
\end{aligned}$$

Fix an arbitrary $T > 0$, and write $D = \operatorname{diag}(d_1, \ldots, d_M)$. Since, for each i, $\langle y_i, u_i - \frac{1}{\gamma_i} y_i \rangle_T \geq 0$, it follows that also $\langle d_i y_i, u_i - \frac{1}{\gamma_i} y_i \rangle_T \geq 0$, or, in vector form:

$$\langle Dy, u - \Gamma y \rangle_T \geq 0.$$

Substituting $u = v + Ay$, we obtain $\langle Dy, v + Ey \rangle_T \geq 0$, from which, using the Cauchy-Schwartz inequality:

$$\beta \|v\|_T \|y\|_T \geq \langle Dy, v \rangle_T \geq -\langle Dy, Ey \rangle_T \geq \alpha \|y\|_T^2$$

for some $\beta > 0$. So $\|y\|_T \leq \rho \|u\|_T$, with $\rho = \frac{\beta}{\alpha}$, as desired. □

The required passivity properties can be checked through dissipation inequalities involving appropriate Lyapunov-like storage functions, as explained in [4, 5] for several classes of systems which arise in biological applications. Although the conclusion provides a purely input/output stability property, state-space global asymptotic stability results may be obtained as corollaries, by combining I/O stability with appropriate detectability and controllability conditions on subsystems, as discussed in [6]. The verifiable state-space conditions given in these papers guarantee the desired passivity properties for the subsystems. These conditions are particularly suitable for systems of biological interest because they are applicable to models with nonnegative state variables, and do not rely on the knowledge of the location of the equilibrium. The state-space approach further made it possible to prove robustness of our stability criterion in the presence of diffusion terms.

3 Recovering the Classical Secant Condition

The classical "secant condition" applies to systems that are obtained as negative feedback cycles. For such systems,

$$E = E_{cyclic} = \begin{bmatrix} -\frac{1}{\gamma_1} & 0 & \cdots & 0 & -1 \\ 1 & -\frac{1}{\gamma_2} & \ddots & & 0 \\ 0 & 1 & -\frac{1}{\gamma_3} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -\frac{1}{\gamma_n} \end{bmatrix} \tag{3}$$

and it is shown in [4] that this matrix is diagonally stable if and only if the following condition is satisfied:

$$\gamma_1 \cdots \gamma_n < \sec(\pi/n)^n. \tag{4}$$

Unlike a *small-gain* condition which would restrict the right-hand side of (4) to be 1, the "secant criterion" (4) also exploits the phase of the loop and allows the right-hand side to be as high as 8 when $n = 3$ (and infinite for $n = 1, 2$). The secant criterion is also necessary for stability of E when the γ_i's are identical.

A classical result [7, 8] in mathematical biology is that for linear systems (as well as for certain restricted classes of nonlinear feedback systems [7]) is that a matrix of the form:

$$A = \begin{bmatrix} -a_1 & 0 & \cdots & 0 & -b_n \\ b_1 & -a_2 & \ddots & & 0 \\ 0 & b_2 & -a_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & b_{n-1} & -a_n \end{bmatrix} \tag{5}$$

$a_i > 0,\ b_i > 0,\ i = 1, \cdots, n$ is A Hurwitz if the following sufficient condition holds:

$$\frac{b_1 \cdots b_n}{a_1 \cdots a_n} < \sec(\pi/n)^n.$$

The matrix A can be interpreted as the closed-loop matrix for a cyclic interconnection of the linear systems with transfer functions $b_i/(s + a_i)$, which are passive with $\gamma_i = \frac{b_i}{a_i}$. Thus, this result is an immediate consequence of Lemma 1 (expressed in state space terms).

4 Branched Structures

It is of interest to ask for characterizations of diagonal stability for the matrices E corresponding to other interconnection structures. In this section, we review recent results from [5] that analyze certain branched structures.

A common form of feedback inhibition in metabolic networks occurs when several end metabolites in different branches of a pathway inhibit a reaction located before the branch point [13, 15]. As an example of this situation we consider the network in Figure 1 where the end metabolites with concentrations x_4 and x_6 inhibit the formation of x_1 from an initial substrate x_0. Assuming that

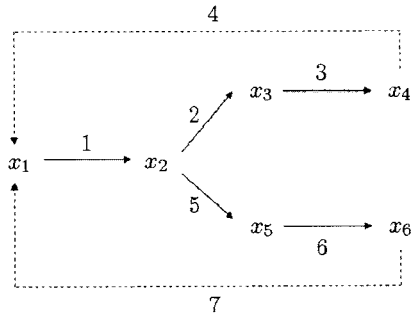


Fig. 1. Feedback inhibition in a branched network. The dashed links 4 and 7 indicate negative (inhibitory) feedback signals. The E matrix for this network is (6).

x_0 is kept constant, and that its conversion to x_1 is regulated by two isofunctional enzymes each of which is selectively sensitive to x_4 or x_6, this example may be seen as an interconnection of several (one-dimensional) subsystems, one for each of the variables, and, under reasonable hypotheses, each of these systems is strictly passive, with appropriate constants γ_i's, as required in Lemma 1 (see [5] for details). Thus, in order to conclude stability, we must study when the matrix

$$E = \begin{bmatrix} -\frac{1}{\gamma_1} & 0 & 0 & -1 & 0 & 0 & -1 \\ 1 & -\frac{1}{\gamma_2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -\frac{1}{\gamma_3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -\frac{1}{\gamma_4} & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -\frac{1}{\gamma_5} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -\frac{1}{\gamma_6} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -\frac{1}{\gamma_7} \end{bmatrix} \tag{6}$$

is diagonally stable. Note that the 4×4 principal submatrices obtained by deleting row-column pairs $\{5,6,7\}$ and $\{2,3,4\}$ each exhibit a cyclic structure for which, as shown in [4], diagonal stability is equivalent to the secant criteria

$$\gamma_1\gamma_2\gamma_3\gamma_4 < \sec(\pi/4)^4 = 4 \quad \text{and} \quad \gamma_1\gamma_5\gamma_6\gamma_7 < 4, \tag{7}$$

respectively. Because principal submatrices of a diagonally stable matrix are also diagonally stable, we conclude that (7) is a necessary condition for the diagonal stability of (6). In fact, we prove the following necessary and sufficient condition:

Lemma 2. The matrix E in (6) is diagonally stable iff

$$\gamma_1\gamma_2\gamma_3\gamma_4 + \gamma_1\gamma_5\gamma_6\gamma_7 < \sec(\pi/4)^4 = 4\,. \tag{8}$$

Proof: We prove the sufficiency of this condition as a consequence of a more general fact. Consider the following diagonal matrix:

$$D = \operatorname{diag}\left(1\,,\ \frac{\gamma_3\gamma_4}{2}\,,\ \frac{\gamma_4}{\gamma_2}\,,\ \frac{2}{\gamma_2\gamma_3}\,,\ \frac{\gamma_6\gamma_7}{2}\,,\ \frac{\gamma_7}{\gamma_5}\,,\ \frac{2}{\gamma_5\gamma_6}\right) \tag{9}$$

and the matrix

$$M := E^T D + DE.$$

We will prove that condition (8) implies that $M \leq 0$. Diagonal stability of E follows from this claim in view of the following argument: Given any γ_i's satisfying the constraint (8), we can find $\tilde{\gamma}_i > \gamma_i$ that still satisfy the constraint, and under this transformation E gets transformed to $\tilde{E} = E + \Delta$, where Δ is some positive diagonal matrix. Now let $\tilde{D}$ be defined for $\tilde{E}$ as in (9) with γ_i's replaced by $\tilde{\gamma}_i$'s. Since $E^T\tilde{D} + \tilde{D}E < \tilde{E}^T\tilde{D} + \tilde{D}\tilde{E} = \tilde{M}$, and since $\tilde{M} \leq 0$, it follows that $E^T\tilde{D} + \tilde{D}E < 0$, which means that E is diagonally stable.

To prove that (8) implies $M \leq 0$, we let $E_\varepsilon := E - \varepsilon I$ for each $\varepsilon > 0$, and show that $M_\varepsilon = E_\varepsilon^T D + DE_\varepsilon$ is negative definite for small enough $\varepsilon > 0$. By continuity, this last property implies that $M \leq 0$. In order to check negative

definiteness of M_ε, we consider the principal minors $\mu_i(\varepsilon)$, $i = 1, \ldots, 7$ of M_ε, and ask that they all have sign $(-1)^i$ for small $\varepsilon > 0$. Each μ_i is a polynomial of degree ≤ 7 on ε and, upon lengthy calculations omitted here, the determinant of M_ε can be expanded as follows:

$$\mu_7(\varepsilon) = \frac{8\gamma_4\gamma_7(\gamma_5 + 2\gamma_6 + \gamma_7)(\gamma_2 + 2\gamma_3 + \gamma_4)}{\gamma_1\gamma_2^3\gamma_3\gamma_5^3\gamma_6}\,\Delta\,\varepsilon^2 + O(\varepsilon^3), \tag{10}$$

where $\Delta = \gamma_1\gamma_2\gamma_3\gamma_4 + \gamma_1\gamma_5\gamma_6\gamma_7 - 4$. Similarly, we have:

$$\mu_6(\varepsilon) = \frac{-2\gamma_4\gamma_7^2(\gamma_2 + 2\gamma_3 + \gamma_4)}{\gamma_1\gamma_2^3\gamma_3\gamma_5^2}\,\Delta\,\varepsilon + O(\varepsilon^2),$$

$$\mu_5(\varepsilon) = \frac{2\gamma_4\gamma_6\gamma_7(\gamma_2 + 2\gamma_3 + \gamma_4)}{\gamma_1\gamma_2^3\gamma_3\gamma_5}\,\Delta\,\varepsilon + O(\varepsilon^2),$$

$$\mu_4(\varepsilon) = \frac{-2\gamma_4(\gamma_2 + 2\gamma_3 + \gamma_4)}{\gamma_1\gamma_2^3\gamma_3}\,\Delta_1\,\varepsilon + O(\varepsilon^2),$$

where $\Delta_1 = \gamma_1\gamma_2\gamma_3\gamma_4 - 4$,

$$\mu_3(\varepsilon) = \frac{\gamma_4^2}{2\gamma_1\gamma_2^2}\,\Delta_1 + O(\varepsilon),$$

$$\mu_2(\varepsilon) = \frac{-\gamma_3\gamma_4}{4\gamma_1\gamma_2}\,(\Delta_1 - 4) + O(\varepsilon),$$

and

$$\mu_1(\varepsilon) = -\frac{2}{\gamma_1} - 2\varepsilon.$$

Since $\Delta_1 < \Delta$, we conclude that the matrix M_ε is negative definite for all small enough $\varepsilon > 0$ if and only if $\Delta < 0$. In particular, condition (8) implies that $M \leq 0$, as claimed.

Finally, we prove the necessity of (8) for the diagonal stability of E in (6). To this end, we define $\hat{E} = \text{diag}\,(\gamma_1, \cdots, \gamma_7)\,E$ which has all diagonal components equal to -1, and characteristic polynomial equal to:

$$(s+1)^3[(s+1)^4 + k],$$

where $k := \gamma_1\gamma_2\gamma_3\gamma_4 + \gamma_1\gamma_5\gamma_6\gamma_7$. For $k \geq 0$, the roots of $(s+1)^4 = -k$ have real part $\pm\sqrt[4]{k/4} - 1$; hence $k < 4$ is necessary for these real parts to be negative. Because (8) is necessary for the Hurwitz property of $\hat{E}$, it is also necessary for its diagonal stability. Since diagonal stability of $\hat{E}$ is equivalent to diagonal stability of E, we conclude that (8) is necessary for the diagonal stability of E.

Extension of our results to more general classes of branched structures are being currently developed.

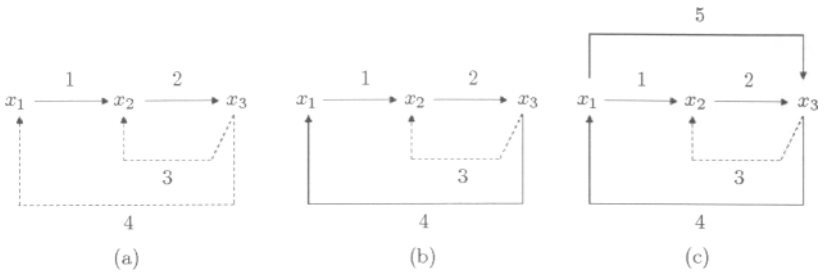


Fig. 2. Three feedback configurations

5 A Signaling Network

We study in this section the diagonal stability of the matrices E associated respectively to the three graphs shown in Figure 2). These interconnection graphs are motivated by the paper [16], which dealt with MAPK (mitogen-activated protein kinase) cascades in PC-12 cells. The nodes x_1, x_2 and x_3 represent the proteins Raf-1, Mek1/2 and Erk1/2, respectively, and dashed lines represent negative feedback signals. The authors of [16] showed that there are topological differences depending on whether the cells are activated with (a) epidermal or (b) neuronal growth factors, leading in particular to a change of sign in the feedback from Erk1/2 to Raf-1. In addition, there is an increased connectivity from Raf-1 to Erk1/2 when neuronal growth factor activation is observed over a longer period. The paper [16] also relates the differences in dynamic behavior to the change in functionality (proliferation or differentiation) of the network.

Assuming once again one-dimensional systems (this may be generalized to a more realistic model that keeps track of phosphorylation states), one may assume that each system is passive with appropriate constants γ_i (see [5] for details), so we study the associated E matrices, which are, for the feedback configurations (a) and (b) in Figure 2:

$$E_a = \begin{bmatrix} -\frac{1}{\gamma_1} & 0 & 0 & -1 \\ 1 & -\frac{1}{\gamma_2} & -1 & 0 \\ 0 & 1 & -\frac{1}{\gamma_3} & 0 \\ 0 & 1 & 0 & -\frac{1}{\gamma_4} \end{bmatrix} \qquad E_b = \begin{bmatrix} -\frac{1}{\gamma_1} & 0 & 0 & 1 \\ 1 & -\frac{1}{\gamma_2} & -1 & 0 \\ 0 & 1 & -\frac{1}{\gamma_3} & 0 \\ 0 & 1 & 0 & -\frac{1}{\gamma_4} \end{bmatrix} \tag{11}$$

and for configuration (c) is:

$$E_c = \begin{bmatrix} -\frac{1}{\gamma_1} & 0 & 0 & 1 & 0 \\ 1 & -\frac{1}{\gamma_2} & -1 & 0 & 0 \\ 0 & 1 & -\frac{1}{\gamma_3} & 0 & 1 \\ 0 & 1 & 0 & -\frac{1}{\gamma_4} & 1 \\ 0 & 0 & 0 & 1 & -\frac{1}{\gamma_5} \end{bmatrix}. \tag{12}$$

The following lemma derives necessary and sufficient conditions for the diagonal stability of E_a and E_b:

Lemma 3. The matrix E_a in (11) is diagonally stable iff $\gamma_1\gamma_2\gamma_4 < 8$, and E_b is diagonally stable iff $\gamma_1\gamma_2\gamma_4 < 1$.

Proof: Note that the 3×3 principal submatrix $\tilde{E}_a$ obtained by deleting the third row and column of E_a exhibits the cyclic form (3) for which diagonal stability is equivalent to $\gamma_1\gamma_2\gamma_4 < 8$ from the secant criterion. Likewise, the corresponding submatrix $\tilde{E}_b$ of E_b is of the form (3) with the upper right element -1 replaced by $+1$. Because all diagonal entries of $\tilde{E}_b$ are negative and off-diagonal entries are nonnegative, it follows from [17, Theorem 2.3] that this submatrix is diagonally stable iff the principal minors of $-\tilde{E}_b$ are all positive. Checking the positivity of these principal minors, we obtain the diagonal stability condition $\gamma_1\gamma_2\gamma_4 < 1$. Because principal submatrices of a diagonally stable matrix are also diagonally stable we conclude that the conditions $\gamma_1\gamma_2\gamma_4 < 8$ and $\gamma_1\gamma_2\gamma_4 < 1$ for the diagonal stability of $\tilde{E}_a$ and $\tilde{E}_b$ are necessary for the diagonal stability of the full matrices E_a and E_b, respectively. To prove that they are also sufficient, we note that both E_a and E_b possess the property that their entries $(2,3)$ and $(3,2)$ are of opposite sign, and all other off-diagonal entries in the third row and column are zero. This means that, if the principal submatrix obtained by deleting the third row and column is diagonally stable then so is the full matrix. (To see this, let the diagonal Lyapunov solution for the submatrix be $\tilde{D} = \text{diag}\{d_1, d_2, d_4\}$, and choose $d_3 = d_2$ in $D = \text{diag}\{d_1, d_2, d_3, d_4\}$ for the full matrix so that all off-diagonal entries in the third rows and columns of $DE_a + E_a^T D$ and $DE_b + E_b^T D$ are zero.) □

The Matrix E_c

It harder to establish conditions for the diagonal stability of the matrix E_c in (12). As a first observation, note that the principal submatrix $\tilde{E}_c$ obtained by deleting the third row and column exhibits nonnegative off-diagonal entries and, thus, its diagonal stability is equivalent [17, Theorem 2.3] to the positivity of the principal minors of $-\tilde{E}_c$, which results in the condition:

$$\gamma_1\gamma_2\gamma_4 + \gamma_4\gamma_5 < 1. \tag{13}$$

Because principal submatrices of a diagonally stable matrix are also diagonally stable, (13) is necessary for the diagonal stability of the full matrix E_c. In contrast to our analysis for E_a and E_b however, we cannot conclude sufficiency of this condition for the diagonal stability of E_c because the entries $(3,5)$ and $(5,3)$ of the deleted row and column do not have opposite signs (*cf.* proof of Lemma 3).

We explored numerically the dependence on γ_3 and γ_4 when the remaining parameters are fixed; Figure 3 is an example of the conditions obtained. Specifically, we sketch the exact diagonal stability region in the parameter plane(γ_3, γ_4) when fixing $\gamma_1 = 1$, $\gamma_2 = \gamma_5 = 0.5$ (so that (13) becomes $\gamma_4 < 1$), plotting the region in the in which diagonal stability is confirmed numerically by a linear matrix inequality (LMI) solver. Observe that there is a gap between the necessary condition (13) and the exact condition: this feasibility region is narrower than $\gamma_4 < 1$

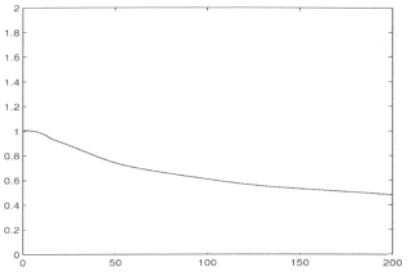


Fig. 3. The region under the curve is the diagonal stability region for (12) in the (γ_3, γ_4)-plane when the other gains are fixed at $\gamma_1 = 1$, $\gamma_2 = \gamma_5 = 0.5$

which means that, unlike the feedback configurations (a) and (b), diagonal stability for the configuration in Figure 2(c) is affected by the magnitude of the gain γ_3. The precise characterization of diagonal stability for matrices of the form E_c is still open.

6 Conclusions

The interconnection approach pioneered by Vidyasagar has long had a deep impact in control and systems theory. Supplemented with explicit characterizations of diagonal stability and with procedures for verifying passivity in reaction network models [4, 5], this approach is now throwing new light onto classical results in mathematical biology, and suggesting new directions for further research in that field. A most noteworthy feature of this approach is its "robustness" to uncertainty in dynamics and parameters. Once that the interconnection structure is known, inserting any subsystems that have appropriate passivity properties (quantified by the γ_i's) will result in a stable interconnection. In systems molecular biology, often precise estimates of parameters are very hard to come by, as discussed in [18]. Approaches like this one, that only require a relatively small amount of quantitative information, are particularly useful in that context.

References

1. Vidyasagar, M.: Input-Output Analysis of Large Scale Interconnected Systems. Springer, Berlin (1981)
2. Sundareshan, M.K., Vidyasagar, M.: l^2-stability of large-scale dynamical systems: Criteria via positive operator theory. IEEE Transactions on Automatic Control AC-22, 396–400 (1977)
3. Moylan, P.J., Hill, D.J.: Stability criteria for large-scale systems. IEEE Trans. Autom. Control 23(2), 143–149 (1978)
4. Arcak, M., Sontag, E.: Diagonal stability of a class of cyclic systems and its connection with the secant criterion. Automatica 42(9), 1531–1537 (2006)

5. Arcak, M., Sontag, E.D.: A passivity-based stability criterion for a class of interconnected systems and applications to biochemical reaction networks. Mathematical Biosciences and Engineering (to appear, 2007) (preprint: arxiv0705.3188v1 [q-bio])
6. Sontag, E.D.: Passivity gains and the "secant condition" for stability. Systems Control Lett. 55(3), 177–183 (2006)
7. Tyson, J.J., Othmer, H.G.: The dynamics of feedback control circuits in biochemical pathways. In: Rosen, R., Snell, F.M. (eds.) Progress in Theoretical Biology, vol. 5, pp. 1–62. Academic Press, London (1978)
8. Thron, C.D.: The secant condition for instability in biochemical feedback control - Parts I and II. Bulletin of Mathematical Biology 53, 383–424 (1991)
9. Goodwin, B.C.: Oscillatory behavior in enzymatic control processes. Adv. Enzyme Reg. 3, 425–439 (1965)
10. Hastings, S.P., Tyson, J., Webster, D.: Existence of periodic orbits for negative feedback cellular control systems. Journal of Differential Equations 25(1), 39–64 (1977)
11. Glass, L., Pasternack, J.S.: Prediction of limit cycles in mathematical models of biological control systems. Bulletin of Mathematical Biology, 27–44 (1978)
12. Morales, M., McKay, D.: Biochemical oscillations in controlled systems. Biophys. J. 7, 621–625 (1967)
13. Stephanopoulos, G.N., Aristidou, A.A., Nielsen, J.: Metabolic Engineering Principles and Methodologies. Academic Press, London (1998)
14. Kholodenko, B.N.: Negative feedback and ultrasensitivity can bring about oscillations in the mitogen-activated protein kinase cascades. Eur. J. Biochem 267, 1583–1588 (2000)
15. Chitour, Y., Grognard, F., Bastin, G.: Equilibria and stability analysis of a branched metabolic network with feedback inhibition. Networks and Heterogeneous Media 1, 219–239 (2006)
16. Santos, S.D.M., Verveer, P.J., Bastiaens, P.I.H.: Growth factor induced MAPK network topology shapes Erk response determining PC-12 cell fate. Nature Cell Biology 9, 324–330 (2007)
17. Berman, A., Plemmons, R.J.: Nonnegative Matrices in the Mathematical Sciences. Society for Industrial and Applied Mathematics, Classics in Applied Mathematics, Philadelphia (1994)(Originally published by Academic Press, New York, 1979)
18. Sontag, E.D.: Some new directions in control theory inspired by systems biology. IET Systems Biology 1, 9–18 (2004)

Identification of Linear Continuous-time Systems Based on Iterative Learning Control

Toshiharu Sugie

Department of Systems Science, Kyoto University, Japan
sugie@i.kyoto-u.ac.jp

Summary. One of the most important issues in control system design is to obtain an accurate model of the plant to be controlled. Though most of the existing identification methods are described in discrete-time, it would be more appropriate to have continuous-time models directly from the sampled I/O data. This paper presents a novel approach for such direct identification of continuous-time systems based on iterations. The method achieves identification through iterative learning control concepts in the presence of heavy measurement noise. The robustness against measurement noise is achieved through (i) projection of continuous-time I/O signals onto a finite dimensional parameter space, and (ii) noise tolerant learning laws. The method can be easily applied to system identification in closed loop. The effectiveness of the method is demonstrated through numerical examples for systems including non-minimum phase one.

Keywords: Identification, linear continuous-time systems, iterative learning control.

1 Introduction

One of the most important issues in control system design is to obtain an accurate model of the plant to be controlled. Though most of the existing identification methods are described in discrete-time, it would often be convenient to have continuous-time models directly from the I/O data. In fact, it is often easier for us to capture the plant dynamics intuitively in continuous-time rather than in discrete-time. A basic difficulty of continuous-time identification is that standard approaches (called *direct* methods) require the time-derivatives of I/O data in the presence of measurement noise. A comprehensive survey on this topic has been given by [18], [17] and [15]. Furthermore, the Continuous-Time System Identification (CONTSID) tool-box has been developed on the basis of these direct methods [6, 7, 5].

On the other hand, iterative learning control (ILC) has attracted much attention over the last two decades as a powerful model-free control methodology [1, 10, 11, 2, 4]. ILC yields the input which achieves perfect output tracking by iteration of trials for uncertain systems. Though ILC can deal with plants having large uncertainty, most ILC approaches need time-derivatives of I/O data in the continuous-time case [16], and therefore it is quite sensitive to measurement noise. Recently, Hamamoto and Sugie [8] [9] proposed an ILC where the

V.D. Blondel et al. (Eds.) Recent Advances in Learning and Control, LNCIS 371, pp. 205–218, 2008.
springerlink.com

learning law works in a certain finite-dimensional subspace and showed that time-derivative of the tracking error is not required to achieve perfect tracking in the their scheme.

Based on this work, Sugie and Sakai, [12] and [13], proposed an ILC which works in the presence of heavy measurement noise, and, moreover, the method was shown to be applicable to the identification of continuous-time systems as well. This identification method has several advantages such as: (i) no time-derivatives of I/O data are required, (ii) it delivers unbiased estimations. This work was followed by Campi et al [3]. They show a way how to deal with plant zeros and provide more noise tolerant learning law which guarantees zero convergence of the parameter estimation error as the number of trials increases. Further, Sakai and Sugie [14] extend this result to closed loop system identification. The purpose of this paper is to describe a basic idea of these ILC based identification methods with some numerical examples which demonstrate its effectiveness.

The following notations will be used. Superscript denotes the trial number and subscript denotes the element of a set or a matrix. For instance, input u at the k-th trial is written as u^k while $\boldsymbol{x}_i$ is the i-th element of the vector $\boldsymbol{x}$.

2 System Description

Consider the continuous-time SISO system described by

$$y(t) = \frac{B^\circ(p)}{A^\circ(p)}u(t) \triangleq \frac{1 + \beta_1^\circ p + \cdots + \beta_m^\circ p^m}{\alpha_0^\circ + \alpha_1^\circ p + \cdots + \alpha_n^\circ p^n}u(t) \tag{1}$$

where $u(t)$ and $y(t)$, $t \in [0, T]$, are the input and the output, respectively, $\alpha_i^\circ \in \mathbb{R}$ $(i = 1, \cdots, n)$ and $\beta_i^\circ \in \mathbb{R}$ $(i = 0, 1, \cdots, m)$ are coefficient parameters, while p is the differential operator, *i.e.*, $pu(t) = \mathrm{d}u(t)/\mathrm{d}t$. We assume the following:

- Many experiments on the system can be repeated with zero initial state on the time interval $[0,\ T]$.
- Though the true parameters α_i° and β_i° are unknown, $A^\circ(p)$ and $B^\circ(p)$ are coprime and their order n and m are known.
- We can measure $\tilde{y}(t)$, the output contaminated with noise,

$$\tilde{y}(t) = y(t) + w(t)$$

where $w(t)$ is zero-mean measurement noise.

The goal is to determine a model in the class

$$\mathcal{M} = \left\{ \frac{B(p)}{A(p)} = \frac{1 + \beta_1 p + \cdots + \beta_m p^m}{\alpha_0 + \alpha_1 p + \cdots \alpha_n p^n} \right\}$$

based on I/O measurements $u(t)$ and $\tilde{y}(t)$.

3 Identification Procedure

This section gives a basic idea of how to identify the system through iteration of trials. For simplicity, we discuss the case where the system has no finite zeros, i.e., $B^\circ(p) = 1$ in this section. The general case will be discussed later.

3.1 Data Generation Scheme and Projection

Choose a smooth signal $r(t)$, $t \in [0, T]$ satisfying

$$r(0) = 0, \quad \dot{r}(0) = 0, \quad \cdots, \quad r^{(n-1)} = 0.$$

At the k-th trial, perform the following experiment on $[0, T]$ which produces the signal $\varepsilon^k(t)$ when the parameter estimates $\alpha_0^k, \cdots, \alpha_n^k$ are given.

(i) Define $A^k(p) := \alpha_0 + \alpha_1^k p + \cdots + \alpha_n^k p^n$

(ii) Inject $u^k(t) = A^k(p)r(t)$ into the system.

(iii) Observe the corresponding output $\tilde{y}^k(t)$.

(iv) Compute the tracking error signal $\varepsilon^k(t)$ by $\varepsilon^k(t) = \tilde{y}^k(t) - r(t)$.

Note that $\varepsilon^k(t)$ is obtained without taking any derivative of noisy measurements, only derivatives of $r(t)$ are required. Note also that $\varepsilon^k(t)$ can also be written as

$$\varepsilon^k(t) = \frac{A^k(p) - A^\circ(p)}{A^\circ(p)} r(t) + w^k(t). \tag{2}$$

Now, we introduce $n+1$ square-integrable functions $f_1(t), \ldots, f_{n+1}(t)$ which satisfy the following condition.

Condition 1: If

$$\int_0^T \left(\frac{A(p) - A^\circ(p)}{A^\circ(p)} r(t) \right) \cdot f_i(t) \mathrm{d}t = 0, \quad \forall\, i \tag{3}$$

are satisfied, then $A(p) - A^\circ(p) = 0$ holds.

Condition 1 requires that if the projection of $\frac{A(p)-A^\circ(p)}{A^\circ(p)} r(t)$ onto $f_i(t)$ is zero for all i's, then $\frac{A(p)-A^\circ(p)}{A^\circ(p)} r(t)$ must be zero. Since $\frac{A(p)-A^\circ(p)}{A^\circ(p)} r(t)$ linearly depends on $n+1$ parameters, determining $f_i(t)$'s such that this condition is satisfied is not difficult.

We next regard $\varepsilon^k(t)$ and $f_1(t), f_2(t), \cdots, f_{n+1}(t)$ as elements of $L_2[0, T]$, and project $\varepsilon^k(t)$ onto the finite-dimensional subspace described by

$$\mathcal{F} \triangleq \mathrm{span}\{f_1(t), f_2(t), \cdots, f_{n+1}(t)\}$$

The projection is written as

$$\varepsilon^k(t)|_{\mathcal{F}} = \delta_1^k f_1(t) + \cdots + \delta_{n+1}^k f_{n+1}(t)$$

and $\boldsymbol{\delta}^k \triangleq [\delta_1^k, \cdots, \delta_{n+1}^k]^T$ is its vector representation. Let for brevity

$$\boldsymbol{\gamma}^\circ = [\alpha_0^\circ, \alpha_1^\circ, \cdots, \alpha_n^\circ]^T, \quad \boldsymbol{\gamma}^k = [\alpha_0^k, \alpha_1^k, \cdots, \alpha_n^k]^T$$

and, for the time being, $w^k(t) = 0$ holds. It is easy to see that the tracking error $\boldsymbol{\delta}^k \in \mathbb{R}^{n+1}$ depends on the parameter estimate $\boldsymbol{\gamma}^k \in \mathbb{R}^{n+1}$ linearly. So there exist a matrix $M \in \mathbb{R}^{(n+1)\times(n+1)}$ and an offset term $\bar{\boldsymbol{\delta}} \in \mathbb{R}^{n+1}$ such that

$$\boldsymbol{\delta}^k = M\boldsymbol{\gamma}^k + \bar{\boldsymbol{\delta}} \tag{4}$$

holds. Note that, from Condition 1, it follows that $\boldsymbol{\delta}^k = 0$ implies $\boldsymbol{\gamma}^k = \boldsymbol{\gamma}^\circ$. This also means that the equation

$$0 = M\boldsymbol{\gamma}^k + \bar{\boldsymbol{\delta}}$$

has the only solution $\boldsymbol{\gamma}^k = \boldsymbol{\gamma}^\circ$, so that M is non-singular and $\bar{\boldsymbol{\delta}} = -M\boldsymbol{\gamma}^\circ$. Thus, (4) can be re-written as

$$\boldsymbol{\delta}^k = M(\boldsymbol{\gamma}^k - \boldsymbol{\gamma}^\circ)$$

with M non-singular. When noise $w^k(t)$ is taken into account, $\boldsymbol{\delta}^k$ becomes

$$\boldsymbol{\delta}^k = M(\boldsymbol{\gamma}^k - \boldsymbol{\gamma}^\circ) + \boldsymbol{\nu}^k \tag{5}$$

where $\boldsymbol{\nu}^k$ accounts for the projection of $w^k(t)$ onto $\mathcal{F}$.

3.2 Update Laws

Three types of the update laws for (5) have been proposed so far.

