

# **Topics in Nonlinear Systems: Stability, Spectra and Optimization**

Robert J. Orsi

Bachelor of Science / Bachelor of Engineering

October 1996

*A thesis submitted for the degree of Master of Engineering  
of the Australian National University*

Department of Systems Engineering  
Research School of Information Sciences and Engineering  
The Australian National University



**Topics in Nonlinear Systems:  
Stability, Spectra and Optimization**



## **Statement of Originality**

These masters studies were conducted with Professor John B. Moore as supervisor, and Dr Robert E. Mahony and Dr Rodney A. Kennedy as advisors.

The work presented in this thesis is the result of original research carried out by myself, in collaboration with others, while enrolled in the Department of Systems Engineering as a Master of Engineering student. It has not been submitted for any other degree or award in any other university or educational institution.

Robert Orsi  
October 1996



## **Acknowledgements**

During my masters I worked closely with Professor John Moore and Dr Robert Mahony and I would like to thank them for all their support, insight and enthusiasm. They made my masters a rich learning experience. I would also like to thank Dr Rodney Kennedy for his support at the start of my masters. In addition, I would like to thank Professor Iven Mareels for many enjoyable discussions, Vaughan Clarkson for his expertise on pulse trains, and lastly all the staff and students from the Department of Systems Engineering for providing such a pleasant atmosphere to work in.

In addition to the financial support of the Australian National University, I am also very grateful for the financial support received from Telstra Australia under the TRL Postgraduate Fellowship scheme.

Lastly, special thanks go to my parents for all their love and support.



## *Abstract*

This thesis considers three topics in the area of nonlinear systems.

**Internal Stability Issues in Output Stabilization of a Class of Nonlinear Control Systems.** Fundamental in the design of practical output stabilizing control laws is the internal stability of the closed loop system. The first part of the thesis considers the issue of output stabilization for a continuous time affine system using a static state output linearising control law. The system considered exhibits no drift for zero output. The control considered exponentially stabilizes the system output provided the closed loop system remains internally stable. In general it is too much to hope for globally well defined closed loop dynamics, however, it is shown that there exists a neighbourhood of the zero output level set for which the closed loop system is internally stable.

**Interleaved Pulse Train Spectrum Estimation.** The second part of the thesis considers a problem closely related to the pulse train deinterleaving problem. Considered are signals consisting of a finite though unknown number of periodic time-interleaved pulse trains. For such signals, a novel nonlinear approach is presented for determining both the number of pulse trains present and the frequency of each pulse train. This approach requires only the time of arrival data of each pulse. It is robust to noisy time of arrival data and missing pulses, and most importantly, is very computationally efficient.

**Equality Constrained Quadratic Optimization.** The last part of the thesis considers the problem of minimizing a quadratic cost subject to purely quadratic equality constraints. This problem is solved by first relating it to a standard semidefinite programming problem. The approach taken leads to a dynamical systems analysis of semidefinite programming

and the formulation of a gradient descent flow which can be used to solve semidefinite programming problems. Though the reformulation of the initial problem as a semidefinite programming problem does not in general lead directly to a solution of the initial problem, the initial problem is solved by using a modified flow incorporating a penalty function.

## *List of Publications*

The following is a list publications in refereed journals and conference proceedings completed while I was a Master of Engineering student.

### **Journal Papers**

1. R. J. Orsi, J. B. Moore and R. E. Mahony. “Spectrum Estimation of Interleaved Pulse Trains”. Submitted to *IEEE Transactions on Signal Processing*, 1996.
2. R. J. Orsi, R. E. Mahony and J. B. Moore. “A Dynamical Systems Analysis of Semi-definite Programming with Application to Quadratic Optimization with Purely Quadratic Equality Constraints”. To be submitted.

### **Conference Papers**

3. R. J. Orsi and R. E. Mahony. “Internal Stability Issues in Output Stabilization of a Class of Nonlinear Control Systems”. Presented at *Mathematical Theory of Networks and Systems*, St. Louis, USA, 1996.
4. R. J. Orsi, J. B. Moore and R. E. Mahony. “Interleaved Pulse Train Spectrum Estimation”. In *Proceedings of the Fourth International Symposium on Signal Processing and its Applications*, pp. 125–128, Gold Coast, Australia, 1996.

Papers [1] and [4] contain overlapping material.



# *Contents*

Statement of Originality . . . . .	iii
Acknowledgements . . . . .	v
<b>Abstract</b>	<b>vii</b>
<b>List of Publications</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Internal Stability Issues in Output Stabilization of a Class of Nonlinear Control Systems</b>	<b>5</b>
2.1 Problem Formulation . . . . .	6
2.2 Main Result . . . . .	8
<b>3 Interleaved Pulse Train Spectrum Estimation</b>	<b>15</b>
3.1 Problem Formulation and Approach . . . . .	16
3.2 A Non-Generic Special Case . . . . .	17
3.2.1 The Non-Generic Case: A Simulation Example . . . . .	21
3.3 The Generic Case . . . . .	22

3.3.1	Additional Processing . . . . .	24
3.3.2	When all else fails.... . . . .	25
3.4	Further Analysis . . . . .	26
3.4.1	Decreasing N . . . . .	26
3.4.2	Noisy Time of Arrival Data . . . . .	27
3.4.3	Missing Pulses . . . . .	30
3.5	Additional Comments and Concluding Remarks . . . . .	30
<b>4</b>	<b>Equality Constrained Quadratic Optimization</b>	<b>33</b>
4.1	Problem Formulation . . . . .	35
4.2	The Geometry of the Feasible Sets . . . . .	38
4.3	Gradient Flow . . . . .	45
4.4	Further Analysis . . . . .	51
4.5	Gradient Flow with Penalty Function . . . . .	56
4.6	Solution Methods . . . . .	58
4.7	Conclusion . . . . .	59
<b>5</b>	<b>Conclusion</b>	<b>63</b>

# *Chapter 1*

## *Introduction*

Many problems and physical systems of interest are inherently nonlinear. Superposition does not hold for nonlinear systems and this tends to make their analysis difficult. Given a nonlinear problem, one would perhaps hope that the problem could be linearized in some (meaningful) manner so that the problem could be analyzed using linear techniques. Unfortunately, this is not always possible, consider for example nonlinear optimization problems. Even for those problems that can be linearized, the linearization process may still fail to solve the problem at hand. For example, given a nonlinear control system, a feedback controller designed to increase the stability of a linearized model of the system may in fact drastically reduce the stability of the actual nonlinear system (Freeman & Kokotović 1996). Hence, linear analysis of some nonlinear problems is either not possible or simply does not work. Many important problems fall into this category and must be analyzed from an inherently nonlinear viewpoint.

In this thesis I look at three topics for which nonlinear behaviour is the essential ingredient. The three topics are considered individually and are presented in Chapters 2, 3 and 4.

The first of the three topics presented considers internal stability in output stabilization of a class of nonlinear control systems. Fundamental in the design of practical output stabilizing control laws is whether the state remains bounded and well defined for all time, that is, whether the closed loop system remains *internally stable*. In general, it is too much to hope

that a nonlinear system will have globally well defined closed loop dynamics (Sussmann 1990, Sussmann & Kokotovic 1991). By assuming a certain structure of the system equations, it is possible to obtain internal stability results valid in a neighbourhood of the zero output level set. In Mahony, Mareels, Campion & Bastin (1993) it is shown that for systems with no drift and linearly appearing input, there exists an output stabilizing control and a neighbourhood of the zero output level set for which the closed loop system is internally stable. In Chapter 2 of this thesis, the result of Mahony et al. (1993) is extended to include a class of systems with non-trivial drift term. Specifically, the result is extended to systems that exhibit no drift for zero output.

The second topic considered is closely related to the *pulse train deinterleaving* problem. A periodic pulse train consists of a sequence of periodically spaced pulses. Often a single channel receiver will receive periodic pulse trains from a number of sources simultaneously. The superposition of all the received pulse trains is known as an *interleaved pulse train*. The process of determining the number of pulse trains present in this signal and associating each received pulse with a source is termed *pulse train deinterleaving*. An important application of pulse train deinterleaving is in radar detection (Wiley 1982).

Typical approaches to the pulse train deinterleaving are sequential search (Mardia 1989) and histogramming (Mardia 1989, Milojević & Popović 1992). A practical disadvantage of these algorithms is the computational effort they require. If  $N$  is the number of pulses being processed, computations are of the order of  $N^2$  (Perkins & Coat 1994).

In Chapter 3 of this thesis, a novel nonlinear approach is presented that, given an interleaved pulse train signal, determines both the number of pulse trains present and the frequency of each pulse train. (This information is termed the *interleaved pulse train spectrum*.) The proposed approach is robust to noisy time of arrival data and missing pulses, and most importantly, is very computationally efficient. If  $N$  is the number of pulses being processed, computations are of the order  $N \log N$ .

The third and final topic is presented in Chapter 4 of this thesis and considers the problem of minimising a quadratic cost subject to purely quadratic equality constraints. Such problems are non-convex and their geometry is such that in many cases the resulting con-

straint set consists of the union of a number of disconnected subsets, each with their own local minima. To overcome these problems, the quadratic optimization problem is related through a number of steps to a semidefinite programming problem. The approach taken leads to a dynamical systems analysis of semidefinite programming and the formulation of a gradient descent flow which can be used (in theory at least) to solve semidefinite programming problems. Though the reformulation of the initial problem as a semidefinite programming problem does not in general lead directly to a solution of the initial problem, the initial problem is solved by using a modified flow incorporating a penalty function.

The thesis ends with some concluding remarks and a discussion of areas of possible further research.

## ***References***

- Freeman, R. & Kokotović, P. (1996). *Robust Nonlinear Control Design: State-Space and Lyapunov Techniques*, Systems and Control: Foundations & Applications, Birkhäuser, Boston, USA.
- Mahony, R. E., Mareels, I. M., Campion, G. & Bastin, G. (1993). Output regulation for systems linear in the input, *Conference on Mathematical Theory of Networks and Systems*, Regensburg, Germany.
- Mardia, H. K. (1989). New techniques for the deinterleaving of repetitive sequences, *IEE Proceedings-F* **136**: 149–154.
- Milojević, D. J. & Popović, B. M. (1992). Improved algorithm for the deinterleaving of radar pulses, *IEE Proceedings-F* **139**: 98–104.
- Perkins, J. & Coat, I. (1994). Pulse train deinterleaving via the Hough transform, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* **3**: 197–200.
- Sussmann, H. J. (1990). Limitations on the stabilizability of globally minimum phase systems, *IEEE Transactions on Automatic Control* **35**(1): 117–119.

Sussmann, H. J. & Kokotovic, P. V. (1991). The peaking phenomenon and the global stabilization of non-linear systems, *IEEE Transactions on Automatic Control* **36**(4): 424–439.

Wiley, R. G. (1982). *Electronic Intelligence: The Analysis of Radar Signals*, Artech House.

## *Chapter 2*

# *Internal Stability Issues in Output Stabilization of a Class of Nonlinear Control Systems*

Fundamental in the design of practical stabilizing control laws is whether the state remains bounded and well defined for all time, that is, whether the closed loop system remains *internally stable*.

In general it is too much to hope that a nonlinear system will have globally well defined closed loop dynamics (Sussmann 1990, Sussmann & Kokotovic 1991). In the case where the zero dynamics are asymptotically stable (at a point  $x^*$ ) it can be shown (Byrnes & Isidori 1991, Lemma 4.2) that there exists a control law and a local neighbourhood around  $x^*$  for which the closed loop system is asymptotically stable to  $x^*$ . In general this neighbourhood does not contain the entire zero output level set and as a consequence this result has limited applicability to output stabilization.

By assuming that there exists some form of energy structure related to zeroing the output (Lin, Sontag & Wang 1994) one can hope to use Lyapunov theory to obtain results valid in a neighbourhood of the zero output level set. An energy based approach is also used in (Byrnes, Isidori & Willems 1991) for full state stabilization and is closely related to recent

work on input to state stability (Sontag 1989). If the system is derived as a Hamiltonian system then a natural energy structure exists which can be exploited for output or state stabilization (Nijmeijer & van der Schaft 1990, Ch. 10).

Alternately it is possible to get similar results by assuming certain structure of the system equations. For example, for systems with no drift and linearly appearing input, there exists a control and a neighbourhood of the zero output level set for which the closed loop system is internally stable (Mahony, Mareels, Campion & Bastin 1993, 1996).

In this chapter we take a similar approach to Mahoney et al. (1993, 1996) and extend their result to include a class of systems with non-trivial drift term. We consider the issue of output stabilization for a continuous time affine system using a static state output linearising control law. The system considered exhibits no drift for zero output, each of its outputs has relative degree one and its input-output decoupling matrix is full rank. We show that the control considered exponentially stabilizes the system output provided the closed loop system remains internally stable. We then show that there exists a neighbourhood of the zero output level set for which the closed loop system is internally stable.

The rest of this chapter consists of a problem formulation section and a main result section. In the problem formulation section, the candidate control law is introduced. The main result section contains the main result.

## ***2.1 Problem Formulation***

Consider a system of the form

$$\begin{aligned}\dot{x}(t) &= f(x(t)) + g(x(t))u, & x(0) &= x_0 \\ y(t) &= h(x(t))\end{aligned}\tag{2.1}$$

where  $x \in \mathbb{R}^n$ ,  $u \in \mathbb{R}^m$  and  $y \in \mathbb{R}^p$ , and  $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ ,  $g : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times m}$  and  $h : \mathbb{R}^n \rightarrow \mathbb{R}^p$

are smooth functions. Let  $D$  denote the derivative operator<sup>1</sup> and assume the system satisfies the following properties:

### System Properties

(P1) The system is input-output square<sup>2</sup>.

(P2)  $Dh(x)g(x)$  is full rank for all  $x \in \mathbb{R}^n$ .

**Remark 2.1** Property (P2) is equivalent to the characteristic number of each output being zero and the input-output decoupling matrix being full rank for all  $x \in \mathbb{R}^n$  (Nijmeijer & van der Schaft 1990, Section 8.1) or alternatively the system having vector relative degree  $\{1, \dots, 1\}$  for all  $x \in \mathbb{R}^n$  (Isidori 1995, pg. 220).  $\square$

Consider the evolution of the output  $y(t)$  of such a system. If  $x(t)$  satisfies (2.1), taking the time derivative of  $y(t)$  yields

$$\begin{aligned} \dot{y}(t) &= Dh(x(t))\dot{x}(t) \\ &= Dh(x(t))f(x(t)) + Dh(x(t))g(x(t))u. \end{aligned}$$

For a system of the type described, the static state feedback control law

$$u(x) = -(Dh(x)g(x))^{-1}(h(x) + Dh(x)f(x)) \quad (2.2)$$

is well defined for all  $x \in \mathbb{R}^n$ . Substituting this control into the dynamics for  $y(t)$  gives

$$\dot{y}(t) = -y(t),$$

provided the closed loop system remains internally stable. If the system does remain intern-

---

<sup>1</sup>For a differentiable map  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ ,

$$Df(x) := \begin{pmatrix} \frac{\partial f^1}{\partial x^1}(x) & \cdots & \frac{\partial f^1}{\partial x^n}(x) \\ \vdots & & \vdots \\ \frac{\partial f^m}{\partial x^1}(x) & \cdots & \frac{\partial f^m}{\partial x^n}(x) \end{pmatrix}.$$

<sup>2</sup>A system is input-output square if it possesses the same number of inputs as outputs, that is, if  $p = m$ .

ally stable

$$y(t) = y(0)e^{-t}$$

and the output converges exponentially to zero.

We term the static control law (2.2) the *output linearising control* for the system. Due to its simplicity and the strong convergence of  $y(t)$  it induces, this control strategy is a desirable candidate for output stabilization of systems of the type described. However, without a more explicit knowledge of (2.1), internal stability of the closed loop is in doubt.

## 2.2 Main Result

In this section we show that there exists a neighbourhood of the set  $\{x \in \mathbb{R} \mid h(x) = 0\}$  for which the closed loop system, consisting of the system (2.1) and control (2.2), is internally stable.

