Springer Proceedings in Mathematics & Statistics

Steve Baigent
Martin Bohner
Saber Elaydi *Editors*

Progress on Difference Equations and Discrete Dynamical Systems

25th ICDEA, London, UK, June 24–28, 2019

Springer

Springer Proceedings in Mathematics & Statistics

Volume 341

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

More information about this series at http://www.springer.com/series/10533

Steve Baigent · Martin Bohner · Saber Elaydi
Editors

Progress on Difference Equations and Discrete Dynamical Systems

25th ICDEA, London, UK, June 24–28, 2019

Springer

Editors
Steve Baigent
Department of Mathematics
University College London
London, UK

Saber Elaydi
Department of Mathematics
Trinity University
San Antonio, TX, USA

Martin Bohner
Department of Mathematics and Statistics
Missouri University of Science
and Technology
Rolla, MO, USA

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-030-60106-5 ISBN 978-3-030-60107-2 (eBook)
https://doi.org/10.1007/978-3-030-60107-2

Mathematics Subject Classification: 34N05, 37Exx, 39Axx, 92Dxx

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

Preface

The 25th International Conference on Difference Equations and Applications was held at UCL (University College London) from June 24–28, 2019, under the auspices of the International Society of Difference Equations (ISDE). Over 120 researchers from over 35 countries participated in the conference which was hosted by the UCL Mathematics Department. There was a very busy programme with 8 plenary talks and over 90 contributed talks spread over four and a half days.

The plenary speakers were Paul Glendinning (UK), Mats Gyllenberg (Finland), Mihàly Pituk (Hungary), Adina Luminiţa Sasu (Romania), Ewa Schmeidel (Poland), Andrey Shilnikov (USA), Horst Thieme (USA) and Patricia Wong (Singapore).

There was a wide variety of topics covered at the conference. Difference equations pervade mathematics and the topics covered included chaos, bifurcation theory, renormalization theory, exponential dichotomies, dynamical systems on time scales, monotone systems theory, stability theory, integrable systems and many other areas. In addition, there were applications of difference equations to a diverse set of subjects such as ecology, neuroscience, epidemiology, economics and control theory to mention a few. There were also special sessions of more of a pure mathematical flavour organized for Nevanlinna theory and discrete integrable dynamics.

This book is composed of contributions from both plenary speakers and conference participants. It reflects well the sheer length and breadth of material covered in just four and a half days. The first part of the book is formed from chapters contributed by plenary speakers, whereas the second part contains articles by attendees on the topics that they spoke on at the meeting.

Any international conference of this size takes a fair amount of organization. So it is entirely appropriate to conclude by offering our gratitude to all of those who contributed to the success of the conference. In particular, at UCL we would like to mention Professors Rodney Halburd and Steven Bishop as fellow organizers, and Belgin Seymenoğlu, Jason Vittis and Jordan Hofmann who helped enormously with the day-to-day running of the conference, and finally Soheni Francis who oversaw over the administration of the entire event.

Finally, we would like to acknowledge the generous support of our sponsors, the UCL Mathematics Department and the Taylor & Francis group.

London, UK — Steve Baigent
Rolla, MO, USA — Martin Bohner
San Antonio, TX, USA — Saber Elaydi
July 2020

Contents*

*Contents are ordered according to the names of speakers who presented talks at ICDEA2019, and whenever there are multiple authors, the speaker's name is capitalised.

Papers by Plenary Speakers

Caputo Nabla Fractional Boundary Value Problems

ALLAN PETERSON and Wei Hu

Abstract We study boundary value problems with the Caputo nabla difference in the context of discrete fractional nabla calculus, especially when the right boundary condition has a fractional order. We first construct the Green's function for the general case and study the properties of the Green's function in several cases. We then apply the cone theory in a Banach space to show the existence of positive solutions to a nonlinear boundary value problem.

Keywords Discrete fractional calculus · Boundary value problems · Green's function

1 Nabla Fractional Calculus

In this chapter, we introduce the notation, definitions, and results concerning nabla fractional calculus. Most of these results can be found in the monograph [12] by Goodrich and Peterson.

1.1 Basic Definitions

Definition 1 For $a, b \in \mathbb{R}$ and $b - a \in \mathbb{Z}^+ := \{1, 2, \dots\}$, the sets $\mathbb{N}_a$ and $\mathbb{N}_a^b$ are defined by

$$\mathbb{N}_a := \{a, a+1, a+2, \dots\} \quad \text{and} \quad \mathbb{N}_a^b := \{a, a+1, a+2, \dots b\}.$$

A. PETERSON (✉)
University of Nebraska-Lincoln, Lincoln, NE 68588, USA
e-mail: apeterson1@unl.edu; tinaandal@gmail.com

W. Hu
Tusculum University, Greeneville, TN 37615, USA
e-mail: whu@tusculum.edu

S. Baigent et al. (eds.), *Progress on Difference Equations and Discrete Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 341,
https://doi.org/10.1007/978-3-030-60107-2_1

Definition 2 The nabla operator (backwards difference operator), ∇, for $f : \mathbb{N}_a \to \mathbb{R}$ is defined by

$$(\nabla f)(t) = f(t) - f(\rho(t)),$$

where $\rho(t) := t - 1$ is the backward jump operator.

The operator ∇^n is defined recursively by

$$\nabla^n f(t) := \nabla(\nabla^{n-1} f(t))$$

for $t \in \mathbb{N}_{a+n}$, $n \in \mathbb{N}_1$, where $f : N_{a-n} \to \mathbb{R}$, and ∇^0 is the identity operator.

Lemma 1 *The **binomial expression** for $\nabla^N f(t)$, where $N \in \mathbb{N}_0$ and $f : \mathbb{R} \to \mathbb{R}$ is given by*

$$\nabla^N f(t) = \sum_{j=0}^{N} (-1)^j \binom{N}{j} f(t-j)$$

for $t \in \mathbb{R}$.

Definition 3 The **rising function** is defined by

$$t^{\overline{n}} := t(t+1)\cdots(t+n-1),$$

for $t \in \mathbb{R}$ and $n \in \mathbb{N}_1$.

Remark 1 Note that

$$\begin{aligned} t^{\overline{n}} &= t(t+1)\cdots(t+n-1) \\ &= \frac{\Gamma(t)\cdot t(t+1)\cdots(t+n-1)}{\Gamma(t)} \\ &= \frac{\Gamma(t+n)}{\Gamma(t)}, \quad t \notin -\mathbb{N}_0, \end{aligned}$$

where Γ is the gamma function.

Definition 4 Motivated by Remark 1, we define the **(generalized) rising function** by

$$t^{\overline{r}} := \frac{\Gamma(t+r)}{\Gamma(t)}$$

for the values of t and r such that the right-hand side of this equation makes sense. We adopt the convention that $t^{\overline{r}} := 0$ for t a nonpositive integer but $t + r$ not a nonpositive integer.

1.2 Nabla Fractional Sums and Differences

Definition 5 For $f : \mathbb{N}_{a+1} \to \mathbb{R}$ and $t \in \mathbb{N}_a$, we define the nabla integral of f from a to t by

$$\int_a^t f(\tau)\nabla\tau := \sum_{\tau=a+1}^{t} f(\tau), \quad t \in \mathbb{N}_a$$

with the convention that the integral is zero if the upper limit of the summation is less than the lower limit.

Definition 6 For $\mu \notin \mathbb{Z}^-$, we define the μ-th order nabla fractional Taylor monomial (based at a) by

$$H_\mu(t, a) := \frac{(t-a)^{\overline{\mu}}}{\Gamma(\mu+1)},$$

whenever the right-hand side is meaningful.

Remark 2 For $\mu \notin \mathbb{Z}^-$ and $a, b \in \mathbb{R}$, we see that

$$H_\mu(t, a) = H_\mu(t+b, a+b)$$

by the definition of Taylor monomials.

In the next theorem, we list several important properties of the Taylor monomials.

Theorem 1 ([12], Theorem 3.57) *For $\mu \notin \mathbb{Z}^-$, the μ-th order nabla fractional Taylor monomial has the following properties:*

(i) $H_\mu(a, a) = 0$;
(ii) $\nabla H_\mu(t, a) = H_{\mu-1}(t, a)$;
(iii) $\displaystyle\int_a^t H_\mu(s, a)\nabla s = H_{\mu+1}(t, a)$;
(iv) $\displaystyle\int_a^t H_\mu(t, \rho(s))\nabla s = H_{\mu+1}(t, a)$;
(v) *for* $k \in \mathbb{N}_1$, $H_{-k}(t, a) = 0$, $t \in \mathbb{N}_a$.

provided the expressions in this theorem are well defined.

Lemma 2 *For $t > a$,*

$$\sum_{k=0}^{N-1} H_k(t, a) = H_{N-1}(t, a-1).$$

Proof We proceed by induction. Let $N = 2$ be the base case. We have

$$\sum_{k=0}^{1} H_k(t, a) = H_0(t, a) + H_1(t, a) = 1 + t - a = H_1(t, a-1).$$

Assume $\sum_{k=0}^{M-1} H_k(t,a) = H_{M-1}(t,a-1)$, then we have

$$\begin{aligned}
\sum_{k=0}^{M} H_k(t,a) &= \sum_{k=0}^{M-1} H_k(t,a) + H_M(t,a) \\
&= H_{M-1}(t,a-1) + H_M(t,a) \\
&= \frac{(t-a+1)^{\overline{M-1}}}{(M-1)!} + \frac{(t-a)^{\overline{M}}}{M!} \\
&= \frac{M(t-a+M)}{M(t-a+M)} \frac{(t-a+1)^{\overline{M-1}}}{(M-1)!} + \frac{(t-a+M)}{(t-a+M)} \frac{(t-a)^{\overline{M}}}{M!} \\
&= \Big(\frac{M}{t-a+M} + \frac{t-a}{t-a+M} \Big) \frac{(t-a+1)^{\overline{M}}}{M!} \\
&= H_M(t,a-1).
\end{aligned}$$

□

Remark 3 The above lemma is also seen in [10], where Gensler gives a proof using Pochhammer polynomials.

Definition 7 Assume $f : \mathbb{N}_{a+1} \to \mathbb{R}$ and $\mu > 0$. Then the nabla fractional sum is defined by

$$\nabla_a^{-\mu} f(t) := \int_a^t H_{\mu-1}(t, \rho(s)) f(s) \nabla s,$$

for $t \in \mathbb{N}_a$, where by convention $\nabla_a^{-\mu} f(a) = 0$.

Definition 8 (*Riemann-Liouville Nabla Fractional Difference*) Let $f : \mathbb{N}_{a+1-N} \to \mathbb{R}$ and $\nu \in \mathbb{R}^+$. We define the ν-th order nabla fractional difference of f by

$$\nabla_a^{\nu} f(t) := \nabla^N \nabla_a^{-(N-\nu)} f(t)$$

for $t \in \mathbb{N}_{a+1}$, where $N = \lceil \nu \rceil$.

The following theorem from [1] shows that the fractional difference $\nabla_a^{\nu} f(t)$ is obtained from the fractional sum $\nabla_a^{\nu} f(t)$ by replacing ν by $-\nu$.

Theorem 2 *Assume $f : \mathbb{N}_a \to \mathbb{R}$, $\nu > 0$, and $\nu \notin \mathbb{N}_1$. Then*

$$\nabla_a^{\nu} f(t) = \int_a^t H_{-\nu-1}(t, \rho(s)) f(s) \nabla s$$

In the following theorem, we list some additional important results for Taylor monomials.

Theorem 3 ([12], Theorem 3.93. Generalized Power Rules) *Let $\nu \in \mathbb{R}^+$ and $\mu \in \mathbb{R}$ such that $\mu, \nu + \nu$, and $\mu - \nu$ are nonnegative integers. Then the following hold for $t \in \mathbb{N}_a$.*

(i) $\nabla_a^{-\nu} H_\mu(t, a) = H_{\mu+\nu}(t, a)$,
(ii) $\nabla_a^{\nu} H_\mu(t, a) = H_{\mu-\nu}(t, a)$,
(iii) $\nabla_a^{-\nu}(t-a)^{\overline{\mu}} = \dfrac{\Gamma(\mu+1)}{\Gamma(\mu+\nu+1)}(t-a)^{\overline{\mu+\nu}}$,
(iv) $\nabla_a^{\nu}(t-a)^{\overline{\mu}} = \dfrac{\Gamma(\mu+1)}{\Gamma(\mu+\nu+1)}(t-a)^{\overline{\mu-\nu}}$.

Later in this paper we will use Laplace transforms in some of our later proofs. We will use the following results.

Definition 9 ([12, Definition 3.64]) Assume $f : \mathbb{N}_{a+1} \to \mathbb{R}$, then we define the Laplace transform of f (based at a) by

$$\mathcal{L}_a\{f\}(s) = \sum_{k=1}^{\infty}(1-s)^{k-1} f(a+k)$$

for those values of s such that the above infinite series converges.

Definition 10 ([12, Definition 3.77]) For $f, g : \mathbb{N}_{a+1} \to \mathbb{R}$, we define the nabla convolution product of f and g by

$$(f * g)(t) := \int_a^t f(t - \rho(s) + a) g(s) \nabla s, \qquad t \in \mathbb{N}_{a+1}.$$

Theorem 4 ([12, Theorem 3.80]) *For $f : \mathbb{N}_{a+1} \to \mathbb{R}$ and ν not a nonpositive integer, we have*

$$\nabla_a^{-\nu} f(t) = (H_{\nu-1}(\cdot, a) * f)(t), \qquad t \in \mathbb{N}_{a+1}.$$

Theorem 5 ([12, Theorem 3.81] Nabla Convolution Theorem) *Assume $f, g : \mathbb{N}_{a+1} \to \mathbb{R}$ and their nabla Laplace transforms converge for $|s-1| < r$. Then*

$$\mathcal{L}_a\{f * g\}(s) = \mathcal{L}_a\{f\}(s)\mathcal{L}_a\{g\}(s)$$

for $|s-1| < r$.

Theorem 6 ([12, Theorem 3.82]) *Assume $\mu > 0$ and the nabla Laplace transform of $f : \mathbb{N}_{a+1} \to \mathbb{R}$ converges for $|s-1| < r$ for some $r > 0$. Then*

$$\mathcal{L}_a\{\nabla_a^{-\mu} f\}(s) = \frac{1}{s^\mu}\mathcal{L}_a\{f\}(s),$$

for $|s-1| < r$.

2 Caputo Nabla Fractional Differences

One of our main interests in this paper is to consider the Caputo fractional equation

$$\nabla_{a*}^{\nu} y(t) = h(t),$$

where the operator ∇_{a*}^{ν} is the ν Caputo fractional difference operator. In the section of our paper we define and give several important properties of this Caputo fractional difference operator, which we will use to prove our main results.

Definition 11 (*Caputo Nabla Fractional Difference*) Assume $f : \mathbb{N}_{a+1-N} \to \mathbb{R}$ and $\mu > 0$. Then the μ-th Caputo nabla fractional difference of f is defined by

$$\nabla_{a^*}^{\mu} f(t) := \nabla_a^{-(N-\mu)} \nabla^N f(t)$$

for $t \in \mathbb{N}_{a+1}$, where $N = \lceil \mu \rceil$.

We note some differences between the Caputo nabla difference and the nabla Riemann-Liouville difference in the following remark.

Remark 4 For $\mu > 0$ and any constant C, we have for the Caputo case

$$\nabla_{a^*}^{\mu} C = \nabla_a^{-(N-\mu)} \nabla^N C = 0.$$

But for the nabla Riemann-Liouville case we get that

$$\begin{aligned} \nabla_a^{\mu} C &= \nabla^N \nabla_a^{-(N-\mu)} C \\ &= \nabla^N \int_a^t H_{N-\mu-1}(t, \rho(s)) C \nabla s \\ &= \nabla^N C H_{N-\mu}(t, a) \qquad \text{(by Theorem 1, (iii))} \\ &= C H_{-\mu}(t, a) \qquad \text{(by Theorem 1, (ii))} \end{aligned}$$

for $t \in \mathbb{N}_a$, which is zero only for μ a positive integer.

3 Composition Rules

In this section we will give several important compositions rules.

Lemma 3 ([12], Lemma 3.108) *Let $k \in \mathbb{N}_0$, $\mu > 0$, and $N = \lceil \mu \rceil$. Then*

$$\nabla^k \nabla_a^{-\mu} f(t) = \nabla_a^{k-\mu} f(t)$$

and

$$\nabla^k \nabla_a^{\mu} f(t) = \nabla_a^{k+\mu} f(t)$$

for $t \in \mathbb{N}_{a+k}$.

The following lemma appears in Ariel Setniker's dissertation [22]. Here we give a different proof.

Lemma 4 *For* $\mu \in \mathbb{R}^+$ *and* $f : \mathbb{N}_{a+1} \to \mathbb{R}$,

$$\nabla_a^{-\mu} \nabla f(t) = \nabla \nabla_a^{-\mu} f(t) - H_{\mu-1}(t, a) f(a)$$

for $t \in \mathbb{N}_a$.

Proof

$$\begin{aligned}
\nabla_a^{-\mu} \nabla f(t) - \nabla \nabla_a^{-\mu} f(t) &= \int_a^t H_{\mu-1}(t, \rho(s)) \nabla f(s) \nabla s - [\nabla_a^{-\mu} f(t) - (\nabla_a^{-\mu} f)(t-1)] \\
&= \sum_{s=a+1}^{t} H_{\mu-1}(t, \rho(s)) \nabla f(s) - [\nabla_a^{-\mu} f(t) - (\nabla_a^{-\mu} f)(t-1)] \\
&= \sum_{s=a+1}^{t} H_{\mu-1}(t, \rho(s))(f(s) - f(s-1)) - \nabla_a^{-\mu} f(t) \\
&\quad + \sum_{s=a+1}^{t-1} H_{\mu-1}(t-1, \rho(s)) f(s) \\
&= -\sum_{s=a+1}^{t} H_{\mu-1}(t, \rho(s)) f(s-1) + \sum_{s=a+1}^{t-1} H_{\mu-1}(t-1, \rho(s)) f(s) \\
&= -H_{\mu-1}(t, a) f(a) - \sum_{s=a+2}^{t} H_{\mu-1}(t, s-1) f(s-1) \\
&\quad + \sum_{s=a+2}^{t} H_{\mu-1}(t-1, s-2) f(s-1) \\
&= -H_{\mu-1}(t, a) f(a) \qquad \text{(by Remark 2)}
\end{aligned}$$

for $t \in \mathbb{N}_a$. Thus the proof is complete. □

We now generalize the above lemma.