(A) Error accumulation type learning law [13]
First one is to use the accumulated error, i.e.,

$$\boldsymbol{\gamma}^{k+1} = H_1\boldsymbol{\gamma}^k + H_2\boldsymbol{\xi}^k \tag{6}$$

$$\boldsymbol{\xi}^{k+1} = \boldsymbol{\xi}^k + \boldsymbol{\delta}^{k+1} \tag{7}$$

for $k = 0, 1, 2, \cdots$ with the initial condition $\boldsymbol{\gamma}^0 = 0$ and $\boldsymbol{\xi}^0 = 0$. where $H_1 \in \mathbb{R}^{(n+1)\times(n+1)}$ and $H_2 \in \mathbb{R}^{(n+1)\times(n+1)}$ are gain matrices. The gain $H := [H_1, H_2]$ are chosen by taking account of both the noise reduction and the learning convergence speed.

Define $\boldsymbol{x}^k \triangleq \left[(\boldsymbol{\gamma}^k)^T,\ (\boldsymbol{\xi}^k)^T\right]^T$, then the update law is rewritten by

$$\boldsymbol{x}^{k+1} = A\boldsymbol{x}^k + B_0\boldsymbol{\gamma}^\circ + B_1\boldsymbol{\nu}^{k+1} + B_2\boldsymbol{\gamma}^{k+1} \tag{8}$$

$$\boldsymbol{\gamma}^{k+1} = [H_1 \ \ H_2]\boldsymbol{x}^k \tag{9}$$

$$A \triangleq \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}, \ B_0 \triangleq \begin{bmatrix} 0 \\ -M \end{bmatrix}, B_1 \triangleq \begin{bmatrix} 0 \\ I \end{bmatrix}, \ B_2 \triangleq \begin{bmatrix} I \\ M \end{bmatrix}.$$

Let

$$\boldsymbol{E}\left[\boldsymbol{\nu}^k \boldsymbol{\nu}^{kT}\right] = \Phi > 0.$$

In order to reduce the noise effect, select an output $\boldsymbol{z}^k$ for the system described by (8) and (9), on which the noise effect should be minimized, as

$$\boldsymbol{z}^k = C_1 \boldsymbol{x}^k + D_{12} \boldsymbol{\gamma}^{k+1}, \tag{10}$$

where C_1 and D_{12} are matrices with appropriate dimension. Let $T_{z\nu}(z)$ denote the transfer matrix from $\boldsymbol{\nu}^{k+1}$ to $\boldsymbol{z}^k$, i.e.,

$$T_{z\nu}(z) = (C_1 + D_{12}H)(zI - A - B_2 H)^{-1} B_1.$$

Then, it is known that

$$\boldsymbol{E}[\|\boldsymbol{z}^\infty\|] = \|T_{zw}(z)\Phi^{1/2}\|_2$$

holds, where $\boldsymbol{z}^\infty \triangleq \lim_{k\to\infty} \boldsymbol{z}^k$ and $\|\cdot\|_2$ denotes the $\mathcal{H}^2$ norm. Furthermore it is easy to verify that if there exist matrices Q, P and W satisfying the following LMI (linear matrix inequality)

$$\begin{bmatrix} P & PA^T + Q^T B_2^T \\ AP + B_2 Q & P - B_1 \Phi B_1^T \end{bmatrix} > 0, \tag{11}$$

$$\begin{bmatrix} P & PC_1^T + Q^T D_{12}^T \\ C_1 P + D_{12} Q & W \end{bmatrix} > 0, \tag{12}$$

$$\text{trace} W < \gamma^2 \tag{13}$$

for given $\gamma > 0$, then

$$\|T_{zw}(z)\Phi^{1/2}\|_2 < \gamma$$

holds for $H = QP^{-1}$.

While, concerning the convergence speed of the learning, it is known that $\rho_s(A + B_2 H) < \rho$ holds for given $\rho > 0$ if there exist a matrix P satisfying

$$\begin{bmatrix} -\rho P & (A + B_2 H)P \\ P(A + B_2 H)^T & -\rho P \end{bmatrix} < 0 \tag{14}$$

Therefore, if there exist matrices M and P satisfying

$$\begin{bmatrix} -\rho P & AP + B_2 Q \\ PA + Q^T B_2^T & -\rho P \end{bmatrix} < 0, \tag{15}$$

then the feedback gain $H = QP^{-1}$ fulfills the specification.

Summarizing the above arguments, we can obtain the following result.

Theorem 1. [13] Given an arbitrary positive scalar $\rho < 1$. Let M, P and W be the matrices which minimize $\epsilon > 0$ subject to (11) $\sim$ (15). Then when the update law (6) and (7) with $H \triangleq [H_1,\ H_2] = QP^{-1}$ is adopted,

$$\boldsymbol{E}[\|\boldsymbol{z}^\infty\|] < \epsilon^* \tag{16}$$

$$\rho_s(A + B_2 H) < \rho \tag{17}$$

hold, where ϵ^* denotes the optimal value of ϵ and $\rho_s(\cdot)$ denotes the spectral radius. □

This update law gives an unbiased estimation with minimum variance.

(B) Kalman filter type learning law [3]
The second one is described by

$$\gamma^{k+1} = \gamma^k + H^k \delta^k, \tag{18}$$

where $H^k \in \mathbb{R}^{(n+1)\times(n+1)}$ is the learning gain to choose. Inserting (5) in the above equation yields

$$\gamma^{k+1} = \gamma^k + H^k M(\gamma^k - \gamma^\circ) + H^k \nu^k$$

Thus, defining $\tilde{\gamma}^k \triangleq \gamma^k - \gamma^\circ$ we have

$$\tilde{\gamma}^{k+1} = (I + H^k M)\tilde{\gamma}^k + H^k \nu^k \tag{19}$$

which describes how the error $\tilde{\gamma}^k$ propagates through trials. Now, let

$$\Phi \triangleq \boldsymbol{E}[\nu^k(\nu^k)^T], \quad P^k \triangleq \boldsymbol{E}[\tilde{\gamma}^k(\tilde{\gamma}^k)^T],$$

and discuss how to select H^k so as to reduce P^k optimally under the assumption that M and Φ are known. The computation is in line with Kalman filtering variance minimization. Noise is assumed to be independent in different experiments. From (19), we have:

$$\begin{aligned} P^{k+1} &= \boldsymbol{E}\left[\{(I + H^k M)\tilde{\gamma}^k + H^k \nu^k\}\{(I + H^k M)\tilde{\gamma}^k + H^k \nu^k\}^T\right] \\ &= (I + H^k M)P^k(I + H^k M)^T + H^k \Phi (H^k)^T \end{aligned}$$

Therefore, P^{k+1} is minimized by the choice

$$H^k = -P^k M^T (M P^k M^T + \Phi)^{-1} \tag{20}$$

With this choice, we obtain

$$P^{k+1} = P^k - P^k M^T (M P^k M^T + \Phi)^{-1} M P^k \tag{21}$$

Eqns. (20) and (21), where (21) is initialized with

$$P^0 = \boldsymbol{E}[\tilde{\gamma}^0(\tilde{\gamma}^0)^T]. \tag{22}$$

Based on the above idea, the following theorem is obtained.

Theorem 2. [3] Suppose we adopt the updating law (18) with (20), (21), (22). Then it holds that

$$\boldsymbol{E}[(\gamma^k - \gamma^\circ)(\gamma^k - \gamma^\circ)^T] \to 0, \quad \text{as} \quad k \to \infty. \tag{23}$$

□

It implies that the method gives us the true parameter γ° in the presence of measurement noise through iteration of trials. Further, an alternative version is given as follows.

Theorem 3: [3] If the updating law (18) is adopted where H^k is given by

$$H^k = -\frac{1}{k+1}M^{-1}, \tag{24}$$

then (23) holds. □

4 Digital Implementation

Now we discuss how to implement the iterative identification method when the I/O data are available only at the sampled times. Namely, we suppose that the I/O data are $\{u(iT_s), \tilde{y}(iT_s)\}$ $(i = 0, 1, \cdots, q)$, where T_s is sampling time satisfying $qT_s = T$, and the input is injected to the plant via the zero-order holder (ZOH).

(A) Representation in the projected space
Suppose that the reference signal $r(t)$ is chosen. Define

$$V_r(t) = \left[r(t), \frac{\mathrm{d}r(t)}{\mathrm{d}t}, \ldots, \frac{\mathrm{d}^n r(t)}{\mathrm{d}t^n}\right],$$

then, we have

$$u(t)^k = A^k(p)r(t) = V_r(t)\gamma^k.$$

Since the data are available only at sampled times, we define

$$\boldsymbol{u}^k \triangleq [u^k(0), u^k(T_s), \cdots, u^k(qT_s)]^T \in \mathbb{R}^{q+1}$$
$$\tilde{\boldsymbol{y}}^k \triangleq [\tilde{y}^k(0), \tilde{y}^k(T_s), \cdots, \tilde{y}^k(qT_s)]^T \in \mathbb{R}^{q+1}.$$

The vectors $\boldsymbol{r}^k \in \mathbb{R}^{q+1}$, $\boldsymbol{y}^k \in \mathbb{R}^{q+1}$, $\boldsymbol{\varepsilon}^k \in \mathbb{R}^{q+1}$ and $\boldsymbol{w}^k \in \mathbb{R}^{q+1}$ are defined in the same way. Similarly, let $V_{dr} \in \mathbb{R}^{(q+1)\times(n+1)}$ be

$$V_{dr} \triangleq \begin{bmatrix} r(0) & \dot{r}(0) & \ldots & r^{(n)}(0) \\ r(T_s) & \dot{r}(T_s) & \ldots & r^{(n)}(T_s) \\ \vdots & \vdots & \ldots & \vdots \\ r(qT_s) & \dot{r}(qT_s) & \ldots & r^{(n)}(qT_s) \end{bmatrix}.$$

Then, we have

$$\boldsymbol{u}^k = V_{dr}\gamma^k. \tag{25}$$

While, let $\{f_1(t), f_2(t), \cdots, f_{n+1}(t)\}$ are given functions which satisfy Condition 1 for given $r(t)$, and define $V_{df} \in \mathbb{R}^{(q+1)\times(n+1)}$ by

$$V_{df} \triangleq \begin{bmatrix} f_1(0) & f_2(0) & \ldots & f_{n+1}(0) \\ f_1(T_s) & f_2(T_s) & \ldots & f_{n+1}(T_s) \\ \vdots & \vdots & \ldots & \vdots \\ f_1(qT_s) & f_2(qT_s) & \ldots & f_{n+1}(qT_s) \end{bmatrix}.$$

Let the QR decomposition of V_{df} be

$$V_{df} = UR, \quad U^T U = I_{n+1},$$

where $U =: [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_{n+1}] \in \mathbb{R}^{(q+1)\times(n+1)}$ and $R \in \mathbb{R}^{(n+1)\times(n+1)}$ is a non-singular upper triangular matrix. These $\boldsymbol{f}_i$'s constitute an orthogonal basis for projection in the digital implementation. Therefore, for example, the vector representation of $\boldsymbol{\varepsilon}^k$ and $\boldsymbol{w}^k$ are given by

$$\boldsymbol{\delta}^k = U^T \boldsymbol{\varepsilon}^k, \quad \boldsymbol{\nu}^k = U^T \boldsymbol{w}^k,$$

(B) Estimate of M

If we inject the input sequence $\boldsymbol{u}^k$ with the ZOH, the corresponding output $\boldsymbol{y}^k$ is given by

$$\boldsymbol{y}^k = G\boldsymbol{u}^k$$

irrespective of k, where G is specified by

$$G = \begin{bmatrix} g_0 & 0 & 0 & \ldots & 0 \\ g_1 & g_0 & 0 & \ldots & 0 \\ g_2 & g_1 & g_0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ g_N & g_{N-1} & g_{N-2} & \cdots & g_0 \end{bmatrix} \in \mathbb{R}^{(q+1)\times(q+1)}. \tag{26}$$

The first column of G is the output $\boldsymbol{y}$ when we inject the input $\boldsymbol{u} = [1, 0, 0, \cdots, 0]$. The tracking error is obtained by

$$\boldsymbol{\varepsilon}^k = G\boldsymbol{u}^k + \boldsymbol{w}^k - \boldsymbol{r}.$$

Through the projection, we have

$$\boldsymbol{\delta}^k = U^T G V_{dr}(\boldsymbol{\gamma}^k - \boldsymbol{\gamma}^\circ) + \boldsymbol{\nu}^k$$

with the aid of $\boldsymbol{u}^k = V_{dr}\boldsymbol{\gamma}^k$ and $\boldsymbol{r} = GV_{dr}\boldsymbol{\gamma}^\circ$. This implies that

$$M = U^T G V_{dr} \tag{27}$$

$$\boldsymbol{\gamma}^\circ = M^{-1} U^T \boldsymbol{r}. \tag{28}$$

Note that we can obtain an estimate $\hat{G}$ of G through simple experiments. For example, observe the unit step responses twice (with independent measurement noises), say, $\tilde{y}_{step1}(t)$ and $\tilde{y}_{step2}(t)$. Then, we obtain $\hat{g}_0 = 0$ and

$$\hat{g}_i = \tilde{y}_{step1}(iT_s) - \tilde{y}_{step2}((i-1)T_s) \quad \text{for } i \geq 1.$$

Once we have $\hat{G}$, we obtain an estimate of M by computing $\hat{M} = U^T \hat{G} V_{dr}$.

(C) Identification steps

The total identification steps are summarized as follows:

(Step 1) Choose the reference signal $r(t)$.

(Step 2) Obtain $\{g_i\}$ in (26) through experiments, and estimate M.

(Step 3) Choose an update law in subsection 3.2.
(Step 4) Perform the experiment in subsection 3.1 iteratively.

In Step 4, inject $\boldsymbol{u}^k = V_{dr}\boldsymbol{\gamma}^k$ and obtain $\boldsymbol{\delta}^k$ at k-th trial. Then calculate $\boldsymbol{\gamma}^{k+1}$ based on the update law with $\boldsymbol{\delta}^k$.

5 Extension to General Cases

So far, we have discussed the case where the system has no finite zeros. This section shows how to extend the proposed identification procedure to more general cases.

5.1 Systems with Poles and Zeros [3]

Consider the plant $B^\circ(p)/A^\circ(p)$ described in (1). The procedure is almost the same except we compute the following error signal

$$\begin{aligned} \varepsilon^k(t) &= \tilde{y}^k(t) - B^k(p)r(t) \\ B^k(p) &:= 1 + \beta_1^k p + \cdots + \beta_m^k p^m \end{aligned} \tag{29}$$

in stead of $\varepsilon^k(t) = \tilde{y}^k(t) - r(t)$, where $\{\beta_1^k, \cdots, \beta_n^k\}$ are the parameter estimates of $\{\beta_1^\circ, \cdots, \beta_n^\circ\}$.

Using the projection onto $\mathcal{F} \triangleq \text{span}\{f_1(t), f_2(t), \cdots, f_{n+m+1}(t)\}$, we have (5) with

$$\boldsymbol{\gamma}^\circ = [\alpha_0^\circ, \cdots, \alpha_n^\circ, \beta_1^\circ, \cdots, \beta_m^\circ]^T, \quad \boldsymbol{\gamma}^k = [\alpha_0^k, \cdots, \alpha_n^k, \beta_1^k, \cdots, \beta_m^k]^T$$

Therefore, the update laws given in section 3 can be used.

Concerning to digital implementation, M should be replaced by

$$M = U^T\{GV_{dr}, -V_{dr}[2 : m+1]\},$$

according to (29). Here $V_{dr}[2 : m+1]$ means the matrix $[v_2, v_3, \cdots, v_{m+1}]$ where v_i denotes the i-th column of V_{dr}.

5.2 Closed Loop Systems [14]

Now consider the closed system shown in Fig. 1, where $K(p)$ is the given stabilizing controller and $P(p) := B^\circ(p)/A^\circ(p)$. In this case, we inject

$$u_a^k(t) = B^k(p)r(t), \qquad u_b^k(t) = A^k(p)r(t).$$

Then the corresponding output is given by

$$\tilde{y}^k(t) = [P(p)K(p)S(p)B^k(p) + P(p)S(p)A^k(p)]r(t) + S(p)w^k(t), \tag{30}$$

$$S(p) := (1 + K(p)P(p))^{-1}. \tag{31}$$

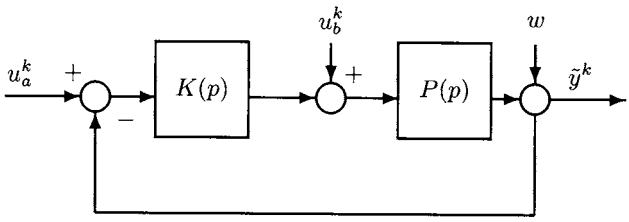


Fig. 1. Closed loop system with measurement noise

It is easy to see that $\tilde{y}^k(t) = B^\circ(p)r(t)$ holds, if $A^k(p) = A^\circ(p)$, $B^k(p) = B^\circ(p)$ and $w(t) = 0$ are satisfied. Therefore, as for the error signal, we compute

$$\begin{aligned}\varepsilon^k(t) &= \tilde{y}^k(t) - B^k(p)r(t) \\ &= [P(p)S(p)A^k(p) - S(p)B^k(p)]r(t) + S(p)w^k(t). \end{aligned} \tag{32}$$

Then, we have (5) again, so that the propose method is available.

6 Numerical Examples

We demonstrate the effectiveness of the proposed method through simulation in this section.

(A) Example 1

Consider the third order plant given by

$$P(s) = \frac{1}{s^3 + 10s^2 + 30s + 8}. \tag{33}$$

The time interval of each trial is $T = 10$[s], and the sampling period is 10[ms]. We add the white noise where the noise to the signal ratio (NSR) is around 30%.

Suppose the reference signal $r(t)$ be the output of the filter $1/(0.5s+1)^6$ when a white noise is injected. As for $\{g_i\}$, an example of the plant output with noise which corresponds to the input $\boldsymbol{u} = [1, 0, \cdots, 0]^T$ is shown in Fig. 2. The thick line shows the true (i.e., noise free) response. From the noisy data, we compute an estimate $\hat{M}$. Then, we employ the error accumulation type update law, and perform the iteration 100 times (i.e., $k = 1 \sim 100$).

Fig. 3 shows the estimated coefficients at each trial k $(= 1, 2, \cdots, 100)$ in the proposed method. From this figure, we see that the denominator coefficients of the plant (i.e., $\alpha_0 = 8$, $\alpha_1 = 30$, $\alpha_2 = 10$, $\alpha_3 = 1$) are almost accurately estimated in the presence of relatively heavy noise.

(B) Example 2

Next, consider the non-minimum phase system described by

$$P(s) = \frac{-T_1 s + 1}{\left(\frac{s^2}{\omega_{n,1}^2} + \frac{2\zeta_1 s}{\omega_{n,1}} + 1\right)\left(\frac{s^2}{\omega_{n,2}^2} + \frac{2\zeta_2 s}{\omega_{n,2}} + 1\right)},$$

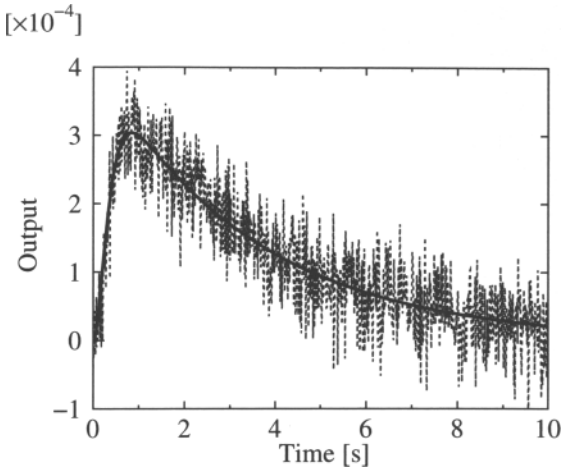


Fig. 2. Impulse like response for estimation of M (NSR = 30%)

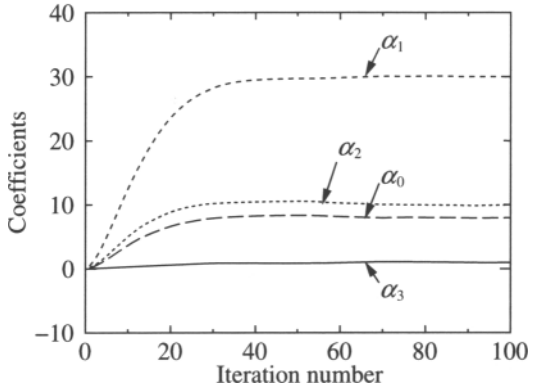


Fig. 3. Identified coefficients in each trial

where $T_1 = 4$[s], $\omega_{n,1} = 20$[rad/s], $\zeta_1 = 0.1$, $\omega_{n,2} = 2$[rad/s], and $\zeta_2 = 0.25$. The time interval of each trial and the sampling period are the same as the above example. We add the white noise with 100 % NSR to evaluate the performance under extremely noisy case.

The signal $r(t)$ is chosen to be the output of the filter $1/(0.1s+1)^6$ with the multiple sinusoidal input given by $\sin(t)+\sin(1.9t)+\sin(2.1t)+\sin(18t)+\sin(22t)$. We estimate M based on the measurement data, and apply the Kalman type update law for the iteration.

Fig. 4 shows the output $\tilde{y}^k(t)$ obtained at the 30th trial. The same figure also displays the signal $B^k(p)r(t)$ by thick line. Though the noise is so intensive, the output looks to track $B^k(p)r(t)$.

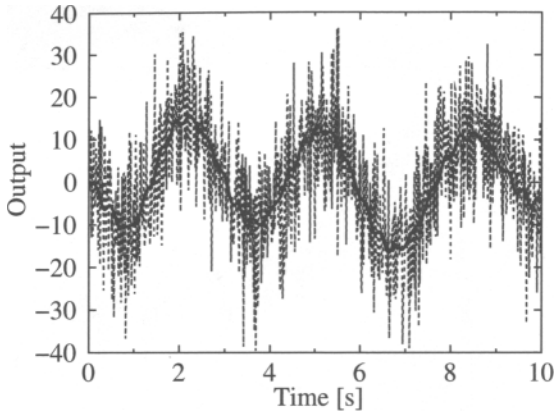


Fig. 4. Measured output and the target signal at 30-th trial(NSR= 100%)

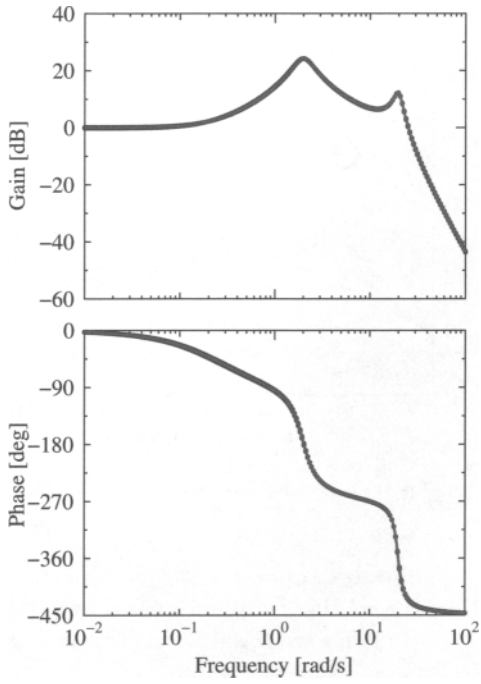


Fig. 5. Bode plots of the estimated systems ($k = 10$)

Fig. 5 shows the Bode plots of the estimated systems at the 3rd trial for 50 runs of the algorithm (bundle of thin lines), where the Bode plot of the true system is also shown in thick line. It tells that the fairly accurate models are obtained with small variance in spite of the heavy measurement noise.

These results demonstrate the effectiveness of the proposed method.

7 Conclusion

This paper described a novel approach for identification of continuous-time systems based on trial iterations. The method enable us to obtain an accurate model in the presence of heavy measurement noise through iterative learning control concepts. The robustness against noise is achieved through (i) projection of continuous-time I/O signals onto an appropriate finite dimensional parameter space, and (ii) noise tolerant learning laws. The method can be applied to closed loop system identification. The effectiveness of the method has been demonstrated through numerical examples for systems including non-minimum phase one.

Acknowledgements. This paper is based on the joint work with Prof. Fumitoshi Sakai and Prof. Marco Campi. The author would like to express his sincere gratitude to them.

References

1. Arimoto, S., Kawamura, S., Miyazaki, F.: Bettering operation of robotics. Journal of Robotic System 1(2), 123–140 (1984)
2. Bein, Z., Xu, J.X.: Iterative learning control – Analysis, design, integration and applications. Kluwer Academic Publishers, Boston (1998)
3. Campi, M.C., Sugie, T., Sakai, F.: Iterative identification method for linear continuous-time systems. In: Proc. of the 45th IEEE Conference on Decision and Control, pp. 817–822 (2006)
4. Chen, Y., Wen, C.: Iterative learning control: convergence, robustness and applications. LNCIS. vol. 248. Springer, Heidelberg (1999)
5. Garnier, H., Gilson, M., Husestein, E.: Developments for the MATLAB® CONTSID toolbox. In: CD-ROM. Proc. of the 13th IFAC Symposium on System Identification (2003)
6. Garnier, H., Mensler, M.: CONTSID: a CONtinuous-Time System IDentification toolbox for MATLAB®. In: Proc. of the 5th European Control Conference (1999)
7. Garnier, H., Mensler, M.: The CONTSID toolbox: a MATLAB® toolbox for CONtinuous-Time System IDentification. In: CD-ROM. Proc. of the 12th IFAC Symposium on System Identification (2000)
8. Hamamoto, K., Sugie, T.: An iterative learning control algorithm within prescribed input-output subspace. Automatica 37(11), 1803–1809 (2001)
9. Hamamoto, K., Sugie, T.: Iterative learning control for robot manipulators using the finite dimensional input subspace. IEEE Trans. Robotics and Automation 18(4), 632–635 (2002)
10. Kawamura, S., Miyazaki, F., Arimoto, S.: Realization of robot motion based on a learning method. IEEE Trans. on Systems, Man and Cybernetics 18(1), 126–134 (1988)
11. Moore, K.L.: Iterative learning control for deterministic systems. Springer-Verlag Series on Advances in Industrial Control. Springer, London (1993)
12. Sakai, F., Sugie, T.: Continuous-time systems identification based on iterative learning control. In: Proc. of the 16th IFAC World Congress (2005)

13. Sakai, F., Sugie, T.: H 2-suboptimal iterative learning control for continuous-time system identification. In: Proc. of American Control Conference, pp. 946–951 (2006)
14. Sakai, F., Sugie, T.: A continuous-time closed-loop identification method based on iterative learning. In: Proc. of the 46th IEEE Conference on Decision and Control (to appear, 2007)
15. Sinha, N.K., Rao, G.P.: Identification of Continuous-Time Systems. Kluwer Academic Publishers, Dordrecht (1991)
16. Sugie, T., Ono, T.: An iterative learning control law for dynamical systems. Automatica 27(4), 729–732 (1991)
17. Unbehauen, H., Rao, G.P.: Continuous-time approaches to system identification - a survey. Automatica 26(1), 23–35 (1990)
18. Young, P.: Parameter estimation for continuous-time models - a survey. Automatica 17(1), 23–39 (1981)

A Pontryagin Maximum Principle for Systems of Flows*

Héctor J. Sussmann

Department of Mathematics, Rutgers University, U.S.A.
sussmann@math.rutgers.edu

Summary. We present a generalization of the Pontryagin Maximum Principle, in which the usual adjoint equation, which contains derivatives of the system vector fields with respect to the state, is replaced by an integrated form, containing only differentials of the reference flow maps. In this form, the conditions of the maximum principle make sense for a number of control dynamical laws whose right-hand side can be nonsmooth, nonlipschitz, and even discontinuous. The "adjoint vectors" that are solutions of the "adjoint equation" no longer need to be absolutely continuous, and may be discontinuous and unbounded. We illustrate this with two examples: the "reflected brachistochrone problem" (RBP), and the derivation of Snell's law of refraction from Fermat's minimum time principle. In the RBP, where the dynamical law is Hölder continuous with exponent 1/2, the adjoint vector turns out to have a singularity, in which one of the components goes to infinity from both sides, at an interior point of the interval of definition of the reference trajectory. In the refraction problem, where the dynamical law is discontinuous, the adjoint vector is bounded but has a jump discontinuity.

1 Introduction

It is well known that the minimum time problem whose solution is Snell's law of refraction was the first link of a long chain of mathematical developments that eventually led to the Pontryagin Maximum Principle (PMP) of optimal control theory: Snell's law was used by Johann Bernoulli's in his solution of the brachistochrone problem; this in turn was a decisive step towards the formulation of the general necessary condition of Euler and Lagrange for the classical Calculus of Variations; the Euler-Lagrange conditions were then strengthened by Legendre, whose second-order condition was later strengthened by Weierstrass; and, finally, Weierstrass' excess function condition led to the Pontryagin Maximum Principle (PMP), stated and proved in [1].

In the five decades since the formulation of the PMP, the result has been generalized in many directions, incorporating high-order conditions(cf. [7], [8]) and various types of nonsmoothness (cf. [2], [3], [4], [5], [6], [9], [13]), and producing intrinsic coordinate-free formulations on manifolds. It is remarkable and somewhat disappointing, however, that the refraction problem that leads to Snell's law

* Research supported in part by NSF Grant DMS-05-09930.

V.D. Blondel et al. (Eds.) Recent Advances in Learning and Control, LNCIS 371, pp. 219–232, 2008.
springerlink.com

does not fit within the framework of any of these generalizations, because even the non-smooth versions of the PMP require Lipschitz conditions on the system vector fields, and for the refraction problem the vector fields are actually discontinuous. A similar phenomenon occurs with the "reflected brachistochrone problem" (RBP), a very natural optimization problem with a Hölder continuous right-hand side.

The purpose of this note is to present a generalization of the PMP that applies to problems such as refraction[1] and the RBP. This result—of which a preliminary announcement was made in 2004 in [12]—is a special case of several far-reaching extensions of the PMP proved by us in other papers (cf. [10, 11, 12]) that are much longer and more technical. We choose to isolate this particular aspect of the general results and present it separately because it lends itself to a relatively simple and self-contained treatment.

In our version of the PMP, the usual adjoint equation, which contains derivatives with respect to the state, is replaced by an integrated form, containing only differentials of the reference flow maps. In this form, the conditions of the maximum principle make sense for a number of control dynamical laws whose right-hand side can be nonsmooth, nonlipschitz, and even discontinuous. The "adjoint vectors" that are solutions of the "adjoint equation" no longer need to be absolutely continuous, and could even be discontinuous and unbounded. In both the refraction problem and the RBP, the state space is $\mathbb{R}^2$, and the system vector fields are smooth everywhere, except along the x axis. For the refraction problem, the system vector fields are discontinuous, and the adjoint vector turns out to be discontinuous as well, but bounded, having a jump discontinuity at the point where the trajectory crosses the x axis. axis. For the RBP, the system vector fields are Hölder continuous with exponent $1/2$, and—somehwat surprisingly, considering that the RBP vector fields are less irregular than those of the refraction problem— the adjoint vector turns out to be discontinuous with a worse singularity: at the point where the trajectory crosses the x axis, the adjoint vector becomes infinite. And for both problems, the adjoint vector cannot possibly be characterized as a solution of an ordinary differential equation.

2 Preliminaries on Sets, Maps, and Flows

Sets and maps. If S is a set, then $\mathbb{I}_S$ will denote the identity map of S. If A, B are sets, then the notations $f : A \hookrightarrow B$, $f : A \mapsto B$ will indicate, respectively, that f is a possibly partially defined (abbr. "ppd") map from A to B and that f

[1] Some readers may object to our inclusion of the refraction example here, on the grounds that the solution can easily be found by elementary means. Our motivation is identical to that of many authors of calculus of variations textbooks, who choose to include, as an application of the Euler-Lagrange equations, the derivation of the fact that the shortest path joining two points is a straight line segment, even though this can also be proved by completely trivial arguments. In both cases, the purpose is to show that the new necessary condition applies to a very old problem that played a key role in the early history of the subject.

is an everywhere defined map from A to B. If $A \subseteq B$ and $f : B \hookrightarrow C$, then $f\lceil A$ is the restriction of f to A, so $f\lceil A : A \hookrightarrow C$, and $f\lceil A : A \mapsto C$ if $f : B \mapsto C$.

Totally ordered sets. If E is a totally ordered set, with ordering $\preceq$, we use $E^{\preceq,2}$ to denote the set of all ordered pairs $(s,t) \in E \times E$ such that $s \preceq t$, and write $E^{\preceq,3}$ to denote the set of all ordered triples $(r,s,t) \in E \times E \times E$ such that $r \preceq s \preceq t$. A *subinterval* of E is a subset I of E such that, whenever $x \preceq y \preceq z$, $x \in I$, $z \in I$, and $y \in E$, it follows that $y \in I$. If $a \in E$, $b \in E$, and $a \preceq b$, then the *E-interval from a to b* is the set $[a,b]_E \stackrel{\text{def}}{=} \{x \in E : a \preceq x \preceq b\}$.

