Volume 10, Number 2, 2010 ISSN 1562-8353

NONLINEAR DYNAMICS AND SYSTEMS THEORY
An International Journal of Research and Surveys

Volume 10 Number 2 2010

CONTENTS

 ISSN 1562-8353 Printed in Ukraine
To receive contents and abstracts by e-mail, visit our Website at: **http://www.e-ndst.kiev.ua**

NONLINEAR DYNAMICS & SYSTEMS THEORY Volume 10, No. 2, 2010

Nonlinear Dynamics and Systems Theory

An International Journal of Research and Surveys

InforMath Publishing Group
http://www.e-ndst.kiev.ua

Nonlinear Dynamics and Systems Theory

An International Journal of Research and Surveys

INSTRUCTIONS FOR CONTRIBUTORS

(1) General. Nonlinear Dynamics and Systems Theory (ND&ST) is an international journal devoted to publishing peer-refereed, high quality, original papers, brief notes and review articles focusing on nonlinear dynamics and systems theory and their practical applications in engineering, physical and life sciences. Submission of a manuscript is a representation that the submission has been approved by all of the authors and by the institution where the work was carried out. It also represents that the manuscript has not been previously published, has not been copyrighted, is not being submitted for publication elsewhere, and that the authors have agreed that the copyright in the article shall be assigned exclusively to InforMath Publishing Group by signing a transfer of copyright form. Before submission, the authors should visit the website:

http://www.e-ndst.kiev.ua

for information on the preparation of accepted manuscripts. Please download the archive Sample_NDST.zip containing example of article file (you can edit only the file Samplefilename.tex).

(2) Manuscript and Correspondence. Manuscripts should be in English and must meet common standards of usage and grammar. To submit a paper, send by e-mail a file in PDF format directly to

Professor A.A. Martynyuk, Institute of Mechanics,
Nesterov str.3, 03057, MSP 680, Kiev-57, Ukraine
e-mail: anmart@stability.kiev.ua; center@inmech.kiev.ua

or to one of the Regional Editors or to a member of the Editorial Board. Final version of the manuscript must typeset using LaTex program which is prepared in accordance with the style file of the Journal. Manuscript texts should contain the title of the article, name(s) of the author(s) and complete affiliations. Each article requires an abstract not exceeding 150 words. Formulas and citations should not be included in the abstract. AMS subject classifications and key words must be included in all accepted papers. Each article requires a running head (abbreviated form of the title) of no more than 30 characters. The sizes for regular papers, survey articles, brief notes, letters to editors and book reviews are: (i) 10-14 pages for regular papers, (ii) up to 24 pages for survey articles, and (iii) 2-3 pages for brief notes, letters to the editor and book reviews.

(3) Tables, Graphs and Illustrations. Each figure must be of a quality suitable for direct reproduction and must include a caption. Drawings should include all relevant details and should be drawn professionally in black ink on plain white drawing paper. In addition to a hard copy of the artwork, it is necessary to attach the electronic file of the artwork (preferably in PCX format).

(4) References. References should be listed alphabetically and numbered, typed and punctuated according to the following examples. Each entry must be cited in the text in form of author(s) together with the number of the referred article or in the form of the number of the referred article alone.

Journal: [1] Poincare, H. Title of the article. *Title of the Journal* **Vol. l** (No.l), Year, Pages. [Language]

Book: [2] Liapunov, A.M. *Title of the book.* Name of the Publishers, Town, Year.

Proceeding: [3] Bellman, R. Title of the article. In: *Title of the book.* (Eds.). Name of the Publishers, Town, Year, Pages. [Language]

(5) Proofs and Sample Copy. Proofs sent to authors should be returned to the Editorial Office with corrections within three days after receipt. The corresponding author will receive a sample copy of the issue of the Journal for which his/her paper is published.

(6) Editorial Policy. Every submission will undergo a stringent peer review process. An editor will be assigned to handle the review process of the paper. He/she will secure at least two reviewers' reports. The decision on acceptance, rejection or acceptance subject to revision will be made based on these reviewers' reports and the editor's own reading of the paper.

NONLINEAR DYNAMICS AND SYSTEMS THEORY
An International Journal of Research and Surveys
Published by InforMath Publishing Group since 2001

Volume 10 Number 2 2010

CONTENTS

Founded by A.A. Martynyuk in 2001.
Registered in Ukraine Number: KB 5267 / 04.07.2001.

NONLINEAR DYNAMICS AND SYSTEMS THEORY

An International Journal of Research and Surveys

Nonlinear Dynamics and Systems Theory (ISSN 1562–8353 (Print), ISSN 1813–7385 (Online)) is an international journal published under the auspices of the S.P. Timoshenko Institute of Mechanics of National Academy of Sciences of Ukraine and Curtin University of Technology (Perth, Australia). It aims to publish high quality original scientific papers and surveys in areas of nonlinear dynamics and systems theory and their real world applications.

AIMS AND SCOPE

Nonlinear Dynamics and Systems Theory is a multidisciplinary journal. It publishes papers focusing on proofs of important theorems as well as papers presenting new ideas and new theory, conjectures, numerical algorithms and physical experiments in areas related to nonlinear dynamics and systems theory. Papers that deal with theoretical aspects of nonlinear dynamics and/or systems theory should contain significant mathematical results with an indication of their possible applications. Papers that emphasize applications should contain new mathematical models of real world phenomena and/or description of engineering problems. They should include rigorous analysis of data used and results obtained. Papers that integrate and interrelate ideas and methods of nonlinear dynamics and systems theory will be particularly welcomed. This journal and the individual contributions published therein are protected under the copyright by International InforMath Publishing Group.

PUBLICATION AND SUBSCRIPTION INFORMATION

Nonlinear Dynamics and Systems Theory will have 4 issues in 2010, printed in hard copy (ISSN 1562–8353) and available online (ISSN 1813–7385), by InforMath Publishing Group, Nesterov str., 3, Institute of Mechanics, Kiev, MSP 680, Ukraine, 03057. Subscription prices are available upon request from the Publisher (`mailto:anmart@stability.kiev.ua`), SWETS Information Services B.V. (`mailto:Operation-Academic@nl.swets.com`), EBSCO Information Services (`mailto:journals@ebsco.com`), or website of the Journal: `http://e-ndst.kiev.ua`. Subscriptions are accepted on a calendar year basis. Issues are sent by airmail to all countries of the world. Claims for missing issues should be made within six months of the date of dispatch.

ABSTRACTING AND INDEXING SERVICES

Papers published in this journal are indexed or abstracted in: Mathematical Reviews / MathSciNet, Zentralblatt MATH / Mathematics Abstracts, PASCAL database (INIST–CNRS) and SCOPUS.

Special Issue
on
Dynamical Systems and Control Theory and Their Applications
in dedication to Professor T.L. Vincent

This is a special issue dedicated to the memory of Professor Thomas L. Vincent. Tom Vincent passed away on 26 October, 2009 at the age of 74. He was the first graduate of the University of Arizona's doctoral program in Aerospace Engineering. He was a Professor of Aerospace and Mechanical Engineering at the University of Arizona for 41 years until his retirement in 2001, after which he became an active Professor Emeritus.

Tom has made outstanding contributions to optimal control theory, game theory, evolutionary theory and applications to aerospace problems, control systems, management of biological populations and Darwinian dynamics. He has published ten books and about 150 papers on his own and with others *). His work has influenced the scientific activity of many researchers in these areas. Tom was a great colleague and a true friend. He will be missed by all those who knew and benefited from him and/or his work.

The editors of this special issue have had the privilege of knowing Professor Vincent, personally and professionally. B.S. Goh first met Tom in 1970 at Berkeley where they were both on sabbatical leave and doing research with George Leitmann. BSG introduced Tom to applications of optimal control in the management of fisheries and pest populations and began a fruitful collaboration in this and other research areas. C.S. Lee studied MSc under BSG and PhD with Tom.

Tom made many visits to Western Australia and first met K.L. Teo in 1983. They had established long and lasting research collaboration in control theory and applications.

Professor Vincent made a significant contribution to the Journal Nonlinear Dynamics and Systems Theory during 2001–2005 as a member of Editorial Board of the Journal.

We wish to express our sincere appreciation to all those who have contributed to the completion of this Special Issue. In particular, we are deeply grateful to our referees who provided prompt and extensive reviews for all the submitted papers. Their valuable constructive comments have contributed to the quality of this Special Issue. We also wish to thank the Editor-in-Chief, Professor A.A. Martynyuk, for his kind cooperation and professional support. Our special thanks go to Mrs. Lisa Holling, Teo's Secretary, for her help during the editing process of the Special Issue. In this Special Issue we are pleased to have contributions from a number of outstanding scholars who have known Tom Vincent, both professionally and personally.

B.S. Goh[1], and K.L. Teo[2],
Guest Editors

*) A list of Tom Vincent's publications is available in the journal, Evolutionary Ecology Research Vol. 11 (2009).

[1]Nanjing University, mailto:bsgoh65@yahoo.com
[2]Curtin University of Technology, mailto:K.L.Teo@curtin.edu.au

Nonlinear Dynamics and Systems Theory, **10** (2) (2010) 103–116

On the Dynamics of a Class of Darwinian Matrix Models †

J. M. Cushing *

Department of Mathematics and the Interdisciplinary Program in Applied Mathematics
University of Arizona, 617 N. Santa Rita, Tucson, AZ 85721 USA

Received: November 15, 2009; Revised: March 24, 2010

Abstract: Using the methodology of evolutionary game theory (EGT), I study a class of Darwinian matrix models which are derived from a class of nonlinear matrix models for structured populations that are known to possess stable (normalized) distributions. Utilizing the limiting equations that result from this ergodic property, I prove extinction and stability results for the limiting equations of the EGT versions of these kinds of structured population models. This is done in a bifurcation theory context. The results provide conditions sufficient for a branch of non-extinction equilibria to bifurcate from the branch of extinction equilibria. When this bifurcation is supercritical (explicit criteria are given), these equilibria are stable and represent stable non-extinction equilibria (which are also candidate ESS equilibria). These kinds of matrix models are motivated by applications to size structured populations, and I give an application of this type. Besides illustrating the formal theory, this application shows the importance of trade-offs among life history parameters that are necessary for the existence of an evolutionarily stable equilibrium.

Keywords: *structured population dynamics; nonlinear matrix model; stable distribution; limiting equation; evolutionary game theory; bifurcation; equilibrium, stability.*

Mathematics Subject Classification (2000): 92D15, 92D25, 39A60.

1 Introduction

Nonlinear matrix models are widely used to describe and study the discrete time dynamics of structured populations. These models take the form

$$x(t+1) = P(x(t))x(t), \tag{1}$$

† Research partially supported by NSF grant DMS 0917435.
* Corresponding author: mailto:cushing@math.arizona.edu

where $P(x)$ is an $m \times m$ non-negative projection matrix that is assumed primitive (irreducible and possessing a strictly positive dominant eigenvalue) for each (column vector) $x \in \Omega$, where Ω is an open set in R^m containing the origin. Here $x(t)$ is a demographic distribution vector at time $t \in Z_+ = \{0, 1, 2, \cdots\}$ that is based on a classification scheme for individuals in the population (chronological age, weight, size, etc.). For more on matrix models in population dynamics see [1, 9, 10].

In general the projection matrix has the form

$$\begin{array}{c} P(x) = F(x) + T(x), \\ F(x) = [f_{ij}(x)], \quad T(x) = [s_{ij}(x)], \end{array} \tag{2}$$

where $f_{ij} \geq 0$ is the amount (number, density, etc.) of surviving i-class offspring per j-class individual in a unit of time and where s_{ij}, $0 \leq s_{ij} \leq 1$, is the fraction of j-class individuals that survive and move to the i-class over one unit of time [9, 10]. In one type of model that arises in population dynamics and theoretical ecology, the projection matrix also has the form

$$P(x) = a(x)I + b(x)L, \tag{3}$$

where I is the $m \times m$ identity matrix, L is an $m \times m$ constant matrix, and a, b are scalar valued functions of x. For examples see [2, 7, 6, 9, 15], Chapter 17 in [3], Chapter 3 in [5], and Section 3.

For models of the form (3) there exists an asymptotically stable (normalized) distribution vector. This is a consequence of the following theorem.

Theorem 1.1 [2, 7, 9] *Consider the equation* $x(t+1) = (\alpha(t)I + \beta(t)L)\,x(t)$ *where (a)* α, β *are real valued functions for which there exist constants* α_0, β_0 *such that* $0 \leq \alpha(t) \leq \alpha_0$, $0 < \beta_0 \leq \beta(t)$ *for all* $t \in Z_+$; *(b) the* $m \times m$ *constant matrix* L *has a strictly dominant, simple eigenvalue* $\theta > 0$ *with a positive eigenvector* $v \in int\left(R_+^m\right)$. *Suppose* $x(t)$ *is a solution satisfying* $0 \neq x(t) \geq 0$ *for all* $t \in Z_+$ *and* $p(t)$ *is a weighted total population size:*

$$p(t) \doteq \omega \cdot x(t), \quad 0 \neq \omega \in R_+^m.$$

Then

$$\lim_{t \to +\infty} \frac{x(t)}{p(t)} = \frac{v}{\omega \cdot v}. \tag{4}$$

We can apply Theorem 1.1 to solutions of the nonlinear matrix equation (1)-(3) with $\alpha(t) = a(x(t))$ and $\beta(t) = b(x(t))$. We then use (4) to replace $x(t)$ and $x(t+1)$ in (1) by their asymptotic equivalents $p(t)v/\omega \cdot v$ and $p(t+1)v/\omega \cdot v$ and obtain the scalar limiting equation

$$p(t+1) = \left[a\left(\frac{v}{\omega \cdot v}p(t)\right) + b\left(\frac{v}{\omega \cdot v}p(t)\right)\theta\right] p(t)$$

for the total population size $p(t)$. Thus, for these kinds of matrix models, the high dimensional dynamics of the original model are replaced by those of the scalar limiting equation for total population size (which depend on the dominant eigenvalue θ of L), and the asymptotic distribution (4) (calculated from the eigenvector v associated with θ). For applications see Section 3 and [9] (and papers cited therein).

Under the assumption that $P(x)$ is nonnegative and primitive for $x \in \Omega$, it has a strictly dominant eigenvalue $r = r(x) > 0$. It is easy to see that under the assumption (b) in Theorem 1.1

$$r(x) = a(x) + b(x)\theta$$

and as a result the limiting equation can be written as

$$p(t+1) = r\left(\frac{v}{\omega \cdot v} p(t)\right) p(t). \tag{5}$$

As a special case, if $P = P(p)$ and hence $a = a(p)$ and $b = b(p)$ are functions of a weighted total population size p (as they frequently are in applications), the limiting equation is $p(t+1) = [a(p(t)) + b(p(t))\,\theta]\,p(t)$ or

$$p(t+1) = r(p(t))\,p(t).$$

Theorems relating the (equilibrium and cycle) dynamics of the limiting equation to the dynamics of the original matrix model appear in [7].

In their book Vincent and Brown [17] provide a methodology for extending matrix models for structured populations to an evolutionary setting. Their methodology involves a dynamically evolving phenotypic trait, which affects demographic parameters in the entries of the projection matrix and whose dynamics are in turn affected by the population dynamics. Vincent and Brown refer to this coupling of the evolutionary and population dynamics as Darwinian dynamics. Our goal here is to study Darwinian matrix models with projection matrices of the particular form (3) by making use of the ergodic Theorem 1.1 and the resulting limiting equation (5). In Section 2 we study, in the context of bifurcation theory, the existence and stability of both extinction and non-extinction equilibria. Section 3 contains an application to a Darwinian model based on a class of structured models studied in the literature which has historical roots in a seminal paper of Leslie on matrix models in population dynamics [15].

2 Darwinian Matrix Models

Let u denote the mean of a phenotypic trait (with a heritable component) that is subject to natural selection. The Darwinian dynamics associated with a matrix equation are

$$\begin{aligned} x(t+1) &= P(x(t), u(t))x(t), \\ u(t+1) &= u(t) + \sigma^2 \frac{\partial \ln r(x(t), u(t))}{\partial u}, \end{aligned} \tag{6}$$

where $P = P(x, u)$ is now assumed a function of u as well as x and $r = r(x, u)$ is its dominant eigenvalue. Here the constant σ^2 is the variance of the phenotypic trait each point in time; it is a measure of the speed of evolution. Let $\Upsilon \subseteq R^1$ be an open interval. We make the following assumptions:

$$A: \begin{cases} \textit{The nonnegative, primitive matrix } P(x,u) \textit{ has the form (3)} \\ \textit{with an } m \times m \textit{ constant matrix } L \textit{ and real valued functions} \\ a, b \in C^2\left(\Omega \times \Upsilon \to R^1_+\right) \textit{ that satisfy the following:} \\ \quad \textit{(a) there exist constants } a_0, b_0 \textit{ such that } 0 \le a(x,u) \le a_0, \\ \qquad \textit{and } 0 < b_0 \le b(x,u) \textit{ for } (x,u) \in \Omega \times \Upsilon; \\ \quad \textit{(b) } L \textit{ has a simple, strictly dominant eigenvalue } \theta > 0 \textit{ with a} \\ \qquad \textit{positive eigenvector } v. \end{cases}$$

Under assumption A, Theorem 1.1 applies to (6) with $\alpha(t) = a(x(t), u(t))$ and $\beta(t) = b(x(t), u(t))$ and implies that solutions have a stable normalized distribution (4). From

(6) we derive the two scalar equations

$$p(t+1) = \omega \cdot P\left(x(t), u(t)\right) x(t), \tag{7a}$$

$$u(t+1) = u(t) + \sigma^2 \frac{\partial \ln r\left(x(t), u(t)\right)}{\partial u}, \tag{7b}$$

for the dynamics of the total population size $p(t) = \omega \cdot x(t)$ and the mean trait $u(t)$. Replacing $x(t)$ by $vp(t)/\omega \cdot v$, we obtain the limiting equations [7, 14]

$$p(t+1) = r\left(\frac{p(t)}{\omega \cdot v} v, u(t), \theta\right) p(t), \tag{8a}$$

$$u(t+1) = u(t) + \sigma^2 \left.\frac{\partial \ln r\left(x, u, \theta\right)}{\partial u}\right|_{(x,u)=\left(\frac{p(t)}{\omega \cdot v} v, u(t)\right)}, \tag{8b}$$

for $p(t)$ and $u(t)$, where for convenience we have added θ to the argument list in the dominant eigenvalue

$$r(x, u, \theta) \triangleq a(x, u) + b(x, u)\theta \tag{9}$$

of $P(x, u) = a(x, u)I + b(x, u)L$. This system of limiting equations is two dimensional and therefore more analytically tractable than the original $m + 1$ dimensional matrix model (6). We now turn our attention to an analysis of the equilibrium states of this limiting system. We will relate these dynamics to those of the original matrix model in Section 2.3.

2.1 The limiting system: existence of equilibria

The equilibrium equations for (8) are

$$p = r\left(\frac{p}{\omega \cdot v} v, u, \theta\right) p,$$
$$0 = r_u\left(\frac{p}{\omega \cdot v} v, u, \theta\right),$$

where the subscript u denotes partial differentiation $\partial/\partial u$. We are interested in the existence of two types of equilibria. An *extinction equilibrium* (p, u) of (8) is one in which $p = 0$ and a *non-extinction equilibrium* is one in which $p > 0$.

We are also interested in the stability of these equilibria, when they exist. We refer to a (locally asymptotically) stable equilibrium as an *evolutionarily stable equilibrium*. (In the language of [17] the associated equilibrium trait has convergent stability.) We say that a population whose orbit tends to a stable extinction equilibrium evolves to extinction, while one whose orbits tend to a non-extinction equilibrium evolutionarily persists and equilibrates.

Definition 2.1 A pair u, θ (with $\theta > 0$) is an **extinction pair** if

$$r_u\left(0, u, \theta\right) = a_u\left(0, u\right) + b_u\left(0, u\right)\theta = 0. \tag{10}$$

An extinction pair u^*, θ^* is a **critical extinction pair** if in addition it satisfies $r\left(0, u^*, \theta^*\right) = 1$. That is to say, a critical extinction pair u^*, θ^* satisfies

$$\begin{aligned} r\left(0, u^*, \theta^*\right) &= a\left(0, u^*\right) + b\left(0, u^*\right)\theta^* = 1, \\ r_u\left(0, u, \theta\right) &= a_u\left(0, u^*\right) + b_u\left(0, u^*\right)\theta^* = 0. \end{aligned} \tag{11}$$

Clearly $(p,u)=(0,u)$ is an extinction equilibrium of (8)-(9) (with parameter value θ) if and only if u,θ is an extinction pair. As we will see, critical extinction pairs serve as bifurcation points for the creation of non-extinction equilibria.

The non-extinction equilibrium equations are

$$\begin{aligned} 1 &= r\left(\frac{p}{\omega\cdot v}v,u,\theta\right),\\ 0 &= r_u\left(\frac{p}{\omega\cdot v}v,u,\theta\right). \end{aligned}$$

If u^*,θ^* is a critical extinction pair, the implicit function theorem implies that these equilibrium equations have a solution $(p,u)=(\pi(\theta),\upsilon(\theta))$ for θ near θ^*, where $\pi(\theta),\upsilon(\theta)$ are twice continuously differentiable functions that satisfy $(\pi(\theta^*),\upsilon(\theta^*))=(0,u^*)$, provided the Jacobian with respect to p and u

$$\begin{pmatrix} \nabla_x r\left(0,u^*,\theta^*\right)\cdot\frac{v}{\omega\cdot v} & 0\\ \nabla_x r_u\left(0,u^*,\theta^*\right)\cdot\frac{v}{\omega\cdot v} & r_{uu}\left(0,u^*,\theta^*\right)\end{pmatrix}$$

is non-singular at $(p,u)=(0,u^*),\theta=\theta^*$, i.e. provided

$$\delta\doteq\nabla_x r\left(0,u^*,\theta^*\right)\cdot v\neq 0\quad\text{and}\quad r_{uu}\left(0,u^*,\theta^*\right)\neq 0.$$

This branch of equilibria $(p,u)=(\pi(\theta),\upsilon(\theta))$ consists of non-extinction equilibria $p=\pi(\theta)>0$ for $\theta>\theta^*$ if $\pi'(\theta)>0$ or for $\theta<\theta^*$ if $\pi'(\theta^*)<0$. An implicit differentiation of $1=r\left(\pi(\theta)v/\omega\cdot v,\upsilon(\theta),\theta\right)$ shows (recall (11))

$$\pi'(\theta^*)=-\frac{\omega\cdot v}{\delta}r_\theta\left(0,u^*,\theta^*\right).$$

Since $r_\theta\left(0,u^*,\theta^*\right)=b\left(0,u^*,\theta^*\right)>0$, the sign of $\pi'(\theta^*)$ is the opposite of the sign of δ.

Theorem 2.1 *Assume A and that u^*, $\theta^*>0$ is a critical extinction pair (i.e., a pair that satisfies (11)) for which*

$$\begin{aligned} &\delta\doteq\left[\nabla_x a\left(0,u^*\right)+\nabla_x b\left(0,u^*\right)\theta^*\right]\cdot v\neq 0,\\ &a_{uu}\left(0,u^*\right)+b_{uu}\left(0,u^*\right)\theta^*\neq 0. \end{aligned}\tag{12}$$

Then there exists a (twice continuously differentiable) branch of non-extinction equilibria $(p,u)=(\pi(\theta),\upsilon(\theta))$ for

$$\begin{aligned} \theta &\gtrapprox \theta^* \quad \textit{if}\quad \delta<0,\\ \theta &\lessapprox \theta^* \quad \textit{if}\quad \delta>0, \end{aligned}$$

such that $(\pi(\theta^*),\upsilon(\theta^*))=(0,u^*)$.

In many applications, the dependency of the projection matrix, and hence a and b, on x is through a dependency on a weighted population size p, i.e., $a(p,u)$ and $b(p,u)$. In that case, $\delta=r_p\left(0,u^*,\theta^*\right)\omega\cdot v$ and the condition $\delta\neq 0$ is equivalent to

$$a_p\left(0,u^*\right)+b_p(0,u^*)\theta^*\neq 0,$$

where a_p and b_p are the partial derivatives of a and b with respect to p.

We can view the existence result in Theorem 2.1 as a bifurcation phenomenon by using θ as a bifurcation parameter. To clarify this, we distinguish two types of extinction pairs.

Definition 2.2 A **type 1 extinction pair** u, θ is one for which

$$b_u(0,u) = a_u(0,u) = 0 \quad \text{and} \quad \theta \in \Upsilon \text{ is arbitrary.}$$

A **type 2 extinction pair** u, θ is one for which

$$b_u(0,u) \neq 0 \quad \text{and} \quad \theta = -\frac{a_u(0,u)}{b_u(0,u)}.$$

Type 1 extinction pairs produce a branch of extinction equilibria $(p,u) = (0,u^*)$ of the limiting system (8) for all values of $\theta \in \Upsilon$ where u^* satisfies $b_u(0,u^*) = a_u(0,u^*) = 0$. The branch of non-extinction equilibria in Theorem 2.1 intersects this branch of extinction equilibria in a transcritical bifurcation at the critical extinction pair $u, \theta = u^*, \theta^*$ where

$$\theta^* = \frac{1 - a(0,u^*)}{b(0,u^*)}. \tag{13}$$

See Figure 1(a,b).

Type 2 extinction pairs produce a branch of extinction equilibria $(p,u) = (0,u)$ for $\theta = -a_u(0,u)/b_u(0,u)$ and those values of u for which $b_u(0,u) \neq 0$. The branch of non-extinction equilibria in Theorem 2.1 intersects this branch of extinction equilibria in a transcritical bifurcation at the critical extinction pair u^*, θ^* value where u^* satisfies

$$-\frac{a_u(0,u^*)}{b_u(0,u^*)} = \frac{1 - a(0,u^*)}{b(0,u^*)}. \tag{14}$$

and θ^* is given by (13). See Figure 1(c,d).

We say that the bifurcation is *supercritical* (or *to the right*) if $\delta < 0$ and *subcritical* (or *to the left*) if $\delta > 0$.

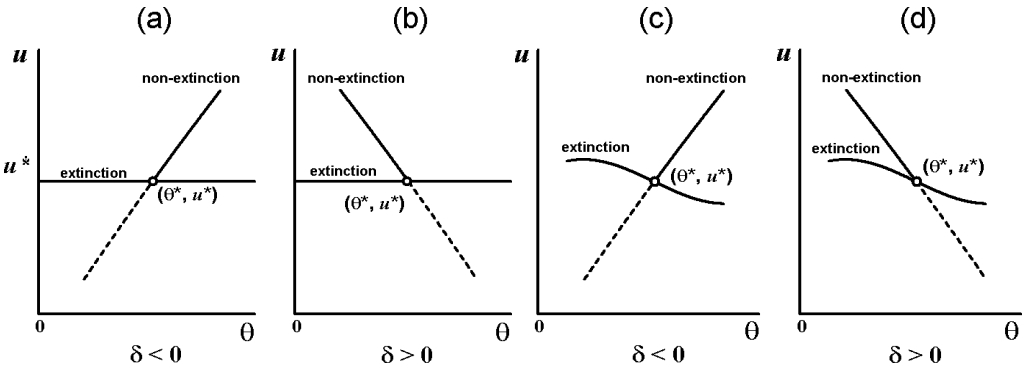

Figure 1. All graphs show intersecting branches of extinction and non-extinction pairs u, θ (which correspond to extinction and non-extinction equilibria of the limiting equations (8) respectively). The dashed lines are pairs that correspond to equilibria with $p < 0$ and therefore are not biologically relevant. The intersection occurs at a critical pair u^*, θ^* with θ^* defined by (13). In graphs (a) and (b) the extinction pairs are of Type 1 and plot as a horizontal straight line where u^* satisfies $b_u(0,u^*) = a_u(0,u^*) = 0$. In graphs (c) and (d) the extinction pairs are of Type 2 and u^* satisfies (14).

2.2 The limiting system: stability of equilibria

The eigenvalues μ_1, μ_2 of the Jacobian of the limiting system (8), which when evaluated at either an extinction or a non-extinction equilibrium has a triangular form (because $r_u\left(\frac{p}{\omega\cdot v}v, u\right)p$ vanishes at either type of equilibrium), are

$$\mu_1 \doteq r\left(\frac{p}{\omega\cdot v}v, u, \theta\right) + p\frac{\nabla_x r\left(\frac{p}{\omega\cdot v}v, u, \theta\right)\cdot v}{\omega\cdot v}, \quad \mu_2 \doteq 1 + \sigma^2 r_{uu}\left(\frac{p}{\omega\cdot v}v, u, \theta\right).$$

For an extinction equilibrium $(p, u) = (0, u)$ these eigenvalues are

$$\mu_1 = r(0, u, \theta), \quad \mu_2 = 1 + \sigma^2 r_{uu}(0, u, \theta).$$

The linearization principle implies the equilibrium is unstable if $r(0, u, \theta) > 1$ or if $r_{uu}(0, u, \theta) > 0$ and is (locally asymptotically) stable if $r(0, u, \theta) < 1$, $r_{uu}(0, u, \theta) < 0$ and $\sigma^2 < -2/r_{uu}(0, u, \theta)$. Note it is necessary for the stability of an extinction equilibrium $(p, u) = (0, u)$ that $r(0, \cdot, \theta)$ have a local maximum at u.

Lemma 2.1 *Assume A and that u, θ is an extinction pair.*
(a) The extinction equilibrium $(0, u)$ of the limiting system (8) is unstable if

$$a(0, u) + b(0, u)\theta > 1 \quad \text{or} \quad a_{uu}(0, u) + b_{uu}(0, u)\,\theta > 0.$$

(b) Assume $a_{uu}(0, u) + b_{uu}(0, u)\,\theta < 0$. Then $(0, u)$ is (locally asymptotically) stable if

$$a(0, u) + b(0, u)\theta < 1 \quad \text{and} \quad \sigma^2 < -2\left(a_{uu}(0, u) + b_{uu}(0, u)\,\theta\right)^{-1}.$$

Let u^*, θ^* be a critical extinction pair for which the conditions (12) hold. This point is a bifurcation point for non-extinction equilibria (as in Figure 1) whose stability properties we now consider. If

$$a_{uu}(0, u^*) + b_{uu}(0, u^*)\,\theta^* > 0,$$

then, because $\mu_2 > 1$, the extinction equilibria for $\theta \approx \theta^*$ are unstable (Lemma 2.1). By continuity, an eigenvalue of the Jacobian evaluated at the bifurcating non-extinction equilibria is also greater than one for $\theta \approx \theta^*$. Thus, in this case equilibria of both types are unstable near the bifurcation point.

If, on the other hand,

$$a_{uu}(0, u^*) + b_{uu}(0, u^*), \theta^* < 0,$$

then by Lemma 2.1 the extinction equilibrium loses stability as the bifurcation parameter θ increases through the critical value θ (assuming $\theta_{cr} > 0$). It follows by the exchange of stability principle for transcritical bifurcations [13] that a supercritical (right) bifurcation results in the stability of the non-extinction equilibria and a subcritical (left) bifurcation results in the instability of the non-extinction equilibria.

We have arrived at our main result concerning the limiting system (8) for the Darwinian matrix model (6).

Theorem 2.2 *Assume A and that u^*, $\theta^* > 0$ is a critical extinction pair for which (12) and $a(0, u^*) < 1$ hold. Then for the limiting system (8) there exist branches of extinction and non-extinction equilibria, parameterized by θ, that transcritically bifurcate (intersect) at $\theta = \theta^*$ given by (13). (a) Assume $a_{uu}(0, u^*) + b_{uu}(0, u^*)\,\theta^* < 0$. Then*

the extinction equilibria lose stability as θ increases through θ^*. Moreover, near the bifurcation point (i.e. for $\theta \approx \theta^*$), and for σ^2 sufficiently small, i.e., for

$$\sigma^2 < -2\left(a_{uu}\left(0,u^*\right) + b_{uu}\left(0,u^*\right)\theta^*\right)^{-1},$$

the bifurcating non-extinction equilibria are (evolutionarily) stable if the bifurcation is supercritical ($\delta < 0$) and are unstable if the bifurcation is subcritical ($\delta > 0$). (b) If $a_{uu}\left(0,u^*\right) + b_{uu}\left(0,u^*\right)\theta^* > 0$, then both the extinction equilibria and the non-extinction equilibria are unstable near the bifurcation point, i.e. for ($\theta \approx \theta^*$).

2.3 The Darwinian matrix model

In Sections 2.1 and 2.2, we obtained existence and stability results for the limiting equations (8) of the dynamic equations (7) for $p(t)$ and $u(t)$ associated with the Darwinian matrix equation (6). Under certain hypotheses the asymptotic dynamics of these two systems are related [7, 14]. (The theorems and the proofs given in [7] are for scalar maps, but remain valid virtually verbatim for systems of scalar maps.) Roughly speaking, if the dynamics of the limiting equations are not too complicated, then no orbit of (7) will approach an unstable equilibrium (or cycle) of the limiting system and if $(p(0),u(0))$ is sufficiently close to a (locally asymptotically) stable equilibrium (or cycle) (p_e,u_e) of the limiting equations and if the initial normalized distribution $x(0)/p(0)$ is sufficiently close to the limiting distribution $v/\omega\cdot v$, then

$$\lim_{t\to+\infty}\frac{x(t)}{p(t)} = \frac{v}{\omega\cdot v} \quad \text{and} \quad \lim_{t\to+\infty}(p(t),u(t)) = (p_e,u_e).$$

The hypotheses required are that the limiting equations have at most a finite number of equilibria (or cycles) in any compact subset of R^2_+, all of which are hyperbolic, and the ω-limit sets of bounded orbits are equilibria (or cycles).

We conclude with some remarks concerning the results in Sections 2.1 and 2.2.

Remark 2.1. The inequality $r_{uu}(0,u^*,\theta^*) = a_{uu}\left(0,u^*\right) + b_{uu}\left(0,u^*\right)\theta^* > 0$ in Theorem 2.2(b) implies that the inherent growth rate $r(0,u,\theta^*)$ has a local minimum (of 1) as a function of the trait u at $u = u^*$ (assuming the eigenvalue $\theta \approx \theta^*$ remains fixed). Since in this case all equilibria on both bifurcating branches (extinction and non-extinction) are unstable, it follows that no population will evolve to have trait $u \approx u^*$, whether the population goes extinct or not.

Remark 2.2. Evolutionarily stable equilibria occur in the transcritical bifurcation when $r_{uu}(0,u^*,\theta^*) < 0$ and hence $r(0,u,\theta^*)$ has a local maximum (of 1) as a function of the trait u. In this case, the extinction equilibria are unstable if the demographic parameters in the matrix L are such that $\theta \lessapprox \theta^*$ and populations evolve to extinction. On the other hand, for $\theta \gtrapprox \theta^*$ the extinction equilibria are unstable and the population will not evolve to extinction. In this case the non-extinction equilibria are stable if $\delta < 0$ and populations evolve to an evolutionarily stable non-extinction equilibrium (p_e,u_e), $p_e > 0$, with a trait $u = u_e$ at which $r(p_e,u,\theta_e)$ has a local maximum. This is because the equilibrium equation is $r_u(p_e,u,\theta_e) = 1$ and, by continuity, $r_{uu}(p_e,u_e,\theta_e) < 0$ for $\theta_e \gtrapprox \theta^*$. If $r(p_e,u,\theta_e)$ in fact has a global maximum on the trait interval Υ at $u = u_e$, then the evolutionary stability of the equilibrium plus $r_{uu}(p_e,u_e,\theta_e) < 0$ implies the equilibrium is an ESS (see the ESS Maximum Principle in [17]). That is to say, the

population evolves to a non-extinction equilibrium state that is resistant to invasion by other mutant species.

Remark 2.3. The condition $\delta \triangleq \nabla_x r\left(0, u^*, \theta^*\right) \cdot v < 0$, required for a supercritical bifurcation of evolutionarily stable non-extinction equilibria, is a negative feedback condition. This is because it requires sufficiently large negative derivatives of the inherent population growth rate r with respect to the components in the distribution vector x. This condition is met under the usual assumptions of so-called density effects in ecology. In order to fail, i.e., in order for $\delta > 0$, positive feedback terms (Allee effects) would have to out weigh the negative density effects. As we have seen, this would lead to a subcritical bifurcation of unstable non-extinction equilibria.

Remark 2.4. Since $r(x,u) = a(x,u) + b(x,u)\theta$, we have the relationship $r(0,u^*) = a(0,u^*) + b(0,u^*)\theta$ between the dominant eigenvalue $r(0,u^*)$ (the inherent population growth rate at the critical trait u^*) and θ. The bifurcation described in Theorems 2.1 and 2.2 in terms of θ can therefore be restated in terms of the magnitude of $r(0,u^*)$. Thus, the bifurcation phenomenon in these theorems (and hence the possibility of a bifurcation from an evolutionary state of extinction state to an evolutionary state of non-extinction) occurs when the magnitude of $r(0,u^*)$ increases through 1. See [11]. As is shown in [12], this phenomenon can also be equivalently stated in terms of the inherent net reproductive number $R_0(0,u^*)$ at the critical trait. See [12]. The quantity $R_0(0,u^*)$, which is generally more analytically tractable than $r(0,u^*)$, is the dominant eigenvalue of $F(0,u^*)(I - T(0,u^*))^{-1}$ [8, 9, 10].

Remark 2.5. The definition of a type 2 extinction pair u^*, θ^*clearly requires that $a(0,u)$ and $b(0,u)$ have opposite monotonicities at $u = u^*$. In specific applications the biological implication of this fact is usually that some kind of trade-off between two demographic parameters occurs as the the trait u is changed. We will see an example of this in Section 3.

3 An Application

Consider a projection matrix (2) in which the matrix of class transitions are

$$s_{jj} = \pi_j\left(1 - \gamma_j\right), \quad s_{ij} = \pi_i \tau_{ij} \gamma_j.$$

Here γ_j is the fraction that leaves the j^{th} size class per unit time, τ_{ij} is the fraction of those who leave that moves to class i, and π_j is the survival rates per unit time. We can put this general model into the form (3) under the following two assumptions: *the fraction of j-class individuals leaving the j-class, γ_j, and the class specific fertility rates, f_{ij}, are proportional to a function of a resource consumption rate $u \geq 0$ and the survival rates π_j are class independent.* Specifically

$$\gamma_j = \tau_j \phi(u), \quad f_{ij} = \pi_i(u)\varphi_{ij}\phi(u), \quad \pi_i = \pi(u),$$

where $0 \leq \phi(u) \leq 1$ for $u \in \Upsilon = [0, u_{\max})$, $u_{\max} \leq +\infty$. For a reproductively obligate resource, we have $\phi(0) = 0$. For this model, the fertility and transition matrices are

$$F = \pi(u)\phi(u)\left[\varphi_{ij}\right], \qquad T = \pi(u)\begin{bmatrix} 1-\tau_1\phi(u) & \tau_{12}\tau_2\phi(u) & \cdots & \tau_{1m}\tau_m\phi(u) \\ \tau_{21}\tau_1\phi(u) & 1-\tau_2\phi(u) & \cdots & \tau_{2m}\tau_m\phi(u) \\ \vdots & \vdots & & \vdots \\ \tau_{m1}\tau_1\phi(u) & \tau_{m2}\tau_2\phi(u) & \cdots & 1-\tau_m\phi(u) \end{bmatrix}.$$

Let $\tau_k = \max\{\tau_i\}$ and re-write T as

$$T = \pi\left(1 - \tau_k\phi(u)\right) I + \pi\phi(u) \begin{bmatrix} \tau_k - \tau_1 & \tau_{12}\tau_2 & \cdots & \tau_{1m}\tau_m \\ \tau_{21}\tau_1 & \tau_k - \tau_2 & \cdots & \tau_{2m}\tau_m \\ \vdots & \vdots & & \vdots \\ \tau_{m1}\tau_1 & \tau_{m2}\tau_2 & \cdots & \tau_k - \tau_m \end{bmatrix}.$$

Then we can write the projection matrix as

$$P = \pi(u)\left(1 - \tau_k\phi(u)\right) I + \pi(u)\phi(u)L, \tag{15}$$

where

$$L = \begin{bmatrix} \tau_k - \tau_1 + \varphi_{11} & \tau_{12}\tau_2 + \varphi_{12} & \cdots & \tau_{1m}\tau_m + \varphi_{1m} \\ \tau_{21}\tau_1 + \varphi_{21} & \tau_k - \tau_2 + \varphi_{22} & \cdots & \tau_{2m}\tau_m + \varphi_{2m} \\ \vdots & \vdots & & \vdots \\ \tau_{m1}\tau_1 + \varphi_{m1} & \tau_{m2}\tau_2 + \varphi_{m2} & \cdots & \tau_k - \tau_m + \varphi_{mm} \end{bmatrix} \tag{16}$$

is a non-negative matrix. This matrix model is motivated by applications in which the classes are based on physiological size of individuals; see [9] and the references cited therein for examples. It has the form (3) with $a = \pi(u)\left(1 - \tau_k\phi(u)\right)$ and $b = \pi\phi(u)$. We assume L has a positive dominant eigenvalue θ which has an associated positive eigenvector.

For density dependence in the fertility and survivorship rates $\pi = \pi(x,u)$, $\phi = \phi(x,u)$, then

$$a\left(x,u\right) = \pi(x,u)\left[1 - \tau_k\phi(x,u)\right], \quad b(x,u) = \pi(x,u)\phi(x,u).$$

In this application we assume density dependence is through a dependency on a weighted total population size p. Then

$$\gamma_j = \tau_j\phi(p,u), \quad f_{ij} = \varphi_{ij}\phi(p,u), \quad \pi_i = \pi(p,u) \tag{17}$$

in the fertility and transition matrices F and T. Theorems 1.1, 2.1 and 2.2 apply to this population model with

$$a(p,u) = \pi(p,u)\left(1 - \tau_k\phi(p,u)\right), \quad b(p,u) = \pi(p,u)\phi(p,u) \tag{18}$$

in the projection matrix (3). Thus, there is a stable normalized distribution and the asymptotic population dynamics are described by the limiting equations (8). We illustrate the application of Theorems 2.1 and 2.2 with a specific example.

Important in the evolution and adaptation of biological species are trade-offs among life history characteristics and strategies [16]. In the model above, *we assume a trade-off between fertility and survivorship as a function of the resource consumption rate* u. Thus, an increase in u results in an increase in fertility but also a decrease in survivorship. A decrease in survivorship can be the result of many causes: the stress and metabolic costs associated with finding and consuming prey, a resulting exposure to predators, etc.

We take

$$\phi(p,u) = \frac{1}{1+cp}f(u), \quad \pi(p,u) = \frac{1}{1+cp}\pi_0\left(1 - f(u)\right), \quad c > 0, \quad 0 < \pi_0 < 1, \tag{19}$$

where $f(u)$ is a twice continuously differentiable, real value function of u on an interval $0 \leq u \leq u_m \leq +\infty$ that satisfies

$$f(0) = 0, \quad f\,'(u) > 0, \quad \lim_{u \to u_m} f(u) = 1.$$

Specific examples include $f(u) = 1 - e^{-\alpha u}$ on $0 \leq u < +\infty$ and $f(u) = u^p$ on $0 \leq u \leq 1$. Here we have taken the dependence of fertility and survivorship on weighted population size p to have the discrete time, logistic form $1/(1 + cp)$ as first consider by Leslie [15]. Note that fertility is 0 at consumption rate $u = 0$ and that survivorship $\pi(p, u)$ is 0 as the consumption rate u approaches u_m. Neither of these two extremes is therefore favorable for the persistence of the population.

Straightforward calculations solving equations (11) show that there exists a (unique) critical extinction pair given by the formulas

$$u^* = f^{\,-1}\left(\frac{-(1-\pi_0) + \sqrt{1-\pi_0}}{\pi_0}\right), \qquad \theta^* = \tau_k + \frac{\pi_0}{2 - \pi_0 - 2\sqrt{1-\pi_0}}. \tag{20}$$

Note that $0 < u^* < u_m$. Moreover, further calculations show

$$\delta = -c\left(1 + \sqrt{1-\pi_0}\right)\omega \cdot v < 0,$$
$$a_{uu}(0, u^*) + b_{uu}(0, u^*)\,\theta^* = -2\left(2 - \pi_0 + 2\sqrt{1-\pi_0}\right)\left(f\,'(u^*)\right)^2 < 0,$$

and, as a result, there is a supercritical bifurcation of evolutionarily stable, non-extinction equilibria as θ increases through θ^* (Theorems 2.1 and 2.2). Since

$$b_u(0, u) = \left(2 - \pi_0 - 2\sqrt{1-\pi_0}\right) f\,'(u^*) > 0,$$

the critical extinction pair is of Type 2 and the bifurcation has the form in Figure 1(c).

As a consequence of these results, the Darwinian model (6) with (15) and (17)-(19) predicts evolution to extinction for $\theta < \theta^*$ and evolution to a non-extinction equilibrium for $\theta \gtrapprox \theta^*$. Note that the bifurcating, evolutionarily stable non-extinction equilibria have traits near u^* and therefore lie between the two unfavorable traits of 0 and u_m.

This bifurcation result is stated in terms of the dominant eigenvalue θ of L the matrix given by (16). Often of interest is how the bifurcation to evolutionarily stable states depends on the class-specific parameters appearing as entries in L. In general, of course, there is no formula that explicitly relates θ to the entries in L (when the number of classes m is large). However, as pointed out in Remark 4, this bifurcation result can be equivalently re-stated in terms of $r(0, u^*) = a(0, u^*) + b(0, u^*)\theta$, namely, that the bifurcation occurs as $r(0, u^*)$ increases through 1 or equivalently as the inherent net reproductive number $R_0(0, u^*)$ (at the critical trait u^*) increases through 1. The quantity $R_0(0, u^*)$ is the dominant eigenvalue of $F(0, u^*)\left(I - T(0, u^*)\right)^{-1}$ and explicit formulas for it in terms of the entries in the projection matrix are often available [8, 9, 10]. This is particularly true, for example, when there is only one newborn class, i.e., when only the first row in F is nonzero.

As an example, suppose the population model is based on an Usher matrix or, as it is called in [1], the standard size-structured model. In this model, individuals either remain in a size class or advance (grow into) the next size class in a unit of time. This means that the transition matrix T is bidiagonal with nonzero entries along the main diagonal and its subdiagonal only. All newborns are assumed to lie in the smallest size class and

hence only the first row of the fertility matrix F is nonzero. This Usher model takes the form

$$F = \pi(p,u)\phi(p,u)\begin{bmatrix} \varphi_{11} & \varphi_{12} & \cdots & \varphi_{1m} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

$$+\,\pi(p,u)\begin{bmatrix} 1-\tau_1\phi(p,u) & 0 & \cdots & 0 & 0 \\ \tau_1\phi(p,u) & 1-\tau_2\phi(p,u) & \cdots & 0 & 0 \\ 0 & \tau_2\phi(p,u) & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1-\tau_{m-1}\phi(p,u) & 0 \\ 0 & 0 & \cdots & \tau_{m-1}\phi(p,u) & 1 \end{bmatrix}.$$

The formula for the inherent net reproductive number of an Usher matrix gives (see [8, 9, 10])

$$R_0(p,u) = \pi(p,u)\phi(p,u)\sum_{i=1}^{m}\varphi_{1i}\prod_{j=1}^{i}\frac{\pi(p,u)\phi(p,u)\tau_{j-1}}{1-\pi(p,u)\left(1-\tau_j\phi(p,u)\right)},$$

where for notational convenience $\tau_0 = 1$ and $\tau_m = 0$. Thus, from (19) and (20) we obtain

$$R_0(0,u^*) = \pi_0\left(1-f(u^*)\right)f(u^*)\sum_{i=1}^{m}\varphi_{1i}\prod_{j=1}^{i}\frac{\pi_0\left(1-f(u^*)\right)f(u^*)\tau_{j-1}}{1-\pi_0\left(1-f(u^*)\right)\left(1-\tau_j f(u^*)\right)},$$

where

$$f(u^*) = \frac{-(1-\pi_0)+\sqrt{1-\pi_0}}{\pi_0}.$$

The bifurcation to evolutionary non-extinction equilibria occurs for $R_0(0,u^*) \gtrapprox 1$ [12]. In this interpretation, the bifurcation phenomenon can now be determined in terms of the any of the size-specific fertilities φ_{1i} or the growth rates τ_i or the survivorship π_0.