Theorem 7 ([22, Theorem 2.8]) *Let* $\mu > 0$, $N \in \mathbb{N}_1$, *and* $f : \mathbb{N}_{a-N+1} \to \mathbb{R}$. *Then*

$$(\nabla_a^{-\mu} \nabla^N f)(t) = (\nabla^N \nabla_a^{-\mu} f)(t) - \sum_{k=0}^{N-1} H_{\mu-N+k}(t, a) \nabla^k f(a),$$

for $t \in \mathbb{N}_a$.

Proof We proceed by induction. The base case when $N = 1$ has been shown by the previous lemma. Now assume the result is true for all $N > 1$ and consider the following for $t \in \mathbb{N}_a$.

$$\begin{aligned}
(\nabla_a^{-\mu}\nabla^{N+1} f)(t) &= (\nabla_a^{-\mu}\nabla^N \nabla f)(t) \\
&= (\nabla^N \nabla_a^{-\mu}\nabla f)(t) - \sum_{k=0}^{N-1} H_{\mu-N+k}(t,a)(\nabla^k \nabla f)(a) \\
&= \nabla^N \left[(\nabla \nabla_a^{-\mu} f)(t) - H_{\mu-1}(t,a) f(a)\right] - \sum_{k=0}^{N-1} H_{\mu-N+k}(t,a)\nabla^{k+1} f(a) \\
&= (\nabla^{N+1}\nabla_a^{-\mu} f)(t) - H_{\mu-N-1}(t,a) f(a) - \sum_{k=0}^{N-1} H_{\mu-N+k}(t,a)\nabla^{k+1} f(a) \\
&= (\nabla^{N+1}\nabla_a^{-\mu} f)(t) - H_{\mu-N-1}(t,a) f(a) - \sum_{k=1}^{N} H_{\mu-(N+1)+k}(t,a)\nabla^{k} f(a) \\
&= (\nabla^{N+1}\nabla_a^{-\mu} f)(t) - \sum_{k=0}^{N} H_{\mu-(N+1)+k}(t,a)\nabla^{k} f(a)
\end{aligned}$$

□

Corollary 1 *Assume $\nu > 0$, and $N := \lceil \nu \rceil$ and $f : \mathbb{N}_{a-N+1} \to \mathbb{R}$. Let $\mu = N - \nu$ in Theorem 7. Then we have*

$$\nabla_{a^*}^{\nu} f(t) = \nabla_a^{\nu} f(t) - \sum_{k=0}^{N-1} H_{k-\nu}(t,a)\nabla^k f(a),$$

for $t \in \mathbb{N}_a$.

Corollary 2 *When $\mu = N$ in Theorem 7 we get that*

$$\nabla_a^{-N}\nabla^N f(t) = f(t) - \sum_{k=0}^{N-1} H_k(t,a)\nabla^k f(a).$$

for $t \in \mathbb{N}_a$.

Proof

$$\begin{aligned}
\nabla_a^{-N}\nabla^N f(t) &= \nabla^N \nabla_a^{-N} f(t) - \sum_{k=0}^{N-1} H_k(t,a)\nabla^k f(a) \\
&= f(t) - \sum_{k=0}^{N-1} H_k(t,a)\nabla^k f(a), \qquad \text{(by Lemma 3)}
\end{aligned}$$

for $t \in \mathbb{N}_a$. □

Theorem 8 ([22, Theorem 2.18]) *Assume $\nu > 0$, $N := \lceil \nu \rceil$ and $f : \mathbb{N}_{a+1-N} \to \mathbb{R}$. Then*

$$\nabla_a^{-\nu} \nabla_{a^*}^{\nu} f(t) = f(t) - \sum_{k=0}^{N-1} H_k(t, a) \nabla^k f(a)$$

for $t \in \mathbb{N}_a$.

Proof By the definition of Caputo difference, we have

$$\begin{aligned} \nabla_a^{-\nu} \nabla_{a^*}^{\nu} f(t) &= \nabla_a^{-\nu} \nabla_a^{-(N-\nu)} \nabla^N f(t) \\ &= \nabla_a^{-(\nu+N-\nu)} \nabla^N f(t) \\ &= \nabla_a^{-N} \nabla^N f(t) \\ &= f(t) - \sum_{k=0}^{N-1} H_k(t, a) \nabla^k f(a), \end{aligned}$$

for $t \in \mathbb{N}_a$, where we used Corollary 2 in the last step. □

Theorem 9 ([12, Theorem 3.107]) *Assume $f : \mathbb{N}_{a+1} \to \mathbb{R}$, and $\nu, \mu > 0$. Then*

$$\nabla_a^{-\nu} \nabla_a^{-\mu} f(t) = \nabla_a^{-\nu-\mu} f(t), \qquad t \in \mathbb{N}_a.$$

The following theorem is a generalization of [12, Theorem 3.107].

Theorem 10 *Let $a \in \mathbb{R}$, $b, \nu, \mu \in \mathbb{R}^+$, and $f : \mathbb{N}_{a-b+1} \to \mathbb{R}$. Then*

$$\nabla_a^{-\nu} \nabla_{a-b}^{-\mu} f(t) = \nabla_a^{-\nu-\mu} f(t), \qquad t \in \mathbb{N}_{a-b}.$$

Proof Applying the Laplace transform, we get that

$$\begin{aligned} \mathcal{L}_a\{\nabla_a^{-\nu} \nabla_{a-b}^{-\mu} f\}(s) &= \frac{1}{s^\nu} \mathcal{L}_a\{\nabla_{a-b}^{-\mu} f\}(s) \\ &= \frac{1}{s^\nu} \mathcal{L}_a\{H_{\mu-1}(\cdot, a-b)\}(s) \mathcal{L}_a\{f\}(s) \quad \text{(by Theorem 5)} \\ &= \frac{1}{s^\nu} \Big\{ \sum_{k=1}^{\infty} (1-s)^{k-1} H_{\mu-1}(a-b+k, a-b) \Big\} \mathcal{L}_a\{f\}(s) \\ &= \frac{1}{s^\nu} \Big\{ \sum_{k=1}^{\infty} (1-s)^{k-1} H_{\mu-1}(a+k, a) \Big\} \mathcal{L}_a\{f\}(s) \quad \text{(by Remark 2)} \\ &= \frac{1}{s^\nu} \mathcal{L}_a\{H_{\mu-1}(\cdot, a)\}(s) \mathcal{L}_a\{f\}(s) \\ &= \frac{1}{s^\nu} \mathcal{L}_a\{\nabla_a^{-\mu} f\}(s) \\ &= \mathcal{L}_a\{\nabla_a^{-\nu} \nabla_a^{-\mu} f\}(s) \\ &= \mathcal{L}_a\{\nabla_a^{-\nu-\mu} f\}(s) \quad \text{(by Theorem 9)} \end{aligned}$$

Then by the uniqueness of Laplace transforms, we have

$$\nabla_a^{-\nu}\nabla_{a-b}^{-\mu} f(t) = \nabla_a^{-\nu-\mu} f(t), \qquad t \in \mathbb{N}_{a-b+1}.$$

Moreover, $\nabla_a^{-\nu}\nabla_{a-b}^{-\mu} f(a-b) = 0 = \nabla_a^{-\nu-\mu} f(a-b)$. □

4 Caputo Nabla Fractional Boundary Value Problems

4.1 Introduction

The study of fractional calculus dates back to the time of Leibniz. Many applications of fractional calculus emerged in the past few decades. Tenreiro Machado et al. [20] investigate some engineering applications of fractional calculus. Valério et al. [24] survey the application of fractional calculus to scientific and engineering problems in the past two centuries. Recently, Graef et al. [14] use fractional differential equations to study bike sharing systems. However, only in recent years, fractional difference equations began to be studied. Miller and Ross published a landmark paper on fractional difference calculus [21] in 1988. Recent results can be found in [2, 4–6, 9–13, 16–18, 23].

This chapter is motivated by Eloe et al. [7], St Goar [23] and Erbe and Peterson [8]. Atıcı and Eloe [3] studied a two-point boundary value problem and then the results were generalized by Goodrich [11] and Eloe et al. [7], where fractional boundary value conditions are considered. Eloe et al. [7] studied the Green's functions for a family of delta fractional boundary value problems and St Goar [23] considered a right focal boundary value theorem problem with a Caputo fractional difference. In [8], Erbe and Peterson discussed the existence of positive solutions to a boundary value problem on time scales, using the cone theory that can be found in [15, 19]. In this chapter we consider a Caputo nabla fractional boundary value problem (FBVP) with a fractional boundary condition of the form

$$\begin{cases} -\nabla_{a^*}^{\nu} y(t) = h(t), & t \in \mathbb{N}_{a+1}^{b} \\ y(a-i) = 0, & 1 \le i \le N-1 \\ (\nabla_{a^*}^{\beta} y)(b) = 0, \end{cases} \tag{1}$$

where $h : \mathbb{N}_{a+1} \to \mathbb{R}$, $\nu > 1$, $0 \le \beta \le N-1 < \nu \le N$, $b-a \in \mathbb{Z}$ and $b-a \ge N-1$.

When $1 < \nu \le 2$ and $\beta = 0$, the FBVP (1) becomes a two-point problem that will be discussed in Sect. 2.5. When $1 < \nu \le 2$ and $\beta = 1$, it becomes the right focal problem in [23].

In Sect. 2.2, we derive the Green's function for solving the FBVP (1) by adopting a construction approach which is similar to the method used by Eloe et al. [7], so that

the Green's function, the solution of the FBVP, and the existence and uniqueness of the solution are all included in the following theorem. In Sect. 2.3, we discuss the behavior of the Green's function with respect to the parameter β and prove that the Green's function is nonnegative in several cases. In Sect. 2.4, we prove two comparison theorems. In Sect. 2.5, we use cone theory in a Banach space to study a two-point problem and discuss the existence of positive solutions to its nonlinear case.

4.2 Green's Function

In this section, we are interested in the Green's function for the homogeneous nabla FBVP

$$\begin{cases} -\nabla_{a^*}^{\nu} y(t) = 0, & t \in \mathbb{N}_{a+1}^{b} \\ y(a-i) = 0, & 1 \le i \le N-1 \\ (\nabla_{a^*}^{\beta} y)(b) = 0, \end{cases} \tag{2}$$

where $\nu > 1$, $0 \le \beta \le N-1 < \nu \le N$, $b-a \in \mathbb{Z}$ and $b-a > N-1$.

Theorem 11 *Assume $\nu > 1$, $0 \le \beta \le N-1 < \nu \le N$, $b-a \in \mathbb{Z}$, $b-a > N-1$ and $h : \mathbb{N}_{a+1} \to \mathbb{R}$. The Green's function $G(t,s) : \mathbb{N}_{a-N+1}^{b} \times \mathbb{N}_{a+1}^{b} \to \mathbb{R}$ for the Caputo nabla FBVP (2) is given by*

$$G(t,s) = \begin{cases} u(t,s), & a+1 \le t \le \rho(s) \le b \\ v(t,s), & a+1 \le \rho(s) \le t \le b \end{cases} \tag{3}$$

where

$$u(t,s) := \frac{H_{\nu-\beta-1}(b,\rho(s))H_{N-1}(t,\rho(a))}{H_{N-\beta-1}(b,\rho(a))}$$

and

$$v(t,s) := \frac{H_{\nu-\beta-1}(b,\rho(s))H_{N-1}(t,\rho(a))}{H_{N-\beta-1}(b,\rho(a))} - H_{\nu-1}(t,\rho(s)).$$

Furthermore, the unique solution of the Caputo nabla FBVP (1) is given by

$$y(t) = \int_a^b G(t,s)h(s)\nabla s,$$

for $t \in \mathbb{N}_a^b$.

Proof Applying the nabla sum operator $\nabla_a^{-\nu}$ to both sides of the Caputo nabla fractional equation in (1) we get that

$$-\nabla_a^{-\nu}\nabla_{a^*}^{\nu} y(t) = \nabla_a^{-\nu} h(t)$$

for $t \in N_a$.

By Theorem 8, the general solution to $\nabla_{a^*}^{\nu} y(t) = h(t)$ is given by

$$y(t) = \sum_{k=0}^{N-1} H_k(t, a)\nabla^k y(a) - \nabla_a^{-\nu} h(t). \tag{4}$$

for $t \in N_a$.

Then we use the binomial expression in Lemma 1 to expand $\nabla^k y(a)$ and apply the boundary conditions on the left in the FBVP (1). Then we use Lemma 2 to obtain the following

$$\begin{aligned} y(t) &= \sum_{k=0}^{N-1} H_k(t, a) \sum_{j=0}^{k} (-1)^j \binom{k}{j} y(a - j) - \nabla_a^{-\nu} h(t) \\ &= \Big(\sum_{k=0}^{N-1} H_k(t, a) \Big) y(a) - \nabla_a^{-\nu} h(t) \\ &= H_{N-1}(t, \rho(a)) y(a) - \nabla_a^{-\nu} h(t) \end{aligned} \tag{5}$$

for $t \in \mathbb{N}_a$.

Now we use the boundary condition on the right in (1) to solve for $y(a)$. We apply the Caputo nabla fractional operator $\nabla_{a^*}^{\beta}$ to both sides of (5) and then evaluate the result at $t = b$. So we have

$$(\nabla_{a^*}^{\beta} y)(b) = 0 = y(a)(\nabla_{a^*}^{\beta} H_{N-1}(\cdot, \rho(a)))(b) - (\nabla_{a^*}^{\beta} \nabla_a^{-\nu} h)(b). \tag{6}$$

Next we use the power rules from Theorem 1 (ii) and Theorem 3 (i) to find that

$$\begin{aligned} (\nabla_{a^*}^{\beta} H_{N-1}(\cdot, \rho(a)))(b) &= (\nabla_a^{-(M-\beta)} \nabla^M H_{N-1}(\cdot, \rho(a)))(b) \\ &= (\nabla_a^{-(M-\beta)} H_{N-1-M}(\cdot, \rho(a)))(b) \\ &= H_{N-\beta-1}(b, \rho(a)), \end{aligned} \tag{7}$$

where $M = \lceil \beta \rceil$.

Note that in the above proof, we used $M \leq N - 1$ so that $H_{N-M-1}(b, a) \neq 0$. Otherwise, we would not be able to solve for $y(a)$.

Next consider the term $(\nabla_{a^*}^{\beta}\nabla_a^{-\nu}h)(b)$ in (6) and find that

$$\begin{aligned}(\nabla_{a^*}^{\beta}\nabla_a^{-\nu}h)(b) &= (\nabla_a^{-(M-\beta)}\nabla^M\nabla_a^{-\nu}h)(b)\\ &= (\nabla_a^{-(M-\beta)}\nabla_a^{M-\nu}h)(b) \qquad \text{(by Lemma 3)}\\ &= (\nabla_a^{\beta-\nu}h)(b), \qquad \text{(by Theorem 9)} \end{aligned} \tag{8}$$

where $M = \lceil\beta\rceil$.

We substitute (7) and (8) into (6) to obtain

$$y(a)H_{N-\beta-1}(b,\rho(a)) - (\nabla_a^{\beta-\nu})h(b) = 0.$$

Solving for $y(a)$, we get

$$y(a) = \frac{(\nabla_a^{\beta-\nu}h)(b)}{H_{N-\beta-1}(b,\rho(a))}.$$

Therefore, the solution of the boundary value problem (1) is given by

$$\begin{aligned} y(t) &= \frac{(\nabla_a^{\beta-\nu}h)(b)}{H_{N-\beta-1}(b,\rho(a))}H_{N-1}(t,\rho(a)) - \nabla_a^{-\nu}h(t)\\ &= \int_a^b \frac{H_{\nu-\beta-1}(b,\rho(s))H_{N-1}(t,\rho(a))}{H_{N-\beta-1}(b,\rho(a))}h(s)\nabla s - \int_a^t H_{\nu-1}(t,\rho(s))h(s)\nabla s\\ &= \int_a^{t+1} \frac{H_{\nu-\beta-1}(b,\rho(s))H_{N-1}(t,\rho(a))}{H_{N-\beta-1}(b,\rho(a))}h(s)\nabla s\\ &\quad + \int_{t+1}^b \frac{H_{\nu-\beta-1}(b,\rho(s))H_{N-1}(t,\rho(a))}{H_{N-\beta-1}(b,\rho(a))}h(s)\nabla s\\ &\quad - \int_a^t H_{\nu-1}(t,\rho(s))h(s)\nabla s - H_{\nu-1}(t,\rho(t+1))h(t+1)\\ &= \int_a^{t+1} \frac{H_{\nu-\beta-1}(b,\rho(s))H_{N-1}(t,\rho(a))}{H_{N-\beta-1}(b,\rho(a))}h(s)\nabla s - \int_a^{t+1} H_{\nu-1}(t,\rho(s))h(s)\nabla s\\ &\quad + \int_{t+1}^b \frac{H_{\nu-\beta-1}(b,\rho(s))H_{N-1}(t,\rho(a))}{H_{N-\beta-1}(b,\rho(a))}h(s)\nabla s\\ &= \int_a^b G(t,s)h(s)\nabla s, \quad t \in \mathbb{N}_{a-N+1}^b, \end{aligned}$$

where we used that $H_{\nu-1}(t,\rho(t+1)) = H_{\nu-1}(t,t) = 0$ by Theorem 1 (i) and $G(t,s)$ is the Green's function defined in (3). □

Remark 5 In [23], St Goar studied a Caputo nabla FBVP with an integer order boundary condition of the form

$$\begin{cases} \nabla_{a^*}^{\nu} y(t) = h(t), & t \in \mathbb{N}_{a+1}^{b} \\ \nabla^k y(a-1) = 0, & 0 \le k \le N-2 \\ (\nabla^{N-1} y)(b) = 0 \end{cases} \tag{9}$$

where $\nu > 1$, $N = \lceil \nu \rceil$, $b - a \in \mathbb{Z}$ and $b - a \ge N - 1$.