Manifolds, tangent and cotangent spaces. "Manifold" will mean "smooth manifold", "smooth" means "of class C^∞," and T_xM, T_x^*M denote, respectively, the tangent and cotangent spaces to a manifold M at a point x of M.

Set separation. Let S_1 and S_2 be subsets of a Hausdorff topological space T, and let p be a point of T. We say that S_1 and S_2 are *separated at p* if $S_1 \cap S_2 \subseteq \{p\}$, i.e. if S_1 and S_2 have no common point other than p. We say that S_1 and S_2 are *locally separated at p* if there exists a neighborhood V of p such that $S_1 \cap V$ and $S_2 \cap V$ are separated.

Flows and their trajectories. Every sufficiently well-behaved vector field gives rise to a flow, but flows are typically less well-behaved than the vector fields that generate them. This is a reason for studying flows independently from their generators, as we now do.

Definition 1. Let E be a totally ordered set with ordering $\preceq$, and let Ω be a set. A *flow* on Ω with time set E (or, more simply, a *flow on* (Ω, E)) is a family $\Phi = \{\Phi_{t,s}\}_{(s,t)\in E^{\preceq,2}}$ of ppd maps from Ω to Ω that satisfy the identities (F1) $\Phi_{t,s} \circ \Phi_{s,r} = \Phi_{t,r}$ *whenever* $(r,s,t) \in E^{\preceq,3}$, and (F2) $\Phi_{t,t} = \mathbb{I}_\Omega$ *whenever* $t \in E$.

A *trajectory* of a flow Φ on (Ω, E) is a map $\xi : I \mapsto \Omega$, defined on a subinterval I of E, such that $\xi(t) = \Phi_{t,s}(\xi(s))$ whenever $(s,t) \in I^{\preceq,2}$. ◇

Real augmentation of sets. If Ω is a set, then we will write $\Omega^\# = \mathbb{R} \times \Omega$. If Ω is a smooth manifold, then $\Omega^\#$ is obviously a smooth manifold as well. In that case, if $x^\# = (x_0, x) \in \Omega^\#$, the tangent space $T_{x^\#}\Omega^\#$ and the cotangent space $T^*_{x^\#}\Omega^\#$ will be identified with the products $\mathbb{R} \times T_x\Omega$ and $\mathbb{R} \times T_x^*\Omega$ using the canonical identification maps.

Augmented flows and their trajectories. In optimal control theory, it is often customary to "add the cost variable to the state of a system," thus transforming the optimization problem into a set separation problem in one higher dimension. This augmentation procedure can be carried out directly for flows.

Definition 2. If Φ is a flow on (Ω, E), a *real augmentation* of Φ is a family $c = \{c_{t,s}\}_{(s,t)\in E^{\preceq,2}}$ of ppd functions from Ω to $\mathbb{R}$ such that

(RA) $c_{t,r}(x) = c_{s,r}(x) + c_{t,s}(\Phi_{s,r}(x))$ *whenever* $x \in \Omega$ *and* $(r,s,t) \in E^{\preceq,3}$.

A *flow-augmentation pair* (abbr. F-A pair) on (Ω, E) is a pair (Φ, c) such that Φ is a flow on Ω with time set E and c is a real augmentation of Φ. ◇

(Notice that (RA) implies, in particular, that $c_{t,t}(x) = 0$, since we can always take $r = s = t$ and use the fact that $\Phi_{t,t}(x) = x$.)

To any F-A pair (Φ, c) on (Ω, E) we can associate a family of mappings $\Phi^{\#,c}_{t,s} : \Omega^\# \hookrightarrow \Omega^\#$, by letting $\Phi^{\#,c}_{t,s}(x_0, x) = (x_0 + c_{t,s}(x), \Phi_{t,s}(x))$ for each $(s,t) \in E^{\preceq,2}$. It is then clear that $\Phi^{\#,c} = \{\Phi^{\#,c}_{t,s}\}_{(s,t)\in E^{\preceq,2}}$ is a flow on $\Omega^\#$. A flow Ψ such that $\Psi = \Phi^{\#,c}$ for some Φ, c is called a *real-augmented flow*. It is easy to see that a flow $\Psi = \{\Psi_{t,s}\}_{(s,t)\in E^{\preceq,2}}$ on $\mathbb{R} \times \Omega$ is a real-augmented flow if and only if—if we write $\Psi_{t,s}(x_0, x) = (\psi_{0,t,s}(x_0, x), \psi_{t,s}(x_0, x))$—the maps $\psi_{0,t,s}$, $\psi_{t,s}$ are such that the point $\psi_{t,s}(x_0, x) \in \Omega$ and the number $\psi_{0,t,s}(x_0, x) - x_0 \in \mathbb{R}$ do not depend on x_0. In that case, the pair (Φ, c) such that $\Psi = \Phi^{\#,c}$ is uniquely determined by Ψ as follows: $\Phi_{t,s}(x) = \psi_{t,s}(x_0, x)$ and $c_{t,s}(x) = \psi_{0,t,s}(x_0, x) - x_0$ if $(x_0, x) \in \Omega^\#$.

Definition 3. An *augmented trajectory* of an F-A pair (Φ, c) on a pair (Ω, E) is a trajectory of the flow $\Phi^{\#,c}$, i. e., a map $I \ni t \mapsto \xi^\#(t) = (\xi_0(t), \xi(t)) \in \Omega^\#$, defined on a subinterval I of E, such that $\xi(t) = \Phi_{t,s}(\xi(s))$ and $\xi_0(t) = \xi_0(s) + c_{t,s}(\xi(s))$ whenever $(s,t) \in I^{\preceq,2}$. ◇

Differentiability of flows. Given a trajectory ξ of a flow Φ, it makes sense to talk about differentiability of Φ along ξ.

Definition 4. Assume that Ω is a manifold, $\Phi = \{\Phi_{t,s}\}_{(s,t)\in E^{\preceq,2}}$ is a flow on (Ω, E), I is a subinterval of E, and $\xi : I \mapsto \Omega$ is a trajectory of Φ. We call Φ

(1) *continuous near* ξ if for each $(s,t) \in I^{\preceq,2}$ the map $\Phi_{t,s}$ is continuous on a neighborhood of $\xi(s)$.
(2) *differentiable along* ξ if for each $(s,t) \in I^{\preceq,2}$ the flow map $\Phi_{t,s}$ is differentiable at $\xi(s)$. ◇

The above definition can be applied to an augmented flow $\Phi^{\#,c}$. If $\Phi^{\#,c}$ is differentiable along an augmented trajectory $\xi^\# = (\xi_0, \xi)$, then the differentials $D\Phi^{\#,c}_{t,s}(\xi^\#(s))$ have a special structure, reflecting the special structure of the maps $\Phi^{\#,c}_{t,s}$. Indeed, if we write $M^\#_{t,s} = D\Phi^{\#,c}_{t,s}(\xi^\#(s))$ (so that $M^\#_{t,s}$ is a linear map from $T_{\xi^\#(s)}\Omega^\#$ to $T_{\xi^\#(t)}\Omega^\#$), then it is easy to see that the result $M^\#_{t,s|} \cdot (v_0, v)$ of applying the linear map $D\Phi^{\#,c}_{t,s}(\xi^\#(s))$ to a tangent vector $(v_0, v) \in T_{\xi^\#(s)}\Omega^\# \sim \mathbb{R} \times T_{\xi(s)}\Omega$ is the vector $(v_0 + m_{t,s} \cdot v, M_{t,s} \cdot v)$, which belongs to $T_{\xi^\#(t)}\Omega^\# \sim \mathbb{R} \times T_{\xi(t)}\Omega$, where

$$m_{t,s} = Dc_{t,s}(\xi(s)), \quad M_{t,s} = D\Phi_{t,s}(\xi(s)), \tag{1}$$

so that $m_{t,s} \in T^*_{\xi(s)}\Omega$ and $M_{t,s}$ is a linear map from $T_{\xi(s)}\Omega$ to $T_{\xi(t)}\Omega$.

Variational fields. The differentials $D\Phi_{t,s}(\xi(s))$, $D\Phi^\#_{t,s}(\xi^\#(s))$, can be used to propagate tangent vectors forwards and cotangent vectors backwards.

Definition 5. A field of tangent vectors $I \ni s \mapsto v(s) \in T_{\xi(s)}\Omega$ such that $v(t) = M_{t,s} \cdot v(s)$ whenever $(s,t) \in I^{\preceq,2}$—where $M_{t,s}$ is defined by (1)—is called a *variational vector field of* (Φ, c) *along* ξ. ◇

Definition 6. A field of vectors $I \ni s \mapsto v^{\#}(s) = (v_0(s), v(s)) \in T_{\xi^{\#}(s)}\Omega^{\#}$ such that $v_0(t) = v_0(s) + m_{t,s} \cdot v(s)$ and $v(t) = M_{t,s} \cdot v(s)$ whenever $(s,t) \in I^{\preceq,2}$—where $m_{t,s}$, $M_{t,s}$ are defined by (1)—is called an *augmented variational vector field of* (Φ, c) *along* ξ. ◇

Adjoint fields. The dual maps $\Lambda^{\#}_{t,s} \stackrel{\text{def}}{=} \left(D\Phi^{\#,c}_{t,s}(\xi(s))\right)^{\dagger} : T^*_{\xi^{\#}(t)}\Omega^{\#} \mapsto T^*_{\xi^{\#}(s)}\Omega^{\#}$ (where, naturally, we use the canonical identification $T^*_{\xi^{\#}(r)}\Omega^{\#} \sim \mathbb{R} \times T^*_{\xi(r)}\Omega$ for every r) are given (if we write the maps as acting on the right on augmented covectors) by $(\omega^{\#}(t) \cdot \Lambda^{\#}_{t,s}) \cdot v^{\#}(s) = \omega^{\#}(t) \cdot M^{\#}_{t,s} v^{\#}(s))$, so that, if we write $\omega^{\#}(t) = (\omega_0(t), \omega(t))$, we see that

$$\begin{aligned}(\omega_0(s), \omega(s)) \cdot (v_0(s), v(s)) &= \left((\omega_0(t), \omega(t)) \cdot \Lambda^{\#}_{t,s}\right) \cdot (v_0(s), v(s)) \\ &= \omega_0(t)v_0(s) + \omega_0(t)m_{t,s} \cdot v(s) + (\omega(t) \circ M_{t,s}) \cdot v(s)\,,\end{aligned}$$

and then $\omega_0(s) = \omega_0(t)$ and $\omega(s) = \omega_0 m_{t,s} + \omega(t) \circ M_{t,s}$.

Definition 7. A field of covectors $I \ni s \mapsto \omega^{\#}(s) = (\omega_0(s), \omega(s)) \in T^*_{\xi^{\#}(s)}\Omega^{\#}$ such that ω_0 is a constant function, and ω satisfies

$$\omega(s) = \omega_0 m_{t,s} + \omega(t) \circ M_{t,s} \text{ whenever } (s,t) \in I^{\preceq,2}\,, \tag{2}$$

is called an *augmented adjoint field of covectors* (or *augmented adjoint vector*) of (Φ, c) along ξ. ◇

The constant $-\omega_0$ is the *abnormal multiplier*, and the identity (2) is the *integrated adjoint equation.*

Approximating cones. A *cone* in a real linear space X is a nonempty subset of X which is closed under multiplication by nonnegative scalars, i.e., such that if $c \in C$, $r \in \mathbb{R}$, and $r \geq 0$, it follows that $rc \in C$. (It then follows automatically that $0 \in C$.) The *polar* of a cone C in X is the set $C^{\dagger} = \{\lambda \in X^{\dagger} : \lambda(c) \leq 0 \text{ whenever } c \in C\}$, where $X^{\dagger}$ is the dual space[2] of X.

Definition 8. If M is a smooth manifold, $S \subseteq M$, and $s \in S$, a *Boltyanskii approximating cone to* S *at* s is a convex cone C in T_sM having the property that there exist m, U, D,F, L, such that

(1) $m \in \mathbb{N}$, U is a neighborhood of 0 in $\mathbb{R}^m$, and D is a convex cone in $\mathbb{R}^m$,
(2) $F : U \cap D \mapsto M$ is a continuous map such that $F(U \cap D) \subseteq S$,
(3) $L : \mathbb{R}^m \mapsto T_sM$ is a linear map,
(4) $F(v) - s - L \cdot v = o(\|v\|)$ as $v \to 0$ via values in $U \cap D$,
(5) $L \cdot D = C$.

[2] In all cases occurring in this paper, X is finite-dimensional, so we do not need to distinguish between algebraic and topological duals.

3 The Main Theorem

We consider optimal control problems arising from an *augmented flow system* $\Psi = \{{}_\eta\Psi\}_{\eta\in\mathcal{U}}$, indexed[3] by a "class of admissible controls" $\mathcal{U}$. We assume that a fixed totally ordered set E is specified, such that the time set E_η of each ${}_\eta\Psi$ is equal to E.

The state space of the system is a smooth manifold Ω. Each ${}_\eta\Psi$ is a flow on the real-augmented space $\Omega^\# = \mathbb{R}\times\Omega$, given by ${}_\eta\Psi = {}_\eta\Phi^{\#,\,{}_\eta c}$, where the pair $({}_\eta\Phi,\, {}_\eta c)$ is a real-augmented flow on Ω with time set E. We use $\preceq$ to denote the ordering of E. We assume we are given an initial state $\hat{x}\in\Omega$, a terminal set S, which is a subset of Ω, and initial and terminal times $\hat{a}\in E$, $\hat{b}\in E$, such that $\hat{a}\preceq\hat{b}$.

The objective is to minimize the cost ${}_\eta c_{\hat{b},\hat{a}}(\hat{x})$ in the class $\mathcal{A}$ of all $\eta\in\mathcal{U}$ such that the terminal point ${}_\eta\Phi_{\hat{b},\hat{a}}(\hat{x})$ belongs to S. Equivalently, we want to minimize the cost $\xi_0(\hat{b})-\xi_0(\hat{a})$ in the class $\tilde{\mathcal{A}}$ of all pairs $(\eta,\xi^\#)$ such that $\eta\in\mathcal{U}$, $\xi^\#=(\xi_0,\xi)$ is an augmented trajectory of $({}_\eta\Phi,\, {}_\eta c)$, $\xi(\hat{a})=\hat{x}$ and $\xi(\hat{b})\in S$.

We assume that we are given data

$$\mathcal{D} = (n,\Omega,E,\preceq,\mathcal{U},\Phi,c,\Psi,\hat{a},\hat{b},\hat{x},S)\,, \tag{3}$$

so that $\Psi=\{{}_\eta\Psi\}_{\eta\in\mathcal{U}}$, $\Phi=\{{}_\eta\Phi\}_{\eta\in\mathcal{U}}$, $c=\{{}_\eta c\}_{\eta\in\mathcal{U}}$, and ${}_\eta\Psi={}_\eta\Phi^{\#,\,{}_\eta c}$ for every $\eta\in\mathcal{U}$. We define $\hat{I}=\{t\in E:\hat{a}\preceq t\preceq\hat{b}\}$.

Precisely, we will assume that $\mathcal{D}$ satisfies

(A1) *$n\in\mathbb{N}$, and Ω is a smooth manifold of dimension n;*
(A2) *E is a totally ordered set, with partial ordering $\preceq$;*
(A3) *$\mathcal{U}$ is a set;*
(A4) *$\Psi=\{{}_\eta\Psi\}_{\eta\in\mathcal{U}}$ is an augmented flow system on Ω with time set E;*
(A5) *${}_\eta\Psi={}_\eta\Phi^{\#,\,{}_\eta c}$, where $({}_\eta\Phi,\, {}_\eta c)$ is a flow-augmentation pair on Ω with time set E;*
(A6) *$\hat{x}\in\Omega$, $\hat{a}\in E$, $\hat{b}\in E$, and $\hat{a}\preceq\hat{b}$;*
(A7) *$\hat{a}\in E$, $\hat{b}\in E$, and $\hat{a}\preceq\hat{b}$.*

We let $\tilde{\mathcal{A}}$ be the class of all pairs $(\eta,\xi^\#)$ such that $\eta\in\mathcal{U}$, $\xi^\#=(\xi_0,\xi)$ is an augmented trajectory of $({}_\eta\Phi,\, {}_\eta c)$, $\xi(\hat{a})=\hat{x}$, and $\xi(\hat{b})\in S$.

We assume that we are given a candidate control η_* and candidate augmented trajectory $\xi^\#_*=(\xi_{*,0},\xi_*)$, such that

(A8) $(\eta_*,(\xi_{*,0},\xi_*))=(\eta_*,\xi^\#_*)\in\tilde{\mathcal{A}}$.

Clearly, then, the three maps $\xi_*:\hat{I}\mapsto\Omega$ (the "reference trajectory"), $\xi_{0,*}:\hat{I}\mapsto\mathbb{R}$ (the "reference running cost"), $\xi^\#_*:\hat{I}\mapsto\Omega^\#$ (the "reference augmented trajectory"), satisfy, for all $t\in\hat{I}$,

$$\xi_*(t)={}_\eta\Phi_{t,\hat{a}}(\hat{x}_*)\,,\qquad \xi_{0,*}(t)={}_\eta c_{t,\hat{a}}(\hat{x}_*)\,,\qquad \xi^\#_*(t)={}_\eta\Phi^{\#,\,{}_\eta c}_{t,a}(0,\hat{x}_*)\,,$$

as well as $\xi^\#_*(t)=(\xi_{0,*}(t),\xi_*(t))$.

[3] We put the subscript η on the left because we will want to write formulas such as ${}_\eta\Psi={}_\eta\Phi^{\#,\,{}_\eta c}$, ${}_\eta\Psi=\{{}_\eta\Psi_{t,s}\}_{(s,t)\in E^\preceq,2}$ and ${}_\eta\Phi^{\#,\,{}_\eta c}=\{{}_\eta\Phi^{\#,\,{}_\eta c}_{t,s}\}_{s,t\in E_\eta}$.

Our key assumption is that the pair $(\eta_*, \xi_*^\#)$ is a solution of our optimal control problem, that is, that

(A9) $\xi_{*,0}(\hat{b}) - \xi_{*,0}(\hat{a}) \leq \xi_0(\hat{b}) - \xi_0(\hat{a})$ *for all* $(\eta, (\xi_0, \xi)) \in \tilde{\mathcal{A}}$.

In addition, we make the crucial technical assumption that

(A10) *The reference flow* ${}_{\eta_*}\Psi$ *is continuous near the reference trajectory* $\xi_*^\#$, *and differentiable along* $\xi_*^\#$.

We define an *impulse vector* for the data 12-tuple $\mathcal{D}$ and the reference control-augmented trajectory pair $(\eta_*, \xi_*^\#)$ to be a pair $(v^\#, t)$ such that $t \in \hat{I}$ and $v^\# \in T_{\xi_*^\#(t)}\Omega^\#$. We use $\mathcal{V}_{max}(\mathcal{D}, (\eta_*, \xi_*^\#))$ to denote the set of all impulse vectors for $\mathcal{D}$, $(\eta_*, \xi_*^\#)$.

Let $\mathbf{v}^\# = ((v_1^\#, t_1), \ldots, (v_m^\#, t_m))$ be a finite sequence of members of $\mathcal{V}_{max}(\mathcal{D}, (\eta_*, \xi_*^\#))$, and assume that (A10) holds. We then define linear maps $L_0^{\mathcal{D},\eta_*,\xi_*^\#,\mathbf{v}^\#} : \mathbb{R}^m \times T_{\hat{x}}\Omega \mapsto \mathbb{R}$, $L^{\mathcal{D},\eta_*,\xi_*^\#,\mathbf{v}^\#} : \mathbb{R}^m \times T_{\hat{x}}\Omega \mapsto T_{\xi_*(\hat{b})}\Omega$, and $L^{\#,\mathcal{D},\eta_*,\xi_*^\#,\mathbf{v}^\#} : \mathbb{R}^m \times T_{\hat{x}}\Omega \mapsto T_{\xi_*^\#(\hat{b})}\Omega^\#$, by first writing $v_j^\# = (v_{0,j}, v_j)$, with $v_{0,j} \in \mathbb{R}$, $v_j \in T_{\xi_*(t)}\Omega$, and then letting

$$L_0^{\mathcal{D},\eta_*,\xi_*^\#,\mathbf{v}^\#}(\varepsilon_1, \ldots, \varepsilon_m, w) = m_{\hat{b},\hat{a}} \cdot w + \sum_{j=1}^{m} \varepsilon_j (v_{0,j} + m_{\hat{b},t_j} \cdot v_j),$$

$$L^{\mathcal{D},\eta_*,\xi_*^\#,\mathbf{v}^\#}(\varepsilon_1, \ldots, \varepsilon_m, w) = M_{\hat{b},\hat{a}} \cdot w + \sum_{j=1}^{m} \varepsilon_j M_{\hat{b},t_j} \cdot v_j,$$

$$L^{\#,\mathcal{D},\eta_*,\xi_*^\#,\mathbf{v}^\#}(\varepsilon_1, \ldots, \varepsilon_m, w) = (L_0^{\mathcal{D},\eta_*,\xi_*^\#,\mathbf{v}^\#}(\varepsilon_1, \ldots, \varepsilon_m, w), L^{\mathcal{D},\eta_*,\xi_*^\#,\mathbf{v}}(\varepsilon_1, \ldots, \varepsilon_m, w)).$$

In the following definition, $\mathbb{R}_+^m$ denotes the nonnegative orthant of $\mathbb{R}^m$, that is, the set $\{(h_1, \ldots, h_m) \in \mathbb{R}^m : h_1 \geq 0, \ldots, h_m \geq 0\}$. Furthermore, for any $a, b \in E$ such that $a \preceq b$, and any $x \in \Omega$, $\mathcal{R}_{a,b}^\#(x)$ denotes the "reachable set from x for the augmented system over the interval from a to b," so that

$$\mathcal{R}_{a,b}^\#(x) \stackrel{\text{def}}{=} \left\{(r, y) : (\exists \eta \in \mathcal{U})\Big(y = {}_\eta\Phi_{b,a}(x), \text{ and } r = {}_\eta c_{b,a}(x)\Big)\right\}.$$

Definition 9. A set $\mathcal{V}$ of impulse vectors is *variational* for $\mathcal{D}$, η_*, $\xi_*^\#$ if for every finite sequence $\mathbf{v}^\# = ((v_1^\#, t_1), \ldots, (v_m^\#, t_m))$ of members of $\mathcal{V}$ it follows that there exist neighborhoods P, Q of 0, $\hat{x}$, in $\mathbb{R}_+^m$, Ω, respectively, and a continuous map $F : P \times Q \mapsto \Omega^\#$, such that

(1) F *is differentiable at* $(0, \hat{x})$ *with differential* $L^{\#,\mathcal{D},\eta_*,\xi_*^\#,\mathbf{v}^\#}$ *(in the precise sense of Remark 1 below).* ◇

(2) $F(P \times \{x\}) \subseteq \mathcal{R}_{\hat{a},\hat{b}}^\#(x)$ for every $x \in Q$.

Remark 1. The precise meaning of the assertion that "F is differentiable at $(0,\hat{x})$ with differential $L^{\#,\mathcal{D},\eta_*,\xi_*^\#,\mathbf{v}^\#}$" is as follows. Let $\hat{y} = \xi_*(\hat{b})$, $\hat{y}_0 = \xi_{0,*}(\hat{b})$, $\hat{y}^\# = (\hat{y}_0, \hat{y})$, so $\hat{y}^\# = F(0,\hat{x})$. Let $\tilde{P}$, $\tilde{Q}$, R, J be open neighborhoods of 0, $\hat{x}$, $\hat{y}_0$, $\hat{y}$, in $\mathbb{R}^m$, Ω, $\mathbb{R}$, Ω, respectively, such that $\tilde{P} \subseteq P$, $\tilde{Q} \subseteq Q$, $F((\tilde{P} \cap \mathbb{R}^m_+) \times \tilde{Q}) \subseteq J \times R$, $\tilde{Q}$ is the domain of a coordinate chart $\kappa : \tilde{Q} \mapsto \mathbb{R}^n$ for which $\kappa(\hat{x}) = 0$, and R is the domain of a coordinate chart $\zeta : R \mapsto \mathbb{R}^n$ for which $\zeta(\hat{y}) = 0$. Use κ and ζ to identify the sets $\tilde{Q}$ and R with their images $\kappa(\tilde{Q})$, $\zeta(R)$, so $\tilde{Q}$ and R are now open subsets of $\mathbb{R}^n$. Then

$$F(\varepsilon_1, \ldots, \varepsilon_m, w) = L^{\#,\mathcal{D},\eta_*,\xi_*^\#,\mathbf{v}^\#}(\varepsilon_1, \ldots, \varepsilon_m, w) + o(\varepsilon_1 + \ldots + \varepsilon_m + \|w\|) \quad (4)$$

as $(\varepsilon_1, \ldots, \varepsilon_m, w)$ goes to 0 via values in $(\tilde{P} \cap \mathbb{R}^m_+) \times \tilde{Q}$.

Our last two assumptions are

(A11) *$\mathcal{V}$ is a variational set of impulse vectors for $\mathcal{D}$, η_*, $\xi_*^\#$.*
(A12) *C is a Boltyanskii approximating cone to S at $\xi_*(\hat{b})$.*

The following is then our main result.

Theorem 1. *Assume that we are given a data 12-tuple $\mathcal{D}$ as in (3), as well as η_*, $\xi_*^\#$, $\mathcal{V}$, C, such that Assumptions (A1) to (A12) hold. Write*

$$\Phi^*_{t,s} = {}_{\eta_*}\Phi_{t,s}\,, \quad c^*_{t,s} = {}_{\eta_*}c_{t,s}\,, \quad M_{t,s} = D\Phi^*_{t,s}(\xi_*(s))\,, \quad m_{t,s} = Dc^*_{t,s}(\xi_*(s))\,,$$

(so that $m_{t,s} = \nabla c^*_{t,s}(\xi_*(s))$). Then there exist a map $\hat{I} \ni t \mapsto \omega(t) \in \mathbb{R}_n$ and a real constant ω_0 such that

(1) $\omega_0 \geq 0$,
(2) $(\omega_0, \omega(t)) \neq (0,0)$ *for all* $t \in I$,
(3) $\omega(s) = \omega(t) \cdot M_{t,s} - \omega_0 m_{t,s}$ *whenever* $s, t \in E$ *and* $s \leq t$,
(4) $\langle \omega(t), v\rangle - \omega_0 v_0 \leq 0$ *whenever* $(v^\#, t) = ((v_0, v), t) \in \mathcal{V}$,
(5) the transversality condition $-\omega(\hat{b}) \in C^\dagger$ *holds.*

Proof. Fix a norm $\|\cdot\|$ on the tangent space $T_{\xi_*(\hat{b})}\Omega$. Let $\mathcal{K}$ be the set of all pairs $(\tilde{\omega}_0, \bar{\omega}) \in \mathbb{R} \times T_{\xi_*(\hat{b})}\Omega$ such that $\tilde{\omega}_0 \leq 0$, $|\tilde{\omega}_0| + \|\bar{\omega}\| = 1$, and $-\bar{\omega} \in C^\dagger$. Then $\mathcal{K}$ is clearly compact. For each subset $\mathcal{W}$ of $\mathcal{V}$, let $\mathcal{K}(\mathcal{W})$ be the subset of $\mathcal{K}$ consisting of all $(\tilde{\omega}_0, \bar{\omega}) \in \mathcal{K}$ such that, for all $((v_0, v), t) \in \mathcal{W}$, the identity $\left\langle \bar{\omega} \cdot D_{\eta_*}\Phi_{\hat{b},t}(\xi_*(t)) + \tilde{\omega}_0 \nabla c_{\hat{b},t}(\xi_*(t)), v \right\rangle + \tilde{\omega}_0 v_0 \leq 0$ holds.

It clearly suffices to prove that the set $\mathcal{K}(\mathcal{V})$ is nonempty. Indeed, if a pair $(\tilde{\omega}_0, \bar{\omega})$ belongs to $\in \mathcal{K}(\mathcal{V})$, we may define $\omega_0 = -\tilde{\omega}_0$ and then, for for $t \in I$, let $\omega(t) = \bar{\omega} \cdot M_{\hat{b},t} + \tilde{\omega}_0 m_{\hat{b},t}$ (that is, $\omega(t) = \bar{\omega} \cdot M_{\hat{b},t} - \omega_0 m_{\hat{b},t}$) for $t \in I$. Write $\omega^\# = (\omega_0, \omega)$. A simple calculation shows that $\omega^\#$ is an augmented adjoint vector that satisfies all our conclusions.

Furthermore, it is evident from the definition of the sets $\mathcal{K}(\mathcal{W})$ that if a subset $\mathcal{W}$ of $\mathcal{V}$ is the union $\cup_{\lambda \in \Lambda} \mathcal{W}_\lambda$ of a family of subsets $\mathcal{V}$, then

$$\mathcal{K}(\mathcal{W}) = \cap_{\lambda \in \Lambda} \mathcal{K}(\mathcal{W}_\lambda)\,.$$

Hence it suffices to prove that $\mathcal{K}(\mathcal{W})$ is nonempty whenever $\mathcal{W}$ is a finite subset of $\mathcal{V}$.

So let $\mathcal{W}$ be a finite subset of $\mathcal{V}$. Let $\mathbf{v}^{\#} = ((v_1^{\#}, t_1), \ldots, (v_m^{\#}, t_m))$ be a finite sequence that contains all the members of $\mathcal{W}$, and write $v_j^{\#} = (v_{0,j}, v_j)$ for $j = 1, \ldots, m$. Since $\mathcal{V}$ is variational, Definition 3 enables us to pick neighborhoods P, Q, R of 0, $\hat{x}(= \xi_*(\hat{a}))$, $\xi_*^{\#}(\hat{b})$, in $\mathbb{R}_+^m$, Ω, $\Omega^{\#}$, respectively, and a continuous map $F : P \times Q \mapsto R$, which is differentiable at $(0, \hat{x})$ with differential $L^{\#,\mathcal{D},\eta_*,\xi_*^{\#},\mathbf{v}^{\#}}$, so that F satisfies—relative to coordinate charts near $\hat{x}$, $\xi_*(\hat{b})$ for which $\hat{x} = 0$ and $\xi_*(\hat{b}) = 0$—the condition

$$F(\varepsilon_1, \ldots, \varepsilon_m, w) = L^{\#,\mathcal{D},\eta_*,\xi_*^{\#},\mathbf{v}^{\#}}(\varepsilon_1, \ldots, \varepsilon_m, w) + o(\varepsilon_1, \ldots, \varepsilon_m + \|w\|) \quad (5)$$

as $(\varepsilon_1, \ldots, \varepsilon_m, w)$ goes to $(0, 0)$ via values in $\mathbb{R}_+^m \times Q$, as well as the property that $F(P \times \{x\}) \subseteq \mathcal{R}_{\hat{a},\hat{b}}^{\#}(x)$ for every $x \in Q$.

In particular, if we let $G : P \mapsto \Omega^{\#}$ be the map given by

$$G(\varepsilon_1, \ldots, \varepsilon_m) = F(\varepsilon_1, \ldots, \varepsilon_m; 0)\,,$$

then G is a continuous map into $\mathcal{R}_{\hat{a},\hat{b}}^{\#}(\hat{x})$ which is differentiable at 0 with differential $\check{L}^{\#,\mathcal{D},\eta_*,\xi_*^{\#},\mathbf{v}^{\#}}$, where $\check{L}^{\#,\mathcal{D},\eta_*,\xi_*^{\#},\mathbf{v}^{\#}}$ is the map

$$(\varepsilon_1, \ldots, \varepsilon_m) \mapsto \Big(\sum_{j=1}^{m} \varepsilon_j (v_{0,j} + m_{\hat{b},t_j} \cdot v_j)\,,\ \sum_{j=1}^{m} \varepsilon_j M_{\hat{b},t_j} \cdot v_j \Big)\,,$$

and $M_{t,s}$, $m_{t,s}$, are defined by

$$M_{t,s} = D\Phi_{t,s}(\xi_*(s))\,, \quad m_{t,s} = \nabla c_{t,s}(\xi_*(s)) \text{ for } s, t \in I,\ s \preceq t\,.$$

Let $w_j^{\#} = \check{L}^{\#,\mathcal{D},\eta_*,\xi_*^{\#},\mathbf{v}^{\#}} \cdot e_j^m$, where $e_j^m = (\delta_j^1, \ldots, \delta_j^m)$, and the δ_j^k are the Kronecker symbols. Then $w_j^{\#} = (v_{0,j} + m_{\hat{b},t_j} \cdot v_j\,,\, M_{\hat{b},t_j} \cdot v_j)$ and and $\check{L}^{\#,\mathcal{D},\eta_*,\xi_*^{\#},\mathbf{v}^{\#}}(\varepsilon_1, \ldots, \varepsilon_m) = \sum_{j=1}^m \varepsilon_j w_j^{\#}$.