For example, if all size classes but the largest consist of juveniles, so that all $\varphi_{1i} = 0$ except $\varphi_{1m} > 0$, then we have the formula

$$R_0(0,u^*) = \left[\pi_0\left(1-f(u^*)\right)f(u^*)\right]^{m+1}\varphi_{1m}\prod_{j=1}^{m}\frac{\tau_{j-1}}{1-\pi_0\left(1-f(u^*)\right)\left(1-\tau_j f(u^*)\right)}$$

and the bifurcation requirement that $R_0(0,u^*) \gtrapprox 1$ can now be re-stated as a threshold for adult fertility $\varphi_{1m} \gtrapprox \varphi_{1m}^*$.

4 Concluding Remarks

Theorems 2.1 and 2.2 describe a fundamental bifurcation phenomenon for a class of nonlinear matrix models that describe the evolutionary dynamics of a structured population. The type of matrix models considered in these theorems (which are motivated by certain size-structured models that arise in applications found in the literature) possess a strong ergodic property: solutions, whatever their dynamics, have a stable (normalized)

class distribution. This property, when applied to the Darwinian matrix models obtained from these population dynamic models by the methods of evolutionary game theory [17], leads to limiting equations for the evolving phenotypic trait and the total (weighted) population size.

The bifurcation phenomenon in Theorems 2.1 and 2.2 is fundamental in the sense that it concerns the fundamental biological question of extinction versus non-extinction, or in the context of the Darwinian models (6) considered here, evolution to extinction versus evolution to a non-extinction equilibrium state. These theorems show that this transition occurs at (and only at) a critical value θ^* of the bifurcation parameter θ and what we have defined to be a critical extinction trait value $u = u^*$. The bifurcation does not always lead to stable non-extinction equilibria, however, and Theorem 2.2 describes when the bifurcation is stable and when it is not.

The requirements for a stable bifurcation turn out to imply (among other things) that the inherent growth rate r of the population dynamics must attain a (local) maximum at the critical value of the trait (a fact that also implies the evolutionarily stable non-extinction equilibria are candidates for an ESS [17]). Although we do not pursue the issue here, the biological interpretation of these requirements is that some kind of a trade-off must occur among vital life history traits as a function of the phenotypic trait u. This is illustrated by the example in Section 3.

There remain several interesting open problems. Theorem 2.1 provides the existence of a local bifurcating branch of non-extinction equilibria. Similar theorems for population dynamic models without evolution assert the global existence of this branch. A global bifurcation theorem for the Darwinian model is lacking. The instability results in 2.2(b), in which the equilibria on both the extinction equilibrium and non-extinction equilibrium branches are unstable, leave open the question of the asymptotic dynamics in this case. The same question arises in 2.2(a) when the bifurcation is subcritical. Also, the methodology of evolutionary game theory is applicable when more than one phenotypic trait evolves. The ergodic Theorem 1.1 would still apply to the Darwinian matrix models for multiple traits and hence permit an analysis by means of lower dimensional limiting equations. Bifurcation theorems for these multi-trait Darwinian models would be of interest.

In this paper the focus is on the special class of Darwinian matrix models with projection matrices of the form (3). A bifurcation theorem for matrix models with more general projection matrices is given in [11].

References

[1] Caswell, H. *Matrix Population Models: Construction, Analysis and Interpretation*, Second Edition. Sinauer Associates, Inc. Publishers, Sunderland, Massachusetts, 2001.

[2] Crowe, K. M. A nonlinear ergodic theorem for discrete systems. *Journal of Mathematical Biology* **32** (1994) 179–191.

[3] Crowe, K. M. Nonlinear ergodic theorems and symmetric versus asymmetric competition. In: *Structured-Population Models in Marine, Terrestrial, and Freshwater Systems* (S. Tuljapurkar & H. Caswell, eds.), Chapman & Hall, New York, 1997.

[4] Cushing, J. M. Some competition models for size-structured populations. *Rocky Mountain Journal of Mathematics* **20** (4) (1990) 879–897.

[5] Cushing, J. M. Competing size-structured species. *Mathematical Population Dynamics* (O. Arino, D. E. Axelrod, and M. Kimmel, eds.). Marcel Dekker, Inc., New York, 1991.

[6] Cushing, J. M. A discrete model of competing size-structured species. *Theoretical Population Biology* **41** (2) (1992) 372–387.

[7] Cushing, J. M. A strong ergodic theorem for some nonlinear matrix models for the dynamics of structured populations. *Natural Resource Modeling* **3**(3) (1989) 331–375.

[8] Cushing, J. M. and Yicang, Z. The net reproductive value and stability in structured population models, *Natural Resource Modeling* **8** (4) (1994) 1–37.

[9] Cushing, J. M. *An Introduction to Structured Population Dynamics*. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 71, SIAM, Philadelphia, 1998.

[10] Cushing, J. M. Matrix models and population dynamics. In: *Mathematical Biology* (Mark Lewis, A. J. Chaplain, James P. Keener, Philip K. Maini, eds.). IAS/Park City Mathematics Series, Vol. 14, American Mathematical Society, Providence, RI, 2009, 47–150.

[11] Cushing, J. M. A bifurcation theorem for Darwinian matrix models. *Nonlinear Studies* **17**(1) (2010) 1–13.

[12] Cushing, J. M. On the relationship between r and R_0 and its role in the bifurcation of stable equilibria of Darwinian matrix models. *Journal of Biological Dynamics* (to appear).

[13] Kielhöfer, H. *Bifurcation Theory: An Introduction with Applications to PDEs*. Applied Mathematical Sciences 156. Springer, New York, 2004.

[14] LaSalle, J. P. *The Stability of Dynamical Systems*. Regional Conference Series in Applied Mathematics, SIAM, Philadelphia, 1976.

[15] Leslie, P. H. Some further notes on the use of matrices in population mathematics. *Biometrika* **35** (1948) 213–245.

[16] Roff, D. A. *Life History Evolution*. Sinauer Associates Inc., Massachusetts, 2002.

[17] Vincent, T. L. and Brown, J. S. *Evolutionary Game Theory, Natural Selection, and Darwinian Dynamics*. Cambridge University Press, Cambridge, 2005.

Nonlinear Dynamics and Systems Theory, **10** (2) (2010) 117–124

A Stochastic Optimal Control Model of Pollution Abatement

D. Dragone [1], L. Lambertini [1*], G. Leitmann [2] and A. Palestini [1]

[1] *Department of Economics, University of Bologna, Strada Maggiore 45, 40125 Bologna, Italy*
[2] *College of Engineering, University of California at Berkeley, Berkeley CA 94720*

Received: November 5, 2009; Revised: March 29, 2010

Abstract: We model a dynamic monopoly with environmental externalities, investigating the adoption of a tax levied on the firm's instantaneous contribution to the accumulation of pollution. The latter process is subject to a shock, which is i.i.d. across instants. We prove the existence of an optimal tax rate such that the monopoly replicates the same steady state welfare level as under social planning. Yet, the corresponding output level, R&D investment for environmental friendly technologies and surplus distribution necessarily differ from the socially optimal ones.

Keywords: *environmental externality; stochastic shock; optimal taxation; differential game.*

Mathematics Subject Classification (2000): 91A23, 49N70, 91A80, 91B76.

1 Introduction

A lively debate is currently taking place on the need to preserve the environment from the negative consequences of pollution generated by industrial activities [⋆]. A crucial aspect is the lack of incentives on the part of profit-seeking firms to carry out R&D projects to generate new environmental-friendly technologies.

To this effect, policy makers may adopt several forms of regulation and taxation/subsidization policies to induce firms to internalize externalities and invest accordingly. The standard approach to this problem consists in introducing a Pigouvian tax or subsidy rule whereby firms pay or receive an amount of money proportional to the aggregate current stock of pollutants generated by the industry as a whole. We propose an alternative policy design, where the tax is levied on the marginal contribution to the

* Corresponding author: mailto:luca.lambertini@unibo.it

⋆ For an exhaustive account, see [12] and [6].

accumulation process followed by pollution. Accordingly, this policy is close in spirit to the adoption of a quality standard, such as the use of filters capturing CO_2, in order to decrease the amount of pollutants emitted by a car per mile.

We evaluate this perspective in a simple optimal control model where the market is monopolistic. To add a realistic feature to our framework, we allow for the presence of a stochastic shock affecting the accumulation of pollution, in such a way that the resulting optimal tax rate depends on the expected value of the shock. Our main result is that there exists a tax policy (i) inducing the firm to invest in R&D for a greener technology and (ii) yielding the same steady state social welfare as under social planning. However, the two allocations characterizing, respectively, the regulated monopoly and the first best differ under all remaining respects, i.e., price, output, R&D investment and surplus distribution.

The remainder of the paper is structured as follows. The model is laid out in Section 2. Section 3 contains the analysis of the regulated monopoly. The first best allocation is described in Section 4, while Section 5 investigates the optimal design of taxation. Concluding remarks are in Section 6.

2 The Setup

Consider a monopolistic single-product firm facing the instantaneous demand function $p(t) = a - q(t)$, where $a > 0$ is the reservation price and $q(t) \in [0, a - c]$ is the output level. The production cost is linear in $q(t)$ with unit cost $c \in (0, a)$. The production process involves a negative environmental externality $S(t)$, that accumulates according to the dynamics

$$\dot{S}(t) = b(t)q(t) - \frac{\delta S(t)}{\theta(t)}. \tag{1}$$

This evolutionary structure features a depreciation rate $\delta > 0$, which is also affected by a stochastic shock on its slope, in the form of a random variable $\theta(t)$, i.i.d. over time, with mean $E(\theta) = 1$ and variance $Var(\theta) = \sigma_\theta^2 > 1$. For future reference, we define the mean of the reciprocal as $E(\theta^{-1}) = w > 1$, by Jensen's inequality *. Our way of modelling uncertainty admittedly differs from the standard approach taken in the existing literature on stochastic differential games, where usually a Wiener process appears (for an overview, see [4], 2000, ch. 8). In place of a Wiener process, we consider the presence of a shock that, being i.i.d. across instants, allows us to take the necessary conditions on the expected value of the Hamiltonian function. Accordingly, by applying Pontryagin's maximum principle to the expected value of the Hamiltonian, one has that the mean and variance of the shock enter the necessary conditions as parameters. That is, this way of formalising the presence of uncertainty has no procedural bearings on the solution techniques needed to characterise the resulting steady state equilibrium. In particular, when deriving the control equations, we will see that these contain the parameters of the distribution of the shock, but not the shock itself as a function of time.

The assumption that the dynamics of the stock of pollution is subject to shocks has been introduced to capture the idea, largely discussed in the current debate on global warming and the anthropic responsibility in its evolution, that our knowledge of this matter is still incomplete and subject to natural factors beyond human control. In particular, our way of modelling (1) refers to uncertainty affecting measures of the rate

* For examples of analogous approaches in literature on industrial organization, see [8] or [10]. Systems of equations with random parameters have also been investigated by [11].

at which the atmosphere can absorb and eliminate CO_2-equivalent emissions, especially if one takes into account deforestation ⋆.

To create an incentive for the monopolist to invest in R&D so as to make its productive technology more environmental-friendly, the government imposes an instantaneous Pigouvian taxation. Usually, the Pigouvian tax is levied on the total amount of the externality (see [7], or [1],[2], *inter alia*). Here, we propose an alternative policy design, whereby the firm is subject to an instantaneous tax equal to $\tau b(t)$, i.e., what is being taxed is indeed the rate $b(t)$ at which a unit of final product contributes to the increase in the stock of pollution. The coefficient $b(t)$ is thus a further state variable whose dynamic equation is a linear one:

$$\dot{b}(t) = -k(t) + \eta b(t), \tag{2}$$

with $\eta > 0$, and decreasing in $k(t) \geq 0$, which is the instantaneous R&D effort carried out by the firm. A plausible economic interpretation of $b(t)$ is to see it as the environmental obsolescence rate of technology, measuring the growth rate of the external damage involved by the use of technologies that become increasingly more polluting as time goes by.

The R&D technology used by the firm involves an instantaneous cost measured by $\Gamma(t) = zk^2(t)$, where z is a positive constant. The problem for the monopolistic firm consists in maximizing w.r.t controls $k(t)$ and $q(t)$ the expected value of the following payoff functional:

$$J \equiv \int_0^\infty e^{-\rho t}((p(t) - c)q(t) - \Gamma(t) - \tau b(t))dt, \tag{3}$$

subject to:

$$\begin{cases} \dot{b}(t) = -k(t) + \eta b(t), \\ \dot{S}(t) = b(t)q(t) - \dfrac{\delta S(t)}{\theta(t)}, \\ b(0) = b_0 > 0, \\ S(0) = S_0 > 0. \end{cases} \tag{4}$$

This is a modified (monopolistic) version of a dynamic oligopoly game with environmental effects examined in [5]. The main differences consist in (i) the presence of a shock affecting the accumulation of the environmental externality; (ii) the functional form of the dynamics of $b(t)$, linear in this case; (iii) the tax is levied on the monopolist's contribution to pollution and not on the overall stock of pollution itself.

3 The Monopoly Optimum

The current value Hamiltonian function reads as:

$$H(\cdot) = (a - c - q(t))q(t) - zk^2(t) - \tau b(t) + \tag{5}$$

$$+\lambda(t)(-k(t) + \eta b(t)) + \mu(t)\left(b(t)q(t) - \frac{\delta S(t)}{\theta(t)}\right),$$

where $\lambda(t)$ is the costate variable associated with the state $b(t)$ and $\mu(t)$ is the one associated with the other state $S(t)$. Because of the aleatory effect, the monopolist is

⋆ According to the Department of Economics and Social Affairs of UNO, "ongoing deforestation accounts for about 8% of the world's annual carbon emissions" (see [3], 2009, p. 86).

supposed to maximize the expected value of the Hamiltonian $E(H)$. From now on, we will drop the time argument for brevity.

What follows is the list of the necessary conditions for the maximization of $E(H)$, adjoint equations and transversality conditions (an application of Pontryagin's Maximum Principle in a stochastic framework can be found in [9]) *:

$$\frac{\partial E(H)}{\partial k} = -2zk - \lambda = 0, \tag{6}$$

$$\frac{\partial E(H)}{\partial q} = a - c - 2q + \mu b = 0, \tag{7}$$

$$\dot{\lambda} = \rho\lambda - \frac{\partial E(H)}{\partial b} = (\rho - \eta)\lambda - \mu q + \tau, \tag{8}$$

$$\dot{\mu} = \rho\mu - \frac{\partial E(H)}{\partial S} = (\rho + \delta w)\,\mu, \tag{9}$$

$$\lim_{t\longrightarrow\infty} e^{-\rho t}\lambda(t) = 0, \qquad \lim_{t\longrightarrow\infty} e^{-\rho t}\mu(t) = 0. \tag{10}$$

Note that in (9) we make use of the expected value $E(\theta^{-1}) = w$. On this basis, we are going to write the control equations as well as the state equations in expected value. By differentiating (6) and (7) w.r.t. time, (4), (8) and (9) amount to the following state-control dynamical system:

$$\begin{cases} E\left(\dot{b}\right) = -k + \eta b, \\ E\left(\dot{S}\right) = bq - \delta w S, \\ E\left(\dot{k}\right) = (\rho - \eta)k - \dfrac{(a - c - 2q)q}{2zb} - \dfrac{\tau}{2z}, \\ E\left(\dot{q}\right) = \dfrac{1}{2}\left(\dfrac{-a + c + 2q}{b}\right)[-k + (\eta + \rho + \delta w)b]. \end{cases} \tag{11}$$

Proposition 3.1 *The model admits a unique steady state* $P^* = (b^*, S^*, k^*, q^*)$, *whose coordinates are, respectively,*

$$b^* = \frac{\tau}{2\eta(\rho - \eta)z}, \quad S^* = \frac{\tau(a - c)}{4\eta(\rho - \eta)z\delta w}, \quad k^* = \frac{\tau}{2(\rho - \eta)z}, \quad q^* = \frac{a - c}{2}.$$

Proof Solving (11) yields a unique stationary point P^*, whose coordinates are all strictly positive if $\rho - \eta > 0$.

Clearly, if $\rho \in [0, \eta)$, then τ must be negative in order for the vector (b^*, S^*, k^*, q^*) to be economically meaningful. In this range, the fact that discounting is lower than the environmental obsolescence rate entails that the only feasible policy takes the form of a subsidy. Conversely, for all $\rho > \eta$, τ must be positive, i.e., the regulator has to tax the firm to induce the entrepreneur to carry out a positive amount of R&D.

* We omit the explicit exposition of second order conditions for a maximum as they are satisfied by construction.

Note that, while k^* is a function of τ, q^* is not. This immediately implies that, by adopting this policy, the regulator is providing the firm with an incentive to carry out R&D while leaving unaffected the choice of the optimal monopoly output (and therefore the corresponding price level).

As a clear consequence of the dynamics of the model, the only coordinate affected by uncertainty is S^*, i.e., the steady state level of the pollution stock is a function of w. The Jacobian matrix of (11) evaluated at P^* is:

$$J(P^*) = \begin{pmatrix} \eta & 0 & -1 & 0 \\ \dfrac{a-c}{2} & -\delta w & 0 & \dfrac{\tau}{2\eta(\rho-\eta)z} \\ 0 & 0 & \rho-\eta & \dfrac{\eta(a-c)(\rho-\eta)}{\tau} \\ 0 & 0 & 0 & \rho+\delta w \end{pmatrix}.$$

Proposition 3.2 *P^* is a saddle point for the system (11).*

Proof $J(P^*)$ has the negative eigenvalue $\lambda_1 = -\delta w$, and the positive eigenvalues $\lambda_2 = \rho + \delta w$, $\lambda_3 = \eta$ and that is sufficient to deduce that P^* represents a saddle point equilibrium for (11).

Note that, if $\rho > \eta$, the stable subspace $E(P^*)$ is spanned by the vector $(0, 1, 0, 0)$, that is, on the S-axis the time trajectory of the stock of pollution asymptotically heads towards the level S^*.

Since the model features a single agent, there obviously exists a unique feedback stationary strategy coinciding with the open-loop solution. In particular, the related optimal value function $V(b, S)$ satisfying the Hamilton-Jacobi-Bellman equation is linear-quadratic in b and linear in S.

3.1 Welfare and profit assessment

Let $E(\pi^*)$, $E(CS^*)$ and $E(SW^*)$ be the profit, the consumer surplus and the social welfare functions evaluated at the steady state P^*. We have that:

$$E(\pi^*) = (a - c - q^*)q^* - z(k^*)^2 - \tau b^* = \frac{(a-c)^2}{4} - \frac{\tau^2(3\eta - 2\rho)}{4\eta(\rho-\eta)^2 z}, \tag{12}$$

independent of w.

Proposition 3.3 *1. If $\rho > \dfrac{3\eta}{2}$, then $E(\pi^*) > 0$ for every τ.*

2. If $0 < \rho < \dfrac{3\eta}{2}$, then $E(\pi^*) > 0$

$$\forall\ \tau \in \left(\min\left\{ \mp(a-c)(\rho-\eta)\sqrt{\frac{\eta z}{3\eta - 2\rho}} \right\}, \max\left\{ \mp(a-c)(\rho-\eta)\sqrt{\frac{\eta z}{3\eta - 2\rho}} \right\} \right).$$

Proof Trivially, the expression (12) is strictly positive irrespective of the value of τ if $\rho > \dfrac{3\eta}{2}$, whereas if $\rho < \dfrac{3\eta}{2}$, the positivity is ensured if τ belongs to the interval $(-\tau_1,\ \tau_1)$, where $\tau_1 = \max \pm(a-c)(\rho-\eta)\sqrt{\dfrac{\eta z}{3\eta - 2\rho}}$.

The consumer surplus at equilibrium reads:

$$E\left(CS^*\right) = \frac{(q^*)^2}{2} + \tau b^* = \frac{(a-c)^2}{8} + \frac{\tau^2}{2\eta(\rho-\eta)z}.$$

Finally, the social welfare at equilibrium follows:

$$E\left(SW^*\right) = E\left(CS^* - S^* + \pi^*\right) = \frac{3(a-c)^2}{8} - \frac{\tau^2}{4(\rho-\eta)^2 z} - \frac{\tau(a-c)}{4\eta(\rho-\eta)z\delta w}.$$

Proposition 3.4 *$E\left(SW^*\right) > 0$ for every τ belonging to the interval:*

$$\left(\min\left\{-\frac{(a-c)(\rho-\eta)}{2}\left(\frac{1}{\eta\delta w} \mp z\sqrt{\frac{1}{\eta^2 z^2\delta^2 w^2} + \frac{6}{z}}\right)\right\},\right.$$

$$\left.\max\left\{-\frac{(a-c)(\rho-\eta)}{2}\left(\frac{1}{\eta\delta w} \mp z\sqrt{\frac{1}{\eta^2 z^2\delta^2 w^2} + \frac{6}{z}}\right)\right\}\right).$$

Proof It suffices to solve the inequality $E(SW^*) > 0$ with respect to τ.

4 The First Best

Now we briefly expose the first best solution that would be attained if the firm were run by a benevolent planner maximizing the discounted flow of social welfare w.r.t. q and k. The planner's Hamiltonian is:

$$H_P(\cdot) = (a-c-q)q - \frac{q^2}{2} - S - zk^2 + \lambda(-k+\eta b) + \mu\left(bq - \frac{\delta S}{\theta}\right). \tag{13}$$

Taking the necessary conditions on the expected value of $H_P(\cdot)$ and following the same procedure as in the previous section, we obtain the following steady state coordinates (the subscript P stands for *planner*):

$$\begin{aligned} b_P &= \frac{(a-c)(\rho+\delta w)}{1+2\eta z(\rho+\delta w)^2(\rho-\eta)}, \\ S_P &= \frac{2(a-c)^2\eta(\rho-\eta)(\rho+\delta w)^3 z}{\delta w\left[1+2\eta z(\rho+\delta w)^2(\rho-\eta)\right]^2}, \\ k_P &= \frac{(a-c)(\rho+\delta w)\eta}{1+2\eta z(\rho+\delta w)^2(\rho-\eta)}, \\ q_P &= \frac{2(a-c)\eta(\rho-\eta)(\rho+\delta w)^2 z}{1+2\eta z(\rho+\delta w)^2(\rho-\eta)}. \end{aligned} \tag{14}$$

Note that all of these coordinates are affected by the shock. The associated profits and consumer surplus are:

$$E\left(\pi_P\right) = \frac{(a-c)^2\eta(2\rho-3\eta)(\rho+\delta w)^2 z}{\left[1+2\eta z(\rho+\delta w)^2(\rho-\eta)\right]^2}, \tag{15}$$

$$E\left(CS_P\right) = \frac{1}{2}\left[\frac{2\left(a-c\right)\eta\left(\rho-\eta\right)\left(\rho+\delta w\right)^2 z}{1+2\eta z\left(\rho+\delta w\right)^2\left(\rho-\eta\right)}\right]^2. \tag{16}$$

Hence, the resulting social welfare level is $E\left(SW_P\right) = E\left(\pi_P + CS_P - S_P\right)$. Before proceeding any further, it is worth stressing the following straightforward result:

Proposition 4.1 *If* $\rho > \dfrac{3\eta}{2}$, *then* $\pi_P > 0$.

We are now in a position to address the following question, i.e., whether the policy maker regulating the behaviour of a profit-maximizing monopolist can design an optimal tax rate τ so as to replicate the same welfare performance associated to the first best allocation we have just sketched. This must be done under the non-negativity constraint concerning the firm's profits, as established in Proposition 3.3.1. In doing so, we shall confine our attention to the parameter range $\rho > 3\eta/2$, in order for the planning equilibrium to be sustainable under our partial equilibrium approach, i.e., in absence of any other industrial sector that could be taxed to raise the money necessary for the survival of the public monopoly for all $\rho \in [0, 3\eta/2)$.

5 Designing the Optimal Taxation

The policy maker's problem consists in solving

$$\Delta E\left(SW\right) = E\left(SW_P - SW^*\right) = 0 \tag{17}$$

w.r.t. τ, with

$$\Delta E\left(SW\right) = \frac{1}{8}\left[\frac{2\tau^2}{\left(\rho-\eta\right)^2 z} + \frac{2\left(a-c\right)\tau}{\delta\eta\left(\rho-\eta\right)wz} - 3\left(a-c\right)^2 + \Psi\right], \tag{18}$$

where

$$\Psi \equiv \frac{8\eta\left(a-c\right)^2\left(\rho+\delta w\right)^2 z\left[2\delta\eta\left(\rho-\eta\right)^2 w\left(\rho+\delta w\right)^2 z - 2\rho\left(\rho-\eta\right) - \delta\eta w\right]}{\delta\left[1+2\eta z\left(\rho+\delta w\right)^2\left(\rho-\eta\right)\right]w}. \tag{19}$$

Equation (17) has two real roots in τ, $\tau_- < 0 < \tau_+$ [⋆]. On this basis, we can state our final result:

Proposition 5.1 *For all* $\rho > \dfrac{3\eta}{2}$, *there exists a tax* (τ_+) *allowing the policy maker to replicate at the monopoly equilibrium the social welfare performance associated with the first best.*

The negative solution must be discarded in view of Proposition 4.1. As a last remark, again recollecting Proposition 3.1, it is worth pointing out that such a policy can only reproduce the aggregate surplus created by this industry, while the output and the R&D effort will necessarily differ across regimes. To see this, it's sufficient to compare q^* against q_P: while the former is constant (and coincides with the standard output that we usually observe in a monopoly equilibrium with the same demand and cost functions), the latter clearly accounts for the shock affecting the accumulation of pollution. Additionally, one may observe that $k^* = k_P$ obtains in correspondence of a value of τ that does not solve (17).

[⋆] We omit the lengthy expressions of the two roots for the sake of brevity.

6 Conclusion

In a dynamic monopoly model with environmental externalities, we have investigated the possibility of using a tax taylored on the firm's instantaneous contribution to the accumulation of pollution, which is subject to a shock, the latter being i.i.d. across instants. There exists an optimal tax rate such that the industry exactly replicates the same steady state welfare performance as in the first best. However, the corresponding expected values of output level, R&D investment for green technologies and surplus distribution necessarily differ from those characterising social planning.

An interesting extension of the foregoing analysis is the design of the same kind of policy in an oligopoly game where each single firm might refrain from investing in environmental friendly technologies due to the usual free riding incentive usually associated with strategic interplay. This is left for future research.

Acknowledgment

We thank B.S. Goh (Editor) and two anonymous referees for helpful comments and suggestions. The usual disclaimer applies.

References

[1] Benchekroun, H. and Long, N.V. Efficiency inducing taxation for polluting oligopolists. *Journal of Public Economics* **70** (1998) 325–342.

[2] Benchekroun, H. and Long, N.V. On the multiplicity of efficiency-inducing tax rules. *Economics Letters* **76** (2002) 331–336.

[3] DESA *Promoting development, saving the planet. World economic and social survey 2009.* Department of Economic and Social Affairs, United Nations Organization, New York, 2009.

[4] Dockner, E. J., Jørgensen, S., Long, N.V. and Sorger, G. *Differential games in economics and management science*. Cambridge University Press, Cambridge, 2000.

[5] Dragone, D., Lambertini, L. and Palestini, A. The incentive to invest in environmental-friendly technologies: dynamics makes a difference. *DSE Working Paper* (658) (2009) http://www2.dse.unibo.it/wp/658.pdf.

[6] Jørgensen, S., Martin-Herran, G. and Zaccour, G. Dynamic games in the economics and management of pollution. Mimeo, GERAD, Montreal, 2009.

[7] Karp, L. and Livernois, J. Using automatic tax changes to control pollution emissions. *Journal of Environmental Economics and Management* **27** (1994) 38–48.

[8] Klemperer, P. and Meyer, M. Price competition vs quantity competition: the role of uncertainty. *RAND Journal of Economics* **17** (1986) 618–638.

[9] Lambertini, L. Stackelberg leadership in a dynamic duopoly with stochastic capital accumulation. *Journal of Evolutionary Economics* **15** (2005) 443–465.

[10] Lambertini, L. Process R&D in monopoly under demand uncertainty. *Economics Bulletin* **15** (2006) 1–9.

[11] Masterkov, Yu. V. and Rodina, L. I. The sufficient conditions of local controllability for linear systems with random parameters. *Nonlinear Dynamics and Systems Theory* **7**(3) (2007) 303–314.

[12] Stern, N. *The economics of climate change: the Stern review*. Cambridge University Press, Cambridge, 2007.

Nonlinear Dynamics and Systems Theory, **10** (2) (2010) 125–160

Gradient Transformation Trajectory Following Algorithms for Equality-Constrained Minimization †

W. J. Grantham *

School of Mechanical and Materials Engineering
Washington State University
Pullman, Washington 99164-2920
United States of America

Received: November 15, 2009; Revised: March 25, 2010

Abstract: For minimizing a scalar-valued function subject to equality constraints, we develop and investigate a family of gradient transformation differential equation algorithms. This family includes, as special cases: Min-Max Ascent, Hestenes' Method of Multipliers, Newton's method, and a Gradient Enhanced Min-Max (GEMM) algorithm that we extend to handle equality constraints. We apply these methods to Rosenbrock's function with a parabolic constraint. We show that Min-Max Ascent is locally and (experimentally) globally asymptotically stable but extremely stiff and has extremely slow convergence. Hestenes' Method of Multipliers is also locally and (experimentally) globally asymptotically stable and has faster convergence, but is still very stiff. Newton's method is not stiff, but does not yield global asymptotic stability. However, GEMM is both globally asymptotically stable and not stiff. We study the stiffness of the gradient transformation family in terms of Lyapunov exponent time histories. Starting from points where all the methods in this paper do work, we show that Min-Max Ascent and Hestenes' Method of Multipliers are very stiff and slow to converge, but with the Method of Multipliers being approximately 2 times as fast as Min-Max Ascent. Newton's method is not stiff and is approximately 900 times as fast as Min-Max Ascent and 400 times as fast as the Method of Multipliers. In contrast, the Gradient Enhanced Min-Max method is globally convergent, is not stiff, and is approximately 100 times faster than Newton's method, 40,000 times faster than the Method of Multipliers, and 90,000 times faster than Min-Max Ascent.

Keywords: *nonlinear programming, Lagrangian Min-Max, stiff differential equations, Lyapunov exponents.*

Mathematics Subject Classification (2000): 90C30, 90C47, 70K70, 34D08.

† In Memorium: Thomas Lange Vincent, Doctor of Philosophy, Professor Emeritus, The University of Arizona: Born, September 16, 1935; Died, October 26, 2009.

* Corresponding author: mailto:grantham@wsu.edu

1 Minimization with Equality Constraints

We consider the nonlinear programming problem of finding a point $\mathbf{x}^* \in \mathcal{R}^n$ to

$$\min \phi(\mathbf{x}) \quad \text{subject to} \quad \boldsymbol{\psi}(\mathbf{x}) = \mathbf{0}, \tag{1}$$

where the functions $\phi(\cdot) : \mathcal{R}^n \to \mathcal{R}^1$ and $\boldsymbol{\psi}(\cdot) : \mathcal{R}^n \to \mathcal{R}^m$ are $\mathcal{C}^2$. We develop differential equation algorithms of the form

$$\dot{\mathbf{x}} = \mathbf{g}(\mathbf{x}),$$

where $\dot{(\cdot)}$ denotes $\mathrm{d}()/\mathrm{d}t$ and t is time. We choose the function $\mathbf{g}(\cdot)$ with the objective of having solutions $\mathbf{x}(t) \to \mathbf{x}^*$ as $t \to \infty$. Such "trajectory following" algorithms have received considerable attention in recent years. In [1] Steepest Descent differential equations are used to design controllers for nonlinear systems. In [2] optimal control differential equations are used to design new discrete minimization algorithms. In [3] and [4] differential equation algorithms are investigated for min-max optimization problems. In [5] and [6] differential equations for Newton's method are used to find all of the stationary points of a function. In [7] a Gradient Enhanced Newton algorithm is developed for finding a stationary proper minimum point. In [8] a Gradient Enhanced Min-Max method is developed for finding a proper stationary min-max saddle point.

In this paper, we extend the stationary min and min-max results of [7] and [8] to include equality constraints. As in these previous papers, we are concerned with differential equation-based algorithms, and with the stiffness and domain of stability of a family of gradient-based numerical update algorithms. These algorithms include, as special cases, Steepest Descent, Min-Max Ascent, Newton's Method, augmented Lagrangians and Hestenes' Method of Multipliers, and the Gradient Enhanced Min-Max algorithm that we extend here for minimization subject to equality constraints. We use Lyapunov exponents to measure the stiffness (*e.g.,* widely separated time scales and eigenvalues) of the various algorithms when applied to an equality constrained version of Rosenbrock's "banana" function.

2 Necessary Conditions at a Minimum Point

The necessary conditions for $\mathbf{x}^* \in \mathcal{R}^n$ to yield a regular [9, p. 35] local minimum can be expressed in terms of the Lagrangian

$$L(\mathbf{x}, \boldsymbol{\lambda}) \triangleq \phi(\mathbf{x}) - \boldsymbol{\lambda}^\mathsf{T} \boldsymbol{\psi}(\mathbf{x}), \tag{2}$$

where $\boldsymbol{\lambda} \in \mathcal{R}^m$ is a vector of Lagrange multipliers and $\boldsymbol{\psi}(\cdot) : \mathcal{R}^n \to \mathcal{R}^m$ represents a system of $m < n$ equality constraints that must be satisfied at $\mathbf{x}^*$.

The first-order Karush–Kuhn–Tucker necessary conditions [9, p.57] are that:

$$\mathbf{0}^\mathsf{T} = \frac{\partial L}{\partial \mathbf{x}} = \frac{\partial \phi}{\partial \mathbf{x}} - \boldsymbol{\lambda}^\mathsf{T} \frac{\partial \boldsymbol{\psi}}{\partial \mathbf{x}}, \quad \mathbf{0}^\mathsf{T} = \frac{\partial L}{\partial \boldsymbol{\lambda}} = -\boldsymbol{\psi}^\mathsf{T}(\mathbf{x}), \tag{3}$$

where $\partial L/\partial \mathbf{x} \triangleq [\partial L/\partial x_1, \ldots, \partial L/\partial x_n]$. Since $\partial L/\partial \boldsymbol{\lambda} = -\boldsymbol{\psi}^\mathsf{T}$ the necessary conditions can be written in terms of $\mathbf{y}^\mathsf{T} = [\mathbf{x}^\mathsf{T}, \boldsymbol{\lambda}^\mathsf{T}] \in \mathcal{R}^p$, $p = n + m$, as

$$\nabla_{\mathbf{y}} L(\mathbf{y}) \triangleq \left[\frac{\partial L}{\partial \mathbf{y}}\right]^\mathsf{T} = \mathbf{0}, \tag{4}$$

that is,

$$\nabla_{\mathbf{x}} L = \left[\frac{\partial \phi}{\partial \mathbf{x}}\right]^{\mathsf{T}} - \left[\frac{\partial \boldsymbol{\psi}}{\partial \mathbf{x}}\right]^{\mathsf{T}} \boldsymbol{\lambda} = \mathbf{0}, \quad \nabla_{\boldsymbol{\lambda}} L = \left[\frac{\partial L}{\partial \boldsymbol{\lambda}}\right]^{\mathsf{T}} = -\boldsymbol{\psi} = \mathbf{0}.$$

The necessary conditions (4) are stationarity conditions, yielding candidates that may be local minima, maxima, or saddle points. Suppose that the constraint qualification conditions [9, p. 55] hold: that at $\mathbf{x}^*$ there exists a nonzero vector $\boldsymbol{\eta} \in \mathcal{R}^n$ tangent to the constraints $\boldsymbol{\psi}(\mathbf{x}^*) = \mathbf{0}$. Then the second-order necessary condition [9, p. 56] for a regular local minimum point is that

$$\boldsymbol{\eta}^{\mathsf{T}} H(\mathbf{x}^*, \boldsymbol{\lambda}^*) \boldsymbol{\eta} \geq 0 \quad \text{for all nonzero } \boldsymbol{\eta} \text{ such that } \frac{\partial \boldsymbol{\psi}(\mathbf{x}^*)}{\partial \mathbf{x}} \boldsymbol{\eta} = \mathbf{0}, \tag{5}$$

where $H(\mathbf{x}, \boldsymbol{\lambda}) = \partial^2 L(\mathbf{x}, \boldsymbol{\lambda}) / \partial \mathbf{x}^2$. A second-order sufficient condition is that $\nabla_{\mathbf{y}} L(\mathbf{y}) = [\partial L(\mathbf{y}^*)/\partial \mathbf{y}]^{\mathsf{T}} = \mathbf{0}$ and

$$\boldsymbol{\eta}^{\mathsf{T}} H(\mathbf{x}^*, \boldsymbol{\lambda}^*) \boldsymbol{\eta} > 0 \quad \text{for all nonzero } \boldsymbol{\eta} \text{ such that } \frac{\partial \boldsymbol{\psi}(\mathbf{x}^*)}{\partial \mathbf{x}} \boldsymbol{\eta} = \mathbf{0}, \tag{6}$$

which would be satisfied, for example, by the stronger condition that $H(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ be positive definite.

3 Numerical Minimization Methods

Numerical minimization methods [10] generally seek a search direction $\mathbf{s}$ and a step size α for a move $\mathbf{x} \leftarrow \mathbf{x} + \alpha \mathbf{s}$. Here, we focus on the instantaneous search direction, using a differential step size with continuous updating of the search direction. Thus we develop "trajectory following" algorithms of the form $\mathrm{d}\mathbf{x}/\mathrm{d}t = \mathbf{g}(\mathbf{x})$. Such differential equations-based algorithms have been very useful in developing new discrete optimization algorithms [2] based on long-term optimal control algorithms. In addition, using a differential step size avoids difficulties such as "chatter" that can occur with discrete step size algorithms such as Steepest Descent applied, for example, to Rosenbrock's function [7].

3.1 Unconstrained minimization

3.1.1 Steepest descent

The simplest algorithm for minimizing an unconstrained function $\phi(\mathbf{x})$ is the Steepest Descent algorithm

$$\dot{\mathbf{x}} = -\boldsymbol{\nabla} \phi,$$

with $\boldsymbol{\nabla} \phi \triangleq [\partial \phi / \partial \mathbf{x}]^{\mathsf{T}}$, which yields

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = \frac{\partial \phi}{\partial \mathbf{x}} \dot{\mathbf{x}} = -\left\| \boldsymbol{\nabla} \phi \right\|^2,$$

where $\|\cdot\|$ denotes the Euclidian norm. If $\mathbf{x}^*$ is a local minimal point for $\phi(\mathbf{x})$ then $V(\mathbf{x}) = \phi(\mathbf{x}) - \phi(\mathbf{x}^*)$ is a local Lyapunov function, establishing that Steepest Descent is at least locally asymptotically stable at a proper local minimum. In addition, if $\mathbf{x}^*$ is unique and $\|\boldsymbol{\nabla} \phi(\mathbf{x})\| \to \infty$ as $\|\mathbf{x}\| \to \infty$ then the minimal point $\mathbf{x}^*$ is globally asymptotically stable.

Steepest Descent may produce stiff systems. Such systems require much more complicated differential equation solvers than Euler's method or Runge–Kutta methods, leading to complicated discrete versions.

3.1.2 Newton's method

From Taylor's theorem applied to the stationarity necessary condition

$$\boldsymbol{\nabla}\phi = \mathbf{0},$$

we get

$$\boldsymbol{\nabla}\phi(\mathbf{x}+\Delta\mathbf{x}) = \boldsymbol{\nabla}\phi(\mathbf{x}) + \boldsymbol{\nabla}^2\phi(\mathbf{x})\Delta\mathbf{x} + O(\|\Delta\mathbf{x}\|^2), \tag{7}$$

where $\boldsymbol{\nabla}^2\phi \triangleq \partial^2\phi/\partial\mathbf{x}^2$ is the Hessian matrix, $\Delta\mathbf{x} = \dot{\mathbf{x}}\Delta t + O(\Delta t^2)$, and $O(\alpha^2)/\alpha \to 0$ as $\alpha \to 0$. Setting the left-hand side equal to zero yields the discrete-time ($\Delta t = 1$, small $\|\Delta\mathbf{x}\|$) version of Newton's method: $\Delta\mathbf{x} = -\left[\boldsymbol{\nabla}^2\phi\right]^{-1}\boldsymbol{\nabla}\phi$.

In the limit as $\Delta t \to 0$, the continuous-time Newton method is given by

$$\dot{\mathbf{x}} = -\left[\boldsymbol{\nabla}^2\phi\right]^{-1}\boldsymbol{\nabla}\phi. \tag{8}$$

The discrete-time version of Newton's method corresponds to applying Euler integration $\Delta\mathbf{x} = \dot{\mathbf{x}}\Delta t$ to (8) with $\Delta t = 1$.

Note that Newton's method is only well defined in a region where the determinant $\left|\boldsymbol{\nabla}^2\phi(\mathbf{x})\right|$ does not change sign and is nonzero, such as some neighborhood of a proper local minimal point $\mathbf{x}^*$, at which $\boldsymbol{\nabla}^2\phi(\mathbf{x}^*) > 0$ (positive definite). Newton's method, in regions where it does work, typically converges much faster than Steepest Descent, and yields non-stiff systems. In particular, in terms of the gradient $\boldsymbol{\nabla}\phi\left[\mathbf{x}\left(t\right)\right]$ along $\mathbf{x}(t)$, Newton's method (8) yields

$$\frac{\mathrm{d}\boldsymbol{\nabla}\phi}{\mathrm{d}t} = \left[\boldsymbol{\nabla}^2\phi\right]\dot{\mathbf{x}} = -\boldsymbol{\nabla}\phi,$$

which is non stiff, with eigenvalues $\mu_k = -1$, $k = 1, \ldots, n$. Note that, along $\mathbf{x}(t)$ we have $\boldsymbol{\nabla}\phi\left[\mathbf{x}\left(t\right)\right] = \boldsymbol{\nabla}\phi\left[\mathbf{x}\left(0\right)\right]\mathrm{e}^{-t} \to \mathbf{0}$ as $t \to \infty$, hence $\boldsymbol{\nabla}\phi\left[\mathbf{x}(t)\right] \to \mathbf{0}$. As with Steepest Descent, Newton's method is at least locally asymptotically stable to a proper local minimal point.

3.2 Constrained minimization

3.2.1 Penalty functions

The earliest approach to handling equality constraints $\boldsymbol{\psi}(\mathbf{x}) = \mathbf{0}$ was to apply unconstrained minimization to a penalty function such as Courant's penalty function

$$\pi(\mathbf{x},\beta) = \phi(\mathbf{x}) + \frac{1}{2}\beta\left\|\boldsymbol{\psi}(\mathbf{x})\right\|^2 = \phi(\mathbf{x}) + \frac{1}{2}\beta\boldsymbol{\psi}^{\intercal}(\mathbf{x})\boldsymbol{\psi}(\mathbf{x}), \tag{9}$$

with a sequence of increasing values for $\beta > 0$. Then Steepest Descent yields

$$\dot{\mathbf{x}} = -\boldsymbol{\nabla}\pi = -\boldsymbol{\nabla}\phi(\mathbf{x}) - \beta\boldsymbol{\Gamma}^{\intercal}(\mathbf{x})\boldsymbol{\psi}(\mathbf{x}),$$

where $\boldsymbol{\Gamma} \in \mathcal{R}^{m\times n}$ is given by

$$\boldsymbol{\Gamma}(\mathbf{x}) = \partial\boldsymbol{\psi}(\mathbf{x})/\partial\mathbf{x}. \tag{10}$$

The main difficulty with this approach is that, for any finite $\beta > 0$, the point that minimizes $\pi(\mathbf{x},\beta)$ is not exactly the same point that minimizes $\phi(\mathbf{x})$ subject to $\boldsymbol{\psi}(\mathbf{x}) = \mathbf{0}$, except in the limit as $\beta \to \infty$. In addition, large values of β yield stiff systems. Note that Newton's method applied to (9) may alleviate the stiffness, but not the "mismatch" between the two minimization solutions, which requires $\beta \to \infty$.

3.2.2 Newton's method

The first-order necessary conditions for a stationary point of $\phi(\mathbf{x})$ subject to $\boldsymbol{\psi}(\mathbf{x}) = \mathbf{0}$ are given by (3) in terms of a Lagrange multiplier vector. Newtons method, applied to (4), is given by

$$\begin{bmatrix} \dot{\mathbf{x}} \\ \dot{\boldsymbol{\lambda}} \end{bmatrix} = -\mathbf{H}^{-1}(\mathbf{y}) \begin{bmatrix} \boldsymbol{\nabla}_{\mathbf{x}} L \\ -\boldsymbol{\psi}(\mathbf{x}) \end{bmatrix}, \tag{11}$$

where

$$H(\mathbf{y}) = \boldsymbol{\nabla}^2_{\mathbf{y}} L \triangleq \frac{\partial^2 L}{\partial \mathbf{y}^2} = \begin{bmatrix} \dfrac{\partial^2 L}{\partial \mathbf{x}^2} & \dfrac{\partial^2 L}{\partial \boldsymbol{\lambda} \partial \mathbf{x}} \\ \dfrac{\partial^2 L}{\partial \mathbf{x} \partial \boldsymbol{\lambda}} & \dfrac{\partial^2 L}{\partial \boldsymbol{\lambda}^2} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial^2 L}{\partial \mathbf{x}^2} & -\boldsymbol{\Gamma}^{\mathsf{T}} \\ -\boldsymbol{\Gamma} & \mathbf{0} \end{bmatrix}. \tag{12}$$

3.2.3 Min-Max Lagrangians

Consider the Lagrangian

$$L(\mathbf{x}, \boldsymbol{\lambda}) = \phi(\mathbf{x}) - \boldsymbol{\lambda}^{\mathsf{T}} \boldsymbol{\psi}(\mathbf{x}).$$

Let $\mathbf{x}(\boldsymbol{\lambda})$ denote the unconstrained minimizer for $L(\mathbf{x}, \boldsymbol{\lambda})$, and let $\mathbf{x}^*$ and $\boldsymbol{\lambda}^*$ be the solution and Lagrange multiplier, respectively, for the constrained minimization problem (1). Then $L(\mathbf{x}(\boldsymbol{\lambda}), \boldsymbol{\lambda}) \leq L(\mathbf{x}, \boldsymbol{\lambda})\ \forall\ \mathbf{x}$, along with $\boldsymbol{\psi}(\mathbf{x}^*) = \mathbf{0}$, yields

$$\begin{aligned} L(\mathbf{x}(\boldsymbol{\lambda}), \boldsymbol{\lambda}) &\leq L(\mathbf{x}^*, \boldsymbol{\lambda}) = \phi(\mathbf{x}^*) - \boldsymbol{\lambda}^{\mathsf{T}} \boldsymbol{\psi}(\mathbf{x}^*) \\ &= \phi(\mathbf{x}^*) = \phi(\mathbf{x}^*) - \boldsymbol{\lambda}^{*\mathsf{T}} \boldsymbol{\psi}(\mathbf{x}^*) \\ &= L(\mathbf{x}^*, \boldsymbol{\lambda}^*) = L(\mathbf{x}\,(\boldsymbol{\lambda}^*), \boldsymbol{\lambda}^*). \end{aligned}$$

Thus

$$L(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \max_{\boldsymbol{\lambda}} \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}) = \min_{\mathbf{x}} \max_{\boldsymbol{\lambda}} L(\mathbf{x}, \boldsymbol{\lambda}), \tag{13}$$

since $L(\mathbf{x}, \boldsymbol{\lambda})$ is linear in $\boldsymbol{\lambda}$.