The Green's function for the corresponding homogeneous FBVP

$$\begin{cases} \nabla_{a^*}^{\nu} y(t) = 0, & t \in \mathbb{N}_{a+1}^{b} \\ \nabla^k y(a-1) = 0, & 0 \le k \le N-2 \\ (\nabla^{N-1} y)(b) = 0 \end{cases} \tag{10}$$

is given by

$$G(t,s) = \begin{cases} u(t,s), & t \le \rho(s) \\ v(t,s), & t \ge \rho(s) \end{cases} \tag{11}$$

where

$$u(t,s) = -H_{N-1}(t, a-1) H_{\nu-N}(b, \rho(s))$$

and

$$v(t,s) = -H_{N-1}(t, a-1) H_{\nu-N}(b, \rho(s)) + H_{\nu-1}(t, \rho(s)).$$

In the Green's function (11), $u(t,s)$ is defined to be the unique solution of the BVP

$$\begin{cases} \nabla_{a^*}^{\nu} u(t,s) = 0, & t \in \mathbb{N}_{a+1}^{b} \\ \nabla^k u(a-1,s) = 0, & 0 \le k \le N-2 \\ (\nabla^{N-1} u)(b,s) = -\nabla^{N-1} x(b,s), \end{cases} \tag{12}$$

where $x(t,s) = H_{\nu-1}(t, \rho(s))$.

We note that the Green's function we have found in Theorem 11 reduces to (11) when $\beta = N - 1$ and satisfy the following BVP that is similar to (12):

$$\begin{cases} \nabla_{a^*}^{\nu} u(t,s) = 0, & t \in \mathbb{N}_{a+1}^{b} \\ u(a-i,s) = 0, & 1 \le i \le N-1 \\ (\nabla_{a^*}^{\beta} u)(b,s) = -(\nabla_{a^*}^{\beta}) x(b,s), \end{cases} \tag{13}$$

where $x(t,s) = H_{\nu-1}(t, \rho(s))$.

St Goar also presented a generalized FBVP where the boundary condition on the right is $(\nabla^i y)(b) = 0$, $i \in \mathbb{N}_0^{N-1}$. Both of the two FBVPs are special cases of our FBVP (1).

5 Properties of the Green's Function

In this section, we examine the behavior of the Green's functions $G(t, s)$ in Theorem 11 for $0 \leq \beta \leq 1$. Note that $G(\beta; t, s) := G(t, s)$.

Theorem 12 *Let $1 < \nu < 2$, and $2 \leq b - a \leq \frac{1}{2-\nu}$. Then the Green's function $G(0; t, s)$ has the following properties:*

(i) $G(0; a-1, s) = G(0; b, s) = 0$.
(ii) $G(0; t, s) \geq 0$ *for* $(t, s) \in \mathbb{N}_{a-1}^{b} \times \mathbb{N}_{a+1}^{b}$,
(iii) $\max_{t \in \mathbb{N}_{a-1}^{b}} G(0; t, s) = G(0; \rho(s), s)$ *for* $(t, s) \in \mathbb{N}_{a-1}^{b} \times \mathbb{N}_{a+1}^{b}$, *and*
(iv) $G(0; t, s) \geq k \cdot G(0; \rho(s), s)$ *for a constant* $k \in (0, 1)$ *and* $(t, s) \in \mathbb{N}_{a}^{b-1} \times \mathbb{N}_{a+1}^{b}$.

Proof (i) Note that $1 < \nu < 2$ implies $N = 2$. When $\beta = 0$, the boundary conditions in the FBVP (1) becomes $y(a-1) = 0$ and $y(b) = 0$.

By direct computation, we have

$$G(0; a-1, s) = \frac{H_{\nu-1}(b, \rho(s)) H_1(a-1, \rho(a))}{H_1(b, \rho(a))} = 0,$$

and

$$G(0; b, s) = \frac{H_{\nu-1}(b, \rho(s)) H_1(b, \rho(a))}{H_1(b, \rho(a))} - H_{\nu-1}(b, \rho(s)) = 0.$$

Hence the Green's function satisfies the boundary conditions.

(ii) For $t \leq \rho(s)$ we have

$$\begin{aligned} G(0; t, s) &= u(t, s) \\ &= \frac{H_{\nu-1}(b, \rho(s)) H_1(t, \rho(a))}{H_1(b, \rho(a))} \\ &= \frac{[\Gamma(b - \rho(s) + \nu - 1)](t - \rho(a))}{[\Gamma(\nu)\Gamma(b - \rho(s))](b - \rho(a))} \\ &\geq 0. \end{aligned}$$

For $t \geq \rho(s)$, we have

$$G(0; t, s) = v(t, s) = \frac{H_{\nu-1}(b, \rho(s)) H_1(t, \rho(a))}{H_1(b, \rho(a))} - H_{\nu-1}(t, \rho(s)).$$

Then the nabla difference of $v(t, s)$ with respect to t is given by

$$\nabla_t v(t, s) = \frac{H_{\nu-1}(b, \rho(s))}{H_1(b, \rho(a))} - H_{\nu-2}(t, \rho(s)). \tag{14}$$

We claim that $H_{\nu-2}(t,\rho(s))$ is decreasing in t for each fixed $s \in \mathbb{N}_{a+1}^{b}$.
For $t = s$, $H_{\nu-2}(t,\rho(s)) = H_{\nu-2}(t,\rho(t)) = 1$.
For $t \geq s+1$, we have

$$\begin{aligned}\nabla_t H_{\nu-2}(t,\rho(s)) &= H_{\nu-3}(t,\rho(s)) \\ &= \frac{\Gamma(t-s+\nu-2)}{\Gamma(\nu-2)\Gamma(t-s+1)} \\ &< 0,\end{aligned}$$

since $t-s+\nu-2 > 0$, $t-s+1 > 0$, and $-1 < \nu-2 < 0$.

Hence we substitute $H_{\nu-2}(t,\rho(s))$ by $H_{\nu-2}(b,\rho(s))$ in (14) and get the inequality

$$\begin{aligned}\nabla_t v(t,s) &\leq \frac{H_{\nu-1}(b,\rho(s))}{H_1(b,\rho(a))} - H_{\nu-2}(b,\rho(s)) \\ &= \frac{\Gamma(b-\rho(s)+\nu-1)}{(b-\rho(a))\Gamma(\nu)\Gamma(b-\rho(s))} - \frac{\Gamma(b-\rho(s)+\nu-2)}{\Gamma(\nu-1)\Gamma(b-\rho(s))} \\ &= \left(\frac{b-\rho(s)+\nu-2}{(\nu-1)(b-\rho(a))} - 1\right)\frac{\Gamma(b-\rho(s)+\nu-2)}{\Gamma(\nu-1)\Gamma(b-\rho(s))}.\end{aligned}$$

We see that

$$\frac{\Gamma(b-\rho(s)+\nu-2)}{\Gamma(\nu-1)\Gamma(b-\rho(s))} > 0$$

and

$$\begin{aligned}\frac{b-\rho(s)+\nu-2}{(\nu-1)(b-\rho(a))} - 1 &\leq \frac{b-a+\nu-2}{(\nu-1)(b-a+1)} - 1 \\ &= \frac{b-a+\nu-2-(\nu-1)(b-a+1)}{(\nu-1)(b-a+1)} \\ &= \frac{b-a+\nu-2-(\nu-1)-(\nu-1)(b-a)}{(\nu-1)(b-a+1)} \\ &= \frac{(b-a)(2-\nu)-1}{(\nu-1)(b-a+1)} \\ &\leq \frac{\frac{1}{2-\nu}(2-\nu)-1}{(\nu-1)(b-a+1)} \\ &= 0.\end{aligned}$$

Hence $\nabla_t v(t,s) \leq 0$. So $G(0;t,s)$ is decreasing for $t \geq \rho(s)$.
We also see that $v(b,s) = 0$ from the proof of (i).
Hence $G(0;t,s) \geq 0$.

(iii) We note that

$$\begin{aligned}\nabla_t u(t,s) &= \nabla_t \left[\frac{H_{\nu-1}(b,\rho(s))H_1(t,\rho(a))}{H_1(b,\rho(a))} \right] \\ &= \frac{H_{\nu-1}(b,\rho(s))}{H_1(b,\rho(a))} > 0.\end{aligned}$$

for $t \le \rho(s)$ and each fixed $s \in \mathbb{N}_{a+1}^{b}$. We have shown $\nabla_t v(t,s) \le 0$ for $t \ge \rho(s)$ and each fixed $s \in \mathbb{N}_{a+1}^{b}$. Hence the maximum of the $G(0;t,s)$ occurs at $t = \rho(s)$.

(iv) For $a \le t \le \rho(s)$, since $G(0;t,s)$ increases in t, we have

$$\begin{aligned}\frac{G(0;t,s)}{G(0;\rho(s),s)} &\ge \frac{G(0;a,s)}{G(0;\rho(s),s)} \\ &= \frac{\dfrac{H_{\nu-1}(b,\rho(s))(a-a+1)}{b-a+1}}{\dfrac{H_{\nu-1}(b,\rho(s))(s-a)}{b-a+1}} \\ &= \frac{1}{s-a} \\ &\ge \frac{1}{b-a}.\end{aligned}$$

For $\rho(s) \le t \le b-1$, since $G(0;t,s)$ decreases in t, we have

$$\begin{aligned}\frac{G(0;t,s)}{G(0;\rho(s),s)} &\ge \frac{G(0;b-1,s)}{G(0;\rho(s),s)} \\ &= \frac{\dfrac{H_{\nu-1}(b,\rho(s))(b-1-a+1)}{b-a+1} - H_{\nu-1}(b-1,\rho(s))}{\dfrac{H_{\nu-1}(b,\rho(s))(s-a)}{b-a+1} - H_{\nu-1}(\rho(s),\rho(s))} \\ &= \frac{b-a}{s-a} - \frac{H_{\nu-1}(b-1,\rho(s))(b-a+1)}{H_{\nu-1}(b,\rho(s))(s-a)} \\ &= \frac{1}{s-a}\Big(b-a-\frac{(b-s)^{\overline{\nu-1}}(b-a+1)}{(b-s+1)^{\overline{\nu-1}}}\Big) \\ &= \frac{1}{s-a}\Big(b-a-\frac{\Gamma(b-s+\nu-1)\Gamma(b-s+1)}{\Gamma(b-s)\Gamma(b-s+\nu)}(b-a+1)\Big) \\ &= \frac{1}{s-a}\Big(b-a-\frac{(b-s)(b-a+1)}{b-s+\nu-1}\Big) \\ &= \frac{1}{s-a}\Big(\frac{(b-a)(b-s+\nu-1)-(b-s)(b-a+1)}{b-s+\nu-1}\Big) \\ &= \frac{1}{s-a}\Big(\frac{(b-a)(\nu-1)+(b-s)}{b-s+\nu-1}\Big)\end{aligned}$$

$$\begin{aligned}
&\geq \frac{1}{b-a}\Big(\frac{(b-a)(\nu-1)+(b-s)}{b-s+\nu-1}\Big)\\
&= \frac{\nu-1}{b-s+\nu-1}+\frac{b-s}{(b-a)(b-s+\nu-1)}\\
&\geq \frac{\nu-1}{b-a+\nu-1}.
\end{aligned}$$

Let $k := \min\left\{\frac{1}{b-a}, \frac{\nu-1}{b-a+\nu-1}\right\}$. It is clear that $0 < k < 1$. □

Theorem 13 *Let $1 < \nu < 2$ such that $2 \leq b-a \leq \frac{1}{2-\nu}$, $N = \lceil \nu \rceil$, $0 \leq \beta \leq 1$, and $G(\beta; t, s) := G(t, s)$ be the Green's function defined in Theorem 11. Then $G(\beta; t, s)$ is increasing in β for all fixed $(t, s) \in \mathbb{N}_{a-1}^{b} \times \mathbb{N}_{a+2}^{b}$.*

Proof The behavior of $G(\beta; t, s)$ in β is determined by

$$\frac{H_{\nu-\beta-1}(b, \rho(s))}{H_{N-\beta-1}(b, \rho(a))} = \frac{\Gamma(b-s+\nu-\beta)}{\Gamma(b-s+1)\Gamma(\nu-\beta)}\frac{\Gamma(b-a+1)\Gamma(N-\beta)}{\Gamma(b-a+N-\beta)}.$$

Since $\Gamma(b-a+1)$ and $\Gamma(b-s+1)$ are both positive, the behavior of $G(\beta; t, s)$ in β as a continuous variable is further determined by

$$f(\beta) := \frac{\Gamma(b-s+\nu-\beta)\Gamma(N-\beta)}{\Gamma(\nu-\beta)\Gamma(b-a+N-\beta)} = \frac{(\nu-\beta)^{\overline{b-s}}}{(N-\beta)^{\overline{b-a}}}.$$

Since $1 < \nu < 2$ implies $N = 2$, we substitute N in the above definition to get

$$f(\beta) = \frac{\Gamma(b-s+\nu-\beta)\Gamma(2-\beta)}{\Gamma(\nu-\beta)\Gamma(b-a+2-\beta)} = \frac{(\nu-\beta)^{\overline{b-s}}}{(2-\beta)^{\overline{b-a}}}. \tag{15}$$

Since $b-s \in \mathbb{Z}$, $b-a \in \mathbb{Z}$ and $b-a \geq b-s$, we use Definition 3 to expand the rising functions (15) and then obtain

$$\begin{aligned}
f(\beta) &= \frac{(\nu-\beta)(\nu-\beta+1)\cdots(\nu-\beta+b-s-1)}{(2-\beta)(2-\beta+1)\cdots(2-\beta+b-s-1)\cdots(2-\beta+b-a-1)}\\
&= \prod_{i=0}^{b-s-1}\frac{\nu-\beta+i}{2-\beta+i}\prod_{j=b-s}^{b-a-1}\frac{1}{2-\beta+j}.
\end{aligned}$$

Then we use the generalized product rule for n functions of the form

$$\frac{d}{dx}\prod_{i=1}^{n} f_i(x) = \prod_{i=1}^{n} f_i(x)\sum_{i=1}^{n}\frac{f_i'(x)}{f_i(x)}$$

to find the derivative of $f(\beta)$

$$\begin{aligned} f'(\beta) &= \left(\prod_{i=0}^{b-s-1} \frac{\nu-\beta+i}{2-\beta+i}\right)\left(\sum_{i=0}^{b-s-1} \frac{\nu-2}{(2-\beta+i)(\nu-\beta+i)}\right)\left(\prod_{j=b-s}^{b-a-1} \frac{1}{2-\beta+j}\right) \\ &\quad + \left(\prod_{i=0}^{b-s-1} \frac{\nu-\beta+i}{2-\beta+i}\right)\left(\prod_{j=b-s}^{b-a-1} \frac{1}{2-\beta+j}\right)\left(\sum_{j=b-s}^{b-a-1} \frac{1}{2-\beta+j}\right) \\ &= f(\beta) \sum_{i=0}^{b-s-1} \frac{\nu-2}{(2-\beta+i)(\nu-\beta+i)} + f(\beta) \sum_{j=b-s}^{b-a-1} \frac{1}{2-\beta+j} \\ &= f(\beta)\left(\sum_{i=0}^{b-a-1} \frac{1}{2-\beta+i} - \sum_{j=0}^{b-s-1} \frac{1}{\nu-\beta+j}\right). \end{aligned}$$

Since $f(\beta) > 0$, the sign of $f'(\beta)$ depends on

$$\phi(\beta) := \sum_{i=0}^{b-a-1} \frac{1}{2-\beta+i} - \sum_{j=0}^{b-s-1} \frac{1}{\nu-\beta+j}.$$

It can be shown that $\phi(\beta)$ increases in ν. Since $2 \le b-a \le \dfrac{1}{2-\nu}$, we have $\nu \ge \dfrac{3}{2}$. When $\nu = \dfrac{3}{2}$, $b-a=2$ and there are two values for s: $a+1$ or $a+2$. $a+1$ is not in the domain of s and $b-s-1=-1$ when $s=a+2$. So we have

$$\phi(\beta) = \frac{1}{2-\beta} + \frac{1}{3-\beta} > 0 \quad \text{for } \beta \in [0,1].$$

Hence $\phi(\beta) > 0$ for all $\nu \in [\frac{3}{2}, 2)$ since $\phi(\beta)$ increases in ν. It follows that $f'(\beta) > 0$ and the proof is complete. □

Theorem 14 *Let* $1 < \nu < 2$, $2 \le b-a \le \dfrac{1}{2-\nu}$, $0 \le \beta \le 1$, *and* $(t,s) \in \mathbb{N}_{a-1}^{b} \times \mathbb{N}_{a+2}^{b}$. *The Green's function* $G(\beta; t, s)$ *has the following properties:*

(i) $G(\beta; t, s) \ge 0$,
(ii) $\max_{t \in \mathbb{N}_{a-1}^{b}} G(\beta; t, s) = G(\beta; \rho(s), s)$.

Proof

(i) By Theorem 12, $G(0; t, s) \ge 0$ for $1 < \nu < 2$, $2 \le b-a \le \dfrac{1}{2-\nu}$, and $(t,s) \in \mathbb{N}_{a-1}^{b} \times \mathbb{N}_{a+1}^{b}$. Then the result is valid for $(t,s) \in \mathbb{N}_{a-1}^{b} \times \mathbb{N}_{a+2}^{b}$. Therefore by Theorem 13, we have $G(\beta; t, s) \ge 0$ for $0 \le \beta \le 1$.

(ii) $G(\beta; t, s)$ is increasing in t for $t \leq \rho(s)$ since

$$\nabla_t u(t, s) = \frac{H_{\nu-\beta-1}(b, \rho(s))}{H_{1-\beta}(b, \rho(a))} > 0.$$

Now we show $G(\beta; t, s)$ is decreasing in t for $t \geq \rho(s)$. Consider

$$\nabla_t v(t, s) = \frac{H_{\nu-\beta-1}(b, \rho(s))}{H_{1-\beta}(b, \rho(a))} - H_{\nu-2}(t, \rho(s)).$$

By the proof of Theorem 13, we have $\dfrac{H_{\nu-\beta-1}(b, \rho(s))}{H_{1-\beta}(b, \rho(a))}$ increases in β for $s \in \mathbb{N}_{a+2}^b$. This implies $\nabla_t v(t, s)$ increases in β for each fixed $s \in \mathbb{N}_{a+2}^b$. So when $\beta = 1$, we have

$$\begin{aligned} \nabla_t v(t, s)\Big|_{\beta=1} &= H_{\nu-2}(b, \rho(s)) - H_{\nu-2}(t, \rho(s)) \\ &\leq 0, \end{aligned}$$

since $H_{\nu-2}(t, \rho(s))$ decreases in t by the proof of Theorem 12. So $\nabla_t v(t, s) \leq 0$ for $0 \leq \beta \leq 1$. Therefore $\max_{t \in \mathbb{N}_{a-1}^b} G(\beta; t, s) = G(\beta; \rho(s), s)$. □

We have been considering $\nu < 2$ in the previous theorems. We treat $\nu = 2$ as a special case in the following theorem.