It is then clear (by applying Definition 8, with $D = \mathbb{R}_+^m$, $L = \check{L}^{\#,\mathcal{D},\eta_*,\xi_*^{\#},\mathbf{v}^{\#}}$) that, if we write $\mathcal{C}$ to denote the convex cone $\mathcal{C}$ generated by the vectors $w_j^{\#}$ (so that $\mathcal{C} = \check{L}^{\#,\mathcal{D},\xi_*^{\#},\eta_*,\mathbf{v}^{\#}}(\mathbb{R}_+^m)$), then $\mathcal{C}$ is a Boltyanskii approximating cone to the augmented reachable set $\mathcal{R}_{\hat{a},\hat{b}}^{\#}(\hat{x})$ at $\xi_*^{\#}(\hat{b})$.

Now, let $S^{\#} = \{(x_0, x) \in \Omega^{\#} : x \in S \text{ and } x_0 \leq \xi_{0,*}(\hat{b}) - \xi_{0,*}(\hat{a}) - \psi(x)\}$, where ψ is a smooth function on Ω that vanishes at $\xi_*(\hat{b})$ and is strictly positive everywhere else. Let $C^{\#} =]-\infty, 0] \times C$. Then $C^{\#}$ is a Boltyanskii approximating cone to $S^{\#}$ at $\xi_*^{\#}(\hat{b})$. Furthermore, it is easy to see that the optimality of $(\eta_*, \xi_*^{\#})$ implies that $\mathcal{R}_{\hat{a},\hat{b}}^{\#}(\hat{x})$ at $\xi_*^{\#}(\hat{b})$ and $S^{\#}$ are separated at $\xi_*^{\#}(\hat{b})$. Then standard set separation theorems tells us that the cones $\mathcal{C}$ and $C^{\#}$ are not strongly transversal. Since $C^{\#}$ is not a linear subspace, the cones $\mathcal{C}$ and $C^{\#}$ are in fact not transversal. This implies that there exists a nonzero covector $\bar{\omega}^{\#} = (\tilde{\omega}_0, \bar{\omega} \in T_{\xi_*^{\#}(\hat{b})}\Omega^{\#}$ such

that $\langle \bar\omega^\#, z\rangle \geq 0$ whenever $z \in C^\#$, and $\langle \bar\omega^\#, z\rangle \leq 0$ whenever $z \in \mathcal{C}$. It follows that $-\bar\omega \in C^\dagger$, and also that $\tilde\omega_0 \leq 0$. $\omega(t) = \bar\omega \cdot D_{\eta_*}\Phi_{\hat b,t}(\xi_*(t)) + \tilde\omega_0 \nabla c_{\hat b,t}(\xi_*(t))$ for $t \in I$. If $j = 1, \ldots, m$, then

$$\begin{aligned} 0 \geq \langle \bar\omega^\#, w_j^\# \rangle &= \Big\langle \Big((\tilde\omega_0, \bar\omega), (v_{0,j} + m_{\hat b,t_j} \cdot v_j \,,\, M_{\hat b,t_j} \cdot v_j) \Big) \Big\rangle \\ &= \tilde\omega_0 v_{0,j} + \tilde\omega_0\, m_{\hat b,t_j} \cdot v_j + \bar\omega \cdot M_{\hat b,t_j} \cdot v_j \\ &= \tilde\omega_0 v_{0,j} + \Big(\tilde\omega_0\, m_{\hat b,t_j} + \bar\omega \cdot M_{\hat b,t_j} \Big) \cdot v_j \\ &= \tilde\omega_0 v_{0,j} + \omega(t_j) \cdot \cdot v_j \,. \end{aligned}$$

This shows that $\bar\omega^\# \in \mathcal{K}(\mathcal{W})$, so $\mathcal{K}(\mathcal{W}) \neq \emptyset$, completing our proof. ◇

4 Variable Time Problems

A minimum time problem is, by its very nature, a variable time-interval problem. Hence such a problem does not fit the framework of our main theorem, if we require that the time set E be a subset of $\mathbb{R}$, and that the time from s to t be precisely $t - s$. It is possible, however, to apply Theorem 1 to minimum time problems, and to more general variable time-interval problems, by means of a simple device. Assume that we start with a situation in which E is a subset of $\mathbb{R}$ and our flow-augmentation pairs $({}_\eta\Phi,\ {}_\eta c)$ are such that that ${}_\eta c_{t,s}(x) = t - s$ whenever $(s,t) \in E^{\preceq,2}$. We want to change our point and think of E as representing a "pseudotime" which is no longer physical time, although it will correspond to physical time along the reference trajectory—for example, in the form of a clock that displays at each $t \in E$ the value t. For this purpose, we allow "insertion variations" in which the reference augmented flow map ${}_{\eta_*}\Phi^{\#,\,\eta_*^c}_{\hat b,\hat a}$ is replaced by the map ${}_{\eta_*}\Phi^{\#,\,\eta_*^c}_{\hat b,t}\ {}_\eta\Phi^{\#,\,\eta^c}_{t+\varepsilon,t}\ {}_{\eta_*}\Phi^{\#,\,\eta_*^c}_{t,\hat a}$ *still regarded mas a transition map from "time"* $\hat a$ *to "time"* $\hat b$, even though the true physical time $\xi_0(\hat b) - \xi_0(\hat a)$ during which this transition occurs is $\hat b - \hat a + \varepsilon$. We also allow "deletion variations" in which the reference augmented flow map ${}_{\eta_*}\Phi^{\#,\,\eta_*^c}_{\hat b,\hat a}$ is replaced by the map ${}_{\eta_*}\Phi^{\#,\,\eta_*^c}_{\hat b,t}\ {}_\eta\Phi^{\#,\,\eta_*^c}_{t-\varepsilon,\hat a}$, again regarded as a transition map from "time" $\hat a$ to "time" $\hat b$, even though the true physical time $\xi_0(\hat b) - \xi_0(\hat a)$ of this transition is $\hat b - \hat a - \varepsilon$. (Naturally, for this to be possible, we need, for example, to be able to regard ${}_{\eta_*}\Phi^{\#,\,\eta_*^c}_{\hat b,t}$ as a "time $t + \varepsilon$ to time $\hat b + \varepsilon$" map. The key condition needed for all this to work is to have a *time-translation invariant* system, that is, a system for which ${}_\eta\Phi_{t,s} = {}_\eta\Phi_{t+\alpha,s+\alpha}$ for all $\alpha \in \mathbb{R}$.)

The variational impulses $(v^\#, t)$ that occur in our main theorem are, in general, of a special form. First of all, for each $t \in E$ that occurs in one of the members $(v^\#, t) \in \mathcal{V}$, there exists a vector $v^\#_{del}(t)$, depending on t but not on $v^\#$, that corresponds to the "deletion of the reference control on intervals of legth ε." Second, for each $v^\# \in \mathcal{V}[t]$—where $\mathcal{V}[t] = \{v^\# : (v^\#, t) \in \mathcal{V}\}$—the vector $v^\#$ corresponds to the "deletion of the reference control on intervals of legth ε

followed by an insertion of some other control on an interval of lentgh ε, " so that $v^{\#}_{ins} \stackrel{\text{def}}{=} v^{\#} - v^{\#}_{del}(t)$ corresponds to an insertion without deletion, and then $v^{\#} = v^{\#}_{ins} + v^{\#}_{del}(t)$. If we allow the insertions to be carried out without a corresponding deletion we get, in addition to the inequalities $\langle \omega^{\#}(t), v^{\#} \rangle \leq 0$ that occur in (4) of the statement of Theorem 1, the new inequalities $\langle \omega^{\#}(t), v^{\#}_{ins} \rangle \leq 0$. If we also allow deletions to be carried out without a corresponding insertion, we get the inequalities $\langle \omega^{\#}(t), v^{\#}_{del}(t) \rangle \leq 0$. On the other hand, one of the controls that can be used in an insertion is the reference control itself, and this insertion corresponds to the vector $-v^{\#}_{del}(t)$, yielding the inequality $\langle \omega^{\#}(t), -v^{\#}_{del}(t) \rangle \leq 0$. So $\langle \omega^{\#}(t), v^{\#}_{del}(t) \rangle = 0$. and $\langle \omega^{\#}(t), -v^{\#}_{del}(t) \rangle = 0$. In other words,

(%) *for a variable time interval problem where the impulses* $(v^{\#}, t)$ *admit the decomposition* $v^{\#} = v^{\#}_{ins} + v^{\#}_{del}(t)$ *as above, the conclusion of Theorem 1, that* $\langle \omega^{\#}(t), v^{\#} \rangle \leq 0$ *—which is equivalent to the inequality* $\langle \omega^{\#}(t), v^{\#}_{ins} \rangle \leq \langle \omega^{\#}(t), -v^{\#}_{del}(t) \rangle$ *—can be strengthened to*

$$\langle \omega^{\#}(t), v^{\#}_{ins} \rangle \leq \langle \omega^{\#}(t), -v^{\#}_{del}(t) \rangle = 0\,. \tag{6}$$

5 The Reflected Brachistochrone

As a our firt example of a nontrivial application of Theorem 1, we briefly outline the results on the "reflected brachistochrone problem" (RBP), studied in detail in [12].

The RBP is the minimum time problem for the dynamical law $\dot{x} = u\sqrt{|y|}$, $\dot{y} = v\sqrt{|y|}$, with state $(x, y) \in \mathbb{R}^2$ and control $(u, v) \in \mathbb{S}^1$ where $\mathbb{S}^1$ is the unit circle $\{(u, v) \in \mathbb{R}^2 : u^2 + v^2 \leq 1\}$. The goal is to characterize the minimum-time trajectory from A to B, for any two given points $A = (x_A, x_B)$, $B = (x_B, y_B)$ in $\mathbb{R}^2$.

The result is as follows. If $y_A y_B \geq 0$, then the optimal trajectory from is obtained by solving the classical (1696-7) *brachistochrone problem* (BP) of Johann Bernoulli. The solution is an arc of cycloid if $x_A \neq x_B$, and a straight-line segment when $x_A = x_B$. The most interesting case is when $y_A y_B < 0$. The classical results on the BP tell us that the solution consists of an arc of cycloid ξ_1 from a to a point C in the x axis, followed by another arc of cycloid ξ_2 from C to B. The point C is determined by applying Theorem 1, and has to be such that *the rolling circles that generate the* ξ_1 *and* ξ_2 *have equal radii.*

6 Snell's Law of Refraction

We consider the minimum time problem for the two-dimensional system $\dot{x} = c(x, y)u$, $\dot{y} = c(x, y)v$, where the control (u, v) takes values in the unit circle $\mathbb{S}^1$, and the function c (the "speed of light") is given by $c(x, y) = c_+$ if $y \geq 0$ and $c(x, y) = c_-$ if $y < 0$. Here c_+ and c_- are two fixed positive constants such that $c_+ > c_-$. We will focus on the problem of finding a time-minimizing arc from a point $A = (x_A, y_A)$ such that $y_A > 0$ to a point $B = (x_B, y_B)$ such that $y_B < 0$.

The solution of this problem—Snell's law of refraction—is well known, and can be derived by very elementary means: first, one shows that the solution must consist of a straight segment from A to a point C lying on the x axis, followed by the segment from C to B; finding C then becomes a rather simple first-year calculus exercise. Here we will show how our version of the Maximum Principle applies to this problem, and leads to Snell's law.

We take our control set U to be the product $\mathbb{S}^1 \times \mathbb{S}^1$, and then define, for each $z = (u_+, v_+, u_-.v_-) \in U$, a discontinuous vector field X_z by letting $X_z(x,y) = (c_+u_+, c_+v_+)$ if $y > 0$, and $X_z(x,y) = (c_-u_-, c_-v_-)$ if $y \leq 0$. We let G be the subset of U consisting of those $(u_+, v_+, u_-.v_-) \in U$ such that $v_+ < 0$ and $v_- < 0$. We use L to denote the x axis.

An elementary argument shows that an optimal trajectory ξ_* must consist of a segment from A to C followed by a segment from C to B, where $C \in L$. That is, we can confine ourselves to a trajectory $\xi_* : [0,T] \mapsto \mathbb{R}^2$ such that (i) $\xi_*(0) = A$, (ii) $\xi_*T) = B$, (iii) $\xi_*(\tau) \in L$ for some τ such that $0 < \tau < T$, (iv) if $\xi_*(t) = (x_*(t), y_*(t))$ for $t \in [0,T]$, then $y_*(t) > 0$ for $0 \leq t < \tau$ and $y_*(t) < 0$ for $\tau < t \leq T$, and (v) the curve ξ_* is a trajectory of a *constant* control $z_* \in U$.

All that is left now is to find a condition that will determine C. With our choice of U, constant controls have two degrees of freedom, but one is removed when we stipulate that ξ_*, starting at A, has to go through B, so we need to find an extra constraint on z_*.

Let us compute the flow of X_z for a $z \in G$. It suffices to compute the maps $\Phi^{X_z}_{t,0}$, since $\Phi^{X_z}_{t,s} = \Phi^{X_z}_{t-s,0}$.

If $y > 0$, $t > 0$, and we let $(\tilde{x}, \tilde{y}) = \Phi^{X_z}_{t,0}(x,y)$, then $y > \tilde{y}$ and, in addition, $\tilde{y} > 0$ as long as $t < \tau_z(x,y)$, where $\tau_z(x,y)$ is the time for which $\Phi^{X_z}_{\tau_z(x,y),0}(x,y) \in L$. It is clear that $\tau_z(x,y) = -\frac{y}{c_+v_+}$, and also that

$$\Phi^{X_z}_{t,0}(x,y) = (x + tc_+u_+\,,\, y + tc_+v_+) \text{ if } 0 < t < \tau_z(x,y)\,,$$
$$\Phi^{X_z}_{\tau_z(x,y)+t,0}(x,y) = (x + \tau_z(x,y)c_+u_+ + tc_-u_-\,,\, tc_-v_-) \text{ if } t > 0\,.$$

In particular, given a t such that $t \neq \tau_z(x_A, y_A)$, the flow map $\Phi^{X_z}_{t,0}$ is of class C^1 near A, and is given, for (x,y) in some neighborhood $N(t)$ of A, by

$$\begin{aligned}
\Phi^{X_z}_{t,0}(x,y) &= (x + tc_+u_+, y + tc_+v_+) \text{ if } 0 < t < \tau_z(x_A, y_A)\,,\\
\Phi^{X_z}_{t,0}(x,y) &= (x + \tau_z(x,y)c_+u_+ + (t - \tau_z(x,y))c_-u_-\,,\, (t - \tau_z(x,y))c_-v_-)\\
&= \Big(x - \frac{y}{v_+}u_+ + (t + \frac{y}{c_+v_+})c_-u_-\,,\, (t + \frac{y}{c_+v_+})c_-v_-\Big)\\
&= \Big(x + \frac{c_-u_- - c_+u_+}{c_+v_+}y + tc_-u_-, \frac{c_-v_-}{c_+v_+}y + tc_-v_-\Big) \text{ if } t > \tau_z(x_A, y_A)\,.
\end{aligned}$$

It follows from this that the differential $D^z(t)$ of $\Phi^{X_z}_{t,0}$ at A is given by $D^z(t) = \mathbb{I}_{\mathbb{R}^2}$ if $t < -\frac{y_A}{c_+v_+}$ and $D^z(t) = M_z$ if $t > -\frac{y_A}{c_+v_+}$, where M_z is the matrix $\begin{bmatrix} 1 & \frac{c_-u_- - c_+u_+}{c_+v_+} \\ 0 & \frac{c_-v_-}{c_+v_+} \end{bmatrix}$.

We now let $\hat{D}(t)$ denote the differential of the flow map $\Phi_{T,t}^{X_{z_*}}$ at $\xi_*(t)$, where $\xi_* : [0,T] \mapsto\in \mathbb{R}^2$ is our reference trajectory $z_* = (u_{*,+}, v_{*,+}, u_{*,-}.v_{*,-})$ is our constant reference control, and It is then clear that $\hat{D}(t) = D^{z_*}(T)D^{z_*}(t)^{-1}$, and then $\hat{D}(t) = \mathbb{I}_{\mathbb{R}^2}$ if $t > -\frac{y_A}{c_+ v_{*,+}}$, and $\hat{D}(t) = M_{z_*}$ if $t < -\frac{y_A}{c_+ v_{*,+}}$.

To apply our flow version of the Maximum Principle, we take the time set E to be $[0,T]\backslash\{\tau_{z_*}(A)\}$. Then each flow map $\Phi_{t,s}^{X_{z_*}}$ is of class C^1 (and, in fact, real analytic) on a neighborhood of $\xi_*(s)$, as long as $s,t \in E$ and $s \leq t$.

Let ω be the adjoint vector given by the Maximum Principle. Then $\omega(t) = \bar{\omega}^-$ if $t > -\frac{y_A}{c_+ v_{*,+}}$, and $\omega(t) = \bar{\omega}^+$ if $t < -\frac{y_A}{c_+ v_{*,+}}$, where $\bar{\omega}^- = (\bar{\omega}_x, \bar{\omega}_y)$, $\bar{\omega}^+ = (\bar{\omega}_x, \hat{\omega}_y)$, and $\hat{\omega}_y = \frac{1}{c_+ v_{*,+}}\Big((c_- u_{*,-} - c_+ u_{*,+})\bar{\omega}_x + c_- v_{*,-}\bar{\omega}_y\Big)$. The Hamiltonian maximization condition of the Maximum Principle implies that ω^- must be a scalar multiple of $(u_{*,-}, v_{*,-})$, and ω^+ has to be a scalar multiple of $(u_{*,+}, v_{*,+})$. This means that $\bar{\omega}_x = k_- u_{*,-} = k_+ u_{*,+}$, $\bar{\omega}_y = k_- v_{*,-}$, and $\hat{\omega}_y = k_+ v_{*,+}$ for some positive constants k_-, k_+.

It follows that $\frac{k_-}{k_+} = \frac{u_{*,+}}{u_{*,-}}$. Let ω_0 be the abnormal multiplier. Then

$$0 = \langle \bar{\omega}^-, c_-(u_{*,-}, v_{*,-})\rangle - \omega_0 = \langle \bar{\omega}^+, c_+(u_{*,+}, v_{*,+})\rangle - \omega_0\,,$$

so $\langle \bar{\omega}^-, c_-(u_{*,-}, v_{*,-})\rangle = \langle \bar{\omega}^+, c_+(u_{*,+}, v_{*,+})\rangle$.

Furthermore, $\bar{\omega}^- = k_-(u_{*,-}, v_{*,-})$, $\bar{\omega}^+ = k_+(u_{*,+}, v_{*,+})$, and both $(u_{*,-}, v_{*,-})$, and $(u_{*,+}, v_{*,+})$ are unit vectors. It follows that $k_- c_- = k_+ c_+$. Hence $\frac{k_-}{k_+} = \frac{c_+}{c_-}$. Therefore $\frac{u_{*,+}}{u_{*,-}} = \frac{c_+}{c_-}$.

Let θ_i be the "angle of incidence," that is, the angle between the line AC and the y axis. Let θ_r be the "angle of refraction," that is, the angle between the line CB and the y axis. It is then clear that $u_{*,+} = \sin\theta_i$ and $u_{*,-} = \sin\theta_r$. Then

$$\frac{\sin\theta_i}{\sin\theta_r} = \frac{c_+}{c_-}\,, \tag{7}$$

which is precisely Snell's law.

References

1. Pontryagin, L.S., Boltyanskii, V.G., Gamkrelidze, R.V., Mischenko, E.F.: The Mathematical Theory of Optimal Processes. Wiley, New York (1962)
2. Clarke, F.H.: The Maximum Principle under minimal hypotheses. SIAM J. Control Optim. 14, 1078–1091 (1976)
3. Clarke, F.H., Ledyaev, Y.S., Stern, R.J., Wolenski, P.R.: Nonsmooth Analysis and Control Theory. Graduate Texts in Mathematics, vol. 178. Springer, New York (1998)
4. Clarke, F.H.: Necessary conditions in dynamic optimization. Memoirs Amer. Math. Soc. 816, 173 (2005)
5. Ioffe, A.: Euler-Lagrange and Hamiltonian formalisms in dynamic optimization. Trans. Amer. Math. Soc. 349, 2871–2900 (1997)
6. Ioffe, A., Rockafellar, R.T.: The Euler and Weierstrass conditions for nonsmooth variational problems. Calc. Var. Partial Differential Equations 4, 59–87 (1996)

7. Knobloch, H.W.: High Order Necessary Conditions in Optimal Control. Springer, Berlin (1975)
8. Krener, A.J.: The High Order Maximal Principle and Its Application to Singular Extremals. SIAM J. Contrl. Optim. 15, 256–293 (1977)
9. Mordukhovich, B.: Variational Analysis and Generalized Differentiation, I: Basic Theory; II: Applications. Grundlehren Series (Fundamental Principles of Mathematical Sciences), vol. 330 and 331. Springer, Berlin (2006)
10. Sussmann, H.J.: A strong version of the Łojasiewicz Maximum Principle. In: Pavel, N.H. (ed.) Optimal Control of Differential Equations, Athens, OH. Lect. Notes in Pure and Applied Math., vol. 160, pp. 293–309. Marcel Dekker, New York (1993)
11. Sussmann, H.J.: New theories of set-valued differentials and new versions of the maximum principle of optimal control theory. In: Isidori, A., Lamnabhi-Lagarrigue, F., Respondek, W. (eds.) Nonlinear Control in the year 2000, pp. 487–526. Springer, London (2000)
12. Sussmann, H.J.: Optimal control of nonsmooth systems with classically differentiable flow maps. In: NOLCOS 2004. Proc. 6th IFAC Symposium on Nonlinear Control Systems, Stuttgart, Germany, September 1-3, 2004, vol. 2, pp. 609–704 (2004)
13. Vinter, R.B.: Optimal Control. Birkhäuser, Boston (2000)

Safe Operation and Control of Diesel Particulate Filters Using Level Set Methods

Stuart Swift, Keith Glover, and Nick Collings

Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, UK
sjs1014@cam.ac.uk, kg@eng.cam.ac.uk, nc@eng.cam.ac.uk

Summary. Ceramic filters used to trap diesel exhaust particulates must be periodically regenerated to remove the deposited carbon. Running the engine to provide hot and reactive feedgas can initiate a reaction to remove this carbon. Under certain conditions the heat released by this reaction can be more than enough to sustain the reaction and can result in very high local temperatures, which may typically be seen to propagate down the filter as a thermal wave. Stresses in the substrate due to its thermal expansion under these conditions can cause permanent damage to the filter. Simple control relevant models have been developed to investigate when it is safe to initiate a regeneration and when it may be necessary to change the operating conditions to avoid damage to the filter. Level set methods have been used in the reachability analysis of these problems with consideration to the design of engine management strategy.

Keywords: Diesel particulate filter, reachability, safe control, level set methods.

1 Introduction

Diesel particulate filtering was first considered around the end of the 1970s [10] due to concerns about the health impacts of inhaled particulates. Many designs of filter have been investigated, in particular for special applications. The most common type of diesel particulate filter (DPF) is a wall flow design based upon the ceramic monolith used for a three-way catalyst. Blocking the channel ends alternately forces the air flow to pass through the wall of the filter. Thus the porosity, pore size and thickness of the DPF wall determines its basic filtering behaviour.

High exhaust temperatures of around 600°C are required to initiate regeneration. These high temperatures must be sustained for the duration of the regeneration, which typically lasts for at least one minute. Such conditions may occur during normal driving for heavy duty vehicles, but light duty vehicles require deliberately inefficient engine operation or an additional source of heat to reach these temperatures. Throttling of the air intake may assist in raising the exhaust temperature.

A DPF reacting ten grammes of soot per litre of DPF over one minute, producing equal quantities of CO and CO_2, will be releasing heat at a rate of 3.8 kW

V.D. Blondel et al. (Eds.) Recent Advances in Learning and Control, LNCIS 371, pp. 233–247, 2008.
springerlink.com

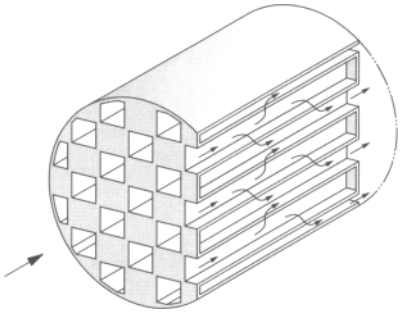


Fig. 1. Diesel particulate filter structure and gas flow paths

per litre, taking the enthalpy of formation for CO and CO_2 as 10 MJ/kg and 35 MJ/kg respectively. Once ignited, the behaviour of this reaction is strongly dependent upon the flow rate and oxygen concentration of the feedgas.

The field of Filtration Combustion has contributed to the development of DPF technology and to the understanding and modelling of DPF behaviour. Modelling of a fibrous filter in Bisset and Shadman [20] was adapted to model DPF behaviour in [5] for favourable high flow rate regeneration conditions. Thermal wave behaviour investigated in Aldushin [1,2] shows similarities to the worst case conditions encountered in DPFs when a sharp exotherm occurs at low flow rates. In this situation the heat released by the burning soot is concentrated in the thermal wave as it travels down the filter, potentially resulting in temperatures high enough to damage the substrate.

In this paper these worst case conditions are represented by a model simple enough for analysis using Level Set Methods. The model parameters are derived using experimental data from a 2.0-litre diesel engine in a dynamometer test cell. These methods were used to simulate the set of all possible trajectories for this model from a bounded set of initial conditions. This is achieved by evolving the boundary of the set over time. This avoids the problems of divergent trajectories and special cases which can be missed by Monte Carlo simulation methods. The set of conditions which lead to excessively high temperatures can be obtained in this way. The boundary of this set can be used as the basis for a supervisory control scheme, indicating when it will be necessary to take action to avoid damage to the filter.

2 Modelling

Early approaches to DPF modelling came from a background of through-flow filtration modelling [5], though the differences between filtration combustion models and the developed DPF models are subtle but significant. Widely referenced, the modelling work of Bisset and Shadman [6] appears to have been generally accepted for the early development of diesel particulate filters. Observations from this paper may have given rise to a variety of modelling directions.

For the case of high gas flow rates it has been argued [12] that the temperature does not vary widely along the DPF channels. This may be somewhat wide of the mark if there is a thick insulating soot layer lining the inlet channels, and will certainly not be true if a thermal wave has already been initiated. This assumption of a small temperature variation through the soot layer also depends upon a high estimate of the soot packing density.

The non-dimensional representative parameter ϵ is assumed to be small ($\epsilon \approx 0.001$) in [6], where

$$\epsilon = c_{p,g} F \omega_b / A \lambda_1$$

with a gas specific heat capacity $c_{p,g}$, mass flow rate of inlet gas F, soot cake layer thickness ω_b, a filtration area A and a bulk thermal conductivity for the soot layer of $\lambda_1 = 0.84$ W/m.K, based upon a bulk density of $\rho_1 = 550$ kg/m^3 for the soot layer. In Haralampous [9] it has been noted that consensus appears to have moved toward much lower soot densities, implying a greater layer thickness and lower thermal conductivity. A value of 75 kg/m^3 is used for the bulk density in [9], corresponding to a soot layer porosity of 0.97 and giving a value of $\lambda_1 = 0.11$ W/m.K for the resulting bulk thermal conductivity. High gas velocities and high soot loadings in conjunction with these very low soot densities give a larger value of $0.1 < \epsilon < 0.4$ corresponding with a noticeable insulating effect of the soot layer. An example case is shown in [9] of a 200 K drop across a 0.35 mm soot cake. Conditions of low flow rate which are less suited to this assumption are more important for the investigation of peak temperatures.

One refinement of this model is the inclusion of oxygen depletion through the soot layer and the resulting effects on the reaction rate. At low temperatures the oxygen will not be completely consumed as it passes through the soot layer and DPF wall, in this case the reaction rate is described as thermally or kinetically limited and is represented by the rate equation (1). Although oxygen depletion through the soot layer is considered in [6], it is assumed that all of the reaction products go with the convective flow into the exhaust channel, whereas arguments from [17] indicate that there may be significant diffusion of gases into and out of the surface layer, contributing to a drop in oxygen concentration down the inlet channel.

The effects of engine operating conditions on the loading of the DPF have been studied in more recent papers, with the presence of partially reacted fuel and oil compounds adsorbed on the soot particles affecting the regeneration behaviour in addition to the soot packing density already mentioned. The presence of this soluble organic fraction (SOF) may increase the reactivity of the soot, lowering the ignition temperature.

2.1 Soot Reaction Model

The form of the soot reaction rate equation has been established for some time [8], it is considered to be in Arrhenius form [6, 11] and can be a reasonably accurate representation of the exothermic reaction which may far exceed the environment temperature once started.

$$K_R = kTy_{O_2} \exp(-E/RT) \tag{1}$$

Where R is the gas constant, E is the activation energy, k is the frequency factor, y_{O_2} is the oxygen concentration, T is the temperature and K_R is the reaction rate (ie. $\dot{m}_c = -K_R \cdot m_c$). The activation energy and frequency factor must be determined empirically and are dependent upon the composition of both the soot deposits and feedgas.

A significant consequence of this type of reaction behaviour is that the reaction temperature may increase rapidly following a drop to low gas flow rates, as the feedgas temperature then has less influence over the local reaction temperatures. Reducing the available reactants, principally oxygen, available in the feedgas can slow the reaction. Increasing exhaust gas recirculation (EGR) alone may not be sufficient to accomplish this as the EGR rate is limited by the flow restriction of the EGR system. Using an intake throttle to further increase the EGR rate can lead to combustion stability issues. An additional injection of fuel late in the combustion cycle can be a more practical way of reducing the exhaust feedgas.

In the development of control strategies the modelling challenge is to extract the significant features and dependencies of the equipment and represent them in a form suitable for the design and analysis of a controller. State-estimation models for on-board calculation may be required by these implemented control strategies. Simplified models are also useful for on-board diagnostics, to determine the status of the DPF from minimal sensor information.

Research in the field of filtration combustion modelling has yielded results of possible interest since the invention of the DPF. The combustion has been modelled as a smouldering wave travelling through the filter [1, 2], with zones representing where the filter is hot enough for ignition, where there is adequate oxygen, and where the filter is still sufficiently loaded to react. Heat transport down the filter is calculated using a representative wave speed. This approach could be useful for producing a DPF model which both captures the important features of regeneration and is simple enough to be useful for control application. A similar modelling approach has been followed here in order to develop this model for level set simulation and reachability analysis.

2.2 Experimental Observations

A catalysed DPF was used in experiments with a 2.0-litre diesel engine to investigate the worst case regeneration behaviour. The filter specifications are shown in table 1 and the layout of installed thermocouples is shown in figure 2.

The *drop-to-idle* experiment studied here shows the effects of a sudden change of engine operating point upon the progress of a steady regeneration. As the feedgas and DPF temperature just rise above 650°C the engine is switched down to idle, reducing the mass flow rate of gas, increasing the oxygen concentration and reducing the feedgas temperature. The increase in oxygen concentration raises the reaction rate of the hot soot, and the decrease in gas flow rate reduces the cooling effect of the gas flow. Available oxygen is rapidly consumed in the local high temperature region of the intake channels, probably reducing the oxygen concentration available down the channel. Once an exotherm becomes established the heat released is carried downstream, where there is available soot

Table 1. Johnson Matthey DPF Specifications

Filter Specifications	Value	Filter Dimensions	Value
Type	Catalysed	Brick Diameter [mm]	140
Substrate Material	SiC	Brick Length [mm]	150
Cell Density [cpsi]	300	Brick Volume [litres]	2.5
DPF Wall Thickness [mm]	0.4	Can Diameter [mm]	160
Porosity [%]	50	Can Length [mm]	340
Specific Heat [J/kg-K]	1120	Can Wall [mm]	1.5

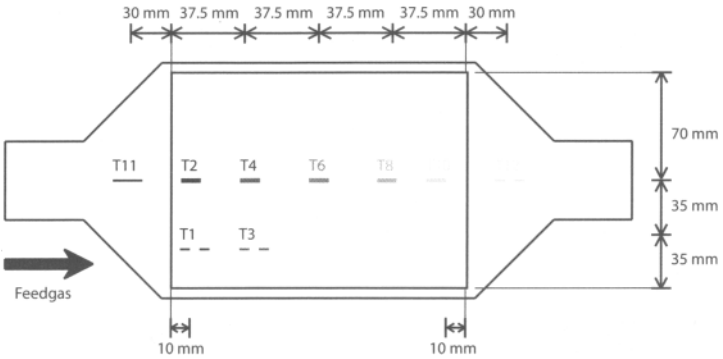


Fig. 2. DPF thermocouple locations

and oxygen the temperature of this exotherm can continue to rise. Figure 3 shows the rapid development of an exotherm and its progression down the filter. The higher initial temperature results in the exotherm becoming established quickly, close to the front of the filter. Heat released in the strongly reacting zone raises the maximum temperature of this hot region to 1040°C by the time it reaches the back of the filter.

2.3 Control Volume Model

Experimental results have shown that the worst-case regeneration conditions occur when an exotherm travels down the filter as a thermal wave. The aim of this modelling work is to represent this behaviour and hence give more accurate results for the worst-case conditions than conventional bulk valued models, whilst keeping the system order low and suitable for use with level set methods. The initial idea was to model a control volume which contains only the reaction zone and moves with this zone as it proceeds down the filter. This control volume has a gas flow in from the engine side and a gas flow out to the exhaust side. It also has a flow of soot coated substrate entering from the exhaust side and a flow of clean substrate leaving towards the engine.