Consider a system of the form (2.1) which satisfies properties (P1) and (P2). Given an input  $u$  and an initial condition  $x_0$ , let  $x(t; x_0)$  denote the state of the system at time  $t$  for the given initial condition and input. Define

$$G(x) := g(x)(Dh(x)g(x))^{-1} \quad \text{and} \quad (2.3)$$

$$F(x) := (I - G(x)Dh(x))f(x). \quad (2.4)$$

The closed loop response of the system (2.1) to the control input (2.2) is given by

$$x(t; x_0) = x_0 + \int_0^t \left( F(x(\tau; x_0)) - G(x(\tau; x_0))h(x(\tau; x_0)) \right) d\tau. \quad (2.5)$$

In the sequel we only consider systems satisfying the following additional property:

### System Property

(P3) The drift term  $f(x) = 0$  whenever the output  $h(x) = 0$ .

**Remark 2.2** Systems for which the output is related to the energy of the system tend to satisfy (P3). For example, if the system output consists of kinetic plus potential energy, with potential energy normalised such that it is greater than or equal to zero, zero output corresponds to zero energy in the system. As long as the applied control is zero then the system is naturally at rest for zero output and hence  $f(x) = 0$  when  $h(x) = 0$ .  $\square$

**Lemma 2.3** Consider a system of the form (2.1) which satisfies properties (P1), (P2) and (P3). Let the input  $u$  be given by (2.2) and let  $F(x)$  be defined as in (2.4). Then given  $x^* \in \{x \in \mathbb{R}^n \mid h(x) = 0\}$  and  $r > 0$ , there exist constants  $M_{x^*,r} > 0$  and  $r'$ ,  $0 < r' \leq r$ , such that

$$|F(x)| \leq M_{x^*,r}|h(x)| \quad \forall x \in B_{x^*}(r').$$

$\diamond$

**PROOF.** As  $Dh(x)g(x)$  is full rank,  $Dh(x)$  must also be full rank. This implies that there exists a function  $z : \mathbb{R}^n \rightarrow \mathbb{R}^{n-m}$  and a constant  $r'$ ,  $0 < r' \leq r$ , such that the function

$$\phi(x) := \begin{pmatrix} y \\ z \end{pmatrix}$$

is a diffeomorphism on the ball  $B_{x^*}(r')$  (Nijmeijer & van der Schaft 1990, Proposition 2.18, pg. 35).

In the ball  $B_{x^*}(r')$  rewrite  $f$  and  $F$  as functions of the new coordinates  $y$  and  $z$ ,

$$\begin{aligned} \tilde{f}(y, z) &:= f \left( \phi^{-1} \begin{pmatrix} y \\ z \end{pmatrix} \right) && \text{and} \\ \tilde{F}(y, z) &:= F \left( \phi^{-1} \begin{pmatrix} y \\ z \end{pmatrix} \right). \end{aligned}$$

For a given  $z$ , if the set  $\{y \mid \phi^{-1}((y^T \ z^T)^T) \in B_{x^*}(r')\}$  is non-empty, let

$$M_z := \sup_{\{y \mid \phi^{-1}((y^T \ z^T)^T) \in B_{x^*}(r')\}} \frac{|\tilde{F}(y, z) - \tilde{F}(0, z)|}{|(y^T \ z^T)^T - (0 \ z^T)^T|}.$$

Otherwise, let  $M_z := 0$ . Define

$$M_{x^*, r} := \sup_z M_z.$$

Let  $x \in B_{x^*}(r')$  and  $(y^T \ z^T)^T = \phi^{-1}(x)$ . There exist two possibilities,  $y = 0$  and  $y \neq 0$ . If  $y = 0$  then by assumption  $f(x) = 0$ . This in turn implies that  $F(x) = 0$  and the result follows. If  $y \neq 0$  then noting that  $\tilde{F}(0, z) = 0$ ,

$$\begin{aligned} |F(x)| &= \frac{|\tilde{F}(y, z) - \tilde{F}(0, z)|}{|(y^T \ z^T)^T - (0 \ z^T)^T|} |(y^T \ z^T)^T - (0 \ z^T)^T| \\ &= \frac{|\tilde{F}(y, z) - \tilde{F}(0, z)|}{|(y^T \ z^T)^T - (0 \ z^T)^T|} |y| \\ &\leq M_{x^*, r} |y|. \end{aligned}$$

■

**Theorem 2.4** *Consider a system of the form (2.1) which satisfies properties (P1), (P2) and (P3). Then, for the closed loop system using the output linearizing control law (2.2), there exists an open neighbourhood  $\Omega \subset \mathbb{R}^n$  of the set  $\{x \in \mathbb{R}^n \mid h(x) = 0\}$  such that for any initial condition  $x_0 \in \Omega$  the solution of the system  $x(t; x_0)$  is well defined and bounded for all time and the output  $h(x(t)) \rightarrow 0$  as  $t \rightarrow \infty$ .* ◇

**PROOF.** Let  $F(x)$  be defined as in (2.4) and let  $x^* \in \{x \in \mathbb{R}^n \mid h(x) = 0\}$  and  $r > 0$ . By Lemma 2.3 there exist constants  $M_{x^*, r} > 0$  and  $r', 0 < r' \leq r$ , such that

$$|F(x)| \leq M_{x^*, r} |h(x)| \quad \forall x \in B_{x^*}(r').$$

Let

$$\Omega_{x^*, r} := \left\{ x \in B_{x^*}(r'/4) \mid (M_{x^*, r} + |G(x)|)|h(x)|e^{\lambda|h(x)|} < \frac{r'}{2} \right\}$$

where  $G(x)$  is given by (2.3) and  $\lambda = \lambda(x^*, r')$  and is given by

$$\lambda(x^*, r') := \sup_{x, y \in B_{x^*}(r')} \frac{|G(x) - G(y)|}{|x - y|}.$$

Further, define the set

$$\Omega := \cup_{x^* \in \{x \mid h(x)=0\}} \cup_{r>0} \Omega_{x^*, r}.$$

Note that each set  $\Omega_{x^*, r}$  is non-empty (as  $x^* \in \Omega_{x^*, r}$ ) and that each of these sets is an open subset of  $\mathbb{R}^n$ . Hence  $\Omega$  is open and  $\{x \in \mathbb{R}^n \mid h(x) = 0\} \subset \Omega$ .

We now proceed to show that  $x_0 \in \Omega$  is sufficient for the state to remain well defined for all time.

Let  $x_0 \in \Omega$ . Then  $x_0 \in \Omega_{x^*, r}$  for some  $x^*$  and  $r$ . For this particular  $x^*$  and  $r$ , let  $M_{x^*, r}$  and  $r'$  be the corresponding constants defined in Lemma 2.3. As  $f(x)$ ,  $g(x)$  and  $h(x)$  are smooth and  $Dh(x)g(x)$  is full rank there exists a unique local solution to the system,  $x(t; x_0)$ , that is well defined on some maximal interval  $[0, t^*)$ .

The proof proceeds by contradiction. Assume that there exists a time for which  $|x(t; x_0) - x_0| \geq 3r'/4$ . Define  $t_1 < t^*$  as

$$t_1 = \inf_{t>0} \{t \mid |x(t; x_0) - x_0| \geq 3r'/4\}.$$

Note that  $t_1$  is defined in such a way that for the given  $x_0 \in \Omega_{x^*, r}$ ,  $x(t; x_0) \in B_{x^*}(r')$  for  $t \in [0, t_1)$ .

The closed loop system response (2.5) may be rewritten as

$$x(t; x_0) - x_0 = \int_0^t \left( F(x(\tau; x_0)) - G(x(\tau; x_0))h(x(\tau; x_0)) \right) d\tau.$$

Computing the norm of this expression and approximating the integral for  $t \in [0, t_1)$  one obtains

$$\begin{aligned}
|x(t; x_0) - x_0| &\leq \int_0^{t_1} \left( |F(x(\tau; x_0))| + |G(x(\tau; x_0))| |h(x(\tau; x_0))| \right) d\tau \\
&\leq \int_0^{t_1} (M_{x^*, r} + |G(x_0)| + \lambda |x(\tau; x_0) - x_0|) |h(x_0)| e^{-\tau} d\tau \\
&= \left( M_{x^*, r} + |G(x_0)| \right) |h(x_0)| (1 - e^{-t_1}) \\
&\quad + \int_0^{t_1} \lambda |x(\tau; x_0) - x_0| |h(x_0)| e^{-\tau} d\tau \\
&\leq \left( M_{x^*, r} + |G(x_0)| \right) |h(x_0)| + \int_0^{t_1} \lambda |x(\tau; x_0) - x_0| |h(x_0)| e^{-\tau} d\tau.
\end{aligned}$$

The above inequality is in a form in which the Bellman-Gronwall Lemma (Sontag 1990, Lemma C.3.1, pg. 346) may be applied. Application of the lemma yields that

$$|x(t; x_0) - x_0| \leq (M_{x^*, r} + |G(x_0)|) |h(x_0)| e^{\lambda |h(x_0)| (1 - e^{-t_1})}.$$

Hence,

$$|x(t; x_0) - x_0| \leq (M_{x^*, r} + |G(x_0)|) |h(x_0)| e^{\lambda |h(x_0)|}$$

and by construction of  $\Omega_{x^*, r}$  one has that

$$|x(t; x_0) - x_0| < r'/2.$$

This contradicts our starting assumption, namely the existence of a time at which  $|x(t; x_0) - x_0| \geq 3r'/4$ . Hence, the state must remain in a ball about  $x^*$  of radius  $3r'/4$  and is well defined for all time. The remainder of the theorem follows directly from the nature of the output function. ■

## References

Byrnes, C. I. & Isidori, A. (1991). Asymptotic stabilization of minimum phase nonlinear

- systems, *IEEE Transactions on Automatic Control* **36**(10): 1122–1137.
- Byrnes, C. I., Isidori, A. & Willems, J. C. (1991). Passivity, feedback equivalence, and the global stabilization of minimum phase nonlinear systems, *IEEE Transactions on Automatic Control* **36**(11): 1228–1240.
- Isidori, A. (1995). *Nonlinear Control Systems*, Communications and Control Engineering Series, third edn, Springer-Verlag, Berlin, Germany.
- Lin, Y., Sontag, E. D. & Wang, Y. (1994). A smooth converse Lyapunov theorem for robust stability, Private communication.
- Mahony, R. E., Mareels, I. M., Bastin, G. & Campion, G. (1996). Output stabilization of square non-linear systems. To appear in *Automatica*.
- Mahony, R. E., Mareels, I. M., Campion, G. & Bastin, G. (1993). Output regulation for systems linear in the input, *Conference on Mathematical Theory of Networks and Systems*, Regensburg, Germany.
- Nijmeijer, H. & van der Schaft, A. (1990). *Nonlinear Dynamical Control Systems*, Springer Verlag, New York, USA.
- Sontag, E. D. (1989). Smooth stabilization implies coprime factorization, *IEEE Transactions on Automatic Control* **34**(4): 435–443.
- Sontag, E. D. (1990). *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, Texts in Applied Mathematics, Springer Verlag, New York, USA.
- Sussmann, H. J. (1990). Limitations on the stabilizability of globally minimum phase systems, *IEEE Transactions on Automatic Control* **35**(1): 117–119.
- Sussmann, H. J. & Kokotovic, P. V. (1991). The peaking phenomenon and the global stabilization of non-linear systems, *IEEE Transactions on Automatic Control* **36**(4): 424–439.



## *Chapter 3*

# *Interleaved Pulse Train Spectrum Estimation*

A periodic pulse train consists of a sequence of periodically spaced pulses. Often a single channel receiver will receive periodic pulse trains from a number of sources simultaneously. The superposition of all the received pulse trains is known as an *interleaved pulse train*. The process of determining the number of pulse trains present in this signal and associating each received pulse with a source is termed *pulse train deinterleaving*. This process relies on the assumption that the different pulse train sources have different characteristics such as period of pulse emission. An important application of pulse train deinterleaving is in radar detection (Wiley 1982). Potential applications include computer communications and neural systems.

Typical approaches to pulse train deinterleaving are sequential search (Mardia 1989) and histogramming (Mardia 1989, Milojević & Popović 1992). A practical disadvantage of these algorithms is the computational effort they require. If  $N$  is the number of pulses being processed, computations are of the order of  $N^2$  (Perkins & Coat 1994).

A recent novel approach to pulse train deinterleaving is given in Moore & Krishnamurthy (1994) where the problem is first formulated as a stochastic discrete-time dynamic linear model. Like sequential search and histogramming, this method is quite computationally

expensive.

Rather than trying to deinterleave a received interleaved pulse train directly, we focus solely on determining the number of pulse trains present and the frequency of each pulse train. We term this information the *interleaved pulse train spectrum*. We present a novel nonlinear approach for estimating the spectrum of a signal consisting of a finite though unknown number of periodic time-interleaved pulse trains. Only the time of arrival data of each pulse is used. No knowledge of any other pulse train characteristics such pulse energy is required nor is prior knowledge of the transmitter characteristics required. An advantage of our scheme is that, if  $N$  is the number of pulses being processed, computations are of the order  $N \log N$ . Note that once the interleaved pulse train spectrum is known it is a relatively easy task to deinterleave the received signal using standard methods such as those used after histogramming (Mardia 1989).

This chapter is structured as follows. A problem formulation is first presented followed by an overview of the proposed scheme. In the remainder of the chapter, aspects of the approach are discussed in greater depth. Firstly an analysis of a special class of non-generic pulse train sequences that satisfy various simplifying assumptions is undertaken. Using insight gained from this non-generic case, some simulation results for the generic case are presented and discussed. These results include considering how the length of the data set, pulse time of arrival noise, and missing pulses effect the accuracy of the scheme. The chapter ends with some concluding remarks.

### ***3.1 Problem Formulation and Approach***

Consider  $M$  periodic pulse train sources. Let  $T_i$ ,  $f_i$  and  $\phi_i$  denote respectively the period, frequency and phase of the  $i$ th source. The received interleaved signal consists of the superposition of the  $M$  pulse trains produced by these sources. Let  $t_0, t_1, \dots, t_N$  denote the times of arrival of  $N + 1$  consecutive pulses in this signal nominally setting  $t_0 := 0$ . The problem is as follows:

**Problem 3.1** Given  $t_0, \dots, t_N$ , determine both the number of pulse trains present and the frequency of each pulse train.  $\square$

The first step in the proposed scheme is to calculate

$$x(n) := e^{j \frac{2\pi}{T} t_n} \quad \text{for } n = 0, \dots, N - 1, \quad (3.1)$$

where  $j := \sqrt{-1}$ . The signal  $x(n)$  can be thought of as taking the interval  $[t_0, t_{N-1}]$ , containing the first  $N$  pulse times of arrival, normalising its length to approximately  $2\pi$  and then wrapping this normalised interval around the unit circle. Note that as mentioned before  $t_0 = 0$ .

The next step is to take the  $N$ -length discrete Fourier transform (DFT) of (3.1). The magnitude of this transformed signal contains the information necessary to determine the interleaved pulse train spectrum. That is, it contains the information necessary to (i) determine how many pulse trains are present and (ii) make a good estimate of their frequencies. (The phase response seems to contain little information.) Redundant information within the magnitude signal can be used to improve confidence of results.

By choosing appropriate data lengths, the proposed scheme can employ the fast Fourier transform. Hence the computational cost of the scheme is of the order of  $N \log N$ .

Lastly, note that the proposed scheme is nonlinear.

### 3.2 *A Non-Generic Special Case*

The nonlinear nature of the proposed scheme makes mathematical analysis of the scheme difficult. In this section we consider a class of non-generic pulse trains that satisfy various simplifying assumptions and make a mathematical analysis of the scheme possible. As discussed in the next section, this analysis provides valuable insight into the generic case.

It is assumed that the pulse trains considered in this section satisfy the follow properties:

(P1) The period of each pulse train is rational, that is,  $T_i \in \mathbb{Q}$ ,  $i = 1, \dots, M$ .

(P2) The phase of each pulse train is zero, that is,  $\phi_i = 0$ ,  $i = 1, \dots, M$ .

The above properties imply that if the received signal contains a sufficiently large number of pulses, it will be periodic. It is assumed that a sufficiently large number of pulses have been received such that this is the case. The *overall signal period* will be denoted by  $T$ .