Theorem 15 *Let $\nu = 2$, $0 \leq \beta \leq 1$ and $(t, s) \in \mathbb{N}_{a-1}^b \times \mathbb{N}_{a+1}^b$. Then the Green's function $G(\beta; t, s)$ increases in β, $G(\beta; t, s) \geq 0$, and $\max_{t \in \mathbb{N}_{a-1}^b} G(\beta; t, s) = G(\beta; \rho(s), s)$.*

Proof Let $\nu = 2$. The Green's Function becomes

$$G(\beta; t, s) = \begin{cases} \dfrac{H_{1-\beta}(b, \rho(s)) H_1(t, \rho(a))}{H_{1-\beta}(b, \rho(a))}, & t \leq \rho(s) \\ \dfrac{H_{1-\beta}(b, \rho(s)) H_1(t, \rho(a))}{H_{1-\beta}(b, \rho(a))} - H_{\nu-1}(t, \rho(s)), & t \geq \rho(s) \end{cases}$$

and its behavior in β depends on

$$\frac{H_{1-\beta}(b, \rho(s)) H_1(t, \rho(a))}{H_{1-\beta}(b, \rho(a))} = \frac{\Gamma(b - s + 2 - \beta)\Gamma(b - a + 1)}{\Gamma(b - a + 2 - \beta)\Gamma(b - s + 1)}(t - a + 1).$$

We note that $\Gamma(b - s + 2 - \beta)$, $\Gamma(b - a + 2 - \beta)$, and $t - a + 1$ are positive. Consider

$$\phi(\beta) := \frac{\Gamma(b - s + 2 - \beta)}{\Gamma(b - a + 2 - \beta)}$$

$$= \prod_{i=b-s}^{b-a-1} \frac{1}{2-\beta+i}.$$

Since $0 \le \beta \le 1$ and $\frac{1}{2-\beta+i}$ is positive and increasing with repect to i for each fixed β we have that $G(\beta; t, s)$ is increasing in β for $\nu = 2$.

For $t \le \rho(s)$, we have

$$\begin{aligned} G(\beta; t, s) &= u(t, s) \\ &= \frac{H_{1-\beta}(b, \rho(s)) H_1(t, \rho(a))}{H_{1-\beta}(b, \rho(a))} \\ &= \frac{(b-\rho(s))^{\overline{1-\beta}}(t-\rho(a))}{(b-\rho(a))^{\overline{1-\beta}}} \\ &= \frac{\Gamma(b-s+2-\beta)\Gamma(b-a+1)}{\Gamma(b-a+2-\beta)\Gamma(b-s+1)}(t-a+1) \\ &\ge 0, \end{aligned}$$

and for $t \ge \rho(s)$, we have

$$G(\beta; t, s) = v(t, s) = \frac{H_{1-\beta}(b, \rho(s)) H_1(t, \rho(a))}{H_{1-\beta}(b, \rho(a))} - H_1(t, \rho(s)).$$

The nabla difference of $v(t, s)$ in t for each fixed s is

$$\begin{aligned} \nabla_t v(t, s) &= \frac{H_{1-\beta}(b, \rho(s)) H_0(t, \rho(a))}{H_{1-\beta}(b, \rho(a))} - H_0(t, \rho(s)) \\ &= \frac{H_{1-\beta}(b, \rho(s))}{H_{1-\beta}(b, \rho(a))} - 1 \\ &= \frac{\Gamma(b-s+2-\beta)\Gamma(b-a+1)}{\Gamma(b-a+2-\beta)\Gamma(b-s+1)} - 1 \\ &= \prod_{i=0}^{s-a-1} \frac{b-a-i}{b-a+1-\beta-i} - 1 \\ &< 0 \end{aligned}$$

for $t \le \rho(s)$ and

$$\begin{aligned} v(b, s) &= \frac{H_{1-\beta}(b, \rho(s)) H_1(b, \rho(a))}{H_{1-\beta}(b, \rho(a))} - H_1(b, \rho(s)) \\ &= \Big(\prod_{i=0}^{s-a-1} \frac{b-a-i}{b-a+1-\beta-i} \Big)(b-a+1) - (b-s+1). \end{aligned}$$

We already have that $v(b, s)$ increases in β. When $\beta = 0$, we have

$$v(b, s)\Big|_{\beta=0} = \frac{H_1(b, \rho(s))H_1(b, \rho(a))}{H_1(b, \rho(a))} - H_1(b, \rho(s)) = 0.$$

It follows that $v(b, s) \geq 0$ for all $\beta \in [0, 1]$. Hence $G(\beta; t, s) \geq 0$.

To see that $\max_{t \in \mathbb{N}_{a-1}^b} G(\beta; t, s) = G(\beta; \rho(s), s)$, we note that

$$\begin{aligned} \nabla_t u(t, s) &= \nabla_t \left[\frac{H_{1-\beta}(b, \rho(s))H_1(t, \rho(a))}{H_{1-\beta}(b, \rho(a))} \right] \\ &= \frac{H_{1-\beta}(b, \rho(s))}{H_{1-\beta}(b, \rho(a))} > 0. \end{aligned}$$

for $t \leq \rho(s)$ and each fixed $s \in \mathbb{N}_{a+1}^b$. Therefore $\max_{t \in \mathbb{N}_{a-1}^b} G(\beta; t, s) = G(\beta; \rho(s), s)$. □

Remark 6 There is no upper limit on $b - a$ in Theorem 15. We can also see this from Theorem 13, where $b - a \to \infty$ as $\nu \to 2$. The Fig. 1 shows the graphs of the Green's function for $\nu = 2$, $b = 10$, $a = 0$, $s = 4$ and $\beta = 0, 0.6, 1$.

Remark 7 The condition $b - a \leq \frac{1}{2-\nu}$ in Theorem 12 is also seen in [23, Theorem 3.11], where it was used to obtain a positive lower bound for $v(t, s)$. In the proof of Theorem 12 this condition is needed for a decreasing $v(t, s)$ in t for fixed s.

Remark 8 Note that the condition $2 \leq b - a \leq \frac{1}{2-\nu}$ implies $\nu \geq 3/2$. If $1 < \nu < 3/2$, we need to find new conditions for the relationship between ν and $b - a$. Otherwise, the Green's function could be negative, as shown in Fig. 2.

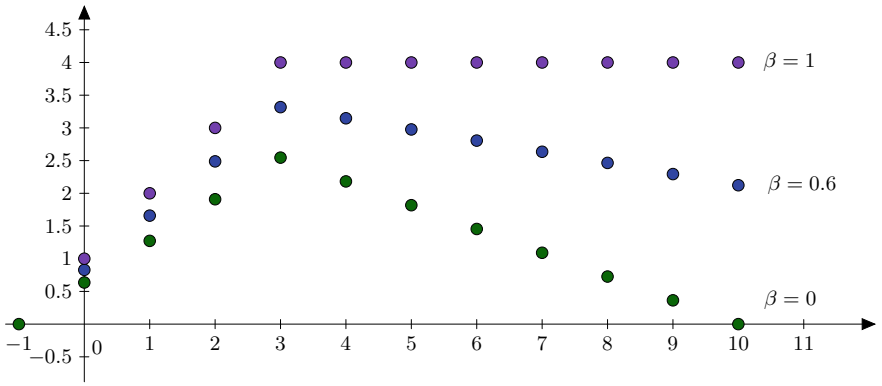

Fig. 1 The Green's function for $\nu = 2$ with $\beta = 0, 0.6, 1$

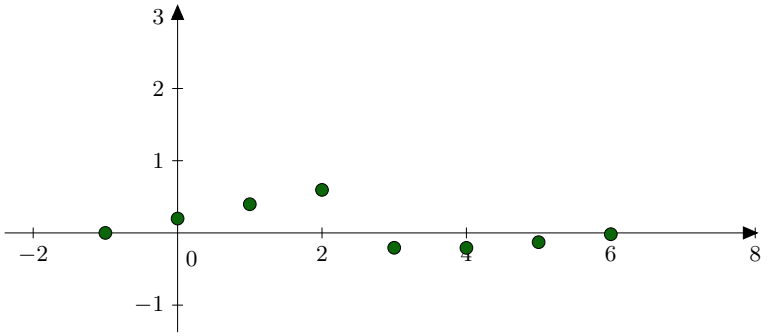

Fig. 2 The Green's function for $\nu = 1.2$, $\beta = 0.2$ $b = 6$, $a = 0$ and $s = 3$

5.1 Comparison Theorems

In this section, we give two comparison theorems as a direct result of the nonnegative property of the Green's function for $1 \leq \nu < 2$,$0 \leq \beta \leq 1$, $2 \leq b - a \leq \frac{1}{2-\nu}$.

Theorem 16 *Assume* $1 \leq \nu < 2$, $0 \leq \beta \leq 1$, $2 \leq b - a \leq \frac{1}{2-\nu}$, *and* $u, v : \mathbb{N}_{a-1}^{b} \to \mathbb{R}$ *such that* $u(t)$ *and* $v(t)$ *satisfy*

$$\begin{aligned} \nabla_{a^*}^{\nu} u(t) &\geq \nabla_{a^*}^{\nu} v(t), \quad t \in \mathbb{N}_{a+1}^{b} \\ u(a-1) &= v(a-1) \\ (\nabla_{a^*}^{\beta} u)(b) &= (\nabla_{a^*}^{\beta} v)(b) \end{aligned}$$

Then $u(t) \leq v(t)$ *for* $t \in \mathbb{N}_{a-1}^{b}$.

Proof Let

$$w(t) = v(t) - u(t), \quad t \in \mathbb{N}_{a-1}^{b},$$

then

$$h(t) := \nabla_{a^*}^{\nu} w(t) = \nabla_{a^*}^{\nu} v(t) - \nabla_{a^*}^{\nu} u(t) \leq 0, \quad t \in \mathbb{N}_{a+1}^{b}$$

by hypothesis. Note that $w(t)$ solves the Caputo FBVP

$$-\nabla_{a^*}^{\nu} w(t) = -h(t), \quad t \in \mathbb{N}_{a+1}^{b} \tag{16}$$

$$w(a-1) = 0, \tag{17}$$

$$(\nabla_{a^*}^{\beta} w)(b) = 0. \tag{18}$$

Hence by Theorem 11

$$w(t) = \int_a^b G(\beta; t, s)[-h(s)]\nabla s \tag{19}$$

$$= \sum_{s=a+1}^{b} G(\beta; t, s)[-h(s)], \quad t \in \mathbb{N}_{a-1}^{b}. \tag{20}$$

Since $h(t) \leq 0$ for $t \in \mathbb{N}_{a+1}^{b}$ and by Theorem 14, $G(\beta; t, s) \geq 0$ for $t \in \mathbb{N}_{a-1}^{b}$ and $s \in \mathbb{N}_{a+1}^{b}$, we have that $w(t) \geq 0$ for $t \in \mathbb{N}_{a}^{b}$. This implies that $v(t) \geq u(t)$ for $t \in \mathbb{N}_{a}^{b}$. But by the hypothesis $u(a-1) = v(a-1)$. Hence

$$v(t) \geq u(t), \quad t \in \mathbb{N}_{a-1}^{b}. \qquad \square$$

Next, we give the solution to the FBVP (1) with nonhomogeneous boundary conditions for $1 < \nu < 2$.

Theorem 17 *Assume $1 < \nu < 2$, $0 \leq \beta \leq 1$, and $h : \mathbb{N}_{a+1}^{b} \to \mathbb{R}$. The solution of the FBVP*

$$\begin{cases} -\nabla_{a^*}^{\nu} y(t) = h(t), & t \in \mathbb{N}_{a+1}^{b} \\ y(a-1) = A, \\ (\nabla_{a^*}^{\beta} y)(b) = B. \end{cases}$$

where A and B are constants, is given by

$$y(t) = z(t) + \int_a^b G(t, s)h(s)\nabla s \quad t \in \mathbb{N}_{a-1}^{b},$$

where $G(t, s)$ is the Green's function defined in Theorem 11 and $z(t)$ is the solution of the FBVP

$$\begin{cases} -\nabla_{a^*}^{\nu} z(t) = 0, & t \in \mathbb{N}_{a+1}^{b} \\ z(a-1) = A, \\ (\nabla_{a^*}^{\beta} z)(b) = B. \end{cases}$$

Proof This theorem is an immediate corollary of Theorem 11 for $1 < \nu < 2$. $\square$

Theorem 18 *Assume $1 < \nu < 2$, $0 \leq \beta \leq 1$ and $z(t)$ solves the FBVP*

$$\begin{cases} -\nabla_{a^*}^{\nu} z(t) = 0, & t \in \mathbb{N}_{a+1}^{b} \\ z(a-1) = A, \\ (\nabla_{a^*}^{\beta} z)(b) = B. \end{cases}$$

If $A, B \geq 0$, then $z(t) \geq 0$ for $t \in \mathbb{N}_{a-1}^{b}$.

Proof Since $1 < \nu < 2$, we have that $N = \lceil \nu \rceil = 2$. Using Theorem 8, the general solution of $-\nabla_{a^*}^{\nu} z(t) = 0$ is given by

$$z(t) = \sum_{k=0}^{1} H_k(t, a) \nabla^k z(a).$$

Using the boundary condition on the left, we have

$$\begin{aligned} z(t) &= H_0(t, a) z(a) + H_1(t, a) \nabla z(a) \\ &= z(a) + H_1(t, a)(z(a) - z(a-1)) \\ &= z(a) + H_1(t, a)(z(a) - A), \end{aligned}$$

for $t \in \mathbb{N}_{a-1}^{b}$. Applying the boundary condition on the right, we have

$$(\nabla_{a^*}^{\beta} z)(b) = H_{1-\beta}(b, a)(z(a) - A) = B,$$

which implies

$$z(a) = A + \frac{B}{H_{1-\beta}(b, a)}.$$

Hence, for $t \in \mathbb{N}_{a-1}^{b}$,

$$z(t) = A + \frac{B}{H_{1-\beta}(b, a)} + H_1(t, a) \frac{B}{H_{1-\beta}(b, a)} \geq 0,$$

since $H_1(t, a) = t - a > 0$ and $H_{1-\beta}(b, a) = \frac{\Gamma(b-a+1-\beta)}{\Gamma(2-\beta)\Gamma(b-a)} > 0$. □

Using Theorems 17 and 18, the following comparison theorem is a generalization of Theorem 16.

Theorem 19 *Assume* $1 < \nu < 2$, $0 \leq \beta \leq 1$, $2 \leq b - a \leq \frac{1}{2-\nu}$, *and* $u, v : \mathbb{N}_{a-1}^{b} \to \mathbb{R}$ *such that* $u(t)$ *and* $v(t)$ *satisfy*

$$\begin{aligned} \nabla_{a^*}^{\nu} u(t) &\geq \nabla_{a^*}^{\nu} v(t), \quad t \in \mathbb{N}_{a+1}^{b} \\ u(a-1) &\geq v(a-1) \\ (\nabla_{a^*}^{\beta}) u(b) &\geq (\nabla_{a^*}^{\beta}) v(b). \end{aligned}$$

Then $u(t) \leq v(t)$ *for* $t \in \mathbb{N}_{a-1}^{b}$.

Proof Let $w(t) := u(t) - v(t)$ for $t \in \mathbb{N}_{a-1}^{b}$, $A := u(a-1) - v(a-1)$, and $B := (\nabla_{a^*}^{\beta}) u(b) - (\nabla_{a^*}^{\beta}) v(b)$. Then we have

$$h(t) := \nabla_{a^*}^{\nu} w(t) = \nabla_{a^*}^{\nu} u(t) - \nabla_{a^*}^{\nu} v(t) \geq 0, \quad t \in \mathbb{N}_{a+1}^{b}.$$

Note that $-w(t)$ is the solution of the FBVP

$$\begin{cases} -\nabla_{a^*}^{\nu} w(t) = h(t), & t \in \mathbb{N}_{a+1}^{b} \\ w(a-1) = A, \\ (\nabla_{a^*}^{\beta} w)(b) = B, \end{cases}$$

where $A, B \geq 0$.

Hence by Theorem 17,

$$-w(t) = z(t) + \int_a^b G(t,s)h(s)\nabla s, \quad t \in \mathbb{N}_{a-1}^{b}$$

where $z(t)$ solves the Caputo FBVP

$$\begin{cases} -\nabla_{a^*}^{\nu} z(t) = 0, & t \in \mathbb{N}_{a+1}^{b} \\ z(a-1) = A, \\ (\nabla_{a^*}^{\beta} z)(b) = B. \end{cases}$$

for $t \in \mathbb{N}_{a-1}^{b}$. Note that $z(t) \geq 0$ for $t \in \mathbb{N}_{a-1}^{b}$ by Theorem 18. We also have $G(t,s) \geq 0$ for $t \in \mathbb{N}_{a-1}^{b}$ and $s \in \mathbb{N}_{a+1}^{b}$ by Theorem 14 and $h(t) \geq 0$ for $t \in \mathbb{N}_{a+1}^{b}$. Therefore, we have $w(t) = u(t) - v(t) \leq 0$, $t \in \mathbb{N}_{a-1}^{b}$. □

5.2 Two-Point Problems

In this section we give the solution to a two-point FBVP in general and then study a nonlinear case.

We note that the Green's function $G(0;t,s)$ provides the unique solution to a two-point problem of the Caputo nabla difference:

$$\begin{cases} -\nabla_{a^*}^{\nu} y(t) = h(t), & t \in \mathbb{N}_{a+1}^{b} \\ y(a-1) = 0, \\ y(b) = 0, \end{cases} \tag{21}$$

where $h : \mathbb{N}_{a+1}^{b} \to \mathbb{R}$, $1 < \nu < 2$, $b - a \in \mathbb{Z}$ and $b - a > N - 1$.

Theorem 20 *The solution of the two-point problem* (21) *is given by*

$$y(t) = \int_a^b G(0;t,s)h(s)\nabla s, \quad t \in \mathbb{N}_{a-1}^{b},$$

where $G(0;t,s)$ *is the Green's function defined by*

$$G(0;t,s)=\begin{cases}\dfrac{H_{\nu-1}(b,\rho(s))(t-\rho(a))}{b-\rho(a)}, & a+1\le t\le\rho(s)\le b\\ \dfrac{H_{\nu-1}(b,\rho(s))(t-\rho(a))}{b-\rho(a)}-H_{\nu-1}(t,\rho(s)), & a+1\le\rho(s)\le t\le b.\end{cases}$$

Proof The two-point problem (21) is a special case of the Caputo nabla FBVP (1) for $\beta=0$. Therefore the result follows from the proof of Theorem 11. □

Next we consider a nonlinear two-point problem when $\dfrac{3}{2}\le\nu<2$

$$\begin{cases}-\nabla^{\nu}_{a^*}y(t)=h(t,y(t-1)), & t\in\mathbb{N}^{b}_{a+1}\\ y(a-1)=0,\\ y(b)=0,\end{cases}\tag{22}$$

where $h:\mathbb{N}^{b}_{a+1}\times\mathbb{R}^{+}\to\mathbb{R}^{+}$ is continuous, $b-a\in\mathbb{Z}$, and $2\le b-a\le\dfrac{1}{2-\nu}$.