The model states are the representative temperature for the control volume and the positions of the front and rear of the reaction zone. The position x_1

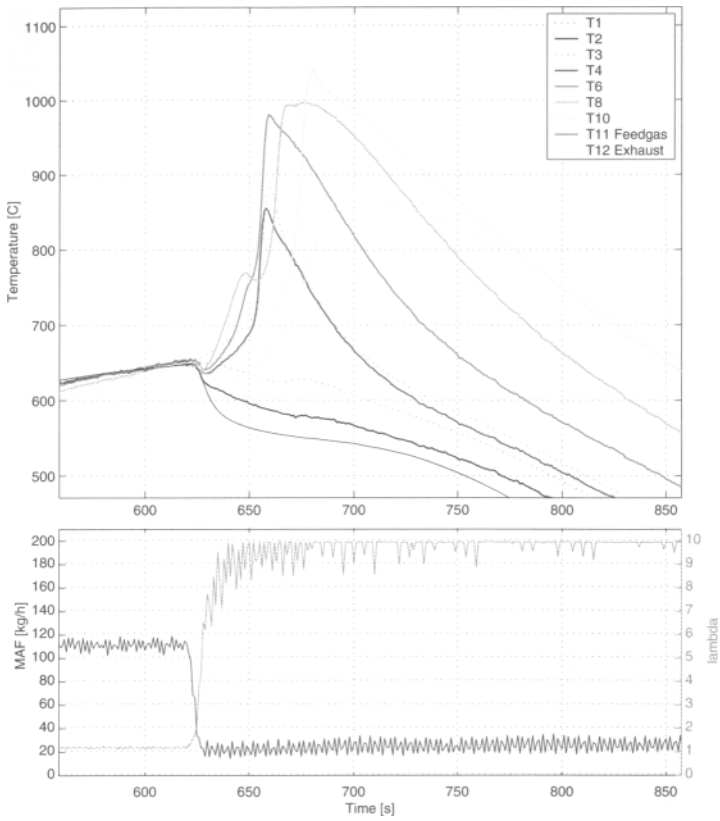


Fig. 3. Drop to idle regeneration experiment

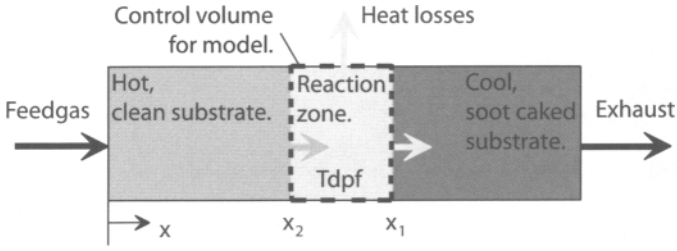


Fig. 4. Control volume model of DPF reaction zone

of the front of the reaction zone is given by the thermal wave speed which is dependent upon various factors including the feedgas flow rate and the insulating effects of the soot lining in the DPF inlet channels. The position x_2 of the rear of the reaction zone is determined by what fraction of the original soot loading

has been removed during the regeneration, thus assuming that the DPF behind the reaction zone has been completely regenerated.

Fitting the model to the experimental data through simulation appears to indicate that the the initial temperature of the DPF substrate is not particularly significant, although it will likely affect the amount of pre-heating from the feedgas required to ignite the soot. Once this heat has been supplied to the DPF by the feedgas it is carried along the filter by heat exchange with the gas and substrate, raising the temperature of the DPF ahead of the reaction zone. Some similarities may be observed with the behaviour of thermal waves in the field of filtration combustion [1,2], where the filter is a packed bed or similar design. As this control volume model assumes that the reaction zone is essentially at a uniform temperature, which typically rises during its movement down the filter, then it may be appropriate to consider the corresponding increase in the peak temperature of the pre-heat region as a result of a heat loss from the reaction zone. Total heat loss from the reaction zone is likely to be much greater however, due to the radial flow of heat to the peripheral substrate, jacket and DPF can. The actual heat loss to the environment from the outside of the can is probably not as significant as the thermal inertia of the can itself during the rapid heating from the reaction zone. These heat losses have been treated as a lumped heat loss from the reaction zone, proportional to the reaction zone temperature, when fitting the model to the observed experimental data.

As the actual variation in oxygen concentration down the inlet channel is unknown, it has been assumed to be uniform. This would be inaccurate if oxygen diffuses out of the inlet channel at a significant rate, or if the available oxygen is consumed by reactions in the inlet channel. The flow rate of gas through the channel wall has been chosen to be constant per unit area in the model, a simplifying assumption as the actual distribution is dependent upon the thickness of the soot coating on the channel wall and the temperature dependent viscosity of the flowing gases. Both the oxygen concentration and the flow rate of gas through the wall in the reaction zone affect the supply of oxygen available to react with the carbon.

A simple calculation for the thermal wave speed, which assumes very good heat transfer between the flowing gas and the substrate in the filter, gives a thermal wave velocity much slower than that observed experimentally.

$$v_{wave} = \frac{\dot{m}_{gas} c_{p,gas}}{A_{filter} \rho_{filter} c_{p,filter}} \tag{2}$$

Where $\dot{m}_{gas}$ is the mass flow rate of the gas, ρ_{filter} is the bulk density of the filter, A_{filter} is the cross sectional area of the filter and $c_{p,gas}$, $c_{p,filter}$ are the specific heat capacities. The model uses this simple estimate multiplied by a factor of $f_C = 3.5$, to give a physically representative thermal wave speed which is still proportional to the gas flow velocity. The precise nature of how the observed wave speed exceeds the simple estimate is not known, but is thought to be due to the insulating soot layer [9] and the exotherm at its surface.

Writing out the dynamic equations for the three model states we have

$$\dot{T} = \frac{-m_{c,cv}(K_R \Delta H)}{m_{c,cv} c_{p,soot} + \rho_{filter} c_{p,filter} V_w}$$

where V_w is the volume of the walls in the control volume and ΔH is the specific heat of the reaction. The position states are given by

$$\dot{x}_1 = v_{wave} \times f_C$$

and

$$\dot{x}_2 = \frac{\dot{m}_c l}{m_{c,init}}$$

with the rate of consumption of soot mass being limited by either the temperature or available oxygen.

$$\dot{m}_c = \min(\dot{m}_{c,Olim}, \dot{m}_{c,Tlim})$$

The temperature limited rate of soot reaction is given by the Arrhenius rate equation

$$\dot{m}_{c,Tlim} = -K_R \cdot m_{c,cv} = -k y_{O_2} T e^{-(\frac{E}{RT})} \cdot m_{c,init} \cdot \frac{x_1 - x_2}{l}$$

and the oxygen limited case is determined by the rate at which oxygen flows through the DPF walls in the reaction zone

$$\dot{m}_{c,Olim} = -\frac{12}{16} \cdot \frac{2}{3} \cdot \dot{m}_{O_2,cv} = -\frac{12}{16} \cdot \frac{2}{3} \cdot \frac{x_1 - x_2}{l} \cdot y_{O_2} \cdot \dot{m}_{feedgas}$$

The initiation of these thermal waves is not well represented by this model, in particular under the drop to idle conditions where the DPF is already hot enough for ignition, but the large gas flow rate through the channel walls prevents the soot temperature rising rapidly above the feedgas temperature.

Additionally, the drop to idle only occurs after the regeneration has begun, so the DPF has already lost some carbon before the change in conditions, this

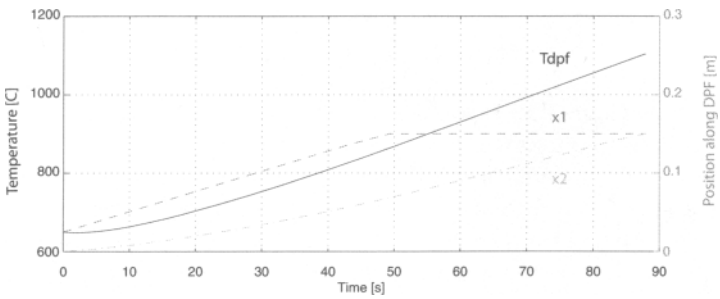


Fig. 5. Drop to idle simulation of control volume model

is not accounted for in the model and may result in higher peak temperatures in the simulations of this model.

Results from a conventional simulation of the control volume model are shown in figure 5, corresponding to the experimental results in figure 3. The positions x_1 and x_2 represent the front of the thermal wave and the back of the reaction zone respectively in these figures. Direct comparison of the simulation results to the experimental results is not completely straightforward as the experimental results are obtained from a limited number of static thermocouples. The estimated growth of the peak temperature in the model is a useful representation of the actual behaviour observed.

3 Level Set Simulation

The theory and methods in this section are taken from [13] and [15], which draw upon the work in [18,19,16], and the software implementation of these algorithms from [14] have been used to compute the results presented in this paper.

An example level set function for a two dimensional set is shown in figure 6 sourced from [13], where the surface of this function can be seen to be an approximate signed distance function. Normally it is not necessary to show the level set function itself, only the set boundaries need to be viewed to work with the level set methods.

In this work the level set function will be evaluated at grid points over the state space. The location of the set boundary may then be specified to less than the grid spacing by interpolating between these points. In fact the accuracy of these level set methods allow the boundary of the state space to be resolved to less than one tenth of the grid spacing over most of the state space. Fortunately it is only strictly necessary to operate on the level set function in the neighbourhood of the the boundary, considerably reducing the overall computational load.

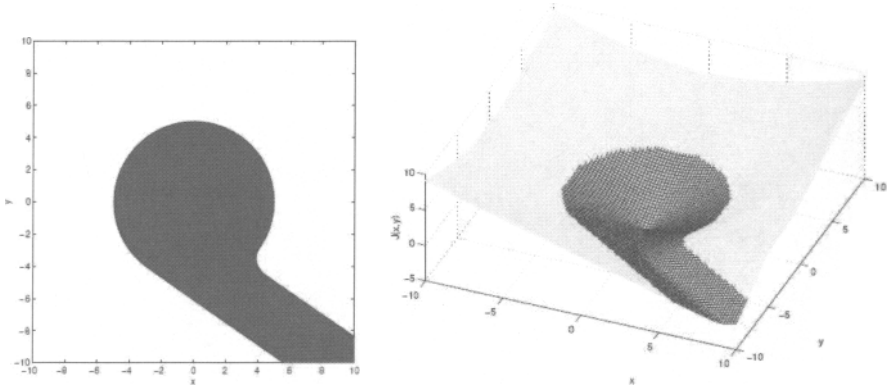


Fig. 6. Level set representation [13]

Simulating systems modelled using non-linear ordinary differential equations $\dot{x} = f(x)$ becomes a *convection* problem involving the level set function $\phi(\cdot)$ in the higher dimensional space. In this case simulations which start within a set of initial conditions $x(0) : \phi(x(0), 0) < 0$ define where and when the level set function will be negative as the simulations flow with time. The following relation is sufficient to express this requirement

$$\phi(x(t), t) = \phi(x(0), 0)$$

Although this is stricter than necessary, as only the sign of $\phi(.)$ is important. Thus the level set function changes over time, following the convective flow of the simulation trajectories. This equation can be differentiated

$$\frac{\partial}{\partial t}\phi(x,t) + \frac{d}{dt}x \cdot \frac{\partial}{\partial x}\phi(x,t) = 0$$

and gives the partial differential equation in the form

$$\frac{\partial}{\partial t}\phi(x,t) + f(x) \cdot \frac{\partial}{\partial x}\phi(x,t) = 0$$

Where the time derivative $\frac{\partial}{\partial t}\phi(x,t)$ is calculated by applying suitable numerical methods to the PDE in the region near to the set boundary, where $\phi(x,t)$ is small. The other term corresponds to the spatial movement of the level set function over the grid as seen from a point on the grid. This movement is given by the convective flow $\dot{x} = f(x)$ calculated at that point. A similar derivation may be found in section 1.2 of [19], and additional terms for the equation to handle other problem formulations are listed in section 1.1 of [14].

Reachability studies for safety take the set of states where it is unsafe to operate the DPF due to thermal limitations and work backwards in time to find the set of states which lead to this region.

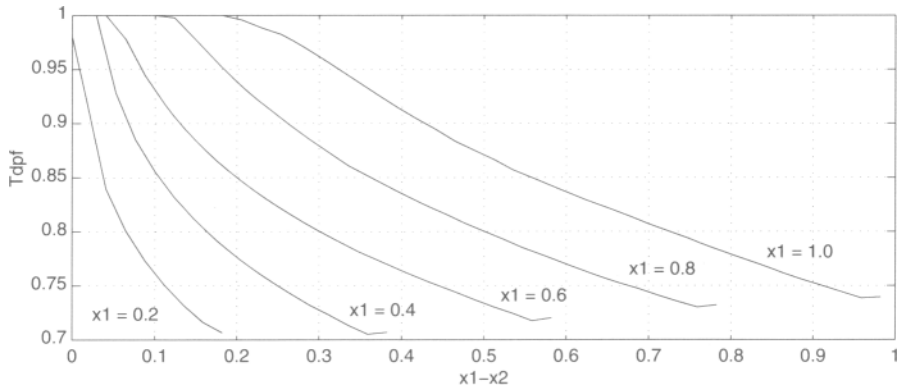


Fig. 7. Backwards reachability analysis for a feedgas flow rate of 25 kg/h, at 18% oxygen (unsafe region above the curves)

The backwards safety analysis for a given engine operating point begins with an initial avoid set consisting of all of the system states above a chosen threshold temperature, in this case 1000°C. Figures 7-8 show the boundary of the reach-avoid set for a range of values of x_1 for two different feedgas conditions. In each of the studied cases the growth of the set has converged within 60 seconds. The differences between these backwards safety analysis results for this model show that the supply of oxygen is the main factor affecting the development of the exotherm.

It is possible to slow down the soot reaction in the DPF by changing the engine operating point to reduce the concentration of available oxygen in the feedgas, however this will typically incur fuel and emissions penalties and hence should only be done when necessary. Figure 9 shows the experimental results of changing the air to fuel ratio to $\lambda = 2$ at idle to reduce the peak temperature of an exotherm during regeneration. Changing to an operating mode like this when the system state is approaching the reach-avoid set can prevent the system from entering the avoid region. Thus the region from which the model can complete a successful regeneration is increased by using the safe low oxygen concentration mode to reduce the exotherm and keep the DPF temperature away from the upper limit.

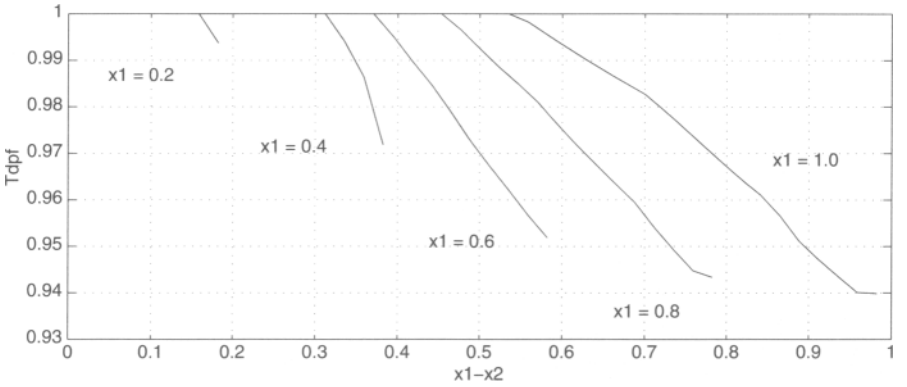


Fig. 8. Backwards reachability analysis for a feedgas flow rate of 18 kg/h, at 11% oxygen (unsafe region above the curves)

4 Discussion

To avoid generating high local temperatures if the engine should drop to idle during a regeneration, it is proposed that the idle conditions could include a post injection to reduce the oxygen concentration in the feedgas. Fuelling and EGR were chosen to give $\lambda = 2$ at idle with no load. Under these conditions the engine produces soot at a much greater rate, but it is only held at this rate until the filter has cooled below the light-off temperature. The filter can capture this soot and it does not significantly contribute to the total soot loading on the filter.

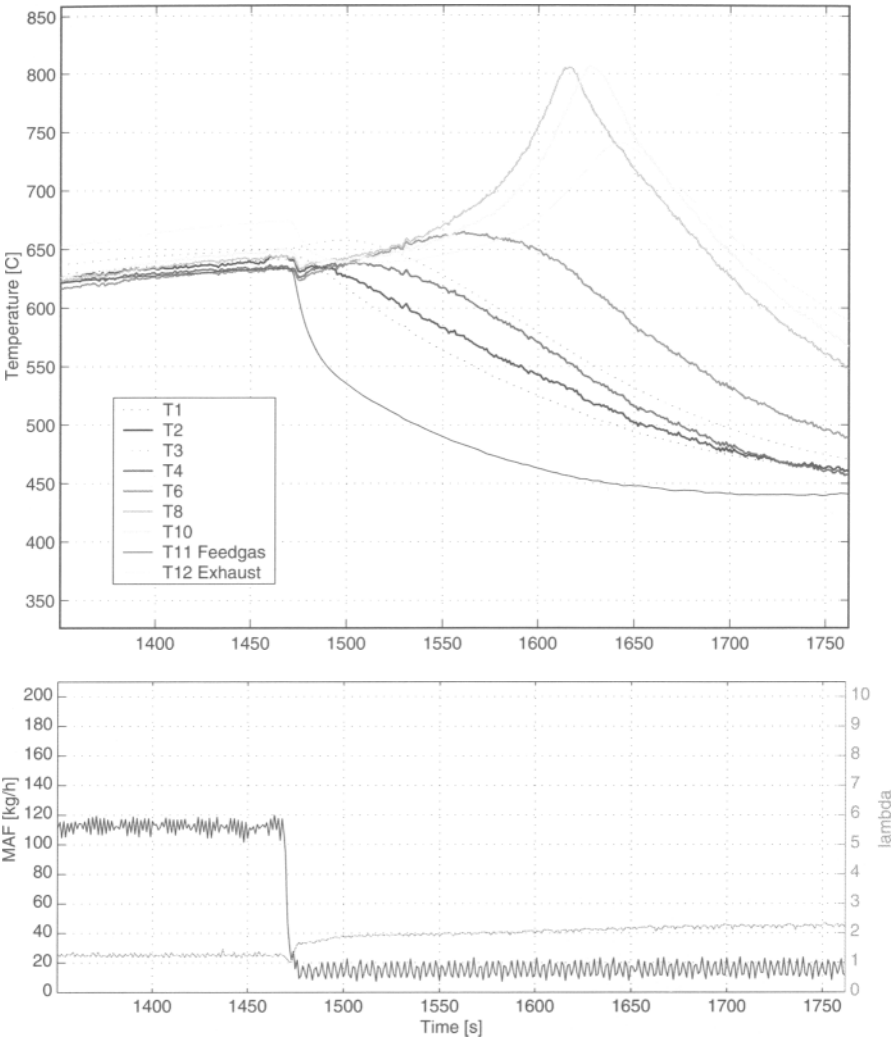


Fig. 9. Drop to $\lambda = 2$ idle regeneration experiment

Figure 9 shows a regeneration where the engine drops to idle at $\lambda = 2$ from initial conditions close to those which led to the very hot conditions before. However in this case the restricted oxygen concentration has resulted in a much lower exotherm, peaking at around 800°C.

The shape of the thermal wave can only be approximately estimated from fixed thermocouples but did show similarity across the experiments, with higher exotherms being shorter in length. Placing more thermocouples along the centre line of the filter could have given more information about the shape of the thermal wave, although the presence of the thermocouples in the exhaust channels

could affect the pattern of gas flow. The development of local hot regions in the filter away from the centre line is difficult to monitor, but was minimised by thoroughly regenerating the filter each time before it was loaded with soot.

As anticipated, the experiments where the engine dropped to idle conditions after initiating a regeneration gave a sharp exotherm. The resulting high temperature region travelled down the filter as a convective thermal wave, changing in magnitude depending upon the feedgas composition. The worst-case conditions observed were given by returning to a reasonable running speed with light fuelling after initiating a thermal wave by dropping to idle conditions. The increase in oxygen supply to the hot DPF caused a considerable increase in the rate of heat released by the exotherm, resulting in the highest peak temperatures and thermal gradients observed. Conversely, high levels of EGR and an additional late injection of fuel can be used to reduce the feedgas oxygen concentration enough to prevent an exotherm developing.

These features of the DPF regeneration have been represented reasonably well using the simple control volume model, although the initiation stage of the thermal wave is not so well captured. The highly non-linear response of the soot oxidation reaction to changes in temperature is a very important feature of the dynamic model, and the ability of level set methods to handle such models has enabled reachability analysis studies to be performed. These studies fit the experimental observations well and give a good indication of the range of conditions where continued regeneration of the filter will lead to excessively high temperatures.

Future work could take various directions, possibly focusing on development of the algorithms and their implementation or upon the modelling work and application areas for these methods. The accuracy of the level set methods used could be improved by extending the software routines to include point cloud methods [7], where the boundary location is additionally tracked by the flow of marker particles. Greater speed and efficiency could be attained by evaluating the level set methods over an adaptive triangulated mesh [3, 4, 22].

The control volume DPF model may be suitable for use with other control and analysis techniques. However it may be necessary to develop a good representation of the ignition of an exotherm in the filter model, which is currently not accurately modelled. Gas sampling from various locations down the intake and exhaust channels would be very useful for indicating how the oxygen and CO_2 concentrations vary down the length of the filter. The model currently assumes that these concentrations are constant down the channels, and that the gas flow through the wall is uniformly distributed. If the density of the soot cake on the channel walls is very low, then it might be possible for turbulence to mix more of the combustion products with flowing gas than would occur otherwise. Incorporation of the temperature dependence of the gas viscosity and pressure variations down the intake channel may also improve the realism of this model by giving a more accurate estimate of the gas flow rate though the channel wall over the length of the control volume.

Acknowledgements

This work has been supported in part by EPSRC, The MathWorks Inc. and with equipment supplied by Johnson Matthey Ltd.

References

1. Aldushin, A.P.: New results in the Theory of Filtration Combustion. Combustion and Flame 94, 308–320 (1993)
2. Aldushin, A.P., Rumanov, I.E., Matkowsky, B.J.: Maximal Energy Accumulation in a Superadiabtic Filtration Combustion Wave. Combustion and Flame 118, 76–90 (1999)
3. Barth, T.J., Sethian, J.A.: Numerical Schemes for the Hamilton-Jacobi and Level Set Equations on Triangulated Domains. Journal of Computational Phys. 145(1), 1–40 (1998)
4. Barth, T.J., Sethian, J.A.: Implementation of Hamilton-Jacobi and Level Set Equations on Triangulated Domains, von Karman Institute Lecture Series, Computational Fluid Mechanics (1998)
5. Bissett, E.J.: Mathematical Model of the Thermal Regeneration of a Wall-Flow Monolith Diesel Particulate Filter. Chemical Engineering Science 39(7/8), 1233–1244 (1984)
6. Bisset, E.J., Shadman, F.: Thermal regeneration of diesel-particulate monolithic filters. American Institute of Chemical Engineers Journal 39(5), 753–758 (1985)
7. Enright, D., Fedkiw, R., Ferziger, J., Mitchell, I.: A Hybrid Particle Level Set Method for Improved Interface Capturing. J. Comput. Phys. 183, 83–116 (2002)
8. Field, M.A., et al.: Combustion of Pulverized Coal, BCURA Leatherhead, 329. Cheroy and Sons Ltd. Banbury, England (1967)
9. Haralampous, O., Koltsakis, G.C.: Intra-layer temperature gradients during regeneration of diesel particulate filters. Chemical Engineering Science 57, 2345–2355 (2002)
10. Howitt, J., Montierth, M.: Cellular Ceramic Diesel Particulate Filter, SAE, Paper 810114 (1981)
11. Kladopoulou, E.A., Yang, S.L., Johnson, J.H., Parker, G.G., Konstandopoulos, A.G.: A Study Describing the Performance fo Diesel Particulate Filters During Loading and Regeneration - A Lumped Parameter Model of Control Applications, SAE, Paper 2003-01-0842 (2003)
12. Koltsakis, G.C., Stamatelos, A.M.: Modes of Catalytic Regeneration in Diesel Particulate Filters. Industrial and Engineering Chemistry Research 36, 4155–4165 (1997)
13. Mitchell, I.M.: Application of Level Set Methods to Control and Reachability Problems in Continuous and Hybrid Systems, Doctoral Thesis, Stanford University (August 2002)
14. Mitchell, I.M.: A Toolbox of Level Set Methods. Technical Report UBC CS TR-2004-09, Department of Computer Science, University of British Columbia (July 1, 2004)
15. Osher, S., Fedkiw, R.: Level Set Methods and Dynamic Implicit Surfaces. Springer, New York (2003)
16. Osher, S., Sethian, J.A.: Fronts Propagating with Curvature-Dependent Speed: Algorithms Based on Hamilton-Jacobi Formulation. Journal of Computational Physics 79, 12–49 (1988)

17. Raithby, G.: Laminar Heat Transfer in the Thermal Entrance Region of Circular Tubes and Two-Dimensional Rectangular Ducts with Wall Suction and Injection. Int. J. Heat Mass Transfer 14, 223–243 (1971)
18. Sethian, J.A.: Level Set Methods. Cambridge University Press, Cambridge (1996)
19. Sethian, J.A.: Level Set Methods and Fast Marching Methods. Cambridge University Press, Cambridge (1999)
20. Shadman, F., Bisset, E.J.: Thermal Regeneration of Diesel Particulate Filters: Development of a Mathematical Model. IEC Proc. Des. Dev., 22, 203 (1993)
21. Tan, J.C.: A Study of the Regeneration Process in Diesel Particulate Taps Using a Copper Fuel Additive, M.S. Thesis, Michigan Technological University, Houghton, MI (1995)
22. Zhang, Y.T., Shu, C.W.: High-order WENO schemes for Hamilton-Jacobi equations on triangular meshes. SIAM J. Sci. Comput., Society for Industrial and Applied Mathematics 24(3), 1005–1030 (2003)

Robust Control of Smart Material-based Actuators

Sina Valadkhan[1], Kirsten Morris[2], and Amir Khajepour[1]

[1] Department of Mechanical Engineering, University of Waterloo, Waterloo, Ontario, Canada
svaladkh@uwaterloo.ca, akhajepour@uwaterloo.ca
[2] Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, Canada
kmorris@uwaterloo.ca

Summary. Actuators incorporating smart materials are competitive choices for micro-positioning devices, but they are difficult to control because of their nonlinearity and hysteresis. These actuators are shown to be passive using several different approaches. Using passivity, a class of stabilizing controllers is identified. A controller for velocity control is designed that optimizes tracking performance and evaluated experimentally. A different approach is used to establish stability and tracking when position control is required. Robust stability is obtained in both approaches since stability is not sensitive to actuator parameters.

1 Introduction

Smart materials, such as magnetostrictives, piezo-ceramics and shape memory alloys, have many advantages for use in micro-positioning. For instance, magnetostrictive actuators display a very large force with a fast response time. Unfortunately, smart materials are highly nonlinear and hysteretic.

Since sub-micron accuracy is expected from micro-positioning actuators, a feedback control system is needed. The controller should stabilize the closed-loop system in all conditions and satisfy performance objectives. In order to stabilize a physical system, robust stability is needed because the system parameters are not exactly known.

A popular approach in control of smart materials is to linearize the system by incorporating the inverse of the actuator model before the actuator. If the model can be inverted, the system can be linearized. This leads to a composite system that is approximately the identity. A second controller is then designed using the identity as the plant model. In this approach, an accurate model for the system is needed. In [15], the Preisach model is coupled to an ordinary differential equation to model a magnetostrictive actuator. The model is inverted and used before the actuator to linearize the system. In [12], the actuator is linearized by an inverse model and $\mathbb{H}_2$ optimal control is used to provide robust stability for the linearized system. The resulting controllers are complex. Also, stability of the closed-loop system in the presence of modelling errors is not established.

V.D. Blondel et al. (Eds.) Recent Advances in Learning and Control, LNCIS 371, pp. 249–262, 2008.
springerlink.com

In this paper, we give a review of two existing approaches for robust control of smart materials: passivity and monotonicity. Controllers are designed and experimentally implemented on an actuator composed of Terfenol-D, the most commonly used magnetostrictive material.

Passivity-based controller design is frequently used in control of many systems, including structural mechanical systems - see, for instance, [1, 4]. The most popular model for smart materials is the Preisach model [3, 14]. In [6], it is shown that the Preisach model is passive. This result was used to develop a controller for a shape memory alloy actuator. In [16] it was shown, using physics, that magnetostrictive materials are passive. In this paper we develop a robustly stabilizing controller for magnetostrictive actuators using passivity. Since passivity is proven using a physical argument for magnetostrictive materials, robust stability of the closed loop is guaranteed regardless of the actuator model. A performance criterion is given and the controller parameters are chosen in order to optimize this criterion while satisfying the robust stability condition. The optimal controller is implemented and evaluated.

Another approach can also be used to obtain robust stability results. Most hysteretic materials are monotonic; that is, increasing inputs cause increasing outputs and the same for decreasing inputs. Monotonicity, along with some stronger assumptions and techniques from differential equations, was used in [10, 11] to establish results on the asymptotic behaviour of systems that include hysteresis. In [18], monotonicity and the counter-clockwise behaviour of a hysteresis loop are used to show that PI control of hysteretic systems leads to a stable closed loop and perfect tracking of step inputs. A controller is designed and then tested on a magnetostrictive actuator.

We first provide a brief overview of the most commonly used model for smart materials.

2 Preisach Model

The Preisach model is widely used to model smart materials because of its general structure, simple computation and application to many materials; see for example [17] for its application to magnetostrictive materials and [5, 8, 9] for other materials. This model is briefly explained here. For details, see [3, 14].

The basis of the Preisach model is the relay shown in Figure 1a. Each relay is denoted by two parameters: half width r and shift s.

The output of this relay is either $+1$ or -1. If the relay is in the $+1$ state and the input becomes less than $s-r$, the relay switches to -1, and if the relay is in the -1 state and the input becomes greater than $s+r$, the relay switches to the $+1$ state, otherwise, the output remains the same. The result is that the (r, s)-plane is separated into two regions by a boundary τ, depending on whether the relay is in the $+1$ or -1 state (Figure 1b). The relay output is only defined for a continuous input $u(t)$ and so the Preisach model is only valid for continuous inputs. An infinite number of relays with different s and r are used. Each relay output, multiplied by a locally integrable weight function $\mu(r, s)$, contributes

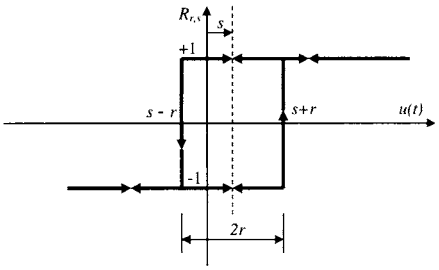


Fig. 1a. The Preisach relay

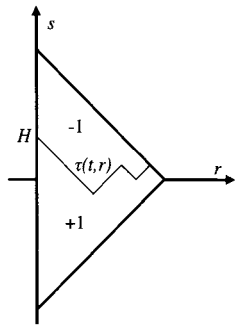


Fig. 1b. Preisach boundary

to form the output of the model. The weight function $\mu(r,s)$ is determined by experimental data for the hysteretic system. Let $R_{r,s}[u(\cdot)](t)$ be the output of the relay with half-width r and shift s. The output $y(t)$ of the model is

$$y(t) = \int_{-\infty}^{\infty} \int_{0}^{\infty} R_{r,s}[u(\cdot)](t)\mu(r,s)drds, \tag{1}$$

or in terms of the boundary τ,

$$y(t) = 2\int_{0}^{\infty} \int_{-\infty}^{\tau(t,r)} \mu(r,s)dsdr - \int_{-\infty}^{\infty} \int_{0}^{\infty} \mu(r,s)drds. \tag{2}$$

The Preisach model can be placed into the standard dynamical system framework with state the boundary τ [7].

3 Passivity of Smart Materials

The standard feedback configuration shown in Figure 2 is used. Consider a function $f : \mathbb{R}_+ \to \mathbb{R}^n$. The truncation f_T of f to the interval $[0,T]$ is

$$f_T(t) = \begin{cases} f(t), & 0 \le t \le T \\ 0, & T < t \end{cases}. \tag{3}$$

Other definitions are standard and can be found in, for instance [19].Passivity is defined as follows.

Definition 1. *[20] Consider a dynamical system with state variables x, an input u and output y. If there is a real-valued function $S(x)$ that is bounded from below satisfying the following relation for any $t_i \le t_f$:*

$$S(x(t_i)) + \int_{t_i}^{t_f} \langle u, y \rangle \, dt \ge S(x(t_f)) \tag{4}$$

the dynamical system is called passive.