In addition to properties (P1) and (P2), the following assumption is also made:

(P3) The received signal consists of exactly an integer number of overall signal periods.

Let  $r_i$  denote the number of pulses from pulse train  $i$  appearing in one period of the received signal. Then

$$T = T_1 r_1 = \dots = T_M r_M \quad (3.2)$$

and the total number of pulses in one period of the received signal is

$$N_T := \sum_{i=1}^M r_i.$$

**Remark 3.2** It is assumed that the pulse time of arrival data,  $t_0, \dots, t_N$ , is noise free and that there are no missing pulses. □

The prior assumptions imply that  $N/N_T \in \mathbb{Z}$ , and that

$$t_N = \frac{N}{N_T} T. \quad (3.3)$$

**Theorem 3.3** Consider a signal consisting of  $M$  interleaved pulse trains satisfying properties (P1), (P2) and (P3). Let  $x(n)$  be defined as in (3.1) and let  $X(k)$ ,  $k = 0, \dots, N-1$ ,

denote its discrete Fourier transform. Then, defining  $k' = k - 1$ ,

$$X(k') = \begin{cases} \frac{N}{N_T} \sum_{l=0}^{N_T-1} e^{j \frac{2\pi}{N} \left( \frac{N_T}{T} t_l - k'l \right)}, \\ \quad \text{if } k' = \frac{pN}{N_T}, p = 0, \dots, N_T - 1, \\ 0, & \text{otherwise.} \end{cases}$$

Furthermore, for  $p = r_j$ ,  $\frac{k'}{N_T}$  equals  $f_j$ , the frequency of the  $j$ th pulse train.  $\diamond$

**PROOF.** By definition

$$\begin{aligned} X(k) &= \sum_{n=0}^{N-1} x(n) e^{-jk \left( \frac{2\pi}{N} \right) n} \\ &= \sum_{n=0}^{N-1} e^{j \frac{2\pi}{N_T} t_n} e^{-jk \left( \frac{2\pi}{N} \right) n}. \end{aligned}$$

Replacing  $n$  by  $mN_T + l$  and noting that (P2) implies that  $t_n = t_{mN_T+l} = mT + t_l$ ,

$$\begin{aligned} X(k) &= \sum_{m=0}^{\frac{N}{N_T}-1} \sum_{l=0}^{N_T-1} e^{j \frac{2\pi}{N_T} (mT+t_l)} e^{-jk \left( \frac{2\pi}{N} \right) (mN_T+l)} \\ &= \sum_{m=0}^{\frac{N}{N_T}-1} \sum_{l=0}^{N_T-1} e^{j 2\pi \frac{N_T}{N} (mT+t_l)} e^{-jk \left( \frac{2\pi}{N} \right) (mN_T+l)} \\ & \hspace{15em} \text{by (3.3)} \\ &= \sum_{m=0}^{\frac{N}{N_T}-1} \sum_{l=0}^{N_T-1} e^{j \frac{2\pi}{N} \left( (1-k)N_T m + \frac{N_T}{T} t_l - kl \right)} \\ &= \sum_{l=0}^{N_T-1} \left[ \sum_{m=0}^{\frac{N}{N_T}-1} \left( e^{j \frac{2\pi}{N} (1-k)N_T} \right)^m \right] e^{j \frac{2\pi}{N} \left( \frac{N_T}{T} t_l - kl \right)}. \end{aligned}$$

Consider the summation in square brackets in the line above. Letting

$$z := e^{j \frac{2\pi}{N} (1-k)N_T},$$

$$\begin{aligned} \sum_{m=0}^{\frac{N}{N_T}-1} \left( e^{j \frac{2\pi}{N} (1-k) N_T} \right)^m &= \sum_{m=0}^{\frac{N}{N_T}-1} z^m \\ &= \begin{cases} \frac{z^{\frac{N}{N_T}} - 1}{z - 1}, & \text{if } z \neq 1, \\ \frac{N}{N_T}, & \text{if } z = 1. \end{cases} \end{aligned}$$

Note that  $z^{\frac{N}{N_T}} = 1$  and hence that

$$\sum_{m=0}^{\frac{N}{N_T}-1} \left( e^{j \frac{2\pi}{N} (1-k) N_T} \right)^m = \begin{cases} 0, & \text{if } z \neq 1, \\ \frac{N}{N_T}, & \text{if } z = 1. \end{cases}$$

Also note that

$$\begin{aligned} z = 1 &\Leftrightarrow \frac{1-k}{N} N_T = -p \\ &\text{where } p \in \mathbb{Z} \text{ and } k \in \{0, \dots, N-1\} \\ &\Leftrightarrow k-1 = \frac{pN}{N_T}, \quad p = 0, \dots, N_T-1. \end{aligned}$$

Note that  $k$  above is indeed always an integer as  $N/N_T \in \mathbb{Z}$ .

Replacing  $k-1$  with  $k'$ , the DFT of  $x(n)$  can now be seen to be the expression given in the theorem statement. Furthermore by (3.3),

$$\frac{k'}{t_N} = \frac{pN/N_T}{NT/N_T} = \frac{p}{T},$$

and for  $p = r_j$ , (3.2) implies that

$$\frac{k'}{t_N} = \frac{r_j}{T_j r_j} = \frac{1}{T_j} = f_j.$$

■

Theorem 3.3 shows that the  $N$ -length DFT of  $x(n)$  is non-zero at at most  $N_T$  points. Furthermore,  $M$  of these possibly non-zero points correspond to the  $M$  pulse train frequen-

cies. Additionally, if  $f$  is a pulse train frequency, the theorem predicts the existence of harmonics at  $2f, 3f, \dots$

### 3.2.1 The Non-Generic Case: A Simulation Example

The proposed methodology was applied to a signal satisfying properties (P1), (P2) and (P3). The signal consisted of  $M = 3$  interleaved pulse trains with respective frequencies  $f_1 = 0.25$  Hz,  $f_2 = 0.75$  Hz and  $f_3 = 0.8$  Hz. The magnitude of the signal produced by the DFT is shown in Figure 3.1.

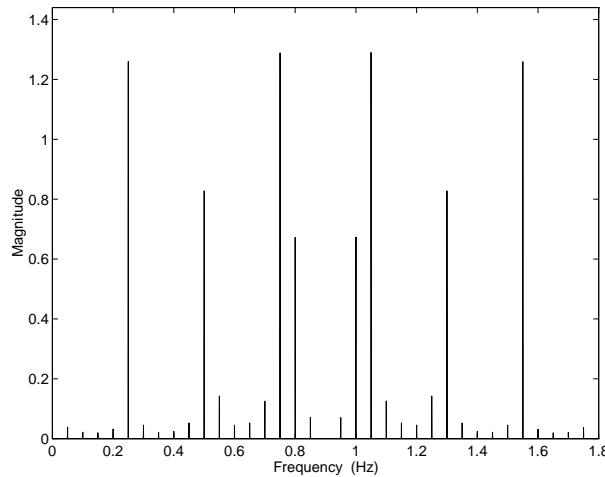


Figure 3.1: A non-generic magnitude plot.

As predicted the magnitude plot contains only a small number of non-zero values uniformly spaced in frequency. How the interleaved pulse train spectrum is identified from such a magnitude plot will be discussed in the next section which deals with the generic case. For the moment notice that the  $M$  largest values produced by the proposed scheme do not necessarily correspond to the  $M$  pulse trains present. Despite this, notice that, for this example at least,  $M = 3$  of the larger magnitudes present do correspond to pulse train frequencies.

Though it is not apparent from Figure 3.1, the DFT magnitude at 0 Hz is very large and is approximately equal to  $N$ . This term is an artifact of the processing method and is

ignored.

### 3.3 *The Generic Case*

In this section a generic case simulation is presented and discussed. The time of arrival data used in this simulation is not corrupted by noise and the data does not contain any missing pulses. (The effects of noisy time of arrival data and missing pulses are discussed in §3.4.)

The simulated signal consists of ten interleaved pulse trains. The frequencies of the pulse trains were chosen arbitrarily and are listed in Table 3.1. Each pulse train has a random phase and the number of pulses used in the simulation is  $N = 2^2 = 4096$ . The magnitude plot produced by applying our approach is shown in Figure 3.2. Figure 3.3 highlights the output in the frequency range 1–6 kHz. Here, it can be seen that the ten largest magnitudes in the spectrum correspond to the ten pulse trains. (As in the non-generic case, the spectrum contains a large term at 0 Hz which is ignored.)

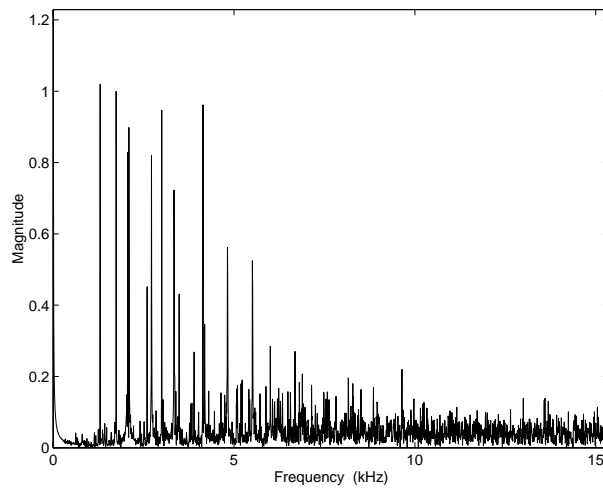


Figure 3.2: A generic magnitude plot ( $N = 4096$ ).

If the original signal consists of  $M$  interleaved pulse trains, the number of pulses processed,  $N$ , will approximately be equal to  $t_N(f_1 + \dots + f_M)$ . The number of pulse trains present is determined by assuming the  $m$  largest magnitudes correspond to pulse trains.

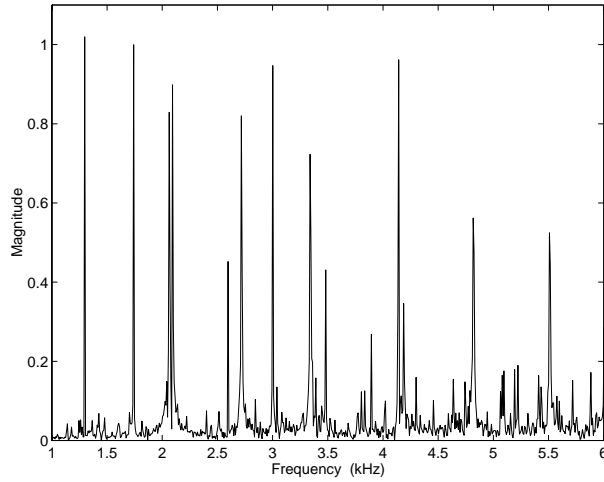


Figure 3.3: The region of interest from Figure 3.2.

Starting with  $m = 1$ ,  $m$  is incremented until  $t_N(\hat{f}_1 + \dots + \hat{f}_m)$  is approximately equal to  $N$ . (The frequencies  $\hat{f}_1, \dots, \hat{f}_m$  are estimates of  $f_1, \dots, f_m$ .) If no such  $m$  can be found it means the  $M$  largest magnitudes in the spectrum do not correspond to the  $M$  pulse trains present.

For our example, actual versus estimated pulse train frequencies are presented in Table 3.1.

PT No.	Actual Freq. (kHz)	Estimated Freq. (kHz)
1	1.2980	1.2981
2	1.7400	1.7409
3	2.0658	2.0635
4	2.0944	2.0935
5	2.7183	2.7164
6	3.0000	3.0015
7	3.3416	3.3392
8	4.1416	4.1421
9	4.8200	4.8174
10	5.5100	5.5077

Table 3.1: Actual vs. estimated pulse train frequencies.

### 3.3.1 Additional Processing

As Figure 3.3 demonstrates, the proposed scheme can produce a magnitude plot that contains magnitudes of a substantial size that do not correspond to pulse trains. Instead of trying to determine the interleaved pulse train spectrum from such a magnitude plot, the magnitude signal can be further processed in such a way that many of the magnitudes that do not correspond to pulse trains can be removed. This processing removes many of the artifacts present and leads to a more reliable estimate of the interleaved pulse train spectrum. We now discuss how this additional processing is done.

As mentioned previously, at least for the non-generic case, if  $f$  is a pulse train frequency, Theorem 3.3 predicts the existence of harmonics at  $2f, 3f, \dots$ . Simulations indicate that such harmonics also appear in the generic case and that in fact they make up the majority of spurious pulse train magnitudes. Simulations also indicate that the largest value in a magnitude plot always corresponds to a pulse train. Additional processing starts by assuming the largest magnitude present corresponds to a pulse train. The estimated frequency of this pulse train,  $\hat{f}_1$ , is taken to be the frequency corresponding to this magnitude. Any magnitudes at  $2\hat{f}_1, 3\hat{f}_1, \dots$  are assumed to be harmonics of this pulse train and are removed from the magnitude plot. (In practice, for reasons of robustness, one or two frequency bins directly either side of the DFT bin corresponding to each harmonic are also removed.) The sum of the magnitudes of these harmonics are then added to the magnitude at frequency  $\hat{f}_1$ . This process is then repeated on the second largest magnitude present and then on the next largest magnitude and so on until  $t_N(\hat{f}_1 + \dots + \hat{f}_m)$  approximately equals  $N$ . (In practice, when deciding which is the next largest magnitude, not only are the bins corresponding to previously identified pulse trains ignored but the two bins either side of such bins are also ignored. This helps to ensure that if a pulse train magnitude is spread over more than one bin, the pulse train is not incorrectly identified as a multiple pulse train.) The result of such additional processing for our example is shown in Figure 3.4.

Importantly, note that the additional processing discussed in this subsection incurs negligible additional computational cost.

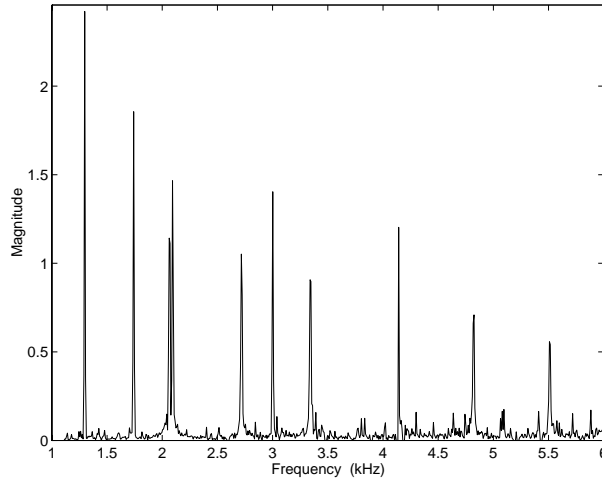


Figure 3.4: Magnitude plot after additional processing.

### 3.3.2 *When all else fails....*

If the proposed scheme (with additional processing) fails to properly identify the interleaved pulse train spectrum, for example, if no  $m$  can be found such that  $t_N(\hat{f}_1 + \dots + \hat{f}_m)$  is approximately equal to  $N$ , the spectrum can be identified in the following manner. As previously mentioned, simulations indicate that the largest magnitude in a magnitude plot always corresponds to a pulse train. Having identified the frequency corresponding to the largest magnitude, standard methods (Mardia 1989) can be used to deinterleave the corresponding pulse train. By deinterleave we mean that all pulses in the received interleaved signal that are members of the identified pulse train can be removed. This produces a new interleaved signal with one less pulse train present than the original. The proposed scheme can then be applied to this new signal and another pulse train can be identified and deinterleaved. This process can be repeated until all pulse trains are identified.

The method described in this subsection involves considerably more computational effort than the approach described earlier. As a consequence it should only be used as a last resort. No results presented here are based on such processing.

### 3.4 Further Analysis

In this section we continue to look at the generic case and consider how the length of the data set, pulse time of arrival noise, and missing pulses effect results. All pulse train data used in this section is based on the interleaved pulse train signal used in §3.3.

#### 3.4.1 Decreasing $N$

In this subsection we consider the effect of decreasing  $N$ , that is, of using a smaller number of pulses.