A Min-Max Ascent algorithm [3] for achieving the Lagrangian saddle point defined by (13) is given by

$$\dot{\mathbf{x}} = -\boldsymbol{\nabla}_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}) = -\boldsymbol{\nabla} \phi(\mathbf{x}) + \boldsymbol{\Gamma}^{\mathsf{T}}(\mathbf{x})\, \boldsymbol{\lambda}, \tag{14}$$

$$\dot{\boldsymbol{\lambda}} = \quad \boldsymbol{\nabla}_{\boldsymbol{\lambda}} L(\mathbf{x}, \boldsymbol{\lambda}) = -\boldsymbol{\psi}(\mathbf{x}). \tag{15}$$

As noted in [11], methods such as this, where $\mathbf{x}^*$ solves the primal problem (1) and $\boldsymbol{\lambda}(\mathbf{u})$ is the Lagrange multiplier vector for an associated dual problem [12, p. 113]:

$$\min_{\substack{\mathbf{u} \in \mathcal{N}_{\mathbf{u}} \\ \mathbf{x}(\mathbf{u}) \in \mathcal{N}_{\mathbf{x}}}} \phi[\mathbf{x}(\mathbf{u})] \qquad \text{subject to} \qquad \boldsymbol{\psi}\,[\mathbf{x}(\mathbf{u})] = \mathbf{u} \tag{16}$$

with $\boldsymbol{\lambda}(\mathbf{0}) = \boldsymbol{\lambda}^*$ and $\mathcal{N}_{\mathbf{u}} \subset \mathcal{R}^m$ and $\mathcal{N}_{\mathbf{x}} \subset \mathcal{R}^n$ being small neighborhoods of $\mathbf{u}^* = \mathbf{0}$ and $\mathbf{x}^*$, respectively, have "*... serious disadvantages. First, problem* (1) *must have a locally convex structure in order for the dual problem* (16) *to be well defined and* (15) *to be meaningful. Second, ..., the ascent iteration* (15) *converges only moderately fast.*"

3.2.4 Augmented Lagrangians

The following results apply to the equality constrained problem (1), but can be extended to the general nonlinear programming problem with equality and inequality constraints.

Consider an augmented Lagrangian [13]

$$\mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\beta}) \quad \triangleq \; L(\mathbf{x},\boldsymbol{\lambda}) + \tfrac{1}{2}\boldsymbol{\psi}(\mathbf{x})^{\mathsf{T}}\mathbf{S}\boldsymbol{\psi}(\mathbf{x}) = \phi(\mathbf{x}) + \tfrac{1}{2}\boldsymbol{\psi}(\mathbf{x})^{\mathsf{T}}\mathbf{S}\boldsymbol{\psi}(\mathbf{x}) - \boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{\psi}(\mathbf{x}), \tag{17}$$

where $\boldsymbol{\beta} \in \mathcal{R}^m$, $\boldsymbol{\beta} \geq \mathbf{0}$, and $\mathbf{S} = \operatorname{diag}[\boldsymbol{\beta}] \in \mathcal{R}^{m\times m}$. The augmented Lagrangian (17) can be viewed either as 1) the Lagrangian plus a penalty term or 2) the Lagrangian for minimizing a weighted Courant penalty function (9) subject to $\boldsymbol{\psi}(\mathbf{x}) = \mathbf{0}$. In the first case view, since a purely $\mathbf{x}$-dependent penalty term has been added to $L(\mathbf{x},\boldsymbol{\lambda})$, we expect that in changing to a $\max_{\boldsymbol{\lambda}} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x},\boldsymbol{\lambda})$ *vs.* $\max_{\boldsymbol{\lambda}} \min_{\mathbf{x}} L(\mathbf{x},\boldsymbol{\lambda})$ process, $\boldsymbol{\lambda}$ has no effect, but $\mathbf{x}$ affords a trade-off between the $\mathbf{x}$ that minimizes $L(\mathbf{x},\boldsymbol{\lambda})$ and the $\mathbf{x}$ that minimizes $\mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\beta})$. However, the penalty weights β_i on the ψ_i do not need to approach infinity for the two solutions to be the same and can be quite moderate in size. We have:

Theorem 3.1 *If second-order sufficient conditions* (6) *hold at* $(\mathbf{x}^*,\boldsymbol{\lambda}^*)$ *then there exists* $\boldsymbol{\beta}' \geq \mathbf{0}$ *such that for any* $\boldsymbol{\beta} > \boldsymbol{\beta}'$, $\mathbf{x}^*$ *is an isolated local minimizer of* $\mathcal{L}(\mathbf{x},\boldsymbol{\lambda}^*,\boldsymbol{\beta})$, *that is,* $\mathbf{x}^* = \mathbf{x}(\boldsymbol{\lambda}^*)$. *Furthermore,* $\boldsymbol{\lambda}^*$ *is a local maximizer of* $\nu(\boldsymbol{\lambda}) \triangleq \mathcal{L}(\mathbf{x}(\boldsymbol{\lambda}),\boldsymbol{\lambda},\boldsymbol{\beta})$.

Proof [10, pp. 289–291].

Hereafter we consider the case where $\mathbf{S} = \beta\mathbf{I}_m$ and drop the $\boldsymbol{\beta}$ argument in $\mathcal{L}(\cdot)$ unless it is expressly needed for the discussion.

4 Gradient Transformation Trajectory Following

From Theorem 3.1 we seek a $\max_{\boldsymbol{\lambda}} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x},\boldsymbol{\lambda})$. We consider the class of *Gradient Transformation algorithms*, of the form

$$\dot{\mathbf{y}} = -\mathbf{P}(\mathbf{y})\boldsymbol{\nabla}_{\mathbf{y}}\mathcal{L}(\mathbf{y}), \tag{18}$$

where $\mathbf{P}(\mathbf{y}) \in \mathcal{R}^{p\times p}$ is a Gradient Transformation matrix to be chosen, $\mathbf{y} \in \mathcal{R}^p$ with $\mathbf{y}^{\mathsf{T}} = [\mathbf{x}^{\mathsf{T}},\boldsymbol{\lambda}^{\mathsf{T}}]$, and

$$\mathbf{h}(\mathbf{y}) = \boldsymbol{\nabla}_{\mathbf{y}}\mathcal{L}(\mathbf{y}) = \begin{bmatrix}\boldsymbol{\nabla}_{\mathbf{x}}\mathcal{L}\\ \boldsymbol{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\nabla}_{\mathbf{x}}L + \beta\boldsymbol{\Gamma}^{\mathsf{T}}\boldsymbol{\psi}\\ \boldsymbol{\nabla}_{\boldsymbol{\lambda}}L\end{bmatrix} = \begin{bmatrix}\boldsymbol{\nabla}\phi - \boldsymbol{\Gamma}^{\mathsf{T}}[\boldsymbol{\lambda} - \beta\boldsymbol{\psi}]\\ -\boldsymbol{\psi}\end{bmatrix}. \tag{19}$$

Thus the Gradient Transformation algorithms are of the form

$$\begin{bmatrix}\dot{\mathbf{x}}\\ \dot{\boldsymbol{\lambda}}\end{bmatrix} = -\begin{bmatrix}\mathbf{P}_{\mathbf{xx}} & \mathbf{P}_{\mathbf{x}\boldsymbol{\lambda}}\\ \mathbf{P}_{\boldsymbol{\lambda}\mathbf{x}} & \mathbf{P}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}\end{bmatrix}\begin{bmatrix}\boldsymbol{\nabla}_{\mathbf{x}}\mathcal{L}\\ \boldsymbol{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}\end{bmatrix} = \begin{bmatrix}\mathbf{P}_{\mathbf{xx}} & \mathbf{P}_{\mathbf{x}\boldsymbol{\lambda}}\\ \mathbf{P}_{\boldsymbol{\lambda}\mathbf{x}} & \mathbf{P}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}\end{bmatrix}\begin{bmatrix}-\boldsymbol{\nabla}\phi + \boldsymbol{\Gamma}^{\mathsf{T}}[\boldsymbol{\lambda} - \beta\boldsymbol{\psi}]\\ \boldsymbol{\psi}\end{bmatrix}. \tag{20}$$

If $\mathbf{P}(\mathbf{y})$ is nonsingular in a region $\mathcal{R} \subseteq \mathcal{R}^p$ containing $\mathbf{y}^* = (\mathbf{x}^*,\boldsymbol{\lambda}^*)$ then for (18) the only equilibrium points in $\mathcal{R}$ are where $\boldsymbol{\nabla}_{\mathbf{y}}\mathcal{L}(\mathbf{y}^*) = \mathbf{0}$. We will be concerned with the uniqueness and local and global stability of the (possibly multiple) equilibria and with the "stiffness" of the resulting system, corresponding to various choices for $\mathbf{P}(\mathbf{y})$.

4.1 Min-Max ascent

The original trajectory following method [3] for seeking $\max_{\boldsymbol{\lambda}} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$ is via steepest descent for $\mathbf{x}$ and steepest ascent for $\boldsymbol{\lambda}$, yielding

$$\begin{array}{rcccccl} \dot{\mathbf{x}} & = & -\boldsymbol{\nabla}_{\mathbf{x}}\mathcal{L} & = & -\left[\boldsymbol{\nabla}_{\mathbf{x}} L + \beta \boldsymbol{\Gamma}^{\intercal}\boldsymbol{\psi}\right] & = & -\nabla_{\mathbf{x}}\phi + \boldsymbol{\Gamma}^{\intercal}\left[\boldsymbol{\lambda} - \beta\boldsymbol{\psi}\right], \\ \dot{\boldsymbol{\lambda}} & = & \boldsymbol{\nabla}_{\boldsymbol{\lambda}}\mathcal{L} & = & \boldsymbol{\nabla}_{\boldsymbol{\lambda}} L & = & -\boldsymbol{\psi}. \end{array}$$

This corresponds to choosing

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{\mathbf{xx}} & \mathbf{P}_{\mathbf{x}\boldsymbol{\lambda}} \\ \mathbf{P}_{\boldsymbol{\lambda}\mathbf{x}} & \mathbf{P}_{\boldsymbol{\lambda}\boldsymbol{\lambda}} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_m \end{bmatrix} \tag{21}$$

in (20), where $\mathbf{I}_p$ denotes the $p \times p$ identity matrix.

4.2 Hestenes' method of multipliers

In a discrete-time setting let $\boldsymbol{\lambda}_k$ denote the current estimate for the Lagrange multiplier $\boldsymbol{\lambda}^*$ and let $\mathbf{x} = \mathbf{x}(\boldsymbol{\lambda}_k)$ denote the minimizer of $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}_k)$. Then

$$\mathbf{0} = \boldsymbol{\nabla}_{\mathbf{x}}\mathcal{L} = \boldsymbol{\nabla}_{\mathbf{x}} L + \beta\boldsymbol{\Gamma}^{\intercal}\boldsymbol{\psi} = \nabla_{\mathbf{x}}\phi - \boldsymbol{\Gamma}^{\intercal}\left[\boldsymbol{\lambda}_k - \beta\boldsymbol{\psi}\right].$$

Hestenes [13] suggests taking $\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k - \beta\boldsymbol{\psi}$. Then if $\boldsymbol{\psi}(\mathbf{x}_{k+1}) = \mathbf{0}$ at the minimizer $\mathbf{x}_{k+1}$ of $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}_{k+1})$

$$\mathbf{0} = \boldsymbol{\nabla}_{\mathbf{x}}\mathcal{L} = \boldsymbol{\nabla}_{\mathbf{x}} L = \nabla_{\mathbf{x}}\phi\left(\mathbf{x}_{k+1}\right) - \boldsymbol{\Gamma}^{\intercal}\left(\mathbf{x}_{k+1}\right)\boldsymbol{\lambda}_{k+1}$$

would yield $(\mathbf{x}_{k+1}, \boldsymbol{\lambda}_{k+1}) = (\mathbf{x}^*, \boldsymbol{\lambda}^*)$ satisfying the first-order necessary conditions (3).

The continuous-time version of Hestenes' Method of Multipliers is $\dot{\boldsymbol{\lambda}} = -\beta\boldsymbol{\psi}$. This, coupled with steepest descent on $\mathbf{x}$, corresponds to choosing in (20):

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{\mathbf{xx}} & \mathbf{P}_{\mathbf{x}\boldsymbol{\lambda}} \\ \mathbf{P}_{\boldsymbol{\lambda}\mathbf{x}} & \mathbf{P}_{\boldsymbol{\lambda}\boldsymbol{\lambda}} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & -\beta\mathbf{I}_m \end{bmatrix}. \tag{22}$$

4.3 Newton's method

For $\max_{\boldsymbol{\lambda}} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$ the first-order necessary conditions are

$$\mathbf{0} = \boldsymbol{\nabla}_{\mathbf{x}}\mathcal{L} = \boldsymbol{\nabla}_{\mathbf{x}} L + \beta\boldsymbol{\Gamma}^{\intercal}\boldsymbol{\psi}, \quad \mathbf{0} = \boldsymbol{\nabla}_{\boldsymbol{\lambda}}\mathcal{L} = \boldsymbol{\nabla}_{\boldsymbol{\lambda}} L = -\boldsymbol{\psi}. \tag{23}$$

Newton's method applied to (23) is given by

$$\begin{bmatrix} \dot{\mathbf{x}} \\ \dot{\boldsymbol{\lambda}} \end{bmatrix} = -\mathcal{H}^{-1}(\mathbf{y}) \begin{bmatrix} \boldsymbol{\nabla}_{\mathbf{x}}\mathcal{L} \\ \boldsymbol{\nabla}_{\boldsymbol{\lambda}}\mathcal{L} \end{bmatrix} = -\mathcal{H}^{-1}(\mathbf{y}) \begin{bmatrix} \boldsymbol{\nabla}_{\mathbf{x}}\mathcal{L} \\ -\boldsymbol{\psi} \end{bmatrix}, \tag{24}$$

where

$$\mathcal{H}(\mathbf{y}) \triangleq \boldsymbol{\nabla}^2_{\mathbf{y}}\mathcal{L} = \frac{\partial^2 \mathcal{L}}{\partial \mathbf{y}^2} = \begin{bmatrix} \dfrac{\partial^2 \mathcal{L}}{\partial \mathbf{x}^2} & -\boldsymbol{\Gamma}^{\intercal} \\ -\boldsymbol{\Gamma} & \mathbf{0} \end{bmatrix}, \tag{25}$$

with $\Gamma(\mathbf{x})$ defined by (10). This corresponds to choosing in (20):

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{\mathbf{xx}} & \mathbf{P}_{\mathbf{x}\boldsymbol{\lambda}} \\ \mathbf{P}_{\boldsymbol{\lambda}\mathbf{x}} & \mathbf{P}_{\boldsymbol{\lambda}\boldsymbol{\lambda}} \end{bmatrix} = \mathcal{H}^{-1}.$$

Instead of assuming $\mathcal{H}^{-1}$ exists, Newton's method can be written as

$$\mathcal{H}(\mathbf{y}) \begin{bmatrix} \dot{\mathbf{x}} \\ \dot{\boldsymbol{\lambda}} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial^2 \mathcal{L}}{\partial \mathbf{x}^2} & -\boldsymbol{\Gamma}^\intercal \\ -\boldsymbol{\Gamma} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \dot{\mathbf{x}} \\ \dot{\boldsymbol{\lambda}} \end{bmatrix} = - \begin{bmatrix} \boldsymbol{\nabla}_{\mathbf{x}} \mathcal{L} \\ \boldsymbol{\nabla}_{\boldsymbol{\lambda}} \mathcal{L} \end{bmatrix}. \tag{26}$$

A geometric interpretation of Newton's method is given by noting that

$$\frac{\mathrm{d}}{\mathrm{d}t} \begin{bmatrix} \boldsymbol{\nabla}_{\mathbf{x}} \mathcal{L} \\ \boldsymbol{\nabla}_{\boldsymbol{\lambda}} \mathcal{L} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial^2 \mathcal{L}}{\partial \mathbf{x}^2} \dot{\mathbf{x}} + \dfrac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\lambda} \partial \mathbf{x}} \dot{\boldsymbol{\lambda}} \\ \dfrac{\partial^2 \mathcal{L}}{\partial \mathbf{x} \partial \boldsymbol{\lambda}} \dot{\mathbf{x}} + \dfrac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\lambda}^2} \dot{\boldsymbol{\lambda}} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial^2 \mathcal{L}}{\partial \mathbf{x}^2} & -\boldsymbol{\Gamma}^\intercal \\ -\boldsymbol{\Gamma} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \dot{\mathbf{x}} \\ \dot{\boldsymbol{\lambda}} \end{bmatrix}. \tag{27}$$

Thus from (26)

$$\frac{\mathrm{d}}{\mathrm{d}t} \begin{bmatrix} \boldsymbol{\nabla}_{\mathbf{x}} \mathcal{L} \\ \boldsymbol{\nabla}_{\boldsymbol{\lambda}} \mathcal{L} \end{bmatrix} = - \begin{bmatrix} \boldsymbol{\nabla}_{\mathbf{x}} \mathcal{L} \\ \boldsymbol{\nabla}_{\boldsymbol{\lambda}} \mathcal{L} \end{bmatrix}. \tag{28}$$

Hence

$$\begin{bmatrix} \boldsymbol{\nabla}_{\mathbf{x}} \mathcal{L} \\ \boldsymbol{\nabla}_{\boldsymbol{\lambda}} \mathcal{L} \end{bmatrix}_t = \mathrm{e}^{-t} \begin{bmatrix} \boldsymbol{\nabla}_{\mathbf{x}} \mathcal{L} \\ \boldsymbol{\nabla}_{\boldsymbol{\lambda}} \mathcal{L} \end{bmatrix}_{t=0} \text{ and we have } \begin{bmatrix} \boldsymbol{\nabla}_{\mathbf{x}} \mathcal{L} \\ \boldsymbol{\psi} \end{bmatrix}_t = \mathrm{e}^{-t} \begin{bmatrix} \boldsymbol{\nabla}_{\mathbf{x}} \mathcal{L} \\ \boldsymbol{\psi} \end{bmatrix}_{t=0} \to \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} \text{ as } t \to \infty.$$

Thus Newton's method: a) is at least locally asymptotically stable to a point $\hat{\mathbf{y}}$ satisfying the necessary conditions (3) provided $\mathcal{H}^{-1}(\hat{\mathbf{y}})$ exists, b) is not stiff (all eigenvalues are $\mu = -1$), and c) has a domain of attraction that is the region containing $\hat{y}$, where $\mathcal{H}^{-1}$ exists. However, Newton's method may not be globally convergent. Furthermore, it only seeks stationary points of the augmented Lagrangian $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$, not specifically those yielding $\max_{\boldsymbol{\lambda}} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$.

5 Stiff Differential Equations

Stiff systems are systems of differential equations which have two or more widely separated time scales, usually specified in terms of eigenvalues. For nonlinear systems we will use Lyapunov exponents.

5.1 Lyapunov exponents

Lyapunov exponents [14, p. 205] are generalizations of eigenvalues and characteristic (Floquet) multipliers that provide information about the (average) rates at which neighboring trajectories converge or diverge in a nonlinear system. Let $\mathbf{y}(t)$ and $\tilde{\mathbf{y}}(t)$ be solutions to

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}), \tag{29}$$

starting from neighboring initial conditions, and let $\rho(t) = \|\tilde{\mathbf{y}}(t) - \mathbf{y}(t)\|$ be the distance between the trajectory $\mathbf{y}(t)$ and the perturbed trajectory $\tilde{\mathbf{y}}(t)$ at time t. If $\rho(0)$ is arbitrarily small and $\rho(t) \to \rho(0)\mathrm{e}^{\sigma t}$ as $t \to \infty$ then σ is called a Lyapunov exponent for the reference trajectory $\mathbf{y}(t)$. The distance between the trajectory points $\mathbf{y}(t)$ and $\tilde{\mathbf{y}}(t)$ grows, shrinks, or remains constant for $\sigma > 0$, $\sigma < 0$, or $\sigma = 0$, respectively. In a p-dimensional state space there are p real Lyapunov exponents, $\sigma_1 \geq \ldots \geq \sigma_p$, corresponding to exponential growth rates in p orthogonal directions. For a given trajectory

$\mathbf{y}(t)$ the Lyapunov exponents are unique, but are functions of the initial state. Arbitrarily close initial states (*e.g.,* on and to either side of a separatrix) may yield trajectories with different Lyapunov exponents, corresponding to different behaviors as $t \to \infty$.

If $\mathbf{f}(\cdot)$ is continuous and continuously differentiable the Lyapunov exponents can be calculated in terms of the state perturbation equations

$$\dot{\boldsymbol{\eta}} = \mathbf{A}(t)\boldsymbol{\eta}, \qquad \mathbf{A}(t) = \frac{\partial \mathbf{f}[\mathbf{y}(t)]}{\partial \mathbf{y}}, \tag{30}$$

where $\mathbf{A}(t)$ is evaluated along a trajectory $\mathbf{y}(t)$ and, for small α, $\tilde{\mathbf{y}}(t) = \mathbf{y}(t) + \alpha\boldsymbol{\eta}(t) + \mathbf{O}(\alpha^2)$ is an initially neighboring trajectory. If $\mathbf{f}(\cdot)$ is discontinuous across some "switching surface" in state space certain "jump conditions" must be imposed to accurately compute Lyapunov exponents [15].

For the special case of an equilibrium $\mathbf{y}(t) =$ constant, so that $\mathbf{A}$ is constant, the Lyapunov exponents σ_k are the real parts of the eigenvalues μ_k, $k = 1, \ldots, p$, of $\mathbf{A}$. The same result holds for trajectories that asymptotically approach an equilibrium.

One way to compute Lyapunov exponents numerically [16] is to integrate the equations of motion (29), along with p copies of the perturbation equations (30), one for each of p initially orthogonal unit perturbations $\boldsymbol{\eta}_k(0)$, corresponding to the semi-axes of an initially spherical p-dimensional ellipsoid in state space. At $t > 0$ we define the instantaneous Lyapunov exponents as

$$\sigma_k(t) = \frac{1}{t} \ln \left[\frac{\|\boldsymbol{\eta}_k(t)\|}{\|\boldsymbol{\eta}_k(0)\|} \right] \tag{31}$$

with the Lyapunov exponents $\sigma_k = \lim_{t\to\infty} \{\sigma_k(t)\}$. We define the instantaneous "stiffness" as $\Sigma(t) \triangleq |\sigma_{\max}(t) - \sigma_{\min}(t)|$. As the trajectory $\mathbf{y}(t)$ moves through state space, the perturbation vectors $\boldsymbol{\eta}_k(t)$ rotate (so they are no longer orthogonal) and stretch or shrink as the axes of the ellipsoid centered at $\mathbf{y}(t)$ change. Over time, the perturbation vectors will all tend to align with the major axis of the ellipse, corresponding to the largest Lyapunov exponent, in a manner analogous to the power method for generating the dominant eigenvalue and eigenvector of a matrix. Since some of the Lyapunov exponents may be positive, particularly in chaotic systems, the algorithm incorporates a periodic discontinuous rescaling of the perturbation vectors, to avoid numerical overflow, using a Gramm-Schmidt orthonormalization procedure [14, p. 207].

5.2 State perturbation equations

Let $\mathbf{p}_k^{\mathsf{T}}(\mathbf{y})$, $k = 1, \ldots, p$, denote the k-th row of $\mathbf{P}(\mathbf{y})$. Then

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) = -\mathbf{P}(\mathbf{y})\boldsymbol{\nabla}_{\mathbf{y}}\mathcal{L}(\mathbf{y}) = -\begin{bmatrix} \mathbf{p}_1^{\mathsf{T}}(\mathbf{y}) \\ \vdots \\ \mathbf{p}_p^{\mathsf{T}}(\mathbf{y}) \end{bmatrix} \boldsymbol{\nabla}_{\mathbf{y}}\mathcal{L}(\mathbf{y}).$$

Along a trajectory $\mathbf{y}(t)$ the state perturbation equations (30), with $\mathbf{A}(\mathbf{y}) = \partial\mathbf{f}(\mathbf{y})/\partial\mathbf{y}$, are given by

$$\mathbf{A}(\mathbf{y}) = -\mathbf{P}(\mathbf{y})\mathcal{H}(\mathbf{y}) - \begin{bmatrix} \frac{\partial \mathcal{L}(\mathbf{y})}{\partial \mathbf{y}} \frac{\partial \mathbf{p}_1(\mathbf{y})}{\partial \mathbf{y}} \\ \vdots \\ \frac{\partial \mathcal{L}(\mathbf{y})}{\partial \mathbf{y}} \frac{\partial \mathbf{p}_p(\mathbf{y})}{\partial \mathbf{y}} \end{bmatrix}, \tag{32}$$

where $\mathcal{H} = \boldsymbol{\nabla}_{\mathbf{y}}^2 \mathcal{L}(\mathbf{y}) = \partial^2 \mathcal{L}/\partial \mathbf{y}^2$. At a stationary point $\mathbf{y}^*$ of $\mathcal{L}(\mathbf{y})$, $\boldsymbol{\nabla}_{\mathbf{y}} \mathcal{L}(\mathbf{y}^*) = \mathbf{0}$ and $\mathbf{A}(\mathbf{y}^*) = -\mathbf{P}(\mathbf{y}^*)\mathcal{H}(\mathbf{y}^*)$. This result also holds for $\mathbf{P}$ constant. The eigenvalues of $\mathbf{A}(\mathbf{y}^*)$ provide a measure of the stiffness of the system (29), at least near $\mathbf{y}^*$. Along a trajectory $\mathbf{y}(t)$ the Lyapunov exponents (31) do so.

6 Unconstrained Min-Max Saddle Point

In [8] a Gradient Enhanced Min-Max (GEMM) algorithm is developed as a variable Levenberg-Marquardt modification [17, p. 145] to Newton's method, designed to find saddle points of a scalar-valued function. The GEMM algorithm specifically seeks min-max saddle points, whereas Newton's method seeks stationary points. As we shall see, GEMM generally has a larger domain of attraction than Newton's method (by keeping the Hessian matrix nonsingular), is not stiff, and is faster than Newton's method.

As background we summarize some results from [8] for GEMM applied to the problem of finding a game-theoretic saddle point in the absence of equality constraints.

Let $\mathbf{M}^\intercal$ denote the transpose of a matrix $\mathbf{M}$. For $\mathbf{y}^\intercal = [\mathbf{u}^\intercal, \mathbf{v}^\intercal]$, with $\mathbf{u} \in \mathcal{U} \subseteq R^n$, $\mathbf{v} \in \mathcal{V} \subseteq R^m$, and $\mathbf{y} \in R^p$, $p = n + m$, we are concerned with finding a point $\mathbf{y}^* = (\mathbf{u}^*, \mathbf{v}^*)$ to yield a min-max for a C^2 scalar-valued function $\phi(\mathbf{y}) = \phi(\mathbf{u}, \mathbf{v})$, such that $\mathbf{u}^*$ minimizes ϕ and $\mathbf{v}^*$ maximizes ϕ. That is, $\phi(\mathbf{u}^*, \mathbf{v}) \le \phi(\mathbf{u}^*, \mathbf{v}^*) \le \phi(\mathbf{u}, \mathbf{v}^*)$ for all $\mathbf{u} \in \mathcal{U}$ and $\mathbf{v} \in \mathcal{V}$. Denote the gradient of ϕ by

$$\mathbf{g} = \left[\frac{\partial \phi}{\partial \mathbf{y}}\right]^\intercal = \begin{bmatrix} \mathbf{g}_u \\ \mathbf{g}_v \end{bmatrix} = \begin{bmatrix} \left[\frac{\partial \phi}{\partial \mathbf{u}}\right]^\intercal \\ \left[\frac{\partial \phi}{\partial \mathbf{v}}\right]^\intercal \end{bmatrix}$$

and the Hessian of ϕ by

$$\mathbf{G} = \frac{\partial^2 \phi}{\partial \mathbf{y}^2} = \begin{bmatrix} \mathbf{G}_{uu} & \mathbf{G}_{uv} \\ \mathbf{G}_{uv}^\intercal & \mathbf{G}_{vv} \end{bmatrix},$$

where $\mathbf{g}_u \in R^n$, $\mathbf{g}_v \in R^m$, $\mathbf{G}_{uu} = \partial^2 \phi/\partial \mathbf{u}^2 \in R^{n \times n}$, $\mathbf{G}_{vv} = \partial^2 \phi/\partial \mathbf{v}^2 \in R^{m \times m}$, and $\mathbf{G}_{uv} = \partial^2 \phi/\partial \mathbf{u} \partial \mathbf{v} \in R^{n \times m}$.

We are particularly concerned with finding a **proper stationary min-max point** $\mathbf{y}^*$, at which:

1. $\phi(\mathbf{u}^*, \mathbf{v}) < \phi(\mathbf{u}^*, \mathbf{v}^*) < \phi(\mathbf{u}, \mathbf{v}^*)$ for all $\mathbf{u} \in \mathcal{U} - \{\mathbf{u}^*\}$ and $\mathbf{v} \in \mathcal{V} - \{\mathbf{v}^*\}$,
2. $\mathbf{g}^* = \mathbf{g}(\mathbf{y}^*) = \mathbf{0}$,
3. $\mathbf{G}_{uu}^* = \mathbf{G}_{uu}(\mathbf{y}^*) \ge 0$ (positive semidefinite),
4. $\mathbf{G}_{vv}^* = \mathbf{G}_{vv}(\mathbf{y}) \le 0$ (negative semidefinite),
5. $|\mathbf{G}^*| = |\mathbf{G}(\mathbf{y}^*)| < 0$,

where $|\cdot|$ denotes the determinant. In addition we assume that $\mathbf{g}(\mathbf{y}) \ne \mathbf{0}$ for $\mathbf{y} \ne \mathbf{y}^*$ and that $\|\mathbf{g}(\mathbf{y})\| \to \infty$ as $\|\mathbf{y} - \mathbf{y}^*\| \to \infty$, where $\|\cdot\|$ denotes the Euclidian norm.

For $\mathbf{u} \in \mathcal{U}$ and $\mathbf{v} \in \mathcal{V}$ let

$$\mathcal{R}_u = \{(\mathbf{u}, \mathbf{v}) : \mathbf{v} \in \mathcal{V} \text{ and } \phi(\mathbf{u}, \mathbf{v}) \le \phi(\bar{\mathbf{u}}, \mathbf{v}) \text{ for all } \bar{\mathbf{u}} \in \mathcal{U}, \}$$

denote the rational reaction set for the minimizing player $\mathbf{u}$, and let

$$\mathcal{R}_v = \{(\mathbf{u}, \mathbf{v}) : \mathbf{u} \in \mathcal{U} \text{ and } \phi(\mathbf{u}, \bar{\mathbf{v}}) \le \phi(\mathbf{u}, \mathbf{v}) \text{ for all } \bar{\mathbf{v}} \in \mathcal{V}\}$$

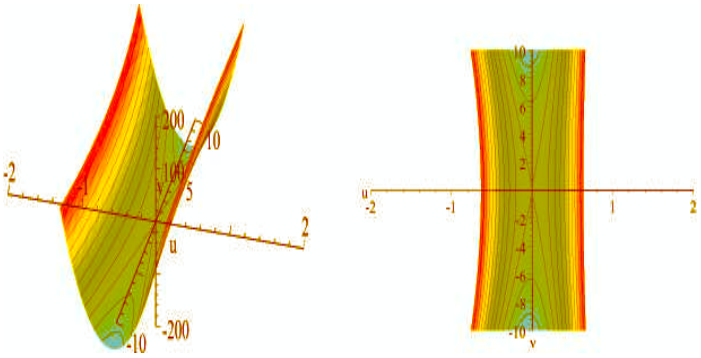

Figure 1: Banana saddle ($a = 1000$, $c = 1$).

denote the rational reaction set for the maximizing player $\mathbf{v}$. On $\mathcal{R}_u$ with $\mathbf{u} \in \overset{o}{\mathcal{U}}$(interior of $\mathcal{U}$) it is necessary [9, p. 149] that

$$\mathbf{0} = \mathbf{g}_u(\mathbf{u}, \mathbf{v}) = \left[\frac{\partial \phi(\mathbf{u}, \mathbf{v})}{\partial \mathbf{u}}\right]^{\mathsf{T}} \tag{33}$$

and

$$\mathbf{G}_{uu}(\mathbf{u}, \mathbf{v}) = \frac{\partial^2 \phi(\mathbf{u}, \mathbf{v})}{\partial \mathbf{u}^2} \geq 0.$$

On $\mathcal{R}_v$ with $\mathbf{v} \in \overset{o}{\mathcal{V}}$ it is necessary that

$$\mathbf{0} = \mathbf{g}_v(\mathbf{u}, \mathbf{v}) = \left[\frac{\partial \phi(\mathbf{u}, \mathbf{v})}{\partial \mathbf{v}}\right]^{\mathsf{T}} \tag{34}$$

and

$$\mathbf{G}_{vv}(\mathbf{u}, \mathbf{v}) = \frac{\partial^2 \phi(\mathbf{u}, \mathbf{v})}{\partial \mathbf{v}^2} \leq 0.$$

6.1 Stingray saddle function

For $a > 0$ and $c > 0$ we consider the "Stingray" saddle function

$$\phi = \frac{a}{2}u^2 + \frac{c}{2}(u-1)\,v^2 \tag{35}$$

with gradient and Hessian

$$\mathbf{g} = \begin{bmatrix} g_u \\ g_v \end{bmatrix} = \begin{bmatrix} au + \frac{c}{2}v^2 \\ c\,(u-1)\,v \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} G_{uu} & G_{uv} \\ G_{uv} & G_{vv} \end{bmatrix} = \begin{bmatrix} a & cv \\ cv & c\,(u-1) \end{bmatrix}.$$

The function has a unique proper min-max point at $\mathbf{y}^* = (u^*, v^*) = (0, 0)$, with $\mathbf{g} \neq \mathbf{0}$ for $\mathbf{y} \neq \mathbf{0}$ and $\|\mathbf{g}\| \to \infty$ as $\|\mathbf{y}\| \to \infty$. Note that $|\mathbf{G}| = ac\,(u-1) - c^2v^2 = 0$ on $u = 1 + \frac{c}{a}v^2$. Also note that $G_{uu} = a > 0$ for all (u, v), but $G_{vv} = c\,(u-1) < 0$ only for $u < 1$. The Stingray function $\phi(u, v)$ is convex in u for each v, but is concave in v only for $u < 1$. For $u > 1$ the function is convex in v.

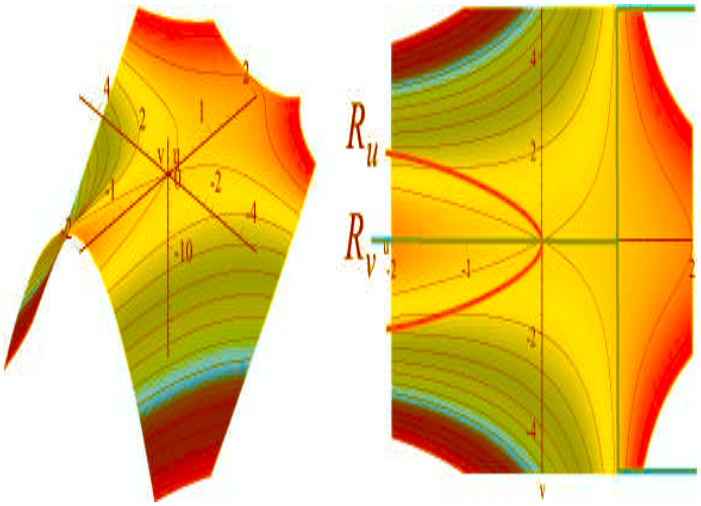

Figure 2: Stingray saddle ($a = 1$, $c = 1$).

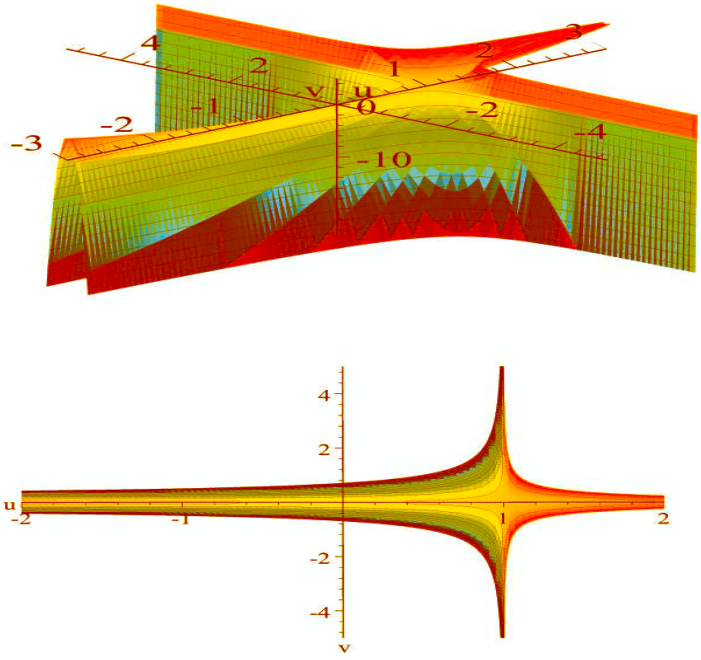

Figure 3: Stingray saddle ($a = 1$, $c = 100$).

Figures 1–3 show three-dimensional and contour plots for various values of a and c. For $a = 1000$ and $c = 1$ the function is similar to Rosenbrock's "banana" function, having a steep-walled canyon with a parabolic valley, except that the stationary point is a saddle point instead of a minimum point. For $a = 1$ and increasing values of c the function looks like a stingray flapping its wings. Unless otherwise specified, we will consider the case where $a = 1$ and $c = 100$. For these parameter values the stingray function has a sharp local $\max_v$ ridge on $v = 0$, $u < 1$, and a local $\max_u$ valley on $u = -\frac{c}{2a}v^2$.

As illustrated in Figure 2, the $\min_u \max_v \phi$ rational reaction sets, for $v_{\min} \le v \le v_{\max}$ with $v_{\max} > 0$ and $v_{\min} < 0$, are

$$\mathcal{R}_u = \left\{(u,v) : u = -\frac{c}{2a}v^2\right\}, \quad \mathcal{R}_v = \left\{(u,v) : v = v^\circ(u)\right\},$$

where

$$v^\circ(u) = \begin{cases} 0 & \text{if} \quad u < 1, \\ \in [v_{\min}, v_{\max}] & \text{if} \quad u = 1, \\ v_{\max} & \text{if} \quad u > 1 \text{ and } v_{\max} \ge |v_{\min}|, \\ v_{\min} & \text{if} \quad u > 1 \text{ and } v_{\max} \le |v_{\min}|. \end{cases}$$

In particular, while the minimizing player u seeks $g_u = 0$, the maximizing player v only seeks $g_v = 0$ for $u < 1$. For $u > 1$ the maximizing player seeks either the upper or lower bound on v. Nevertheless, the intersection $\mathcal{R}_u \cap \mathcal{R}_v$ of the reaction sets is the min-max point $u^* = v^* = 0$, where both $g_u = 0$ and $g_v = 0$.

6.2 Gradient enhanced Newton (GEN) minimization

Consider, for a moment, Newton's method applied to the problem of finding a unique proper *minimum* point for a function $\phi(\mathbf{y})$. For the case where $\mathbf{G}(\mathbf{y}) = \partial^2\phi/\partial\mathbf{y}^2$ is not positive definite everywhere, the Levenberg–Marquardt modification to Newton's method [17, pp. 145–149] is given by $(\alpha\mathbf{I} + \mathbf{G})\dot{\mathbf{y}} = -\mathbf{g}$, where $\alpha \ge 0$ and $\mathbf{I}$ denotes the $p \times p$ identity matrix. If $\mathbf{F} = \alpha\mathbf{I} + \mathbf{G}$ is positive definite, then let $\dot{\mathbf{y}} = -\mathbf{P}(\mathbf{y})\mathbf{g}$, with $\mathbf{P}(\mathbf{y}) = \mathbf{F}^{-1} = (\alpha\mathbf{I} + \mathbf{G})^{-1}$. Then $\dot{\phi} = \mathbf{g}^\intercal\dot{\mathbf{y}} = -\mathbf{g}^\intercal\mathbf{P}\mathbf{g} < 0$ for $\mathbf{g} \ne \mathbf{0}$ establishes (global) asymptotic stability.

Let μ_i and $\boldsymbol{\xi}_i$, $i = 1, \ldots, p$, denote the eigenvalues and eigenvectors of $\mathbf{G}$, respectively. For symmetric $\mathbf{G}$ the eigenvalues are all real, but may not all be positive. The matrix $\mathbf{F} = \alpha\mathbf{I} + \mathbf{G}$ has eigenvalues $\omega_i = \mu_i + \alpha$ and eigenvectors $\boldsymbol{\xi}_i$, since $\mathbf{F}\boldsymbol{\xi}_i = (\mu_i + \alpha)\,\boldsymbol{\xi}_i$. Thus, at a point $\mathbf{y}$, if α is sufficiently large all of the eigenvalues of $\mathbf{F}$ will be positive. As $\alpha \to 0$ the method approaches Newton's method applied to $\phi(\mathbf{y})$, and as $\alpha \to \infty$ the method approaches Steepest Descent applied to $\phi(\mathbf{y})/\alpha$.

The Levenberg–Marquardt minimization method generally will not work with constant α. If $|\mathbf{G}(\mathbf{y})|$ changes sign somewhere then for constant α the determinant $|\mathbf{F}| = |\alpha\mathbf{I} + \mathbf{G}|$ will also generally change sign, although at a different place than $|\mathbf{G}(\mathbf{y})|$.

In [7] we develop a **Gradient Enhanced Newton** (GEN) minimization method, in which $\alpha = \gamma\,\|\mathbf{g}\| = \gamma\sqrt{\mathbf{g}^\intercal\mathbf{g}}$ with constant $\gamma \ge 0$, yielding

$$\dot{\mathbf{y}} = -\mathbf{P}(\mathbf{y})\mathbf{g} = -\left[\gamma\,\|\mathbf{g}\|\,\mathbf{I} + \mathbf{G}\right]^{-1}\mathbf{g}. \tag{36}$$

The ideas behind this minimization method are: 1) at points where $\|\mathbf{g}\| \ne 0$ we can make $\mathbf{F}$ be positive definite for sufficiently large $\gamma \ge 0$; 2) for small γ or near places where $\mathbf{g} = \mathbf{0}$ the method behaves like Newton's method; 3) the speed $\|\dot{\mathbf{y}}\| \approx 1/\gamma$. In [7] it is shown that, for sufficiently large $\gamma \ge 0$, GEN is globally asymptotically stable for functions that

have a single proper stationary minimum point and satisfy a Lyapunov growth condition. In addition, when applied to Rosenbrock's "banana" function, GEN is uniformly nonstiff and approximately 25 times faster than Newton's method and approximately 2500 times faster than Steepest Descent.

A very recent paper [18] shows that, for long-term optimization algorithms, Levenberg–Marquardt, especially in the form (36), is more fundamental than Newton's method and that Newton's method should be viewed as a special case of Levenberg–Marquardt, rather than the other way around.

6.3 Gradient enhanced min-max

The Levenberg–Marquardt modification of Newton's method can not be used for min-max problems, but a variation of it can. Consider the Hessian

$$\mathbf{G}(\mathbf{y}) = \frac{\partial^2 \phi}{\partial \mathbf{y}^2} = \begin{bmatrix} \frac{\partial^2 \phi}{\partial u^2} & \frac{\partial^2 \phi}{\partial u \partial v} \\ \frac{\partial^2 \phi}{\partial v \partial u} & \frac{\partial^2 \phi}{\partial v^2} \end{bmatrix} = \begin{bmatrix} \mathbf{G}_{uu} & \mathbf{G}_{uv} \\ \mathbf{G}_{uv}^\intercal & \mathbf{G}_{vv} \end{bmatrix},$$

which is positive definite at a proper minimum point. But at a proper min-max point $\mathbf{y}^*$ we have $\mathbf{G}^*_{uu} = \mathbf{G}_{uu}(\mathbf{y}^*) \geq 0$, $\mathbf{G}^*_{vv} = \mathbf{G}_{vv}(\mathbf{y}^*) \leq 0$, and $|\mathbf{G}^*| = |\mathbf{G}(\mathbf{y}^*)| < 0$. Thus the eigenvalues of $\mathbf{G}^*_{uu}$ are ≥ 0, the eigenvalues of $\mathbf{G}^*_{vv}$ are ≤ 0, and the product of the eigenvalues of $\mathbf{G}^*$ is negative. When $|\mathbf{G}(\mathbf{y})|$ passes through zero, so does one or more of its eigenvalues. The Levenberg-Marquardt matrix $\mathbf{F} = \alpha \mathbf{I} + \mathbf{G}$ could be used to make *all* of its eigenvalues be positive (or *all* of them negative, for $\alpha < 0$) at any given point $\hat{\mathbf{y}}$. But if $\alpha = \alpha(\mathbf{y}) \geq 0$, with $\alpha(\mathbf{y}^*) = 0$ and $|\mathbf{G}^*| = |\mathbf{G}(\mathbf{y}^*)| < 0$, then somewhere between $\mathbf{y}^*$ and $\hat{\mathbf{y}}$ we would have $|\mathbf{F}(\mathbf{y})| = 0$, as one of the positive eigenvalues goes negative or one of the negative eigenvalues goes positive. What we need to do, to ensure that the replacement matrix $\mathbf{F}(\mathbf{y})$ for $\mathbf{G}(\mathbf{y})$ is nonsingular, is to keep the positive eigenvalues positive and the negative eigenvalues negative, yielding $|\mathbf{F}(\mathbf{y})| < 0$.