We are going to apply the framework from Erbe and Peterson [8], where the authors discussed the existence of positive solutions to a boundary value problem on time scales.

Definition 12 Let $\mathcal{B}$ be a Banach space. Then $\mathcal{P}\subset\mathcal{B}$ is called a cone provided $\mathcal{P}$ is a nonempty, closed, convex subset of $\mathcal{B}$ satisfying

(i) $\lambda\ge0$ and $x\in\mathcal{P}$ implies $\lambda x\in\mathcal{P}$;
(ii) $x,-x\in\mathcal{P}$ implies $x=0$,

where 0 is the identity element in $\mathcal{B}$.

The following fixed point theorem concerning cone expansion and cone compression appears in [15, 19]. It has been a useful tool in the analysis of nonlinear problems in both differential and difference equations. See [3, 8, 23].

Theorem 21 *Let $\mathcal{B}$ be a Banach space and let $\mathcal{P}\subset\mathbb{E}$ be a cone. Assume Ω_1 and Ω_2 are open subsets of $\mathcal{B}$ with $0\in\Omega_1$ and $\overline{\Omega}_1\subset\Omega_2$, and assume that*

$$A:\mathcal{P}\cap(\overline{\Omega}_2/\Omega_1)\to\mathcal{P}$$

is a completely continuous operator such that either

(i) $\|Ax\|\le\|x\|$, $x\in\mathcal{P}\cap\partial\Omega_1$, *and* $\|Ax\|\ge\|x\|$, $x\in\mathcal{P}\cap\partial\Omega_2$, *or*
(ii) $\|Ax\|\ge\|x\|$, $x\in\mathcal{P}\cap\partial\Omega_1$, *and* $\|Ax\|\le\|x\|$, $x\in\mathcal{P}\cap\partial\Omega_2$.

Then A has a fixed point in $\mathcal{P}\cap(\overline{\Omega}_2\setminus\Omega_1)$.

For the analysis of the nonlinear two-point problem (22), we define a Banach space

$$\mathcal{E}:=\{y:\mathbb{N}^{b}_{a-1}\to\mathbb{R}:y(a-1)=y(b)=0\}$$

with the norm $\|\cdot\|$ defined by

$$\|y\| := \max\{|y(t)|, t \in \mathbb{N}_{a-1}^{b}\}.$$

We define a cone $\mathcal{K}$ in $\mathcal{E}$ by

$$\mathcal{K} := \{y \in \mathbb{E} : y(t) \geq 0 \text{ for } t \in \mathbb{N}_{a-1}^{b} \text{ and } y(t) \geq k\|y\| \text{ for } t \in \mathbb{N}_{a}^{b-1}\},$$

where $k \in (0, 1]$ is as defined in Theorem 12.

We define an operator A by

$$Ay(t) = \int_{a}^{b} G(0; t, s)h(s, y(s-1))\nabla s = \sum_{a+1}^{b} G(0; t, s)h(s, y(s-1))$$

for $t \in \mathbb{N}_{a}^{b-1}$ and $y \in \mathcal{K}$.

Now we show that $A : \mathcal{K} \to \mathcal{K}$. We have $G(0; t, s) \geq 0$ for $(t, s) \in \mathbb{N}_{a-1}^{b} \times \mathbb{N}_{a+1}^{b}$ from Theorem 12 and $h : \mathbb{N}_{a+1}^{b} \times \mathbb{R}^{+} \to \mathbb{R}^{+}$ from the two-point problem (22). So we have $Ay(t) \geq 0$. Then by Theorem 12 (ii),

$$\begin{aligned} Ay(t) &\geq \int_{a}^{b} kG(0; \rho(s), s)h(s, y(s-1))\nabla s \\ &\geq \int_{a}^{b} k \max_{t \in \mathbb{N}_{a}^{b-1}} G(0; t, s)h(s, y(s-1))\nabla s \\ &\geq k \max_{t \in \mathbb{N}_{a}^{b-1}} \int_{a}^{b} G(0; t, s)h(s, y(s-1))\nabla s \\ &= k\|y\|. \end{aligned}$$

So $Ay(t) \in \mathcal{K}$. Therefore, $A : \mathcal{K} \to \mathcal{K}$.

We also note that the operator A is completely continuous since it is a sum of finite discrete terms.

We will give sufficient conditions related to the behavior of $h(t, y)$ so that the nonlinear two-point problem (22) has a positive solution. We define γ and δ such that

$$\begin{aligned} \frac{1}{\gamma} &:= \int_{a}^{b} G(0; \rho(s), s)\nabla s, \text{ and} \\ \frac{1}{\delta} &:= k \int_{a}^{b-1} G(0; t_0, s)\nabla s, \end{aligned}$$

for fixed $t_0 \in \mathbb{N}_{a}^{b-1}$.

Theorem 22 *If there exist* $\mu_1, \mu_2 \in (0, \infty)$ *such that*

$$h(t, y) \leq \gamma\mu_1 \ \forall\ t \in \mathbb{N}_{a-1}^{b} \text{ and } 0 \leq y \leq \mu_1, \text{ and}$$
$$h(t, y) \geq \delta y \ \forall\ t \in \mathbb{N}_{a}^{b-1} \text{ and } k\mu_2 \leq y \leq \mu_2,$$

then the two-point FBVP (22)

$$\begin{cases} -\nabla_{a^*}^{\nu} y(t) = h(t, y(t-1)), & t \in \mathbb{N}_{a+1}^{b} \\ y(a-1) = 0, \\ y(b) = 0 \end{cases}$$

has a positive solution.

Proof Case 1: $\mu_1 < \mu_2$. Let $\Omega_1, \Omega_2 \subset \mathcal{E}$ such that Ω_1 be the ball centered at the origin with radius μ_1 and Ω_2 be the ball centered at the origin with radius μ_2.

If $y \in \mathcal{K}$ and $\|y\| = \mu_1$, we have

$$\begin{aligned} Ay(t) &= \int_a^b G(0; t, s)h(s, y(s-1))\nabla s \\ &\leq \int_a^b G(0; \rho(s), s)h(s, y(s-1))\nabla s \\ &\leq \gamma\mu_1 \int_a^b G(0; \rho(s), s)\nabla s \\ &= \mu_1. \end{aligned}$$

So $\|Ay\| \leq \|y\|$ for $y \in \mathcal{K} \cap \partial\Omega_1$.

If $y \in \mathcal{K}$ and $\|y\| = \mu_2$, then $y \geq k\mu_2$ implies $y \geq k\|y\|$ and we have

$$\begin{aligned} Ay(t_0) &= \int_a^b G(0; t_0, s)h(s, y(s-1))\nabla s \\ &\geq \int_a^{b-1} G(0; t_0, s)h(s, y(s-1))\nabla s \\ &\geq \delta \int_a^{b-1} G(0; t_0, s)y(s-1)\nabla s \\ &\geq \delta k\|y\| \int_a^{b-1} G(0; t_0, s)\nabla s \\ &= \|y\|. \end{aligned}$$

So $\|Ay\| \geq \|y\|$ for $y \in \mathcal{K} \cap \partial\Omega_2$. Hence by Theorem 21 (i), the operator A has a fixed point in $\mathcal{K} \cap (\overline{\Omega}_2 \setminus \Omega_1)$. Therefore the FBVP (22) has a positive solution $y(t)$ such that $\mu_1 \leq \|y\| \leq \mu_2$.

Case 2: $\mu_2 < \mu_1$. Let $\Omega_1, \Omega_2 \subset \mathcal{E}$ such that Ω_1 be the ball centered at the origin with radius μ_2 and Ω_2 be the ball centered at the origin with radius μ_1.

If $y \in \mathcal{K}$ and $\|y\| = \mu_2$, then by the similar argument as in Case 1, we have $\|Ay\| \geq \|y\|$ for $y \in \mathcal{K} \cap \partial\Omega_1$. If $y \in \mathcal{K}$ and $\|y\| = \mu_1$, then by the similar argument as in Case 1, we have $\|Ay\| \leq \|y\|$ for $y \in \mathcal{K} \cap \partial\Omega_2$. Hence by Theorem 21 (ii), the operator A has a fixed point in $\mathcal{K} \cap (\overline{\Omega}_2 \setminus \Omega_1)$. Therefore the FBVP (22) has a positive solution $y(t)$ such that $\mu_2 \leq \|y\| \leq \mu_1$. □

Now we consider more restrictions to the behavior of $h(t, y)$. We have assumed that $h : \mathbb{N}_{a+1}^b \times \mathbb{R}^+ \to \mathbb{R}^+$. We further assume that the limits

$$\lambda_0 := \lim_{y \to 0^+} \frac{h(t, y)}{y} \quad \text{and} \quad \lambda_\infty := \lim_{y \to \infty} \frac{h(t, y)}{y}$$

exist uniformly in $\mathbb{R} \cup \{-\infty, \infty\}$.

Theorem 23 *If either $\lambda_0 = 0$ and $\lambda_\infty = \infty$ or $\lambda_0 = \infty$ and $\lambda_\infty = 0$, the two-point FBVP (22)*

$$\begin{cases} -\nabla_{a^*}^{\nu} y(t) = h(t, y(t-1)), & t \in \mathbb{N}_{a+1}^b \\ y(a-1) = 0, \\ y(b) = 0, \end{cases}$$

has a positive solution.

Proof Assume $\lambda_0 = 0$ and $\lambda_\infty = \infty$.

Since $\lambda_0 = \lim_{y \to 0^+} \frac{h(t,y)}{y} = 0$, we pick $r_1 > 0$ such that

$$h(t, y) \leq \gamma y$$

for $0 \leq y \leq r_1$ and $t \in \mathbb{N}_{a-1}^b$. Let $\Omega_1 := \{y \in \mathcal{K} : \|y\| < r_1\}$ be the open ball in $\mathcal{E}$ centered at the origin with radius r_1. If $y \in \mathcal{K} \cap \partial\Omega_1$, then $\|y\| = r_1$ and

$$\begin{aligned} Ay(t) &= \int_a^b G(0; t, s) h(s, x(s-1)) \nabla s \\ &\leq \gamma \int_a^b G(0; t, s) x(s-1) \nabla s \\ &\leq \gamma r_1 \int_a^b G(0; t, s) \nabla s \\ &\leq \gamma r_1 \int_a^b G(0; \rho(s), s) \nabla s \\ &= r_1. \end{aligned}$$

Hence $\|Ay\| \leq \|y\|$.

Since $\lambda_\infty = \lim_{y \to \infty} \frac{h(t,y)}{y} = \infty$, there exists an r_1' such that

$$h(t, y) \geq \delta r_1'$$

for $y \geq r_1'$. Then we define

$$R_1 := \max\{2r_1, r_1'\}$$

and $\Omega_2 := \{y \in \mathcal{K} : \|y\| < R_1\}$. Now for $y \in \mathcal{K} \cap \partial\Omega_2$ and $t_0 \in N_a^{b-1}$, we have

$$\begin{aligned} Ay(t_0) &= \int_a^b G(0; t_0, s)h(s, y(s-1))\nabla s \\ &\geq \int_a^{b-1} G(0; t_0, s)h(s, y(s-1))\nabla s \\ &\geq \delta \int_a^{b-1} G(0; t_0, s)y(s-1)\nabla s \\ &\geq \delta k\|y\| \int_a^{b-1} G(0; t_0, s)\nabla s \\ &= \|y\|. \end{aligned}$$

Hence $\|Ay\| \geq Ay(t_0) \geq \|y\|$. Therefore by Theorem 21 (i), A has a fixed point in $\mathcal{K} \cap (\overline{\Omega}_2 \setminus \Omega_1)$. It follows that the FBVP (22) has a positive solution.

Next we assume $\lambda_0 = \infty$ and $\lambda_\infty = 0$. Since $\lambda_0 = \lim_{y\to 0^+} \frac{h(t,y)}{y} = \infty$, we may pick $r_2 > 0$ such that

$$h(t, y) \geq \delta y$$

for $0 \leq y \leq r_2$ and $t \in \mathbb{N}_{a-1}^b$. Let $\Omega_1 := \{y \in \mathcal{K} : \|y\| < r_2\}$ be the open ball in $\mathcal{E}$ centered at the origin with radius r_2. If $y \in \mathcal{K} \cap \partial\Omega_1$, then $\|y\| = r_2$ and

$$\begin{aligned} Ay(t_0) &= \int_a^b G(0; t_0, s)h(s, y(s-1))\nabla s \\ &\geq \int_a^{b-1} G(0; t_0, s)h(s, y(s-1))\nabla s \\ &\geq \delta \int_a^{b-1} G(0; t_0, s)y(s-1)\nabla s \\ &\geq \delta k\|y\| \int_a^{b-1} G(0; t_0, s)\nabla s \\ &= \|y\|. \end{aligned}$$

Hence $\|Ay\| \geq Ay(t_0) \geq \|y\|$. Since $\lambda_\infty = \lim_{y\to\infty} \frac{h(t,y)}{y} = 0$, there exists an r_2' such that

$$h(t, y) \leq \gamma r_2'$$

for $y \geq r_2'$. Then we define

$$R_2 := \max\{2r_2, r_2'\}$$

and $\Omega_2 := \{y \in \mathcal{K} : \|y\| < R_2\}$. Now for $y \in \mathcal{K} \cap \partial\Omega_2$, we have

$$\begin{aligned} Ay(t) &= \int_a^b G(0; t, s)h(s, x(s-1))\nabla s \\ &\leq \gamma \int_a^b G(0; t, s)x(s-1)\nabla s \\ &\leq \gamma\|y\| \int_a^b G(0; t, s)\nabla s \\ &\leq \gamma\|y\| \int_a^b G(0; \rho(s), s)\nabla s \\ &= \|y\| = R_2. \end{aligned}$$

Hence $\|Ay\| \leq \|y\|$.

Therefore by Theorem 21 (ii), A has a fixed point in $\mathcal{K} \cap (\overline{\Omega}_2 \setminus \Omega_1)$. It follows that the FBVP (22) has a positive solution. □

References

1. Ahrendt, K., Castle, L., Holm, M., Yochman, K.: Laplace transforms for the nabla-difference operator and a fractional variation of parameters formula. Commun. Appl. Anal. **16**(3), 317 (2012)
2. Ahrendt, K.: The Existence of Solutions for a Nonlinear, Fractional Self-Adjoint Difference Equation. Ph.D. thesis, University of Nebraska-Lincoln, 2017
3. Atıcı, F.M., Eloe, P.W.: Two-point boundary value problems for finite fractional difference equations. J. Diff. Equ. Appl. **17**(04), 445–456 (2011)
4. Atıcı, F.M., Uyanik, M.: Analysis of discrete fractional operators. Appl. Anal. Discret. Mat., 139–149 (2015)
5. Atıcı, F.M., Belcher, M., Marshall, D.: A new approach for modeling with discrete fractional equations. Fundam. Inf. **151**(1–4), 313–324 (2017)
6. Brackins, A.M.: Boundary Value Problems of Nabla Fractional Difference Equations. Ph.D. thesis, University of Nebraska-Lincoln (2015)
7. Eloe, P.W., Kublik C.M., Neugebauer, J.T.: Comparison of green's functions for a family of boundary value problems for fractional difference equations. J. Diff. Equ. Appl., 1–12 (2018)
8. Erbe, L., Peterson, A.: Positive solutions for a nonlinear differential equation on a measure chain. Math. Comput. Model. **32**(5–6), 571–585 (2000)
9. Erbe, L., Goodrich, C.S., Jia, B., Peterson, A.: Survey of the qualitative properties of fractional difference operators: monotonicity, convexity, and asymptotic behavior of solutions. Adv. Diff. Equ. **2016**(1), 43 (2016)

10. Gensler, S.C.: Fractional Difference Operators and Relatedlems. Ph.D. thesis, University of Nebraska-Lincoln (2018)
11. Goodrich, C.S.: On a fractional boundary value problem with fractional boundary conditions. Appl. Math. Lett. **25**(8), 1101–1105 (2012)
12. Goodrich, C., Peterson, A.C.: Discrete Fractional Calculus. Springer (2015)
13. Goodrich, C.S.: A note on convexity, concavity, and growth conditions in discrete fractional calculus with delta difference. Math. Inequal. Appl. **19**, 769–779 (2016)
14. Graef, J.R., Ho, S., Kong, L., Wang, M.: A fractional differential equation model for bike share systems. The Joint Mathematics Meetings, Baltimore, MD, January 2019
15. Guo, D., Lakshmikantham, V.: Nonlinear Problems in Abstract Cones. Academic Press (1988)
16. Holm, M.T.: The theory of Discrete Fractional Calculus: Development and Application. Ph.D. thesis, University of Nebraska-Lincoln (2011)
17. Ikram, A.: Green's Functions and Lyapunov Inequalities for Nabla Caputo Boundary Value Problems. Ph.D. thesis, University of Nebraska-Lincoln (2018)
18. Jia, B., Erbe, L., Goodrich, C., Peterson, A.: The relation between nabla fractional differences and nabla integer differences. Filomat **31**(6), 1741–1753 (2017)
19. Krasnosel'skiĭ, M.A.: Positive solutions of operator equations. P. Noordhoff (1964)
20. Machado, J.T., Silva, M.F., Barbosa, R.S., Jesus, I.S., Reis, C.M., Marcos, M.G., Galhano, A.F.: Some applications of fractional calculus in engineering. Math. Probl. Eng. (2010)
21. Miller, K.S., Ross, B.: Fractional difference calculus. In: Proceedings of the International Symposium on Univalent Functions, Fractional Calculus and Their Applications, pp. 139–152 (1988)
22. Setniker, A.: Sequential Differences in Nabla Fractional Calculus. Ph.D. thesis, University of Nebraska-Lincoln (2019)
23. St Goar, J.: A Caputo Boundary Value Problem in Nabla Fractional Calculus. Ph.D. thesis, University of Nebraska-Lincoln (2016)
24. Valério, D., Machado, J.T., Kiryakova, V.: Some pioneers of the applications of fractional calculus. Fract. Calc. Appl. Anal. **17**(2), 552–578 (2014)

A Note on Ergodicity for Nonautonomous Linear Difference Equations

Mihály Pituk

Abstract For a class of nonautonomous linear difference equations with bounded, nonnegative and uniformly primitive coefficients it is shown that the normalized positive solutions are asymptotically equivalent to the Perron vectors of the transition matrix at infinity.