The simulation presented in §3.3 used  $N = 2^{12} = 4096$  pulses (see Figure 3.3). The output produced by using a smaller number of pulses,  $N = 2^{10} = 1024$ , is shown in Figure 3.5. This figure is a plot of the results without additional processing. As would be expected, using a smaller number of pulses leads to a loss in resolution. Overall the results are still very good though pulse trains 3 and 4, which are very close in frequency (2.0658 kHz and 2.0944 kHz respectively), have not been distinguished and have been incorrectly identified as a single pulse train.

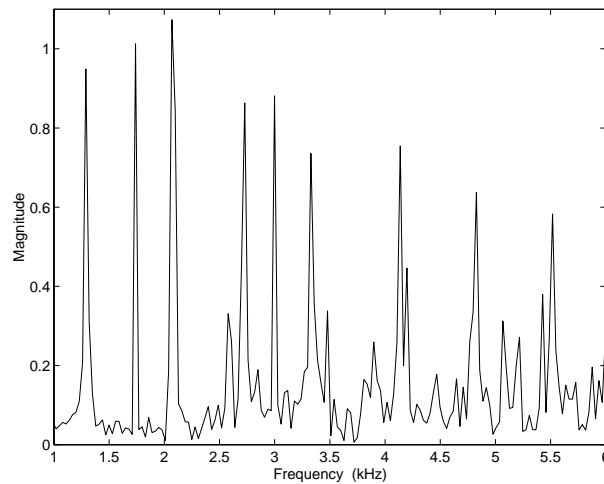


Figure 3.5: A magnitude plot using a decreased number of pulses:  $N = 1024$ .

The effect of further reducing  $N$  is demonstrated in Figure 3.6 in which  $N = 2^8 = 256$ .

As can be seen resolution has been greatly reduced and in practice a larger value of  $N$  would be required if accurate estimates of the pulse train frequencies were required. Notice however that despite the poor resolution, 9 prominent spikes are present, representing 9 of the 10 pulse trains present.

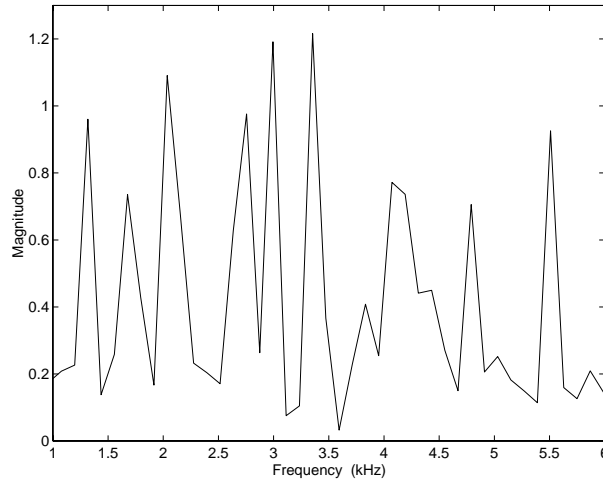


Figure 3.6: A magnitude plot with  $N = 256$ .

Overall, Figures 3.3, 3.5 and 3.6 demonstrate that performance degrades gracefully as  $N$  is decreased.

### 3.4.2 *Noisy Time of Arrival Data*

In this subsection we consider the effect of noisy time of arrival data. Time of arrival noise is modelled as zero mean Gaussian noise and, as in §3.3, we use  $N = 4096$ .

The output produced using our approach for a noise standard deviation of 0.025 seconds is shown in Figure 3.7. Figure 3.8 shows the results after additional processing. All 10 pulse trains are correctly identified and the estimated pulse train frequencies produced are given in Table 3.2. Table 3.2 also lists actual pulse train frequencies for convenience of comparison. Observe that for this level of noise, results are just as good as in the noise free case given in §3.3.

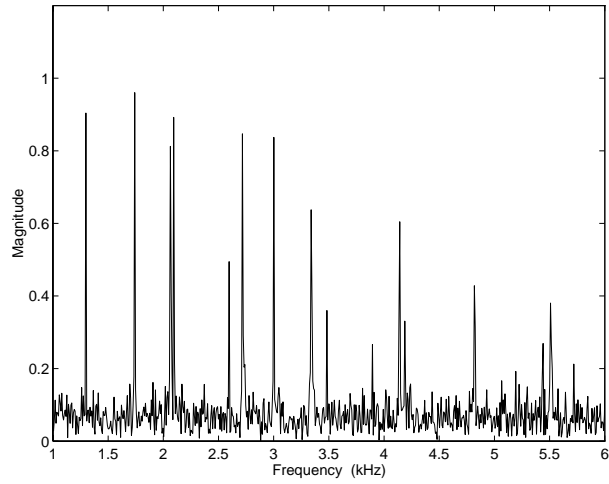


Figure 3.7: Magnitude plot for data with noisy times of arrival. Noise std. dev. = 0.025 seconds.

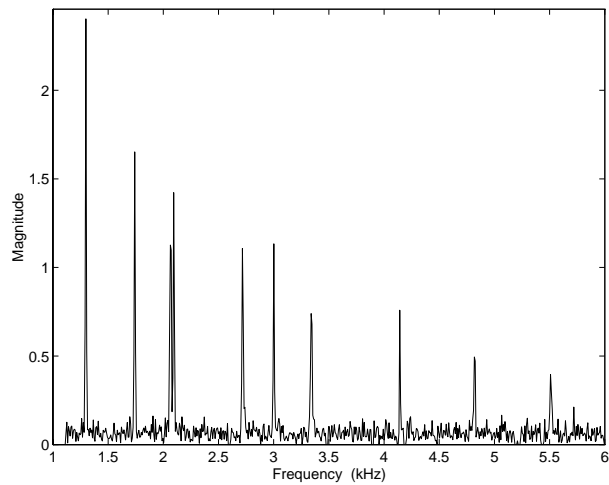


Figure 3.8: The magnitude plot given in Figure 3.7 after additional processing.

PT No.	Actual Freq. (kHz)	Estimated Freq. (kHz)
1	1.2980	1.2983
2	1.7400	1.7411
3	2.0658	2.0638
4	2.0944	2.0938
5	2.7183	2.7167
6	3.0000	3.0019
7	3.3416	3.3396
8	4.1416	4.1426
9	4.8200	4.8180
10	5.5100	5.5085

Table 3.2: Actual vs. estimated pulse train frequencies: time of arrival noise std. dev. = 0.025 seconds.

Figure 3.9 shows the output produced for a noise standard deviation of 0.05 seconds. In this case, 9 of the 10 pulse train frequencies were correctly identified however 4 spurious frequencies were also identified. The frequency estimates in kHz were 1.2979, 1.4179, 1.5529, 1.7405, 2.0256, 2.0631, 2.0931, 2.7158, 2.8508, 3.0008, 3.3384, 4.1412 and 4.8238.

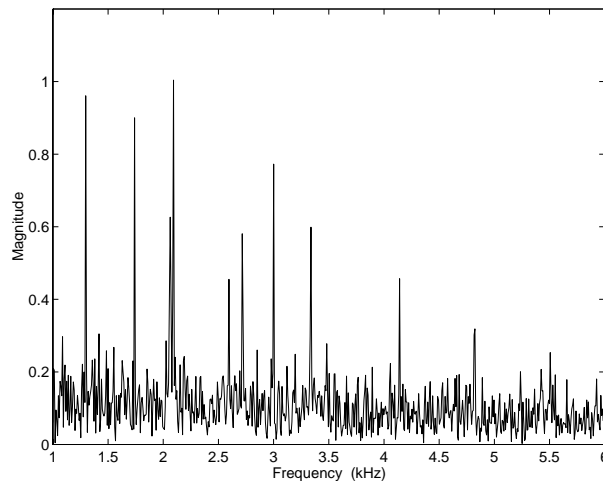


Figure 3.9: Magnitude plot for data with noisy times of arrival. Noise std. dev. = 0.05 seconds.

From the figures discussed above it can be seen that increased time of arrival noise leads

to greater noise in the pulse train magnitude plots. This increase in noise in turn leads to a decrease in performance. Note that, though it is not shown here, when processing noisy data, results can be improved by increasing  $N$ .

### 3.4.3 *Missing Pulses*

In this subsection we consider the effect of missing pulses. The signal processed was the same as the one used in §3.3 except that each pulse was given a probability of 1% of not being present. As before,  $N = 4096$ . Figure 3.10 is a plot of the output produced after additional processing. Though pulse trains 3 and 4 cannot visually be distinguished from Figure 3.10, our approach correctly identifies the 10 pulse trains present and the estimates of their frequencies are given in Table 3.3.

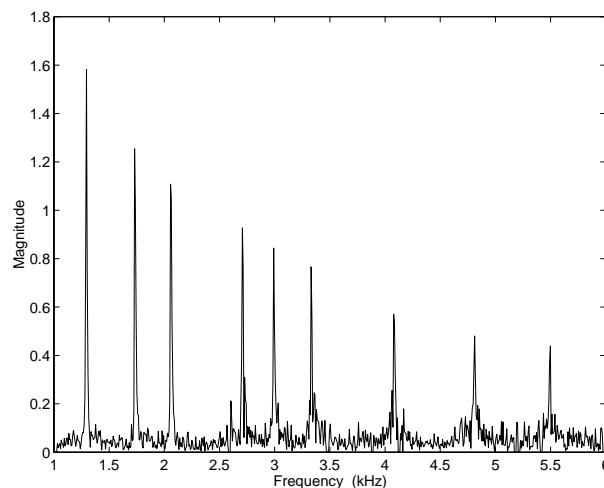


Figure 3.10: A magnitude plot of data with 1% of pulses missing after additional processing.

## 3.5 *Additional Comments and Concluding Remarks*

The most important property of the proposed scheme is that it is computationally efficient. Computations are of the order of  $N \log N$ . Other typical deinterleaving methods such as sequential search (Mardia 1989) and histogramming (Mardia 1989, Milojević & Popović

PT No.	Actual Freq. (kHz)	Estimated Freq. (kHz)
1	1.2980	1.3018
2	1.7400	1.7406
3	2.0658	2.0679
4	2.0944	2.0977
5	2.7183	2.7225
6	3.0000	3.0052
7	3.3416	3.3474
8	4.1416	4.1656
9	4.8200	4.8277
10	5.5100	5.5195

Table 3.3: Actual vs. estimated pulse train frequencies for data with 1% missing pulses.

1992) require order  $N^2$  computations (Perkins & Coat 1994).

The proposed methodology is also quite robust to noise. In §3.4 it was shown that the performance of the proposed scheme degrades gracefully as pulse time of arrival noise is introduced and increased and that it is robust to missing pulses.

Simulations also indicate that magnitudes corresponding to lower frequency pulse trains tend to be larger than the magnitudes of pulse trains with comparatively higher frequencies. In fact, if the ratio of largest to smallest pulse train frequencies present in an interleaved signal is too large, the spectrum magnitudes corresponding to the high frequency pulse trains become submerged in noise. How large this ratio can be is dependent on  $N$  and its size increases as  $N$  is increased. If the proposed method is having trouble detecting high frequency pulse trains, one option to try to improve detection would be to increase  $N$ . Note that increasing  $N$  also increases accuracy of frequency estimation.

Commonly used algorithms such as sequential search also suffer significant degradation of performance when the ratio of pulse train frequencies becomes too large. Since these existing methods identify high frequency pulse trains most effectively it is believed the proposed scheme could be used to compliment an existing algorithm for deinterleaving signals with pulse train frequency ratios exceeding these levels.

## ***References***

- Mardia, H. K. (1989). New techniques for the deinterleaving of repetitive sequences, *IEE Proceedings-F* **136**: 149–154.
- Milojević, D. J. & Popović, B. M. (1992). Improved algorithm for the deinterleaving of radar pulses, *IEE Proceedings-F* **139**: 98–104.
- Moore, J. B. & Krishnamurthy, V. (1994). Deinterleaving pulse trains using discrete-time stochastic dynamic-linear models, *IEEE Transactions on Signal Processing* **42**(11): 3092–3103.
- Perkins, J. & Coat, I. (1994). Pulse train deinterleaving via the Hough transform, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* **3**: 197–200.
- Wiley, R. G. (1982). *Electronic Intelligence: The Analysis of Radar Signals*, Artech House.

## *Chapter 4*

# *Equality Constrained Quadratic Optimization*

Constrained quadratic optimization problems form an important area of research and arise in many practical applications. Many of the problems that have been studied in this area fall into the category of linearly constrained, convex quadratic programming problems. A wealth of effective techniques are available for solving such problems, in particular we mention active set methods (Fletcher 1987) and various interior point methods (Faybusovich 1991, Tits & Zhou 1993, Nesterov & Nemirovskii 1994). The situation is not as tractable when one allows nonlinear equality constraints, see for example Thng, Cantoni & Leung (1996). Such constraints inherently lead to non-convex feasible sets which often consist of a number of disconnected components.

An important area of current research closely related to constrained quadratic optimization is that of semidefinite programming. Semidefinite programming is a convex optimization method that unifies a number of standard problems such as linear and quadratic programming and has a wide variety of applications from engineering to combinatorial optimization. Importantly, there exist many effective interior-point methods to solve semidefinite programming problems. These methods have polynomial worst-case complexity and perform well in practice (Vandenberghe & Boyd 1996).

In this chapter, we consider the problem of minimising a quadratic cost subject to purely quadratic equality constraints. Such problems are non-convex and their geometry is such that in many cases the resulting constraint set consists of the union of a number of disconnected subsets, each with their own local minima. To overcome the problem of multiple minima, we reformulate the problem in a novel manner. The reformulation involves the consideration of a sequence of linear optimization problems on the boundary of the positive definite matrices. Each of these problems is nested together in a manner that leads to a standard semidefinite programming problem on the interior of the positive definite matrices. The approach taken leads to the formulation of a gradient descent flow which can be used (in theory at least) to solve semidefinite programming problems. Though our reformulation of the initial problem as a semidefinite programming problem does not in general lead directly to a solution of the initial problem, the initial problem is solved by using a modified flow incorporating a penalty function.

Our aims in this chapter are twofold. We present both a method for minimising a quadratic cost subject to quadratic equality constraints and we provide an analysis of semidefinite programming from the non-standard though very interesting viewpoint of dynamical systems. Though it is unlikely that the gradient flow developed will provide a practical approach to solving semidefinite programming problems, the analysis undertaken provides an interesting new perspective into the geometry of such problems.

The chapter is structured as follows. In §4.1, a quadratic optimization problem subject to pure quadratic equality constraints is introduced and then related through a number of steps into a semidefinite programming problem. In §4.2, the geometry of this problem is analyzed. A gradient flow to solve semidefinite programming problems is developed and analyzed in §4.3; §4.4 contains some further analysis. In §4.5, a modified version of the gradient flow incorporating a penalty function is introduced. In §4.6, various methods of solving the original quadratic optimization problem based on the flow of §4.5 are discussed and a simulation example for one of the methods is presented. The chapter ends with some concluding remarks.

## 4.1 Problem Formulation

In this section a quadratic optimization problem subject to purely quadratic equality constraints is presented. This problem is then reformulated as a sequence of linear optimization problems on the boundary of the positive definite matrices. Each of these problems is nested together in a manner that leads to a standard semidefinite programming problem.

Consider the quadratic optimization problem:

**Problem 4.1** Given  $A_0, A_1, \dots, A_m \in \mathbb{R}^{n \times n}$  and  $c_1, \dots, c_m \in \mathbb{R}$ ,

$$\begin{aligned} \text{minimize} \quad & \phi(x) := x^T A_0 x \\ \text{subject to} \quad & x \in \mathbb{R}^n, \end{aligned} \tag{4.1}$$

$$\psi_i(x) := x^T A_i x = c_i, \quad i = 1, \dots, m. \tag{4.2}$$

□

The feasible set, those points which satisfy the constraints (4.1, 4.2), will certainly not be convex, and in general will have a number of separate connected components. Indeed, without considerably more knowledge of the matrices  $A_1, \dots, A_m \in \mathbb{R}^{n \times n}$  and the scalars  $c_1, \dots, c_m \in \mathbb{R}$  it is unclear whether the feasible set is non-empty. To avoid dealing with a null problem of this form the following assumptions are made:

### Assumption 4.2

- i) The matrices  $A_0, A_1, \dots, A_m$  are symmetric.
- ii) The set of points satisfying the constraints (4.1, 4.2) is non-empty.
- iii) The matrices  $A_1, \dots, A_m$  are linearly independent.