Consider

$$\dot{\mathbf{y}} = -\mathbf{P}\mathbf{g} \tag{37}$$

with

$$\mathbf{P} = \mathbf{F}^{-1} = \begin{bmatrix} \alpha_u \mathbf{I}_u + \mathbf{G}_{uu} & \mathbf{G}_{uv} \\ \mathbf{G}_{uv}^\intercal & -\alpha_v \mathbf{I}_v + \mathbf{G}_{vv} \end{bmatrix}^{-1}. \tag{38}$$

For $\alpha_u = \alpha_v = \alpha \to \infty$ the method approaches Min-Max Ascent (see Section 6.4.1) applied to ϕ/α. For $\alpha \to 0$ the method approaches Newton's method applied to ϕ. The **Gradient Enhanced Min-Max** (GEMM) method is given by (37)–(38) with $\alpha_u = \gamma_u \|\mathbf{g}\|$ and $\alpha_v = \gamma_v \|\mathbf{g}\|$ for constants $\gamma_u \geq 0$ and $\gamma_v \geq 0$. That is, $\mathbf{P} = \mathbf{F}^{-1}$, with

$$\mathbf{F} = \begin{bmatrix} \gamma_u \|\mathbf{g}\| \mathbf{I}_u + \mathbf{G}_{uu} & \mathbf{G}_{uv} \\ \mathbf{G}_{uv}^\intercal & -\gamma_v \|\mathbf{g}\| \mathbf{I}_v + \mathbf{G}_{vv} \end{bmatrix}. \tag{39}$$

In [8] we prove that for sufficiently large constants $\gamma_u \geq 0$ and $\gamma_v \geq 0$ the matrix $\mathbf{F}$ in (39) is nonsingular for all $\mathbf{y}$. Hence the only equilibrium for (37)–(39) is at $\mathbf{y}^*$. A Lyapunov approach can be used to investigate whether the unique equilibrium at $\mathbf{y}^*$ is (globally) asymptotically stable. However, note that using $W(\mathbf{y}) = \mathbf{g}^\intercal \mathbf{g}$ as a descent function [14, p. 276] would not work, since $\dot{W} = \mathbf{g}^\intercal \dot{\mathbf{g}} + \dot{\mathbf{g}}^\intercal \mathbf{g} = \mathbf{g}^\intercal \mathbf{G} \dot{\mathbf{y}} + \dot{\mathbf{y}}^\intercal \mathbf{G} \mathbf{g} = -\mathbf{g}^\intercal \mathbf{Q} \mathbf{g}$, with $\mathbf{Q} = \mathbf{G}\mathbf{P} + \mathbf{P}^\intercal \mathbf{G}$ not being positive definite if $|\mathbf{G}|$ changes sign (see Lyapunov's lemma [14, p. 223]). Also note that replacing the $\min_u \max_v \phi$ problem with Newton's

method (or the Levenberg–Marquardt modification) applied to the least squares problem [17, pp. 146–148] of minimizing $W(\mathbf{y})$, via $\dot{\mathbf{y}} = -\mathbf{H}^{-1}(\mathbf{y})\boldsymbol{\nabla} W$, where $\boldsymbol{\nabla} W = [\partial W/\partial \mathbf{y}]^{\mathsf{T}}$ and $\mathbf{H}(\mathbf{y}) = \partial^2 W/\partial \mathbf{y}^2$, would involve third derivatives of $\phi(\mathbf{y})$.

6.4 Gradient transformation results for the stingray saddle function

6.4.1 Min-max ascent

Since $\mathbf{u}$ seeks $\min_u \phi(\mathbf{u}, \mathbf{v})$ and $\mathbf{v}$ seeks $\max_v \phi(\mathbf{u}, \mathbf{v})$, the first min-max algorithm investigated by researchers [3] was steepest descent on $\mathbf{u}$ and steepest ascent on $\mathbf{v}$.

Let $\mathbf{I}_u$ and $\mathbf{I}_v$ denote the $n \times n$ and $m \times m$ identity matrices, respectively. Taking

$$\mathbf{P}(\mathbf{y}) = \operatorname{diag}[\mathbf{I}_u, -\mathbf{I}_v] = \begin{bmatrix} \mathbf{I}_u & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_v \end{bmatrix}$$

yields the **Min-Max Ascent** algorithm

$$\dot{\mathbf{u}} = -\mathbf{g}_u, \quad \dot{\mathbf{v}} = \mathbf{g}_v \tag{40}$$

with the state perturbation equations

$$\begin{bmatrix} \dot{\boldsymbol{\eta}}_u \\ \dot{\boldsymbol{\eta}}_v \end{bmatrix} = \begin{bmatrix} -\mathbf{G}_{uu} & -\mathbf{G}_{uv} \\ \mathbf{G}_{uv}^{\intercal} & \mathbf{G}_{vv} \end{bmatrix} \begin{bmatrix} \boldsymbol{\eta}_u \\ \boldsymbol{\eta}_v \end{bmatrix}.$$

For the Stingray saddle function

$$\phi = \frac{a}{2}u^2 + \frac{c}{2}(u-1)\,v^2$$

the Min-Max Ascent system is given by

$$\dot{u} = -g_u = -au - \frac{c}{2}v^2, \quad \dot{v} = g_v = c\,(u-1)\,v$$

with the state perturbation equations

$$\begin{bmatrix} \dot{\boldsymbol{\eta}}_u \\ \dot{\boldsymbol{\eta}}_v \end{bmatrix} = \begin{bmatrix} -a & -cv \\ cv & c\,(u-1) \end{bmatrix} \begin{bmatrix} \boldsymbol{\eta}_u \\ \boldsymbol{\eta}_v \end{bmatrix}.$$

At the stationary point the state perturbation matrix

$$\mathbf{A}(\mathbf{y}^*) = \begin{bmatrix} -a & 0 \\ 0 & -c \end{bmatrix}$$

has eigenvalues $\{-a, -c\}$. For $a = 1$ and $c = 100$ Min-Max Ascent yields a very stiff system.

Figure 4 shows Min-Max Ascent trajectories for the case where $a = 1$ and $c = 100$. For numerical integration we use fixed step size ($\Delta t = 10^{-5}$, because of stiffness) standard $4th$-order Runge-Kutta. Trajectories for $u < 1$ rapidly approach the $v = 0$ ($g_v = 0$) surface (the sharp local maximum ridge of the Stingray) and then slowly move along the ridge toward the saddle point at the origin. This is caused by the stiffness of the system. Notice the tendency, in the region $u > 1$, for trajectories to diverge from the $g_v = 0$ surface rather than converge to it. This is caused by G_{vv} not being negative definite everywhere.

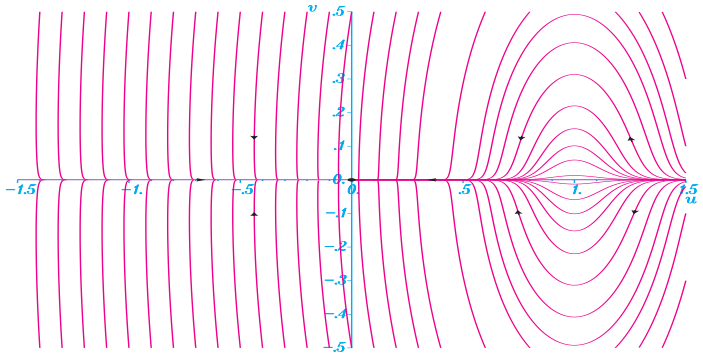

Figure 4: Min-Max Ascent ($a = 1$, $c = 100$).

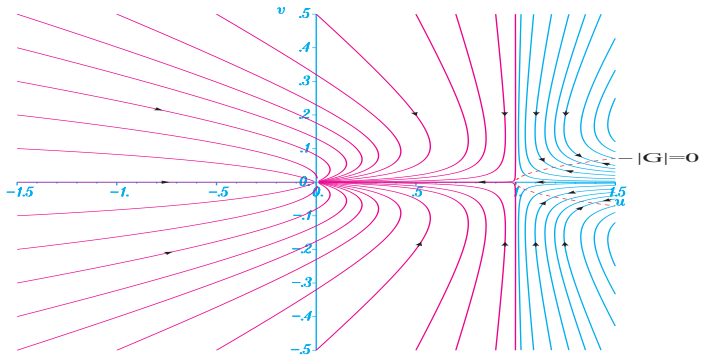

Figure 5: Newton's method ($a = 1$, $c = 100$).

6.4.2 Newton's method

Newton's method, in which $\mathrm{d}\mathbf{g}/\mathrm{d}t = -\mathbf{g}$, [hence, $\mathbf{g}(t) = \mathbf{g}(0)\mathrm{e}^{-t} \to \mathbf{0}$ as $t \to \infty$], corresponds to $\mathbf{P}(\mathbf{y}) = \mathbf{G}^{-1}(\mathbf{y})$. Applied to the Stingray saddle function, Newton's method is given by

$$\begin{aligned}\dot{\mathbf{y}} &= -\mathbf{G}^{-1}\mathbf{g} \qquad (41)\\ &= -\begin{bmatrix} a & cv \\ cv & c\,(u-1) \end{bmatrix}^{-1} \begin{bmatrix} au + \frac{c}{2}v^2 \\ c\,(u-1)\,v \end{bmatrix} = -\frac{c}{|\mathbf{G}|}\begin{bmatrix} (u-1)\left(au + \frac{1}{2}cv^2\right) - cv^2\,(u-1) \\ -v\left(au + \frac{1}{2}cv^2\right) + a\,(u-1)\,v \end{bmatrix},\end{aligned}$$

where $|\mathbf{G}| = ac\,(u-1) - c^2v^2$. Figure 5 shows trajectories for Newton's method applied to the Stingray saddle function ($a = 1$, $c = 100$) using $4th$-order Runge-Kutta ($\Delta t = 10^{-3}$).

At $\mathbf{y}^* = (u^*, v^*) = (0, 0)$ the state perturbation equations yield

$$\mathbf{A}(\mathbf{y}^*) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

with eigenvalues $\{-1, -1\}$. This is clearly not a stiff system near $\mathbf{y}^*$. Trajectories move at a much better speed than in Min-Max Ascent, as indicated by the step size. However,

Newton's method is not globally asymptotically stable to $\mathbf{y}^*$. Note that solutions to (41) only exist for $|\mathbf{G}| \neq 0$ and that $|\mathbf{G}| = 0$ on $v^2 = (u-1)\,a/c$. The domain of attraction to $\mathbf{y}^*$ is only the region $u < 1$, that is, the region where $\mathbf{G}_{vv} < 0$.

6.4.3 GEMM

For a and $c > 0$ in the Stingray saddle function (35) consider

$$\mathbf{F} = \begin{bmatrix} \alpha_u \mathbf{I}_u + \mathbf{G}_{uu} & \mathbf{G}_{uv} \\ \mathbf{G}_{uv}^{\intercal} & -\alpha_v \mathbf{I}_v + \mathbf{G}_{vv} \end{bmatrix} = \begin{bmatrix} \alpha_u + a & cv \\ cv & -\alpha_v + c\,(u-1) \end{bmatrix}.$$

The determinant $|\mathbf{F}| = (\alpha_u + a)\,[-\alpha_v + c\,(u-1)] - c^2v^2$ is zero on $c^2v^2 = (\alpha_u + a)\,[-\alpha_v + c\,(u-1)]$ provided $-\alpha_v + c\,(u-1) \geq 0$. Since $\alpha_u + a > 0$ for all $\alpha_u \geq 0$ with $a > 0$ and $c > 0$, a necessary and sufficient condition for $|\mathbf{F}(u,v)| < 0$ for all u, v is that $-\alpha_v \mathbf{I}_v + \mathbf{G}_{vv} = -\alpha_v + c\,(u-1) < 0$ for all u. We can ensure that $|\mathbf{F}(u,v)| < 0$ for all u, v by taking

$$\alpha_v = \gamma_v \,\|\mathbf{g}\| = \gamma_v \sqrt{\left(au + \frac{c}{2}v^2\right)^2 + c^2\,(u-1)^2\,v^2}$$

with sufficiently large $\gamma_v > 0$. Then

$$|\mathbf{F}| = -\gamma_v\,(\alpha_u + a)\sqrt{\left(au + \frac{c}{2}v^2\right)^2 + c^2\,(u-1)^2\,v^2} + (\alpha_u + a)\,c\,(u-1) - c^2v^2.$$

The $\max_v |\mathbf{F}|$ occurs on $v = 0$, with

$$|\mathbf{F}|_{v=0} = -\gamma_v\,(\alpha_u + a)\sqrt{(au)^2} + (\alpha_u + a)\,c\,(u-1) = (\alpha_u + a)\,[-\gamma_v a\,|u| + c\,(u-1)]\,.$$

For $u \leq 0$ we have $|\mathbf{F}|_{v=0} < 0$. For $u > 0$ we have

$$0 = |\mathbf{F}|_{v=0} = (\alpha_u + a)\,[-\gamma_v au + c\,(u-1)] = (\alpha_u + a)\,[(c - \gamma_v a)\,u - c]$$

at

$$u = \frac{1}{1 - \gamma_v \frac{a}{c}}$$

which yields $u < 0$ (a contradiction) for $\gamma_v > c/a$. We conclude that $|\mathbf{F}(u,v)| < 0$ for all u, v if we take $\alpha_u = \gamma_u \,\|\mathbf{g}\|$ and $\alpha_v = \gamma_v \,\|\mathbf{g}\|$, with $\gamma_u \geq 0$ and $\gamma_v > c/a$.

Applied to the Stingray function, the Gradient Enhanced Min-Max algorithm is given by

$$\begin{bmatrix} \dot{u} \\ \dot{v} \end{bmatrix} = -\mathbf{F}^{-1}\mathbf{g} = -\frac{1}{|\mathbf{F}|}\begin{bmatrix} [-\gamma_v \,\|\mathbf{g}\| + c\,(u-1)]\left(au + \frac{1}{2}cv^2\right) - c^2v^2\,(u-1) \\ -cv\left(au + \frac{1}{2}cv^2\right) + (\gamma_u \,\|\mathbf{g}\| + a)\,c\,(u-1)\,v \end{bmatrix},$$

where $0 > |\mathbf{F}| = (\gamma_u \,\|\mathbf{g}\| + a)\,[-\gamma_v \,\|\mathbf{g}\| + c\,(u-1)] - c^2v^2$ for all u, v, provided $\gamma_u \geq 0$ and $\gamma_v > c/a$. For $a = 1$, $c = 100$, and $\gamma_v = 101$, Figures 6–7 show trajectories for the Gradient Enhanced Min-Max algorithm for $\gamma_u = 1$ and 10, respectively.

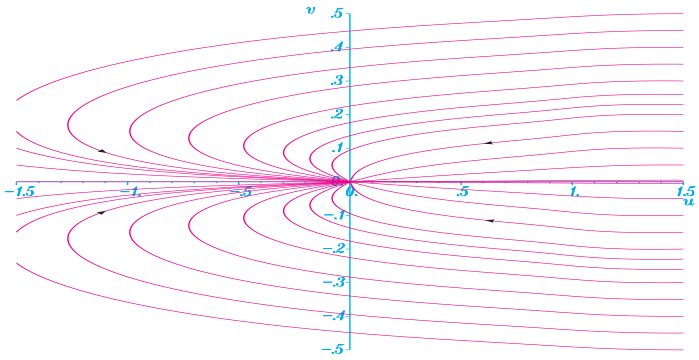

Figure 6: GEMM trajectories ($\gamma_u = 1$, $\gamma_v = 101$).

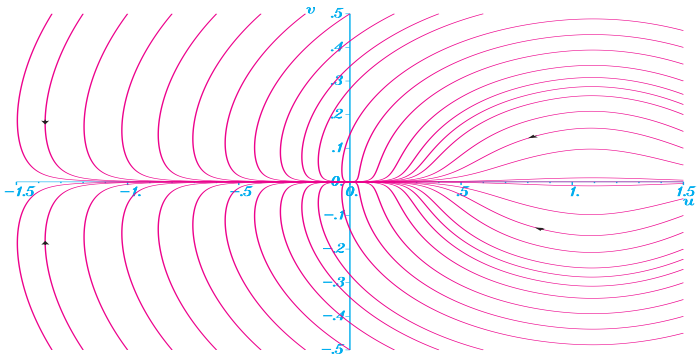

Figure 7: GEMM trajectories ($\gamma_u = 10$, $\gamma_v = 101$).

6.4.4 Unconstrained trajectory following performance comparisons

For comparison of Min-Max Ascent, Newton's method, and the Gradient Enhanced Min-Max (GEMM) method, we consider the trajectories starting from $(u, v) = (-1.5, 0.5)$ for the Stingray saddle function. We use fixed time step standard $4th$-order Runge-Kutta with the time step Δt chosen to control the approximate initial displacement $\Delta s = \|\dot{\mathbf{y}}(0)\| \, \Delta t$. The trajectories are terminated when $\|\mathbf{g}\| < 10^{-3}$. We consider two cases: Table 1 shows results for the "Banana saddle" ($a = 1000$, $c = 1$, $\gamma_u = \gamma_v = 1$, stiffness ≈ 1000), and Table 2 shows results for the "Stingray saddle" ($a = 1$, $c = 100$, $\gamma_u = 1$, $\gamma_v = 101$, stiffness ≈ 100). The results indicate that Newton's method is about 60 to 440 times faster than Min-Max Ascent, and that the Gradient Enhanced Min-Max method is about 2 to 3 times faster than Newton's method and about 175 to 1000 times faster than Min-Max Ascent. These results are consistent with the results [7] for the Gradient Enhanced Newton (GEN) minimization method. When applied to Rosenbrock's function, GEN is approximately 25 time faster than Newton's method and approximately 2500 times faster than Steepest Descent.

In [8] we show that the Gradient Enhanced Min-Max method provides global asymptotic stability to the saddle point for functions such as the Stingray saddle function, which have a single proper stationary min-max point and satisfy a Lyapunov growth

Table 1: Banana saddle results ($a = 1000$, $c = 1$).

Method	Δt	$\|\dot{\mathbf{x}}(0)\|\,\Delta t$	**Final** t	**# Steps**	**Ratio**
Min-Max Ascent	10^{-6}	1.4999×10^{-3}	6.210995	6,210,995	980.2
Newton	10^{-3}	1.5133×10^{-3}	14.221	14221	2.24
GEMM	2.5×10^{-3}	1.4999×10^{-3}	15.84	6336	1

Table 2: Stingray saddle results ($a = 1$, $c = 100$).

Method	Δt	$\|\dot{\mathbf{x}}(0)\|\,\Delta t$	**Final** t	**# Steps**	**Ratio**
Min-Max Ascent	10^{-5}	1.2548×10^{-3}	7.32973	732,973	175.5
Newton	10^{-3}	1.2962×10^{-3}	11.74	11,740	2.81
GEMM	1.5×10^{-2}	1.2543×10^{-3}	62.625	4,175	1

condition. For the Stingray function Newton's method is not stiff but does not provide global asymptotic stability. Min-Max Ascent, applied to the Stingray function, provides global asymptotic stability [8] but is very stiff. When applied to the Stingray function, the Gradient Enhanced Min-Max method is very fast and is not stiff, whereas Min-Max Ascent is very slow and very stiff. The Gradient Enhanced Min-Max method is approximately 3 times faster than Newton's method and approximately 175 to 1000 times faster than Min-Max Ascent.

7 Min-Max Saddle Point with Equality Constraints

In this section we extend the results in [8] to the problem of finding Lagrangian saddle points $\mathbf{y} = (\mathbf{x}, \boldsymbol{\lambda})$ for the problem of minimizing a scalar-valued function $\phi(\mathbf{x})$ subject to equality constraints $\boldsymbol{\psi}(\mathbf{x}) = \mathbf{0}$. In particular we expand our previous results to handle the fact that $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$ is linear in $\boldsymbol{\lambda}$.

For the augmented Lagrangian $\mathcal{L}$, and its gradient $\mathbf{h}(\mathbf{y})$ and Hessian $\mathcal{H}(\mathbf{y})$ given by (19) and (25), respectively, we are particularly concerned with finding a **proper Lagrangian saddle point** $\mathbf{y}^* = (\mathbf{x}^*, \boldsymbol{\lambda}^*)$, at which:

i) $\mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}) \leq \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) < \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*)$ for all $(\mathbf{x}, \boldsymbol{\lambda}) \neq (\mathbf{x}^*, \boldsymbol{\lambda}^*)$,

ii) $\mathbf{h}^* = [\partial \mathcal{L}(\mathbf{y}^*)/\partial \mathbf{y}]^\mathsf{T} = \mathbf{0}$, where $\mathbf{h}(\mathbf{y}) = \boldsymbol{\nabla}_\mathbf{y}\, \mathcal{L}(\mathbf{y})$,

iii) $\mathcal{H}^*_{\mathbf{xx}} = \partial^2 \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*)/\partial \mathbf{x}^2 \geq 0$ (positive semidefinite),

where, for $\beta \geq 0$, the augmented Lagrangian is

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \phi(\mathbf{x}) - \boldsymbol{\lambda}^\mathsf{T} \boldsymbol{\psi}(\mathbf{x}) + \beta \frac{1}{2} \boldsymbol{\psi}^\mathsf{T}(\mathbf{x}) \boldsymbol{\psi}(\mathbf{x}). \tag{42}$$

In addition we assume that $\mathbf{h}(\mathbf{y}) = \boldsymbol{\nabla}_\mathbf{y} \mathcal{L}(\mathbf{y}) = [\partial \mathcal{L}(\mathbf{y})/\partial \mathbf{y}]^\mathsf{T} \neq \mathbf{0}$ for $\mathbf{y} \neq \mathbf{y}^*$ and that $\|\mathbf{h}(\mathbf{y})\| \to \infty$ as $\|\mathbf{y} - \mathbf{y}^*\| \to \infty$, where $\|\cdot\|$ denotes the Euclidian norm.

As a modification to Newton's method (26) we consider a gradient transformation algorithm of the form

$$\dot{\mathbf{y}} = -\mathbf{P}(\mathbf{y}) \boldsymbol{\nabla}_\mathbf{y} \mathcal{L}(\mathbf{y}) = -\mathcal{F}^{-1}(\mathbf{y}) \boldsymbol{\nabla}_\mathbf{y} \mathcal{L}(\mathbf{y}), \tag{43}$$

where

$$\mathcal{F}(\mathbf{y}) = \mathcal{H}(\mathbf{y}) + \|\mathbf{h}\| \begin{bmatrix} \gamma_\mathbf{x} \mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & -\gamma_{\boldsymbol{\lambda}} \mathbf{I}_m \end{bmatrix} = \begin{bmatrix} \gamma_\mathbf{x} \|\mathbf{h}\| \mathbf{I}_n + \dfrac{\partial^2 \mathcal{L}}{\partial \mathbf{x}^2} & -\boldsymbol{\Gamma}^\mathsf{T} \\ -\boldsymbol{\Gamma} & -\gamma_{\boldsymbol{\lambda}} \|\mathbf{h}\| \mathbf{I}_m \end{bmatrix}, \tag{44}$$

with

$$\mathbf{h}(\mathbf{y}) = \boldsymbol{\nabla}_{\mathbf{y}}\mathcal{L}(\mathbf{y}) = \begin{bmatrix} \boldsymbol{\nabla}_{\mathbf{x}}\mathcal{L} \\ \boldsymbol{\nabla}_{\boldsymbol{\lambda}}\mathcal{L} \end{bmatrix} = \begin{bmatrix} \left[\dfrac{\partial \mathcal{L}}{\partial \mathbf{x}}\right]^{\mathsf{T}} \\ -\boldsymbol{\psi} \end{bmatrix}, \tag{45}$$

$$\mathcal{H}(\mathbf{y}) = \frac{\partial^2 \mathcal{L}(\mathbf{y})}{\partial \mathbf{y}^2} = \begin{bmatrix} \mathcal{H}_{\mathbf{xx}} & \mathcal{H}_{\boldsymbol{\lambda}\mathbf{x}} \\ \mathcal{H}_{\boldsymbol{\lambda}\mathbf{x}}^{\mathsf{T}} & \mathcal{H}_{\boldsymbol{\lambda}\boldsymbol{\lambda}} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial^2 \mathcal{L}}{\partial \mathbf{x}^2} & -\boldsymbol{\Gamma}^{\mathsf{T}} \\ -\boldsymbol{\Gamma} & \mathbf{0} \end{bmatrix}, \tag{46}$$

$$\boldsymbol{\Gamma} = \frac{\partial \boldsymbol{\psi}}{\partial \mathbf{x}}. \tag{47}$$

7.1 Nonsingularity of $\mathcal{F}(\mathbf{y})$

We will show that for a sufficiently large constant $\gamma_{\mathbf{x}} \geq 0$ and any constant $\gamma_{\boldsymbol{\lambda}} > 0$ the matrix $\mathcal{F}(\mathbf{y})$ in (44) is nonsingular for all $\mathbf{y}$. Hence the only equilibrium for (18) with (43)–(44) is at $\mathbf{y}^*$. To prove that $\mathcal{F}(\mathbf{y})$ is nonsingular, we have the following results:

Lemma 7.1 *For $\mathbf{y} \in \mathcal{R}^p$ let $\mathbf{M}(\mathbf{y})$ be an $s \times s$ matrix whose elements are functions of class C^q, $q \geq 0$, in a neighborhood of $\hat{\mathbf{y}} \in \mathcal{R}^p$, with distinct eigenvalues at $\hat{\mathbf{y}}$. Then the eigenvalues $\mu_k(\mathbf{y})$, $k = 1, \ldots, s$, of $\mathbf{M}(\mathbf{y})$ are of class C^q in a neighborhood of $\hat{\mathbf{y}}$.*

Proof The characteristic equation is $0 = \mathcal{P}(\mu, \mathbf{y}) = |\mu\mathbf{I} - \mathbf{M}(\mathbf{y})| = \mu^s + p_{s-1}\mu^{s-1} + \ldots + p_1\mu + p_0$, where $\mathbf{I}$ denotes the $s \times s$ identity matrix. The coefficients $p_k(\mathbf{y})$ are C^q since they can be determined from Newton's identities [14, p. 227] in terms of the trace$(\mathbf{M}^k)$, $k = 1, \ldots, s$, of powers of $\mathbf{M}(\mathbf{y})$, which only involves products and sums of the elements of $\mathbf{M}(\mathbf{y})$. Then the lemma follows from the implicit function theorem [9, p. 21], with Jacobian $\mathrm{d}\mathcal{P}\left(\mu_k, \hat{\mathbf{y}}\right)/\mathrm{d}\mu \neq 0$ for the case where the eigenvalues μ_k, $k = 1, \ldots, s$, are distinct.

For repeated eigenvalues, the elements of $\mathbf{M}(\mathbf{y})$ can be perturbed by an arbitrarily small amount $\epsilon > 0$ to yield distinct eigenvalues [19, p. 89]. For a more detailed analysis of the case of repeated eigenvalues, see [20, p. 134]. Henceforth, we will consider only the case of distinct eigenvalues.

Theorem 7.1 *For $\mathbf{y} \in \mathcal{R}^p$ let $\mathbf{M}(\mathbf{y}) \in \mathcal{R}^{s \times s}$ be a continuous symmetric matrix with $\mathbf{M}(\mathbf{y}^*) \geq 0$ (≤ 0) and let $\mathcal{L}(\mathbf{y})$ be a scalar-valued function of class C^q, $q \geq 1$. Let $\mathbf{h}(\mathbf{y}) = [\partial \mathcal{L}/\partial \mathbf{y}]^{\mathsf{T}}$. If $\mathbf{h}(\mathbf{y}^*) = \mathbf{0}$, with $\mathbf{h}(\mathbf{y}) \neq \mathbf{0}$ for $\mathbf{y} \neq \mathbf{y}^*$ and $\|\mathbf{h}(\mathbf{y})\| \to \infty$ as $\|\mathbf{y} - \mathbf{y}^*\| \to \infty$, then for $\gamma \geq 0$ (≤ 0) with $|\gamma|$ sufficiently large, the $s \times s$ matrix $\mathbf{N}(\mathbf{y}) = \gamma \|\mathbf{h}(\mathbf{y})\| \mathbf{I} + \mathbf{M}(\mathbf{y})$ is positive definite (negative definite) for all $\mathbf{y} \neq \mathbf{y}^*$.*

Proof We consider the positive semidefinite case for $\mathbf{M}(\mathbf{y}^*)$. The proof for the negative semidefinite case is analogous. At $\mathbf{y}$ let $\mu(\mathbf{y})$ denote the smallest (possibly negative) eigenvalue of $\mathbf{M}(\mathbf{y})$, with corresponding unit eigenvector $\boldsymbol{\xi}(\mathbf{y})$. For $\gamma \geq 0$ let $\omega(\mathbf{y}) = \mu(\mathbf{y}) + \gamma \|\mathbf{h}(\mathbf{y})\|$ denote the corresponding smallest eigenvalue of $\mathbf{N}(\mathbf{y})$, with corresponding unit eigenvector $\boldsymbol{\xi}(\mathbf{y})$, where $\boldsymbol{\xi}^{\mathsf{T}}\mathbf{N}(\mathbf{y})\boldsymbol{\xi} = \omega(\mathbf{y})\boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{\xi} = \omega(\mathbf{y}) = \mu(\mathbf{y}) + \gamma \|\mathbf{h}(\mathbf{y})\|$. Let $\mathcal{B}_r = \{\mathbf{y} : \|\mathbf{y} - \mathbf{y}^*\| \leq r\}$. From Lemma 7.1 $\mu(\mathbf{y})$ is continuous on $\mathcal{R}^p$, with $\mu(\mathbf{y}^*) \geq 0$ and all the other eigenvalues of $\mathbf{M}(\mathbf{y}^*)$ are positive. For arbitrarily small $\epsilon > 0$ let $\bar{\mathbf{y}}$ be a minimal point for $\mu(\mathbf{y})$ on $\mathcal{B}_\epsilon$. If $\bar{\mathbf{y}} = \mathbf{y}^*$ choose any $\bar{\gamma} > 0$. If $\bar{\mathbf{y}} \neq \mathbf{y}^*$ choose $\bar{\gamma} > \max\{0, -\mu(\bar{\mathbf{y}})/\|\mathbf{h}(\bar{\mathbf{y}})\|\}$. Then for $\gamma > \bar{\gamma}$, $\mu(\mathbf{y}) > 0 \; \forall \, \mathbf{y} \in \mathcal{B}_\epsilon - \{\mathbf{y}^*\}$.

For any $r \geq \epsilon$ let $\mathcal{X}_r = \{\mathbf{y} : \epsilon \leq \|\mathbf{y} - \mathbf{y}^*\| \leq r\}$, with $\|\mathbf{h}(\mathbf{y})\| > 0 \; \forall \; \mathbf{y} \in \mathcal{X}_r$. From the theorem of Weierstrass $\mu(\mathbf{y})/\|\mathbf{h}(\mathbf{y})\|$ takes on a minimum value at some point $\hat{\mathbf{y}} \in \mathcal{X}_r$. Let $\hat{\gamma}(r) = \max\{0, -\mu(\hat{\mathbf{y}})/\|\mathbf{h}(\hat{\mathbf{y}})\|\} \geq 0$. Then for $\gamma > \hat{\gamma}(r)$ we have $\omega(\mathbf{y})/\|\mathbf{h}(\mathbf{y})\| = \gamma + \mu(\mathbf{y})/\|\mathbf{h}(\mathbf{y})\| \geq \gamma + \mu(\hat{\mathbf{y}})/\|\mathbf{h}(\hat{\mathbf{y}})\| \geq \gamma - \hat{\gamma}(r) > 0 \; \forall \; \mathbf{y} \in \mathcal{X}_r$. The conditions on $\mathbf{h}(\mathbf{y})$ ensure that $\|\mathbf{h}(\mathbf{y})\| \not\to 0$ as $\|\mathbf{y} - \mathbf{y}^*\| \to \infty$. Thus $\hat{\gamma} = \lim_{r\to\infty}\{\hat{\gamma}(r)\}$ exists. Then $\omega(\mathbf{y}) > 0 \; \forall \; \mathbf{y} \neq \mathbf{y}^*$ provided $\gamma > \max(\bar{\gamma}, \hat{\gamma})$.

Lemma 7.2 *Let* $\mathbf{A} \in \mathcal{R}^{n\times n}$ *be symmetric and* $\mathbf{B} \in \mathcal{R}^{n\times m}$. *If* $\mathbf{A}$ *is positive definite* ($\mathbf{A} > 0$) *then* $\mathbf{B}^\intercal\mathbf{A}\mathbf{B}$ *is at least positive semidefinite* ($\mathbf{B}^\intercal\mathbf{A}\mathbf{B} \geq 0$).

Proof For $\mathbf{z} \in \mathcal{R}^m$ and $\mathbf{s} \in \mathcal{R}^n$, let $\mathbf{s} = \mathbf{B}\mathbf{z}$. Then $\mathbf{s}^\intercal\mathbf{A}\mathbf{s} > 0$ for $\mathbf{s} \neq \mathbf{0}$. Hence $\mathbf{z}^\intercal\mathbf{B}^\intercal\mathbf{A}\mathbf{B}\mathbf{z} \geq 0$ for $\mathbf{z} \neq \mathbf{0}$.

Lemma 7.3 *Let* $\mathbf{A} \in \mathcal{R}^{n\times n}$ *be symmetric and* $\mathbf{B} \in \mathcal{R}^{n\times n}$. *If* $\mathbf{A} > 0$ (< 0) *and* $\mathbf{B} \geq 0$ (≤ 0) *then* $\mathbf{A} + \mathbf{B} > 0$ (< 0).

Proof For $\mathbf{s} \in \mathcal{R}^n$ we have $\mathbf{s}^\intercal(\mathbf{A} + \mathbf{B})\mathbf{s} = \mathbf{s}^\intercal\mathbf{A}\mathbf{s} + \mathbf{s}^\intercal\mathbf{B}\mathbf{s} > 0$ (< 0) for all $\mathbf{s} \neq \mathbf{0}$.

Theorem 7.2 *For* $\mathbf{A} \in \mathcal{R}^{n\times n}$ *symmetric,* $\mathbf{B} \in \mathcal{R}^{n\times m}$, *and* $\mathbf{D} \in \mathcal{R}^{m\times m}$ *symmetric, the matrix*

$$\mathcal{F} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\intercal & \mathbf{D} \end{bmatrix}$$

is nonsingular, with $|\mathcal{F}| < 0$, *if* $\mathbf{A} > 0$ *and* $\mathbf{D} < 0$ (*or if* $\mathbf{A} < 0$ *and* $\mathbf{D} > 0$).

Proof Pre-multiplying the first block row by $\mathbf{B}^\intercal\mathbf{A}^{-1}$ and subtracting from the second block row yields

$$|\mathcal{F}| = \begin{vmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{D} - \mathbf{B}^\intercal\mathbf{A}^{-1}\mathbf{B} \end{vmatrix} = |\mathbf{A}| \left|\mathbf{D} - \mathbf{B}^\intercal\mathbf{A}^{-1}\mathbf{B}\right|.$$

For $\mathbf{A} > 0$, we have [21, p. 128] $|\mathbf{A}| = \mu_1 \cdots \mu_n > 0$, where μ_k, $k = 1, \ldots, n$, are the eigenvalues of $\mathbf{A}$. From Lemma 7.2 $\mathbf{B}^\intercal\mathbf{A}^{-1}\mathbf{B} \geq 0$. Thus from Lemma 7.3 with $\mathbf{D} < 0$ we have $\mathbf{D} - \mathbf{B}^\intercal\mathbf{A}^{-1}\mathbf{B} < 0$. Hence $\left|\mathbf{D} - \mathbf{B}^\intercal\mathbf{A}^{-1}\mathbf{B}\right| = \omega_1 \cdots \omega_m < 0$, where ω_j, $j = 1, \ldots, m$, are the eigenvalues of $\mathbf{D} - \mathbf{B}^\intercal\mathbf{A}^{-1}\mathbf{B}$. Thus, $|\mathcal{F}| = \mu_1 \cdots \mu_n\omega_1 \cdots \omega_m < 0$.

Theorem 7.3 *If* $\mathbf{y}^* = (\mathbf{x}^*, \boldsymbol{\lambda}^*)$ *is a proper Lagrangian* $\max_{\boldsymbol{\lambda}}$-$\min_{\mathbf{x}}$ *saddle point for a scalar-valued* C^2 *function*

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \phi(\mathbf{x}) - \boldsymbol{\lambda}^\intercal\boldsymbol{\psi}(\mathbf{x}) + \frac{1}{2}\beta\boldsymbol{\psi}^\intercal(\mathbf{x})\boldsymbol{\psi}(\mathbf{x}),$$

with $\mathbf{h}(\mathbf{x}, \boldsymbol{\lambda}) = [\partial\mathcal{L}/\partial\mathbf{x}, \partial\mathcal{L}/\partial\boldsymbol{\lambda}]^\intercal \neq \mathbf{0}$ *for* $\mathbf{y} \neq \mathbf{y}^*$ *and* $\|\mathbf{h}(\mathbf{y})\| \to \infty$ *as* $\|\mathbf{y} - \mathbf{y}^*\| \to \infty$, *and with*

$$\mathcal{H}(\mathbf{y}) = \frac{\partial^2\mathcal{L}(\mathbf{y})}{\partial\mathbf{y}^2} = \begin{bmatrix} \mathcal{H}_{\mathbf{x}\mathbf{x}} & \mathcal{H}_{\mathbf{x}\boldsymbol{\lambda}} \\ \mathcal{H}_{\mathbf{x}\boldsymbol{\lambda}}^\intercal & \mathcal{H}_{\boldsymbol{\lambda}\boldsymbol{\lambda}} \end{bmatrix},$$

then for sufficiently large $\gamma_{\mathbf{x}} \geq 0$ *and any* $\gamma_{\boldsymbol{\lambda}} > 0$ *the matrix*

$$\mathcal{F} = \begin{bmatrix} \gamma_{\mathbf{x}}\|\mathbf{h}\|\mathbf{I}_n + \mathcal{H}_{\mathbf{x}\mathbf{x}} & \mathcal{H}_{\mathbf{x}\boldsymbol{\lambda}} \\ \mathcal{H}_{\mathbf{x}\boldsymbol{\lambda}}^\intercal & -\gamma_{\boldsymbol{\lambda}}\|\mathbf{h}\|\mathbf{I}_m \end{bmatrix}$$

is nonsingular, with $|\mathcal{F}| < 0$, *for all* $\mathbf{y} \neq \mathbf{y}^*$.

Proof From Theorem 7.1, for $\mathbf{y} \neq \mathbf{y}^*$, $\gamma_{\mathbf{x}}\|\mathbf{h}\|\mathbf{I}_n + \mathcal{H}_{\mathbf{x}\mathbf{x}} > 0$ for sufficiently large $\gamma_{\mathbf{x}} \geq 0$ and $-\gamma_{\boldsymbol{\lambda}}\|\mathbf{h}\|\mathbf{I}_m < 0$ for any $\gamma_{\boldsymbol{\lambda}} > 0$. Then $|\mathcal{F}| < 0$ follows from Theorem 7.2 with $\mathbf{A} = \gamma_{\mathbf{x}}\|\mathbf{h}\|\mathbf{I}_n + \mathcal{H}_{\mathbf{x}\mathbf{x}}$, $\mathbf{B} = \mathcal{H}_{\mathbf{x}\boldsymbol{\lambda}}$, and $\mathbf{D} = -\gamma_{\boldsymbol{\lambda}}\|\mathbf{h}\|\mathbf{I}_m$.

7.2 Speed of GEMM

Using $\mathcal{F}$ from (44) write the GEMM algorithm in the form

$$\left\{ \|\mathbf{h}\| \begin{bmatrix} \gamma_{\mathbf{x}}\mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & -\gamma_\lambda \mathbf{I}_m \end{bmatrix} + \begin{bmatrix} \dfrac{\partial^2 \mathcal{L}}{\partial \mathbf{x}^2} & -\mathbf{\Gamma}^{\mathsf{T}} \\ -\mathbf{\Gamma} & \mathbf{0} \end{bmatrix} \right\} \begin{bmatrix} \dot{\mathbf{x}} \\ \dot{\boldsymbol{\lambda}} \end{bmatrix} = - \begin{bmatrix} \boldsymbol{\nabla}_{\mathbf{x}} \mathcal{L} \\ \boldsymbol{\nabla}_{\boldsymbol{\lambda}} \mathcal{L} \end{bmatrix}. \tag{48}$$

Thus, using (45), (27), and (28), we have

$$\|\mathbf{h}\| \begin{bmatrix} \gamma_{\mathbf{x}}\mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & -\gamma_\lambda \mathbf{I}_m \end{bmatrix} \begin{bmatrix} \dot{\mathbf{x}} \\ \dot{\boldsymbol{\lambda}} \end{bmatrix} + \frac{\mathrm{d}\mathbf{h}}{\mathrm{d}t} = -\mathbf{h}.$$

For small $\|\mathbf{h}\|$ or small $\gamma_{\mathbf{x}}$, $\gamma_{\boldsymbol{\lambda}}$ GEMM approaches Newton's method applied to $\mathcal{L}$. For large $\|\mathbf{h}\|$ or large $\gamma_{\mathbf{x}}$, $\gamma_{\boldsymbol{\lambda}}$ GEMM approaches Hestenes's Method of Multipliers applied to $\mathcal{L}/(\gamma_{\mathbf{x}} \|\mathbf{h}\|)$ with $\beta \to \gamma_{\mathbf{x}}/\gamma_\lambda$, that is,

$$\|\mathbf{h}\| \begin{bmatrix} \gamma_{\mathbf{x}}\mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & -\gamma_\lambda \mathbf{I}_m \end{bmatrix} \begin{bmatrix} \dot{\mathbf{x}} \\ \dot{\boldsymbol{\lambda}} \end{bmatrix} \approx -\mathbf{h}$$

yields

$$\begin{bmatrix} \dot{\mathbf{x}} \\ \dot{\boldsymbol{\lambda}} \end{bmatrix} \approx - \begin{bmatrix} \frac{1}{\gamma_{\mathbf{x}}}\mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & -\dfrac{1}{\gamma_\lambda}\mathbf{I}_m \end{bmatrix} \frac{1}{\|\mathbf{h}\|}\mathbf{h} = \frac{1}{\|\mathbf{h}\|} \begin{bmatrix} -\dfrac{1}{\gamma_{\mathbf{x}}}\boldsymbol{\nabla}_{\mathbf{x}}\mathcal{L} \\ \dfrac{1}{\gamma_\lambda}\boldsymbol{\nabla}_{\boldsymbol{\lambda}}\mathcal{L} \end{bmatrix}$$

with "speed"

$$\|\dot{\mathbf{y}}\| \to \begin{cases} 1/\gamma & \text{if} \quad \gamma_{\mathbf{x}} = \gamma_\lambda = \gamma, \\ 1/\gamma_{\mathbf{x}} & \text{if} \quad \gamma_{\mathbf{x}} << \gamma_\lambda, \\ 1/\gamma_\lambda & \text{if} \quad \gamma_\lambda << \gamma_{\mathbf{x}}, \end{cases}$$

for large $\|\mathbf{h}\|$, $\gamma_{\mathbf{x}}$, or $\gamma_{\boldsymbol{\lambda}}$.

7.3 Stability of GEMM

For $\dot{\mathbf{y}} = \mathcal{F}^{-1}\mathbf{h}$ the only equilibrium is at $\mathbf{h} = \boldsymbol{\nabla}_{\mathbf{y}}\mathcal{L} = \mathbf{0}$. As $\|\mathbf{h}\| \to 0$ GEMM approaches Newton's method ($\mathcal{F} \to \mathcal{H}$). Thus $\mathbf{y}^*$ is at least locally asymptotically stable and non-stiff, with all eigenvalues $\mu = -1$. From Theorem 7.3 $\mathcal{F}^{-1}(\mathbf{y})$ exists for all $\mathbf{y}$, provided $\gamma_{\boldsymbol{\lambda}} > 0$ and $\gamma_{\mathbf{x}} \geq 0$ is sufficiently large. Therefore the domain of attraction is all of $\mathcal{R}^p$ and GEMM is globally asymptotically stable to $\mathbf{y}^*$.

8 Rosenbrock's Function with Constraint

As an Example we consider the problem of minimizing Rosenbrock's function

$$\phi(\mathbf{x}) = 100(x_1^2 - x_2)^2 + (1 - x_1)^2, \tag{49}$$

subject to the parabolic constraint

$$\psi(\mathbf{x}) = (x_1 - 2)^2 + x_2 - 1 = 0. \tag{50}$$

Figure 8 shows contours of constant $\phi(\mathbf{x})$, along with the constraint $\psi(\mathbf{x}) = 0$.

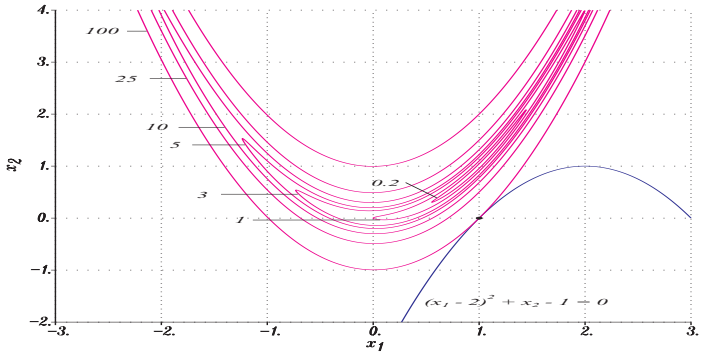

Figure 8: Rosenbrock's function with parabolic constraint.

The gradient and the Hessian matrix of ϕ are given by

$$\boldsymbol{\nabla}\phi(\mathbf{x}) = \left[\frac{\partial\phi}{\partial\mathbf{x}}\right]^{\mathsf{T}} = \begin{bmatrix} 400x_1\left(x_1^2 - x_2\right) + 2\left(x_1 - 1\right) \\ -200\left(x_1^2 - x_2\right) \end{bmatrix}, \tag{51}$$

$$\boldsymbol{\nabla}^2\phi(\mathbf{x}) = \frac{\partial^2\phi}{\partial\mathbf{x}^2} = \begin{bmatrix} 1200x_1^2 - 400x_2 + 2 & -400x_1 \\ -400x_1 & 400 \end{bmatrix}. \tag{52}$$

Rosenbrock's function ϕ is analogous to a curved canyon with very steep walls and a shallow sloping parabolic valley floor, defined by $x_2 = x_1^2$. The function has a single proper unconstrained global minimum at $\hat{\mathbf{x}} = [1, 1]^{\mathsf{T}}$, with $\phi(\mathbf{x}) > 0$ for all $\mathbf{x} \neq \hat{\mathbf{x}}$. Note that $\boldsymbol{\nabla}\phi(\mathbf{x}) \neq \mathbf{0}$ except at $\hat{\mathbf{x}}$ and $\|\boldsymbol{\nabla}\phi(\mathbf{x})\| \to \infty$ as $\|\mathbf{x} - \hat{\mathbf{x}}\| \to \infty$. Hence contours of constant ϕ are topologically equivalent to spheres [14, p. 215]. On the other hand, $\phi(\mathbf{x})$ is not a convex function, that is, it does not satisfy $\phi\left[\theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2\right] \leq \theta\phi(\mathbf{x}_1) + (1-\theta)\phi(\mathbf{x}_2)$ for all $\mathbf{x}_1$, $\mathbf{x}_2$, and $0 \leq \theta \leq 1$. This follows [17, p. 425] from the fact that $\boldsymbol{\nabla}^2\phi(\mathbf{x})$ is positive definite only in the region $x_2 < x_1^2 + 1/2$. Some numerical optimization algorithms have trouble with Rosenbrock's function because they exhibit "stiff" dynamics. For example, applied to the unconstrained problem, the discrete version of Steepest Descent "chatters" along the valley floor.