Keywords Difference equation · Ergodicity · Asymptotic behaviour

1 Introduction

Let $\mathbb{R}$, $\mathbb{R}_+$ and $\mathbb{Z}_+$ denote the set of real numbers, the set of nonnegative numbers and the set of nonnegative integers, respectively. For a positive integer d, $\mathbb{R}^d$ and $\mathbb{R}^{d\times d}$ denote the d-dimensional space of real column vectors and the space of $d \times d$ real matrices, respectively.

Let $\leq$ be the partial order on $\mathbb{R}^d$ induced by the nonnegative cone $\mathbb{R}^d_+$, the set of those vectors in $\mathbb{R}^d$ which have nonnegative components. For $x = (x_1, \ldots, x_d)^T$ and $y = (y_1, \ldots, y_d)^T \in \mathbb{R}^d$, we have $x \leq y$ if and only if $x_i \leq y_i$ for all $i = 1, \ldots, d$. We write $x < y$ if $x \leq y$ and $x_i < y_i$ for some $i \in \{1, \ldots, d\}$ and we write $x \ll y$ if $x_i < y_i$ for all $i = 1, \ldots, d$. A vector x is called *nonnegative*, *positive* and *strongly positive* if $0 \leq x$, $0 < x$ and $0 \ll x$, respectively. A similar notation and terminology is used for matrices. The set of nonnegative matrices in $\mathbb{R}^{d\times d}$ is denoted by $\mathbb{R}^{d\times d}_+$.

A norm $\|\cdot\|$ on $\mathbb{R}^d$ is called *monotone* if $0 \leq x \leq y$ implies $\|x\| \leq \|y\|$. In the sequel, $\|\cdot\|$ denotes any monotone norm on $\mathbb{R}^d$ and the associated induced norm on $\mathbb{R}^{d\times d}$. A vector $x \in \mathbb{R}^d$ is called *normalized* if $\|x\| = 1$.

Consider the nonautonomous linear difference equation

$$x(t+1) = B(t)x(t), \qquad t \in \mathbb{Z}_+, \tag{1}$$

M. Pituk (✉)
Department of Mathematics, University of Pannonia,
Egyetem út 10, 8200 Veszprém, Hungary
e-mail: pitukm@almos.uni-pannon.hu

S. Baigent et al. (eds.), *Progress on Difference Equations and Discrete Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 341,
https://doi.org/10.1007/978-3-030-60107-2_2

where $B\colon \mathbb{Z}_+ \to \mathbb{R}_+^{d\times d}$. Throughout the paper, we shall assume that there exist nonnegative matrices $P,\ Q \in \mathbb{R}_+^{d\times d}$ such that

$$P \le B(t) \le Q \qquad \text{for all } t \in \mathbb{Z}_+ \tag{2}$$

and

$$P \text{ is a primitive matrix.} \tag{3}$$

Recall that a matrix $M \in \mathbb{R}_+^{d\times d}$ is *primitive* if there exists a positive integer q such that $0 \ll M^q$. It is known [1] that every primitive matrix $M \in \mathbb{R}_+^{d\times d}$ has a unique strongly positive normalized eigenvector, the so-called *Perron vector of* M, which will be denoted by $p(M)$. The Perron vector corresponds to the spectral radius $r(M)$ of M so that $Mp(M) = r(M)p(M)$, $0 \ll p(M)$ and $\|p(M)\| = 1$.

Note that assumptions (2) and (3) imply that $B(t)$ is nonnegative and primitive for every $t \in \mathbb{Z}_+$. Therefore each $B(t)$ has a unique Perron vector denoted by $p(B(t))$, $t \in \mathbb{Z}_+$.

The following result from [2] shows that if B in Eq. (1) is slowly varying at infinity, then the normalized positive solutions of (1) are asymptotically equivalent to the Perron vectors of the coefficient matrices $B(t)$ as $t \to \infty$.

Theorem 1 [2, Theorem 1] *Suppose* (2) *and* (3) *hold. Assume also that*

$$B(t+1) - B(t) \longrightarrow 0 \qquad \text{as } t \to \infty. \tag{4}$$

Then for every solution $x\colon \mathbb{Z}_+ \to \mathbb{R}^d$ *of* (1) *with initial value* $x(0) \in \mathbb{R}_+^d \setminus \{0\}$, *we have*

$$\frac{x(t)}{\|x(t)\|} - p(B(t)) \longrightarrow 0 \qquad \text{as } t \to \infty, \tag{5}$$

where $p(B(t))$ *is the Perron vector of* $B(t)$ *for* $t \in \mathbb{Z}_+$.

The asymptotic relation (5) shows that in the long run the behavior of the normalized positive solutions is independent of the initial data. This fact is sometimes called an *ergodic property* of Eq.(1). For the origin of the terminology and further related results, see [3–6] and the references therein.

The purpose of this note is to give an asymptotic description of the normalized positive solutions of Eq. (1) without assuming the slowly varying condition (4). Our main result is formulated in Sect. 3 after presenting some notations and lemmas in Sect. 2. In Sect. 4, we illustrate the main theorem by two examples.

2 Notations and Lemmas

The proof of our main result will be based on the properties of Hilbert's projective metric. Let $\mathbb{R}_{++}^d$ denote the set of strongly positive vectors in $\mathbb{R}^d$. For $x,\ y \in \mathbb{R}_{++}^d$, we define *Hilbert's projective metric* by

$$d(x, y) = \ln \frac{\max_{1\le i\le n} \frac{x_i}{y_i}}{\min_{1\le i\le n} \frac{x_i}{y_i}} = \max_{1\le i,j\le n} \ln \frac{x_i y_j}{x_j y_i}. \tag{6}$$

In the following lemma, we list some basic facts about the projective metric.

Lemma 1 [8, Theorem 2.1, p. 7] *For all x, y and $z \in \mathbb{R}^d_{++}$, we have*

(i) $d(x, y) \ge 0$,
(ii) $d(x, y) = 0$ *if and only if* $y = \beta x$ *for some positive constant* β,
(iii) $d(x, y) = d(y, x)$,
(iv) $d(x, y) \le d(x, z) + d(z, y)$,
(v) $d(\beta x, \gamma y) = d(x, y)$ *for any positive constants* β *and* γ.

A matrix $M \in \mathbb{R}^{d\times d}_+$ is *row-allowable* if it has a positive entry in each of its rows. The next result shows that linear mappings generated by nonnegative row-allowable matrices are nonexpansive, while strongly positive matrices act as contractions in the projective metric.

Lemma 2 [8, Theorem 2.6, p. 22] *Let $M = (m_{ij}) \in \mathbb{R}^{d\times d}_+$ be a nonnegative row-allowable matrix. Then for any x and $y \in \mathbb{R}^d_{++}$, we have*

$$d(Mx, My) \le \tau_B(M) d(x, y), \tag{7}$$

where $\tau_B(M)$ is Birkhoff's contractivity coefficient *given by*

$$\tau_B(M) = \frac{1 - \sqrt{\phi(M)}}{1 + \sqrt{\phi(M)}} \quad \text{with} \quad \phi(M) = \min_{1\le i,j,k,l\le n} \frac{m_{ik} m_{jl}}{m_{jk} m_{il}} \quad \text{if } M \gg 0 \tag{8}$$

and $\tau_B(M) = 1$ if M has at least one 0 *entry.*

By Lemma 2, $\tau_B(M) \le 1$ whenever M is nonnegative and row-allowable and $\tau_B(M) < 1$ if $M \gg 0$. Furthermore, the explicit expression (8) implies that the function τ_B is continuous on the open the set of strongly positive matrices.

We shall also need a result which shows that for strongly positive normalized sequences the convergence in the projective metric implies convergence in any monotone norm.

Lemma 3 [9, Lemma 6.4, p. 217] *For any monotone norm $\|\cdot\|$ on $\mathbb{R}^d$, we have*

$$\|x - y\| \le 3(1 - e^{-d(x,y)}) \quad \text{whenever } x, y \in \mathbb{R}^d_{++} \text{ and } \|x\| = \|y\| = 1. \tag{9}$$

3 Main Result and Proof

The solutions of Eq. (1) can be written as

$$x(t) = X(t, s)x(s) \qquad \text{for } t \ge s \ge 0, \tag{10}$$

where $X(t,s)$, $t \geq s \geq 0$, is the *transition matrix* defined by

$$X(t,s) = B(t-1)B(t-2)\cdots B(s) \qquad \text{for } t \geq s \geq 0. \tag{11}$$

(By definition, $X(s,s) = I$ for $s \geq 0$, where I denotes the $d \times d$ identity matrix.)

Our main result is the following theorem.

Theorem 2 *Suppose (2) and (3) hold. Then for every solution $x\colon \mathbb{Z}_+ \to \mathbb{R}^d$ of (1) with initial value $x(0) \in \mathbb{R}^d_+ \setminus \{0\}$, we have*

$$\frac{x(t)}{\|x(t)\|} - p(X(t,0)) \longrightarrow 0 \qquad \text{as } t \to \infty, \tag{12}$$

where $p(X(t,0))$ is the Perron vector of the transition matrix $X(t,0)$ given by (11).

Theorem 2 may be viewed as a refinement of the Weak Ergodic Theorem by Golubitsky et al. [3, Theorem 2.2] which states that if we take the l_1-norm on $\mathbb{R}^d$, then under conditions (2) and (3) for every pair of solutions x and y of (1) with positive initial data $x(0)$ and $y(0)$,

$$\frac{x(t)}{\|x(t)\|} - \frac{y(t)}{\|y(t)\|} \longrightarrow 0 \qquad \text{as } t \to \infty. \tag{13}$$

For the continuous analogue of Theorem 2, see [7, Theorem 2.2].

Now we give a proof of Theorem 2 which is an appropriate modification of the proof of the Weak Ergodic Theorem [3, Theorem 2.2].

Proof By virtue of (2) and (11), we have

$$P^{t-s} \leq X(t,s) \leq Q^{t-s} \qquad \text{for } t \geq s \geq 0. \tag{14}$$

Since P is nonnegative and primitive, there exists $q > 0$ such that $P^q \gg 0$. From (14), we find that

$$0 \ll P^q \leq X(jq,(j-1)q) \leq Q^q \qquad \text{for } j = 1,2,\ldots. \tag{15}$$

Since P and hence its powers are primitive and every primitive matrix is evidently row-allowable, (14) implies that $X(t,s)$ is nonnegative and row-allowable for $t \geq s \geq 0$. This, together with (15), implies that

$$X(t,0) = X(t,q)X(q,0) \gg 0 \qquad \text{for } t \geq q.$$

Hence

$$x(t) = X(t,0)x(0) \gg 0 \qquad \text{for } t \geq q.$$

Furthermore, by Lemma 2, we have

$$\tau_B(X(t,s)) \le 1 \qquad \text{for } t \ge s \ge 0. \tag{16}$$

Define

$$C = \{\, M \in \mathbb{R}_+^{d\times d} \mid P^q \le M \le Q^q \,\}.$$

Evidently, C is a compact set of strongly positive matrices. As noted before, the Birkhoff contractivity function τ_B is continuous on the set of strongly positive matrices and hence it attains its maximum θ on the compact set C. Using Lemma 2 again, we conclude that

$$\theta = \max_{M\in C} \tau_B(M) < 1.$$

By virtue of (15), we have that $X(jq,(j-1)q) \in C$ for $j = 1, 2, \ldots$. Hence

$$\tau_B(X(jq,(j-1)q)) \le \theta < 1 \qquad \text{for } j = 1, 2, \ldots. \tag{17}$$

Let $t \ge q$ and write $p(t) = p(X(t,0))$ and $r(t) = r(X(t,0))$ for the Perron vector and the spectral radius of $X(t,0)$, respectively, so that

$$X(t,0)p(t) = r(t)p(t).$$

In view of Lemma 3, in order to prove (12) it is enough to show that

$$d\left(\frac{x(t)}{\|x(t)\|}, p(t)\right) = d(x(t), r(t)p(t)) = d(X(t,0)x(0), X(t,0)p(t)) \longrightarrow 0 \tag{18}$$

as $t \to \infty$. (Note that the last and the last but one equalities in (18) follow from Lemma 1 (v).) Let $t \ge q$ be fixed and $k = [t/q]$, the greatest integer part of t/q. By the application of Lemma 2, we obtain

$$\begin{aligned} d(X(t,0)x(0), X(t,0)p(t)) &= d(X(t,kq)X(kq,0)x(0), X(t,kq)X(kq,0)p(t)) \\ &\le \tau_B(X(t,kq))d(X(kq,0)x(0), X(kq,0)p(t)). \end{aligned}$$

This, together with (16), implies

$$d(X(t,0)x(0), X(t,0)p(t)) \le d(X(kq,0)x(0), X(kq,0)p(t)). \tag{19}$$

Taking into account that

$$X(kq,0) = X(kq,(k-1)q)X((k-1)q,(k-2)q)\cdots X(2q,q)X(q,0),$$

a repeated use of Lemma 2, combined with (17), yields

$$d(X(kq,0)x(0), X(kq,0)p(t)) \le \theta^{k-1} d(X(q,0)x(0), X(q,0)p(t)). \tag{20}$$

Let

$$D = \{ X(q,0)v \mid v \in \mathbb{R}^d_+, \ \ \|v\| = 1 \}.$$

Since $X(q,0) \gg 0$, D is a compact subset of $\mathbb{R}^d_{++}$. As noted before, $x(q) = X(q,0)x(0) \gg 0$ and therefore (6) implies that the mapping $u \longmapsto d(x(q),u)$ is continuous on $\mathbb{R}^d_{++}$. As a consequence, the above mapping is bounded on the compact set $D \subset \mathbb{R}^d_{++}$. This implies the existence of $\gamma > 0$ such that for all $u \in D$, we have that

$$d(X(q,0)x(0),u) = d(x(q),u) \le \gamma.$$

From this, taking into account that $X(q,0)p(t) \in D$, we find that

$$d(X(q,0)x(0), X(q,0)p(t)) \le \gamma.$$

This, together with (19) and (20), yields

$$d(X(t,0)x(0), X(t,0)p(t)) \le \gamma\theta^{[t/q]-1} \longrightarrow 0 \qquad \text{as } t \to \infty.$$

Thus, (18) holds. □

4 Examples

We will illustrate Theorem 2 by two examples.

Example 1 We give an asymptotic description of the normalized positive solutions of Eq. (1), where $B\colon \mathbb{Z}_+ \to \mathbb{R}^{2\times 2}_+$ is a 2-periodic matrix function defined by

$$B(t) = \frac{1}{2}\begin{pmatrix} \sqrt{3} & 2+(-1)^t \\ 2-(-1)^t & \sqrt{3} \end{pmatrix} \qquad \text{for } t \in \mathbb{Z}_+.$$

We will use the l_2-norm on $\mathbb{R}^2$. Assumptions (2) and (3) of Theorem 2 are satisfied with

$$P = \frac{1}{2}\begin{pmatrix} \sqrt{3} & 1 \\ 1 & \sqrt{3} \end{pmatrix} \qquad \text{and} \qquad Q = \frac{1}{2}\begin{pmatrix} \sqrt{3} & 3 \\ 3 & \sqrt{3} \end{pmatrix},$$

but the slowly varying condition (4) of Theorem 1 is violated. Thus, Theorem 1 does not apply. As shown in [2, Example 1], in this case conclusion (5) of Theorem 1 does not hold. We shall establish the asymptotic behaviour of the normalized positive solutions of Eq. (1) by applying Theorem 2. In view of the 2-periodicity of B, we have for $t \in \mathbb{Z}_+$,

$$X(2t,0) = M^t, \qquad \text{where } M = B(1)B(0) = \begin{pmatrix} 1 & \sqrt{3} \\ \sqrt{3} & 3 \end{pmatrix}$$

and

$$X(2t+1,0) = N^t B(0), \qquad \text{where } N = B(0)B(1) = \begin{pmatrix} 3 & \sqrt{3} \\ \sqrt{3} & 1 \end{pmatrix}.$$

An easy calculation shows that the Perron vectors and the spectral radii of M and N are

$$p(M) = \frac{1}{2}\begin{pmatrix} 1 \\ \sqrt{3} \end{pmatrix}, \qquad r(M) = 4,$$

and

$$p(N) = \frac{1}{2}\begin{pmatrix} \sqrt{3} \\ 1 \end{pmatrix}, \qquad r(N) = 4,$$

respectively. Since $Mp(M) = 4p(M)$ implies $M^t p(M) = 4^t p(M)$ for $t \geq 1$, in view of uniqueness, M and M^t share the same Perron vector. Hence

$$p(X(2t,0)) = p(M^t) = p(M) \qquad \text{for } t \geq 1. \tag{21}$$

We claim that

$$p(X(2t+1,0)) = p(N^t B(0)) = p(N) \qquad \text{for } t \geq 1. \tag{22}$$

Indeed, if we write $p(t) = p(N^t B(0))$ and $r(t) = r(N^t B(0))$ for brevity, then

$$N^t B(0)p(t) = r(t)p(t) \qquad \text{and} \qquad N^t p(N) = 4^t p(N) \qquad \text{for } t \geq 1.$$

From this, using Lemma 1 (v), we find for $t \geq 1$,

$$\begin{aligned} d(p(t), p(N)) = d(r(t)p(t), 4^t p(N)) &= d(N^t B(0)p(t), N^t p(N)) \\ &\leq (\tau_B(N))^t d(B(0)p(t), p(N)), \end{aligned}$$

where the last inequality follows from Lemma 2. By virtue of (8), we have that $\tau_B(N) = 0$ and hence $d(p(t), p(N)) = 0$ for $t \geq 1$. In view of Lemma 3, this implies (22). Finally, from (21) and (22), by the application Theorem 2, we conclude that for every solution of Eq. (1) with initial value $x(0) > 0$,

$$\frac{x(t)}{\|x(t)\|} - s(t) \longrightarrow 0 \qquad \text{as } t \to \infty, \tag{23}$$

where $s : \mathbb{Z}_+ \to \mathbb{R}^d_+$ is a 2-periodic sequence defined by $s(t) = p(M)$ if $t \in \mathbb{Z}_+$ is even and $s(t) = p(N)$ if $t \in \mathbb{Z}_+$ is odd.