The first of these assumptions can be made without loss of generality due to the symmetry of the functions  $\phi$  and  $\psi_1, \dots, \psi_m$ . The second assumption is for convenience while the third assumption ensures that the constraints (4.2) are non-redundant.

**Remark 4.3** It is important not to implicitly require the structure of the feasible set to be known prior to the solution of Problem 4.1 being undertaken. Computing the set of feasible points is itself a difficult and time consuming task.  $\square$

The approach taken is to reformulate Problem 4.1 as a matrix optimization problem on the boundary of the positive definite matrices. Let  $\text{tr}$  denote the trace operator. Then, given any matrix  $A \in \mathbb{R}^{n \times n}$  and any vector  $x \in \mathbb{R}^n$ , one has

$$x^T A x = \text{tr}(x^T A x) = \text{tr}(A x x^T) = \text{tr}(A X)$$

where  $X := x x^T$ . The set of real  $n \times n$  matrices that can be written in the form  $X = x x^T$ ,  $x \neq 0$ , is the set of symmetric, positive semidefinite matrices of rank 1. Let

$$S(1, n) = \{X \in \mathbb{R}^{n \times n} \mid X^T = X \geq 0, \text{rank}(X) = 1\}$$

denote this set of matrices. Consider the set

$$\mathcal{M}_1 = \{X \in S(1, n) \mid \Psi_i(X) := \text{tr}(A_i X) = c_i, i = 1, \dots, m\}.$$

$\mathcal{M}_1$  is the set of all rank 1 matrices of the form  $x x^T$  where  $x$  is a feasible point for Problem 4.1. This leads to the following optimization problem:

**Problem 4.4** Given  $A_0, A_1, \dots, A_m \in \mathbb{R}^{n \times n}$  and  $c_1, \dots, c_m \in \mathbb{R}$  satisfying Assumption 4.2,

$$\begin{aligned} & \text{minimize} && \Phi(X) := \text{tr}(A_0 X) \\ & \text{subject to} && X \in \mathcal{M}_1. \end{aligned}$$

$\square$

Observe that in the new formulation both the cost and the explicit constraint functions,  $\Psi_i(X)$ , are linear in  $X$ . The nonlinearity in the problem is confined to the geometry of the set  $S(1, n)$ . Much is known about the geometry of  $S(1, n)$ . In particular,  $S(1, n)$  can be

thought of as a homogeneous orbit of the general linear group under congruence transformation (Helmke & Moore 1994). The addition of linear constraints in the definition of  $\mathcal{M}$  will generally divide the set into a number of separate connected components. However, the reformulation allows one to consider the generalized sets

$$S(r, n) = \{X \in \mathbb{R}^{n \times n} \mid X^T = X \geq 0, \text{rank}(X) = r\}, \quad (4.3)$$

and

$$\mathcal{M}_r = \{X \in S(r, n) \mid \text{tr}(A_i X) = c_i, i = 1, \dots, m\}. \quad (4.4)$$

This leads directly to the nested set of optimization problems:

**Problem 4.5** Given  $A_0, A_1, \dots, A_m \in \mathbb{R}^{n \times n}$  and  $c_1, \dots, c_m \in \mathbb{R}$  satisfying Assumption 4.2 and  $r$  some integer  $1 \leq r \leq n$ ,

$$\begin{aligned} & \text{minimize} && \Phi(X) \\ & \text{subject to} && X \in \mathcal{M}_r. \end{aligned}$$

□

In fact, each  $\mathcal{M}_r$  suffers from the same difficulty as  $\mathcal{M}_1$  with potentially several connected components. As the number  $r$  is increased the number of potential separate components reduces until  $r = n$ . In this final case then it is easily seen that  $\mathcal{M}_n$  is simply the intersection of a set of affine constraints with the convex cone of positive definite matrices. Thus,  $\mathcal{M}_n$  consists of only a single connected component and by solving Problem 4.5 for  $r = n$  one avoids the complication of local minima due to the geometry of the constraint sets. Unfortunately,  $\mathcal{M}_n$  is not a closed set and hence the problem could be ill posed.

To avoid this problem we consider the set

$$\mathcal{M} := \bigcup_{r=1, \dots, n} \overline{\mathcal{M}_r}. \quad (4.5)$$

comprising the topological closure of the union of all the sets  $\mathcal{M}_r$ . Then  $\mathcal{M}$  is a closed subset of  $\mathbb{R}^{n \times n}$  with a single connected component. This leads to the well posed optimization problem:

**Problem 4.6** Given  $A_0, A_1, \dots, A_m \in \mathbb{R}^{n \times n}$  and  $c_1, \dots, c_m \in \mathbb{R}$  satisfying Assumption 4.2,

$$\begin{aligned} & \text{minimize} && \Phi(X) \\ & \text{subject to} && X \in \mathcal{M}. \end{aligned}$$

□

Problem 4.6 is a standard semidefinite programming problem (Alizadeh 1995). There exist many practical interior point methods to solve such problems, see for example the review paper Vandenberghe & Boyd (1996). Unfortunately, as will be discussed later, the minimizing solution of Problem 4.6 is not always rank 1 and hence solving Problem 4.6 does not directly solve Problem 4.1. Nevertheless, the solution of Problem 4.6 should lie close to the desired solution of Problem 4.1. In the next sections, we introduce a flow to solve Problem 4.6. Problem 4.1 is then solved using a modified version of this flow incorporating a penalty function designed to penalize solutions of rank greater than 1.

## 4.2 *The Geometry of the Feasible Sets*

In this section, the geometry of the sets  $\mathcal{M}_r$  is investigated. It is shown that, excluding a set of singular points of zero measure, each set  $\mathcal{M}_r$  is a Riemannian manifold. Background material on differential geometry, Lie groups and related material used in this chapter can be found in Boothby (1986) and Helmke & Moore (1994). The homogeneous space structure of  $S(r, n)$  is also discussed in Chapter 5 of Helmke & Moore (1994).

An advantage of dealing with semi-algebraic Lie groups and group actions (such as the general linear group and its group action on  $S(r, n)$ ) is that the linearization of the group

action can be used to provide an explicit algebraic representation of the geometric properties of the homogeneous spaces considered. Following the notation presented in Helmke & Moore (1994, Ch 5), denote the symmetric bracket of two matrices  $A, B \in \mathbb{R}^{n \times n}$  by

$$\{A, B\} := AB + B^T A^T.$$

**Theorem 4.7** *The set*

$$S(r, n) = \{X \in \mathbb{R}^{n \times n} \mid X^T = X \geq 0, \text{rank}(X) = r\}$$

(as previously defined, see (4.3)) is a smooth manifold whose tangent space at an element  $X \in S(r, n)$  is the vector space

$$T_X S(r, n) = \{\{\Delta, X\} \mid \Delta \in \mathbb{R}^{n \times n}\}.$$

◇

**PROOF.** Consider the map

$$\alpha : GL(n, \mathbb{R}) \times \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n \times n}, \quad \alpha(Z, X) = ZXZ^T.$$

It is straight forward to show that

(i) If  $I$  is the identity matrix of  $GL(n, \mathbb{R})$ , then

$$\alpha(I, X) = X \quad \text{for all } X \in \mathbb{R}^{n \times n}.$$

(ii) If  $A, B \in GL(n, \mathbb{R})$ , then

$$\alpha(A, \alpha(B, X)) = \alpha(AB, X) \quad \text{for all } X \in \mathbb{R}^{n \times n}.$$

Hence  $\alpha$  is a left group action. Let  $I_r$  denote the  $r \times r$  identity matrix and let  $E_r$  be the

$n \times n$  block matrix  $E_r = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}$ . Then,

$$S(r, n) = \{ZE_rZ^T \mid Z \in GL(n, \mathbb{R})\}$$

and hence  $S(r, n)$  is an orbit of the Lie group action  $\alpha$ . As this group action is semialgebraic, it follows that  $S(r, n)$  is a smooth submanifold of  $\mathbb{R}^{n \times n}$  (Gibson 1979, pg. 224).

For any  $X \in S(r, n)$ , consider the map

$$\alpha_X : GL(n, \mathbb{R}) \rightarrow S(r, n), \quad Z \mapsto \alpha(Z, X) = ZXZ^T.$$

As  $\alpha$  is a smooth action of a Lie group on a smooth manifold and the orbit  $S(r, n)$  is a smooth submanifold, it follows that the map  $\alpha_X$  is a submersion (Gibson 1979, pg. 74). Hence,  $D\alpha_X|_I$ , the differential of  $\alpha_X$  evaluated at  $Z = I$ , is a linear map from  $T_I GL(n, \mathbb{R})$  onto  $T_X S(r, n)$  and

$$D\alpha_X|_I(\Delta) = \Delta X + X\Delta^T.$$

Noting that  $T_I GL(n, \mathbb{R}) = \mathbb{R}^{n \times n}$  and that  $X$  is arbitrary completes the proof.  $\blacksquare$

Consider the map

$$F : S(r, n) \rightarrow \mathbb{R}^m, \quad X \mapsto (\text{tr}(A_1 X) \dots \text{tr}(A_m X))^T.$$

The set  $\mathcal{M}_r$  (cf. equation (4.4)) is a fiber of this map given by  $\mathcal{M}_r = F^{-1}(c_1, \dots, c_m)$ .

The derivative of  $F$  in direction  $\{\Delta, X\} \in T_X S(r, n)$  is<sup>1</sup>

---

<sup>1</sup>Let  $A$  and  $B$  be real  $n \times n$  matrices. The *vec* of the matrix  $A \in \mathbb{R}^{n \times n}$  is the  $n^2$  length column vector  $\text{vec}(A) := [A(:, 1); \dots; A(:, n)]$ . It is easily verified that

$$\text{tr}(AB) = (\text{vec}(A^T))^T \text{vec}(B).$$

Let  $A_{ij}$  denote the  $ij$ 'th entry of the matrix  $A$ . The *Kronecker product* of the matrices  $A$  and  $B$  is defined by

$$A \otimes B = \begin{pmatrix} A_{11}B & \dots & A_{1n}B \\ \vdots & & \vdots \\ A_{n1}B & \dots & A_{nn}B \end{pmatrix} \in \mathbb{R}^{n^2 \times n^2}.$$

Some readily verified identities involving the *vec* operation and the Kronecker product are (Helmke & Moore

$$\begin{aligned}
DF|_X(\{\Delta, X\}) &= (\text{tr}(A_1\{\Delta, X\}) \dots \text{tr}(A_m\{\Delta, X\}))^T \\
&= 2(\text{tr}(A_1\Delta X) \dots \text{tr}(A_m\Delta X))^T \\
&= 2(\text{vec}(A_1) \dots \text{vec}(A_m))^T (X \otimes I) \text{vec}(\Delta).
\end{aligned}$$

The Fiber Theorem (Helmke & Moore 1994, pg. 346) implies that  $\mathcal{M}_r = F^{-1}(c_1, \dots, c_m)$  is a smooth submanifold of  $S(r, n)$  if the derivative of  $F$  is full rank at every point in the fiber. That is, if

$$(\text{vec}(A_1) \dots \text{vec}(A_m))^T (X \otimes I) \quad (4.6)$$

is full rank for all  $X \in \mathcal{M}_r$ . In addition, at every point  $X$  where the derivative of  $F$  is full rank,  $\mathcal{M}_r$  is locally a manifold and the tangent space of  $\mathcal{M}_r$  at such points is

$$\begin{aligned}
T_X \mathcal{M}_r &= \ker DF|_X \\
&= \{\{\Delta, X\} \mid \Delta \in \mathbb{R}^{n \times n}, \text{tr}(A_i\{\Delta, X\}) = 0, i = 1, \dots, m\}.
\end{aligned}$$

**Definition 4.8** Any point  $X \in \mathcal{M}_r$  for which (4.6) is not full rank, is termed a *singular point*. □

Unfortunately in practice, it is difficult to know in advance when singular points may arise. It follows from Sard's Theorem (Hirsch 1976, pg. 69) that the set of points  $(c_1, \dots, c_m) \in \mathbb{R}^m$  for which  $\mathcal{M}_r$  is not a manifold has measure zero in  $\mathbb{R}^m$ . Consequently, for an arbitrary choice of matrices  $A_1, \dots, A_m$  and scalars  $c_1, \dots, c_m$ , it is unlikely that  $\mathcal{M}_r$  will contain singularities.

The following is an example of the geometry of a set  $\mathcal{M}_r$  that contains singular points. For  $r = 1$ ,  $n = 3$  and  $m = 2$  constraints, let  $\mathcal{M}_1$  be defined by  $A_1 = I$ ,  $A_2 = \text{diag}(1, \frac{1}{4}, 4)$

---

1994, pg. 314),

$$\text{vec}(AB) = (I \otimes A) \text{vec}(B) = (B^T \otimes I) \text{vec}(A)$$

and

$$(A \otimes B)^T = (A^T \otimes B^T).$$

and  $c_1 = c_2 = 1$ . It is easily verified that the matrix  $(\text{vec}(A_1) \text{vec}(A_2))^T (X \otimes I)$  is rank degenerate at  $X = \text{diag}(1, 0, 0) \in \mathcal{M}_1$ . In local coordinates any  $X \in \mathcal{M}_1$  can be represented by  $x \in \mathbb{R}^n$  satisfying  $X = xx^T$ . The fact that the set is not a manifold is clearly demonstrated in Figure 4.1 which is a plot of the constraint set in local coordinates. The plot shows that the points  $(\pm 1, 0, 0)$  are degenerate and hence that the constraint set does not form a manifold.

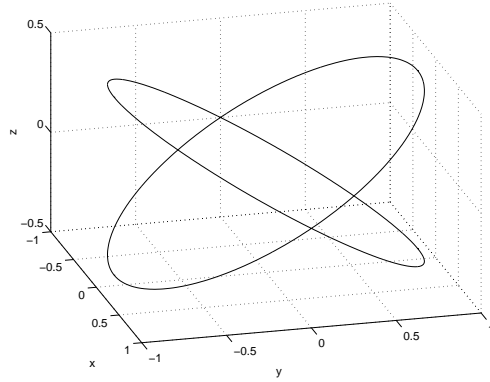


Figure 4.1: A constraint set that is not a manifold.

Before proceeding we make the following definition,

**Definition 4.9**

$$W := (\text{vec}(A_1) \dots \text{vec}(A_m)) \in \mathbb{R}^{n^2 \times m}.$$

□

Assumption 4.2 implies  $W$  has rank  $m$ . Using the definition of  $W$ , a singular point now becomes a point  $X \in \mathcal{M}_r$  for which  $W^T(X \otimes I)$  is not full rank.

**Lemma 4.10** *Given  $A_0, A_1, \dots, A_m \in \mathbb{R}^{n \times n}$  and  $c_1, \dots, c_m \in \mathbb{R}$  satisfying Assumption 4.2, and  $W$  as in Definition 4.9, for each  $r = 1, \dots, n$ , the set*

$$\mathcal{N}_r = \{X \in \mathcal{M}_r \mid W^T(X \otimes I) \text{ is full rank}\}$$

is a smooth submanifold of  $S(r, n)$  that differs from the set  $\mathcal{M}_r$  by at most a set of measure zero. The tangent space of  $\mathcal{N}_r$  at a point  $X$  can be represented by the vector space

$$T_X \mathcal{N}_r = \{ \{\Delta, X\} \mid \Delta \in \mathbb{R}^{n \times n}, \operatorname{tr}(A_i \{\Delta, X\}) = 0, i = 1, \dots, m \}.$$

◇

**PROOF.** The fact that  $\mathcal{N}_r$  differs from  $\mathcal{M}_r$  by at most a set of measure zero is a consequence of the fact that  $\mathcal{M}_r - \mathcal{N}_r$  is a proper algebraic subset, defined by

$$\det (W^T (X \otimes I)^2 W) = 0,$$

of the semi-algebraic set  $\mathcal{M}_r$ . The remainder of the theorem follows from discussion above using the Fiber Theorem (Helmke & Moore 1994, pg. 346). ■

The manifolds  $\mathcal{N}_r, r = 1, \dots, n$ , are submanifolds of the homogeneous spaces  $S(r, n)$ . It is possible to give  $S(r, n)$  a Riemannian structure derived from the normal metric on the general linear group (Helmke & Moore 1994, Ch 5). This metric is known as the normal metric on  $S(r, n)$ . A key property of the metric used is that the algebraic structure of the metric is related for each of the manifolds  $S(r, n), r = 1, \dots, n$ . The manifolds  $\mathcal{N}_r$  inherit this Riemannian structure as submanifolds of  $S(r, n)$ . The explicit form of the normal metric is given in the following discussion.