For the constrained optimization problem the augmented Lagrangian is

$$\begin{aligned} \mathcal{L} &= \phi - \lambda\psi + \frac{1}{2}\beta\psi^2 \\ &= 100(x_1^2 - x_2)^2 + (1 - x_1)^2 - \lambda\left[(x_1 - 2)^2 + x_2 - 1\right] \\ &\quad + \frac{1}{2}\beta\left[(x_1 - 2)^2 + x_2 - 1\right]^2. \end{aligned}$$

With $\mathbf{y}^\intercal = [\mathbf{x}^\intercal, \lambda]$, the gradient of $\mathcal{L}$ is

$$\boldsymbol{\nabla}_{\mathbf{y}}\mathcal{L} = \left[\frac{\partial \mathcal{L}}{\partial \mathbf{y}}\right]^\intercal = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial x_1} \\ \frac{\partial \mathcal{L}}{\partial x_2} \\ \frac{\partial \mathcal{L}}{\partial \lambda} \end{bmatrix} = \begin{bmatrix} \frac{\partial \phi}{\partial x_1} + (\beta\psi - \lambda)\frac{\partial \psi}{\partial x_1} \\ \frac{\partial \phi}{\partial x_2} + (\beta\psi - \lambda)\frac{\partial \psi}{\partial x_2} \\ -\psi \end{bmatrix} \tag{53}$$

$$= \begin{bmatrix} 400(x_1^2 - x_2)x_1 - 2(1 - x_1) + \left(\beta\left[(x_1 - 2)^2 + (x_2 - 1)\right] - \lambda\right) 2(x_1 - 2) \\ -200(x_1^2 - x_2) + \left(\beta\left[(x_1 - 2)^2 + (x_2 - 1)\right] - \lambda\right) \\ -(x_1 - 2)^2 - x_2 + 1 \end{bmatrix}$$

and the Hessian of $\mathcal{L}$ is

$$\mathcal{H} \triangleq \boldsymbol{\nabla}_{\mathbf{y}}^2\mathcal{L} = \frac{\partial^2 \mathcal{L}}{\partial \mathbf{y}^2} = \begin{bmatrix} \frac{\partial^2 \mathcal{L}}{\partial \mathbf{x}^2} & -\left[\frac{\partial \psi}{\partial \mathbf{x}}\right]^\intercal \\ -\frac{\partial \psi}{\partial \mathbf{x}} & 0 \end{bmatrix} = [\mathcal{H}_{ij}], \tag{54}$$

where

$$\mathcal{H}_{11} = \frac{\partial^2 \mathcal{L}}{\partial x_1^2} = \frac{\partial^2 \phi}{\partial x_1^2} + (\beta\psi - \lambda)\frac{\partial^2 \psi}{\partial x_1^2} + \beta\left(\frac{\partial \psi}{\partial x_1}\right)^2 \tag{55}$$

$$= 1200x_1^2 - 400x_2 + 2 + 2\left(\beta\left[(x_1 - 2)^2 + (x_2 - 1)\right] - \lambda\right) + 4\beta(x_1 - 2)^2,$$

$$\mathcal{H}_{12} = \mathcal{H}_{21} = \frac{\partial^2 \mathcal{L}}{\partial x_1 \partial x_2} = \frac{\partial^2 \phi}{\partial x_1 \partial x_2} + (\beta\psi - \lambda)\frac{\partial^2 \psi}{\partial x_1 \partial x_2} + \beta\frac{\partial \psi}{\partial x_1}\frac{\partial \psi}{\partial x_2} \tag{56}$$

$$= -400x_1 + 2\beta(x_1 - 2),$$

$$\mathcal{H}_{13} = \mathcal{H}_{31} = \frac{\partial^2 \mathcal{L}}{\partial x_1 \partial \lambda} = -\frac{\partial \psi}{\partial x_1} = -2(x_1 - 2), \tag{57}$$

$$\mathcal{H}_{22} = \frac{\partial^2 \mathcal{L}}{\partial x_2^2} = \frac{\partial^2 \phi}{\partial x_2^2} + (\beta\psi - \lambda)\frac{\partial^2 \psi}{\partial x_2^2} + \beta\left(\frac{\partial \psi}{\partial x_2}\right)^2 = 200 + \beta, \tag{58}$$

$$\mathcal{H}_{23} = \mathcal{H}_{32} = \frac{\partial^2 \mathcal{L}}{\partial x_2 \partial \lambda} = -\frac{\partial \psi}{\partial x_2} = -1, \tag{59}$$

$$\mathcal{H}_{33} = \frac{\partial^2 \mathcal{L}}{\partial \lambda^2} = 0. \tag{60}$$

The first-order necessary conditions for $\max_\lambda \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda, \beta)$ are:

$$0 = \frac{\partial \mathcal{L}}{\partial x_1} = 400(x_1^2 - x_2)x_1 - 2(1 - x_1) - 2\lambda(x_1 - 2) \tag{61}$$

$$+ \beta\left[(x_1 - 2)^2 + x_2 - 1\right] 2(x_1 - 2),$$

$$0 = \frac{\partial \mathcal{L}}{\partial x_2} = -200(x_1^2 - x_2) - \lambda + \beta\left[(x_1 - 2)^2 + x_2 - 1\right], \tag{62}$$

$$0 = \frac{\partial \mathcal{L}}{\partial \lambda} = -\psi = -(x_1 - 2)^2 - x_2 + 1. \tag{63}$$

At $(\mathbf{x}, \lambda) = (\mathbf{x}^*, \lambda^*)$, with Lagrangian $L(\mathbf{x}, \lambda) = \mathcal{L}(\mathbf{x}, \lambda, \beta)|_{\beta=0}$, using (63) in (62) with $\beta = 0$ yields

$$\lambda = -200\left\{x_1^2 + \left[(x_1 - 2)^2 - 1\right]\right\}. \tag{64}$$

Then using (63) and (64) in (61) yields

$$0 = \frac{\partial L}{\partial x_1} = 800x_1^3 - 2400x_1^2 + 2801x_1 - 1201.$$

This cubic polynomial has roots $x_1 = 1,\ 1 \pm \frac{1}{40}i\sqrt{802}$.

Thus the unique constrained global minimal point is

$$\mathbf{y}^* = (x_1^*, x_2^*, \lambda^*) = (1, 0, -200), \tag{65}$$

with $\phi^* = \phi(\mathbf{x}^*) = 100$ and $\psi^* = \psi(\mathbf{x}^*) = 0$, at which $\boldsymbol{\nabla}_{\mathbf{y}} L(\mathbf{x}^*, \lambda^*) = \boldsymbol{\nabla}_{\mathbf{y}} \mathcal{L}(\mathbf{x}^*, \lambda^*, \beta) = \mathbf{0}$, with

$$\boldsymbol{\nabla}_{\mathbf{x}} L(\mathbf{y}) \neq \mathbf{0} \text{ and } \boldsymbol{\nabla}_{\mathbf{y}} \mathcal{L}(\mathbf{y}) \neq \mathbf{0} \text{ for } \mathbf{y} \neq \mathbf{y}^* \tag{66}$$

and

$$\|\boldsymbol{\nabla}_{\mathbf{x}} L(\mathbf{y})\| \to \infty \text{ and } \|\boldsymbol{\nabla}_{\mathbf{y}} \mathcal{L}(\mathbf{y})\| \to \infty \text{ as } \|\mathbf{y} - \mathbf{y}^*\| \to \infty \tag{67}$$

for all $\beta \geq 0$.

Note that at $\mathbf{y}^*$ the Hessian matrix (12) for the Lagrangian L,

$$H(\mathbf{y}^*) = \frac{\partial^2 L(\mathbf{x}^*, \lambda^*)}{\partial \mathbf{x}^2} = \begin{bmatrix} 1602 & -400 \\ -400 & 200 \end{bmatrix} \tag{68}$$

is positive definite. Thus the second-order sufficient condition (6) is satisfied by the stronger condition that $H(\mathbf{x}^*, \lambda^*)$ is positive definite. For this Example switching from $\max_\lambda \min_{\mathbf{x}} L(\mathbf{x}, \lambda)$ to $\max_\lambda \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda, \beta)$ is not mandatory. However, we will continue using $\mathcal{L}(\mathbf{x}, \lambda, \beta)$, with $L(\mathbf{x}, \lambda) = \mathcal{L}(\mathbf{x}, \lambda, 0)$.

9 Augmented Lagrangian Trajectory Following

9.1 Min-max ascent

Choosing

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{\mathbf{xx}} & \mathbf{P}_{\mathbf{x}\boldsymbol{\lambda}} \\ \mathbf{P}_{\boldsymbol{\lambda}\mathbf{x}} & \mathbf{P}_{\boldsymbol{\lambda}\boldsymbol{\lambda}} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_m \end{bmatrix} \tag{69}$$

in (20) yields

$$\dot{\mathbf{x}} = -\boldsymbol{\nabla}_{\mathbf{x}} \mathcal{L} = -\boldsymbol{\nabla}\phi + \boldsymbol{\Gamma}^{\mathsf{T}} [\boldsymbol{\lambda} - \beta\boldsymbol{\psi}], \tag{70}$$

$$\dot{\boldsymbol{\lambda}} = \boldsymbol{\nabla}_{\boldsymbol{\lambda}} \mathcal{L} = -\boldsymbol{\psi}, \tag{71}$$

which corresponds to steepest descent for $\mathbf{x}$ on $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\beta})$ and steepest ascent for $\boldsymbol{\lambda}$ on $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\beta})$. We will set $\beta = 0$, yielding the Min-Max Ascent algorithm considered in [3].

9.1.1 Simulation results

For the Example problem in Section 8 the Min-Max Ascent equations of motion are

$$\begin{aligned}
\dot{x}_1 &= -400(x_1^2 - x_2)x_1 + 2(1 - x_1) + \left(\lambda - \beta\left((x_1 - 2)^2 + (x_2 - 1)\right)\right) 2\,(x_1 - 2)\,, \\
\dot{x}_2 &= 200(x_1^2 - x_2) + \left(\lambda - \beta\left((x_1 - 2)^2 + (x_2 - 1)\right)\right), \\
\dot{\lambda} &= -\,(x_1 - 2)^2 - x_2 + 1.
\end{aligned} \tag{72}$$

The state perturbation equations $\dot{\boldsymbol{\eta}} = \mathbf{A}\boldsymbol{\eta}$ are

$$\begin{bmatrix} \dot{\boldsymbol{\eta}}_{\mathbf{x}} \\ \dot{\boldsymbol{\eta}}_{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} -\dfrac{\partial^2 \mathcal{L}}{\partial \mathbf{x}^2} & -\dfrac{\partial^2 \mathcal{L}}{\partial \mathbf{x} \partial \boldsymbol{\lambda}} \\ \dfrac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\lambda} \partial \mathbf{x}} & \dfrac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\lambda}^2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\eta}_{\mathbf{x}} \\ \boldsymbol{\eta}_{\mathbf{y}} \end{bmatrix} \tag{73}$$

and yield

$$\begin{bmatrix} \dot{\boldsymbol{\eta}}_1 \\ \dot{\boldsymbol{\eta}}_2 \\ \dot{\boldsymbol{\eta}}_3 \end{bmatrix} = \begin{bmatrix} a_{11} & 400x_1 - 2\beta\,(x_1 - 2) & 2\,(x_1 - 2) \\ 400x_1 - 2\beta\,(x_1 - 2) & -200 - \beta & 1 \\ -2\,(x_1 - 2) & -1 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \\ \boldsymbol{\eta}_3 \end{bmatrix},$$

where

$$a_{11} = -1200x_1^2 + 400x_2 - 2 - 2\left(\beta\left((x_1 - 2)^2 + (x_2 - 1)\right) - \lambda\right) - 4\beta\,(x_1 - 2)^2\,.$$

At $(\mathbf{x}^*, \lambda^*) = (1, 0, -200)$

$$\begin{bmatrix} \dot{\boldsymbol{\eta}}_1 \\ \dot{\boldsymbol{\eta}}_2 \\ \dot{\boldsymbol{\eta}}_3 \end{bmatrix} = \begin{bmatrix} -1602 - 4\beta & 400 + 2\beta & -2 \\ 400 + 2\beta & -200 - \beta & 1 \\ 2 & -1 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \\ \boldsymbol{\eta}_3 \end{bmatrix}.$$

For $\beta = 0$ the eigenvalues (μ_k) and eigenvectors $(\boldsymbol{\xi}_k)$ of $\mathbf{A}$ are

$$\begin{aligned}
\mu_1 &= -5.00 \times 10^{-3}, & \boldsymbol{\xi}_1^{\intercal} &= \begin{bmatrix} 6.23 \times 10^{-8} & 5.00 \times 10^{-3} & 1.0 \end{bmatrix}, \\
\mu_2 &= -93.90, & \boldsymbol{\xi}_2^{\intercal} &= \begin{bmatrix} -0.256 & -0.967 & -4.83 \times 10^{-3} \end{bmatrix}, \\
\mu_3 &= -1708.09, & \boldsymbol{\xi}_3^{\intercal} &= \begin{bmatrix} -0.967 & 0.256 & 1.28 \times 10^{-3} \end{bmatrix}.
\end{aligned} \tag{74}$$

The Lyapunov exponents are $(\sigma_1, \sigma_2, \sigma_3) = (\mu_1, \mu_2, \mu_3)$. This is a very stiff system, with "stiffness" $\Sigma \triangleq |\sigma_{\max} - \sigma_{\min}| \approx 1,700$.

Figure 9 shows Min-Max Ascent trajectories for $\lambda(0) = 0$ and $\beta = 0$, starting from initial $\mathbf{x}(0)$ at the edges of the plot region. The trajectories were generated using standard 4-*th* order Runge-Kutta integration with a fixed step size $\Delta t = 2 \times 10^{-4}$. The trajectories in Figure 9 rapidly approach the valley of Rosenbrock's function then move more slowly along the valley until they reach a region just below the unconstrained minimal point $\hat{\mathbf{x}} = (1, 1)$. Then they move agonizingly slowly down to the constrained minimal point $\mathbf{x}^* = (1, 0)$. For example, the trajectory from $\mathbf{y}(0) = (-2, 4, 0)$ takes approximately 0.1 sec. of simulation time to reach the valley, approximately 3 sec. more to reach a neighborhood

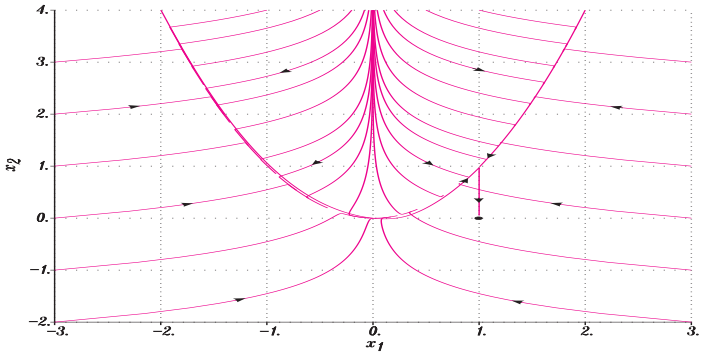

Figure 9: Min-Max Ascent ($\lambda(0) = 0$, $\beta = 0$).

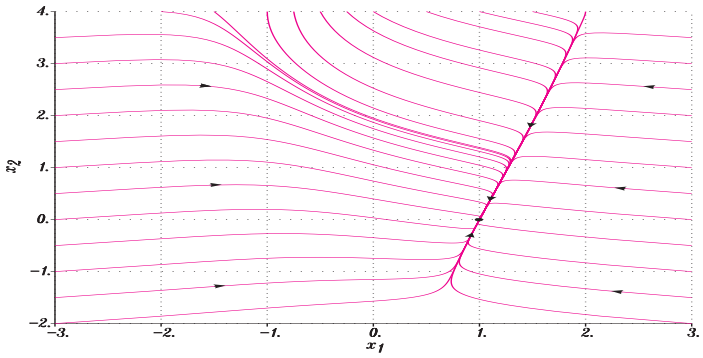

Figure 10: Min-Max Ascent ($\lambda(0) = \lambda^* = -200$, $\beta = 0$).

of $\hat{\mathbf{x}}$, and then more than 1300 sec. longer to converge to $\mathbf{y}^* = (1, 0, -200)$, at $t_f \approx 1400$ sec., with stopping criterion $\|\boldsymbol{\nabla}_{\mathbf{y}} L\| \leq 10^{-3}$. All other trajectories, which are x_1, x_2 projections of the three-dimensional $\mathbf{y} = (x_1, x_2, \lambda)$ trajectories, behaved similarly and converged to $\mathbf{y}^*$, but for plotting purposes were terminated after $t = 0.1$ sec., to illustrate that they overshoot or undershoot the valley.

The extremely slow convergence is associated with λ [due to μ_1 in (74)] and is a result of the choice $\lambda(0) = 0$. The trajectories are essentially steepest descent on $\phi(\mathbf{x})$ until $\lambda(t)$ very slowly converges to λ^*. Figure 10 shows trajectories from the same initial conditions as in Figure 9, except with $\lambda(0) = \lambda^*$. All trajectories, with step size $\Delta t = 10^{-4}$, were terminated when $\|\boldsymbol{\nabla}_{\mathbf{y}} L\| \leq 10^{-3}$, but only required a total of approximately 0.1 sec. of simulation time to converge to $\mathbf{y}^*$.

Figure 11 shows the Lyapunov exponent time histories for Min-Max Ascent on L, starting from $(x_1, x_2\, \lambda) = (-2.5, 0, 0)$ with $\beta = 0$. The system is uniformly very stiff, with $\Sigma(t) = |\sigma_{\max}(t) - \sigma_{\min}(t)|$ varying between approximately 250 and 1,700 and converging to $\Sigma \approx 1,700$.

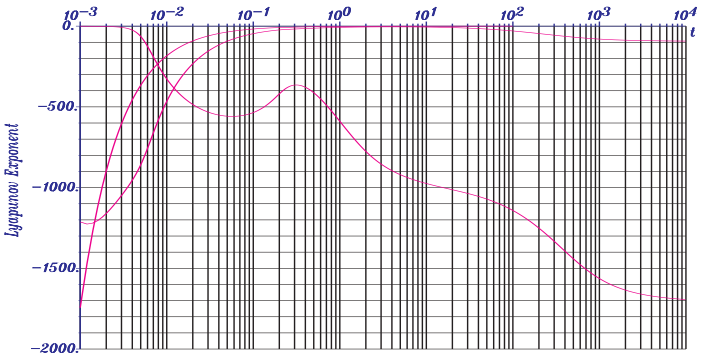

Figure 11: Lyapunov exponents for Min-Max Ascent ($\lambda(0) = 0$, $\beta = 0$).

9.1.2 Stability analysis

Lyapunov's first method establishes that the Min-Max Ascent system (72) is locally asymptotically stable to $\mathbf{y}^*$ for $\beta = 0$. Alternatively, in [4] a min-max sufficiency condition is developed using

$$V(\mathbf{y}) = \frac{1}{2}\left[\mathbf{y} - \mathbf{y}^*\right]^\intercal \left[\mathbf{y} - \mathbf{y}^*\right] = \frac{1}{2}\left[\mathbf{x} - \mathbf{x}^*\right]^\intercal \left[\mathbf{x} - \mathbf{x}^*\right] + \frac{1}{2}\left[\boldsymbol{\lambda} - \boldsymbol{\lambda}^*\right]^\intercal \left[\boldsymbol{\lambda} - \boldsymbol{\lambda}^*\right]. \tag{75}$$

Along $\mathbf{y}(t)$

$$\dot{V}(\mathbf{y}) = \frac{\partial V}{\partial \mathbf{x}}\dot{\mathbf{x}} + \frac{\partial V}{\partial \boldsymbol{\lambda}}\dot{\boldsymbol{\lambda}} = -\left[\mathbf{x} - \mathbf{x}^*\right]^\intercal \boldsymbol{\nabla}_{\mathbf{x}}\mathcal{L} + \left[\boldsymbol{\lambda} - \boldsymbol{\lambda}^*\right]^\intercal \boldsymbol{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}. \tag{76}$$

If

$$\dot{V}(\mathbf{y}) < 0 \quad \text{provided } \dot{\mathbf{y}} \neq \mathbf{0} \tag{77}$$

in a neighborhood of $\mathbf{y}^*$, then $V(\mathbf{y})$ is a Lyapunov function, establishing at least local asymptotic stability [14, p. 217]. The function (75), with $\beta = 0$ so $\mathcal{L} = L$, is used in [3] to establish local asymptotic stability of $\mathbf{y}^*$ for Min-Max Ascent applied to finding a saddle point of $L(\mathbf{x}, \boldsymbol{\lambda})$ under the conditions that L is linear in $\boldsymbol{\lambda}$ and $H = \partial^2 L/\partial \mathbf{x}^2$ is positive definite at $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$.

Unfortunately, for our Example the function (75) does not satisfy (77) everywhere and can not be used to establish global asymptotic stability for our Example. In fact, for some saddle-point problems, Min-Max Ascent can produce Hamiltonian systems [8], which can not be asymptotically stable. However, simulation experiments indicate that Min-Max Ascent is globally asymptotically stable to $\mathbf{y}^*$ for our Example.

9.2 Hestenes' method of multipliers

Choosing

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{\mathbf{xx}} & \mathbf{P}_{\mathbf{x}\boldsymbol{\lambda}} \\ \mathbf{P}_{\boldsymbol{\lambda}\mathbf{x}} & \mathbf{P}_{\boldsymbol{\lambda}\boldsymbol{\lambda}} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & -\beta\mathbf{I}_m \end{bmatrix} \tag{78}$$

in (20) yields

$$\dot{\mathbf{x}} = -\boldsymbol{\nabla}_{\mathbf{x}}\mathcal{L} = -\boldsymbol{\nabla}\phi + \boldsymbol{\Gamma}^\intercal\left[\boldsymbol{\lambda} - \beta\boldsymbol{\psi}\right], \tag{79}$$

$$\dot{\boldsymbol{\lambda}} = \ \boldsymbol{\nabla}_{\boldsymbol{\lambda}}\mathcal{L} = -\beta\boldsymbol{\psi}, \tag{80}$$

corresponding to Hestenes' Method of Multipliers [13].

For our Example the equations of motion for Hestenes' Method of Multipliers are

$$\begin{aligned}
\dot{x}_1 &= -400(x_1^2 - x_2)x_1 + 2(1 - x_1) + \left(\lambda - \beta\left((x_1 - 2)^2 + (x_2 - 1)\right)\right) 2\,(x_1 - 2)\,, \\
\dot{x}_2 &= 200(x_1^2 - x_2) + \left(\lambda - \beta\left((x_1 - 2)^2 + (x_2 - 1)\right)\right), \\
\dot{\lambda} &= -\beta\left[(x_1 - 2)^2 + x_2 - 1\right].
\end{aligned}$$

The state perturbation equations are

$$\begin{bmatrix} \dot{\boldsymbol{\eta}}_1 \\ \dot{\boldsymbol{\eta}}_2 \\ \dot{\boldsymbol{\eta}}_3 \end{bmatrix} = \begin{bmatrix} a_{11} & 400x_1 - 2\beta\,(x_1 - 2) & 2\,(x_1 - 2) \\ 400x_1 - 2\beta\,(x_1 - 2) & -200 - \beta & 1 \\ -2\beta\,(x_1 - 2) & -\beta & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \\ \boldsymbol{\eta}_3 \end{bmatrix},$$

where

$$a_{11} = -1200x_1^2 + 400x_2 - 2 - 2\left(\beta\left((x_1 - 2)^2 + (x_2 - 1)\right) - \lambda\right) - 4\beta\,(x_1 - 2)^2\,.$$

At $(\mathbf{x}^*, \lambda^*)$

$$\begin{bmatrix} \dot{\boldsymbol{\eta}}_1 \\ \dot{\boldsymbol{\eta}}_2 \\ \dot{\boldsymbol{\eta}}_3 \end{bmatrix} = \begin{bmatrix} -1602 - 4\beta & 400 + 2\beta & -2 \\ 400 + 2\beta & -200 - \beta & 1 \\ 2\beta & -\beta & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \\ \boldsymbol{\eta}_3 \end{bmatrix}.$$

For $\beta = 5$ the state perturbation equations have Lyapunov exponents $(\sigma_1, \sigma_2, \sigma_3)$ equal to the eigenvalues $(\mu_1, \mu_2, \mu_3) = \left(-2.44 \times 10^{-2}, -94.91\,, -1732.07\right)$. This is a very stiff system, with stiffness $\Sigma = |\sigma_{\max} - \sigma_{\min}| \approx 1,700$, which is approximately that of Min-Max Ascent. For $\beta = 100$ the eigenvalues are $(\mu_1, \mu_2, \mu_3) = (-0.33, -109.64, -2192.03)$, with stiffness $\Sigma = |\sigma_{\max} - \sigma_{\min}| \approx 2,200$. For very large β the state perturbation matrix

$$\mathbf{A} \approx \begin{bmatrix} -4\beta & 2\beta & -2 \\ 2\beta & -\beta & 1 \\ 2\beta & -\beta & 0 \end{bmatrix}$$

has eigenvalues $(\mu_1, \mu_2, \mu_3) = \left(0, -\frac{5}{2}\beta + \frac{1}{2}\sqrt{25\beta^2 - 20\beta}, -\frac{5}{2}\beta - \frac{1}{2}\sqrt{25\beta^2 - 20\beta}\right) \rightarrow (0, 0, -5\beta)$. Thus the stiffness increases with increasing β. However, as we shall show later, even for $\beta = 5$ the convergence for the Method of Multipliers is much faster than for Min-Max Ascent.

As with Min-Max Ascent, Lyapunov's first method establishes local asymptotic stability of $\mathbf{y}^*$, but no suitable Lyapunov function is known to establish global asymptotic stability. However, experimental simulation results indicate that Hestenes' Method of Multipliers is globally asymptotically stable for our Example.

Figure 12 shows trajectories for Hestenes' Method of Multipliers applied to the augmented Lagrangian $\mathcal{L}$ with $\lambda(0) = 0$ and $\beta = 5$, using step size $\Delta t = 10^{-4}$ with termination when $\|\boldsymbol{\nabla}_{\mathbf{y}}\mathcal{L}\| \leq 10^{-3}$. The behavior is similar to Min-Max Ascent, except for faster convergence to the constrained minimum point, at approximately $t_f = 300$ sec.

Figure 13 shows the Lyapunov exponent time histories for Hestenes Method of Multipliers applied to $\mathcal{L}$, starting from $(x_1, x_2\,\lambda) = (-2.5, 0, 0)$ with $\beta = 5$. The stiffness is similar to Min-Max Ascent, with $\Sigma(t)$ varying between approximately 100 and $1,700$ and converging to $\Sigma \approx 1,700$.

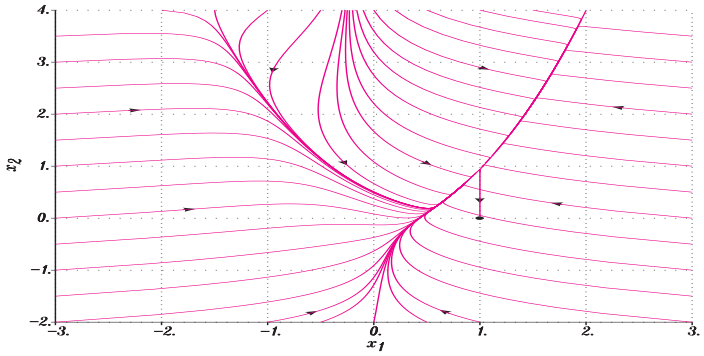

Figure 12: Method of Multipliers ($\lambda(0) = 0$, $\beta = 5$).

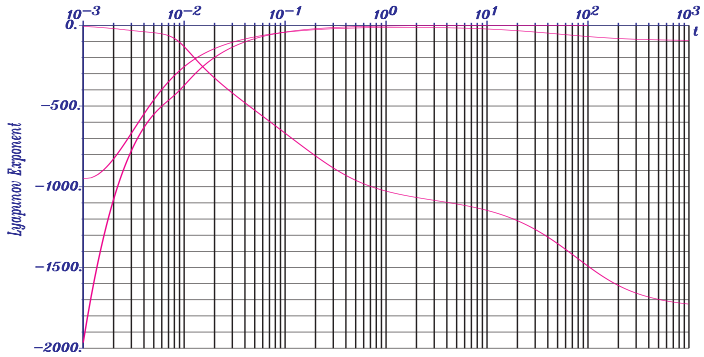

Figure 13: Method of Multipliers Lyapunov exponents ($\lambda(0) = 0$, $\beta = 5$).

9.3 Newton's method

Choosing

$$\mathbf{P} = \begin{bmatrix} \mathbf{P_{xx}} & \mathbf{P_{x\lambda}} \\ \mathbf{P_{\lambda x}} & \mathbf{P_{\lambda\lambda}} \end{bmatrix} = \mathcal{H}^{-1} = \begin{bmatrix} \dfrac{\partial^2 \mathcal{L}}{\partial \mathbf{y}^2} \end{bmatrix}^{-1} = \begin{bmatrix} \dfrac{\partial^2 \mathcal{L}}{\partial \mathbf{x}^2} & -\mathbf{\Gamma}^{\mathsf{T}} \\ -\mathbf{\Gamma} & \mathbf{0} \end{bmatrix}^{-1}, \tag{81}$$

in (20), where $\mathbf{\Gamma} = \partial \boldsymbol{\psi} / \partial \mathbf{x}$, yields Newton's method

$$\begin{bmatrix} \dot{\mathbf{x}} \\ \dot{\boldsymbol{\lambda}} \end{bmatrix} = -\mathcal{H}^{-1}(\mathbf{y}) \begin{bmatrix} \boldsymbol{\nabla}_{\mathbf{x}} \mathcal{L} \\ \boldsymbol{\nabla}_{\boldsymbol{\lambda}} \mathcal{L} \end{bmatrix} = -\mathcal{H}^{-1}(\mathbf{y}) \begin{bmatrix} \boldsymbol{\nabla}_{\mathbf{x}} \mathcal{L} \\ -\boldsymbol{\psi} \end{bmatrix}. \tag{82}$$

At the constrained minimal point the state perturbation equations (30) and (32) have $\mathbf{A}(\mathbf{y}^*) = -\mathbf{I}$, with eigenvalues $\mu = -1$, yielding a non stiff system.

For our Example problem,

$$\mathcal{H} = \begin{bmatrix} \mathcal{H}_{11} & -400x_1 + 2\beta\,(x_1 - 2) & -2\,(x_1 - 2) \\ -400x_1 + 2\beta\,(x_1 - 2) & 200 + \beta & -1 \\ -2\,(x_1 - 2) & -1 & 0 \end{bmatrix},$$

where

$$\mathcal{H}_{11} = 1200x_1^2 - 400x_2 + 2 + 2\left(\beta\left((x_1-2)^2 + (x_2-1)\right) - \lambda\right) + 4\beta\,(x_1-2)^2\,.$$

Then

$$\mathbf{P} = \mathcal{H}^{-1} = \frac{1}{|\mathcal{H}|}\,\mathrm{adj}(\mathcal{H}), \tag{83}$$

where

$$|\mathcal{H}| = -\,(3600+2\beta)\,x_1^2 + (6400+8\beta)\,x_1 + (400-2\beta)\,x_2 - 3202 - 6\beta + 2\lambda \tag{84}$$

and the adjugate matrix is

$$\mathrm{adj}(\mathcal{H}) = \begin{bmatrix} -1 & 2\,(x_1-2) & 800(x_1-1) \\ 2\,(x_1-2) & -4\,(x_1-2)^2 & c_{23} \\ 800(x_1-1) & c_{32} & c_{33} \end{bmatrix} \tag{85}$$

with

$$\begin{aligned} c_{23} = c_{32} &= (2\beta+2000)\,x_1^2 - (8\beta+1600)\,x_1 + (2\beta-400)\,x_2 + 2 + 6\beta - 2, \\ c_{33} &= \left(2\beta^2 + 4000\beta + 80\,000\right)x_1^2 - 8\beta\,(1000+\beta)\,x_1 \\ &\quad + \left(2\beta^2 - 80\,000\right)x_2 + 6\beta^2 + 4402\beta - \lambda\,(400+2\beta) + 400. \end{aligned}$$

At points where $|\mathcal{H}| = 0$ the inverse $\mathcal{H}^{-1}$ fails to exist. As a result, Newton's method is not globally asymptotically stable to the solution point $(x_1^*, x_2^*, \lambda^*) = (1, 0, -200)$ for our Example problem. Specifically, we have $|\mathcal{H}| = 0$ on the parabola

$$(400-2\beta)\,x_2 = (3600+2\beta)\,x_1^2 - (6400+8\beta)\,x_1 + 3202 - 2\lambda + 6\beta.$$

At the optimal point $\mathbf{y}^* = (1, 0, -200)$ with $\beta = 0$

$$|H| = -3600x_1^2 + 6400x_1 + 400x_2 - 3402 = -602.$$

Furthermore,

$$H_{\mathbf{xx}}^* = \frac{\partial^2 L(\mathbf{x}^*, \boldsymbol{\lambda}^*)}{\partial \mathbf{x}^2} = \begin{bmatrix} 802 & -400 \\ -400 & 200 \end{bmatrix}$$

has $|H_{\mathbf{xx}}^*| = 400$ and is positive definite. However, $L(\mathbf{x}^*, \boldsymbol{\lambda}) = L(\mathbf{x}^*, \boldsymbol{\lambda}^*)\,\forall\,\boldsymbol{\lambda}$, since $\boldsymbol{\psi}(\mathbf{x}^*) = \mathbf{0}$. Thus $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ is not a *proper* saddle point, since $H_{\lambda\lambda} = \partial^2 L/\partial\lambda^2 \equiv 0$ instead of $H_{\lambda\lambda} < 0$. However, at $\mathbf{y}^*$ with $\beta = 0$

$$\mathcal{H}^*|_{\beta=0} = \begin{bmatrix} \frac{\partial^2 L}{\partial x_1^2} & \frac{\partial^2 L}{\partial x_1 \partial x_2} & -\frac{\partial \psi}{\partial x_1} \\ \frac{\partial^2 L}{\partial x_1 \partial x_2} & \frac{\partial^2 L}{\partial x_2^2} & -\frac{\partial \psi}{\partial x_2} \\ -\frac{\partial \psi}{\partial x_1} & -\frac{\partial \psi}{\partial x_2} & 0 \end{bmatrix}_{\substack{\mathbf{y}^* \\ \beta=0}} = \begin{bmatrix} 802 & -400 & 2 \\ -400 & 200 & -1 \\ 2 & -1 & 0 \end{bmatrix}$$

is indefinite, with $|\mathcal{H}^*|_{\beta=0} = -2 \neq 0$. Thus $\mathbf{x}^*$ is a "nonsingular" point [13], so there exists $\beta \geq 0$ such that $\mathcal{H}_{\mathbf{xx}}^* = \partial^2 \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*)/\partial \mathbf{x}^2$ is positive definite. In our case $\beta = 0$ suffices, since $H_{\mathbf{xx}}^*$ is already positive definite.

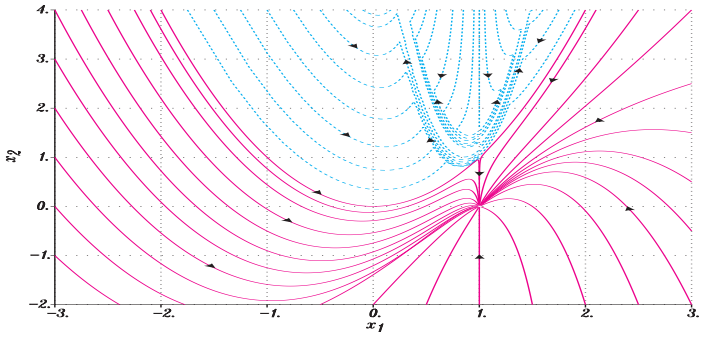

Figure 14: Newton's method ($\lambda(0) = 0$, $\beta = 0$).

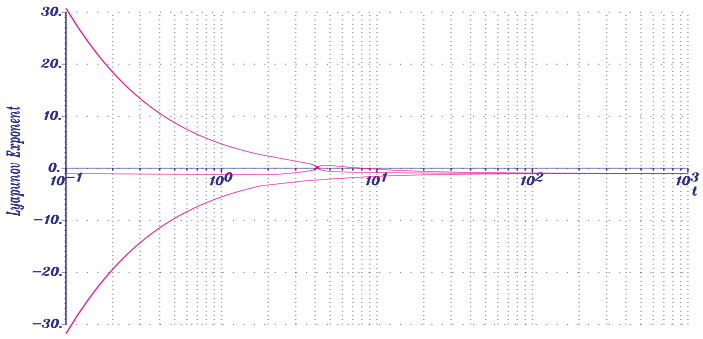

Figure 15: Lyapunov exponents for Newton's method on L.

Figure 14 shows trajectories for Newton's method for $\lambda(0) = 0$ and $\beta = 0$. All of the solid trajectories, generated with standard 4-*th* order Runge–Kutta with fixed step size $\Delta t = 2 \times 10^{-3}$, rapidly converge to $(x_1^*, x_2^*, \lambda^*) = (1, 0, -200)$. The dashed trajectories all reach points where $|\mathcal{H}| = 0$ and do not converge to the constrained minimal point. These trajectories were generated using Branin's method [5], [6], in which $|\mathcal{H}| = 0$ problems are avoided by replacing $\mathcal{H}^{-1}$ in (83) with $\text{adj}(\mathcal{H})$ from (85). The resulting $[\mathbf{x}(t), \lambda(t)]$ trajectories are the same as for Newton's method except for the plot speed and the direction of motion when $|\mathcal{H}| = 0$ surfaces are "crossed".

Figure 15 shows the Lyapunov exponent time histories for Newton's method applied to L, starting from $(x_1, x_2\, \lambda) = (-2.5, 0, 0)$ with $\beta = 0$. The system is initially moderately stiff but achieves $\Sigma(t) = |\sigma_{\max}(t) - \sigma_{\min}(t)| < 10$ in approximately 1 sec., with $\Sigma(t) \to 0$ as all of the Lyapunov exponents converge to $\sigma_k = -1$.

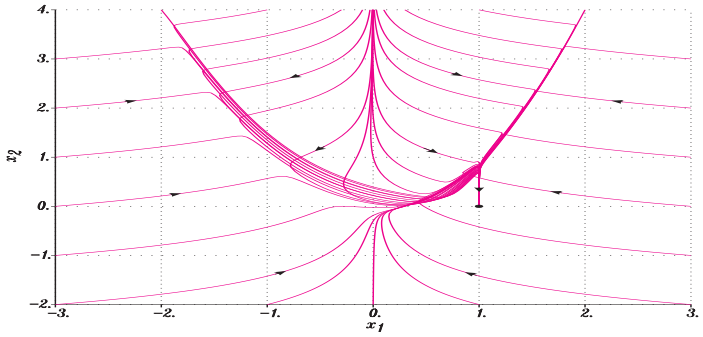

Figure 16: Gradient Enhanced Min-Max ($\lambda(0) = 0$, $\beta = 0$).

9.4 Gradient enhanced min-max

For GEMM applied to our Example problem of Rosenbrock's function with a parabolic constraint, we have from (44)

$$\mathcal{F} = \begin{bmatrix} \gamma_x \|\mathbf{h}\| + \mathcal{H}_{11} & \mathcal{H}_{12} & -2(x_1 - 2) \\ \mathcal{H}_{21} & \gamma_x \|\mathbf{h}\| + \mathcal{H}_{22} & -1 \\ -2(x_1 - 2) & -1 & -\gamma_\lambda \|\mathbf{h}\| \end{bmatrix},$$

where

$$\mathcal{H} = [\mathcal{H}_{ij}] = \begin{bmatrix} \mathcal{H}_{11} & -400x_1 + 2\beta(x_1 - 2) & -2(x_1 - 2) \\ -400x_1 + 2\beta(x_1 - 2) & 200 + \beta & -1 \\ -2(x_1 - 2) & -1 & 0 \end{bmatrix},$$

$$\mathbf{h} = \boldsymbol{\nabla}_{\mathbf{y}} \mathcal{L} = \begin{bmatrix} 400(x_1^2 - x_2)x_1 - 2(1 - x_1) + \left(\beta\left[(x_1 - 2)^2 + (x_2 - 1)\right] - \lambda\right) 2(x_1 - 2) \\ -200(x_1^2 - x_2) + \left(\beta\left[(x_1 - 2)^2 + (x_2 - 1)\right] - \lambda\right) \\ -(x_1 - 2)^2 - x_2 + 1 \end{bmatrix}$$

with $\mathcal{H}_{11} = 1200x_1^2 - 400x_2 + 2 + 4\beta(x_1 - 2)^2 + 2\left(\beta\left[(x_1 - 2)^2 + (x_2 - 1)\right] - \lambda\right)$.

From Theorem 7.2, for $\mathbf{y} \neq \mathbf{y}^*$ and $\gamma_\lambda > 0$, we have $|\mathcal{F}| < 0$ provided $\gamma_{\mathbf{x}} > 0$ is sufficiently large so that $\gamma_{\mathbf{x}} \|\mathbf{h}\| \mathbf{I}_n + \mathcal{H}_{\mathbf{xx}}$ is positive definite for all $\mathbf{y} \neq \mathbf{y}^*$. We choose $\gamma_{\mathbf{x}} = 10$ and $\gamma_\lambda = 0.1$. For our Example system Figure 16 shows Gradient Enhanced Min-Max (GEMM) trajectories applied to the Lagrangian L with step size $\Delta t = 1$. Figure 17 shows Gradient Enhanced Min-Max trajectories applied to the augmented Lagrangian $\mathcal{L}$, with $\beta = 5$ and step size $\Delta t = 1$.

Figure 18 shows the Lyapunov exponent time histories for GEMM applied to L, starting from $(x_1, x_2\, \lambda) = (-2.5, 0, 0)$ with $\beta = 0$, $\gamma_{\mathbf{x}} = 10$, and $\gamma_\lambda = 0.1$. The system is uniformly non stiff, with $\max \Sigma(t) < 1$ and $\Sigma(t) \to 0$ as all of the Lyapunov exponents converge to $\sigma_k = -1$.

10 Constrained Trajectory Following Performance

For each Gradient Transformation algorithm in Section 4 simulation experiments were conducted for a variety of parameter combinations, with the algorithms being applied to

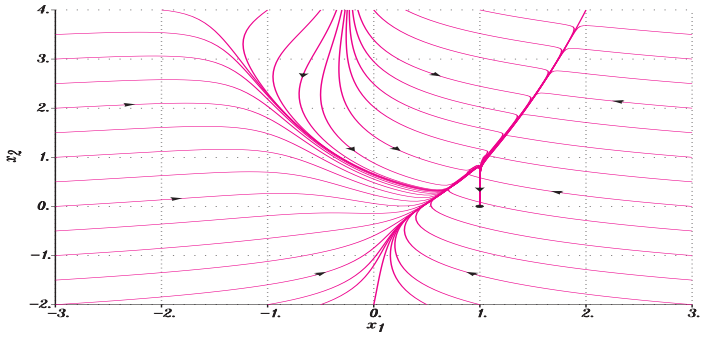

Figure 17: Gradient Enhanced Min-Max ($\lambda(0) = 0$, $\beta = 5$).

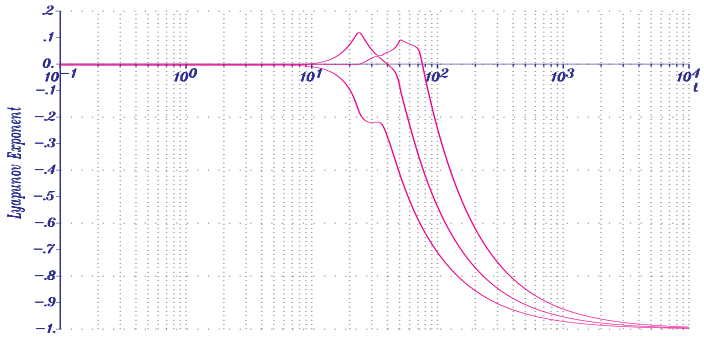

Figure 18: Lyapunov exponents for GEMM on L.

both the Lagrangian L and the augmented Lagrangian $\mathcal{L}$. Table 3 shows the parameter values for each Gradient Transformation trajectory following method that we studied. For comparison, all trajectories were started at a point $\mathbf{x}(0) = (-2.5, 0)$ from which all the algorithms converged to the constrained minimal point $\mathbf{y}^* = (1, 0, -200)$.

Table 4 shows step sizes and simulation results for the methods in Table 3. For each algorithm a trajectory $\mathbf{y}(t)$ was computed starting from $\mathbf{x}(0) = (-2.5, 0)$ using standard 4-*th* order Runge–Kutta with fixed step size Δt, determined to control the approximate initial single step displacement $\Delta s = \|\mathbf{y}(\Delta t) - \mathbf{y}(0)\|$. The trajectories were terminated when $\|\boldsymbol{\nabla}_{\mathbf{y}}\mathcal{L}\| \leq 10^{-3}$. For reference, we include results for Min-Max Ascent starting with $\lambda(0) = \lambda^*$, which yields fairly fast convergence to $\mathbf{y}^*$. All other simulations were started with $\lambda(0) = 0$.

11 Summary

For the problem of minimizing a scalar-valued function subject to equality constraints, the Gradient Transformation family of differential equation algorithms includes, as special cases: Min-Max Ascent, Newton's method, Hestenes' Method of Multipliers, and a Gradient Enhanced Min-Max (GEMM) algorithm extended to handle equality constraints. Applied to Rosenbrock's function with a parabolic constraint, we find that

Table 3: Gradient Transformation algorithm parameters.

Method	$\lambda(0)$	β	γ_x	γ_λ
Min-Max L	0	0	0	0
Min-Max L^*	-200	0	0	0
Hestenes $\mathcal{L}_1$	0	5	0	0
Hestenes $\mathcal{L}_2$	0	100	0	0
Newton L	0	0	0	0
Newton $\mathcal{L}_1$	0	5	0	0
Newton $\mathcal{L}_2$	0	100	0	0
GEMM L	0	0	10	0.1
GEMM $\mathcal{L}_1$	0	5	10	0.1
GEMM $\mathcal{L}_2$	0	100	10	0.1

Table 4: Simulation results for Gradient Transformation algorithms.

Method	**Speed** $\|\dot{\mathbf{y}}(0)\|$	Δt	$\|\mathbf{y}(\Delta t) - \mathbf{y}(0)\|$	**Final** t	**# Steps**	**Ratio**
Min-Max L	6.381×10^3	2×10^{-4}	0.747	1380.364	6901820	92024
Min-Max L^*	8.125×10^3	1×10^{-4}	0.590	0.095	950	13
Hestenes $\mathcal{L}_1$	8.060×10^3	1×10^{-4}	0.568	288.016	2880161	38402
Hestenes $\mathcal{L}_2$	4.104×10^4	2×10^{-5}	0.630	24.965	1248253	16643
Newton L	3.055×10^2	2×10^{-3}	0.610	15.670	7835	104
Newton $\mathcal{L}_1$	3.494×10^2	2×10^{-3}	0.698	15.904	7952	106
Newton $\mathcal{L}_2$	1.053×10^3	5×10^{-4}	0.526	17.529	35058	467
GEMM L	9.394×10^{-2}	1	0.094	75	75	1
GEMM $\mathcal{L}_1$	9.303×10^{-2}	1	0.093	75	75	1
GEMM $\mathcal{L}_2$	9.300×10^{-2}	1	0.093	78	78	1

Min-Max Ascent is globally asymptotically stable but very stiff and has very slow convergence. Hestenes' Method of Multipliers is also globally asymptotically stable and has faster convergence, but is still very slow and very stiff. Newton's method is not stiff, but does not yield global asymptotic stability. However, GEMM is both globally asymptotically stable and not stiff. The stiffness of the Gradient Transformation family is studied in terms of Lyapunov exponent time histories. Starting from points where all the methods in this paper do work, we show that Min-Max Ascent and Hestenes' Method of Multipliers are very stiff and slow to converge, but with the Method of Multipliers being approximately 2 times as fast as Min-Max Ascent. Newton's method, where it works, is not stiff and is approximately 900 times as fast as Min-Max Ascent and 400 times as fast as the Method of Multipliers. In contrast, the Gradient Enhanced Min-Max method is globally convergent, is not stiff, and is approximately 100 times faster than Newton's method, $40,000$ times faster than the Method of Multipliers, and $90,000$ times faster than Min-Max Ascent

References

[1] Vincent, T. L. and Grantham, W. J. Trajectory Following Methods in Control System Design. *J. of Global Optimization* **23** (2002) 267–282.