Example 2 Consider Eq. (1), where $B : \mathbb{Z}_+ \to \mathbb{R}^{2\times 2}_+$ is defined by

$$B(t) = \begin{pmatrix} 1 + e^{-t} & 2 - e^{-t} \\ 1 + \sin^2\sqrt{t} & 1 + \cos^2\sqrt{t} \end{pmatrix} \qquad \text{for } t \in \mathbb{Z}_+.$$

Assumptions (2) and (3) of Theorem 2 are satisfied with

$$P = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad Q = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}.$$

We will use the l_1-norm $\|\cdot\|$ on $\mathbb{R}^2$. It is easy to verify that

$$B(t)u = 3u \qquad \text{for } t \in \mathbb{Z}_+, \qquad \text{where } u = \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{24}$$

Hence $\frac{1}{2}u$ is a normalized strongly positive eigenvector of $B(t)$ for all $t \in \mathbb{Z}_+$. Since $B(t)$ is nonnegative and primitive, in view of the uniqueness, $\frac{1}{2}u$ must be the Perron vector of $B(t)$, i.e. $p(B(t)) = \frac{1}{2}u$ for all $t \in \mathbb{Z}_+$. From (11) and (24), it follows by easy induction that

$$X(t, 0)u = 3^t u \qquad \text{for } t \in \mathbb{Z}_+$$

which implies by a similar argument as before that $p(X(t, 0)) = \frac{1}{2}u$ for all $t \in \mathbb{Z}_+$. By the application Theorem 2, we conclude that for every solution of Eq. (1) with initial value $x(0) > 0$,

$$\frac{x(t)}{\|x(t)\|} \longrightarrow \frac{1}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad \text{as } t \to \infty. \tag{25}$$

Acknowledgements This work was supported in part by the Hungarian National Research, Development and Innovation Office grant no. KH130513 and Széchenyi 2020 under the EFOP-3.6.1-16-2016-00015.

References

1. Berman, A., Plemmons, R.J.: Nonnegative Matrices in the Mathematical Sciences, Classics in Applied Mathematics, vol. 9. SIAM, Philadelphia (1994)
2. Pituk, M., Pötzsche, C.: Ergodicity beyond asymptotically autonomous linear difference equations. Appl. Math. Lett. **86**, 149–156 (2018). https://doi.org/10.1016/j.aml.2018.06.030
3. Golubitsky, M., Keeler, E.B., Rothschild, M.: Convergence of the age structure: Application of the projective metric. Theoret. Population Biol. **7**, 84–93 (1975). https://doi.org/10.1016/0040-5809(75)90007-6
4. Cushing, J.M.: An Introduction to Structured Population Dynamics. SIAM, Philadelphia (1998)
5. Abu-Saris, R., Elaydi, S., Jang, S.: Poincaré type solutions of systems of difference equations. J. Math. Anal. Appl. **275**, 69–83 (2002). https://doi.org/10.1016/S0022-247X(02)00239-1
6. Krause, U.: Positive Dynamical Systems in Discrete Time. De Gruyter, Berlin (2015)
7. Pituk, M., Pötzsche, C.: Ergodicity in nonautonomous linear ordinary differential equations. J. Math. Anal. Appl. **479**, 149–156 (2019). https://doi.org/10.1016/j.jmaa.2019.07.005
8. Hartfiel, D.J.: Nonhomogeneous Matrix Products. World Scientific, New Jersey (2002)
9. Krause, U., Nesemann, T.: Differenzengleichungen und diskrete dynamische Systeme (2. Auflage, in German), de Gruyter, Berlin (2012)

Poincaré Return Maps in Neural Dynamics: Three Examples

Marina L. Kolomiets and ANDREY L. SHILNIKOV

Abstract Understanding of the onset and generic mechanisms of transitions between distinct patterns of activity in realistic models of individual neurons and neural networks presents a fundamental challenge for the theory of applied dynamical systems. We use three examples of slow-fast neural systems to demonstrate a suite of new computational tools to study diverse neuronal systems.

Keywords Neurodynamics · Poincaré return maps · Neural model · Networks

1 Introduction

Most neurons demonstrate oscillations of the membrane potential either endogenously or due to external perturbations. Deterministic description of primary oscillatory activities, such as tonic spiking and bursting, of neuronal dynamics is based on models following the Hodgkin-Huxley formalism [1]. Mathematically, such conductance based models belong to a special class of dynamical systems with at least two distinct time scales, the so-called slow—fast systems [2–8]. Bursting is a manifestation of slow–fast dynamics possessing subcomponents operating at distinct time scales. Neural bursting is a modular activity composed of various limiting branches, corresponding to oscillatory and equilibrium regimes of the fast subsystem, and

M. L. Kolomiets
Department of Mathematics, Academy of Agricultural Sciences, Nizhniy, Novgorod 603107, Russia

A. L. SHILNIKOV (✉)
Neuroscience Institute, Department of Mathematics and Statistics, Georgia State University, Atlanta, Georgia 30303, USA
e-mail: ashilnikov@gsu.edu
URL: https://labs.ni.gsu.edu/ashilnikov/

S. Baigent et al. (eds.), *Progress on Difference Equations and Discrete Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 341,
https://doi.org/10.1007/978-3-030-60107-2_3

connected by transients between them. Using the common mathematical we can better understand the basic onset of bursting oscillations in models of individual and coupled neurons. The study of mechanisms of bursting and its transformations requires nonlocal bifurcation analysis, which is based on the derivation and further examination of Poincaré return maps.

2 Hodgkin-Huxley Type Model of a Leech Heart Interneuron

Our first example is the "reduced" model of heart interneuron model [9–13] derived through the Hodgkin-Huxley gated variables formalism [1] that not every mathematician may be familiar with. Its equations do look too detailed and overwhelming:

$$
\begin{aligned}
&C\,\frac{dV}{dt} = -I_{\mathrm{Na}} - I_{\mathrm{K2}} + I_{\mathrm{L}} - I_{\mathrm{app}} - I_{\mathrm{syn}}, \qquad (1)\\
&I_L = \bar{g}_{\mathrm{L}}\,(V - E_{\mathrm{L}}), \qquad I_{\mathrm{K2}} = \bar{g}_{\mathrm{K2}}\, m_{\mathrm{K2}}^2 (V - E_{\mathrm{K}}),\\
&I_{\mathrm{Na}} = \bar{g}_{\mathrm{Na}}\, m_{\mathrm{Na}}^3\, h_{\mathrm{Na}}\,(V - E_{\mathrm{Na}}), \quad m_{\mathrm{Na}} = m_{\mathrm{Na}}^{\infty}(V),\\
&\tau_{\mathrm{Na}}\,\frac{dh_{\mathrm{Na}}}{dt} = h_{\mathrm{Na}}^{\infty}(V) - h, \quad \tau_{\mathrm{K2}}\frac{dm_{\mathrm{K2}}}{dt} = m_{\mathrm{K2}}^{\infty}(V) - m_{\mathrm{K2}},
\end{aligned}
$$

where $C = 0.5$ nF is the membrane capacitance; V is the membrane potential; $I_{\mathrm{N}a}$ is the fast voltage gated sodium current with slow inactivation h_{Na} and fast activation m_{Na}; I_{K2} is the persistent potassium current with activation m_{K2}; I_{L} is leak current and I_{app} is a constant polarization or external applied current. The maximal conductances are $\bar{\mathrm{g}}_{\mathrm{K2}} = 30$nS, $\bar{g}_{\mathrm{Na}} = 200$nS and $\mathrm{g_L} = 8$nS, and the reversal potentials are $\mathrm{E_{Na}} = 0.045$ V, $\mathrm{E_K} = -0.070$V and $E_{\mathrm{L}} = -0.046$V. The time constants of gating variables are $\tau_{\mathrm{K2}} = 0.25$ sec and $\tau_{\mathrm{Na}} = 0.0405$ s. The steady state values of gating variables, $h_{\mathrm{Na}}^{\infty}(V)$, $m_{\mathrm{Na}}^{\infty}(V)$, $m_{\mathrm{K2}}^{\infty}(V)$, are given by the following sigmoidal functions:

$$
\begin{aligned}
h_{\mathrm{Na}}^{\infty}(V) &= [1 + \exp(500(0.0333 - V))]^{-1}\\
m_{\mathrm{Na}}^{\infty}(V) &= [1 + \exp(-150(0.0305 - V))]^{-1} \qquad (2)\\
m_{\mathrm{K2}}^{\infty}(V) &= [1 + \exp\,(-83(0.018 - V + \mathrm{V_{K2}^{shift}}))]^{-1}.
\end{aligned}
$$

The quantity $\mathrm{V_{K2}^{shift}}$ is a genuine bifurcation parameter for this model: it is the deviation from experimentally averaged voltage value $V_{1/2} = 0.018$ V corresponding to semi-activated potassium channel, i.e. $m_{\mathrm{K2}}^{\infty}(0.018) = 1/2$. Variations of $\mathrm{V_{K2}^{shift}}$ move the slow nullcline $\frac{dm_{\mathrm{K2}}}{dt} = 0$ in the V-direction in the 3D phase, see Fig. 1. Due to the disparity of the time constants of the phase variables, the fast-slow system paradigm is applicable to system (1): its first two differential equations form a fast subsystem, while the last equation is the slow one. The dynamics of such a system are known [14] to be determined by, and centered around, attracting pieces of the slow motion

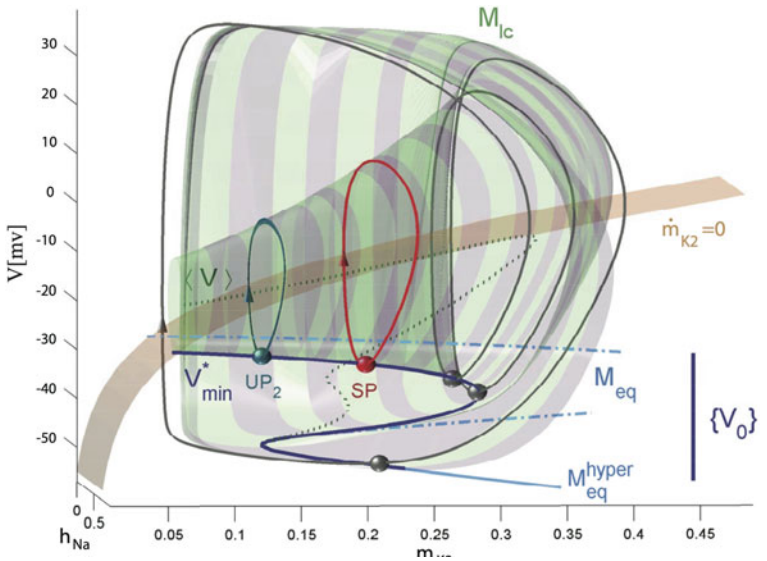

Fig. 1 Slow motion manifolds and nullclines of the model (1): the 2D spiking manifold M_{lc} is foliated by the periodic orbits continued, from the left to the right, as the parameter V_{K2}^{shift} is increased from -0.026 through 0.0018. The space curves V_{min} and $\langle V \rangle$ are made of minimal and average coordinates of the periodic orbits. M_{lc} glues to the hyperpolarized fold of the quiescent manifold, M_{eq}, comprised of the equilibrium states of (2), where the curve of the averaged values $\langle V \rangle$ terminates. An equilibrium state of Eqs. (2) is the intersection point of M_{eq} with the slow (yellow) nullcline $\dot{m}_{K2} = 0$ for given V_{K2}^{shift}. Also shown (in red) is the curve of the v-minimal coordinate values of the periodic orbits making M_{lc}. This curve is used to define the Poincaré map taking it onto itself after one revolution around M_{lc}

manifolds that constitute a skeleton of activity patterns. These manifolds are formed by the limit sets, such as equilibria and limit cycles, of the fast subsystem where the slow variable becomes a parameter in the singular limit.

A typical Hodgkin-Huxley model possesses a pair of such manifolds [15]: quiescent and tonic spiking, denoted by M_{eq} and M_{lc}, correspondingly. A solution of (2) that repeatedly switches between the low, hyperpolarized branch of M_{eq} and the spiking manifold M_{lc} represents a busting activity in the model. Whenever the spiking manifold M_{lc} is transient for the solutions of (1), like those winding around it in Figs. 2, the models exhibits regular or chaotic bursting. Otherwise, the model (1) has a spiking periodic orbit that has emerged on M_{lc} through the saddle-node bifurcation thereby terminating the bursting activity [16] or both regimes may co-exist as in [17, 18].

To determine what makes the spiking and bursting attractors change their shapes and stability, we construct numerically a V_{K2}^{shift} - parameter family of 1D Poincaré maps taking an interval of membrane potentials onto itself. This interval is comprised of the minimal values, denoted by (V_0), of the membrane potential on the found periodic orbits foliating densely the spiking manifold M_{lc}, see Fig. 1. Then, for some V_{K2}^{shift} -

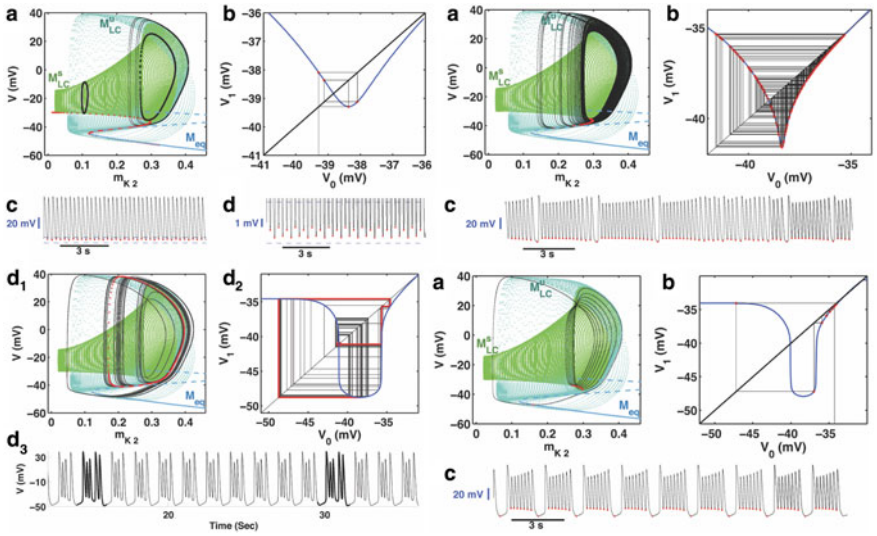

Fig. 2 (Top-left) Four v-minimums of the stable spiking periodic orbit spiking at $V_{K2}^{shift} = 0.0255$ corresponding to the period-4 orbit of the Poincaré map. Insets (C) and (D) show the voltage waveforms.(Top-right) Chaotic spiking of the model and in the map at $V_{K2}^{shift} = -0.0254$. (Bottom) Chaotic bursting at the spike adding transition becomes more regularized with a large number of spikes per burst

values, we integrate numerically the outgoing solutions of (2) starting from the initial conditions corresponding to each (V_0) to find the consecutive minimum (V_1) in the voltage time series. All found pairs (V_0, V_1) constitute the graph of the Poincaré map for given V_{K2}^{shift}.

Figure 2 is a showcase of such 1D unimodal maps with the distinctive U-shape. A fixed point of map would correspond to a single V-minimum on the periodic orbit on the 2D tonic spiking manifold, while period-2 orbit of the map corresponds to the periodic orbit of the model and so forth. A bursting orbit with multiple turns around M_{lc} and switching to and back from M_{lc} is represented by a more complex orbit of a longer period. Moreover, the bursting orbit may become even chaotic at spike adding transition, and as the map reveals that is caused by a homoclinic orbit (red trajectory) of an unstable fixed point corresponding to a saddle periodic orbit of the neural model (1). The shape of the 1D return map infers that as it becomes steeper with a characteristic cusp shape the model would move into the chaotic regime.

3 FitzHugh-Nagumo-Rinzel Model

Our next example is the FitzHugh-Nagumo-Rinzel (FNR) model which is a mathematical model of an elliptic burster (see Fig. 3B); its equations given by [19]:

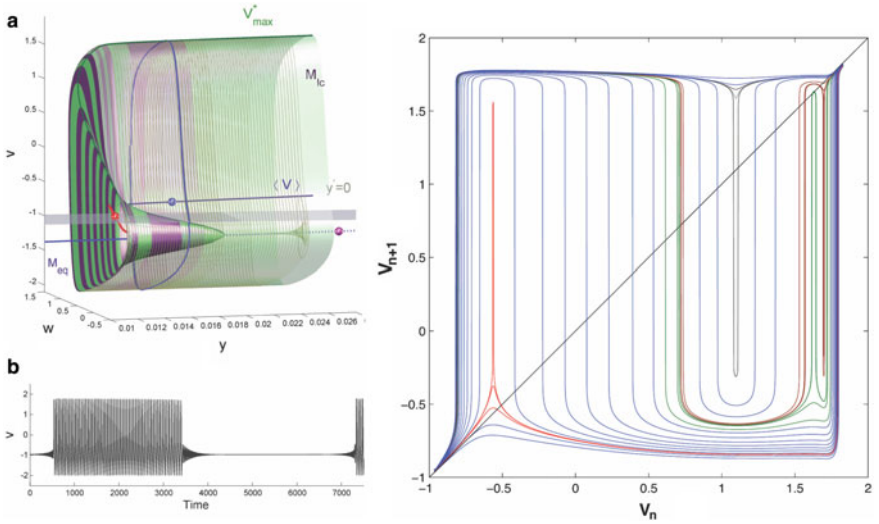

Fig. 3 (A) Topology of the tonic spiking, M_{lc}, and quiescent, M_{eq}, manifolds. The fold on M_{lc}, corresponds to a saddle-node bifurcation where the stable (outer) and saddle (inner) branches, comprised of periodic orbits, merge. The vertex, where the unstable branch of M_{lc} collapses at M_{eq}, corresponds to a subcritical Andronov-Hopf bifurcation. Space curves, labeled by V^*_{max} (in green) and $\langle V^{s,u} \rangle$ (in blue and red, respectively), correspond to the V-maximal and the averaged, over the period, coordinates of the periodic orbits composing M_{lc}. The plane, $y' = 0$, is the slow nullcline, above (below) which the y-component of a solution of the model increases (decreases). The plane is elevated/lowered as the c-parameter is increased/decreased. (right) The "continuously" reshaping family of the 1D Poincaré return maps $T : V_n \to V_{n+1}$ for the FHN-model at $\mu = 0.002$ as c increases from $c = -1$ through $c = -0.55$. Lower graphs correspond to quiescence and subthreshold oscillations in the model; upper graphs correspond to tonic spiking dynamics, while the middle graphs describe bifurcations of bursting. An intersection point of a graph with the bisectrix is a fixed point of the map. The stability of the fixed point is determined by the slope of the graph, i.e. it is stable if $|T'| < 1$

$$\begin{aligned} v' &= v - v^3/3 - w + y + I, \\ w' &= \delta(0.7 + v - 0.8w), \\ y' &= \mu(c - y - v). \end{aligned} \tag{3}$$

Here, $\delta = 0.08$, $I = 0.3125$ is an "external current", and we set $\mu = 0.002$ determining the pace of the slow variable y; the bifurcation parameter of the model is c.