The proof of Theorem 4.7 indicates that the tangent space  $T_X S(r, n)$  can be considered as the image of the surjective linear map

$$D\alpha_X|_I : \mathbb{R}^{n \times n} \rightarrow T_X S(r, n), \quad \Delta \mapsto \{\Delta, X\}.$$

The kernel of  $D\alpha_X|_I$  is

$$K = \ker D\alpha_X|_I = \{ \Delta \in \mathbb{R}^{n \times n} \mid \{\Delta, X\} = 0 \}.$$

With respect to the standard inner product on  $\mathbb{R}^{n \times n}$ ,

$$\langle A, B \rangle = \text{tr}(A^T B),$$

the orthogonal complement of  $\ker D\alpha_X|_I$  is

$$K^\perp = \{Z \in \mathbb{R}^{n \times n} \mid \text{tr}(Z^T \Delta) = 0 \ \forall \Delta \in K\}.$$

This leads to the following orthogonal decomposition of  $\mathbb{R}^{n \times n}$ ,

$$\mathbb{R}^{n \times n} = K \oplus K^\perp.$$

Hence, every element  $\Delta \in \mathbb{R}^{n \times n}$  has a unique decomposition

$$\Delta = \Delta_X + \Delta^X \tag{4.7}$$

where  $\Delta_X \in K$  and  $\Delta^X \in K^\perp$ .

The map  $D\alpha_X|_I$  is surjective with kernel  $K$  and hence induces an isomorphism of  $K^\perp \subset \mathbb{R}^{n \times n}$  onto  $T_X S(r, n)$ . Thus defining an inner product on  $T_X S(r, n)$  is equivalent to defining an inner product on  $K^\perp$ . For  $\{\Delta_1, X\}, \{\Delta_2, X\} \in T_X S(r, n)$ , set

$$\langle\langle \{\Delta_1, X\}, \{\Delta_2, X\} \rangle\rangle := 2\text{tr}((\Delta_1^X)^T \Delta_2^X) \tag{4.8}$$

where  $\Delta_1^X$  and  $\Delta_2^X$  are defined by (4.7). The factor of 2 is added purely for convenience. This defines a positive definite, inner product on  $T_X S(r, n)$ . Since all the constructions are algebraic it is easily verified that the construction depends smoothly on  $X$  and generates a Riemannian metric on  $S(r, n)$  (Helmke & Moore 1994). This metric is referred to as the normal Riemannian metric on  $S(r, n)$ .

Finally, as  $\mathcal{N}_r$  is a submanifold of  $S(r, n)$ , the restriction of this metric to  $\mathcal{N}_r$  is a Riemannian metric on  $\mathcal{N}_r$ .

### 4.3 Gradient Flow

In this section, Problem 4.5 is considered and a gradient descent flow of the cost  $\Phi$  on the smooth manifolds  $\mathcal{N}_r$  introduced. Existence and uniqueness of solutions of the flow are established along with some convergence properties.

**Theorem 4.11** *Given  $A_0, A_1, \dots, A_m \in \mathbb{R}^{n \times n}$  and  $c_1, \dots, c_m \in \mathbb{R}$  satisfying Assumption 4.2, let  $X \in \mathcal{N}_r$  for some  $r$ ,  $1 \leq r \leq n$ . Then there is a unique solution  $(d_1, \dots, d_m)^T$  to the linear equation*

$$\begin{pmatrix} \text{tr}(A_1 A_1 X X) & \cdots & \text{tr}(A_1 A_m X X) \\ \vdots & & \vdots \\ \text{tr}(A_m A_1 X X) & \cdots & \text{tr}(A_m A_m X X) \end{pmatrix} \begin{pmatrix} d_1 \\ \vdots \\ d_m \end{pmatrix} = - \begin{pmatrix} \text{tr}(A_1 A_0 X X) \\ \vdots \\ \text{tr}(A_m A_0 X X) \end{pmatrix} \quad (4.9)$$

and the gradient of  $\Phi(X) = \text{tr}(A_0 X)$  with respect to the normal Riemannian metric (4.8) is given by

$$\begin{aligned} \text{grad}\Phi(X) &:= \{A_0 X + d_1 A_1 X + \cdots + d_m A_m X, X\} \\ &= A_0 X X + X X A_0 + \sum_{i=1}^m d_i (A_i X X + X X A_i). \end{aligned} \quad (4.10)$$

◇

**PROOF.** For a unique solution to (4.9) to exist it is sufficient to show that

$$D(X) = \begin{pmatrix} \text{tr}(A_1 A_1 X X) & \cdots & \text{tr}(A_1 A_m X X) \\ \vdots & & \vdots \\ \text{tr}(A_m A_1 X X) & \cdots & \text{tr}(A_m A_m X X) \end{pmatrix}$$

is full rank. Observing that

$$\text{tr}(A_i A_j X X) = \text{tr}((A_i X)^T A_j X) = (\text{vec}(A_i X))^T \text{vec}(A_j X),$$

it can be verified that

$$D(X) = W^T(X \otimes I)(X \otimes I)W.$$

Recall that  $W^T(X \otimes I)$  is full rank for all  $X \in \mathcal{N}_r$  and hence  $D(X)$  is full rank.

The gradient of  $\Phi : \mathcal{N}_r \rightarrow \mathbb{R}$  with respect to the normal Riemannian metric is the unique vector field  $\text{grad}\Phi$  which satisfies the conditions

$$(i) \quad \text{grad}\Phi(X) \in T_X \mathcal{N}_r \text{ for all } X \in \mathcal{N}_r.$$

$$(ii) \quad D\Phi|_X(\{\Delta, X\}) = \langle \langle \text{grad}\Phi(X), \{\Delta, X\} \rangle \rangle \text{ for all } \{\Delta, X\} \in T_X \mathcal{N}_r.$$

The first of these conditions implies that for all  $X \in \mathcal{N}_r$ ,

$$\text{grad}\Phi(X) = \{\Omega, X\}$$

for some  $\Omega \in \mathbb{R}^{n \times n}$  which possibly depends on  $X$ . In addition  $\text{grad}\Phi(X)$  must also satisfy

$$\text{tr}(A_i \text{grad}\Phi(X)) = 0, \quad \text{for } i = 1, \dots, m. \quad (4.11)$$

Consider

$$\Omega = A_0 X + d_1 A_1 X + \dots + d_m A_m X \quad (4.12)$$

where  $d_1, \dots, d_m$  are given by (4.9). With  $\Omega$  defined by (4.12) it is straightforward to show that  $\text{grad}\Phi(X) = \{\Omega, X\}$  satisfies (4.11) and hence that  $\text{grad}\Phi(X) = \{\Omega, X\}$  satisfies condition (i).

The derivative of  $\Phi$  at  $X$  is

$$D\Phi|_X(\{\Delta, X\}) = \text{tr}(A_0 \{\Delta, X\}).$$

Condition (ii) requires

$$\begin{aligned}
\text{tr}(A_0\{\Delta, X\}) &= \langle\langle \text{grad}\Phi(X), \{\Delta, X\} \rangle\rangle \\
&= \langle\langle \{\Omega, X\}, \{\Delta, X\} \rangle\rangle \\
&= 2\text{tr}((\Omega^X)^T \Delta^X)
\end{aligned}$$

for all  $\{\Delta, X\} \in T_X \mathcal{N}_r$ .

We now show that  $\Omega = \Omega^X$ . Let  $\Lambda \in K$ . Then

$$\begin{aligned}
\text{tr}(\Omega^T \Lambda) &= \text{tr}((XA_0 + d_1XA_1 + \cdots + d_mXA_m)\Lambda) \\
&= \text{tr}(\Lambda XA_0 + d_1\Lambda XA_1 + \cdots + d_m\Lambda XA_m) \\
&= \frac{1}{2}\text{tr}((\Lambda X + X\Lambda^T)A_0 + d_1(\Lambda X + X\Lambda^T)A_1 + \cdots + d_m(\Lambda X + X\Lambda^T)A_m) \\
&= \frac{1}{2}\text{tr}(\{\Lambda, X\}A_0 + d_1\{\Lambda, X\}A_1 + \cdots + d_m\{\Lambda, X\}A_m) \\
&= 0 \quad \text{as } \Lambda \in K
\end{aligned}$$

and hence  $\Omega \in K^\perp$  and  $\Omega = \Omega^X$ . This implies  $\text{tr}((\Omega^X)^T \Delta^X) = \text{tr}(\Omega^T \Delta)$ . Finally we show that  $2\text{tr}(\Omega^T \Delta) = \text{tr}(A_0\{\Delta, X\})$  for all  $\{\Delta, X\} \in T_X \mathcal{N}_r$ . Let  $\{\Delta, X\} \in T_X \mathcal{N}_r$ .

Then

$$\begin{aligned}
2\text{tr}(\Omega^T \Delta) &= 2\text{tr}((XA_0 + d_1XA_1 + \cdots + d_mXA_m)\Delta) \\
&= \text{tr}(A_0(\Delta X + X\Delta^T) + d_1A_1(\Delta X + X\Delta^T) + \cdots + d_mA_m(\Delta X + X\Delta^T)) \\
&= \text{tr}(A_0\{\Delta, X\} + d_1A_1\{\Delta, X\} + \cdots + d_mA_m\{\Delta, X\}) \\
&= \text{tr}(A_0\{\Delta, X\}) \quad \text{as } \text{tr}(A_i\{\Delta, X\}) = 0 \text{ for } i = 1, \dots, m.
\end{aligned}$$

This completes the proof. ■

An important aspect of this construction is that the algebraic representation of the gradient, equation (4.10), as a function from  $\mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n \times n}$  is independent of the rank of  $X$ .

Thus, apart from at singular points, it is possible to consider the algebraic equation

$$\dot{X} = -\left\{A_0 X + \sum_{i=1}^M d_i A_i X, X\right\} \quad (4.13)$$

as a differential equation on  $\mathcal{M}$ , the closure of the union of all the sets  $\mathcal{M}_i$  (cf. (4.5)). There are several advantages in this interpretation of the problem. In particular, the fact that the flow will be defined on a closed (and in most cases of interest, compact) set. However, before proceeding with the analysis it is necessary to consider how to deal with singular points should they occur.

Observe that, whenever the matrices  $A_1, \dots, A_m$  satisfy the linear independence requirement of Assumption 4.2, if  $X \succ 0$  the matrix given by (4.6) is always full rank<sup>2</sup>. Thus, if a singular point does occur it will always occur on the boundary of the positive definite cone.

Suppose  $X$  is a non-singular point that approaches a singular point  $X_s$  via some continuous path. Consider the coefficients  $d_1, \dots, d_m$  defined by (4.9). These  $d_i$ 's are in fact simply projection coefficients that ensure  $\text{tr}(A_i \text{grad}\Phi(X)) = 0$ ,  $i = 1, \dots, m$ . Hence the functions  $d_1, \dots, d_m$  must be continuous. Now, even though the matrix  $D(X)$  becomes singular as  $X \rightarrow X_s$ , the gradient direction  $-\text{grad}\Phi(X)$  remains bounded and has a well defined limit. Consequently, along any solution  $X(t)$  of the gradient flow (4.13), where there exists a time  $t_s > 0$  such that  $X(t_s) = X_s$ , the continuous limit of the gradient

$$\text{grad}\Phi(X_s) = \lim_{t \rightarrow t_s} \text{grad}\Phi(X(t)) \quad (4.14)$$

exists. Since the solution up to this point is unique, then the extension of the gradient field in this manner is unique for a given initial condition. Of course if a different initial condition is chosen, the gradient extension may be different. Considering a single initial condition and applying classical existence and uniqueness theory of ordinary differential equations, it follows that a solution of (4.13), extended via (4.14), must continue to exist beyond time

---

<sup>2</sup>If  $A, B \in \mathbb{R}^{n \times n}$  and  $\lambda_1, \dots, \lambda_n$  and  $\mu_1, \dots, \mu_n$  are the eigenvalues of  $A$  and  $B$  respectively, the eigenvalues of  $A \otimes B$  are  $\lambda_i \mu_j$  for all  $i, j$ . Hence if  $X$  is invertible, so is  $X \otimes I$ .

$t_s$ . The convention of choosing the gradient extension as mentioned above ensures that the solution obtained in this way is unique and (due to continuity) will continue to satisfy the constraints that preserve the solution in  $\mathcal{M}_r$ . Since the cost is analytic, one has that the solution must pass through the singular surface instantaneously at time  $t$  and then continues evolving in  $\mathcal{N}_r$ .

Recall the definition of  $\mathcal{M}$  (cf. (4.5)). The set  $\mathcal{M}$  is a closed subset of the positive semi-definite matrices. Indeed, if in addition to Assumption 4.2 one requires at least one of the constraint matrices  $A_1, \dots, A_m$  to be positive definite then the set  $\mathcal{M}$  is compact. In the sequel, if one or more of the constraint matrices are positive definite, without loss of generality it is assumed that  $A_1$  is positive definite.

**Lemma 4.12** *Given  $A_0, A_1, \dots, A_m \in \mathbb{R}^{n \times n}$  and  $c_1, \dots, c_m \in \mathbb{R}$  satisfying Assumption 4.2, assume in addition that  $A_1$  is positive definite. Then the set  $\mathcal{M}$ , given by (4.5), is a compact, convex subset of the positive semi-definite matrices.  $\diamond$*

**PROOF.** Since  $A_1 > 0$ ,  $A_1$  can be written in spectral form as  $A_1 = \sum_{i=1}^n \lambda_i v_i v_i^T$  where the eigenvalues  $\lambda_1, \dots, \lambda_n > 0$  and eigenvectors  $v_1, \dots, v_n$  are orthonormal. Let  $X \in \mathcal{M}$ . Then similarly  $X$  can be written as  $X = \sum_{j=1}^n \mu_j x_j x_j^T$  where  $\mu_1, \dots, \mu_n \geq 0$  and  $x_1, \dots, x_n$  are orthonormal. This implies that  $\text{tr}(A_1 X) = \sum_{i=1}^n \sum_{j=1}^n \lambda_i \mu_j (v_i^T x_j)^2$ . Let  $\lambda := \min\{\lambda_1, \dots, \lambda_n\}$ . Then, for each  $j$ ,  $1 \leq j \leq n$ ,

$$\begin{aligned} \text{tr}(A_1 X) &\geq \sum_{i=1}^n \lambda_i \mu_j (v_i^T x_j)^2 \\ &\geq \lambda \mu_j \sum_{i=1}^n (v_i^T x_j)^2. \end{aligned} \tag{4.15}$$

As the  $v_i$ 's form an orthonormal basis for  $\mathbb{R}^n$  and  $x_j^T x_j = 1$ , equation (4.15) implies that  $\text{tr}(A_1 X) \geq \lambda \mu_j$  and hence that  $0 < \mu_j \leq c_1 / \lambda$ . Now  $\|X\|^2 = \sum_{j=1}^n \mu_j^2$  and hence

$$\|X\|^2 \leq n \left( \frac{c_1}{\min \text{eig}(A_1)} \right)^2$$

and  $\mathcal{M}$  is bounded. Moreover, due to the construction of  $\mathcal{M}$  it is a closed subset of the positive semi-definite matrices and compactness follows.