The real valued function $S(x)$ is called the storage function. For physical systems the storage function is often the energy.

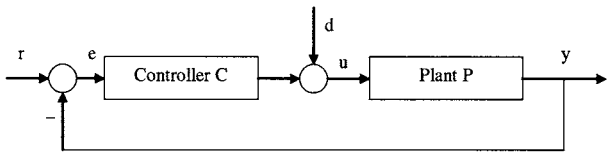


Fig. 2. The standard feedback configuration

If the plant P is passive and the controller C satisfies certain conditions, the passivity theorem can be used to establish stability of the controlled system. The following is one form of this theorem.

Theorem 1. *[19, sec. 6.6.2] Consider the feedback system shown in Figure 2 where C and P map U to U. The set U is a subset of L_{2e}. Assume that for any r and d in L_2, there are solutions e and u in L_{2e} and there are constants α_1, α_2 , α_3, β_1, β_2 and β_3 such that for every real T and $x \in L_{2e}$, the following conditions hold:*

$$\begin{aligned} &\text{I} && \|(Cx)_T\|_2 \leq \alpha_1 \|x_T\|_2 + \beta_1, \\ &\text{II} && \int_0^T \langle x, Cx \rangle \, dt \geq \alpha_2 \|x_T\|_2^2 + \beta_2, \\ &\text{III} && \int_0^T \langle Px, x \rangle \, dt \geq \alpha_3 \|(Px)_T\|_2^2 + \beta_3. \end{aligned} \tag{5}$$

If $\alpha_2 + \alpha_3 > 0$, then the closed loop with inputs r and d is L_2-stable.

A passive system will satisfy condition III with $\alpha_3 = 0$ and β determined by a lower bound on S. Thus, if a controller has finite gain (condition I) and is strictly passive (condition II with $\alpha_2 > 0$) it will stabilize the system. Condition II follows if the controller is linear with a strictly positive real transfer function [19, sec. 6.6.2]. Many smart materials exhibit saturation [6,13,17]; that is, the output does not change if the absolute value of the input is larger than some limit $u_{sat} > 0$. In this case, the weight function μ has compact support; that is, $\mu(r, s) = 0$ for all $r + s$ and $r - s$ greater than u_{sat}. In all physical situations, the value of the input is constrained by actuator limitations to $|u| \leq u_{sat}$. In this situation as well, we can assume the weight function to be zero for all $r + s$ and $r - s$ greater than some limit u_{sat}. For many hysteretic systems, the weight function $\mu(r, s)$ is also nonnegative [6, 13, 17]. These conditions imply passivity of smart materials described by the Preisach model. The storage function has an interpretation as the energy stored in the Preisach relays.

Theorem 2. *[6] If μ has compact support, is bounded, piecewise continuous, and non-negative then the model (2) with input u and output $\dot{y}$ is passive.*

The passivity theorem can be used to conclude stability of controllers for smart material systems described by the Preisach model. This approach was used successfully in velocity control of a shape-memory-alloy in [6].

3.1 Magnetostrictive Materials

Terfenol-D, an alloy of iron, terbium and dysprosium, is the most commonly used magnetostrictive material. For a Terfenol-D rod, if a magnetic field parallel to the axis of the rod is applied, it becomes slightly longer. A magnetization M is produced when a magnetic field is applied to the Terfenol-D sample. The magnetization is not only a function of instantaneous magnetic field H but also of the history of the magnetic field. In Figure 3a, magnetization versus magnetic field is shown for an experiment where the magnetic field H is a decaying sinusoidal signal and stress is 7.18 MPa.

In most applications, it is desired to control the displacement produced by the actuator, not the magnetization. In Figure 3b, the displacement produced by the actuator is plotted versus magnetization M at different stress levels. Unlike the hysteresis curves, the displacement is a function of instantaneous magnetization and stress. It does not depend on the history of the material directly. Except for a dead zone in the middle of the curve, an almost linear relationship is seen between elongation and magnetization. This is used to relate elongation to magnetization.

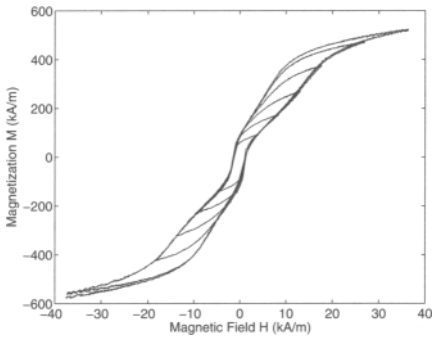


Fig. 3a. Magnetization vs magnetic field

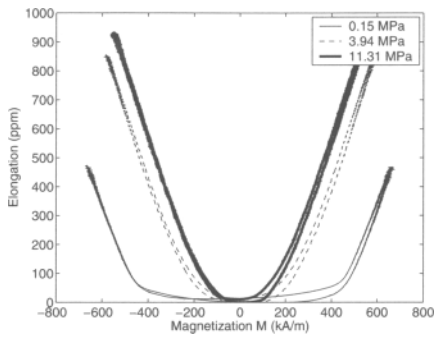


Fig. 3b. Elongation vs magnetization

The passivity of Preisach models can be used to deduce passivity of a magnetostrictive actuator with input u, the magnetic field H, and output y, the rate of magnetization change $\dot{M}$ or elongation $\dot{\varepsilon}$.

For a magnetostrictive system, magnetic field H and stress σ are often both varying as system inputs. The magnetic field is determined by the electrical current sent to the coil generating the magnetic field and the stress is determined by the load applied to the actuator. The magnetization M and strain (displacement) ε are the system outputs that are determined by the inputs. Define the input and output to be

$$u = \begin{pmatrix} \mu_0 H \\ \sigma \end{pmatrix} \tag{6}$$

$$y = \begin{pmatrix} \dot{M} \\ \dot{\varepsilon} \end{pmatrix} \tag{7}$$

where μ_0 is a constant, H is magnetic field, σ is stress, M is magnetization and ε is strain. It is assumed that the magnetostrictive system is under a one-dimensional stress-strain. This is the case for most existing magnetostrictive actuators.

In [16], a physical approach is used to show that magnetostrictive materials are passive, even when both applied field and stress are varying. The storage function is the Helmholtz free energy

$$\psi = E - TS \tag{8}$$

where E is internal energy, T is temperature and S is entropy. It is assumed that there is a good thermal connection between the magnetostrictive material and the surrounding environment so that the temperature of the material T is always close to the room temperature and hence, constant.

We distinguish between a varying stress or magnetic field, used as part of the control input u, and a constant stress or field, regarded as part of the system. In this latter situation, the storage function is modified, just as the total energy of a spring is modified to include the effect of a constant imposed potential field such as gravity. When the stress and magnetic field applied to the system include a constant component, the input is defined as $\bar{u} = \binom{\mu_0(H - H_{const})}{\sigma - \sigma_{const}}$. The output is unchanged. This situation occurs when a constant stress or magnetic field, perhaps due to noise, is present.

Theorem 3. *[16] Define*

$$\psi^F = \psi - \mu_0 \langle H_{const}, M \rangle - \sigma_{const}\varepsilon, \tag{9}$$

where ψ is the system Helmholtz free energy and M is the system magnetization. If temperature is constant, a magnetostrictive material is passive:

$$\psi_i^F + \int_{t_i}^{t_f} \langle \bar{u}, y \rangle \, dt \geq \psi_f^F \tag{10}$$

where subscripts i and f denote initial and final conditions, respectively.

In the next section, we discuss the implementation of passivity-based velocity control of magnetostrictive actuators.

4 Velocity Control of Magnetostrictive Actuators

In Figure 2, plant P represents a magnetostrictive actuator. The plant inputs and outputs are defined by equations (6) and (7), respectively. The plant outputs are the time-derivatives of magnetization and strain- this feedback system is used to control velocity. The reference signal r is equal to $\begin{pmatrix} \dot{M}_r \\ 0 \end{pmatrix}$, where $\dot{M}_r$ is the desired trajectory. Define $\dot{M}_e$ to be $\dot{M}_r - \dot{M}$. The error signal is

$$e = r - y = \begin{pmatrix} \dot{M}_e \\ -\dot{\varepsilon} \end{pmatrix}. \tag{11}$$

The signal $d = \begin{pmatrix} \mu_0 H_{noise} \\ \sigma_{in} \end{pmatrix}$ represents external disturbances. The signal H_{noise} is the controller noise and σ_{in} is the resultant stress from the external load applied to the actuator.

The controller uses $\dot{M}_e$ to determine the magnetic field applied to the magnetostrictive material. The magnetic field has two components: The noise H_{noise} and the magnetic field from the controller. If the relation between $\dot{M}_e$ and the magnetic field H from the controller is denoted by C_1, we have:

$$H = H_{noise} + C_1 \dot{M}_e \tag{12}$$

The stress applied to the magnetostrictive material σ is determined by the external load applied to the actuator and the internal friction of the actuator. If a viscous friction is assumed, the friction force would be $-\kappa\dot{\varepsilon}$, where κ is a positive constant. In this case, the stress σ is:

$$\sigma = \sigma_{in} - \kappa\dot{\varepsilon} \tag{13}$$

By combining equations (12) and (13), we have:

$$Ce = \begin{pmatrix} \mu_0 C_1 \dot{M}_e \\ -\kappa\dot{\varepsilon} \end{pmatrix}. \tag{14}$$

By combining with equation (11), the controller structure is obtained:

$$C = \begin{pmatrix} \mu_0 C_1 & 0 \\ 0 & \kappa \end{pmatrix}. \tag{15}$$

The following stability result now follows.

Theorem 4. *The controller (15) provides an L_2-stable closed loop for a magnetostrictive system if C_1 is a linear mapping with a proper transfer function, finite L_2-gain, and such that for some $\alpha_2 > 0$,*

$$\inf_{\omega \in \mathbb{R}} \operatorname{Re} \hat{C}_1(j\omega) = \alpha_2. \tag{16}$$

Since a physical approach was used to establish passivity of the plant, errors in a particular plant model will not affect stability. Theorem 4 is a stability result for $\dot{M}$ controlled by H. Using the linear relation shown in Figure 3b, M is proportional to ε. Velocity $\dot{\varepsilon}$ can be controlled by controlling $\dot{M}$.

An ideal PID controller with transfer function $\hat{C}_1(s) = \frac{K_I}{s} + K_P + K_D s$ does not satisfy condition I in equation (5) because it does not have a finite L_2 gain due the integral and derivative terms. Furthermore, approximate integration and derivative with transfer functions $\frac{1}{s+\alpha}$ and $\frac{s}{\delta s+1}$, respectively, are generally used in numerical simulations and experimental implementations, due to ease of implementation and also, in the case of approximate differentiation, reduced noise level. A "practical" PID controller with transfer function

$$\hat{C}_1(s) = \frac{K_I}{s+\alpha} + K_P + \frac{K_D s}{\delta s + 1}$$

satisfies condition I if α, $\delta > 0$. If K_I, K_P, $K_D > 0$, we have $\alpha_2 > 0$ in (16) and condition II is also satisfied. Thus, the closed-loop system will be stable with this controller. However, a PID controller cannot provide zero steady-state error unless the reference input is zero. To see this, assume that the steady-state error is zero. The controller output is a constant and a constant input is applied to the actuator. This means zero output velocity. This implies that the reference input is zero. Thus, this PID controller cannot provide perfect tracking of non-zero velocities. This is illustrated by the experimental results in Figure 4.

To provide better performance, the derivative term of the PID controller is replaced with an approximate double integration term. This gives the following controller:

$$\hat{C}_1(s) = \frac{K_I}{(s+\alpha)^2} + \frac{K_P}{s+\delta} + K_D \tag{17}$$

where K_P, K_I, and K_D are the controller gains and α and δ are positive constants.

We now choose the controller gains to optimize performance, subject to constraints on the parameters that ensure stability. The cost function

$$J = \int_{t_1}^{t_2} (y-r)^2 dt$$

where r is the desired velocity, y is the actual actuator velocity, and $[t_1, t_2]$ is the time range of interest, is used. The closed-loop is simulated by using a Preisach model for the actuator as explained in [17]. The cost function $J(K_P, K_I, K_D, \alpha, \delta)$ is minimized using the Nelder-Mead method [2] with a step input r. Provided that $\alpha > 0$, $\delta > 0$ and (16) are satisfied, Theorem 4 implies closed-loop stability. The optimum parameter values were found to be: $\alpha = 0.0106\frac{1}{s}$, $\delta = 0.0231\frac{1}{s}$, $K_I = 2.9881 \times 10^8 \frac{A}{m^2 s}$, $K_P = 1.4854 \times 10^{10} \frac{A}{m^2}$, and $K_D = 9.3098 \times 10^7 \frac{As}{m^2}$. This controller satisfies (16) with $\alpha_2 = 9.3098 \times 10^7 \frac{As}{m^2}$.

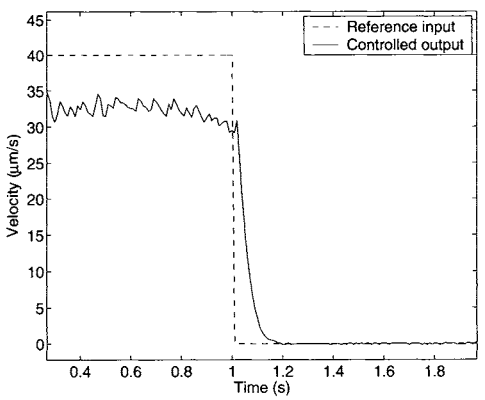


Fig. 4. PID control for a step reference signal.($K_P = 21.62\frac{As}{m^2}$, $K_I = 1.1344 \times 10^9 \frac{A}{m^2}$, $K_D = 2.6602 \times 10^{-8} \frac{As^2}{m^2}$, $\alpha = 1.249 \times 10^{-6} \frac{1}{s}$, and $\delta = 0.0229s$).

The test apparatus, described in detail in [17], consists of a 10cm long Terfenol-D rod, surrounded by a magnetic coil. The coil is connected to a current source controlled by a computer. The elongation is measured using an optical encoder with 10nm resolution. Other sensors monitor temperature. The controller was implemented using MATLAB® Real-Time Workshop.

Figure 5a displays the closed loop performance when the controller designed in this section is implemented and the reference signal is a step. The system settled within about 0.08s after the step, with a steady-state error of $0.09\frac{\mu m}{s}$ or 3% error. It was also observed that the current applied to the actuation unit was always less than the maximum, and the temperature did not vary. Figure 5b compares the reference signal and measured velocity for the closed-loop system when the reference signal is sinusoidal with varying amplitude. Good tracking is seen. The RMS tracking errors is $0.71\frac{\mu m}{s}$, a relative RMS error of about 3.5%.

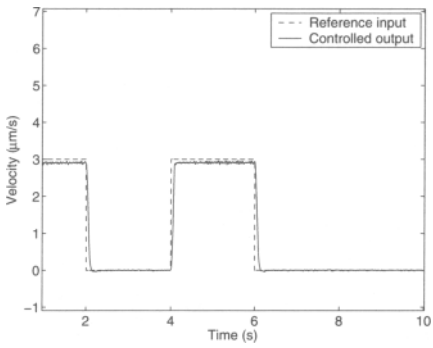


Fig. 5a. Velocity controller, step input (experimental data)

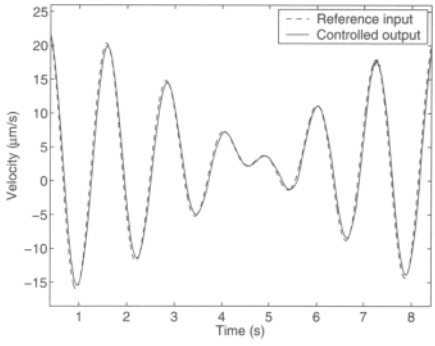


Fig. 5b. Velocity controller, sinusoidal signal (experimental data)

5 Monotonicity

We now describe a different approach to robust control of smart materials that uses the fact that these systems are generally monotonic: an increasing input implies an increasing output, and the same for decreasing inputs. This property can be used to derive useful stability results. Some definitions and assumptions are first needed. Define $\mathbb{R}_+$ to be the set of non-negative real numbers. For any interval $I \subset \mathbb{R}_+$, let $Map(I)$ indicate the set of real-valued functions defined on I. Define $\mathcal{C}(I)$ to be the set of continous functions on a closed interval I. The norm of a function in $\mathcal{C}(I)$ is

$$\|f\|_\infty = \sup_{t \in I} |f(t)|.$$

Definition 2. *[3] An operator* $\Gamma : Map(\mathbb{R}_+) \to Map(\mathbb{R}_+)$ *has the* Volterra property *if, for any* $v, w \in Map(\mathbb{R}_+)$ *and any non-negative* T, $v_T = w_T$ *implies that* $(\Gamma v)_T = (\Gamma w)_T$.

Definition 3. *[3] An operator $\Gamma : Map(\mathbb{R}_+) \to Map(\mathbb{R}_+)$ is* rate independent *if*

$$(\Gamma v) \circ \varphi = \Gamma(v \circ \varphi) \tag{18}$$

for all $v \in Map(\mathbb{R}_+)$ and all continuous monotone time transformations $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ satisfying $\varphi(0) = 0$ and $\lim_{t\to\infty} \varphi(t) = \infty$.

The Volterra property states that the hysteretic system output does not depend on future inputs; that is, determinism. A deterministic, rate independent operator is a hysteresis operator:

Definition 4. *[3] An operator $\Gamma : Map(\mathbb{R}_+) \to Map(\mathbb{R}_+)$ is a* hysteresis operator *if it is rate independent and has the Volterra property.*

Denote the hysteresis model input and output by u and y, respectively. The following assumptions are used throughout this paper. First, for any $\delta > 0$, $0 < s < t$, and any $w \in \mathcal{C}([0, s])$ define

$$B(w, t, \delta) = \{u \in \mathcal{C}([0, t]) |\ u(\tau) = w(\tau), 0 \le \tau \le s, \max_{s\le\tau\le t} |u(\tau) - w(s)| < \delta\}.$$

(A1) (continuity) If $u(t)$ is continuous then $y(t)$ is continuous. That is, $\Gamma : \mathcal{C}(I) \to \mathcal{C}(I)$.

(A2) (monotonicity) Consider an arbitrary interval $[t_i, t_f]$. If for every $t \in [t_i, t_f]$, $u(t_i) \ge u(t)$, then $y(t_i) \ge y(t_f)$. Also, if for every $t \in [t_i, t_f]$, $u(t_i) \le u(t)$, then $y(t_i) \le y(t_f)$.

(A3) (saturation) There exists some $u_{sat} > 0$, y_+ and y_- such that if $|u(t)| \ge u_{sat}$ then $(\Gamma u)(t) = y_+$ and $(\Gamma(-u))(t) = y_-$.

(A4) (Lipshitz property) There exists $\lambda > 0$ such that for each $s > 0$, $w \in \mathcal{C}([0, s])$, there is $\delta > 0$ and $T > s$ such that for all $u_1, u_2 \in B(w, T, \delta)$,

$$\sup_{0\le\tau\le T} |\Gamma(u_1)(\tau) - \Gamma(u_2)(\tau)| \le \lambda \sup_{0\le\tau\le T} |u_1(\tau) - u_2(\tau)|.$$

There is a close connection between assumption (A2) and monotonicity of the hysteretic system, in a sense that an increasing input results in increasing output and the same for decreasing inputs/outputs. By setting $t = t_f$, it is seen that if assumption (A2) holds, the hysteretic system is monotonic. The converse is not true [18].

The Preisach model (2) is a hysteresis operator [3]. The following theorems show that the Preisach model in general satisfies assumptions (A1)-(A4).

Theorem 5. *[3, Prop. 2.4.9,Prop. 2.4.11] If the Preisach weight function $\mu(r, s)$ satisfies*

$$\lambda := \int_0^\infty \sup_{s\in R} |\mu(r, s)| dr < \infty \tag{19}$$

then for initial Preisach boundaries in

$$\{\phi \in Map(\mathbb{R}_+) \mid |\phi(r_1) - \phi(r_2)| \le |r_1 - r_2| \text{ for all } r_1, r_2 \ge 0\}$$

the Preisach operator maps inputs in $C([s,T])$ *to outputs in* $C([s,T])$ *for any interval* $[s,T]$ *where* $0 \leq s < T$. *For any* $w \in C([0,s])$

$$\max_{s \leq t \leq T} |y_1(t) - y_2(t)| \leq \lambda \max_{s \leq t \leq T} |u_1(t) - u_2(t)| \tag{20}$$

for all $u_1, u_2 \in B(w,T)$ *where*

$$B(w,T) = \{u \in C([0,T]) |\ u(\tau) = w(\tau), 0 \leq \tau \leq s\}.$$

The inequality (20) means that a global Lipshitz inequality, stronger than (A4), is satisfied.

As discussed above, the weight function μ in the Preisach model (2) is generally non-negative with compact support.

Theorem 6. *[18] If* $\mu(r,s)$ *is bounded and non-negative with compact support then assumptions (A1)- (A4) are satisfied with* λ *given by (19) and*

$$y_+ = \int_{-\infty}^{\infty} \int_{0}^{\infty} \mu(r,s) dr ds, \tag{21}$$

$$y_- = \int_{-\infty}^{\infty} \int_{0}^{\infty} -\mu(r,s) dr ds. \tag{22}$$

6 Position Control

The monotonicity of hysteretic systems, such as those described by a Preisach model, can be used to show stability and tracking of PI control. The results in this section are given in detail in [18]. The following PI controller is used here for position control:

$$\hat{C}(s) = \frac{K_I}{s} + K_P \tag{23}$$

where K_I and K_P are constants. The closed-loop system shown in Figure 2 with external input $d = 0$ is described by

$$e(t) = r(t) - y(t), \tag{24}$$

$$f(t) = \int_0^t e(\tau) d\tau, \tag{25}$$

$$u(t) = K_P e(t) + K_I f(t), \tag{26}$$

$$y(t) = \Gamma[u(\cdot)](t). \tag{27}$$

The controller parameters are assumed only to satisfy the following general assumptions.

(B1) For the controller in (23), $0 \leq K_P \lambda < 1$ and $K_I > 0$ where $\lambda > 0$ is the Lipchitz constant in assumption (A4).

(B2) The reference signal $r(t)$ is a continuous function of time.

Theorem 7. *Assume that assumptions (A1), (A4), (B1) and (B2) hold. Then (24)-(27) have a unique solution for $u \in \mathcal{C}[0,\infty)$ and $y \in \mathcal{C}[0,\infty)$. If in addition, assumption (A2) holds, $u(0) = 0$ and $|y(0)| \leq \|r\|_\infty$, then $\|y\|_\infty \leq \|r\|_\infty$.*

Theorem 7 implies not only stability, but also that an overshoot cannot occur. For hysteretic systems satisfying the saturation assumption (A3), boundedness of the output can be shown [18]. Theorem 7 extends this result to hysteretic systems that do not satisfy the saturation assumption. More importantly, regardless of saturation, the closed-loop system has a gain of 1 with zero bias.

From the following theorems we can conclude that PI controllers provide a closed loop system that tracks a constant input with zero steady-state error and no overshoot. A bound on the time required to achieve a specified error is obtained.

Theorem 8. *Assume that r is a constant in some interval $[t_0,\infty)$ and assumptions (A1), (A2), (A4) , (B1) and (B2) hold. If for some nonnegative ρ, $|r - y(t_0)| \leq \rho$, then $|r - y(t_1)| \leq \rho$ for all $t_1 \geq t_0$.*

Theorem 8 states that in a constant-input period, the absolute value of the error is never increased. As a result, an oscillatory response or an overshoot cannot be seen. The following theorem proves that under certain conditions, the error can be made arbitrarily small.

Theorem 9. *Let t_0 be a non-negative real number. Assume that $r(t)$ is a constant, r, in $[t_0,\infty)$ and that assumptions (A1)-(A4), and (B1)-(B2) hold. If $y_- \leq r \leq y_+$, then for every $\varepsilon > 0$,*

$$|r - y(t)| \leq \varepsilon, \forall t \geq \bar{t} + t_0, \tag{28}$$

where

$$\bar{t} = \frac{\frac{u_{sat}}{K_I} + |f(t_0)|}{\varepsilon}. \tag{29}$$

Consequently, $\lim_{t\to\infty} y(t) = r$.

Theorem 9 gives an upper limit for the time required to achieve any accuracy ε. Also, it states that if the input to the closed loop is reasonable, zero steady-state error is guaranteed. The condition $y_- \leq r \leq y_+$ is just that the desired reference point is is consistent with the saturation points. Theorem 7 can be used to design a controller for position control. The controller must only satisfy assumptions $0 \leq \lambda K_P < 1$ and $K_I > 0$.

Position control of a magnetostrictive actuator with PI control was implemented experimentally. The same apparatus used in implementation of a velocity controller discussed above, described in [17], was used. An encoder with $10nm$ accuracy measured position.

Theorem 7 states that any PI controller with positive gains provides stability. The gains are chosen to minimize the cost function

$$J = \int_{t_1}^{t_2} (y - r)^2 dt$$

where r is the reference input, y is the closed-loop position and $[t_1, t_2]$ is the time range of interest, subject to the parameter constraints (B1). The magnetostrictive material is simulated by using a Preisach model with a weight function identified in [17]. Using the Preisach model, y is computed as a function of controller parameters. The cost function J is numerically minimized using the Nelder-Mead method [2] with a reference signal r chosen as a series of step inputs. (The ideal version of such a reference input is not continuous. However, in the experiment there is a rapid but continuous change between values and so the signal is continuous.) The optimum values for the controller gains are $K_I = 38.02\ \mathrm{s}^{-1}$ and $K_P = 0.0785$.

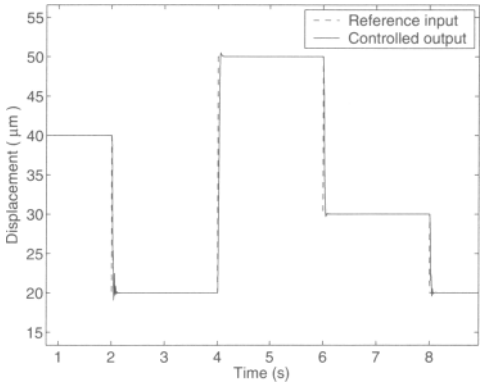


Fig. 6. The closed-loop response to various steps

In Figure 6, the closed-loop response of the system to step input is shown for the optimized controller. Excellent tracking is seen. As predicted by Theorem 9, there is no steady-state error. The system settles to ± 10 nm of the reference signal in less than $0.2s$. This is within the accuracy of the sensor. A small overshoot and some oscillations are also observed. These are likely caused by some unmodeled mass in the system.

7 Conclusions

In this paper, the closed loop stability of magnetorestrictive materials was studied and position and velocity controllers were designed and experimentally evaluated. Passivity and monotonicity of magnetorestrictive materials were established and used to arrive at a class of stabilizing controllers. The controllers are robust to any errors in material parameters. Using passivity, it was shown that an approximate (practical) PID controller with a double integral would provide small steady state error and good tracking performance in velocity control of magnetorestrictive actuators. Using the monotonicity of the material, a PI controller was shown to provide stability and also a satisfactory performance in the position control of magnetorestrictive actuators.

References

1. Akella, P., Wen, J.T.: Synthesized passive feedback control of sensor-rich smart structures - experimental results. In: Proceedings, IEEE conference on control applications, pp. 1098–1103 (1995)
2. Bertsekas, D.P.: Nonlinear Programming. Athena Scientific, Belmont, Massachusetts (1999)
3. Brokate, M., Sprekels, J.: Hysteresis and phase transitions. Springer, New York (1996)
4. Damaren, C.J., Oguamanam, D.C.: Vibration control of spacecraft box structures using a collocated piezo-actuator/sensor. Journal of Intelligent Material Systems and Structures 15, 369–374 (2004)
5. Gorbet, R.B., Wang, D.W.L., Morris, K.A.: Preisach Model Identification of a Two-Wire SMA Actuator. In: Proceedings of the 1998 IEEE International Conference on Robotics and Automation, vol. 3, pp. 2161–2167 (1998)
6. Gorbet, R.B., Morris, K.A., Wang, D.W.L.: Passivity-based stability and control of hysteresis in smart actuators. IEEE Transactions on control systems technology 9, 5–16 (2001)
7. Gorbet, R.B., Morris, K.A., Wang, D.W.L.: Control of Hysteretic Systems: A State-space Approach. In: Yamamoto, Y., Hara, S. (eds.) Learning, Control and Hybrid Systems, pp. 432–451. Springer, Heidelberg (1999)
8. Hughes, D., Wen, J.T.: Preisach Modelling of Piezocramic and Shape Memory Alloy Hysteresis. In: Proceedings of the 1995 IEEE Control Conference on Applications (1995)
9. Iyer, R.V., Tan, X., Krishnaprasad, P.S.: Approximate Inversion of the Preisach Hysteresis Operator with Application to Control of Smart Actuators. IEEE Transactions on Automatic Control 50, 798–810 (2005)
10. Logemann, H., Ryan, E.P., Shvartsman, I.: A class of differential-delay systems with hysteresis: asymptotic behaviour of solutions. Nonlinear analysis (to appear, 2007)
11. Logemann, H., Ryan, E.P., Shvartsman, I.: Integral control of infinite-dimensional systems in the presence of hysteresis: an input-output approach. ESAIM - control, optimisation and calculus of variations 13, 458–483 (2007)
12. Nealis, J.M., Smith, R.C.: Robust control of a magnetostrictive actuator. In: Proceedings of SPIE, The International Society for Optical Engineering, vol. 5049, pp. 221–232 (2003)
13. Smith, R.C., Dapino, M.J., Seelecke, S.: Free energy model for hysteresis in magnetostrictive transducers. J. Appl. Phys. 93, 458–466 (2003)
14. Smith, R.C.: Smart Material Systems: Model Development. SIAM, Philadelphia (2005)
15. Tan, X., Baras, J.S.: Modeling and control of hysteresis in magnetostrictive actuators. Automatica 40, 1469–1480 (2004)
16. Valadkhan, S., Morris, K.A., Khajepour, A.: Passivity of magnetostrictive materials. SIAM Journal of applied mathematics 67, 667–686 (2007)
17. Valadkhan, S., Morris, K.A., Khajepour, A.: A review and comparison of hysteresis models for magnetostrictive materials. Journal of intelligent material systems and structures (submitted, 2007)
18. Valadkhan, S., Morris, K.A., Khajepour, A.: Stability and robust position control of hysteretic systems. IEEE Trans. Auto. Control (submitted, 2007)
19. Vidyasagar, M.: Nonlinear systems analysis. Prentice hall, Englewood Cliffs, New Jersey (1993)
20. Willems, J.C.: Dissipative dynamical systems, part I: General theory. Archives for Rational Mechanics and Analysis 45, 321–351 (1972)

Behaviors Described by Rational Symbols and the Parametrization of the Stabilizing Controllers

Jan C. Willems[1] and Yutaka Yamamoto[2]

[1] K.U. Leuven, B-3001 Leuven, Belgium
Jan.Willems@esat.kuleuven.be
www.esat.kuleuven.be/~jwillems
[2] Kyoto University, Kyoto 606-8501, Japan
yy@i.kyoto-u.ac.jp
www-ics.acs.i.kyoto-u.ac.jp/~yy

Summary. We present a behavioral theory of linear systems described by differential equations involving matrices of rational functions. Representations of controllable and stabilizable systems that are left coprime over the ring of proper, stable, or proper stable rational functions are discussed. These representations lead to effective parametrizations of the set of stabilizing controllers for a plant.

Keywords: Behaviors, rational symbols, controllability, stabilizability, observability, regular controllers, superregular controllers, parametrization of stabilizing controllers.

1 Introduction

It is a pleasure to contribute an article to this Festschrift in honor of Mathukumalli Vidyasagar on the occasion of his 60-th birthday. As the subject of our article, we have chosen the parametrization of stabilizing controllers for linear systems. This topic goes back to the pioneering contributions of Kučera [2] and Youla-Bongiorgio-Jabr [7], and is commonly known as the Kučera-Youla parametrization of the set of stabilizing controllers. This parametrization issue and the algebraic structure that underpins its solution are main topics discussed in Vidyasagar's book [4], one of the few books in the field of Systems & Control that can truly be termed 'Algebraic System Theory'. This book served as the inspiration for the present paper.

Our approach is somewhat different from the usual one in that we do not view a linear system as defined by a transfer function. Rather, we view a system in the behavioral sense, that is, as a family of trajectories. All relevant system properties, such as controllability, stabilizability, observability, and detectability, are defined in terms of the behavior. Control is viewed as restricting the plant behavior by intersecting it with the controller behavior.

V.D. Blondel et al. (Eds.) Recent Advances in Learning and Control, LNCIS 371, pp. 263–277, 2008.
springerlink.com

The behavior of a linear time-invariant differential system is defined as the set of solutions of a system of linear constant-coefficient differential equations. However, these behaviors can be represented in many other ways, for example, as the set of solutions of a system of equations involving a differential operator in a matrix of rational functions, rather than in a matrix of polynomials. The problem of parametrizing the set of stabilizing controllers leads to the question of determining all controller behaviors which, when intersected with the given plant behavior, yield a stable system. The representation of behaviors in terms of rational symbols turns out to be an effective representation that leads to a parametrization of the set of stabilizing controllers.