[2] Goh, B. S. Algorithms for Unconstrained Optimization Problems Via Control Theory. *J. of Optimization Theory and Applications* **92** (3) (1997) 581–604.

[3] Arrow, K. J., Hurwicz, L., and Uzawa, H. *Studies in Linear and Non-Linear Programming*. Stanford Univ. Press, Stanford, California, 1958.

[4] Vincent, T. L., Goh, B. S., and Teo, K. L. Trajectory-Following Algorithms for Min-Max Optimization Problems. *J. of Optimization Theory and Applications* (75) (3) (1992) 501–519.

[5] Gomulka, J. Remarks on Branin's Method for Solving Nonlinear Equations. *Towards Global Optimization* (L. C. W. Dixon and G. P. Szegö, eds.). North-Holland, Amsterdam, 1972, pp. 96–106.

[6] Gomulka, J. Two Implementations of Branin's Method: Numerical Experience. *Towards Global Optimization 2* (L. C. W. Dixon and G. P. Szegö, eds.). North-Holland, Amsterdam, 1978, pp. 151–164.

[7] Grantham, W. J. Trajectory Following Optimization by Gradient Transformation Differential Equations. In: *Proc. 42nd I.E.E.E. Conf. on Decision and Control*. Maui, HI, December 2003, pp. 5496–5501.

[8] Grantham, W. J. Gradient Transformation Trajectory Following Algorithms for Determining Stationary Min-Max Saddle Points. *Advances in Dynamic Game Theory and Applications* (S. J. Jørgensen, T. L. Vincent and M. Quincampoix, eds.). Birkhaüser, Boston, 2006.

[9] Vincent, T. L. and Grantham, W. J. *Optimality in Parametric Systems*. Wiley-Interscience, New York., 1981.

[10] Fletcher, R. *Practical Methods of Optimization*. Wiley, New York, 1987.

[11] Bertsekas, D. P. Multiplier Methods: A Survey. *Automatica* **12** (2) (1976) 133–145.

[12] Bertsekas, D. P. *Constrained Optimization and Lagrange Multiplier Methods*. Athene Scientific, Belmont, MA, 1996.

[13] Hestenes, M. R. Multiplier and Gradient Methods. *J. of Optimization Theory and Applications* **4** (1969) 303–320.

[14] Vincent, T. and Grantham, W. *Nonlinear and Optimal Control Systems*. Wiley, New York, 1997.

[15] Grantham, W. J. and Lee, B. S. A Chaotic Limit Cycle Paradox. *Dynamics and Control* **3** (2) (1993) 159–173.

[16] Wolf, A., Swift, J. B., Swinney, H. L., and Vastano, J. A. Determining Lyapunov Exponents from a Time Series. *Physica 16D* (1985) 285–317.

[17] Chong, E. K. P. and Zak, S. H. *An Introduction to Optimization*. Wiley, New York, 2nd ed., 2001.

[18] Goh, B. S. Greatest Descent Algorithms in Unconstrained Optimization. *J. of Optimization Theory and Applications* **142** (2009) 275–289.

[19] Luenberger, D. G. *An Introduction to Dynamic Systems*. Wiley, New York, 1979.

[20] Kato, T. *A Short Introduction to Perturbation Theory for Linear Operators*. Springer-Verlag, New York, 1982.

[21] Brogan, W. *Modern Control Theory*, Prentice-Hall, Englewood Cliffs, 1982.

Nonlinear Dynamics and Systems Theory, **10** (2) (2010) 161–174

Practical Stability and Controllability for a Class of Nonlinear Discrete Systems with Time Delay ⋆

Zhan Su [1,2], Qingling Zhang [2] and Wanquan Liu [3*]

[1] *College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning province, 110004, PR China*
[2] *Institute of Systems Science, Northeastern University, Shenyang, Liaoning province, 110004, PR China*
[3] *Department of Computing, Curtin University of Technology, Perth WA 6102, Australia*

Received: October 27, 2009; Revised: March 26, 2010

Abstract: In this paper we first study the problem of practical asymptotic stability for a class of discrete-time time-delay systems via Razumikhin-type Theorems. Further the estimations of solution boundary and arrival time of the solution into a region are also investigated based on the practical stability results. Finally, the result on practical asymptotic stability is used to analyze the practical controllability of a general class of nonlinear discrete systems with input time delay. Some explicit criteria for the uniform practical asymptotic stability are derived via Lyapunov function and Razumikhin technique. For illustration, a numerical example is given to show the effectiveness of the proposed results.

Keywords: *practical stability; practical controllability; Razumikhin techniques; discrete systems; time delay.*

Mathematics Subject Classification (2000): 70K20, 93C55, 39A10, 03C45, 93D20, 93B05, 37B25, 39A22.

⋆ This work was supported by a grant from National Natural Science Foundation of China with grant number (60574011).
* Corresponding author: mailto:W.Liu@curtin.edu.au

1 Introduction

Since Lasalle first introduced the concept of practical stability in [1], it attracts much attention in control community. Many works on practical stability have been published with broad applications in different areas. Being much different from stability in terms of Lyapunov functions, practical stability, which stabilizes a system into a region of phase space, is a significant performance specification from engineering point of view, and this idea is quite satisfactory in many applications for quality analysis. In practice, a system is actually unstable, just because the stable domain or the domain of the desired attractor is not large enough; or sometimes, the desired state of a system may be mathematically unstable, yet the system may oscillate sufficiently near to a state, in which the performance is still acceptable, i.e., it is stable in practice. For example, in practical communication or digital control systems, the signals of controller states, measurement outputs, and control inputs are quantized and then encoded for transmission. A feedback law, which global asymptotically stabilizes a given system without quantization, will in general fail to guarantee global asymptotic stability of the closed-loop system, which arises in the presence of a quantizer with a finite number of values. Instead of using the global asymptotic stability, the practice stability can be used to analyze such systems, where there is a region of attraction in the state and the steady state converges to a small limit cycle [2]–[6]. On the other hand, it is well known that for more than one hundred years, Lyapunov's direct method has been the primary technique for dealing with stability problems in difference equations. However, the construction of Lyapunov's function is much more difficult for time-delay systems than for non-delay systems. Such difficulties can be overcome via using Lyapunov functions and Razumikhin techniques. It should be pointed out that the Razumikhin-type method can deal with the time-delay system effectively and is easier to apply in general, therefore such a method has been a main technique for analyzing the stability for time-delay systems [7]–[10].

Though there are several results on the practical stability for hybrid and descriptor systems [11]–[17], to the best of our knowledge, the Razumikhin-type method on practical stability for discrete time-delay systems has not been investigated. Motivated by results in [9], we will study the Razumikhin-type theorem on practical asymptotic stability for a class of discrete time-delay system in this paper. Also estimations of the solution boundary and arrival time of the solution are discussed. Consequently, the proposed theorems are used to study the practical controllability of a general class of nonlinear discrete systems with input time delay. Some explicit criteria for the uniform practical asymptotic stability are obtained via using the Lyapunov function and Razumikhin technique.

This paper is organized as follows. In Section 2, some definitions and preliminaries are introduced. In Section 3, some criteria for uniform practical asymptotic stability of discrete-time systems with finite delay are derived via using the Lyapunov functions and Razumikhin-technique. In Section 4, estimation of the solution boundary and arrival time of the solution are investigated in terms of practical stability. In Section 5, the proposed theorems are used to analyze the practical controllability for a general class of nonlinear discrete systems with input time delay. In Section 6, a numerical example is given to illustrate the effectiveness of main results obtained in Section 5. The last section gives some conclusions.

2 Preliminaries

To describe the main result of this paper, we include some preliminary knowledge on the practical stability for the following general class of nonlinear discrete systems with finite time delay:

$$x(k+1) = F(k, x_d(k)), \quad k \in \mathbb{Z}^+, \tag{1}$$

where $\mathbb{Z}^+$ is the set of nonnegative integers, $d \geq 0$ is an integer, $x(k) \in \mathbb{R}^n$, $x_d(k) = (x^T(k), x^T(k-1), \ldots, x^T(k-d))^T$, $\mathbb{R}^n$ is the n-dimensional Euclidean space. Denote

$$I_d = \{-d, -d+1, \ldots, -1, 0\}, \;\; I_d^1 = I_d \cup \{1\},$$

$$\Xi(I_d, \mathbb{R}^n) = \{\xi_d = (\xi^T(0), \xi^T(-1), \ldots, \xi^T(-d))^T \,|\, \xi : I_d \to \mathbb{R}^n\},$$

$$\Xi_B(I_d, \mathbb{R}^n) = \Xi(I_d, \mathbb{R}^n) \cap \{\xi_d : \xi(s) \in B,\, s \in I_d\},$$

where B is an open ball. Assume $F : \mathbb{Z}^+ \times \Xi_B(I_d, \mathbb{R}^n) \to \mathbb{R}^n$ with $F(k, 0) = 0$ for $k \in \mathbb{Z}^+$, and satisfies certain conditions to guarantee the global existence and uniqueness of solutions. Thus system (1) has zero solution $x(\cdot) \equiv 0$. For any $k_0 \in \mathbb{Z}^+$ and any given initial function $\phi \in \Xi_B(I_d, \mathbb{R}^n)$, the solution of the systems (1) denoted by $x(k; k_0, \phi)$ satisfies (1) for all integers $k \geq k_0$, and $x(k_0 + s; k_0, \phi) = \phi(s)$ for all $s \in I_d$.

For all $\xi_d \in \Xi(I_d, \mathbb{R}^n)$, define the norm of ξ_d as $\|\xi_d\| = \max\limits_{s \in I_d} |\xi(s)|$, where $|\cdot|$ stands for any norm in $\mathbb{R}^n$. We further assume that there exists a constant $L > 0$ such that for all $\xi_d \in \Xi_B(I_d, \mathbb{R}^n)$,

$$|F(k, \xi_d)| \leq L\|\xi_d\|, \quad \forall k \in \mathbb{Z}^+. \tag{2}$$

Now we introduce the following definitions.

Definition 2.1 [9] A wedge function is a continuous strictly increasing function $W : \mathbb{R}^+ \to \mathbb{R}^+$ with $W(0) = 0$.

Definition 2.2 System (1) is said to be:

Practically Stable (P.S.): For given (α, β) with $0 < \alpha < \beta$ and some $k_0 \in \mathbb{Z}^+$, if $\|\phi\| < \alpha$ then $|x(k; k_0, \phi)| < \beta$, $k \geq k_0$;

Uniformly Practically Stable (U.P.S.): If P.S. holds for all $k_0 \in \mathbb{Z}^+$;

Practically Asymptotically Stable (P.A.S.): If P.S. holds, and for each $\varepsilon \in (0, \beta)$, there exists a positive number $K = K(k_0, \alpha, \varepsilon)$ such that $\|\phi\| < \alpha$ implies $|x(k; k_0, \phi)| < \varepsilon$, $k \geq k_0 + K$;

Uniformly Practically Asymptotically Stable (U.P.A.S.): If P.A.S. holds for all $k_0 \in \mathbb{Z}^+$.

Definition 2.3 For a function $V : \mathbb{Z}^+ \times \mathbb{R}^n \to \mathbb{R}^+$, define:

$$\Delta V(k, x(k)) \triangleq V(k+1, x(k+1)) - V(k, x(k)).$$

3 Razumikhin-type Theorems

In this section we will prove some Razumikhin-type theorems with the aim of analyzing the uniform practical asymptotical stability (U.P.A.S.) for a general class of nonlinear discrete systems with finite time delay. We first denote the balls B_0, B_1 and B_2 as the following forms, which will be used in main theorems:

$$\begin{aligned} B_0 &= \{x(k):\ V(k,x(k)) < W_2(\alpha)\};\\ B_1 &= \{x(k):\ V(k,x(k)) < W_1(\beta)\};\\ B_2 &= \{x(k):\ V(k,x(k)) < W_1(\varepsilon)\}. \end{aligned}$$

Theorem 3.1 *Given positive scalars α and β. Assume that scalars ϖ_1, ϖ_2, ϖ_3 with $0 < \varpi_1 \le \varpi_2$, $\varpi_3 > 0$ are all arbitrary. If there exist a scalar $\eta > 0$, a Lyapunov function $V : \mathbb{Z}^+ \times \mathbb{R}^n \to \mathbb{R}^+$, and wedge functions $W_i(\cdot)(i = 1, 2, 3)$, such that*

$$\begin{aligned} &(i)\quad W_1(|x(k)|) \le V(k,x(k)) \le W_2(|x(k)|);\\ &(ii)\quad \Delta V(k,x(k)) \le -W_3(|x(k+1)|) + \varpi_3 \text{ for } \varepsilon_0 \le \|x_d\| \le \rho_0, \end{aligned}$$

provided $\varepsilon_0 \le \rho_0$, $V(k+s, x(k+s)) \le \min\{\varpi_2, V(k+1, x(k+1)) + \eta\}$ for $s \in I_d^1$, and $\varpi_1 \le V(k+1, x(k+1))$. Here $\varepsilon_0 = L^{-1}\alpha$, $\rho_0 = \max\{\beta, W_1^{-1}(W_2(\alpha))\}$, L is defined by (2). Then, we have (1) B_0 is an invariable set; (2) If $W_2(\alpha) < W_1(\beta)$, then B_1 is an invariable set and there exists a positive number $K = K(\alpha, \beta)$ such that for any $k_0 \in \mathbb{Z}$, $\phi \in \Xi_{B_1}(I_d, \mathbb{R}^n)$ implies $\forall\ k \ge k_0 + K$, $x(k; k_0, \phi) \in B_0$.

Proof (1) For each $\phi \in \Xi_{B_0}(I_d, \mathbb{R}^n)$, we have $x(k; k_0, \phi) \in B_0$ for $k_0 - d \le k \le k_0$. Now we claim that for all $k \ge k_0$, $x = x(k; k_0, \phi) \in B_0$.

Suppose this is not true. Then there exist some $k^1 \ge k_0$ such that $x \in B_0$ for all $k_0 - d \le k \le k^1$, and

$$V(k^1 + 1, x(k^1 + 1)) \ge W_2(\alpha), \tag{3}$$

and consequently,

$$\Delta V(k^1, x(k^1)) = V(k^1 + 1, x(k^1 + 1)) - V(k^1, x(k^1)) > 0.$$

On the other hand, by condition (i), we have $W_1(|x(k)|) < W_2(\alpha)$ for $k_0 - d \le k \le k^1$, which implies $\|x_d(k)\| \le \rho_0$ for $k_0 \le k \le k^1$. It follows from (2), (3) and condition (i) that $\alpha \le |x(k^1 + 1)| \le L\|x_d(k^1)\| \le L\rho_0$, which implies $\varepsilon_0 \le \|x_d(k^1)\| \le \rho_0$, $\varepsilon_0 \le \rho_0$. Let $0 < \varpi_1 \le W_2(\alpha) \le W_2(L\rho_0) \le \varpi_2$, and $0 < \varpi_3 < W_3(\alpha)$. Then, it follows from (3) that $\varpi_1 \le V(k^1 + 1, x(k^1 + 1))$, and for $\eta > 0$, $\forall\ s \in I_d^1$,

$$\begin{aligned} &\begin{cases} V(k^1 + s, x(k^1 + s)) < \varpi_2\\ V(k^1 + s, x(k^1 + s)) < V(k^1 + 1, x(k^1 + 1)) + \eta \end{cases}\\ \Longrightarrow\quad &V(k^1 + s, x(k^1 + s)) \le \min\{\varpi_2, V(k^1 + 1, x(k^1 + 1)) + \eta\}. \end{aligned}$$

By condition (ii), we have $\Delta V(k^1, x(k^1)) \le -W_3(|x(k^1 + 1)|) + \varpi_3 < 0$. This is a contradiction. Thus for all $k \ge k_0$, $x(k) \in B_0$, i.e., B_0 is an invariable set.
(2) If $W_2(\alpha) < W_1(\beta)$, we first prove that B_1 is an invariable set. In fact, $\rho_0 = \beta$,

and $\varepsilon_0 = L^{-1}\alpha < L^{-1}W_2^{-1}(W_1(\beta))$. Similar to the proof of (1), one can derive that, $\phi \in \Xi_{B_1}(I_d, \mathbb{R}^n)$ implies $x(k) \in B_1$ for all $k \geq k_0$.

Next, we will find an integer $K = K(\alpha, \beta) > 0$ such that for all $k_0 \in \mathbb{Z}^+$, $\phi \in \Xi_{B_1}(I_d, \mathbb{R}^n)$ implies $x(k; k_0, \phi) \in B_0$ for all $k \geq k_0 + K$.

Assume that $0 < \varpi_1 \leq W_2(\alpha) < W_1(\beta) \leq \varpi_2$, $0 < \varpi_3 < (1/2)W_3(\alpha)$. Let N_η be the first positive integer satisfying

$$W_1(\beta) < W_2(\alpha) + \eta N_\eta. \tag{4}$$

For each $i \in \{0, 1, \ldots, N_\eta\}$, let $k_i = k_0 + i(d + \left[\frac{W_1(\beta)}{\varpi_3}\right])$, where $[\cdot]$ denotes the greatest integer function, η is dependent on ϖ_1 and ϖ_3. We show that for all $i \in \{0, 1, \ldots, N_\eta\}$,

$$V(k, x(k)) < W_2(\alpha) + \eta(N_\eta - i), \quad \forall\, k \geq k_i. \tag{5}$$

Obviously, it follows (4) that (5) holds for $i = 0$ since $x(k) \in B_1$ for all $k \geq k_0$. Suppose (5) holds for some $i \in \{0, 1, \ldots, N_\eta - 1\}$, we aim to show that (5) also holds for $i + 1$, i.e.,

$$V(k, x(k)) < W_2(\alpha) + \eta(N_\eta - i - 1), \quad \forall\, k \geq k_{i+1}.$$

Next we present proof in two steps for clarity.

Step 1. We show that there does exist some $k' \in [k_i + d, k_{i+1}]$ such that

$$V(k', x(k')) < W_2(\alpha) + \eta(N_\eta - i - 1). \tag{6}$$

Suppose this is not true, for all $k \in [k_i + d, k_{i+1}]$, we would have

$$V(k, x(k)) \geq W_2(\alpha) + \eta(N_\eta - i - 1). \tag{7}$$

Noting the assumption that (5) holds for some $i \in \{0, 1, \ldots, N_\eta - 1\}$, then, for all $k \in [k_i + d, k_{i+1} - 1]$, $s \in I_d^1$, from (7) we have

$$V(k + s, x(k + s)) < W_2(\alpha) + \eta(N_\eta - i) \leq V(k + 1, x(k + 1)) + \eta.$$

On the other hand, for all $k \in [k_i + d, k_{i+1} - 1]$, it follows from condition (i), (2) and (7) that $W_2(\alpha) \leq V(k+1, x(k+1)) \leq W_2(|x(k+1)|)$, which implies that $\alpha \leq |x(k+1)| \leq L\rho_0$, $\varepsilon_0 \leq \|x_d(k)\| \leq \rho_0$, $\varepsilon_0 \leq \rho_0$. Then, for all $k \in [k_i + d, k_{i+1} - 1]$, $V(k + s, x(k + s)) \leq \varpi_2$, $s \in I_d^1$, and it follows from (7) that $V(k + 1, x(k + 1)) \geq \varpi_1$. By condition (ii), for all $k \in [k_i + d, k_{i+1} - 1]$,

$$\Delta V(k, x(k)) \leq -W_3(|x(k + 1)|) + \varpi_3 < -\varpi_3.$$

Hence, we have

$$\begin{aligned} V(k_{i+1}, x(k_{i+1})) &\leq V(k_i + d, x(k_i + d)) - \varpi_3(k_{i+1} - k_i - d) \\ &< W_1(\beta) - \varpi_3\left[\frac{W_1(\beta)}{\varpi_3}\right] < 0. \end{aligned}$$

This is a contradiction to the definition of Lyapunov function $V(k, x(k))$. Thus, there does exist some $k' \in [k_i + d, k_{i+1}]$ such that (6) holds.

Step 2. We need to show that

$$V(k, x(k)) < W_2(\alpha) + \eta(N_\eta - i - 1), \quad \forall\, k \geq k'. \tag{8}$$

In fact, suppose this is not true, there must be some $k_1' \geq k'$ such that

$$\begin{aligned} V(k_1', x(k_1')) &< W_2(\alpha) + \eta(N_\eta - i - 1), \\ V(k_1' + 1, x(k_1' + 1)) &\geq W_2(\alpha) + \eta(N_\eta - i - 1). \end{aligned} \tag{9}$$

Hence we have $\Delta V(k_1', x(k_1')) > 0$. On the other hand, $\varpi_1 \leq W_2(\alpha) \leq V(k_1' + 1, x(k_1' + 1))$, $V(k_1' + s, x(k_1' + s)) \leq \varpi_2$. Noting the assumption that (5) holds for some $i \in \{0, 1, \ldots, N_\eta - 1\}$, then, we have for $s \in I_d^1$,

$$V(k_1' + s, x(k_1' + s)) < W_2(\alpha) + \eta(N_\eta - i) \leq V(k_1' + 1, x(k_1' + 1)) + \eta.$$

From condition (i), (2) and (9), we have $W_2(\alpha) \leq V(k_1' + 1, x(k_1' + 1)) \leq W_2(|x(k_1' + 1)|)$, and hence, $\alpha \leq |x(k_1' + 1)| \leq L\rho_0$, $\varepsilon_0 \leq \|x_d(k_1')\| \leq \rho_0$, $\varepsilon_0 \leq \rho_0$. With condition (ii), one can derive that

$$\Delta V(k_1', x(k_1')) \leq -W_3(|x(k_1' + 1)|) + \varpi_3 \leq -\varpi_3 < 0.$$

This is a contradiction again to the definition of Lyapunov function $V(k, x(k))$. Thus (8) holds, and consequently, (5) holds for all $i \in \{0, 1, \ldots, N_\eta\}$. Therefore, we obtain that $x(k) \in B_0$ for all $k \geq k_{N_\eta} = k_0 + K$, where $K = N_\eta(d + \left[\frac{W_1(\beta)}{\varpi_3}\right])$ is independent of k_0 and ϕ. □

Corollary 3.1 *Given positive scalars α and β and assume that $P_V(s) \in C(\mathbb{R}^+, \mathbb{R}^+)$ with $P_V(s) > s$ for $s > 0$. If there exist a Lyapunov function $V : \mathbb{Z}^+ \times \mathbb{R}^n \to \mathbb{R}^+$, and wedge functions $W_i(\cdot)(i = 1, 2, 3)$, satisfying the conditions (i) in Theorem 3.1 and the following condition :*

$$\textit{(ii)}' \quad \Delta V(k, x(k)) \leq -W_3(|x(k + 1)|) \textit{ for } \varepsilon_0 \leq \|x_d(k)\| \leq \rho_0,$$

provided $\varepsilon_0 \leq \rho_0$, $V(k+s, x(k+s)) < P_V(V(k+1, x(k+1)))$ for $s \in I_d^1$, where $\varepsilon_0 = L^{-1}\alpha$, $\rho_0 = \max\{\beta, W_1^{-1}(W_2(\alpha))\}$, L is defined by (2). Then, the conclusion of Theorem 3.1 still holds.

Proof For any $0 < \varpi_1 \leq \varpi_2$, and any $\varpi_3 > 0$, choose $\eta \in (0, \inf\{P_V(s) - s : \varpi_1 \leq s \leq \varpi_2\})$. Then, if $V(k+s, x(k+s)) \leq \min\{\varpi_2, V(k+1, x(k+1)+\eta)\}$ for $s \in I_d^1$, and $\varpi_1 \leq V(k + 1, x(k + 1))$, we have

$$V(k + s, x(k + s)) \leq V(k + 1, x(k + 1)) + \eta < P_V(V(k + 1, x(k + 1))),$$

for $s \in I_d^1$. Hence, by condition (ii)$'$, we have

$$\Delta V(k, x(k)) \leq -W_3(|x(k + 1)|) \leq -W_3(|x(k + 1)|) + \varpi_3.$$

Then, the conditions (i) and (ii) in Theorem 3.1 are both satisfied. Therefore, the result follows. □

By using Theorem 3.1 and Corollary 3.1, we obtain the following Razumikhin-type theorem for the U.P.A.S. with regard to the zero solution of systems (1).

Theorem 3.2 *For given scalar pair (α, β) with $0 < \alpha < \beta$, $\varepsilon \in (0, \beta)$ is arbitrary. Assume that scalars ϖ_1, ϖ_2, ϖ_3 with $0 < \varpi_1 \leq \varpi_2$, $\varpi_3 > 0$ are all arbitrary, $P_V(s) \in$*

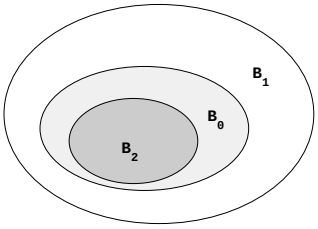

Figure 1: The relationship of the balls B_0, B_1 and B_2.

$C(\mathbb{R}^+,\mathbb{R}^+)$ *with* $P_V(s) > s$ *for* $s > 0$. *If there exist a Lyapunov function* $V : \mathbb{Z}^+ \times \mathbb{R}^n \to \mathbb{R}^+$, *wedge functions* $W_i(\cdot)(i = 1, 2, 3)$, *satisfying*

$$\begin{aligned} &(i) \quad W_2(\alpha) \le W_1(\beta);\\ &(ii) \quad W_1(|x(k)|) \le V(k, x(k)) \le W_2(|x(k)|); \end{aligned}$$

and either of the following conditions $(iii)_a$ *or* $(iii)_b$ *for* $\varepsilon_0 \le \|x_d(k)\| \le \rho_0$, $\varepsilon_0 \le \rho_0$:

$$\begin{aligned} (iii)_a \quad & \Delta V(k, x(k)) \le -W_3(|x(k+1)|) + \varpi_3, \text{ provided}\\ & V(k+s, x(k+s)) \le \min\{\varpi_2, V(k+1, x(k+1)) + \eta\}\\ & \text{for } s \in I_d^1, \text{ and } \varpi_1 \le V(k+1, x(k+1));\\ (iii)_b \quad & \Delta V(k, x(k)) \le -W_3(|x(k+1)|), \text{ provided for } s \in I_d^1,\\ & V(k+s, x(k+s)) < P_V(V(k+1, x(k+1))), \end{aligned}$$

where $\varepsilon_0 = L^{-1}W_2^{-1}(W_1(\varepsilon))$, $\rho_0 = \beta$, L *is defined by (2). Then the zero solution of systems (1) is U.P.A.S.*

Proof By condition (i), $B_0 \subseteq B_1$, as shown in Fig 1. Since

$$\varepsilon_0 = L^{-1}W_2^{-1}(W_1(\varepsilon)) < L^{-1}W_2^{-1}(W_1(\beta)),$$

Then, by Theorem 3.1 and Corollary 3.1, we can assert that, both B_1 and B_2 are invariant sets, and there exists a positive number $K = K(\alpha, \varepsilon)$ such that for any $k_0 \in \mathbb{Z}$, $\phi \in \Xi_{B_0}(I_d, \mathbb{R}^n)$ implies $\forall\ k \ge k_0 + K$, $x(k; k_0, \phi) \in B_2$. By condition (ii), $|x(k)| < \alpha$ implies $x(k) \in B_0$; $x(k) \in B_1$ implies $|x(k)| < \beta$; $x(k) \in B_2$ implies $|x(k)| < \varepsilon$. Then, for any $k_0 \in \mathbb{Z}$, $\|\phi\| < \alpha$ implies $\forall\ k \ge k_0 + K$, $|x(k; k_0, \phi)| < \varepsilon$, i.e., the zero solution of the systems (1) is U.P.A.S. □

Remark 3.1 In Theorem 3.1 and Corollary 3.1, whenever $\varepsilon_0 \le L^{-1}W_2^{-1}(W_1(\beta))$, $\rho_0 \ge \beta$ and $\Delta V(k, x(k)) \le 0$ in the conditions (ii) and (ii)′, one can obtain the result that B_1 is an invariable set. Here, the conditions of Theorem 3.1 and Corollary 3.1 are corresponding to the case that u, v, w are wedge functions in the conditions of Theorem 1 and Corollary 1 in [9]. Moreover, it is more convenient to apply Theorem 3.1 and Corollary 3.1 in this paper to estimate relations between balls B_0 and B_1 in the light of information on $\varepsilon_0 \le \|x_d(k)\| \le \rho_0$, which are not mentioned in Theorem 1, Corollary 1 and Corollary 2 in [9].

4 Estimation of the Solution Boundary and Arrival Time

Now let us consider Theorem 3.2, Corollary 3.1 and Theorem 3.2 from previous section without the condition $W_2(\alpha) < W_1(\beta)$. If $\varepsilon_0 = L^{-1}W_2^{-1}(W_1(\beta))$, $\rho_0 = \max\{\beta, W_1^{-1}(W_2(\alpha))\}$, then we can assert that B_1 is an invariant set. In addition, the trajectory of the solution of system (1) starting from B_0, will fall into B_1 in finite time when $B_0 \supset B_1$, or stay in the region of B_1 when $B_0 \subseteq B_1$. On the other hand, with the assumption of $W_2(\alpha) < W_1(\beta)$, all trajectories which exit from the ball B_0, will take the ball B_1 to be their boundary and can not get out of the region of B_1. Thus, by the proposed theorems and Remark 3.1, as long as $\varepsilon_0 \leq L^{-1}W_2^{-1}(W_1(\beta))$, and $\Delta V(k, x(k)) \leq 0$ in conditions $(\text{iii})_a \sim (\text{iii})_b$, the system is U.P.S.. Following the above analysis, one can observe that it is more convenient to apply Theorem 3.1 and Corollary 3.1 in this paper to estimate relations between the balls B_0 and B_1 by using the information on $\varepsilon_0 \leq \|x_d(k)\| \leq \rho_0$, which are not discussed in Theorem 1, Corollary 1 and Corollary 2 in [9]. We give the following theorem to estimate both the boundary of the solution of system (1) and arrival time K.

Theorem 4.1 *Given scalars α, ε with $0 < \varepsilon < \alpha$, $\sigma_1 > 1$. If there exist a Lyapunov function $V : \mathbb{Z}^+ \times \mathbb{R}^n \to \mathbb{R}^+$, wedge functions $W_i(\cdot)(i = 1, 2, 3)$, satisfying*

$$\begin{aligned} &(i) \quad W_1(|x(k)|) \leq V(k, x(k)) \leq W_2(|x(k)|); \\ &(ii) \quad \Delta V(k, x(k)) \leq -W_3(|x(k+1)|) \, for \|x_d(k)\| \leq \rho_0, \, provided \\ &\qquad V(k+s, x(k+s)) < \sigma_1(V(k+1, x(k+1))) \, for \, s \in I_d^1, \end{aligned}$$

then
(1) $\beta_\alpha = W_1^{-1}(W_2(\alpha))$;
(2) $K = k_0 + m_1 \, (d + m_2)$,
where

$$m_1 = \begin{cases} \frac{W_2(\alpha)+(\sigma_1-2)W_1(\varepsilon)}{(\sigma_1-1)W_1(\varepsilon)}, & \frac{W_2(\alpha)-W_1(\varepsilon)}{(\sigma_1-1)W_1(\varepsilon)} \text{ is integer}; \\ \left[\frac{W_2(\alpha)-W_1(\varepsilon)}{(\sigma_1-1)W_1(\varepsilon)}\right], & \text{otherwise}, \end{cases}$$

$$m_2 = \begin{cases} \frac{2W_2(\alpha)}{W_3(W_2^{-1}(W_1(\varepsilon)))} + 1, & \frac{2W_2(\alpha)}{W_3(W_2^{-1}(W_1(\varepsilon)))} \text{ is integer}; \\ \left[\frac{2W_2(\alpha)}{W_3(W_2^{-1}(W_1(\varepsilon)))}\right], & \text{otherwise}, \end{cases}$$

$[\cdot]$ denotes the greatest integer function, $\rho_0 = W_1^{-1}(W_2(\alpha))$, β_α is the estimation of the solution boundary of system (1), and K is the time that the solution exists from the given ball $\{\phi : \|\phi\| < \alpha\}$ and falls into the region $\{x(k) : \|x(k)\| < \varepsilon\}$.

Proof (1) In Theorem 3.1, let $W_1(\beta_\alpha) = W_2(\alpha)$. Then, $\varepsilon_0 = L^{-1}\alpha$, $\rho_0 = \beta_\alpha$, and $B_1 = B_2 = \{x(k) : V < W_1(\beta_\alpha)\}$. It follows from Theorem 3.1 that the solution starting from B_2 can not exits from B_1, which implies that the solution starting from set $\{\phi : \|\phi\| < \alpha\}$ will have a boundary $\beta_\alpha = W_1^{-1}(W_2(\alpha))$.

(2) In Theorem 3.1, let $B_1 = \{x(k) : V(k, x(k)) < W_2(\alpha)\}$, and $B_2 = \{x(k) : V(k, x(k)) < W_1(\varepsilon)\}$. Notice that $\varepsilon_0 = L^{-1}W_2^{-1}(W_1(\varepsilon))$ and $\rho_0 = W_1^{-1}(W_2(\alpha))$, let $P_V(s) = \sigma_1 s$, then $P_V(s)$ has the required property in Corollary 3.1 and there exist two scalars $\delta_1 > 0$ and $\delta_2 \in (0, 1/2)$, such that $\frac{W_2(\alpha)-W_1(\varepsilon)}{(\sigma_1-1)W_1(\varepsilon)} < \frac{W_2(\alpha)-W_1(\varepsilon)}{(\sigma_1-1)W_1(\varepsilon)-\delta_1} < m_1$, and $\frac{2W_2(\alpha)}{W_3(W_2^{-1}(W_1(\varepsilon)))} < \frac{W_2(\alpha)}{\delta_2 W_3(W_2^{-1}(W_1(\varepsilon)))} < m_2$. With the similar analysis process in the proofs of Theorem 3.1 and Corollary 3.1, one can derive the conclusion of (2) for $\varepsilon \leq \|x(k)\| < \alpha$ with $\eta = (\sigma_1 - 1)W_1(\varepsilon) - \delta_1 \in (0, \inf(P_V(V) - V))$ and $\varpi_3 = \delta_2 W_3(W_2^{-1}(W_1(\varepsilon)))$. □

5 Practical Controllability

In this section we will use the results from previous sections to study the practical controllability for a general class of nonlinear discrete systems with input time delay. Consider the following system:

$$x(k+1) = f(k, x(k)) + \sum_{i=0}^{d} B(k-i)u(k-i), \tag{10}$$

where $f : \mathbb{Z}^+ \times \mathbb{R}^n \to \mathbb{R}^n$, $B : \mathbb{Z}^+ \times \mathbb{R}^n \to \mathbb{R}^{n\times m}$, $i = 1, \ldots, d$, $u(k) \in \mathbb{R}^m$ is input, and is supposed to guarantee the existence and uniqueness of the solution. This type of model is generally studied in networked control systems (NCSs). We first introduce the following definitions:

Definition 5.1 System (10) is called to be:

Uniformly Practically Controllable (U.P.C.) with respect to (α, β), $0 < \alpha < \beta$, if there exist finite time K and a control $u(\cdot)$ defined on $[k_0, K]$ such that all the solutions $x(k) = x(k; k_0, \phi, u)$ that exit from $\{\phi \in \mathbb{R}^n : \|\phi\| < \alpha\}$ will enter into a bounded region $\{x(k) \in \mathbb{R}^n : \|x_d(k)\| < \beta\}$ at time K instant for all $k_0 \in \mathbb{Z}^+$;

Uniformly Practically Asymptotically Controllable (U.P.A.C.) with respect to (α, β), $0 < \alpha < \beta$, if U.P.C. holds, and for each $\varepsilon \in (0, \beta)$, there exists a positive number $K = K(k_0, \alpha, \varepsilon)$ such that $\|\phi\| < \alpha$ implies $|x(k; k_0, \phi, u)| < \varepsilon$ for all $k \geq k_0 + K$.

Theorem 5.1 *Assume that there exists a control law $u(k)$ such that system (10) can be expressed by the form of (1), and the conditions of Theorem 3.2 are satisfied. Then, system (10) is U.P.A.C.with respect to (α, β).*

For system (10), adopt the feedback control law $u(k) = F(k, x(k))x(k)$. Assume $f_u(k, x(k)) = f(k, x(k)) + B(k)u(k)$ and

$$\|f_u(k, x(k))\| \leq \|\Psi_0(k)\|\|x(k)\|.$$

Let $\Psi_i(k) = B(k-i)F(k-i, x(k-i))$, where $F(k, x(k))$ is the control gain matrix, $\Psi_0(k)$ and $\Psi_i(k)$ are of compatible dimensions. Consequently, the closed-loop system of (10) has the following form:

$$x(k+1) = f_u(k, x(k)) + \sum_{i=1}^{d} \Psi_i(k)x(k-i). \tag{11}$$

Let $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ be the maximum eigenvalue and the minimum eigenvalue of a real symmetric matrix, respectively. $\|\cdot\|_2$ stands for the Euclidean vector norm or the 2-norm of a matrix. Then, we have the following corollary.

Corollary 5.1 *If there exists $F(k)$ such that*

$$\sup_{k\in\mathbb{Z}} \sum_{i=0}^{d} \|\Psi_i(k)\|_2^2 < 1 - (\frac{\alpha}{\beta})^2 \tag{12}$$

then, the closed-loop system (11) is U.P.A.S., and system (10) is U.P.A.C. with respect to (α, β) with $0 < \alpha < \beta$.

Proof In fact, by (12), noting that $0 < \alpha < \beta$, then, $\forall \epsilon \in (0, \alpha^2/\beta^2)$, there exist scalars $\delta_1 \in [\alpha^2/\beta^2, 1]$ and $\delta_2 > 1$ such that

$$\sup_{k\in\mathbb{Z}} \|\Psi_0(k)\|_2^2 + \delta_2 \sup_{k\in\mathbb{Z}} \sum_{i=1}^{d} \|\Psi_i(k)\|_2^2 < \delta_1 - (\frac{\alpha}{\beta})^2 + \epsilon < \delta_1.$$

Thus, there exists a positive definite matrix P such that $\lambda_{\min}(P) = \delta_1 \lambda_{\max}(P)$. Choose $V(k, x(k)) = x^T(k)Px(k)$, $W_1(|x(k)|) = \lambda_{\min}(P)x^T(k)x(k)$, and $W_2(|x(k)|) = \lambda_{\max}(P)x^T(k)x(k)$. It is obvious that

$$W_1(|x(k)|) \le V(k, x(k)) \le W_2(|x(k)|).$$

Let $P_V(s) = \delta_2 s$ for $s \ge 0$. Then $P_V(s) > s$ for $s \ge 0$. For all $i \in \{1, \ldots, d\}$, if $V(k-i, x(k-i)) < P_V(V(k+1, x(k+1)))$, then, $\|x(k-i)\|_2^2 < \|x(k+1)\|_2^2 \delta_2/\delta_1$, and it follows (11) that

$$\begin{aligned} \|x(k+1)\|_2^2 &\le \sup_{k\in\mathbb{Z}} \|\Psi_0(k)\|_2^2 \|x(k)\|_2^2 \\ &\quad + \sup_{k\in\mathbb{Z}} \sum_{i=1}^{d} \|\Psi_i(k)\|_2^2 \|x(k-i)\|_2^2 \\ &\le \sup_{k\in\mathbb{Z}} \|\Psi_0(k)\|_2^2 \|x(k)\|_2^2 \\ &\quad + \frac{\delta_2 \sup_{k\in\mathbb{Z}} \sum_{i=1}^{d} \|\Psi_i(k)\|_2^2}{\delta_1} \|x(k+1)\|_2^2. \end{aligned}$$

Consequently,

$$-\|x(k)\|_2^2 \le \frac{\delta_2 \sup_{k\in\mathbb{Z}} \sum_{i=1}^{d} \|\Psi_i(k)\|_2^2 - \delta_1}{\delta_1 \sup_{k\in\mathbb{Z}} \|\Psi_0(k)\|_2^2} \|x(k+1)\|_2^2.$$

Let $\tilde{e} = \frac{\delta_1 - \sup_{k\in\mathbb{Z}} \|\Psi_0(k)\|_2^2 - \delta_2 \sup_{k\in\mathbb{Z}} \sum_{i=1}^{d} \|\Psi_i(k)\|_2^2}{\sup_{k\in\mathbb{Z}} \|\Psi_0(k)\|_2^2}$. Since scalar $\epsilon \in (0, \alpha^2/\beta^2)$ is arbitrary, thus, $\tilde{e} > \frac{\alpha^2}{\beta^2 - \alpha^2} > 0$, and

$$\begin{aligned} \Delta V(k, x(k)) &= x^T(k+1)Px(k+1) - x^T(k)Px(k) \\ &\le -\lambda_{\max}(P) \frac{\alpha^2}{\beta^2 - \alpha^2} \|x(k+1)\|_2^2. \end{aligned}$$

Then, conditions (i), (ii) and $(\text{iii})_b$ of Theorem 3.2 are all satisfied, and hence, the conclusion follows. □

Remark 5.1 In Theorem 3.1, Corollary 3.1, Theorem 3.2 and Theorem 4.1, there is a relation between $V(k+s, x(k+s))(s \in I_d^1)$ and $V(k+1, x(k+1))$, namely, "provided $\mathcal{R}(V(k+s, x(k+s)), V(k+1, x(k+1)))$", where $\mathcal{R}(\cdot, \cdot)$ defines a relation. We call this relation as the $\mathcal{R}$-relation. The conditions (ii), $(\text{ii})'$, $(\text{iii})_a$ and $(\text{iii})_b$ describe the constraint on $\Delta V(k, x(k))$ under the $\mathcal{R}$-relation, but no constrain on $\Delta V(k, x(k))$ without

$\mathcal{R}$-relation. Thus, the condition that the constraint on $\Delta V(k, x(k))$ holds not only with but also without the $\mathcal{R}$-relation, is more restrictive than the condition that the constraint on $\Delta V(k, x(k))$ holds only with the $\mathcal{R}$-relation. Therefore, we can obtain a class of particular cases of Theorem 3.2 with conditions (i), (ii), either $(\text{iii})_a$ or $(\text{iii})_b$, which in fact are corresponding to the well-known Lyapunov-like theorems.

6 Illustrative Numerical Example

To illustrate the effectiveness of the results obtained in previous sections, we consider the following nonlinear discrete system with input time delay:

$$x(k+1) \quad = \quad 1.44x(k) - x^3(k) + 0.069u(k) + 0.031u(k-1), \quad x(k) \in [-1.2, 1.2]. \tag{13}$$

Assume that $\alpha = 0.45$ and $\beta = 0.60$. To obtain the zero solution $x(k) = 0$ in U.P.A.S with (α, β), adopt the following fuzzy control law:

$$\begin{aligned} R_1: \quad & \text{IF } x \text{ is about } \pm 1.2, \text{ THEN} \\ & u = F_1 x(k), \\ R_2: \quad & \text{IF } x \text{ is about } 0, \text{ THEN} \\ & u = F_2 x(k). \end{aligned}$$

The references on fuzzy control can be found in [18, 19]. Then, the overall control law is

$$u(k) = \sum_{i=1}^{2} \mu_i F_i x(k), \tag{14}$$

where $\mu_1 = \frac{x^2}{1.44}$ and $\mu_2 = 1 - \mu_1$ are both membership functions, as shown in Figure 2. The control gain matrices are designed to be $F_1 = -0.0694$ and $F_2 = -18.9114$. Then, the closed-loop system can be expressed as follows:

$$x(k+1) \quad = \quad \big(1.44 - x^2(k) + 0.069 \sum_{i=1}^{2} \mu_i F_i\big) x(k) + 0.031 \sum_{i=1}^{2} \mu_i F_i x(k-1).$$

Denote discriminant function by

$$g(x(k)) = \big(1.44 - x^2(k) + 0.069 \sum_{i=1}^{2} \mu_i F_i\big)^2 + (0.031 \sum_{i=1}^{2} \mu_i F_i)^2.$$

The profile of $g(x)$ is illustrated in Figure 3. We can calculate that $g(x) \leq 0.3619 < 1 - \alpha^2/\beta^2 = 0.4375$ for $x \in [-1.2, 1.2]$. By (12) and Corollary 5.1, system (13) is U.P.A.C. with respect to (α, β). The state curve with initial values $x(-1) = 0.3$, $x(0) = 0.4$ of system (13) with and without fuzzy controller (14) are shown in Figure 4. Without fuzzy controller, i.e., $u(k) = 0$, the zero solution is unstable, and the nonlinear discrete system converges to $x(k) \approx 0.6633 > \beta$; whereas, with fuzzy controller (14), the closed-loop system is U.P.A.S. with (α, β).

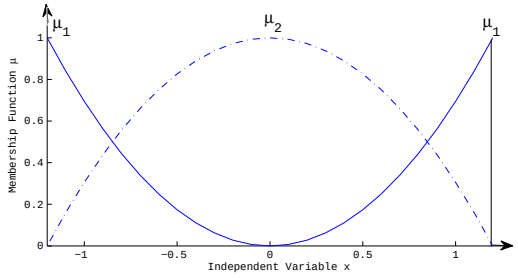

Figure 2: The membership functions of μ_1 and μ_2

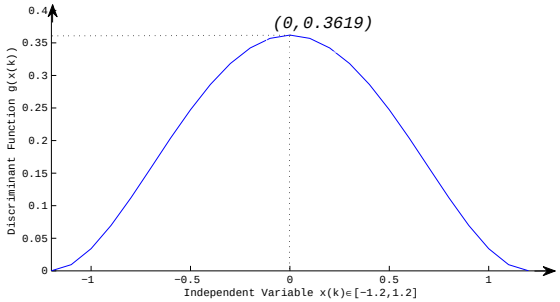

Figure 3: The profile of $g(x(k))$

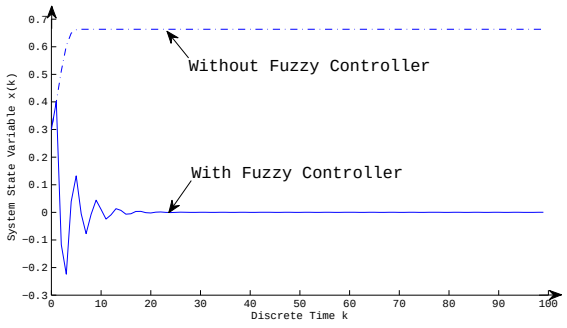

Figure 4: The state curve of system (13) with and without fuzzy controller (14)

7 Conclusions

Motivated by the idea in [9], practical asymptotic stability and controllability are studied for a class of nonlinear discrete systems with time delay. Some explicit criteria for the uniform practical asymptotic stability are established by means of Lyapunov function and Razumikhin technique. Estimations of the solution boundary and arrival time of the solution are also investigated. In addition, the proposed theorems are used to study the practical controllability for a general class of nonlinear discrete systems with input time delay. Finally, a numerical example is presented to illustrate the effectiveness of the proposed results. We believe the results in this paper are useful for the study networked control systems.

References

[1] Lasalle, J. and Lefshetz, S. *Stability by Lyapunov direct method and application*. Academic Press, New York, 1961.

[2] Lakshmikantham, V., Leela, S. and Martynyuk, A. A. *Practical Stability of Nonlinear Systems*. World Scientific, Singapore, 1990.

[3] Delchamps, D. F. Stabilizing a linear system with quantized state feedback. *IEEE Transactions on Automatic Control* **35** (8) (1990) 916–924.