Convexity follows by observing that for any two points  $X_1, X_2 \in \mathcal{M}$ ,  $sX_1 + (1-s)X_2$  is positive semi-definite and

$$\text{tr}(A_i(sX_1 + (1-s)X_2)) = sc_i + (1-s)c_i = c_i, \quad i = 1, \dots, m,$$

for  $0 \leq s \leq 1$ . Thus,  $sX_1 + (1-s)X_2$  is an element of  $\mathcal{M}$  and  $\mathcal{M}$  is convex.  $\blacksquare$

**Theorem 4.13** *Given  $A_0, A_1, \dots, A_m \in \mathbb{R}^{n \times n}$  and  $c_1, \dots, c_m \in \mathbb{R}$  satisfying Assumption 4.2, let  $X(0) = X_0 \in \mathcal{M}$  be any non-singular point (cf. Definition 4.8). Assume that  $A_1$  is positive definite. Then the solution  $X(t)$  to*

$$\begin{aligned} \dot{X} &= -\text{grad}\Phi(X) \\ &= -A_0XX - XXA_0 - \sum_{i=1}^m d_i(A_iXX + XXA_i), \end{aligned} \tag{4.16}$$

extended to all points in  $\mathcal{M}$  via the extension (4.14) satisfies:

- i) *The solution  $X(t)$  exists and is unique for all time  $t \geq 0$  and remains in  $\mathcal{M}$ .*
- ii) *The rank of the solution  $\text{rank}(X(t))$  remains constant for all time.*
- iii) *The equilibria of (4.16) are characterised by those points  $X \in \mathcal{M}$  such that*

$$(A_0 + d_1A_1 + \dots + d_mA_m)X = 0. \tag{4.17}$$

- iv) *The cost  $\Phi(X(t))$  is a monotonically decreasing function of time. The solutions  $X(t)$  converge to a connected component of the set of equilibria given by (4.17).*

$\diamond$

**PROOF.** Existence and uniqueness of the solution of (4.16) over a small time around a point  $X(t)$  is guaranteed by classical ODE theory and the discussion of the extension of the flow onto singular points. (Note that singular initial conditions, for which the extension (4.14) is

unclear, are excluded from consideration.) Existence and uniqueness of the flow for all time is a consequence of the compactness of the set  $\mathcal{M}$  along with the local existence result.

The preservation of the rank of  $X(t)$  for all time is also a direct consequence of the local existence results and the fact that locally  $X(t)$  remains in  $\mathcal{M}$ .

To verify the characterization of the critical points consider a critical point  $X \in \mathcal{M}$ . Then,

$$\text{grad}\Phi(X) = \{\mathcal{A}X, X\} = 0 \quad (4.18)$$

where  $\mathcal{A} := A_0 + d_1A_1 + \dots + d_mA_m$ . Expanding equation (4.18) and multiplying by  $\mathcal{A}$  on the left implies

$$\mathcal{A}\mathcal{A}X X + \mathcal{A}X X\mathcal{A} = 0.$$

Taking the trace of the above equation and using properties of the trace operator implies that

$$\text{tr}((\mathcal{A}X)^T(\mathcal{A}X)) = 0$$

and hence that  $\mathcal{A}X = 0$ . Substituting for  $\mathcal{A}$  in this equation recovers the characterisation in the theorem statement. Sufficiency of the condition is easily verified.

Part iv) is a consequence of the gradient nature of the flow on all but a set of measure zero in  $\mathcal{M}$ . The convergence result follows by consider  $\Phi(X)$  as a Lyapunov function for the flow. ■

## 4.4 Further Analysis

In this section a phase portrait analysis of the flow (4.16) is developed.

The following result establishes that all the critical points of (4.16) occur on the boundary of the positive definite cone, except in the case where the cost is constant on  $\mathcal{M}$ .

**Lemma 4.14** *In the generic case that  $\Phi$  is not constant over  $\mathcal{M}$ , any critical point of Problem 4.6 is rank degenerate.  $\diamond$*

**PROOF.** The proof proceeds by contradiction. Assume  $X_{opt} > 0$  is a minimum of  $\Phi$  and let  $X_1 \in \mathcal{M}$  be a point that does not minimise the cost. Now the positive definite symmetric matrices are open in the symmetric matrices. Hence there exist  $t > 0$  such that  $X_2 := X_{opt} + t(X_{opt} - X_1)$  is positive definite. Note that  $X_2 \in \mathcal{M}$  and that

$$\begin{aligned}\Phi(X_2) &= \Phi(X_{opt}) + t(\Phi(X_{opt}) - \Phi(X_1)) \\ &< \Phi(X_{opt}) \quad \text{as } \Phi(X_{opt}) < \Phi(X_1) \text{ and } t > 0.\end{aligned}$$

This contradicts the fact that  $X_{opt}$  is a global minima and completes the proof.  $\blacksquare$

**Remark 4.15** The situation that the cost functional  $\Phi$  is constant over  $\mathcal{M}$  will occur if one has a degenerate situation such as  $A_0$  being a linear combination of the the constraint matrices  $A_1, \dots, A_m$ .  $\square$

In addition to this, the set of minima for Problem 4.6 form a single connected component in  $\mathcal{M}$ .

**Lemma 4.16** *Given  $A_0, A_1, \dots, A_m \in \mathbb{R}^{n \times n}$  and  $c_1, \dots, c_m \in \mathbb{R}$  satisfying Assumption 4.2, assume in addition that  $A_1$  is positive definite. All minima of the cost  $\Phi(X) = \text{tr}(A_0 X)$  on  $\mathcal{M}$  are global minima. Moreover, the set of all such minima form a convex subset of  $\mathcal{M}$ .  $\diamond$*

**PROOF.** Suppose  $X_{min}$  minimises  $\Phi$  in any neighbourhood  $N \subset \mathcal{M}$ . As  $\mathcal{M}$  is convex (cf. Lemma 4.12), for any  $X_1 \in \mathcal{M}$  there exists  $t, 0 < t < 1$ , such that  $X_2 := tX_{min} + (1 - t)X_1 \in N$ . Linearity of the cost implies  $\Phi(X_{min}) \leq \Phi(X_2) = t\Phi(X_{min}) + (1 - t)\Phi(X_1)$ . This implies that  $\Phi(X_{min}) \leq \Phi(X_1)$  and hence that  $X_{min}$  is a global minima.

Suppose  $X_1, X_2 \in \mathcal{M}$  are both global minima of  $\Phi$ . For any  $t, 0 \leq t \leq 1$ , consider  $X := tX_1 + (1 - t)X_2$ . Then  $X \geq 0$  and by linearity of the constraint equations satisfies

$\text{tr}(A_i X) = c_i, i = 1, \dots, m$ . Hence  $X \in \mathcal{M}$  and by linearity of the cost  $\Phi(X) = \Phi(X_1) = \Phi(X_2)$ . Hence,  $X$  is also a global minima of the cost and the set of all minima is convex. ■

**Remark 4.17** In fact, generically, there is an unique solution to Problem 4.6 (Alizadeh, Haeberly & Overton 1996). □

Classical dynamical systems theory now ensures that the attractive basin of the set of global minima is almost all of the set  $\mathcal{M}$ . Indeed, the authors believe that the attractive sets of the non-minimal critical points are confined to the boundary of the positive definite cone, however, we have no satisfactory proof for this claim.

It is of interest to study more carefully the characteristics of the critical points. Equation (4.17) provides one perspective on the critical points of (4.16) which is that the eigenvectors of  $X$  are elements of the nullspace of

$$A_0 + d_1 A_1 + \dots + d_m A_m.$$

In the case where there is a single constraint, the critical point condition becomes a standard generalized eigenvalue problem,

$$A_0 X = -d_1 A_1 X.$$

Furthermore, when  $A_1$  is positive definite, the one constraint case can be solved analytically. Suppose  $A_1 > 0$  and recall the original problem statement, Problem 4.1, with  $m = 1$  constraints: *minimize*  $x^T A_0 x$  *subject to*  $x^T A_1 x = c_1$ . As  $A_1 > 0$ , there exists an invertible, symmetric square root of  $A_1$ , denoted  $A_1^{\frac{1}{2}}$ . Define  $y := c_1^{-\frac{1}{2}} A_1^{\frac{1}{2}} x$  and  $A'_0 := c_1 A_1^{-\frac{1}{2}} A_0 A_1^{-\frac{1}{2}}$ . Then the optimization problem can be rewritten as *minimize*  $y^T A'_0 y$  *subject to*  $y^T y = 1$ . The solution to this problem is well known: it is the unit length eigenvector corresponding to the minimal eigenvalue of  $A'_0$ . But this is just the same as claiming that  $x$  is the solution corresponding to the smallest eigenvalue of the generalised eigenvalue problem for the matrix pair  $A_0, A_1$ .

Denoting the minimal generalized eigenvalue of  $A_0, A_1$  by  $\lambda$ , one has the following

conditions for  $X = xx^T$  to be a global minimum of Problem 4.6,

- i)  $(A_0 - \lambda A_1)x = 0,$
- ii)  $x^T A_1 x = c_1,$
- iii)  $A_0 - \lambda A_1 \geq 0.$

Equations i) and iii) are solved by choosing  $x$  as the eigenvector corresponding to the minimal generalised eigenvalue of  $A_0, A_1$ . Equation ii) can be satisfied by scaling  $x$ .

In fact, in the semidefinite programming literature it is shown that this characterisation of the global minima of the single constraint case generalises to conditions for a global minima in the multi-constraint case.

**Theorem 4.18**  $X \in \mathcal{M}$  is a optimal solution to Problem 4.6 if and only if

- i)  $(A_0 + d_1 A_1 + \dots + d_m A_m)X = 0,$
- ii)  $\text{tr}(A_i X) = c_i, i = 1, \dots, m,$
- iii)  $A_0 + d_1 A_1 + \dots + d_m A_m \geq 0.$

◇

**PROOF.** See Alizadeh et al. (1996). ■

The final issue that needs to be considered is the question of whether a minimum of Problem 4.6 on the full set  $\mathcal{M}$  relates to a minima of Problem 4.1, the original problem on  $\mathbb{R}^n$ . Unfortunately, a minimum of Problem 4.6 will not always be rank 1 and hence one is not always able to solve Problem 4.1 directly from the solution of Problem 4.6.

The next theorem gives upper and lower bounds on the rank of  $X$ . The result is taken from Alizadeh et al. (1996). Before proceeding we introduce some notation. Let

$$n^{\overline{2}} := n(n + 1)/2.$$

For  $h \geq 0$ , let  $\lfloor h \rfloor$  denote the largest integer less than or equal to  $h$ . Define

$$\sqrt[3]{k} = \lfloor h \rfloor \quad \text{where } h \text{ is the positive real root of } h^3 = k.$$

**Theorem 4.19** *If  $X$  is an minima of Problem 4.6, then generically,*

$$n - \sqrt[3]{n^2 - m} \leq \text{rank}(X) \leq \sqrt[3]{m}.$$

◇

**PROOF.** See Alizadeh et al. (1996). ■

**Corollary 4.20** *Generically, the solution to Problem 4.6 is rank 1 if  $m \leq 2$ .* ◇

**PROOF.** To guarantee that the solution of Problem 4.6 is rank 1 (generically) we require the upper bound given in Theorem 4.19 to be 1. That is, we require  $\sqrt[3]{m} = \lfloor h \rfloor = 1$ . This is equivalent to requiring that

$$h = \frac{-1 + \sqrt{1 + 8m}}{2} < 2.$$

This is true if and only if  $m \leq 2$ . ■

An example of the sorts of bounds produced by Theorem 4.19 are given in Table 4.1 (Alizadeh et al. 1996). Bounds are given for  $n = 20$  and various values of  $m$ .

m	Bounds on $r = \text{rank}(X)$
10	$1 \leq r \leq 4$
20	$1 \leq r \leq 5$
30	$2 \leq r \leq 7$
40	$3 \leq r \leq 8$
50	$3 \leq r \leq 9$

Table 4.1: Generic bounds on  $\text{rank}(X)$  for  $n = 20$ .

## 4.5 Gradient Flow with Penalty Function

In this section we consider a modified version of the flow that incorporates a penalty function designed to encourage the solution of the flow to converge to a rank one matrix.

Consider the cost function

$$\Omega(X) = \|X\|_F^2 - \|X\|_2^2, \quad (4.19)$$

where  $\|X\|_F = \text{tr}(X^2)^{\frac{1}{2}}$  is the Frobenius norm of  $X$  and  $\|X\|_2 = \max_{\|v\|=1} \|Xv\|$  is the 2-norm of  $X$ . If  $0 \leq \lambda_1 \leq \dots \leq \lambda_n$  are the eigenvalues of  $X$ , then  $\Omega(X) = \sum_{i=1}^{n-1} \lambda_i^2$  and can be thought of as a measure of how close the matrix  $X$  is to being rank 1. If  $\Omega(X) \geq 0$  is small then  $X$  is close to being rank 1 and is indeed rank 1 if and only if  $\Omega(X) = 0$ .

Consider the following optimization problem,

**Problem 4.21** Given  $A_0, A_1, \dots, A_m \in \mathbb{R}^{n \times n}$  and  $c_1, \dots, c_m \in \mathbb{R}$  satisfying Assumption 4.2, and  $\epsilon > 0$ ,

$$\begin{aligned} \text{minimize} \quad & \Theta(X) := \Phi(X) + \epsilon \log(\Omega(X)) \\ \text{subject to} \quad & X \in \mathcal{M}, \end{aligned}$$

where  $\Phi(X) = \text{tr}(A_0 X)$ , the cost of Problem 4.6, and  $\Omega(X)$  is defined by (4.19).  $\square$

The term  $\epsilon \log(\Omega(X))$  can be thought of as a penalty function that penalises solutions of rank larger than 1. Let  $X_{opt}$  denote the optimal solution of Problem 4.6 and  $X_\epsilon$  denote a solution of Problem 4.21 for a given  $\epsilon > 0$ . By varying  $\epsilon$  one can trade off how close  $\Phi(X_\epsilon)$  is to  $\Phi(X_{opt})$  versus how close  $X_\epsilon$  is to being rank 1. (Note that  $\Phi(X_{opt}) \leq \Phi(X_\epsilon)$  for all  $\epsilon > 0$ .)

In order to solve Problem 4.21 one may consider developing a gradient descent flow in the same way a gradient flow was developed to solve Problem 4.6. We now proceed to do

this. Consider again equation (4.19) which can be rewritten as

$$\Omega(X) = \text{tr}(X^2) - v^T X^2 v$$

where  $v = v(X)$  is the unit length eigenvector corresponding to the maximum eigenvalue of  $X$ . In the derivation that follows, we are required to take the derivative of  $v$  with respect to  $X$ . Though  $v$  will generally not be differentiable everywhere,  $v$  is differentiable almost everywhere and this is sufficient for our purposes. The derivative of  $\Omega(X)$  in direction  $\{\Delta, X\} \in T_X \mathcal{M}$  is

$$D\Theta|_X (\{\Delta, X\}) = \text{tr}(A_0 \{\Delta, X\}) + \epsilon \frac{D\Omega|_X (\{\Delta, X\})}{\Omega(X)}$$

where

$$\begin{aligned} D\Omega|_X (\{\Delta, X\}) &= 2\text{tr}(X \{\Delta, X\}) - 2v^T X \{\Delta, X\} v - 2v^T X^2 Dv|_X (\{\Delta, X\}) \\ &= 2\text{tr}(X \{\Delta, X\}) - 2v^T X \{\Delta, X\} v - 2\lambda_{\max}^2(X) v^T Dv|_X (\{\Delta, X\}). \end{aligned}$$

As  $v^T v = 1$ , it follows that  $v^T Dv|_X (\{\Delta, X\}) = 0$ . Hence

$$D\Omega|_X (\{\Delta, X\}) = 2\text{tr}((X - vv^T X) \{\Delta, X\})$$

and

$$D\Theta|_X (\{\Delta, X\}) = \text{tr} \left( \left( A_0 + \frac{2\epsilon(X - vv^T X)}{\Omega(X)} \right) \{\Delta, X\} \right).$$

This implies the gradient flow that solves Problem 4.21 is the same as one that solves Problem 4.6 if one replaces  $A_0$  with

$$A_0^\epsilon := A_0 + \frac{2\epsilon(X - vv^T X)}{\Omega(X)}.$$

Note that  $A_0^\epsilon$  is a function of  $X$ .