In the classical approach [2, 4, 7], systems with the same transfer function are identified. By taking a trajectory-based definition of a system, the behavioral point of view is able to carefully keep track of all trajectories, also of the non-controllable ones. Loosely speaking, the stable coprime factorizations of the transfer-function based approach manage to avoid unstable pole-zero cancellations. Our approach avoids introducing, as well as cancelling, common poles and zeros. Since the whole issue of coprime factorizations over the ring of proper stable rational functions started from a need to deal carefully with pole-zero cancellations, we feel that our trajectory-based mode of thinking offers a useful point of view.

A few words about the notation and nomenclature used. We use standard symbols for the sets $\mathbb{R}, \mathbb{N}, \mathbb{Z}$, and $\mathbb{C}$. $\overline{\mathbb{C}}_+ := \{s \in \mathbb{C} \mid \mathrm{Re}\,(s) \geq 0\}$ denotes the closed right-half of the complex plane. We use $\mathbb{R}^{\mathtt{n}}$, $\mathbb{R}^{\mathtt{n}\times\mathtt{m}}$, etc. for vectors and matrices. When the number of rows or columns is immaterial (but finite), we use the notation $^\bullet$, $^{\bullet\times\bullet}$, etc. Of course, when we then add, multiply, or equate vectors or matrices, we assume that the dimensions are compatible. $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathtt{n}})$ denotes the set of infinitely differentiable functions from $\mathbb{R}$ to $\mathbb{R}^{\mathtt{n}}$. The symbol I denotes the identity matrix, and 0 the zero matrix. When we want to emphasize the dimension, we write $I_{\mathtt{n}}$ and $0_{\mathtt{n}_1\times\mathtt{n}_2}$. A matrix is said to be of *full row rank* if its rank is equal to the number of rows. Full column rank is defined analogously.

$\mathbb{R}[\xi]$ denotes the set of polynomials with real coefficients in the indeterminate ξ, and $\mathbb{R}(\xi)$ denotes the set of real rational functions in the indeterminate ξ. $\mathbb{R}[\xi]$ is a ring and $\mathbb{R}[\xi]^{\mathtt{n}}$ a finitely generated $\mathbb{R}[\xi]$-module. $\mathbb{R}(\xi)$ is a field and $\mathbb{R}(\xi)^{\mathtt{n}}$ is an $\mathtt{n}$-dimensional $\mathbb{R}(\xi)$-vector space. The polynomials $p_1, p_2 \in \mathbb{R}[\xi]$ are said to be *coprime* if they have no common zeros. $p \in \mathbb{R}[\xi]$ is said to be *Hurwitz* if it has no zeros in $\overline{\mathbb{C}}_+$. The *relative degree* of $f \in \mathbb{R}(\xi), f = n/d$, with $n, d \in \mathbb{R}[\xi]$, is the degree of the denominator d minus the degree of the numerator n; $f \in \mathbb{R}(\xi)$ is said to be *proper* if the relative degree is ≥ 0, *strictly proper* if the relative degree is > 0, and *biproper* if the relative degree is equal to 0. The rational function $f \in \mathbb{R}(\xi)$, $f = n/d$, with $n, d \in \mathbb{R}[\xi]$ coprime, is said to be *stable* if d is Hurwitz, and *miniphase* if n and d are both Hurwitz.

We only discuss the main ideas. Details and proofs may be found in [6]. The results can easily be adapted to other stability domains, but in this article, we only consider the Hurwitz domain for concreteness.

2 Rational Symbols

We consider behaviors $\mathcal{B} \subseteq (\mathbb{R}^\bullet)^{\mathbb{R}}$ that are the set of solutions of a system of linear-constant coefficient differential equations. In other words, $\mathcal{B}$ is the solution set of

$$R\left(\tfrac{d}{dt}\right) w = 0, \tag{$\mathcal{R}$}$$

where $R \in \mathbb{R}[\xi]^{\bullet\times\bullet}$. We shall deal with infinitely differentiable solutions only. Hence $(\mathcal{R})$ defines the dynamical system $\Sigma = (\mathbb{R}, \mathbb{R}^\bullet, \mathcal{B})$ with

$$\mathcal{B} = \left\{ w \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^\bullet) \;\middle|\; R\left(\tfrac{d}{dt}\right) w = 0 \right\}.$$

We call this system (or its behavior) a *linear time-invariant differential system.* Note that we may as well denote this behavior as $\mathcal{B} = \mathtt{kernel}\left(R\left(\frac{d}{dt}\right)\right)$, since $\mathcal{B}$ is actually the kernel of the differential operator

$$R\left(\tfrac{d}{dt}\right) : \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathtt{column\ dimension}(R)}) \to \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathtt{row\ dimension}(R)}).$$

We denote the set of linear time-invariant differential systems or their behaviors by $\mathcal{L}^\bullet$ and by $\mathcal{L}^{\mathtt{w}}$ when the number of variables is $\mathtt{w}$.

We will extend the above definition of a behavior defined by a differential equation involving a polynomial matrix to a 'differential equation' involving a matrix of rational functions. In order to do so, we first recall the terminology of factoring a matrix of rational functions in terms of polynomial matrices. The pair (P, Q) is said to be a *left factorization over* $\mathbb{R}[\xi]$ of $M \in \mathbb{R}(\xi)^{\mathtt{n}_1\times\mathtt{n}_2}$ if (i) $P \in \mathbb{R}[\xi]^{\mathtt{n}_1\times\mathtt{n}_1}$ and $Q \in \mathbb{R}[\xi]^{\mathtt{n}_1\times\mathtt{n}_2}$, (ii) $\mathtt{determinant}(P) \neq 0$, and (iii) $M = P^{-1}Q$. (P, Q) is said to be a *left-coprime factorization over* $\mathbb{R}[\xi]$ of M if, in addition, (iv) P and Q are left coprime over $\mathbb{R}[\xi]$. Recall that P and Q are said to be *left coprime over* $\mathbb{R}[\xi]$ if for every factorization $\begin{bmatrix} P & Q \end{bmatrix} = F \begin{bmatrix} P' & Q' \end{bmatrix}$ with $F \in \mathbb{R}[\xi]^{\mathtt{n}_1\times\mathtt{n}_1}$, F is $\mathbb{R}[\xi]$-unimodular. It is easy to see that a left-coprime factorization over $\mathbb{R}[\xi]$ of $M \in \mathbb{R}(\xi)^{\mathtt{n}_1\times\mathtt{n}_2}$ is unique up to premultiplication of P and Q by an $\mathbb{R}[\xi]$-unimodular polynomial matrix $U \in \mathbb{R}[\xi]^{\mathtt{n}_1\times\mathtt{n}_1}$.

Consider the system of 'differential equations'

$$G\left(\tfrac{d}{dt}\right) w = 0, \tag{$\mathcal{G}$}$$

with $G \in \mathbb{R}(\xi)^{\bullet\times\bullet}$, called the *symbol* of $(\mathcal{G})$. Since G is a matrix of rational functions, it is not clear when $w : \mathbb{R} \to \mathbb{R}^\bullet$ is a solution of $(\mathcal{G})$. This is not a matter of smoothness, but a matter of giving a meaning to the equality, since $G\left(\frac{d}{dt}\right)$ is not a differential operator, and not even a map.

We define solutions as follows. Let (P, Q) be a left-coprime matrix factorization over $\mathbb{R}[\xi]$ of $G = P^{-1}Q$. Define

$$[\![w : \mathbb{R} \to \mathbb{R}^\bullet \text{ is a solution of } (\mathcal{G})]\!] :\Leftrightarrow [\![Q\left(\tfrac{d}{dt}\right) w = 0]\!].$$

Hence $(\mathcal{G})$ defines the system

$$\Sigma = \left(\mathbb{R}, \mathbb{R}^\bullet, \mathtt{kernel}\left(Q\left(\tfrac{d}{dt}\right)\right)\right) \in \mathcal{L}^\bullet.$$

It follows from this definition that $G\left(\frac{d}{dt}\right)$ is *not* a map on $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^\bullet)$. Rather, $w \mapsto G\left(\frac{d}{dt}\right) w$ is the point-to-set map that associates with $w \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^\bullet)$ the set $v' + v$, with $v' \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^\bullet)$ a particular solution of $P\left(\frac{d}{dt}\right) v' = Q\left(\frac{d}{dt}\right) w$ and $v \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^\bullet)$ any function that satisfies $P\left(\frac{d}{dt}\right) v = 0$. This set of v's is a finite-dimensional linear subspace of $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^\bullet)$ of dimension equal to the degree of $\mathtt{determinant}(P)$. Hence, if P is not an $\mathbb{R}[\xi]$-unimodular polynomial matrix, equivalently, if G is not a polynomial matrix, $G\left(\frac{d}{dt}\right)$ is not a point-to-point map. Viewing $G\left(\frac{d}{dt}\right)$ as a point-to set map leads to the definition of its kernel as

$$\mathtt{kernel}\left(G\left(\tfrac{d}{dt}\right)\right) := \{w \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^\bullet) \mid 0 \in G\left(\tfrac{d}{dt}\right) w\},$$

i.e. $\mathtt{kernel}\left(G\left(\frac{d}{dt}\right)\right)$ consists of the set of solutions of $(\mathcal{G})$, and of its image as

$$\mathtt{image}\left(G\left(\tfrac{d}{dt}\right)\right) := \{v \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^\bullet) \mid v \in G\left(\tfrac{d}{dt}\right) w \text{ for some } w \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^\bullet)\}.$$

Hence $(\mathcal{G})$ defines the system

$$\Sigma = \left(\mathbb{R}, \mathbb{R}^\bullet, \mathtt{kernel}\left(G\left(\tfrac{d}{dt}\right)\right)\right) := \left(\mathbb{R}, \mathbb{R}^\bullet, \mathtt{kernel}\left(Q\left(\tfrac{d}{dt}\right)\right)\right) \in \mathcal{L}^\bullet.$$

Three main theorems in the theory of linear time-invariant differential systems are (i) the elimination theorem, (ii) the one-to-one relation between annihilators and submodules or subspaces, and (iii) the equivalence of controllability and existence of an image representation. Results involving (ii) and (iii) are discussed in later sections.

The *elimination theorem* states that if $\mathcal{B} \in \mathcal{L}^{\mathtt{w}_1 + \mathtt{w}_2}$, then

$$\mathcal{B}_1 := \{w_1 \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathtt{w}_1}) \mid \exists\, w_2 \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathtt{w}_2}) \text{ such that } (w_1, w_2) \in \mathcal{B}\}$$

belongs to $\mathcal{L}^{\mathtt{w}_1}$. In other words, $\mathcal{L}^\bullet$ is closed under projection. The elimination theorem implies that $\mathcal{L}^\bullet$ is closed under addition, intersection, projection, and under action and inverse action with $F\left(\frac{d}{dt}\right)$, where $F \in \mathbb{R}(\xi)^{\bullet \times \bullet}$.

3 Input, Output, and State Cardinality

The integer invariants $\mathsf{w}, \mathsf{m}, \mathsf{p}, \mathsf{n}$ are maps from $\mathcal{L}^\bullet$ to $\mathbb{Z}_+$ that play an important role in the theory of linear time-invariant differential systems. Intuitively,

$\mathsf{w}(\mathcal{B})$ equals the number of variables in $\mathcal{B}$,
$\mathsf{m}(\mathcal{B})$ equals the number of input variables in $\mathcal{B}$,
$\mathsf{p}(\mathcal{B})$ equals the number of output variables in $\mathcal{B}$, and
$\mathsf{n}(\mathcal{B})$ equals the number of state variables in $\mathcal{B}$.

The integer invariant w is defined by $[\![\mathsf{w}(\mathcal{B}) := \mathtt{w}]\!] \iff [\![\mathcal{B} \in \mathcal{L}^{\mathtt{w}}]\!]$.

The other integer invariants are most easily captured by means of representations. A behavior $\mathcal{B} \in \mathcal{L}^\bullet$ admits an *input/output representation*

$$P\left(\tfrac{d}{dt}\right) y = Q\left(\tfrac{d}{dt}\right) u, \quad w = \Pi \begin{bmatrix} u \\ y \end{bmatrix} \tag{i/o}$$

with $P \in \mathbb{R}(\xi)^{\mathsf{p}(\mathcal{B})\times\mathsf{p}(\mathcal{B})}$, $\mathtt{determinant}(P) \neq 0$, $Q \in \mathbb{R}(\xi)^{\mathsf{p}(\mathcal{B})\times\mathsf{m}(\mathcal{B})}$, and $\Pi \in \mathbb{R}^{\mathsf{w}(\mathcal{B})\times\mathsf{w}(\mathcal{B})}$ a permutation matrix. This input/output representation of $\mathcal{B}$ defines $\mathsf{m}(\mathcal{B})$ and $\mathsf{p}(\mathcal{B})$ uniquely. It follows from the conditions on P and Q that u is free, that is, that for any $u \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathsf{m}(\mathcal{B})})$, there exists a $y \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^{\mathsf{p}(\mathcal{B})})$ such that $P\left(\frac{d}{dt}\right) y = Q\left(\frac{d}{dt}\right) u$. The permutation matrix Π shows how the input and output components are chosen from the components of w, and results in an input/output partition of w.

The matrix $G = P^{-1}Q \in \mathbb{R}(\xi)^{\mathsf{p}(\mathcal{B})\times\mathsf{m}(\mathcal{B})}$ is called the *transfer function* corresponding to this input/output partition. In fact, it is possible to choose this partition such that G is proper. It is worth mentioning that in general $P\left(\frac{d}{dt}\right) y = Q\left(\frac{d}{dt}\right) u$ has a different behavior than $y = P^{-1}Q\left(\frac{d}{dt}\right) u$. The difference is due to the fact that $\mathcal{B}$ may not be controllable, as discussed in the next section.

A behavior $\mathcal{B} \in \mathcal{L}^\bullet$ also admits an observable *input/state/output representation*

$$\frac{d}{dt}x = Ax + Bu,\ y = Cx + Du,\quad w = \Pi \begin{bmatrix} u \\ y \end{bmatrix}, \tag{i/s/o}$$

with $A \in \mathbb{R}^{\mathsf{n}(\mathcal{B})\times\mathsf{n}(\mathcal{B})}$, $B \in \mathbb{R}^{\mathsf{n}(\mathcal{B})\times\mathsf{m}(\mathcal{B})}$, $C \in \mathbb{R}^{\mathsf{p}(\mathcal{B})\times\mathsf{n}(\mathcal{B})}$, $D \in \mathbb{R}^{\mathsf{p}(\mathcal{B})\times\mathsf{m}(\mathcal{B})}$, $\Pi \in \mathbb{R}^{\mathsf{w}(\mathcal{B})\times\mathsf{w}(\mathcal{B})}$ a permutation matrix, and (A, C) an observable pair. By eliminating x, the (u, y)-behavior defines an linear time-invariant differential system, with behavior denoted by $\mathcal{B}'$. This behavior is related to $\mathcal{B}$ by $\mathcal{B} = \Pi\mathcal{B}'$. It can be shown that this input/state/output representation of $\mathcal{B}$, including the observability of (A, C), defines $\mathsf{m}(\mathcal{B})$, $\mathsf{p}(\mathcal{B})$, and $\mathsf{n}(\mathcal{B})$ uniquely.

4 Controllability, Stabilizability, Observability, and Detectability

The behavior $\mathcal{B} \in \mathcal{L}^\bullet$ is said to be *controllable* if for all $w_1, w_2 \in \mathcal{B}$, there exists $T \geq 0$ and $w \in \mathcal{B}$, such that $w(t) = w_1(t)$ for $t < 0$, and $w(t) = w_2(t - T)$ for $t \geq T$. $\mathcal{B}$ is said to be *stabilizable* if for all $w \in \mathcal{B}$, there exists $w' \in \mathcal{B}$, such that $w'(t) = w(t)$ for $t < 0$ and $w'(t) \to 0$ as $t \to \infty$.

In words, controllability means that it is possible to switch between any two trajectories in the behavior, and stabilizability means that every trajectory can be steered to zero asymptotically.

Until now, we have dealt with representations involving the variables w only. However, many models, such as first principles models obtained by interconnection and state models, include auxiliary variables in addition to the variables the model aims at. We call the latter *manifest variables*, and the auxiliary variables *latent variables*. In the context of rational models, this leads to the model class

$$R\left(\frac{d}{dt}\right) w = M\left(\frac{d}{dt}\right) \ell \tag{LV}$$

with $R, M \in \mathbb{R}(\xi)^{\bullet\times\bullet}$. By the elimination theorem, the *manifest behavior* of (LV), defined as

$$\{w \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^\bullet) \mid \exists\, \ell \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^\bullet) \text{ such that (LV) holds}\},$$

belongs to $\mathcal{L}^\bullet$.

The latent variable system (LV) is said to be *observable* if, whenever (w, ℓ_1) and (w, ℓ_2) satisfy (LV), then $\ell_1 = \ell_2$. (LV) is said to be *detectable* if, whenever (w, ℓ_1) and (w, ℓ_2) satisfy (LV), then $\ell_1(t) - \ell_2(t) \to 0$ as $t \to \infty$.

In words, observability means that the latent variable trajectory can be deduced from the manifest variable trajectory, and detectability means that the latent variable trajectory can be deduced from the manifest variable trajectory asymptotically. The notions of observability and detectability apply to more general situations, but here we use them only in the context of latent variable systems.

It is easy to derive tests to verify these properties in terms of kernel representations and the zeros of the associated symbol. We first recall the notion of poles and zeros of a matrix of rational functions.

$M \in \mathbb{R}(\xi)^{\mathtt{n}_1 \times \mathtt{n}_2}$ can be brought into a simple canonical form, called the *Smith-McMillan form* by pre- and postmultiplication by $\mathbb{R}[\xi]$-unimodular polynomial matrices. Let $M \in \mathbb{R}(\xi)^{\mathtt{n}_1 \times \mathtt{n}_2}$. There exist $U \in \mathbb{R}[\xi]^{\mathtt{n}_1 \times \mathtt{n}_1}$, $V \in \mathbb{R}[\xi]^{\mathtt{n}_2 \times \mathtt{n}_2}$, both $\mathbb{R}[\xi]$-unimodular, $\Pi \in \mathbb{R}[\xi]^{\mathtt{n}_1 \times \mathtt{n}_1}$, and $Z \in \mathbb{R}[\xi]^{\mathtt{n}_1 \times \mathtt{n}_2}$ such that $M = U\Pi^{-1}ZV$, with

$$\Pi = \mathtt{diagonal}(\pi_1, \pi_2, \cdots, \pi_{\mathtt{n}_1}), Z = \begin{bmatrix} \mathtt{diagonal}(\zeta_1, \zeta_2, \cdots, \zeta_{\mathtt{r}}) & 0_{\mathtt{r} \times (\mathtt{n}_2 - \mathtt{r})} \\ 0_{(\mathtt{n}_1 - \mathtt{r}) \times \mathtt{r}} & 0_{(\mathtt{n}_1 - \mathtt{r}) \times (\mathtt{n}_2 - \mathtt{r})} \end{bmatrix}$$

with $\zeta_1, \zeta_2, \cdots, \zeta_{\mathtt{r}}, \pi_1, \pi_2, \cdots, \pi_{\mathtt{n}_1}$ non-zero monic elements of $\mathbb{R}[\xi]$, the pairs $\zeta_{\mathtt{k}}, \pi_{\mathtt{k}}$ coprime for $\mathtt{k} = 1, 2, \ldots, \mathtt{r}$, $\pi_{\mathtt{k}} = 1$ for $\mathtt{k} = \mathtt{r}+1, \mathtt{r}+2 \ldots, \mathtt{n}_1$, and with $\zeta_{\mathtt{k}-1}$ a factor of $\zeta_{\mathtt{k}}$ and $\pi_{\mathtt{k}}$ a factor of $\pi_{\mathtt{k}-1}$, for $\mathtt{k} = 2, \cdots, \mathtt{r}$. Of course, $\mathtt{r} = \mathtt{rank}(M)$. The roots of the $\pi_{\mathtt{k}}$'s (hence of π_1, disregarding multiplicity issues) are called the *poles* of M, and those of the $\zeta_{\mathtt{k}}$'s (hence of $\zeta_{\mathtt{r}}$, disregarding multiplicity issues) are called the *zeros* of M. When $M \in \mathbb{R}[\xi]^{\bullet \times \bullet}$, the $\pi_{\mathtt{k}}$'s are absent (they are equal to 1). We then speak of the *Smith form.*

Proposition 1

1. $(\mathcal{G})$ is controllable if and only if G has no zeros.
2. $(\mathcal{G})$ is stabilizable if and only if G has no zeros in $\overline{\mathbb{C}}_+$.
3. (LV) is observable if and only if M has full column rank and has no zeros.
4. (LV) is detectable if and only if M has full column rank and has no zeros in $\overline{\mathbb{C}}_+$. ∎

Consider the following special case of (LV)

$$w = M\left(\tfrac{d}{dt}\right)\ell \qquad (\mathcal{M})$$

with $M \in \mathbb{R}(\xi)^{\bullet \times \bullet}$. Note that, with $M\left(\frac{d}{dt}\right)$ viewed as a point-to-set map, the manifest behavior of $(\mathcal{M})$ is equal to $\mathtt{image}\left(M\left(\frac{d}{dt}\right)\right)$. $(\mathcal{M})$ is hence called an *image representation* of its manifest behavior. In the observable case, that is, if M is of full column rank and has no zeros, M has a polynomial left inverse, and hence $(\mathcal{M})$ defines a differential operator mapping w to ℓ. In other words, in the observable case, there exists an $F \in \mathbb{R}[\xi]^{\bullet \times \bullet}$ such that $(\mathcal{M})$ has the representation

$$w = M\left(\tfrac{d}{dt}\right)\ell, \quad \ell = F\left(\tfrac{d}{dt}\right)w.$$

The well-known relation between controllability and image representations for polynomial symbols remains valid in the rational case.

Theorem 2. The following are equivalent for $\mathcal{B} \in \mathcal{L}^\bullet$.

1. $\mathcal{B}$ is controllable.
2. $\mathcal{B}$ admits an image representation $(\mathcal{M})$ with $M \in \mathbb{R}(\xi)^{\bullet\times\bullet}$.
3. $\mathcal{B}$ admits an observable image representation $(\mathcal{M})$ with $M \in \mathbb{R}(\xi)^{\bullet\times\bullet}$. ∎

Let $\mathcal{B} \in \mathcal{L}^\bullet$. The *controllable part* of $\mathcal{B}$ is defined as

$$\mathcal{B}_{\text{controllable}} := \{w \in \mathcal{B} \mid \forall\, t_0, t_1 \in \mathbb{R},\ t_0 \le t_1,\ \exists\, w' \in \mathcal{B} \text{ with compact support such that } w(t) = w'(t) \text{ for } t_0 \le t \le t_1\}.$$

In words, $\mathcal{B}_{\text{controllable}}$ consists of the trajectories in $\mathcal{B}$ that can be steered to zero in finite time. It is easy to see that $\mathcal{B}_{\text{controllable}} \in \mathcal{L}^\bullet$ and that it is controllable. In fact, $\mathcal{B}_{\text{controllable}}$ is the largest controllable behavior contained in $\mathcal{B}$.

The controllable part induces an equivalence relation on $\mathcal{L}^\bullet$, called *controllability equivalence*, by setting

$$[\![\mathcal{B}' \sim_{\text{controllability}} \mathcal{B}'']\!] :\Leftrightarrow [\![\mathcal{B}'_{\text{controllable}} = \mathcal{B}''_{\text{controllable}}]\!].$$

It is easy to prove that $\mathcal{B}' \sim_{\text{controllable}} \mathcal{B}''$ if and only if $\mathcal{B}'$ and $\mathcal{B}''$ have the same compact support trajectories, or, for that matter, the same square integrable trajectories. Each equivalence class modulo controllability contains exactly one controllable behavior. This controllable behavior is contained in all the other behaviors that belong to the equivalence class modulo controllability.

The system $G\left(\frac{d}{dt}\right) w = 0$, where $G \in \mathbb{R}(\xi)^{\bullet\times\bullet}$, and $F\left(\frac{d}{dt}\right) G\left(\frac{d}{dt}\right) w = 0$ are controllability equivalent if $F \in \mathbb{R}(\xi)^{\bullet\times\bullet}$ is square and nonsingular. In particular, two input/output systems (i/o) have the same transfer function if and only if they are controllability equivalent.

If $G_1, G_2 \in \mathbb{R}(\xi)^{\bullet\times\bullet}$ have full row rank, then the behavior defined by $G_1\left(\frac{d}{dt}\right) w = 0$ is equal to the behavior defined by $G_2\left(\frac{d}{dt}\right) w = 0$ if there exists a $\mathbb{R}[\xi]$-unimodular matrix $U \in \mathbb{R}[\xi]^{\bullet\times\bullet}$ such that $G_2 = UG_1$. On the other hand, the behavior defined by $G_1\left(\frac{d}{dt}\right) w = 0$ has the same controllable part as the behavior defined by $G_2\left(\frac{d}{dt}\right) w = 0$ if and only if there exists an $F \in \mathbb{R}(\xi)^{\bullet\times\bullet}$, square and nonsingular, such that $G_2 = FG_1$. If G_1 and G_2 are full row rank polynomial matrices, then equality of the behaviors holds if and only if $G_2 = UG_1$. This illustrates the subtle distinction between equations that have the same behavior, versus behaviors that are controllability equivalent.

5 Rational Annihilators

Obviously, for $n \in \mathbb{R}(\xi)^\bullet$ and $w \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^\bullet)$, the statements $n\left(\frac{d}{dt}\right)^\top w = 0$, and, hence, for $\mathcal{B} \in \mathcal{L}^\bullet$, $n\left(\frac{d}{dt}\right)^\top \mathcal{B} = 0$, meaning $n\left(\frac{d}{dt}\right)^\top w = 0$ for all $w \in \mathcal{B}$, are well-defined, since we have given a meaning to $(\mathcal{G})$.

Call $n \in \mathbb{R}[\xi]^{\bullet}$ a *polynomial annihilator* of $\mathcal{B} \in \mathcal{L}^{\bullet}$ if $n\left(\frac{d}{dt}\right)^{\top} \mathcal{B} = 0$, and call $n \in \mathbb{R}(\xi)^{\bullet}$ a *rational annihilator* of $\mathcal{B} \in \mathcal{L}^{\bullet}$ if $n\left(\frac{d}{dt}\right)^{\top} \mathcal{B} = 0$.

Denote the set of polynomial and of rational annihilators of $\mathcal{B} \in \mathcal{L}^{\bullet}$ by $\mathcal{B}^{\perp_{\mathbb{R}[\xi]}}$ and $\mathcal{B}^{\perp_{\mathbb{R}(\xi)}}$, respectively. It is well known that for $\mathcal{B} \in \mathcal{L}^{\mathbb{w}}$, $\mathcal{B}^{\perp_{\mathbb{R}[\xi]}}$ is an $\mathbb{R}[\xi]$-module, indeed, a finitely generated one, since all $\mathbb{R}[\xi]$-submodules of $\mathbb{R}[\xi]^{\mathbb{w}}$ are finitely generated. However, $\mathcal{B}^{\perp_{\mathbb{R}(\xi)}}$ is also an $\mathbb{R}[\xi]$-module, but a submodule of $\mathbb{R}(\xi)^{\mathbb{w}}$ viewed as an $\mathbb{R}[\xi]$-module (rather than as an $\mathbb{R}(\xi)$-vector space). The $\mathbb{R}[\xi]$-submodules of $\mathbb{R}(\xi)^{\mathbb{w}}$ are not necessarily finitely generated.

The question occurs when $\mathcal{B}^{\perp_{\mathbb{R}(\xi)}}$ is a vector space. This question has a nice answer, given in the following theorem.

Theorem 3. *Let $\mathcal{B} \in \mathcal{L}^{\mathbb{w}}$.*

1. *$\mathcal{B}^{\perp_{\mathbb{R}(\xi)}}$ is an $\mathbb{R}[\xi]$-submodule of $\mathbb{R}(\xi)^{\mathbb{w}}$.*
2. *$\mathcal{B}^{\perp_{\mathbb{R}(\xi)}}$ is an $\mathbb{R}(\xi)$-vector subspace of $\mathbb{R}(\xi)^{\mathbb{w}}$ if and only if $\mathcal{B}$ is controllable.*
3. *Denote the $\mathbb{R}[\xi]$-submodules of $\mathbb{R}[\xi]^{\mathbb{w}}$ by $\mathfrak{M}^{\mathbb{w}}$. There is a bijective correspondence between $\mathcal{L}^{\mathbb{w}}$ and $\mathfrak{M}_{\mathbb{w}}$, given by*

$$\mathcal{B} \in \mathcal{L}^{\mathbb{w}} \mapsto \mathcal{B}^{\perp_{\mathbb{R}[\xi]}} \in \mathfrak{M}^{\mathbb{w}},$$

$$\mathbb{M} \in \mathfrak{M}^{\mathbb{w}} \mapsto \{w \in \mathcal{C}^{\infty}(\mathbb{R}, \mathbb{R}^{\mathbb{w}}) \mid n\left(\tfrac{d}{dt}\right)^{\top} w = 0 \ \forall n \in \mathbb{M}\}.$$

4. *Denote the linear $\mathbb{R}(\xi)$-subspaces of $\mathbb{R}(\xi)^{\mathbb{w}}$ by $\mathfrak{L}^{\mathbb{w}}$. There is a bijective correspondence between $\mathcal{L}^{\mathbb{w}}_{\text{controllable}}$, the controllable elements of $\mathcal{L}^{\mathbb{w}}$, and $\mathfrak{L}_{\mathbb{w}}$ given by*

$$\mathcal{B} \in \mathcal{L}^{\mathbb{w}}_{\text{controllable}} \mapsto \mathcal{B}^{\perp_{\mathbb{R}(\xi)}} \in \mathfrak{L}^{\mathbb{w}},$$

$$\mathbb{L} \in \mathfrak{L}^{\mathbb{w}} \mapsto \{w \in \mathcal{C}^{\infty}(\mathbb{R}, \mathbb{R}^{\mathbb{w}}) \mid n\left(\tfrac{d}{dt}\right)^{\top} w = 0 \ \forall n \in \mathbb{L}\}. \qquad \blacksquare$$

This theorem shows a precise sense in which a linear time-invariant system can be identified by a module, and a controllable linear time-invariant differential system (an infinite dimensional subspace of $\mathcal{C}^{\infty}(\mathbb{R}, \mathbb{R}^{\mathbb{w}})$ whenever $\mathcal{B} \neq \{0\}$) can be identified with a *finite-dimensional* vector space (of dimension $\mathsf{p}(\mathcal{B})$). Indeed, through the polynomial annihilators, $\mathcal{L}^{\mathbb{w}}$ is in one-to-one correspondence with the $\mathbb{R}[\xi]$-submodules of $\mathbb{R}[\xi]^{\mathbb{w}}$, and, through the rational annihilators, $\mathcal{L}^{\mathbb{w}}_{\text{controllable}}$ is in one-to-one correspondence with the $\mathbb{R}(\xi)$-subspaces of $\mathbb{R}(\xi)^{\mathbb{w}}$.

Consider the system $\mathcal{B} \in \mathcal{L}^{\mathbb{w}}$ and its rational annihilators $\mathcal{B}^{\perp_{\mathbb{R}(\xi)}}$. In general, this is an $\mathbb{R}[\xi]$-submodule, but not $\mathbb{R}(\xi)$-vector subspace of $\mathbb{R}(\xi)^{\mathbb{w}}$. Its polynomial elements, $\mathcal{B}^{\perp_{\mathbb{R}[\xi]}}$ always form an $\mathbb{R}[\xi]$-submodule over $\mathbb{R}[\xi]^{\mathbb{w}}$, and this module determines $\mathcal{B}$ uniquely. Therefore, $\mathcal{B}^{\perp_{\mathbb{R}(\xi)}}$ also determines $\mathcal{B}$ uniquely. Moreover, $\mathcal{B}^{\perp_{\mathbb{R}(\xi)}}$ forms an $\mathbb{R}(\xi)$-vector space if and only if $\mathcal{B}$ is controllable. More generally, the $\mathbb{R}(\xi)$-span of $\mathcal{B}^{\perp_{\mathbb{R}(\xi)}}$ is exactly $\mathcal{B}^{\perp_{\mathbb{R}(\xi)}}_{\text{controllable}}$. Therefore the $\mathbb{R}(\xi)$-span of the rational annihilators of two systems are the same if and only if they have the same controllable part. We state this formally.