[4] Chou, J. H., Chen, S. H. and Horng, I. R. Robust stability bound on linear time-varying uncertainties for linear digital control systems under finite wordlength effects. *JSMS International Journal: Series C* **39** (4) (1996) 767–771.

[5] Elia, N. and Mitter, S. K. Stabilization of linear systsems with limited information. *IEEE Transactions on Automatic Control* **46** (9) (2001) 1384–1400.

[6] Fagnani, F. and Zampieri, S. Stability analysis and synthesis for scalar linear systems with a quantized feedback. *IEEE Transactions on Automatic Control* **48** (9) (2003) 1569–1584.

[7] Hou, C. and Qian, J. Decay estimates for applictions of Razumikhin-type theorems. *Automatica* **34** (7) (1998) 921–924.

[8] Blanchini, F. and Ryan, E. P. A Razumikhin-type lemma for functional differential equations with application to adaptive control. *Automatica* **35** (1999) 809–818.

[9] Zhang, S. and Chen, M. P. A new Razumikhin theorem for delay difference equations. *Computers and Mathematics with Applications* **36** (1998) 405–412.

[10] Zhang, S. A new technique in stability of infinite delay differential equations. *Computers and Mathematics with Applications* **44** (2002) 1275–1287.

[11] Zhai, G. and Michel, A.N. On practical stability of switched systems. In: *Proceeding of the 41st IEEE Conference on Decision and Control*. Las Vegas, Nevada USA, December 2002, 3488–3493.

[12] Xu, X. and Zhai, G. Practical stability and stabilization of hybrid and switched systems. *IEEE Transactions on automatic control* **50** (11) (2005) 1897–1903.

[13] Yang, C., Zhang, Q. and Zhou, L. Practical stabilization and controllability of descriptor systems. *International Journal of Information and Systems Sciences* (1) (3–4) (2005) 455–465.

[14] Yang, C., Zhang, Q., Lin, Y. and Zhou, L. Practical stability of closed-loop descriptor systems. *International Journal of Systems Science* **37** (14) (2006) 1059–1067.

[15] Peng, S. and Chen, C. Estimation of the practical stability region of a class of robust controllers with input constraint. *Journal of Franklin Institute* **335B** (7) (1998) 1271–1281.

[16] Kapitaniak, T. and Brindley, J. Practical stability of chaotic attractors. *Chaos, Solitons & Fractals* **9** (7) (1998) 43–50.

[17] Anabtawi, M.J. and Sathananthan, S. Quantitative analysis of hybrid parabolic systems with Markovian regime switching via practical stability. *Nonlinear Analysis: Hybrid Systems* **2** (2008) 980–992.

[18] Takagi, T. and Sugeno, M. Fuzzy identification of system and its applictions to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics* **15** (1) (1985) 116–132.

[19] Babuska, R. *Fuzzy Modeling For Control*. Kluwer Academic Publishers, 1998.

Nonlinear Dynamics and Systems Theory, **10** (2) (2010) 175–188

Global Optimization Method for Continuous-Time Sensor Scheduling

S.F. Woon [1,2*], V. Rehbock [1] and R.C. Loxton [1]

[1] *Department of Mathematics and Statistics, Curtin University of Technology, Bentley, Western Australia 6102, Australia;*
[2] *Physical Science, College of Arts and Sciences, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia.*

Received: June 24, 2009; Revised: March 26, 2010

Abstract: We consider a situation in which several sensors are used to collect data for signal processing. Since operating multiple sensors simultaneously causes system interference, only one sensor can be active at any one time. The problem of scheduling the operation of the sensors to minimize signal estimation error is formulated as a discrete-valued optimal control problem. This problem cannot be solved using conventional optimization techniques. We instead transform it into an equivalent mixed discrete optimization problem. The transformed problem is then decomposed into a bi-level optimization problem, which is solved using a discrete filled function method in conjunction with a conventional optimal control algorithm. Numerical results show that our algorithm is robust, efficient, and reliable in attaining a near globally optimal solution.

Keywords: *sensor scheduling; time scaling transformation; discrete filled function; optimal control; mixed discrete optimization.*

Mathematics Subject Classification (2000): 37N35, 37N40, 90-04, 90-08, 68W01, 90C06, 90C27, 90C30.

* Corresponding author: mailto:woonsiewfang@yahoo.com

1 Introduction

Sensors are used in various applications, including military surveillance, ground mapping, tracking and recognition of targets, instrumentation, air traffic control, imaging, and robotics [1]. Information collected by the sensors is used to design activities that evolve over time in the underlying system [2]. For example, in a defense system, surveillance sensors are used to detect, identify, and localize targets, assess levels of threat, and deduce enemy intent [3]. In some applications, such as robotics, operating several sensors simultaneously causes interference in the system and thus affects the measurement accuracy [4]. Consequently, it is impossible to operate all of the sensors at once. Instead, we need to schedule the operation of sensors over a given time frame so that the signal estimation error is minimized. We assume in this paper that only one sensor may be active at any one time. The accuracy of the estimation obtained by the sensors increases with a decrease in measurement of noise in a stochastic environment. The work presented here was motivated by [5] and [6]. In [6], the optimal scheduling policy is obtained by solving a quasi-variational inequality. However, the complexity of the model in [6] makes it difficult to compute an optimal solution. On the other hand, [5] considers open-loop policies with switches from one sensor to another. This reference proposes a time scaling transformation, which aims to capture a large variety of possible switching sequences. The sensor scheduling problem, which is formulated as a discrete-valued optimal control problem, is first transformed into an optimal parameter selection problem, and then solved using existing optimal control software. The optimal control for the original problem is determined through a reverse transformation. However, this approach introduces a large number of artificial switches, many of which are not utilized in the optimal solution. As a consequence, the resulting optimization problem has many local minima. A study similar to that considered in [5] is performed in [7], where a combination of a branch and cut technique and a gradient-based method is applied to solve the continuous-time sensor scheduling problem.

We consider a general optimal sensor scheduling problem, which is similar to the one discussed in [5] and [7], and propose a transformation to convert it into an equivalent mixed discrete optimization problem, as discussed in Section 3. Then, we propose a novel global optimization algorithm in Section 4, which incorporates a discrete filled function method and a gradient-based method, to avoid local solutions and speed up the computation. To evaluate the effectiveness of our algorithm, we solve a numerical example from the literature and compare the results with those obtained from the methods in [5] and [7] in Section 5.

2 Problem Formulation

Consider the following system of linear stochastic differential equations on a given probability space $(\Omega, \mathcal{F}, \mathcal{P})$:

$$d\mathbf{x}(t) = A(t)\mathbf{x}(t)dt + B(t)dK(t), \qquad t \in [0, T],$$

with initial condition

$$\mathbf{x}(0) = \mathbf{x}_0.$$

Here, $\{\mathbf{x}(t),\ t \in [0, T]\}$ is a $\mathbb{R}^n$-valued state process representing a signal of interest. It is assumed to be square integrable. The initial state, $\mathbf{x}_0$, is a $\mathbb{R}^n$-valued Gaussian random vector on $(\Omega, \mathcal{F}, \mathcal{P})$ with mean $\bar{\mathbf{x}}_0$ and covariance matrix P_0. Furthermore, $A : [0, T] \to$

$\mathbb{R}^{n\times n}$ and $B : [0,T] \to \mathbb{R}^{n\times p}$ are continuous functions. The process $\{K(t),\ t \in [0,T]\}$ is a standard $\mathbb{R}^p$-valued Brownian motion on $(\Omega, \mathcal{F}, \mathcal{P})$ with mean zero and given covariance matrix $Q \in \mathbb{R}^{p\times p}$, where Q is symmetric and positive semi-definite.

Suppose that there are M sensors for detecting the state process. Only one of these sensors may be operated at any one time. A *sensor schedule* is a function $\phi : [0,T] \to \{1,\dots,M\}$ that returns the active sensor at time t. In other words, $\phi(t) = i$ means sensor i is active at time t. Let Φ be the set of all measurable sensor schedules and let $\mathbf{y}$ be the observation process associated with the scheduling policy ϕ. For any $\phi \in \Phi$, we have the following output equation:

$$d\mathbf{y}(t) = \sum_{i=1}^{M} \chi_{\{t:\phi(t)=i\}}(t)\big\{C_i(t)x(t)dt + D_i(t)dW_i(t)\big\}, \qquad t \in [0,T],$$

and

$$\mathbf{y}(0) = \mathbf{0},$$

where, for each $\mathcal{I} \subset [0,T]$,

$$\chi_{\mathcal{I}}(t) = \begin{cases} 1, & t \in \mathcal{I}, \\ 0, & \text{otherwise}, \end{cases}$$

and $\{W_i(t),\ t \in [0,T]\}$ is a standard $\mathbb{R}^m$-valued Brownian motion with mean zero and covariance matrix $R \in \mathbb{R}^{m\times m}$, where R is symmetric and positive definite, $C_i : [0,T] \to \mathbb{R}^{m\times n}$ and $D_i : [0,T] \to \mathbb{R}^{m\times m}$ are continuous functions.

Each sensor makes an observation of the state process that is contaminated by noise. The history of such observation processes is denoted by $\{\mathbf{y}(s), 0 \le s \le t\}$. The data collected from the M sensors are used to estimate the state $\mathbf{x}$ at time t. The best estimate of $\mathbf{x}(t)$ is known as $\hat{\mathbf{x}}(t)$. Since $\mathbf{y}$ is corrupted by noise, the history observed is uncertain. Let the history of such a process be denoted by the smallest σ-algebra, $\mathcal{F}_t^{\mathbf{y}} = \sigma\{\mathbf{y}(s), 0 \le s \le t\}$. Hence, the optimal mean-square estimate of the state given $\mathcal{F}_t^{\mathbf{y}}$ is $\hat{\mathbf{x}}(t)$, and the associated error covariance is $P(t)$. Then, for a given $\phi \in \Phi$, the optimal $\hat{\mathbf{x}}(t)$ is given by the following theorem. The proof of this theorem may be found in [8].

Theorem 2.1 *For each sensor schedule $\phi \in \Phi$, the optimal mean-square estimate of the state $\hat{\mathbf{x}}(t)$ is the unique solution of the following stochastic differential equation:*

$$\begin{aligned} d\hat{\mathbf{x}}(t) &= \left[A(t) - P(t)\sum_{i=1}^{M} \chi_{\{t:\phi(t)=i\}}(t)C_i^{\top}(t)\bar{R}_i^{-1}(t)C_i(t)\right]\hat{\mathbf{x}}(t)dt \\ &\quad + \left[P(t)\sum_{i=1}^{M} \chi_{\{t:\phi(t)=i\}}(t)C_i^{\top}(t)\bar{R}_i^{-1}(t)\right]d\mathbf{y}(t), \quad t \in [0,T], \end{aligned} \tag{1}$$

and

$$\hat{\mathbf{x}}(0) = \bar{\mathbf{x}}_0, \tag{2}$$

where

$$\bar{R}_i^{-1}(t) = \big[D_i(t)R_i(t)D_i^{\top}(t)\big]^{-1}, \tag{3}$$

and the error covariance matrix $P : [0, T] \to \mathbb{R}^{n\times n}$ *is the unique solution of the matrix Riccati differential equation*

$$\dot{P}(t) = A(t)P(t)+P(t)A^{\top}(t)+B(t)QB^{\top}(t)-P(t)\sum_{i=1}^{M}\chi_{\{t:\phi(t)=i\}}(t)C_i^{\top}(t)\bar{R}_i^{-1}(t)C_i(t)P(t) \tag{4}$$

with initial condition

$$P(0) = P_0. \tag{5}$$

Clearly, the solution of (4)-(5) depends on the sensor schedule that is chosen. Let $P(\cdot|\phi)$ be the solution corresponding to $\phi \in \Phi$. We formulate the following sensor scheduling problem.

Problem (P). Choose $\phi \in \Phi$ to minimize

$$g_0(\phi) = \alpha\text{trace}\{P(T|\phi)\} + \int_0^T \text{trace}\{P(t|\phi)\}dt, \tag{6}$$

subject to (4) and (5), where α is a non-negative constant.

The objective function (6) is designed to minimize the estimation error during the operation of the system. Note that Problem (P) is a discrete-valued optimal control problem. The main challenge in solving Problem (P) is that the control ϕ is constrained to take values in the discrete set $\{1, \ldots, M\}$. Each sensor schedule is completely determined by specifying the values in $\{1, \ldots, M\}$ that it assumes and the times when it switches from one value in $\{1, \ldots, M\}$ to another. Clearly, only a finite number of switches are able to be implemented in practice, and hence ϕ is a piecewise constant function with a finite number of switches. In other words, to solve Problem (P), we need to determine both the optimal switching sequence and the optimal switching times. Thus, we transform Problem (P) into an equivalent and solvable form in the next section.

3 Problem Transformation

Recall that only one sensor is active at each time and that only a finite number of switches are allowed. Suppose that we allow a sensor schedule ϕ to switch N times during the time horizon. Let $V = \{\mathbf{v} = [v_1, \ldots, v_{N+1}]^{\top} : v_i \in \{1, \ldots, M\}\}$ be the set of all possible switching sequence vectors. Let $\boldsymbol{\sigma} = [\sigma_1, \ldots, \sigma_{N+1}]^{\top}$, where $\sigma_i \geq 0,\ i = 1, \ldots, N + 1$, denote the duration for which the corresponding sensor v_i in the sequence is active. Clearly,

$$\sum_{i=1}^{N+1}\sigma_i = T.$$

Let Σ denote the set of all such $\boldsymbol{\sigma}$. Note that under the assumption of a finite number of switches, N, any $\phi \in \Phi$ is completely determined by an element $(\mathbf{v}, \boldsymbol{\sigma}) \in V \times \Sigma$, where

$$\phi(t) = v_i, \qquad t \in \left[\sum_{j=1}^{i-1}\sigma_j,\ \sum_{j=1}^{i}\sigma_j\right], \qquad i = 1, \ldots, N + 1.$$

We introduce a new time variable $\tau \in [0, N+1]$ and consider the fixed partition $\{0, 1, \ldots, N+1\}$. The original time horizon $[0, T]$ is transformed into the new time horizon $[0, N+1]$ as follows:

$$\dot{t}(\tau) = \sigma_i, \qquad \tau \in [i-1, i), \qquad i = 1, \ldots, N+1, \tag{7}$$

with the boundary conditions

$$t(0) = 0 \tag{8}$$

and

$$t(N+1) = T. \tag{9}$$

The original dynamics (4)-(5) are transformed into

$$\begin{aligned} \dot{\tilde{P}}(\tau) &= \sigma_i \Big[A(\tau)\tilde{P}(\tau) + \tilde{P}(\tau)A^\top(\tau) + B(\tau)QB^\top(\tau) \\ &\quad - \tilde{P}(\tau)C_{v_i}^\top(\tau)\bar{R}_{v_i}^{-1}(\tau)C_{v_i}(\tau)\tilde{P}(\tau) \Big], \ \tau \in [i-1, i), \ i = 1, \ldots, N+1, \end{aligned} \tag{10}$$

and

$$\tilde{P}(0) = P_0. \tag{11}$$

Hence, the transformed problem is stated formally below. Let $\tilde{P}(\cdot|\mathbf{v}, \boldsymbol{\sigma})$ be the solution of (10)-(11) corresponding to $(\mathbf{v}, \boldsymbol{\sigma}) \in V \times \Sigma$.

Problem (R). Choose $\mathbf{v} \in V$ and $\boldsymbol{\sigma} \in \Sigma$ to minimize

$$g_0(\mathbf{v}, \boldsymbol{\sigma}) = \alpha \text{trace}\{\tilde{P}(N+1|\mathbf{v}, \boldsymbol{\sigma})\} + \sum_{i=1}^{N+1} \int_{i-1}^{i} \text{trace}\{\tilde{P}(\tau|\mathbf{v}, \boldsymbol{\sigma})\}\sigma_i \, d\tau, \tag{12}$$

subject to (7)-(9) and the dynamics (10)-(11), where α is a non-negative constant.

Problem (R), an equivalent problem to Problem (P), is a mixed discrete optimization problem with the discrete variable $\mathbf{v}$ representing the switching sequence and the continuous variable $\boldsymbol{\sigma}$ representing the time length of each mode. We propose to solve Problem (R) by first decomposing it into two levels. Note that for a fixed $\mathbf{v} \in V$, Problem (R) reduces to the following problem.

Problem ($\mathbf{R_1}$). Given $\mathbf{v} \in V$, find a $\boldsymbol{\sigma} \in \Sigma$ to minimize

$$g_0(\boldsymbol{\sigma}|\mathbf{v}) = \alpha \text{trace}\{\tilde{P}(N+1|\boldsymbol{\sigma}, \mathbf{v})\} + \sum_{i=1}^{N+1} \int_{i-1}^{i} \text{trace}\{\tilde{P}(\tau|\boldsymbol{\sigma}, \mathbf{v})\}\sigma_i \, d\tau, \tag{13}$$

subject to (7)-(9) and dynamics (10)-(11), where α is a non-negative constant.
Problem (R_1) is a standard optimal parameter selection problem in a canonical form suitable for the application of a standard algorithm based on the control parameterization concept. For each given $\mathbf{v}$, the optimal value of g_0 in (13) can be determined using an optimal control software, such as MISER3.3 [9], since the switching sequence is fixed. Note that in MISER3.3, the optimal parameter selection problem is solved using a sequential quadratic programming algorithm. The second problem in the proposed

decomposition is defined as follows.

Problem ($\mathbf{R_2}$). Choose $\mathbf{v} \in V$ to minimize the objective function

$$J(\mathbf{v}), \tag{14}$$

where

$$J(\mathbf{v}) = \min_{\boldsymbol{\sigma} \in \Sigma} \quad g_0(\boldsymbol{\sigma}|\mathbf{v}).$$

Note that Problem (R_2) is a purely discrete optimization problem, but computing the value of $J(\mathbf{v})$ requires solving the corresponding Problem (R_1). Hence, Problem (R_1) is a subproblem of Problem (R_2). To obtain a near globally optimal solution for Problem (R), we propose a combined algorithm where Problem (R_2) will be solved using a discrete filled function method and, at each iteration, Problem (R_1) is solved using MISER3.3. For our numerical computations, we have been able to incorporate the discrete filled function method within the MISER3.3 software. The details of the discrete filled function approach are discussed in the next section.

Remark 3.1 Note that the early time scale transformation proposed in [5] introduces a large number of artificial switching instants, typically $N \times M$, most of which are not used in the final optimal solution. As a result, the transformed problem yields many local minima, many of which have high objective values. Our method avoids this difficulty because only N switches are needed.

4 Discrete Filled Function Method

The filled function approach is a global optimization method which was initiated by Ge in the late 1980s [10, 11] to solve continuous global optimization problems. Zhu [12] appears to be the first researcher to adapt the continuous filled function approach in solving discrete optimization problems. However, the filled function proposed by Zhu contains an exponential term, making it difficult to determine an improved point [13] in practice. Since then, various discrete filled functions with improved theoretical properties have been proposed in [13–17] to enhance computational efficiency.

In this paper, we employ a discrete filled function method, which was recently developed in [13], as a part of our proposed algorithm. The basic idea of this method is as follows. We choose an initial sequence and then perform a local search (see Algorithm 4.1 below) to find an initial local minimizer. Then, we construct an auxiliary function, called a filled function, at this local minimizer. By minimizing the filled function, either an improved local minimizer is found or one of the vertices is reached. This process is repeated until no improved local minimizer of the corresponding filled function can be found. The final local minimizer is then taken as an approximation of the global minimizer.

Definition 4.1 For any $\mathbf{v} \in V$, the *neighbourhood* of $\mathbf{v}$ is defined by $N(\mathbf{v}) = \{\mathbf{w} \in V |\ \mathbf{w} = \mathbf{v} \pm \mathbf{e}_i : i = 1, 2, \ldots, N+1\}$. Here, $\mathbf{e}_i$ denotes the i-th standard unit basis vector of $\mathbb{R}^{N+1}$: its i-th component is equal to one and its other components are equal to zero. The set of all feasible directions at $\mathbf{v} \in V$ is defined by $\mathcal{D}(\mathbf{v}) = \{\mathbf{d} \in \mathbb{R}^{N+1} : \mathbf{v} + \mathbf{d} \in N(\mathbf{v})\} \subset \{\pm\mathbf{e}_i,\ i = 1, \ldots, N+1\}$.

Definition 4.2 The sequence $\mathbf{v}^* \in V$ is a *local minimizer* of J if $J(\mathbf{v}^*) \leq J(\mathbf{v})$ for all $\mathbf{v} \in N(\mathbf{v}^*)$. If $J(\mathbf{v}^*) < J(\mathbf{v})$ for all $\mathbf{v} \in N(\mathbf{v}^*) \setminus \{\mathbf{v}^*\}$, then $\mathbf{v}^*$ is a *strict local minimizer*

of J. The sequence $\mathbf{v}^*$ is a *global minimizer* of J if $J(\mathbf{v}^*) \leq J(\mathbf{v})$ for all $\mathbf{v} \in V$. If $J(\mathbf{v}^*) < J(\mathbf{v})$ for all $\mathbf{v} \in V \setminus \{\mathbf{v}^*\}$, then $\mathbf{v}^*$ is a *strict global minimizer* of J.

Definition 4.3 $\mathbf{v}$ is a vertex of V if for each $\mathbf{d} \in \mathcal{D}(\mathbf{v})$, $\mathbf{v} + \mathbf{d} \in V$ and $\mathbf{v} - \mathbf{d} \notin V$. Let $\tilde{V}$ denote the set of vertices of V.

Algorithm 4.1 Discrete Steepest Descent Method

1. Choose an initial switching sequence $\mathbf{v} \in V$.
2. If $\mathbf{v}$ is a local minimizer of J, then stop. Otherwise, find a discrete steepest descent direction $\mathbf{d}^* \in \mathcal{D}(\mathbf{v})$ of J.
3. Let $\mathbf{v} = \mathbf{v} + \mathbf{d}^*$. Go to Step 2.

Based on Definitions 4.1-4.3, we call a function $G_{\mathbf{v}^*} : V \mapsto \mathbb{R}$ a *discrete filled function* of J at $\mathbf{v}^*$ if it satisfies the following conditions:
(a) $\mathbf{v}^*$ is a strict local maximizer of $G_{\mathbf{v}^*}$;
(b) Let $\hat{V}(\mathbf{v}^*) = \{\mathbf{v} \in V : \mathbf{v} \neq \mathbf{v}^*, J(\mathbf{v}) \geq J(\mathbf{v}^*)\}$. $G_{\mathbf{v}^*}$ has no local minimizer in the set $\hat{V}(\mathbf{v}^*) \setminus \tilde{V}$;
(c) $\mathbf{v}^{**} \in V \setminus \tilde{V}$ is a local minimizer of J if and only if $\mathbf{v}^{**}$ is a local minimizer of $G_{\mathbf{v}^*}$.

Define

$$G_{\mu,\rho,\mathbf{v}^*}(\mathbf{v}) = A_\mu(J(\mathbf{v}) - J(\mathbf{v}^*)) - \rho \parallel \mathbf{v} - \mathbf{v}^* \parallel, \tag{15}$$

where

$$A_\mu(y) = y \cdot \mu \left[(1-c) \left(\frac{1-c\mu}{\mu - c\mu} \right)^{-y/\omega} + c \right],$$

$\omega > 0$ is a sufficiently small number, and $0 < c \leq 1$ is a constant. The function $G_{\mu,\rho,\mathbf{v}^*}(\mathbf{v})$ is a discrete filled function when certain conditions on the parameters μ and ρ are satisfied. Hence, it has properties (a)-(c) when those conditions on μ and ρ are met. Note that the discrete filled function is constructed based on the following theorems found in [13]. A detailed convergence analysis for this method has also been given in [13].

Definition 4.4 Let $\mathcal{K}$ be a constant satisfying

$$1 \leq \max_{\substack{\mathbf{v}_1, \mathbf{v}_2 \in V \\ \mathbf{v}_1 \neq \mathbf{v}_2}} \parallel \mathbf{v}_1 - \mathbf{v}_2 \parallel \leq \mathcal{K} < \infty,$$

where $\parallel \cdot \parallel$ is the Euclidean norm. Let $0 < \mathcal{L} < \infty$ be the Lipschitz constant such that $|J(\mathbf{v}_1) - J(\mathbf{v}_2)| \leq \mathcal{L} \parallel \mathbf{v}_1 - \mathbf{v}_2 \parallel$, for any distinct $\mathbf{v}_1, \mathbf{v}_2 \in V$.

Theorem 4.1 *If $\rho > 0$ and $0 < \mu < min\{1, \frac{\rho}{\mathcal{L}}\}$, then $\mathbf{v}^*$ is a strict local maximizer of $G_{\mu,\rho,\mathbf{v}^*}$. If $\mathbf{v}^*$ is a global minimizer of J, then $G_{\mu,\rho,\mathbf{v}^*}(\mathbf{v}) < 0$ for all $\mathbf{v} \in V \setminus \{\mathbf{v}^*\}$.*

Theorem 4.2 *Let $\mathbf{v}^{**}$ be a strict local minimizer of J with $J(\mathbf{v}^{**}) < J(\mathbf{v}^*)$. If $\rho > 0$ is sufficiently small and $0 < \mu < 1$, then $\mathbf{v}^{**}$ is a strict local minimizer of $G_{\mu,\rho,\mathbf{v}^*}$.*

Theorem 4.3 *Let $\acute{\mathbf{v}}$ be a strict local minimizer of $G_{\mu,\rho,\mathbf{v}^*}$ and let $\bar{\mathbf{d}} \in \mathcal{D}(\acute{\mathbf{v}})$ be a feasible direction at $\acute{\mathbf{v}}$ such that $\parallel \acute{\mathbf{v}} + \bar{\mathbf{d}} - \mathbf{v}^* \parallel > \parallel \acute{\mathbf{v}} - \mathbf{v}^* \parallel$. If $\rho > 0$ is sufficiently small and $0 < \mu < min\{1, \frac{\rho}{2\mathcal{K}^2\mathcal{L}}\}$, then $\acute{\mathbf{v}}$ is a local minimizer of J.*

Corollary 4.1 *Assume that every local minimizer of J is strict. Suppose that $\rho > 0$ is sufficiently small and $0 < \mu < min\{1, \frac{\rho}{2K^2\mathcal{L}}\}$. Then, $\mathbf{v}^{**} \in V \setminus \tilde{V}$ is a local minimizer of J with $J(\mathbf{v}^{**}) < J(\mathbf{v}^*)$ if and only if $\mathbf{v}^{**}$ is a local minimizer of $G_{\mu,\rho,\mathbf{v}^*}$.*

Interested readers are referred to [13] for proofs of Theorem 4.1-4.3. Clearly, from Corollary 4.1, $G_{\mu,\rho,\mathbf{v}^*}$ must satisfy the condition (c) of the discrete filled function definition if every local minimizer of J is strict under certain conditions on μ and ρ. If the local minimizer of the discrete filled function $G_{\mu,\rho,\mathbf{v}^*}$ found is an improved point, it is also a local minimizer of the original function J. Based on the theoretical framework described above, a discrete filled function algorithm for global optimization can be stated as follows.

Algorithm 4.2 Discrete Filled Function Method

1. Initialize $\mathbf{v}_0 \in V$, ρ_0, μ_0, $\rho_L > 0$, $0 < \hat{\rho} < 1$, and $0 < \hat{\mu} < 1$.
   Let $\rho := \rho_0$ and $\mu := \mu_0$.
   Choose an initial sequence $\mathbf{v}_0 \in V$.

2. Starting from $\mathbf{v}_0$, minimize $J(\mathbf{v})$ using Algorithm 4.1 to obtain a local minimizer $\mathbf{v}^*$ of J.

3. (a) List the neighbouring sequences of $\mathbf{v}^*$ as $N(\mathbf{v}^*) = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_q\}$. Set $\ell := 1$.
   (b) Set the current switching sequence, $\mathbf{v}_c := \mathbf{w}_\ell$.

4. (a) If there exists a direction $\mathbf{d} \in \mathcal{D}(\mathbf{v}_c)$ such that $J(\mathbf{v}_c + \mathbf{d}) < J(\mathbf{v}^*)$, then set $\mathbf{v}_0 := \mathbf{v}_c + \mathbf{d}$ and go to Step 2. Otherwise, go to (b) below.
   (b) Let $\mathcal{D}_1 = \{\mathbf{d} \in \mathcal{D}(\mathbf{v}_c) : \ J(\mathbf{v}_c + \mathbf{d}) < J(\mathbf{v}_c)$ and $G_{\mu,\rho,\mathbf{v}}(\mathbf{v}_c + \mathbf{d}) < G_{\mu,\rho,\mathbf{v}}(\mathbf{v}_c)\}$.
   If $\mathcal{D}_1 \neq \emptyset$, set $\mathbf{d}^* := \arg\min_{\mathbf{d}\in\mathcal{D}(\mathbf{v}_c)}\{J(\mathbf{v}_c + \mathbf{d}) + G_{\mu,\rho,\mathbf{v}^*}(\mathbf{v}_c + \mathbf{d})\}$.
   Then, set $\mathbf{v}_c := \mathbf{v}_c + \mathbf{d}^*$ and go to Step 4(a). Otherwise, go to (c) below.
   (c) Let $\mathcal{D}_2 = \{\mathbf{d} \in \mathcal{D}(\mathbf{v}_c) : \ G_{\mu,\rho,\mathbf{v}}(\mathbf{v}_c + \mathbf{d}) < G_{\mu,\rho,\mathbf{v}}(\mathbf{v}_c)\}$.
   If $\mathcal{D}_2 \neq \emptyset$, set $\mathbf{d}^* := \arg\min_{\mathbf{d}\in\mathcal{D}(\mathbf{v}_c)}\{G_{\mu,\rho,\mathbf{v}^*}(\mathbf{v}_c + \mathbf{d})\}$.
   Then, set $\mathbf{v}_c := \mathbf{v}_c + \mathbf{d}^*$ and go to Step 4(a). Otherwise, go to Step 5.

5. Let $\acute{\mathbf{v}}$ be the obtained local minimizer of $G_{\mu,\rho,\mathbf{v}^*}$.
   (a) If $\acute{\mathbf{v}} \in \tilde{V}$, set $\ell := \ell + 1$. If $\ell > q$, go to Step 6. Otherwise, go to Step 3(b).
   (b) If $\acute{\mathbf{v}} \notin \tilde{V}$, reduce μ by setting $\mu := \hat{\mu}\mu$ and go to Step 4(b).

6. Reduce ρ by setting $\rho := \hat{\rho}\rho$. If $\rho < \rho_L$, terminate the algorithm. The current $\mathbf{v}^*$ is taken as a global minimizer of the problem. Otherwise, set $\ell := 1$ and go to Step 3(b).

The mechanism behind this algorithm can be illustrated as follows. Firstly, the parameters of the discrete filled function $G_{\mu,\rho,\mathbf{v}^*}$ in (15) are initialized to suitable values. These parameters will be reduced gradually in Steps 5 and 6, to ensure that $G_{\mu,\rho,\mathbf{v}^*}$ eventually satisfies properties (a)-(c). The reduction factor of each parameter is also specified at Step 1.

Secondly, we choose an initial sequence $\mathbf{v}_0$ in the feasible region and minimize the original function J. Recall that the value of J is computed using MISER3.3 according to the discussion in the previous section. The objective function value at each sequence in the neighbourhood of $\mathbf{v}_0$ is calculated. The search direction leading to the most improved objective function value in this neighbourhood is chosen according to Algorithm 4.1. The

process is repeated until a local minimizer of J, namely $\mathbf{v}^*$, is found. Next, we identify the neighbourhood of $\mathbf{v}^*$ in Step 3. One of the neighbouring points of $\mathbf{v}^*$, denoted by $\mathbf{v}_c$, is set to be an initial point to minimize the discrete filled function $G_{\mu,\rho,\mathbf{v}^*}$ in the following step. Note that $\mathbf{v}^*$ is a local maximizer of $G_{\mu,\rho,\mathbf{v}^*}$ here.

In Step 4, we first check to see if there exists a neighbouring sequence of $\mathbf{v}_c$ that is an improvement over the current minimizer. If such a sequence can be found, then we use it as a starting point to minimize the function J using Algorithm 4.1. Otherwise, if we can find a direction that results in an improvement of both J and G compared with the values at $\mathbf{v}_c$, then we choose the direction which gives the greatest such improvement. If such a direction does not exist, then find a steepest descent direction such that $G_{\mu,\rho,\mathbf{v}^*}(\mathbf{v}_c + \mathbf{d}^*) < G_{\mu,\rho,\mathbf{v}^*}(\mathbf{v}_c)$. If none of these directions exists, then $\mathbf{v}_c$ must be a local minimizer of $G_{\mu,\rho,\mathbf{v}^*}$, so we go to Step 5.

If the local minimizer of $G_{\mu,\rho,\mathbf{v}^*}$ is found to be a vertex of the feasible region, choose the next point in $N(\mathbf{v}^*)$ as a starting point to minimize $G_{\mu,\rho,\mathbf{v}^*}$ in Step 5(a). Note that the minimizer of $G_{\mu,\rho,\mathbf{v}^*}$ should be either an improved point or a vertex. Thus, μ is adjusted suitably to satisfy this criteria in Step 5(b).

If no improved sequence is found with the minimization process starting from all neighbouring sequences ending up at the vertices, we reduce ρ, reset $\ell = 1$, and minimize $G_{\mu,\rho,\mathbf{v}^*}$ again with the new value of ρ. The algorithm is repeated until the termination criteria is reached, where ρ reaches its lower bound, ρ_L. In other words, we have minimized the discrete filled function from every search direction from $\mathbf{v}^*$ and failed to find an improved point, even the parameters are small. We repeat Algorithm 4.2 twice, reducing the value of ρ each time to confirm that no better solution can be found. Thus, the final local minimizer $\mathbf{v}^*$ found is taken to be the global solution of J.

To increase the efficiency, we construct a look-up table to store each value of the objective function J computed so far. Thus, we avoid repeated application of the subproblems solution algorithm at the same point. This is essential because computing $J(\mathbf{v})$ involves solving a complex optimal control problem, which takes considerable computational time.

Remark 4.1 Note that a sequential quadratic programming method is employed within MISER3.3 to solve the subproblem (R_1). This is a local search method and thus cannot guarantee the global optimality for the solution of the subproblem. In other words, although we aim to solve the upper level problem globally, the lower level problem may only yield a locally optimal solution. Therefore, we consider our approach to be a heuristic global optimization method with no implied guarantee of finding the overall global optimum. Nevertheless, numerical results demonstrate that good quality solutions can be determined effectively compared with other methods in the literature, such as [5] and [7].

5 Illustrative Example

Consider a sensor scheduling problem with six sensors and seven switches as discussed in [7]. Let $N = 7,\ M = 6,\ n = 2,\ m = 1,\ p = 2,\ T = 8,\ \alpha = 0,\ c = 0.5,\ \mu_0 = 0.1,\ \rho_0 = 0.1,\ \omega = 1,\ \rho_L = 0.001,\ \hat{\rho} = 0.1,\ \hat{\mu} = 0.1$ and consider the following dynamics:

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_1(t) \end{bmatrix} = \begin{bmatrix} 0.5 & 1.0 \\ 1.0 & 0.5 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 2.0 \\ 2.0 \end{bmatrix} K(t), \quad \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

where

$$P_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$C_1(t) = \begin{bmatrix} 1 + 1.2\sin(2t) & 0 \\ 1 + 1.2\sin(2t) & 0 \end{bmatrix}, \qquad D_1(t) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad R_1(t) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$C_2(t) = \begin{bmatrix} 1 + 0.5\cos(2t) & 1 + 0.5\cos(2t) \\ 0 & 0 \end{bmatrix}, \qquad D_2(t) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad R_2(t) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$C_3(t) = \begin{bmatrix} 1 + 0.5\sin(2t) & 0 \\ 0 & 1 + 0.5\cos(2t) \end{bmatrix}, \qquad D_3(t) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad R_3(t) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$C_4(t) = \begin{bmatrix} 0 & 1 + 0.5\cos(2t) \\ 1 + 0.5\sin(2t) & 0 \end{bmatrix}, \qquad D_4(t) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad R_4(t) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$C_5(t) = \begin{bmatrix} 0 & 0 \\ 1 + 0.5\cos(2t) & 1 + 0.5\sin(2t) \end{bmatrix}, \qquad D_5(t) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad R_5(t) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$C_6(t) = \begin{bmatrix} 0 & 1 + 1.8\sin(2t) \\ 0 & 1 + 1.8\cos(2t) \end{bmatrix}, \qquad D_6(t) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad R_6(t) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

For the ease of computation, we are able to embed the filled function algorithm into the MISER3.3 program. The algorithm is terminated when $\mu = 1 \times 10^{-41}$ and $\rho = 1 \times 10^{-3}$, at which stage the best local minimizer found cannot be improved. The computation is performed using the modified version of MISER3.3 on a Windows-based PC, with a CPU speed of 2.4GHz and 2GB RAM. We solve Problem (R), which has a total number of 1,679,616 potential switching sequences, using $\mathbf{v_0} = [6, 5, 2, 6, 5, 2, 6, 1]^\top$ as the initial sequence and $\boldsymbol{\sigma}_0 = [1, 1, 1, 1, 1, 1, 1, 1]^\top$ as the initial guess for $\boldsymbol{\sigma}$. Note that P_0 is initialized as a 2×2 identity matrix. Relevant results obtained are summarized in Table 1. The entries in the $\mathbf{v}^*$ column indicate the optimal solutions for the local searches. From Table 1, $\boldsymbol{\sigma}^* = [0.23501973, 0, 0, 7.7649803, 0, 0, 0, 0]^\top$ for the assumed global minimum indicates that sensors 2, 3, 4, and 5 are not used in the final optimal solution during the tenth iteration. Hence, only two out of six sensors are turned on. The assumed global optimal switching sequence is to turn on sensor 1, followed by sensor 6, with the objective function 14.33176. The number of original function evaluations and filled function evaluations are 5293 and 8517, respectively. This represents 0.32% of the total number of potential sequences. Note that the objective function evaluations do not include those that were obtained from the look-up table.

We tested the problem with five different initial sequences. These are $[1, 2, 3, 4, 5, 6, 1, 2]^\top$, $[6, 5, 4, 3, 2, 1, 6, 5]^\top$, $[1, 6, 3, 2, 4, 5, 3, 1]^\top$, $[1, 6, 1, 6, 1, 6, 1, 6]^\top$, and $[6, 6, 1, 2, 5, 4, 2, 1]^\top$ using the same $P_\mathbf{0}$ and $\boldsymbol{\sigma}_0$ as in the first computation. As many as fifty local minima are found during the searches from the various initial sequences. Starting at each initial sequence, the algorithm successfully identified the same assumed discrete global minimum sequence of Problem (R) observed in the first experiment, that is, sensor 1 is followed by sensor 6, with the cost function value $J = 14.33176$. Again, computational results show that only up to 0.32% of the total number of potential sequences are evaluated. The optimal operating schedule for the control and states are depicted in Figure 1. In addition, several different choices of P_0 are tested in our experimentation with various initial switching sequences. The optimal operating schemes for $P_0 = \mathbf{0}$, $P_0 = 6I$, $P_0 = 10I$ are illustrated by Figures 2, 3, and 4, respectively. From

Table 1: Numerical Results for $P_0 = I$

$\mathbf{v}^*$	$\boldsymbol{\sigma}^*$	J
$[1,6,1,1,6,1,6,6]^\top$	$[0.24035917,0,0,0,0,0,7.7525870,0.0070538593]^\top$	14.649680367412879
$[6,1,6,6,1,6,1,1]^\top$	$[0.17566501,0.18470974,0,7.6396253,0,0,0,0]^\top$	14.504334985710470
$[1,6,1,6,6,1,1,6]^\top$	$[0.23511799,0,0,7.7648820,0,0,0,0]^\top$	14.331763146735220
$[1,6,2,6,6,2,2,6]^\top$	$[0.23501894,0,0,7.7649811,0,0,0,0]^\top$	14.331763102479558
$[1,6,6,6,6,3,3,5]^\top$	$[0.23502083,0,0,7.7649792,0,0,0,0]^\top$	14.331763102474610
$[1,6,6,6,6,6,5,5]^\top$	$[0.23502039,0,0,7.7649796,0,0,0,0]^\top$	14.331763102473506
$[1,6,6,6,1,5,6,2]^\top$	$[0.23501994,0,0,7.7649801,0,0,0,0]^\top$	14.331763102471598
$[1,1,6,6,6,6,5,1]^\top$	$[0.23501894,0,0,7.7649811,0,0,0,0]^\top$	14.331763102445281
$[1,1,6,6,5,6,6,2]^\top$	$[0.23501979,0,0,7.7649802,0,0,0,0]^\top$	14.331763102440952
$[1,1,6,6,6,5,2,1]^\top$	$[0.23501973,0,0,7.7649803,0,0,0,0]^\top$	14.331763102437696

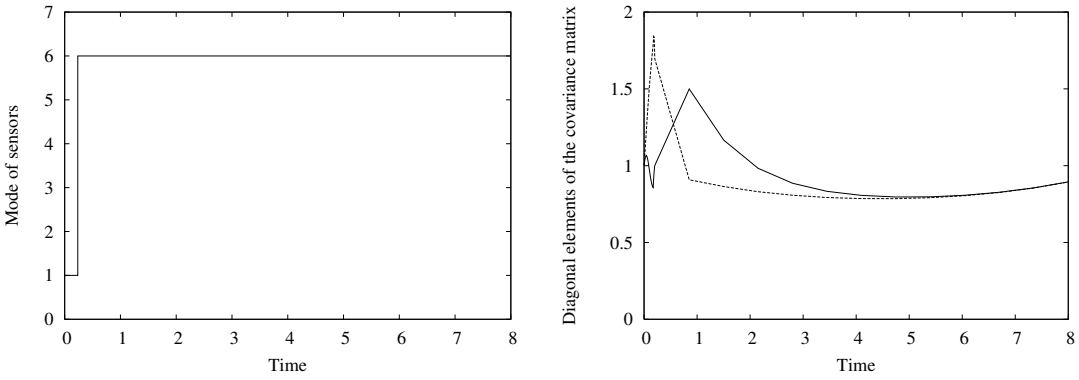

Figure 1: Optimal Sensor Operating Scheme with $P_0 = I$.

these graphs, only the first and sixth sensors are ever used, while the other four are not utilized in any optimal solution.

We also compare the solutions obtained here with those obtained from the other methods proposed in [5] and [7]. These results are summarized in Table 2. Note that the error estimation that we sought is lower than 19.6553, the optimal solution reported in [7], which was obtained using a combination of a branch and bound technique with a gradient-based method. To the best of our knowledge, $P_0 = \mathbf{0}$ is used in [7]. Note that non-zero choices of P_0 lead to even higher objective values when used in conjunction with the solution in [7].

6 Conclusions

A sensor scheduling problem is considered in this paper. It was formulated as a discrete-valued optimal control problem and then transformed into a mixed discrete optimization problem. Then, it was decomposed into a bi-level problem. A new heuristic approach,

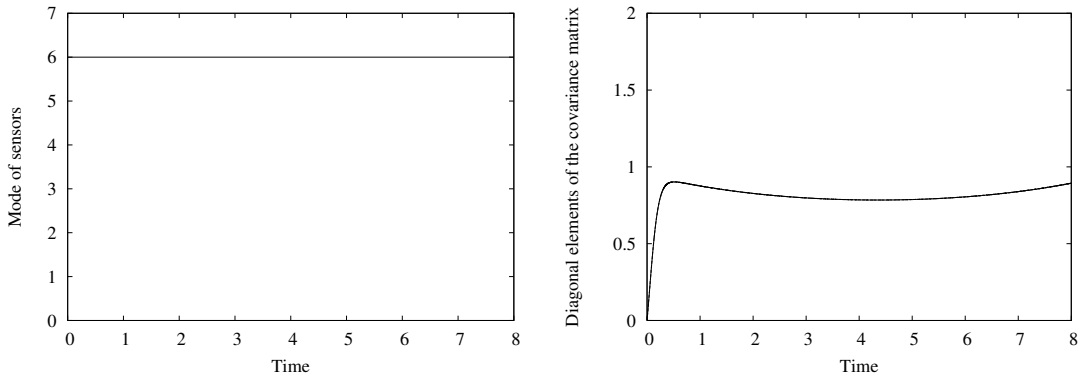

Figure 2: Optimal Sensor Operating Scheme with $P_0 = \mathbf{0}$.

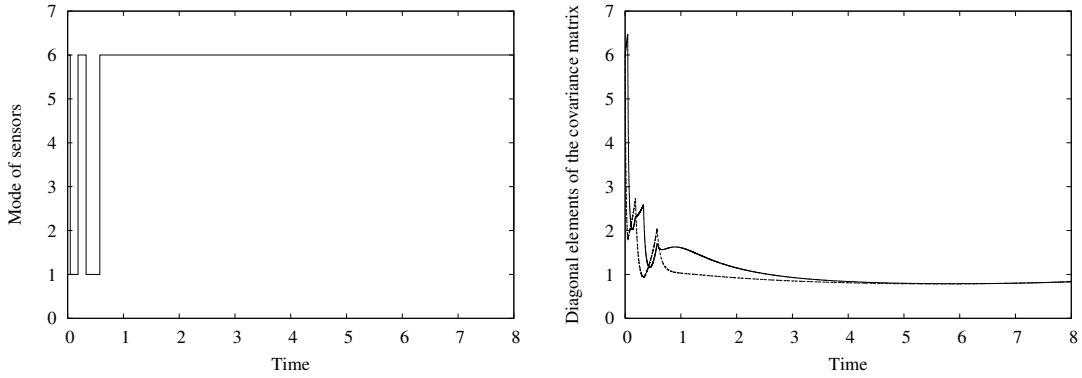

Figure 3: Optimal Sensor Operating Scheme with $P_0 = 6I$.

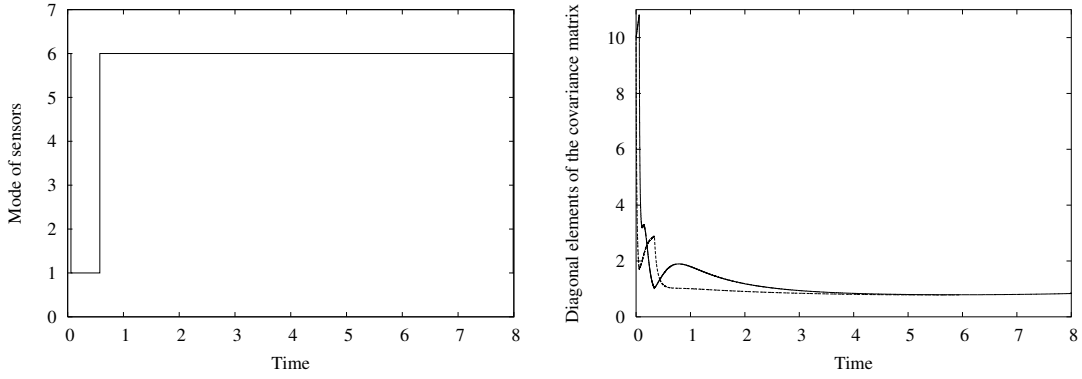

Figure 4: Optimal Sensor Operating Scheme with $P_0 = 10I$.

Table 2: A Comparison of Numerical Results with Other Methods.