## 4.6 Solution Methods

If Problem 4.6 has a rank one optimal solution, a standard semidefinite programming algorithm can be used to efficiently find the optimum solution of Problem 4.6 and hence solve Problem 4.1. In this section we discuss how the modified flow developed in §4.5 can be used to solve Problem 4.1 in the case that Problem 4.6 does not have a rank one solution.

Before proceeding we consider the problem of initial conditions. Before a semidefinite program or a gradient flow can be used, a matrix  $X = X^T > 0$  satisfying  $\text{tr}(A_i X) = c_i$ ,  $i = 1, \dots, m$ , must be found. This is a standard problem in semidefinite programming and many methods of solving this problem exist, see for example Vandenberghe & Boyd (1996, pp. 86–88).

Let  $X_{opt}$  denote the optimum solution of Problem 4.6. One method of solving Problem 4.1 is to find  $X_{opt}$  using semidefinite programming techniques and then to use  $X_{opt}$  as an initial condition for the flow developed in §4.5. Starting from  $X_{opt}$  is an intuitively appealing idea and this method has been found to work well in practice. (Note that  $\epsilon$  can be chosen so that all but the dominate eigenvalue of  $X$  end up being sufficiently small and that the flow of §4.5 can be solved using a standard ODE solver.) If the limit of the flow of §4.5 is denoted by  $X_{rank\ 1}$ , a nice feature of this method is that, having calculated  $X_{opt}$ ,  $\Phi(X_{opt})$  gives us a lower bound on  $\Phi(X_{rank\ 1})$ .

**Simulation Example.** The following is a typical simulation example with  $n = 20$  and  $m = 10$ . We denote the initial feasible point by  $X_0$ , the optimal solution of Problem 4.6 by  $X_{opt}$ , and the limit of the flow by  $X_{rank\ 1}$ . In this particular simulation the optimal solution was found to be rank 2. The costs and eigenvalue spectrums of  $X_0$ ,  $X_{opt}$  and  $X_{rank\ 1}$  are displayed in Table 4.2 and Figure 4.2 respectively. It is interesting to note that the flow does not appear to change the eigenvalue spectrum of  $X_{opt}$  except to reduce those eigenvalues that were associated with additional rank.

The method described above is only one way of using the flow of §4.5. Another method would be the following. Starting with an initial feasible solution, one could start the flow with  $\epsilon \approx 0$ . Once the flow is close to converging for this value of  $\epsilon$  (convergence could be

$X$	Cost $\Phi(X)$
$X_0$	1.6000
$X_{opt}$	-3.8126
$X_{rank\ 1}$	-3.1908

Table 4.2: A cost comparison.

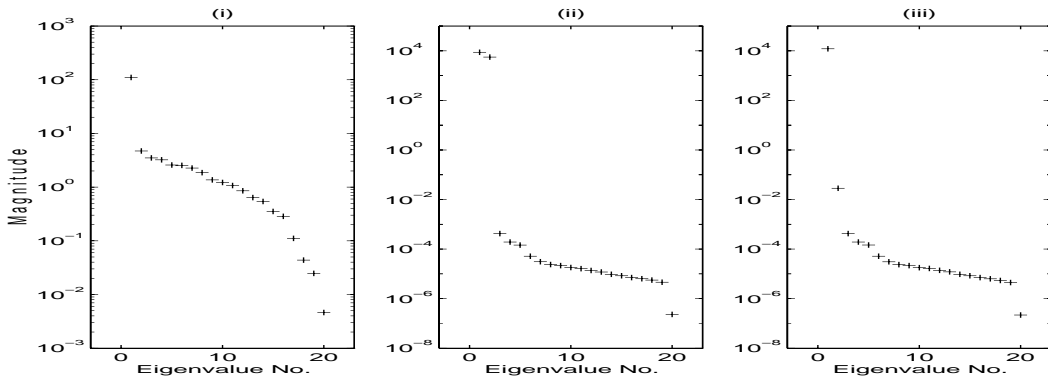


Figure 4.2: The eigenvalue spectrums of (i)  $X_0$ , (ii)  $X_{opt}$ , and (iii)  $X_{rank\ 1}$ .

monitored by checking how close the norm of the gradient of the flow is to zero),  $\epsilon$  could be increased by a certain small amount and the flow allowed to evolve further. Again, once the flow is close to converging,  $\epsilon$  could be increased further and the whole process repeated until a solution was obtained with the  $n - 1$  smallest eigenvalues of  $X$  sufficiently small. In this manner the penalty function can be thought of as a sort of pseudo barrier function. This method has not been tried in practice.

## 4.7 Conclusion

In this chapter we have provided a dynamical systems analysis of semidefinite programming. We have also developed methods of minimizing a quadratic cost subject to purely quadratic constraints based on a gradient descent flow incorporating a penalty function. One of these methods was simulated and found to work well in practice. Despite the encouraging results, at present it is not known whether the methods developed will always

find the optimal solution. Further analysis is required in this area.

## ***References***

Alizadeh, F. (1995). Interior point methods in semidefinite programming with application to combinatorial optimization, *SIAM Journal on Optimization* **5**: 13–51.

Alizadeh, F., Haeberly, J. P. & Overton, M. (1996). Complementarity and nondegeneracy in semidefinite programming, *Math. Programming* (Series B). To appear.

Boothby, W. (1986). *An Introduction to Differentiable Manifolds and Riemannian Geometry*, second edn, Academic Press, San Diego, USA.

Faybusovich, L. E. (1991). Dynamical systems which solve optimisation problems with linear constraints, *IMA Journal of Information and Control* **8**: 135–149.

Fletcher, R. (1987). *Practical Methods of Optimization*, second edn, Wiley, Chichester, UK.

Gibson, C. G. (1979). *Singular Points of Smooth Mappings*, Vol. 25 of *Research Notes in Mathematics*, Pitman, London, UK.

Helmke, U. & Moore, J. B. (1994). *Optimization and Dynamical Systems*, Communications and Control Engineering Series, Springer-Verlag, London, UK.

Hirsch, M. W. (1976). *Differential Topology*, number 33 in *Graduate Texts in Mathematics*, Springer-Verlag, New York, USA.

Nesterov, Y. & Nemirovskii, A. (1994). *Interior-Point Polynomial Algorithms in Convex Programming*, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, USA.

Thng, I., Cantoni, A. & Leung, Y. (1996). Analytical solutions to the optimization of a quadratic cost function subject to linear and quadratic equality constraints, *Applied Mathematics & Optimization* **34**(2): 161–182.

- Tits, A. L. & Zhou, J. L. (1993). A simple, quadratically convergent interior point algorithm for linear programming and convex quadratic programming, *in* W. W. Hager, D. W. Hearn & P. M. Pardalos (eds), *Large Scale Optimization: State of the Art*, Kluwer Academic Publishers B. V.
- Vandenberghe, L. & Boyd, S. (1996). Semidefinite programming, *SIAM Review* **38**(1): 49–95.



## *Chapter 5*

# *Conclusion*

In recent years, the re-analysis of traditional problems using modern mathematical techniques has in many cases lead to new, superior solutions to these problems. In numerical optimization, for example, the analysis of problems using ideas from such areas of mathematics as differential geometry and dynamical systems analysis, has lead to new insights that have in turn resulted in more computationally efficient and better behaved numerical algorithms.

In this thesis we have looked at three important problems in nonlinear systems. Each of these problems has been analyzed using what could perhaps be considered non-standard techniques. The analysis of each of these problems has lead to some interesting results with potential practical applications. The first of the topics covered used ideas from differential geometry and presented some internal stability results for a class of nonlinear control systems. In the second topic covered, we presented a novel nonlinear approach for solving the pulse train deinterleaving problem. The approach we developed was found to perform well in practice and was more computationally efficient than the traditional methods of pulse train deinterleaving such as histogramming. Lastly, the the final topic provided a dynamical systems analysis of semidefinite programming. The insight gained from this non-standard analysis allowed us to develop an innovative method of solving an equality constrained optimization problem.

In the rest of this chapter we give a fuller overview of the three topics discussed in this

thesis and for each of these topics present some areas for possible future research.

**Internal Stability Issues in Output Stabilization of a Class of Nonlinear Control Systems.** In Chapter 2 of this thesis we considered the issue of internal stability in output stabilization of a class of nonlinear control systems. It was shown for continuous time affine control systems, exhibiting no drift for zero output and having vector relative degree  $\{1, \dots, 1\}$ , that there exists an output stabilizing control and a neighbourhood of the zero output level set for which the closed loop system is internally stable.

**Future research.** It should not be too difficult to extend the results of Chapter 2 to nonlinear control systems with arbitrary vector relative degrees. Consider a nonlinear control system with vector relative degree  $\{\rho_1, \dots, \rho_m\}$ . Let  $L_X\phi$  denote the Lie derivative of a function  $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$  with respect to a vector field  $X : \mathbb{R}^n \rightarrow \mathbb{R}^n$  (Isidori 1995, pg. 496), and consider the equation

$$\begin{aligned} \begin{pmatrix} y_1^{\rho_1}(x) \\ \vdots \\ y_m^{\rho_m}(x) \end{pmatrix} &= \begin{pmatrix} L_f^{\rho_1} h_1(x) \\ \vdots \\ L_f^{\rho_m} h_m(x) \end{pmatrix} + \begin{pmatrix} L_{g_1} L_f^{\rho_1-1} h_1(x) & \cdots & L_{g_m} L_f^{\rho_1-1} h_1(x) \\ \vdots & & \vdots \\ L_{g_1} L_f^{\rho_m-1} h_m(x) & \cdots & L_{g_m} L_f^{\rho_m-1} h_m(x) \end{pmatrix} u \\ &:= H(x) + A(x)u. \end{aligned}$$

If the matrix  $A(x)$  is full rank for all  $x \in \mathbb{R}^n$ , a candidate output stabilizing control law is

$$u = A(x)^{-1} \left( -H(x) + \begin{pmatrix} P_1(y_1, \dot{y}_1, \dots, y_1^{(\rho_1-1)}) & \cdots & P_m(y_m, \dot{y}_m, \dots, y_m^{(\rho_m-1)}) \end{pmatrix}^T \right)$$

where, for  $i = 1, \dots, m$ ,  $P_i$  is a polynomial in  $y_i, \dot{y}_i, \dots, y_i^{(\rho_i-1)}$  chosen in such a way that

$$y_i^{\rho_i} = P_i(y_i, \dot{y}_i, \dots, y_i^{(\rho_i-1)})$$

has stable linear dynamics. Using such a control, internal stability results should generalise directly. In the case that  $A(x)$  is not full rank, we expect that the result could be extended by using the dynamic extension algorithm and arguments similar to those that appear in Mahony, Mareels, Bastin & Campion (1996).

Additional future research could also look at more general systems that do not necessarily exhibit zero drift for zero output.

**Interleaved Pulse Train Spectrum Estimation.** In Chapter 3 of this thesis we looked at a problem closely related to the pulse train deinterleaving problem. Considered were signals consisting of a finite though unknown number of periodic time-interleaved pulse trains. For such signals, a nonlinear approach for determining both the number of pulse trains present and the frequency of each pulse train was presented. The approach was found to be robust to noisy time of arrival data and missing pulses, and most importantly, to be more computationally efficient than traditional pulse train deinterleaving methods.

**Future research** in this area could involve achieving a better understanding of practical implementation issues and dealing with these issues by making appropriate modifications to the proposed approach. Additionally, it would be desirable to achieve a better understanding of the magnitude plots and to develop some sort of test, perhaps statistical in nature, to determine if a given magnitude corresponded to a pulse train. (It would also be desirable to determine if the largest magnitude present does indeed always correspond to a pulse train.) Yet another area for possible future research would be to undertake a statistical analysis of the relationship between the generic and non-generic cases. A fundamental property used in the non-generic analysis was that the pulse train frequencies were rational. The rationals are dense in the reals and perhaps this fact could be used in some manner of analysis to achieve further insight into the generic case.

**Equality Constrained Quadratic Optimization.** In Chapter 4 of this thesis we considered the problem of minimizing a quadratic cost subject to purely quadratic equality constraints. This problem was approached by first relating it to a standard semidefinite programming problem. The approach taken lead to a dynamical systems analysis of semidefinite programming and the formulation of a gradient descent flow which could be used to solve semidefinite programming problems. Though the reformulation of the initial problem as a semidefinite programming problem was found in general not to lead directly to a solution of the initial problem, the initial problem was solved by using a modified version of the flow incorporating a penalty function.

**Future research** could involve an analysis of the modified gradient flow and the associated methods proposed for solving the initial quadratic optimization problem. While the method simulated was found to work well in practice and converged to a likely optimum, at present it is not known whether the methods developed will always find the optimal solution. It would be of interest to investigate the performance of the proposed schemes on a wide variety of problems and especially to investigate their performance on some real engineering problems.

Another area for possible future research could be to try to develop a more efficient method of solving the modified gradient flow. Using a numerical ODE solver to solve this flow is computationally quite expensive. Ideally, one would like to develop an explicit numerical scheme for solving the flow. For the present, consider again the following flow from Chapter 4,

$$\dot{X} = \{\mathcal{A}X, X\} \quad (5.1)$$

where  $\mathcal{A} = A_0 + d_1 A_1 + \dots + d_m A_m$ . Exploiting the homogeneous structure of  $S(r, n)$  (cf. §4.2), the matrix  $X$  can be written as  $X = SX_0S^T$  where  $S \in GL(n, \mathbb{R})$  and  $X_0 \in \mathbb{R}^{n \times n}$  satisfies  $X_0 = X_0^T \geq 0$ . Substituting  $X = SX_0S^T$  into (5.1), produces

$$\{\dot{S}, X_0S^T\} = \{\mathcal{A}XS, X_0S^T\}.$$

Instead of a flow on  $X$ , one can now consider the following flow,

$$\dot{S} = \mathcal{A}SX_0S^T S$$

on  $S(t) \in GL(n, \mathbb{R})$ . This  $S$  flow can be thought of a sort of square root version of the original flow and one would expect it to be numerically better conditioned. In fact, it has a number of other numerical advantages over the old flow, for while in theory the solution of equation (5.1),  $X(t)$ , will always be positive semidefinite, numerical inaccuracies may possibly lead to  $X(t)$  leaving the positive semidefinite cone. (Negative eigenvalues of  $X(t)$  often lead to unstable numerical behaviour.) Conversely, the  $S$  flow guarantees that  $X(t)$

will remain positive semidefinite. Furthermore, if the solution method encounters numerical problems due to  $S(t)$  (and hence  $X(t)$ ) becoming singular, it could be re-initialised. The new value of  $X_0$  could be set to the current value of  $X(t) = S(t)X_0S(t)^T$  and  $S(t)$  could be re-initialised to the identity matrix.

In future research, it would also be desirable to be able to generalise the results presented in Chapter 4 to more general cost and constraint equations. Given  $b \in \mathbb{R}^n$  and  $c \in \mathbb{R}$ , consider the purely linear constraint,

$$b^T x = c. \quad (5.2)$$

Squaring both sides of equation (5.2) gives

$$x^T b b^T x = c^2$$

or

$$\text{tr}(BX) = c^2 \quad (5.3)$$

where  $B := b b^T$  and  $X = x x^T$ . Equation (5.3) is in the standard constraint equation form we have been dealing with and hence results can certainly be generalised to include problems that have a single purely linear constraint. (Note that loss of sign information due to squaring (5.2) does not cause a problem as  $X = x x^T = (-x)(-x)^T$ .) Ideally, one would like to extend results not only to problems with multiple, purely linear constraints, but to problems with linear-quadratic cost and constraint terms of the form

$$x^T A x + b^T x$$

where  $A = A^T \in \mathbb{R}^{n \times n}$  and  $b \in \mathbb{R}^n$ . At present, it is not immediately clear how such an extension could be achieved.

## ***References***

Isidori, A. (1995). *Nonlinear Control Systems*, Communications and Control Engineering Series, third edn, Springer-Verlag, Berlin, Germany.

Mahony, R. E., Mareels, I. M., Bastin, G. & Campion, G. (1996). Output stabilization of square non-linear systems. To appear in *Automatica*.