Theorem 4. *Let $\mathcal{B}_1$ be given by $G_1\left(\frac{d}{dt}\right) w = 0$ and $\mathcal{B}_2$ by $G_2\left(\frac{d}{dt}\right) w = 0$, with $G_1, G_2 \in \mathbb{R}(\xi)^{\bullet \times \mathbb{w}}$. The rows of G_1 and G_2 span the same $\mathbb{R}[\xi]$-submodule of*

$\mathbb{R}(\xi)^{\mathbf{w}}$ *if and only if* $\mathcal{B}_1 = \mathcal{B}_2$. *The rows of* G_1 *and* G_2 *span the same* $\mathbb{R}(\xi)$*-vector subspace of* $\mathbb{R}(\xi)^{\mathbf{w}}$ *if and only if* $\mathcal{B}_1$ *and* $\mathcal{B}_2$ *have the same controllable part, that is, if and only if* $\mathcal{B}_1 \sim_{\text{controllable}} \mathcal{B}_2$. ∎

6 Left-prime Representations

In order to express system properties and to parametrize the set of stabilizing controllers effectively, we need to consider representations with matrices of rational functions over certain special rings. We now introduce the relevant subrings of $\mathbb{R}(\xi)$.

1. $\mathbb{R}(\xi)$ itself, the rational functions,
2. $\mathbb{R}[\xi]$, the polynomials,
3. $\mathbb{R}(\xi)_{\mathcal{P}}$, the set elements of $\mathbb{R}(\xi)$ that are proper,
4. $\mathbb{R}(\xi)_{\mathcal{S}}$, the set elements of $\mathbb{R}(\xi)$ that are stable,
5. $\mathbb{R}(\xi)_{\mathcal{PS}} = \mathbb{R}(\xi)_{\mathcal{P}} \cap \mathbb{R}(\xi)_{\mathcal{S}}$, the proper stable rational functions.

We can think of these subrings in terms of poles. Indeed, these subrings are characterized by, respectively, arbitrary poles, no finite poles, no poles at $\{\infty\}$, no poles in $\overline{\mathbb{C}}_+$, and no poles in $\overline{\mathbb{C}}_+ \cup \{\infty\}$. It is easy to identify the unimodular elements (that is, the elements that have an inverse in the ring) of these rings. They consist of, respectively, the non-zero elements, the non-zero constants, the biproper elements, the miniphase elements, and the biproper miniphase elements of $\mathbb{R}(\xi)$.

We also consider matrices over these rings. Call an element of $\mathbb{R}(\xi)^{\bullet\times\bullet}$ *proper*, *stable*, or *proper stable* if each of its entries is. The square matrices over these rings are unimodular if and only if the determinant is unimodular. For $M \in \mathbb{R}(\xi)_{\mathcal{P}}^{\bullet\times\bullet}$, define $M^\infty := \lim_{x\in\mathbb{R}, x\to\infty} M(x)$. Call the matrix $M \in \mathbb{R}(\xi)_{\mathcal{P}}^{\mathrm{n}\times\mathrm{n}}$ *biproper* if it has an inverse in $\mathbb{R}(\xi)_{\mathcal{P}}^{\mathrm{n}\times\mathrm{n}}$, that is, if $\mathtt{determinant}(M^\infty) \neq 0$, and call $M \in \mathbb{R}(\xi)_{\mathcal{S}}^{\mathrm{n}\times\mathrm{n}}$ *miniphase* if it has an inverse in $\mathbb{R}(\xi)_{\mathcal{S}}^{\mathrm{n}\times\mathrm{n}}$, that is, if $\mathtt{determinant}(M^\infty) \neq 0$ is miniphase.

Let $\mathcal{R}$ denote any of the rings $\mathbb{R}(\xi)$, $\mathbb{R}[\xi]$, $\mathbb{R}(\xi)_{\mathcal{P}}$, $\mathbb{R}(\xi)_{\mathcal{S}}$, $\mathbb{R}(\xi)_{\mathcal{PS}}$. $M \in \mathcal{R}^{\mathrm{n}_1\times\mathrm{n}_2}$ is said to be *left prime* over $\mathcal{R}$ if for every factorization of M the form $M = FM'$ with $F \in \mathcal{R}^{\mathrm{n}_1\times\mathrm{n}_1}$ and $M' \in \mathcal{R}^{\mathrm{n}_1\times\mathrm{n}_2}$, F is unimodular over $\mathcal{R}$. It is easy to characterize the left-prime elements. $M \in \mathbb{R}(\xi)^{\mathrm{n}_1\times\mathrm{n}_2}$ is the prime over $\mathcal{R}$ if and only if

1. M is of full row rank when $\mathcal{R} = \mathbb{R}(\xi)$,
2. $M \in \mathbb{R}[\xi]^{\mathrm{n}_1\times\mathrm{n}_2}$ and $M(\lambda)$ is of full row rank for all $\lambda \in \mathbb{C}$ when $\mathcal{R} = \mathbb{R}[\xi]$,
3. $M \in \mathbb{R}(\xi)_{\mathcal{P}}^{\mathrm{n}_1\times\mathrm{n}_2}$ and M^∞ is of full row rank when $\mathcal{R} = \mathbb{R}(\xi)_{\mathcal{P}}$,
4. M is of full row rank and has no poles and no zeros in $\overline{\mathbb{C}}_+$ when $\mathcal{R} = \mathbb{R}(\xi)_{\mathcal{S}}$,
5. $M \in \mathbb{R}(\xi)_{\mathcal{P}}^{\mathrm{n}_1\times\mathrm{n}_2}$, M^∞ is of full row rank, and M has no poles and no zeros in $\overline{\mathbb{C}}_+$, when $\mathcal{R} = \mathbb{R}(\xi)_{\mathcal{PS}}$.

Controllability and stabilizability can be linked to the existence of left-prime representations over these subrings of $\mathbb{R}(\xi)$.

1. $\mathcal{B} \in \mathcal{L}^{\bullet}$ admits a representation $(\mathcal{R})$ with R of full row rank, and a representation $(\mathcal{G})$ with G of full row rank and $G \in \mathbb{R}(\xi)^{\bullet \times \bullet}_{\mathcal{PS}}$, that is, with all its elements proper and stable, meaning that they have no poles in $\overline{\mathbb{C}}_+$.
2. $\mathcal{B}$ admits a representation $(\mathcal{G})$ with G left prime over $\mathbb{R}(\xi)$, that is, with G of full row rank.
3. $\mathcal{B}$ is controllable if and only if it admits a representation $(\mathcal{G})$ with G left prime over $\mathbb{R}(\xi)$, that is, G has full row rank and has no zeros.
4. $\mathcal{B}$ is controllable if and only if it admits a representation $(\mathcal{R})$ with $R \in \mathbb{R}[\xi]^{\bullet \times \bullet}$ left prime over $\mathbb{R}[\xi]$, that is, with $R(\lambda)$ of full row rank for all $\lambda \in \mathbb{C}$.
5. $\mathcal{B}$ is controllable if and only if it admits a representation $(\mathcal{G})$ that is left prime over $\mathbb{R}(\xi)_{\mathcal{P}}$, that is, all elements of G are proper and G^{∞} of full row rank, and G has no zeros.
6. $\mathcal{B}$ admits a representation $(\mathcal{G})$ with G left prime over $\mathbb{R}(\xi)_{\mathcal{P}}$, that is, all elements of G are proper and G^{∞} has full row rank.
7. $\mathcal{B}$ is stabilizable if and only if it admits a representation $(\mathcal{G})$ with $G \in \mathbb{R}(\xi)^{\bullet \times \bullet}_{\mathcal{S}}$ left prime over $\mathbb{R}(\xi)_{\mathcal{S}}$, that is, G has full row rank and no poles and no zeros in $\overline{\mathbb{C}}_+$.
8. $\mathcal{B}$ is stabilizable if and only if it admits a representation $(\mathcal{G})$ with $G \in \mathbb{R}(\xi)^{\bullet \times \bullet}_{\mathcal{PS}}$ left prime over $\mathbb{R}(\xi)_{\mathcal{PS}}$, that is, G^{∞} has full row rank and G has no poles and no zeros in $\overline{\mathbb{C}}_+$.

These results illustrate how system properties can be translated into properties of rational symbols. Roughly speaking, every $\mathcal{B} \in \mathcal{L}^{\bullet}$ has a full row rank polynomial and a full row rank proper and/or stable representation. As long as we allow a non-empty region where to put the poles, we can obtain a representation with a rational symbol with poles confined to that region. The zeros of the representation are more significant. No zeros correspond to controllability. No unstable zeros correspond to stabilizability. In [6] an elementary proof is given that does not involve complicated algebraic arguments of the characterization of stabilizability in terms of a representation that is left-prime over the ring of proper stable rational functions. Analogous results can also be obtained for image representations.

Note that a left-prime representation over $\mathbb{R}(\xi)_{\mathcal{PS}}$ exists if and only if the behavior is stabilizable. This result can be compared with the classical result obtained by Vidyasagar in his book [4], where the aim is to obtain a proper stable left-prime representation of a system that is given as a transfer function, $y = F\left(\frac{d}{dt}\right) u$, where $F \in \mathbb{R}(\xi)^{\mathrm{p} \times \mathrm{m}}$. This system is a special case of $(\mathcal{G})$ with $G = \begin{bmatrix} I_{\mathrm{p}} & -F \end{bmatrix}$, and, since it has no zeros, $y = F\left(\frac{d}{dt}\right) u$ is controllable, and hence stabilizable. Therefore, a system defined by a transfer function admits a representation $G_1\left(\frac{d}{dt}\right) y = G_2\left(\frac{d}{dt}\right) u$ with $G_1, G_2 \in \mathbb{R}(\xi)^{\bullet \times \bullet}_{\mathcal{PS}}$, and $\begin{bmatrix} G_1 & G_2 \end{bmatrix}$ left coprime over $\mathbb{R}(\xi)_{\mathcal{PS}}$. This is an important, classical, result. However, in the controllable case, we can obtain a representation that is left prime over $\mathbb{R}(\xi)_{\mathcal{P}}$, and such that $\begin{bmatrix} G_1 & G_2 \end{bmatrix}$ has no zeros at all. The main difference of our result from the classical left-coprime factorization results over $\mathbb{R}(\xi)_{\mathcal{PS}}$ is that we faithfully preserve the exact behavior and not only the controllable part of a behavior,

whereas in the classical approach all stabilizable systems with the same transfer function are identified. We thus observe that the behavioral viewpoint provides a more intrinsic approach for discussing pole-zero cancellation. Indeed, since the transfer function is a rational function, poles and zeros can — by definition — be added and cancelled *ad libitum*. Transfer functions do not provide the correct framework in which to discuss pole-zero cancellations. Behaviors defined by rational functions do.

7 Control

We refer to [5, 1] for an extensive treatment of control in a behavioral setting. In terms of the notions introduced in these references, we shall be concerned with full interconnection only, meaning that the controller has access to all the system variables. We refer to [1] for a nice discussion of the concepts involved.

In the behavioral approach, control is viewed as the interconnection of a plant and a controller. Let $\mathcal{P}$ (henceforth $\in \mathcal{L}^{\mathbb{w}}$) be called the *plant*, $\mathcal{C}$ (henceforth $\in \mathcal{L}^{\mathbb{w}}$) the *controller*, and their interconnection $\mathcal{P} \cap \mathcal{C}$ (hence also $\in \mathcal{L}^{\mathbb{w}}$), the *controlled system*. This signifies that in the controlled system, the trajectory w has to obey both the laws of $\mathcal{P}$ and $\mathcal{C}$, which leads to the point of view that control means restricting the plant behavior to a subset, the intersection of the plant and the controller.

The controller $\mathcal{C}$ is said to be a *regular controller* for $\mathcal{P}$ if

$$\mathsf{p}(\mathcal{P} \cap \mathcal{C}) = \mathsf{p}(\mathcal{P}) + \mathsf{p}(\mathcal{C}).$$

and *superregular* if, in addition,

$$\mathsf{n}(\mathcal{P} \cap \mathcal{C}) = \mathsf{n}(\mathcal{P}) + \mathsf{n}(\mathcal{C}).$$

The origin and the significance of these concepts is discussed in, for example, [1, section VII]. The classical input/state/output based sensor-output-to-actuator-input controllers that dominate the field of control are superregular. Controllers that are regular, but not superregular, are relevant in control, much more so than is appreciated, for example as PID controllers, or as control devices that do not act as sensor-output-to-actuator-input feedback controllers.

Superregularity means that the interconnection of the plant with the controller can take place at any moment in time. The controller $\mathcal{C} \in \mathcal{L}^{\mathbb{w}}$ is superregular for $\mathcal{P} \in \mathcal{L}^{\mathbb{w}}$ if and only if for all $w_1 \in \mathcal{P}$ and $w_2 \in \mathcal{C}$, there exists a $w \in (\mathcal{P} \cap \mathcal{C})^{\text{closure}}$ such that w_1' and w_2' defined by

$$w_1'(t) = \begin{cases} w_1(t) & \text{for } t \leq 0 \\ w(t) & \text{for } t > 0 \end{cases},$$

and

$$w_2'(t) = \begin{cases} w_2(t) & \text{for } t \leq 0, \\ w(t) & \text{for } t > 0 \end{cases}$$

belongs to $\mathcal{P}$ and $\mathcal{C}$, respectively. Hence, for a superregular interconnection, any distinct past histories in $\mathcal{P}$ and $\mathcal{C}$ can be continued as one and the same future trajectory in $\mathcal{P} \cap \mathcal{C}$. In [5] it has been shown that superregularity can also be viewed as feedback.

The controller $\mathcal{C}$ is said to be *stabilizing* if $\mathcal{P} \cap \mathcal{C}$ is stable, that is, if $w \in \mathcal{P} \cap \mathcal{C}$ implies $w(t) \to 0$ as $t \to \infty$. Note that we consider stability as a property of an autonomous behavior (a behavior $\mathcal{B}$ with $\mathsf{m}(\mathcal{B}) = 0$). In the input/output setting, as in [4], the interconnection of $\mathcal{P}$ and $\mathcal{C}$ is defined to be stable if the system obtained by injecting artificial arbitrary inputs at the interconnection terminals is bounded-input/bounded-output stable. Our stability definition requires that $w(t) \to 0$ for $t \to \infty$ in $\mathcal{P} \cap \mathcal{C}$. It turns out that bounded-input/bounded-output stability requires (i) our stability, combined with (ii) superregularity. Interconnections that are not superregular cannot be bounded-input/bounded-output stable. However, for physical systems these concepts (stability and superregularity) are quite unrelated. For example, the harmonic oscillator $M\frac{d^2}{dt^2}w_1 + Kw_1 = w_2$, with $M, K > 0$, is stabilized by the damper $w_2 = -D\frac{d}{dt}w_1$ if $D > 0$. In our opinion, it makes little sense to call this interconnection unstable, just because the interconnection is not superregular.

Regularity and superregularity can be expressed in terms of left-prime kernel representations with rational symbols.

Proposition 5. *Consider the plant $\mathcal{P} \in \mathcal{L}^{\mathtt{w}}$. Assume that $\mathcal{P}$ is stabilizable. Let $\mathcal{P}$ be described by $P\left(\frac{d}{dt}\right)w = 0$ with $P \in \mathbb{R}(\xi)^{\bullet \times \mathtt{w}}$ left prime over $\mathbb{R}(\xi)_{\mathcal{S}}$. By stabilizability of $\mathcal{P}$ such a representation exists.*

1. *$\mathcal{C} \in \mathcal{L}^{\mathtt{w}}$ is a regular stabilizing controller if and only if $\mathcal{C}$ admits a representation $C\left(\frac{d}{dt}\right)w = 0$ with $C \in \mathbb{R}(\xi)^{\bullet \times \mathtt{w}}$ left prime over $\mathbb{R}(\xi)_{\mathcal{S}}$, and such that*
$$G = \begin{bmatrix} P \\ C \end{bmatrix}$$
is square and $\mathbb{R}(\xi)_{\mathcal{S}}$-unimodular, that is, with $\mathtt{determinant}(G)$ *miniphase.*
2. *$\mathcal{C} \in \mathcal{L}^{\mathtt{w}}$ is a superregular stabilizing controller if and only if $\mathcal{C}$ admits a representation $C\left(\frac{d}{dt}\right)w = 0$ with $C \in \mathbb{R}(\xi)^{\bullet \times \mathtt{w}}$ left prime over $\mathbb{R}(\xi)_{\mathcal{PS}}$, and such that*
$$G = \begin{bmatrix} P \\ C \end{bmatrix}$$
is square and $\mathbb{R}(\xi)_{\mathcal{PS}}$-unimodular, that is, with $\mathtt{determinant}(G)$ *biproper and miniphase.* ∎

The equivalence of the following statements can be shown:

$$[\![\mathcal{P} \text{ is stabilizable}]\!] \Leftrightarrow [\![\exists \text{ a regular controller } \mathcal{C} \text{ that stabilizes } \mathcal{P}]\!]$$
$$\Leftrightarrow [\![\exists \text{ a superregular controller } \mathcal{C} \text{ that stabilizes } \mathcal{P}]\!].$$

Combining this with the previous theorem leads to the following result on matrices of rational functions.

Corollary 6. *1. Assume that* $G \in \mathbb{R}(\xi)_{\mathcal{S}}^{n_1 \times n_2}$ *is left prime over* $\mathbb{R}(\xi)_{\mathcal{S}}$. *Then there exists* $F \in \mathbb{R}(\xi)_{\mathcal{S}}^{(n_2-n_1)\times n_2}$ *such that*

$$\begin{bmatrix} G \\ F \end{bmatrix}$$

is $\mathbb{R}(\xi)_{\mathcal{S}}$*-unimodular.*

2. Assume that $G \in \mathbb{R}(\xi)_{\mathcal{PS}}^{n_1 \times n_2}$ *is left prime over* $\mathbb{R}(\xi)_{\mathcal{PS}}$. *Then there exists* $F \in \mathbb{R}(\xi)_{\mathcal{PS}}^{(n_2-n_1)\times n_2}$ *such that*

$$\begin{bmatrix} G \\ F \end{bmatrix} \text{ is}$$

$\mathbb{R}(\xi)_{\mathcal{PS}}$*-unimodular.*

8 Parametrization of the Set of Regular Stabilizing, Superregular Stabilizing, and Dead-beat Controllers

In this section, we parametrize the set of regular and superregular controllers that stabilize a given stabilizable plant $\mathcal{P} \in \mathcal{L}^{\bullet}$.

8.1 Regular Stabilizing Controllers

Step 1. The parametrization starts from a kernel representation $P\left(\frac{d}{dt}\right) w = 0$ of $\mathcal{P}$, with $P \in \mathbb{R}(\xi)^{\mathsf{p}(\mathcal{P})\times \mathsf{w}(\mathcal{P})}$ left prime over $\mathbb{R}(\xi)_{\mathcal{S}}$. By stabilizability of $\mathcal{P}$, such a representation exists.

Step 2. Construct a $P' \in \mathbb{R}(\xi)_{\mathcal{S}}^{\mathsf{m}(\mathcal{P})\times \mathsf{w}(\mathcal{P})}$ such that

$$\begin{bmatrix} P \\ P' \end{bmatrix}$$

is $\mathbb{R}(\xi)_{\mathcal{S}}$-unimodular. By corollary 6, such a P' exists.

Step 3. The set of regular stabilizing controllers $\mathcal{C} \in \mathcal{L}^{\mathsf{w}(\mathcal{P})}$ is given as the systems with kernel representation $C(\frac{d}{dt})w = 0$, where

$$C = F_1 P + F_2 P',$$

with $F_1 \in \mathbb{R}(\xi)_{\mathcal{S}}^{\mathsf{m}(\mathcal{P})\times \mathsf{p}(\mathcal{P})}$ is free and $F_2 \in \mathbb{R}(\xi)_{\mathcal{S}}^{\mathsf{m}(\mathcal{P})\times \mathsf{m}(\mathcal{P})}$ is $\mathbb{R}(\xi)_{\mathcal{S}}$-unimodular, that is, with $\texttt{determinant}(F_2)$ miniphase.

Step 3'. This parametrization may be further simplified using controllability equivalence, by identifying controllers that have the same controllable part, that is, by considering controllers up to controllability equivalence. The set of controllers $\mathcal{C} \in \mathcal{L}^{\mathsf{w}(\mathcal{P})}$ with kernel representation $C(\frac{d}{dt})w = 0$ and C of the form

$$C = FG + G',$$

with $F \in \mathbb{R}(\xi)_{\mathcal{S}}^{\mathsf{m}(\mathcal{P})\times \mathsf{p}(\mathcal{P})}$ free, consists of regular stabilizing controllers, and contains an element of the equivalence class modulo controllability of each regular stabilizing controller for $\mathcal{P}$.

8.2 Superregular Stabilizing Controllers

Step 1. The parametrization starts from a kernel representation $P\left(\frac{d}{dt}\right) w = 0$ of $\mathcal{P}$, with $P \in \mathbb{R}(\xi)^{\mathsf{p}(\mathcal{P})\times \mathsf{w}(\mathcal{P})}$ left prime over $\mathbb{R}(\xi)_{\mathcal{PS}}$. By stabilizability of $\mathcal{P}$, such a representation exists.

Step 2. Construct a $P' \in \mathbb{R}(\xi)_{\mathcal{S}}^{\mathsf{m}(\mathcal{P})\times \mathsf{w}(\mathcal{P})}$ such that

$$\begin{bmatrix} P \\ P' \end{bmatrix}$$

is $\mathbb{R}(\xi)_{\mathcal{PS}}$-unimodular. By corollary 6, such a P' exists.

Step 3. The set of superregular stabilizing controllers $\mathcal{C} \in \mathcal{L}^{\mathsf{w}(\mathcal{P})}$ is given as the systems with kernel representation $C(\frac{d}{dt})w = 0$, where

$$C = F_1 P + F_2 P',$$

with $F_1 \in \mathbb{R}(\xi)_{\mathcal{PS}}^{\mathsf{m}(\mathcal{P})\times \mathsf{p}(\mathcal{P})}$ free and $F_2 \in \mathbb{R}(\xi)_{\mathcal{PS}}^{\mathsf{m}(\mathcal{P})\times \mathsf{m}(\mathcal{P})}$ $\mathbb{R}(\xi)_{\mathcal{PS}}$-unimodular, that is, with $\mathtt{determinant}(F_2)$ biproper and miniphase.

Step 3'. This parametrization may be further simplified using controllability equivalence, by identifying controllers that have the same controllable part, that is, by considering controllers up to controllability equivalence. The set of controllers $\mathcal{C} \in \mathcal{L}^{\mathsf{w}(\mathcal{P})}$ with kernel representation $C(\frac{d}{dt})w = 0$ and C of the form

$$C = FG + G,'$$

with $F \in \mathbb{R}(\xi)_{\mathcal{PS}}^{\mathsf{m}(\mathcal{P})\times \mathsf{p}(\mathcal{P})}$ free, consists of superregular stabilizing controllers, and contains an element of the equivalence class modulo controllability of each superregular stabilizing controller for $\mathcal{P}$.

It is of interest to compare these parametrizations with the one obtained in [3]. We now show a very simple example to illustrate the difference between the parametrizations obtained in step 3 and step 3'.

Example: Consider the plant $y = 0\,u$, hence $P = \begin{bmatrix} 1 & 0 \end{bmatrix}$, and the superregular stabilizing controller $u + \alpha\frac{d}{dt}u = 0$, with $\alpha \geq 0$. Take $P' = \begin{bmatrix} 0 & 1 \end{bmatrix}$ in the parametrizations. The set of (super)regular stabilizing controllers is given by $C\left(\frac{d}{dt}\right) u = 0$, with $C \in \mathbb{R}(\xi)$ miniphase in the regular case, and miniphase and biproper in the superregular case. Taking $F_2(\xi) = (1 + \alpha\xi)/(1 + 2\alpha\xi)$, for example, yields the controller $u + \alpha\frac{d}{dt}u = 0$, with $\alpha \geq 0$. The parametrization in step 3' yields only the controller $u = 0$, which is indeed the controllable part of $u + \alpha\frac{d}{dt}u = 0$.

This example illustrates that the parametrization in step 3' does not yield all the (super)regular stabilizing controllers, although it yields all the stabilizing controller transfer functions. Note that the parametrization of step 3 does exclude the destabilizing controller $u + \alpha\frac{d}{dt}u = 0$, with $\alpha < 0$.

The trajectory-based parametrization is not only more general, but it also give sharper results. It yields all stabilizing controllers, without having to resort to equivalence modulo controllability.

Acknowledgments

The SISTA Research program is supported by the Research Council KUL: GOA AMBioRICS, CoE EF/05/006 Optimization in Engineering, several PhD/postdoc & fellow grants; by the Flemish Government: FWO: PhD/postdoc grants, projects, G.0407.02 (support vector machines), G.0197.02 (power islands), G.0141.03 (Identification and cryptography), G.0491.03 (control for intensive care glycemia), G. 0120.03 (QIT), 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, research communities (ICCoS, ANMMM, MLDM); by IWT: PhD Grants, McKnow-E, Eureka-Flite2; and by the Belgian Federal Science Policy Office: IUAP P6/04 (Dynamical systems, Control and Optimization, 2007-2011).

This research is also supported by the Japanese Government under the 21st Century COE (Center of Excellence) program for research and education on complex functional mechanical systems, and by the JSPS Grant-in-Aid for Scientific Research (B) No. 18360203, and also by Grand-in-Aid for Exploratory Research No. 17656138.

References

1. Belur, M., Trentelman, H.L.: Stabilization, pole placement, and regular implementability. IEEE Transactions on Automatic Control 47, 735–744 (2002)
2. Kučera, V.: Stability of discrete linear feedback systems, paper 44.1. In: Proceedings of the 6-th IFAC Congress, Boston, Massachusetts, USA (1975)
3. Kuijper, M.: Why do stabilizing controllers stabilize? Automatica 34, 621–625 (1995)
4. Vidyasagar, M.: Control System Synthesis. The MIT Press, Cambridge (1985)
5. Willems, J.C.: On interconnections, control and feedback. IEEE Transactions on Automatic Control 42, 326–339 (1997)
6. Willems, J.C., Yamamoto, Y.: Behaviors defined by rational functions. Linear Algebra and Its Applications 425, 226–241 (2007)
7. Youla, D.C., Bongiorno, J.J., Jabr, H.A.: Modern Wiener-Hopf design of optimal controllers, Part I: The single-input case, Part II: The multivariable case. IEEE Transactions on Automatic Control 21, 3–14, 319–338 (1976)

Author Index

Lecture Notes in Control and Information Sciences

Edited by M. Thoma, M. Morari

Further volumes of this series can be found on our homepage:
springer.com

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]

Vol. 356: Karatkevich A.:
Dynamic Analysis of Petri Net-Based Discrete Systems
166 p. 2007 [978-3-540-71464-4]

Vol. 355: Zhang H.; Xie L.:
Control and Estimation of Systems with Input/Output Delays
213 p. 2007 [978-3-540-71118-6]

Vol. 354: Witczak M.:
Modelling and Estimation Strategies for Fault Diagnosis of Non-Linear Systems
215 p. 2007 [978-3-540-71114-8]

Vol. 353: Bonivento C.; Isidori A.; Marconi L.; Rossi C. (Eds.)
Advances in Control Theory and Applications
305 p. 2007 [978-3-540-70700-4]

Vol. 352: Chiasson, J.; Loiseau, J.J. (Eds.)
Applications of Time Delay Systems
358 p. 2007 [978-3-540-49555-0]

Vol. 351: Lin, C.; Wang, Q.-G.; Lee, T.H., He, Y.
LMI Approach to Analysis and Control of Takagi-Sugeno Fuzzy Systems with Time Delay
204 p. 2007 [978-3-540-49552-9]

Vol. 350: Bandyopadhyay, B.; Manjunath, T.C.; Umapathy, M.
Modeling, Control and Implementation of Smart Structures 250 p. 2007 [978-3-540-48393-9]

Vol. 349: Rogers, E.T.A.; Galkowski, K.; Owens, D.H.
Control Systems Theory and Applications for Linear Repetitive Processes
482 p. 2007 [978-3-540-42663-9]

Vol. 347: Assawinchaichote, W.; Nguang, K.S.; Shi P.
Fuzzy Control and Filter Design for Uncertain Fuzzy Systems
188 p. 2006 [978-3-540-37011-6]

Vol. 346: Tarbouriech, S.; Garcia, G.; Glattfelder, A.H. (Eds.)
Advanced Strategies in Control Systems with Input and Output Constraints
480 p. 2006 [978-3-540-37009-3]

Vol. 345: Huang, D.-S.; Li, K.; Irwin, G.W. (Eds.)
Intelligent Computing in Signal Processing and Pattern Recognition
1179 p. 2006 [978-3-540-37257-8]

Vol. 344: Huang, D.-S.; Li, K.; Irwin, G.W. (Eds.)
Intelligent Control and Automation
1121 p. 2006 [978-3-540-37255-4]

Vol. 341: Commault, C.; Marchand, N. (Eds.)
Positive Systems
448 p. 2006 [978-3-540-34771-2]

Vol. 340: Diehl, M.; Mombaur, K. (Eds.)
Fast Motions in Biomechanics and Robotics
500 p. 2006 [978-3-540-36118-3]

Vol. 339: Alamir, M.
Stabilization of Nonlinear Systems Using Receding-horizon Control Schemes
325 p. 2006 [978-1-84628-470-0]

Vol. 338: Tokarzewski, J.
Finite Zeros in Discrete Time Control Systems
325 p. 2006 [978-3-540-33464-4]

Vol. 337: Blom, H.; Lygeros, J. (Eds.)
Stochastic Hybrid Systems
395 p. 2006 [978-3-540-33466-8]

Vol. 336: Pettersen, K.Y.; Gravdahl, J.T.; Nijmeijer, H. (Eds.)
Group Coordination and Cooperative Control
310 p. 2006 [978-3-540-33468-2]

Vol. 335: Kozłowski, K. (Ed.)
Robot Motion and Control
424 p. 2006 [978-1-84628-404-5]

Vol. 334: Edwards, C.; Fossas Colet, E.; Fridman, L. (Eds.)
Advances in Variable Structure and Sliding Mode Control
504 p. 2006 [978-3-540-32800-1]

Vol. 333: Banavar, R.N.; Sankaranarayanan, V.
Switched Finite Time Control of a Class of Underactuated Systems
99 p. 2006 [978-3-540-32799-8]

Vol. 332: Xu, S.; Lam, J.
Robust Control and Filtering of Singular Systems
234 p. 2006 [978-3-540-32797-4]

Vol. 331: Antsaklis, P.J.; Tabuada, P. (Eds.)
Networked Embedded Sensing and Control
367 p. 2006 [978-3-540-32794-3]

Vol. 330: Koumoutsakos, P.; Mezic, I. (Eds.)
Control of Fluid Flow
200 p. 2006 [978-3-540-25140-8]

Vol. 329: Francis, B.A.; Smith, M.C.; Willems, J.C. (Eds.)
Control of Uncertain Systems: Modelling, Approximation, and Design
429 p. 2006 [978-3-540-31754-8]

Vol. 328: Loría, A.; Lamnabhi-Lagarrigue, F.; Panteley, E. (Eds.)
Advanced Topics in Control Systems Theory
305 p. 2006 [978-1-84628-313-0]

Vol. 327: Fournier, J.-D.; Grimm, J.; Leblond, J.; Partington, J.R. (Eds.)
Harmonic Analysis and Rational Approximation
301 p. 2006 [978-3-540-30922-2]

Vol. 326: Wang, H.-S.; Yung, C.-F.; Chang, F.-R.
H_∞ Control for Nonlinear Descriptor Systems
164 p. 2006 [978-1-84628-289-8]

Vol. 325: Amato, F.
Robust Control of Linear Systems Subject to Uncertain
Time-Varying Parameters
180 p. 2006 [978-3-540-23950-5]

Vol. 324: Christofides, P.; El-Farra, N.
Control of Nonlinear and Hybrid Process Systems
446 p. 2005 [978-3-540-28456-7]

Vol. 323: Bandyopadhyay, B.; Janardhanan, S.
Discrete-time Sliding Mode Control
147 p. 2005 [978-3-540-28140-5]

Vol. 322: Meurer, T.; Graichen, K.; Gilles, E.D. (Eds.)
Control and Observer Design for Nonlinear Finite and Infinite Dimensional Systems
422 p. 2005 [978-3-540-27938-9]

Vol. 321: Dayawansa, W.P.; Lindquist, A.; Zhou, Y. (Eds.)
New Directions and Applications in Control Theory
400 p. 2005 [978-3-540-23953-6]

Vol. 320: Steffen, T.
Control Reconfiguration of Dynamical Systems
290 p. 2005 [978-3-540-25730-1]

Vol. 319: Hofbaur, M.W.
Hybrid Estimation of Complex Systems
148 p. 2005 [978-3-540-25727-1]

Vol. 318: Gershon, E.; Shaked, U.; Yaesh, I.
H_∞ Control and Estimation of State-multiplicative Linear Systems
256 p. 2005 [978-1-85233-997-5]

Vol. 317: Ma, C.; Wonham, M.
Nonblocking Supervisory Control of State Tree Structures
208 p. 2005 [978-3-540-25069-2]