Methods	Objective values
Method in [7] with $P_0 = \mathbf{0}$	19.6553
Method in [5] with $P_0 = 10I$	19.2353622
Proposed method with $P_0 = 10I$	16.5697177
Proposed method with $P_0 = 6I$	15.8781106
Proposed method with $P_0 = I$	14.3317631
Proposed method with $P_0 = \mathbf{0}$	12.9949699

which incorporates the discrete filled function algorithm into standard optimal control software, is proposed for finding a global solution of this problem. Numerical results show that the method is efficient, reliable, and robust in solving a complex discrete-valued optimal control problem. The proposed method successfully identified significantly improved solutions compared with other methods available in the literature.

Acknowledgment

The authors would like to thank David Packer from Curtin University of Technology for proof reading the early draft of this manuscript.

References

[1] Hovanessian, S.A. *Introduction to sensor systems*. Artech House, Inc., Norwood, 1988.

[2] Castanon, D. and Carin, L. Stochastic control theory for sensor management. In: *Foundations & Applications of Sensor Management* (A. Hero, D. Castanon, D. Cochran, and K. Kastella, eds.). SpringerLink, Boston, 2008, 7–32.

[3] Musick, S.H. Defence applications. In: *Foundations & Applications of Sensor Management* (A. Hero, D. Castanon, D. Cochran, and K. Kastella, eds.). SpringerLink, Boston, 2008, 257–268.

[4] Chung, T.H., Gupta, V., Hassibi, B., Burdick, J., and Murray, R.M. Scheduling for distributed sensor networks with single sensor measurement per time step. In: *IEEE International Conference on Robotics & Automation*, New Orleans, April 26 – May 1, 2004, 187–192.

[5] Lee, H.W.J., Teo, K.L., and Lim, A.E.B. Sensor scheduling in continuous time. *Automatica* **37**(12) (2001) 2017–2023.

[6] Baras, J.S. and Bensoussan, A. Optimal sensor scheduling in nonlinear filtering of diffusion process. *SIAM Journal of Control and Optimization* **27** (1989) 786–813.

[7] Feng, Z.G., Teo, K.L., and Rehbock, V. Optimal sensor scheduling in continuous time. *Dynamic Systems and Applications* **17** (2008) 331–350.

[8] Ahmed, N.U. *Linear and nonlinear filtering for scientists and engineers*. World Scientific, Singapore, 1998.

[9] Jennings, L.S., Fisher, M.E., Teo, K.L., and Goh, C.J. *MISER3.3–optimal control software: theory and user manual*. http://www.maths.uwa.edu.au/~les/MISER3.3/ch1.pdf. University of Western Australia, Perth, 2004.

[10] Ge, R. A filled function method for finding a global minimizer of a function of several variables. *Mathematical Programming* **46**(1) (1990) 191–204.

[11] Ge, R. and Huang, C. A continuous approach to nonlinear integer programming. *Applied Mathematics and Computation* **34**(1) (1989) 39–60.

[12] Zhu, W. An approximate algorithm for nonlinear integer programming. *Applied Mathematics and Computations* **93**(2-3) (1998) 183–193.

[13] Ng, C.K., Li, D., and Zhang, L.S. Discrete global descent method for discrete global optimization and nonlinear integer programming. *Journal of Global Optimization* **37**(3) (2007) 357–379.

[14] Ng, C.K., Zhang, L.S., Li, D., and Tian, W.W. Discrete filled function method for discrete global optimization. *Computational Optimization & Applications* **31**(1) (2005) 87–115.

[15] Shang, Y. and Zhang, L. A filled function method for finding a global minimizer on global integer optimization. *Journal of Computational and Applied Mathematics* **181**(1) (2005) 200–210.

[16] Yang, Y. and Liang, Y. A new discrete filled function algorithm for discrete global optimization. *Journal of Computational and Applied Mathematics* **202**(2) (2007) 280–291.

[17] Gu, Y.H. and Wu, Z.Y. A new filled function method for nonlinear integer programming problem. *Applied Mathematics and Computation* **173**(2) (2007) 938–950.

Nonlinear Dynamics and Systems Theory, **10** (2) (2010) 189–201

Optimal Guidance for Lunar Module Soft Landing

J.Y. Zhou [1*], K.L. Teo [1], D. Zhou [2] and G.H. Zhao [3]

[1] *Department of Mathematics and Statistics, Curtin University of Technology, Perth, 6102, Australia*
[2] *Department of Control Science and Engineering, Harbin Institute of Technology, Harbin, 150001, China*
[3] *Institute of Mathematical Sciences, Dalian University of Technology, DaLian, 116624, China*

Received: August 4, 2009; Revised: March 25, 2010

Abstract: In this paper, we consider an optimal control problem arising from the optimal guidance of a lunar module to achieving soft landing, where the description of the system dynamics is in a three-dimensional coordinate system. Our aim is to construct an optimal guidance law to realize the soft landing of the lunar module with the terminal attitude of the module to be within a small deviation from being vertical with respect to lunar surface, such that the fuel consumption and the terminal time are minimized. The optimal control problem is solved by applying the control parameterization technique and a time scaling transform. In this way, the optimal guidance law and the corresponding optimal descent trajectory are obtained. We then move on to consider an optimal trajectory tracking problem, where a desired trajectory is tracked such that the fuel consumption and the minimum time are minimized. This optimal tracking problem is solved using the same approach to the first optimal control problem. Numerical simulations demonstrate that the approach proposed is highly efficient.

Keywords: *optimal guidance law; lunar module; soft landing; optimal control with bounds on control and terminal states; control parameterization; time scaling transform; optimal trajectory tracking.*

Mathematics Subject Classification (2000): 49J15, 93C10, 93C15.

* Corresponding author: mailto:zhouhit@gmail.com

1 Introduction

Exploration of the moon, the nearest celestial body to the earth, is becoming more and more attractive for space scientists in recent years. Satellites and probes have been sent out to the moon for investigations. Generally speaking, there are three kinds of flight motions, i.e., flying over, circling or landing on the moon. Those missions aiming to land the lunar module safely on the surface of the moon are the most important ones.

The soft landing of the lunar module starts from the parking orbit of the moon, after Hohmann transfer the module enters into an elliptical orbit with the aposelene and perilune, which are, respectively, 110km and 15km away from the moon surface. When the module reaches the perilune, the powered descent soft landing begins. Normally, the lunar soft landing process from the perilune to the moon surface can mainly be divided into three phases. The first part is the powered deceleration phase, from 15km to 2km above the lunar surface, the module velocity is reduced to 0m/s by the propellant of the main thruster. The second part, from 2km to 100m above the lunar surface, is the attitude adjustment phase, and the module attitude is adjusted so that it is vertical to the moon surface. The last part is the vertical descent phase, a set of small thrusters is employed to cancel the moon gravity to ensure the module soft landing on the lunar surface vertically. In view of the fact that the surrounding circumstance of the moon is vacuum, lunar soft landing can not be performed in the same way as landing on the earth or mars. This is because the module can not depend on the lunar atmosphere for deceleration. One way of realizing soft landing is to use the reverse force thruster which will, however, consume much of the fuel that the lunar module is carrying. Clearly, if the fuel consumption can be reduced, then more payloads can be equipped. Thus, the optimal control strategy that guarantees the soft landing with least fuel consumption is highly in demand. Consequently, there are now many papers devoted to this area in the literature [1–6]. Meditch [7] discussed the problem of vertical lunar soft landing, where the thruster is operated at its maximum force. In this way, the mission is equivalent to a time optimal control problem and hence can be solved by existing theory. Wang [8] proposed a control scheme for achieving lunar soft landing, where the optimal control theory is used in combination with nonlinear neuro-control. Xi [9] presented an optimal control law obtained by utilizing Pontryagin Maximum Principle for the soft landing of a lunar module. Here, it is assumed that some of the control variables are not bounded. Liu [10] designed an optimal control strategy for the soft landing of a lunar module with a pre-specified terminal time by using the control parameterization technique and a time scaling transform. [1–3] and [7] studied the vertical descent phase of the lunar landing. In [4–6] and [8–10], the soft landing from the perilune to the moon surface is taken as a continuously powered descent process. However, none of these papers take into consideration the terminal angle constraint between the longitudinal axis of the module and the moon surface. In fact, among these research articles, the terminal angle of the module between its longitudinal axis and the plumb line is about fifty degree, which means that the module can not maintain a vertical attitude when it touches down on the ground. Furthermore, the dynamical system considered in most of these articles is in the two-dimensional polar coordinate system. The descent trajectory of the lunar module is assumed to remain in a vertical plane without consideration of the lateral movement. Neither the influence of the moon rotation is taken into account. However the lunar module, in practice, does not descend along such a vertical plane. To be realistic, the motion of the lunar module, which takes into consideration moon rotation, should be

described in a three-dimensional coordinate system [11].

The problem of the soft landing of a lunar module at the minimum time with the least fuel consumption can be formulated as an optimal control problem with constraints on the control and the terminal states. However, it is much too complex to be solved by using Pontryagin Maximum Principle. In this paper, we calculate the optimal descent trajectory of the lunar module by using the control parameterization technique in conjunction with a time scaling transform [12]. The lunar soft landing is treated as a continuously powered descent process with a constraint on the angle of the module between its longitudinal axis and the moon surface. During the entire process of the lunar landing, only the main reverse force thruster is needed for deceleration. Therefore, the design complexity of the guidance control law is reduced substantially. By applying the control parameterization technique and the time scaling transform, the optimal control problem is approximated by a sequence of optimal parameter selection problems. Each of which is basically a mathematical programming problem and hence can be solved by existing gradient-based optimization methods [13–15]. A general purpose optimal control software package, called MISER 3.3 [15], was developed based on these methods. We make use of this optimal control software package to solve our problem in this paper. The optimal trajectory tracking problem, where a desired trajectory is to be tracked with the least fuel consumption in the minimum time, is also considered and the same approach to the first optimal control problem is utilized to solve such an optimal trajectory tracking problem.

2 Problem Formulation

For continuously powered descent soft landing, the reverse force thruster begins to work, starting from the perilune to decelerate the initial velocity of the module. With the cooperation of the attitude control thrusters, the module is guided to reach the landing target vertically with a small and safe final velocity. In this paper, we study the optimal guidance scheme for ensuring the soft landing of the lunar module from the perilune to the moon surface.

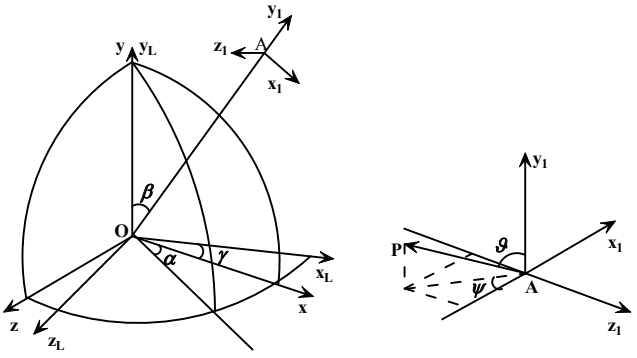

Figure 2.1: Coordinate systems.

As the influences of other celestial bodies on the lunar module are small, compared with the moon gravity, the lunar module soft landing can be treated in a two-body system [16]. The motion of the lunar module soft landing is described in a three-dimensional coordinate system (Figure 2.1). Suppose $oxyz$ and $ox_Ly_Lz_L$ are, respectively, the Lunar

Central Inertial Coordinate and Lunar Fixed Coordinate with the moon equator as the reference plane. $Ax_1y_1z_1$ is the orbit coordinate, A is the position of the lunar module. The three coordinates form a right handed system. α and β represent, respectively, the rotation angles between $oxyz$ and $Ax_1y_1z_1$. The direction of the thrust force P in the coordinate $Ax_1y_1z_1$ can be described in terms of ϑ and ψ. γ is the rotation angle between $oxyz$ and $ox_Ly_Lz_L$. Without lose of generality, we assume that $oxyz$ and $ox_Ly_Lz_L$ coincide with each other at the beginning of the soft landing. Based on Newton's second law, system dynamic equations can be derived to give [11]

$$\begin{cases} \dot{x}_L = V_{xL}, \\ \dot{y}_L = V_{yL}, \\ \dot{z}_L = V_{zL}, \\ \dot{V}_{xL} = BQV_r/m + g_{xL} - 2\omega_L V_{zL}, \\ \dot{V}_{yL} = CQV_r/m + g_{yL}, \\ \dot{V}_{zL} = DQV_r/m + g_{zL} + 2\omega_L V_{xL}, \\ \dot{m} = -Q, \end{cases} \tag{2.1}$$

where

$$\begin{aligned} B &= (\cos\alpha\cos\beta\cos\gamma - \sin\alpha\sin\gamma)\sin\vartheta\cos\psi \\ &\quad -(\sin\alpha\cos\beta\cos\gamma + \cos\alpha\sin\gamma)\sin\vartheta\sin\psi + \sin\beta\cos\gamma\cos\vartheta, \\ C &= -\cos\alpha\sin\beta\sin\vartheta\cos\psi + \cos\beta\cos\vartheta + \sin\alpha\sin\beta\sin\vartheta\sin\psi, \\ D &= (\cos\alpha\cos\beta\sin\gamma + \sin\alpha\cos\gamma)\sin\vartheta\cos\psi \\ &\quad -(\sin\alpha\cos\beta\sin\gamma - \cos\alpha\cos\gamma)\sin\vartheta\sin\psi + \sin\beta\sin\gamma\cos\vartheta. \end{aligned}$$

while x_L, y_L, z_L and V_{xL}, V_{yL}, V_{zL} are the positions and velocities in the Lunar Fixed Coordinate. m is the mass of the lunar module, Q and V_r represent, respectively, the fuel consumption rate and the specific impulse of the thruster, g_{xL}, g_{yL}, and g_{zL} denote the components of lunar gravity in $ox_Ly_Lz_L$, and ω_L is the angular velocity of the moon rotation.

Introduce two new state equations

$$\dot{\vartheta} = v, \tag{2.2}$$

$$\dot{\psi} = w \tag{2.3}$$

and let

$$\begin{aligned} \boldsymbol{x} &= [x_L, y_L, z_L, V_{xL}, V_{yL}, V_{zL}, \vartheta, \psi, m]^T \\ &= [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9]^T, \end{aligned}$$

$$\boldsymbol{u} = [Q, v, w]^T = [u_1, u_2, u_3]^T.$$

The original system dynamics (2.1) can be rewritten in the form of an affine nonlinear system given below.

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t)) + \boldsymbol{B}(\boldsymbol{x}(t))\boldsymbol{u}(t), \tag{2.4}$$

where

$$\boldsymbol{f}(\boldsymbol{x}) = [x_4, x_5, x_6, g_{xL} - 2\omega_L x_6, g_{yL}, g_{zL} + 2\omega_L x_4, 0, 0, 0]^T, \tag{2.5}$$

$$\boldsymbol{B}(\boldsymbol{x}) = \begin{bmatrix} 0 & 0 & 0 & M_1 & M_2 & M_3 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}^T \tag{2.6}$$

while

$$\begin{aligned} M_1 &= [(\cos\alpha\cos\beta\cos\gamma - \sin\alpha\sin\gamma)\sin x_7\cos x_8 \\ &\quad -(\sin\alpha\cos\beta\cos\gamma + \cos\alpha\sin\gamma)\sin x_7\sin x_8 + \sin\beta\cos\gamma\cos x_7]V_r/x_9, \\ M_2 &= [-\cos\alpha\sin\beta\sin x_7\cos x_8 + \cos\beta\cos x_7 + \sin\alpha\sin\beta\sin x_7\sin x_8]V_r/x_9, \end{aligned}$$

and

$$\begin{aligned} M_3 &= [(\cos\alpha\cos\beta\sin\gamma + \sin\alpha\cos\gamma)\sin x_7\cos x_8 \\ &\quad -(\sin\alpha\cos\beta\sin\gamma - \cos\alpha\cos\gamma)\sin x_7\sin x_8 + \sin\beta\sin\gamma\cos x_7]V_r/x_9. \end{aligned}$$

The boundedness constraints on the control vector $\boldsymbol{u} = [u_1, u_2, u_3]^T$ are specified below:

$$\boldsymbol{\alpha} \le \boldsymbol{u}(t) \le \boldsymbol{\beta},\ \forall\ t \ge 0, \tag{2.7}$$

where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \alpha_3]^T$ and $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3]^T$, while $\alpha_i,\ i = 1, 2, 3$, and $\beta_i,\ i = 1, 2, 3$, are given constants. Let $\mathcal{U}$ be the set of all such controls. Elements from $\mathcal{U}$ are called admissible controls and $\mathcal{U}$ is referred to as the class of admissible controls.

The initial conditions of the soft landing are determined by the state of the lunar module in the perilune at the initial time $t_0 = 0$. The terminal constraints are specified by the requirement of the soft landing, i.e., when the lunar module reaches the target at the terminal time t_f which is free, its velocity should be close to zero and its longitudinal axis should be close to vertical to the moon surface. So the initial conditions and terminal state constraints can be expressed as:

$$\boldsymbol{x}(t_0) = [x_{L0}, y_{L0}, z_{L0}, V_{xL0}, V_{yL0}, V_{zL0}, \vartheta_0, \psi_0, m_0]^T \tag{2.8}$$

and

$$\boldsymbol{\Phi} = \begin{bmatrix} x_L(t_f) - x_{Lr} \\ y_L(t_f) - y_{Lr} \\ z_L(t_f) - z_{Lr} \\ V_{xL}(t_f) - 0 \\ V_{yL}(t_f) - 0 \\ V_{zL}(t_f) - 0 \end{bmatrix} = 0, \tag{2.9}$$

$$\vartheta_{t_f} \le x_7(t_f) \le 0, \tag{2.10}$$

where $(x_{Lr},\ y_{Lr},\ z_{Lr})$ represents the position of the landing target in the Lunar Fixed Coordinate, ϑ_{t_f} is the terminal separation angle of the module between its longitudinal axis and the direction of the plumb line. Our aim is to design an optimal control strategy to achieve the task of soft landing of the lunar module such that conditions (2.9) and (2.10) are satisfied and the fuel consumption and the flying time are minimized. The task of minimizing the fuel consumption and the flying time is formulated as the task of minimizing the following cost function

$$J = m_0 - x_9(t_f) + t_f. \tag{2.11}$$

We may now formally state our optimal control problem as follows.
Problem (P): Given system (2.4), find a control $\boldsymbol{u} \in \mathcal{U}$ such that the cost function (2.11) is minimized subject to the control constraint (2.7), the initial condition (2.8) and the terminal state constraints (2.9) and (2.10).

3 Parameterization of the Control

To solve Problem (P), we shall utilize the control parameterization technique to approximate the control vector $\boldsymbol{u}$ with piecewise constant functions over the time interval $[0, t_f]$ as:

$$u_1^p(t) = \sum_{k=1}^{n_p} \sigma_1^k \chi_{[\tau_{k-1}, \tau_k)}(t), \tag{3.1}$$

$$u_2^p(t) = \sum_{k=1}^{n_p} \sigma_2^k \chi_{[\tau_{k-1}, \tau_k)}(t), \tag{3.2}$$

$$u_3^p(t) = \sum_{k=1}^{n_p} \sigma_3^k \chi_{[\tau_{k-1}, \tau_k)}(t), \tag{3.3}$$

where

$$\tau_0, \tau_1, \ldots, \tau_{n_p},\ \tau_{k-1} < \tau_k,\ k = 1, 2, \ldots, n_p \tag{3.4}$$

(with $\tau_0 = 0$ and $\tau_{n_p} = t_f$) are partition points of the time interval $[0, t_f]$, and $\chi_I(t)$ denotes the indicator function of I defined by

$$\chi_I(t) = \begin{cases} 1,\ t \in I, \\ 0,\ \text{elsewhere.} \end{cases} \tag{3.5}$$

Let $\boldsymbol{\tau} = [\tau_1, \ldots, \tau_{n_p}]^T$ and let Υ^p be the set which consists of all such $\boldsymbol{\tau}$. For each $j = 1, 2, 3$, and $k = 1, 2, \ldots, n_p$, σ_j^k is a constant control parameter, and τ_k , $k = 1, \ldots, n_p - 1$, are the switching times. Let $\boldsymbol{\sigma}_j = [\sigma_j^1, \cdots, \sigma_j^{n_p}]^T$, $j = 1, 2, 3$, and let $\boldsymbol{\sigma} = [(\boldsymbol{\sigma}_1)^T, (\boldsymbol{\sigma}_2)^T, (\boldsymbol{\sigma}_3)^T]^T$. Define $\boldsymbol{u}^p = [u_1^p, u_2^p, u_3^p]^T$.

As $\boldsymbol{u}^p \in \boldsymbol{\mathcal{U}}$, it is clear that

$$\alpha_j \le \sigma_j^k \le \beta_j \tag{3.6}$$

for $j = 1, 2, 3$, and $k = 1, 2, \ldots, n_p$. Let Ξ^p denote the set containing all such $\boldsymbol{\sigma}$. Here, for the soft landing of a lunar module, the terminal time $\tau_{n_p} = t_f$ is unknown and regarded as a decision variable.

We shall map all these variable time points τ_k, $k = 1, \ldots, n_p$, into fixed time points ς_k, $k = 1, \ldots, n_p$, in a new time horizon $[0,\ 1]$, such that

$$0 = \varsigma_0 < \varsigma_1 < \cdots < \varsigma_{n_p - 1} < \varsigma_{n_p} = 1. \tag{3.7}$$

For this, we introduce a new state equation defined on $[0,\ 1]$

$$\frac{dt(s)}{ds} = \mu^p(s), \tag{3.8}$$

where $t(0) = 0$, $t(1) = t_f$,

$$\mu^p(s) = \sum_{k=1}^{n_p} \delta_k \chi_{[\varsigma_{k-1}, \varsigma_k)}(s). \tag{3.9}$$

Here,

$$\delta_k \ge 0,\ k = 1, \ldots, n_p, \tag{3.10}$$

are decision variables. $\mu^p(s)$ is called the time scaling control. It is a nonnegative piecewise constant function with possible discontinuities at the pre-fixed knots ς_k, $k = 1, \ldots, n_p - 1$. Let $\boldsymbol{\delta} = [\delta_1, \cdots, \delta_{n_p}]^T$.

By applying the time scaling transform (3.8), system equations (2.4) and (3.8) are transformed into

$$\frac{d\tilde{\boldsymbol{x}}(s)}{ds} = \begin{bmatrix} \mu^p(s)[\boldsymbol{f}(t(s),\hat{\boldsymbol{x}}(s)) + \boldsymbol{B}(t(s),\hat{\boldsymbol{x}}(s))\hat{\boldsymbol{u}}^p(s)] \\ \mu^p(s) \end{bmatrix}, \tag{3.11}$$

where $\tilde{\boldsymbol{x}} = [\tilde{x}_1, \cdots, \tilde{x}_9, \tilde{x}_{10}]^T = [(\hat{\boldsymbol{x}})^T, t]^T$, $\hat{\boldsymbol{x}}(s) = \boldsymbol{x}(t(s))$, and $\hat{\boldsymbol{u}}^p(s) = \boldsymbol{u}(t(s))$ given by

$$\hat{\boldsymbol{u}}^p(s) = \sum_{k=1}^{n_p} \boldsymbol{\sigma}^k \chi_{[\varsigma_{k-1},\varsigma_k)}(s). \tag{3.12}$$

The initial condition is

$$\tilde{\boldsymbol{x}}(0) = [x_{L0}, y_{L0}, z_{L0}, V_{xL_0}, V_{yL_0}, V_{zL0}, \vartheta_0, \psi_0, m_0, 0]^T. \tag{3.13}$$

The cost function (2.11) and the terminal constraints (2.9) and (2.10) become

$$\tilde{J} = m_0 - \tilde{x}_9(1) + \tilde{x}_{10}(1) \tag{3.14}$$

and

$$\tilde{\boldsymbol{\Phi}} = \begin{bmatrix} \tilde{x}_1(1) - x_{Lr} \\ \tilde{x}_2(1) - y_{Lr} \\ \tilde{x}_3(1) - z_{Lr} \\ \tilde{x}_4(1) - 0 \\ \tilde{x}_5(1) - 0 \\ \tilde{x}_6(1) - 0 \\ \tilde{x}_{10}(1) - t_f \end{bmatrix} = 0, \tag{3.15}$$

$$\vartheta_{t_f} \le \tilde{x}_7(1) \le 0, \tag{3.16}$$

respectively. They can be written in canonical form as:

$$\tilde{g}_0(\boldsymbol{\sigma},\boldsymbol{\delta}) = \tilde{\Phi}_0(\tilde{x}(1|\boldsymbol{\sigma},\boldsymbol{\delta}),\ \boldsymbol{\sigma},\boldsymbol{\delta}) + \int_0^1 \tilde{\ell}_0(s,\ \tilde{x}(s|\boldsymbol{\sigma},\boldsymbol{\delta}),\ \boldsymbol{\sigma},\boldsymbol{\delta})ds \tag{3.17}$$

and

$$\tilde{g}_i(\boldsymbol{\sigma},\boldsymbol{\delta}) = \tilde{\Phi}_i(\tilde{x}(1|\boldsymbol{\sigma},\boldsymbol{\delta}),\ \boldsymbol{\sigma},\boldsymbol{\delta}) + \int_0^1 \tilde{\ell}_i(s,\ \tilde{x}(s|\boldsymbol{\sigma},\boldsymbol{\delta}),\ \boldsymbol{\sigma},\boldsymbol{\delta})ds = 0, \quad i = 1, ..., 7, \tag{3.18}$$

$$\tilde{g}_i(\boldsymbol{\sigma},\boldsymbol{\delta}) = \tilde{\Phi}_i(\tilde{x}(1|\boldsymbol{\sigma},\boldsymbol{\delta}),\ \boldsymbol{\sigma},\boldsymbol{\delta}) + \int_0^1 \tilde{\ell}_i(s,\ \tilde{x}(s|\boldsymbol{\sigma},\boldsymbol{\delta}),\ \boldsymbol{\sigma},\boldsymbol{\delta})ds \le 0, \quad i = 8, 9, \tag{3.19}$$

where $\tilde{\ell}_i = 0$, for $i = 0, 1, ..., 9,$, while $\tilde{\Phi}_i$, $i = 0, 1, ..., 9$, are defined by (3.14), (3.15) and (3.16), respectively.

The original optimal control problem is now approximated by a sequence of optimal parameter selection problems depending on p, the number of the partition points of the time horizon $[0,\ t_f]$, given below.

Problem ($\tilde{\mathrm{P}}(p)$): Given system (3.11) with the initial condition (3.13) on the time interval $s \in [0\ ,\ 1]$, find a control parameter vector $\boldsymbol{\sigma} \in \Xi^p$ and a switching time vector $\boldsymbol{\delta} \in \Upsilon^p$, such that the cost function (3.14) is minimized subject to the terminal constraints (3.15) and (3.16).

For each p, Problem $(\tilde{\mathrm{P}}(p))$ can be solved as a nonlinear optimization problem where the cost function (3.14) is minimized subject to the terminal constraints (3.15) and (3.16) and the constraints on the decision vectors $\boldsymbol{\sigma}$ and $\boldsymbol{\delta}$ given by (3.6) and (3.10), where the dynamical system (3.11) is used to generate the values of the cost function (3.14) and the constraint functions (3.15) and (3.16). Existing gradient-based optimization methods can be used to solve Problem $(\tilde{\mathrm{P}}(p))$. For this, we need the gradient formulas of the objective function and the constraint functions. For the constraints (3.6) and (3.10), their gradient formulas are straightforward to calculate. The gradient formulas of the objective function (3.14) and the constraint functions (3.15) and (3.16) are given bellow.

Theorem 3.1 [12] *For each $i = 0, 1, \ldots, 9$, the gradient of the function $\tilde{g}_i$ with respect to $\boldsymbol{\sigma}$ and $\boldsymbol{\delta}$ are given by*

$$\frac{\partial \tilde{g}_i(\boldsymbol{\sigma}, \boldsymbol{\delta})}{\partial \boldsymbol{\sigma}} = \int_0^1 \frac{\partial \tilde{H}_i(s, \tilde{\boldsymbol{x}}(s), \boldsymbol{\sigma}, \boldsymbol{\delta}, \tilde{\boldsymbol{\lambda}}^i(s|\boldsymbol{\sigma}, \boldsymbol{\delta}))}{\partial \boldsymbol{\sigma}} ds \tag{3.20}$$

and

$$\frac{\partial \tilde{g}_i(\boldsymbol{\sigma}, \boldsymbol{\delta})}{\partial \boldsymbol{\delta}} = \int_0^1 \frac{\partial \tilde{H}_i(s, \tilde{\boldsymbol{x}}(s), \boldsymbol{\sigma}, \boldsymbol{\delta}, \tilde{\boldsymbol{\lambda}}^i(s|\boldsymbol{\sigma}, \boldsymbol{\delta}))}{\partial \boldsymbol{\delta}} ds, \tag{3.21}$$

where

$$\tilde{H}_i(s, \tilde{\boldsymbol{x}}, \boldsymbol{\sigma}, \boldsymbol{\delta}, \tilde{\boldsymbol{\lambda}}^i) = \tilde{\ell}_i(s, \tilde{\boldsymbol{x}}, \boldsymbol{\sigma}, \boldsymbol{\delta}) + (\tilde{\boldsymbol{\lambda}}^i)^T \tilde{\boldsymbol{f}}(s, \tilde{\boldsymbol{x}}, \boldsymbol{\sigma}, \boldsymbol{\delta}) \tag{3.22}$$

and, for each $i = 0, 1, \ldots, 9$, $\tilde{\boldsymbol{\lambda}}^i(s|\boldsymbol{\sigma}, \boldsymbol{\delta})$ is the solution of the following co-state system corresponding to $(\boldsymbol{\sigma}, \boldsymbol{\delta})$:

$$\frac{d(\tilde{\boldsymbol{\lambda}}(s))^T}{ds} = -\frac{\partial \tilde{H}_i(s, \tilde{\boldsymbol{x}}(s|\boldsymbol{\sigma}, \boldsymbol{\delta}), \boldsymbol{\sigma}, \boldsymbol{\delta}, \tilde{\boldsymbol{\lambda}}(s))}{\partial \tilde{\boldsymbol{x}}}, \ s \in [0, 1) \tag{3.23}$$

with

$$(\tilde{\boldsymbol{\lambda}}(1))^T = \frac{\partial \tilde{\Phi}_i(\tilde{\boldsymbol{x}}(1|\boldsymbol{\sigma}, \boldsymbol{\delta}))}{\partial \tilde{\boldsymbol{x}}}. \tag{3.24}$$

Proof The proof of Theorem 3.1 is similar to that given for Theorem 5.2.1 of [12].

For each p, Problem $(\tilde{\mathrm{P}}(p))$ is an optimal parameter selection problem, which can be viewed as a nonlinear optimization problem. The gradient formulas of the cost function (3.17) and the constraint functions (3.18) and (3.19) are given in Theorem 3.1, while the constraints (3.6) are just the bounds for these control parameter vectors.

Thus, any existing gradient-based optimization method, such as sequential quadratic programming algorithm [17], can be used to solve Problem $(\tilde{\mathrm{P}}(p))$. The optimal control software MISER 3.3 was implemented based on these ideas. It is used in this paper to solve our optimal control problem. Intuitively, the larger the p, the closer Problem $(\tilde{\mathrm{P}}(p))$ is to Problem (P). This intuition is true. We shall briefly discuss the convergence issue as follows. Let $(\boldsymbol{\sigma}^{p,*}, \boldsymbol{\delta}^{p,*})$ be the optimal parameter vector of Problem $(\tilde{\mathrm{P}}(p))$, and let $\tilde{\boldsymbol{u}}^{p,*}$ be the corresponding piecewise constant control given by

$$\tilde{\boldsymbol{u}}^{p,*}(s) = \sum_{k=1}^{n_p} \boldsymbol{\sigma}^{p,*} \chi_{[\frac{k-1}{n_p}, \frac{k}{n_p})}(s), \tag{3.25}$$

where

$$\tilde{\boldsymbol{u}}^{p,*} = [\tilde{u}_1^{p,*}, \tilde{u}_2^{p,*}, \tilde{u}_3^{p,*}]^T, \tag{3.26}$$

$$\boldsymbol{\sigma}^{p,*} = [(\boldsymbol{\sigma}_1^{p,*})^T, (\boldsymbol{\sigma}_2^{p,*})^T, (\boldsymbol{\sigma}_3^{p,*})^T]^T, \tag{3.27}$$

$$\boldsymbol{\delta}^{p,*} = [\delta_1^{p,*}, \ldots, \delta_{n_p}^{p,*}]^T. \tag{3.28}$$

In the original time horizon $[0, t_f]$, we have

$$\boldsymbol{u}^{p,*}(t) = \sum_{k=1}^{n_p} \boldsymbol{\sigma}^{p,*} \chi_{[\tau_{k-1}^{p,*}, \tau_k^{p,*})}(t), \tag{3.29}$$

where

$$\tau_i^{p,*} = \sum_{k=1}^{i} \delta_k^{p,*}, \quad i = 1, \ldots, n_p. \tag{3.30}$$

Furthermore, let $\boldsymbol{u}^*$ be the optimal control of Problem (P). Then, by virtue of the discussion presented in Section 5 on Convergence Analysis of [18], it holds that
(i) $g_0(\boldsymbol{u}^{p,*}) \to g_0(\boldsymbol{u}^*)$;
(ii) if $\boldsymbol{u}^{p,*} \to \tilde{\boldsymbol{u}}$ almost everywhere in $[0, t_f]$, then $\tilde{\boldsymbol{u}}$ is an optimal control of Problem (P).

From our extensive simulation study experience, we observe that p does not need to be chosen to be too large. In fact, the difference in the cost values between $p = 20$ and those with larger p is, in general, very insignificant. Thus, $p = 20$ is chosen in our numerical simulation.

4 Optimal Trajectory Tracking

We now move on to consider a situation for which the spacecraft is required to track a desired trajectory, such that the fuel consumption and the terminal time are minimized. To realize such an optimal tracking control problem, we only need to modify the cost function J of Problem (P) as:

$$J = m_0 - x_9(t_f) + t_f + \int_0^{t_f} [(x_1(t) - \bar{x}_r(t))^2 + (x_2(t) - \bar{y}_r(t))^2 + (x_3(t) - \bar{z}_r(t))^2]dt, \tag{4.1}$$

where $(\bar{x}_r,\ \bar{y}_r,\ \bar{z}_r)$ denotes the desired reference trajectory. Let this optimal trajectory control problem be referred to as Problem (Q). Using the control parameterization technique and the time scaling transform as described in Section 3, Problem (Q) is transformed into Problem $(\tilde{\mathrm{Q}}(p))$, where the transformed cost function

$$\tilde{J} = m_0 - \tilde{x}_9(1) + \tilde{x}_{10}(1) + \int_0^1 [(\tilde{x}_1(s) - \hat{x}_r(s))^2 + (\tilde{x}_2(s) - \hat{y}_r(s))^2 + (\tilde{x}_3(s) - \hat{z}_r(s))^2]ds \tag{4.2}$$

is to be minimized over $(\Xi^p \times \Upsilon^p)$ subject to the system dynamic (3.11) with initial condition (3.13) and the terminal state constraints (3.15) and (3.16), where $\hat{x}_r(s) = \bar{x}_r(t(s))$, $\hat{y}_r(s) = \bar{y}_r(t(s))$, $\hat{z}_r(s) = \bar{z}_r(t(s))$.

The gradient formulas of the cost function (4.2) and constraint functions (3.15) and (3.16) can be derived in the same way as those of Problem $(\tilde{\mathrm{P}}(p))$ given in Theorem 3.1. The optimal control parameter selection problem $(\tilde{\mathrm{Q}}(p))$ is thus solved utilizing the optimal control software MISER 3.3.

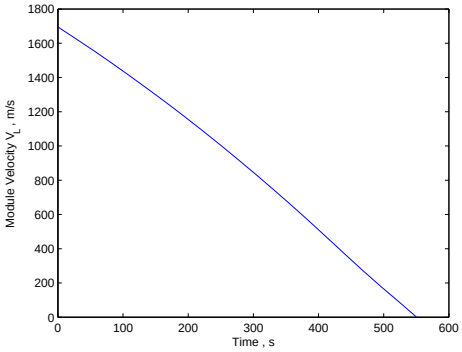

Figure 5.1: Module velocity V_L.

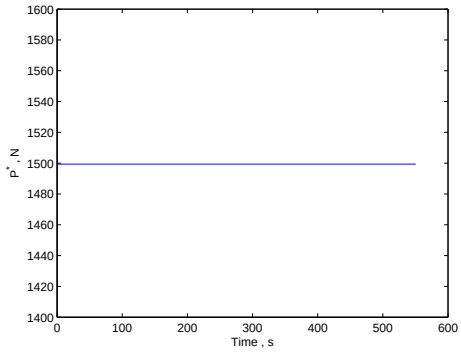

Figure 5.2: Thrust force P^*.

5 Numerical Simulations

The initial conditions of the lunar module are given as: $x_{L0} = 8.19371 \times 10^5$m, $y_{L0} = 1.428867 \times 10^6$m, $z_{L0} = 5.996306 \times 10^5$m, $V_{xL0} = 1115$m/s, $V_{yL0} = -981.82$m/s, $V_{zL0} = 816$m/s, $m_0 = 600$kg. At the initial time of the soft landing, the rotation angle $\gamma(t_0) = 0°$. Specific impulse $V_r = 300 \times 9.8$m/s and angular velocity of the moon rotation $\omega_L = 2.661699 \times 10^{-6}$rad/s.

We first consider the task of achieving the soft landing of the lunar module. The landing target is in Mare Imbrium on the moon surface, which is located at 38.628° North latitude and 36.806° West longitude. Control variables are chosen subject to the bounds: $0 \text{ kg/s} \le \sigma_1^k \le 0.51 \text{ kg/s}$, $\left|\sigma_2^k\right| \le 1 \ °/\text{s}$, $\left|\sigma_3^k\right| \le 1 \ °/\text{s}$, $k = 1, 2, \ldots, n_p$. Terminal separation angle of the module between its longitudinal axis and the plumb line is $\vartheta_{t_f} = 5°$. The scaled time interval is $s \in [0, 1]$ partitioned into 20 equal subintervals. Terminal time of the soft landing is free to vary. The corresponding optimal parameter selection problem is then solved by using the software MISER 3.3. Terminal conditions of the lunar module obtained are listed below.

$$x_L(t_f) = 1.0871218 \times 10^6 m, \ y_L(t_f) = 1.0849749 \times 10^6 m, \ z_L(t_f) = 8.134568 \times 10^5,$$

$$V_{xL}(t_f) = 1 \times 10^{-4}\text{m/s}, \ V_{yL}(t_f) = 0\text{m/s}, \ V_{zL}(t_f) = 2 \times 10^{-4}\text{m/s}.$$

Figure 5.1 shows the time history in the original time horizon $[0, \ t_f]$ of the lunar module velocity. We see that it converges smoothly to zero as the module lands on the moon. Figures 5.2, 5.4 and 5.6 are optimal control outputs during the period of soft landing, also in the original time horizon $[0, \ t_f]$. Here, we see that the reverse force thruster works at its maximum thrust force all the time, while the two angular velocity controllers are operating within their bounds. Under the optimal control law, the lunar module is guided to the target precisely, and the optimal descent trajectory is shown in Figure 5.3. Terminal mass of the module is 319.2728kg. Figure 5.5 depicts the time scaling control. Lunar module lands on the moon surface vertically after 550.4455s, with the terminal separation angle between the module longitudinal axis and the plumb line $\vartheta(t_f) = -4.998°$.

Our next task is to investigate the mission of the optimal trajectory tracking. Suppose the desired trajectory is the one obtained from the solution of Problem (P). Suppose that the initial position of the lunar module is given as $x_{L0} = 8.18348 \times 10^5$m, $y_{L0} = 1.428821 \times 10^6$m, $z_{L0} = 6.01136 \times 10^5$m, which are different from those for Problem (P).

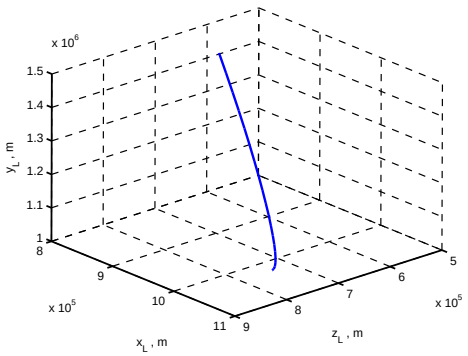

Figure 5.3: Optimal descent trajectory.

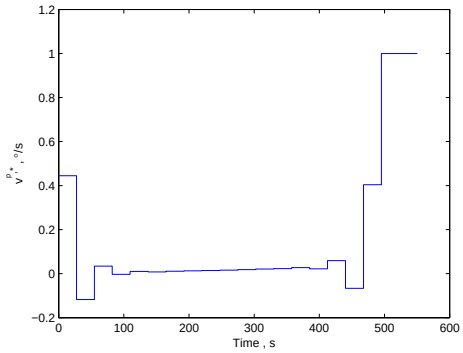

Figure 5.4: Angular velocity $v^{p,*}$.

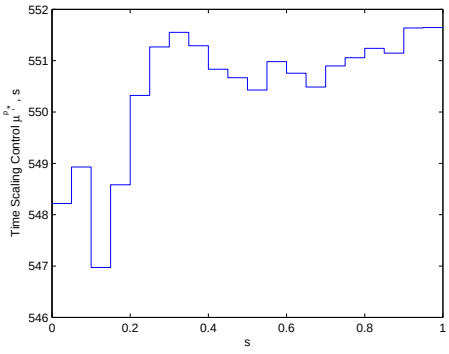

Figure 5.5: Time scaling control $\mu^{p,*}$.

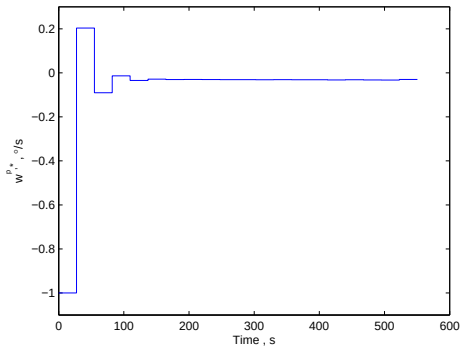

Figure 5.6: Angular velocity $w^{p,*}$.

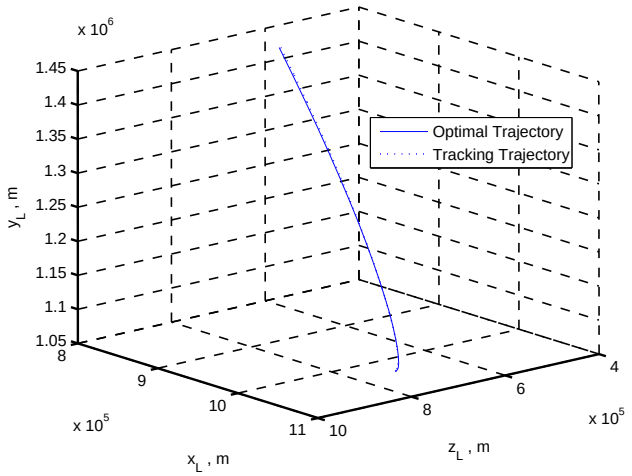

Figure 5.7: Optimal trajectory tracking.

Let this optimal tracking problem be referred to as Problem (Q). It is solved by using the approach detailed in Section 4, where the optimal control software MISER 3.3 is utilized. The optimal control obtained for Problem (P) is used as the initial guess for the search of the optimal control of Problem (Q). Let the optimal control of Problem (Q) obtained be denoted as $\boldsymbol{\upsilon}^*$. Then, under this control, the Lunar module is guided to the target at the terminal time $t_f = 572.8$s. The terminal velocity is 6.2e-4m/s, while the terminal mass is 315.43kg. From Figure 5.7, we see that the optimal trajectory tracks the desired trajectory satisfactorily.

6 Conclusions

This paper studied the soft landing of the lunar module, where its system dynamics is described in a three-dimensional coordinate system. The constraints on the control and the terminal state are also taken into consideration. By using the control parameterization technique and the time scaling transform, the optimal control problem is solved as an optimal parameter selection problem by the optimal control software package MISER 3.3, yielding an optimal control law. This optimal control law steers the lunar module to achieve the pre-specified landing target precisely such that the fuel consumption and the terminal time are minimized. The module touches down on the moon vertically with reference to lunar surface. The task of optimal trajectory tracking was also formulated and solved. The proposed approach is highly effective.

References

[1] Ma, K.M., Chen, L.J. and Wang, Z.C. Practical design of control law for flight vehicle soft landing. *Missiles and Space Vehicles* (2) (2001) 39–43.

[2] Ruan, X.G. A nonlinear nuerocontrol scheme for lunar soft landing. *Journal of Astronautics* **19**(1) (1998) 35–43.

[3] Hebertt, S.R. Soft landing on a planet: a trajectory planning approach for the liouvillian model. *Proceedings of American Control Conference* (1999) 2936–2940.

[4] Wang, D.Y., Li, T.S. and Ma, X.R. Numerical solution of TPBVP in optimal lunar soft landing. *Aerospace Control* (3) (2000) 44–49.

[5] Xu, M. and Li, J.F. Optimal control of lunar soft landing. *Journal of Tsinghua University (Science and Technology)* **41**(8) (2001) 87–89.

[6] Wang, Z., Li, J.F., Cui, N.G. and Liu, T. Genetic algorithm optimization of lunar probe soft landing trajectories. *Journal of Tsinghua University (Science and Technology)* **43**(8) (2003) 1056–1059.

[7] Meditch, S.J. On the problem of optimal thrust programming for a lunar soft landing. *IEEE Transaction on Automatic Control* **9**(4) (1964) 477–484.

[8] Wang, D.Y., Li, T.S., Yan, H. and Ma, X.R. Neuro-optimal guidance control for lunar module soft landing. *Journal of Systems Engineering and Electronics* **10**(3) (1999) 22–31.

[9] Xi, X.N., Zeng, G.Q. Ren, X. and Zhao, H.Y. *Orbit design of lunar probe*. National Defence Industry Press, Beijing, 2001.

[10] Liu, X.L., Duan, G.R. and Teo, K.L. Optimal soft landing control for moon lander. *Automatica* **44**(4) (2008) 1097–1103.

[11] Zhou, J.Y. and Zhou, D. Precise modeling and optimal orbit design of lunar modules soft landing. *Journal of Astronautics* **28**(6) (2007) 1462–1466.

[12] Teo, K.L., Goh, C.J. and Wong, K.H. *A unified computational approach to optimal control problems*. Longman Scientific and Technical, Harlow, 1991.

[13] John, B.T. *Practical methods for optimal control using nonlinear programming*. Society for Industrial and Applied Mathematics, Philadelphia, 2001.

[14] Kenneth, H., Frode, M. and Edvall, M.M. User's guide for tomlab/scos, *(http://tomopt.com/tomlab/products/socs/)*, 2006.

[15] Jennings, L.S., Fisher, M.E., Teo, K.L. and Goh, C.J. MISER3: Solving optimal control problems - an update, *Advance Engineering Software* (1991) 190–196.

[16] Prado, A.F.B.A. A survey on space trajectories in the model of three bodies. *Nonlinear Dynamics and Systems Theory* **6**(4) (2006) 389–400.

[17] Luenberger, D.G. The Gradient Projection Method Along Geodesics. *Management Science* **18**(11) (1972) 620–631.

[18] Teo, K.L., Jennings, L.S., Lee, H.W.J. and Rehbock, V. The control parameterization enhancing transform for constraint optimal control problems. *J. Austral. Math. Soc. Ser. B* 40 (1999) 314–335.