The slow variable y becomes frozen when $\mu = 0$. The first two fast equations in (3) compose the FitzHugh-Nagumo fast subsystem model describing a relaxation oscillator, provided δ is small. This subsystem exhibits either tonic spiking on a stable limit cycle, or quiescence on a stable equilibrium state for some fixed values of y. Stability loss of the equilibrium state in the fast subsystem gives rise to a stable limit cycle through a sub-critical Andronov-Hopf bifurcation when an unstable limit cycle collapses into the equilibrium state. The stable and unstable limit cycle emerge in the FNR-model through a saddle-node bifurcation. Both bifurcations, Andronov-Hopf

and saddle-node, are key to the description of an elliptic burster. Using a traditional slow-fast dissection, one can locate the corresponding branches of the limit cycle and equilibrium states by varying the frozen y-variable in the extended phase space of the fast subsystem. The topology of the tonic spiking, M_{lc}, and quiescent, M_{eq}, in the phase space the FNR-model is revealed in Fig. 3.

4 1D Voltage Maps

Recall that a feature of a slow-fast system is that its solutions are constrained to stay near the slow-motion manifolds, composed of equilibria and periodic obits of the fast subsystem. If both manifolds are transient for the solutions of the corresponding neuron model, it exhibits a bursting behavior, which is a repetitive alternation of tonic spiking and quiescent periods. Otherwise, the model demonstrates the tonic spiking activity if there is a stable periodic orbit on the tonic spiking manifold, or it shows no oscillations when solutions are attracted to a stable equilibrium state on the quiescent manifold.

The core of the methods is a reduction to, and a derivation of, a low dimensional Poincaré return map, with an accompanying analysis of the limit solutions: fixed, periodic and homoclinic orbits, representing various oscillations in the original model. Maps have been actively employed in computational neuroscience, see [20–23] and referenced therein. It is customary that such a map is sampled from voltage traces, for example by singling out successive voltage maxima or minima, or interspike intervals. A drawback of a map generated by time series is a sparseness, as the construction algorithm reveals only a single periodic attractor of a model, unless the latter demonstrates chaotic or mixing dynamics producing a large variety of densely wandering points.

A new, computer assisted method for constructing a complete family of Poincaré maps for an interval of membrane potentials for slow-fast Hodgkin-Huxley models of neurons was proposed in [12] following [24], see above. Having such maps we are able to elaborate on bifurcations in the question of tonic spiking and bursting, detect bistability, as well examine unstable sets, which are the organizing centers of complex dynamics in any model. Using this approach we have studied complex bursting transformations in a leech heart interneuron model and revealed that the cause of complex behaviors at transitions is homoclinic tangles of saddle periodic orbits which can be drastically amplified by small noise [11, 25]. Examination of the maps will help us make qualitative predictions about transitions *before* they actually occur in the models.

The construction of the voltage interval maps is a two stage routine. First, we need to accurately single out the slow motion manifold M_{lc} in the neuronal model using the parameter continuation technique. The manifold is formed by the tonic-spiking periodic orbits as a control parameter in the *slow* equation is varied. Recall, that its variations, raising or lowering the slow nullcline in the phase space of the model, do not alter the fast subsystem and hence do keep the manifold intact. Next a space

curve V^*_{max} on M_{lc} is detected, which corresponds to maximal voltage values of the membrane potentials V_n found on all periodic orbits constituting the tonic spiking manifold, see Fig. 3.

We use this data to further amend the set $\{V_n\}$, by integrating the solutions of the model in the vicinity of each maxima to find the exact locations of the turning points, determined by the condition $V'_{max} = 0$. Next, the points defining $\{V_n\}$ are employed as the initial conditions to compute outgoing solutions of (3) that will stay on or close to M_{lc}. The integration is stopped when a successive maximal value $\{V_{n+1}\}$ of the voltage is reached in the voltage trace. Figure 4 demonstrates how the shape of the 1D maps changes in a complex predictable way as the c-parameter is varied. One can see from the end points, that the map has initially a stable fixed point at the top-right corner that corresponds to the stable tonic spiking orbit on the outer surface of the 2D manifold M_{lc} in Fig. 3(left). One can also foresee from the map at the bottom-right corner in Fig. 3(right) the neural model will undergo a cascade of period-doubling bifurcations of sub-threshold oscillations followed by complex mixed-mode oscillations involving sub-threshold ones and bursting. Our predictions are illustrated and confirmed by Fig. 4 that samples four characteristic 1D Poincaré return maps out of Fig. 3. In it the shape of the 1D Poincaré return maps reveals the underlying cause of chaotic mixed mode oscillations (MMOs) at the transition from tonic spiking to bursting in the in the FNR-model (3) that next become periodic MMOs, and further transition to chaotic and regular sub-threshold oscillations en a route to the quiescent phase in generic elliptic bursters.

5 Example 3: 2D Recurrent Maps in Multifunctional 3-Cell Networks

Many rhythmic motor behaviors such as respiration, chewing, locomotion on land and in water, and heartbeat (in leeches) are produced by networks of cells called central pattern generators (CPGs). A CPG is a neural microcircuit of cells whose synergetic, nonlinear interactions can autonomously generate an array of multicomponent/polyrhythmic bursting patterns of activity that determine motor behaviors in animals, including humans [26–32]. Modeling studies, phenomenologically mathematical and exhaustively computational, have proven useful to gain insights into operational principles of CPGs [33–40]. Although various models, reduced and feasible, of specific CPGs, have been developed, it remains unclear how the CPGs achieve the level of robustness and stability observed in nature [41–45]. Understanding the key universal mechanisms of the functional evolution of neural connectivity, bifurcation mechanisms underlying transitions between different neural activities, and accurate modeling of these processes presents opportunity and challenge for applied mathematics in particular and for all computational sciences in general.

Whereas a dedicated CPG generates a single pattern robustly, a multifunctional or polymorphic CPG can flexibly produce distinct rhythms, such as temporally dis-

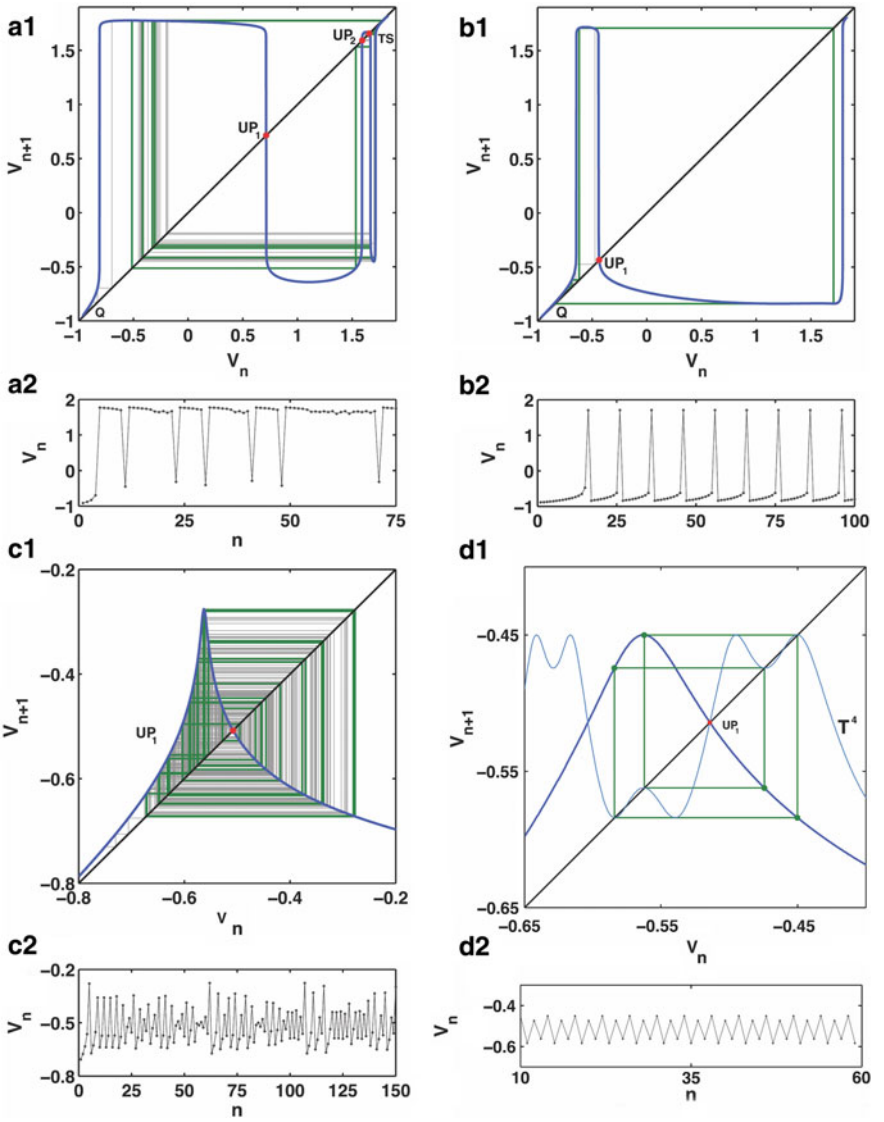

Fig. 4 (A1/2) The shape of the 1D Poincaré return map reveals the underlying cause of chaotic mixed mode oscillations (MMOs) at the transition from tonic spiking to bursting in the in the FNR-model (3) that become periodic MMOs with a single burst followed by nine sub-threshold oscillations (B1/2). (C1/2) The unimodal map corresponding to chaotic and period-4 sub-threshold oscillations (D1/2)

tinct swimming versus crawling locomotions, and alternation of directions of blood circulation in leeches [46–48]. Switching between various attractors of a CPG network causes switching between locomotion behaviors. Each attractor is associated with a definite rhythm running on a specific time scale with well-defined and robust phase lags among the constituting neurons. The emergence of synchronous rhythms in neural networks is closely related to temporal characteristics of coupled neurons due to intrinsic properties and types of synaptic coupling, which can be inhibitory, excitatory and electrical, fast and slow [49–53].

We developed a computational toolkit for oscillatory networks that reduces the problem of the occurrence of bursting and spiking rhythms generated by a CPG network to the bifurcation analysis of attractors in the corresponding Poincaré return maps for the phase lags between oscillatory neurons. The structure of the phase space of the map is an individual signature of the CPG as it discloses all characteristics of the functional space of the network. Recurrence of rhythms generated by the CPG (represented by a system of coupled Hodgkin-Huxley type neurons [54]) lets us employ Poincaré return maps defined for phase lags between spike/burst initiations in the constituent neurons (Fig. 5) [41, 49–51, 55]. Forward trajectories $\left\{\phi_{21}^{(n)}, \phi_{31}^{(n)}\right\}$ of phase points $\mathbf{M}_n = \left(\phi_{21}^{(n)}, \phi_{31}^{(n)}\right)$ of the Poincaré map $\Pi : \mathbf{M}_n \to \mathbf{M}_{n+1}$ are defined through the time delays $\Delta\phi_{j1}^{(n)} = \dfrac{\tau_{j1}^{(n+1)} - \tau_{j1}^{(n)}}{\tau_1^{(n+1)} - \tau_1^{(n)}}$ (on mod 1) between the burst initiations in each cycle normalized over the network period, can converge to several co-existing stable fixed points, thus indicating the given network is multistable, or a single stable invariant circle wrapping around the torus that corresponds to a unique rhythmic outcome with periodically varying phase lags. These are attractors, single or multiple, of the return map on a 2D torus, which are associated with multifunctional or dedicated neural circuits, respectively (Fig. 5). The 2D return map, $\Pi : \mathbf{M}_n \to \mathbf{M}_{n+1}$, for the phase lags can be written as follows:

$$\phi_{21}^{(n+1)} = \phi_{21}^{(n)} + \mu_1 f_1\left(\phi_{21}^{(n)}, \phi_{31}^{(n)}\right), \qquad \phi_{31}^{(n+1)} = \phi_{31}^{(n)} + \mu_2 f_2\left(\phi_{21}^{(n)}, \phi_{31}^{(n)}\right) \tag{4}$$

with μ_i representing the coupling strength, and f_i being some undetermined coupling functions such that $f_1 = f_2 = 0$ corresponds to its fixed points: $\phi_{j1}^{*} = \phi_{j1}^{(n+1)} = \phi_{j1}^{(n)}$. These functions, similar to phase-resetting curves, can be assessed from the simulated data collected for known all trajectories $\left\{\phi_{21}^{(n)}, \phi_{31}^{(n)}\right\}$. By treating f_i as partials $\partial F/\partial\phi_{ij}$, we can restore a "phase potential" $F\left(\phi_{21}, \phi_{31}\right) = C$ that determines the dynamics of the coupled neurons, find its critical points associated with FPs—attractors, repellers and saddles of the map, and by scaling f_i predict their bifurcations due to loss of stability, and hence transformations of rhythmic outcomes of the network as a whole.

With such return maps, we can predict and identify the set of robust outcomes in a CPG with mixed, inhibitory and excitatory, slow or/and fast synapses, which are differentiated by phase-locked or periodically varying lags corresponding, respec-

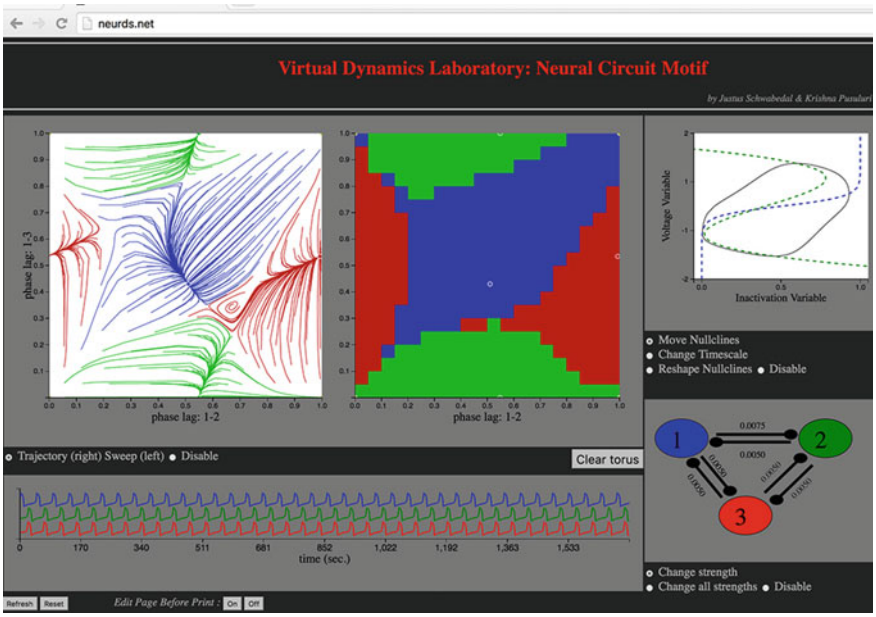

Fig. 5 GPU-based interactive motif-toolbox [56, 57] for computational studies of rhythmogenesis in 3-cell circuits comprised of synaptically coupled FitzhHugh-Nagumo, Hodgkin-Huxley, and 2Θ-neurons, which can generate up to 6 (3 in this figure) robust patterns corresponding to the stable fixed points in the 2D Poincaré return map for the phase lags between constituent cells.

tively, to stable fixed points and invariant circles of the return map. The toolkit lets us predict bifurcations and transformations of rhythmic outcomes before they actually occur in the network. The approach also reveals the capacity of the network and the dependence of its outcomes on coupling strength, wiring circuitry, and synapses, thereby letting one quantitatively and qualitatively identify necessary and sufficient conditions for rhythmic outcomes to occur. Using graphics processor units (GPUs) for parallel simulations of multistable neural networks using multiple initial conditions (as depicted in Fig. 5) can drastically speed up the bifurcation analysis and reduce a simulation time to merely few seconds.

Acknowledgements This work was funded in part by the NSF grant IOS-1455527 and the RSF grant 14-41-00044 at Lobachevsky University of Nizhny Novgorod. We thank the Brains and Behavior initiative of Georgia State University for providing pilot grant support. We acknowledge the support of NVIDIA Corporation with the Tesla K40 GPUs used in this study. Finally, we are grateful to all the current and past members of the Shilnikov NeurDS lab for productive discussions.

References

1. Hodgkin, A.L., Huxley, A.F.: A quantitative description of membrane current and its application to conduction and excitation in nerve. J. Physiol. **117**(4), 500–544 (1952)
2. Arnold, V.I., Afraimovich, V.S., Ilyashenko, Yu.S., Shilnikov, L.P.: Bifurcation Theory, Vol. V of Dynamical Systems. Encyclopaedia of Mathematical Sciences. Springer (1994)
3. Bertram, R., Butte, M.J., Kiemel, T., Sherman, A.: Topological and phenomenological classification of bursting oscillations. Bull. Math. Biol. **57**(3), 413–439 (1995)
4. Izhikevich, E.M.: Dynamical systems in neuroscience. MIT Press, Cambridge, Mass, The geometry of excitability and bursting (2007)
5. Jones, C.K.R.T., Kopell, N.: Tracking invariant-manifolds with differential forms in singularly perturbed systems. J. Differ. Equ. **108**(1), 64–88 (1994)
6. Rinzel, J.: Bursting oscillations in an excitable membrane model. Lect. Notes Math. **1151**, 304–316 (1985)
7. Rinzel, J., Ermentrout, B.: Analysis of neural excitability and oscillations. In: Koch, C., Segev, I. (eds.) Computational neuroscience, pp. 135–169. MIT Press, Cambridge, Mass (1998)
8. Rinzel, J., Wang, X.J.: Oscillatory and bursting properties of neurons. In Arbib, M. (ed.) The Handbook of Brain Theory and Neural Networks, pp. 686–691. MIT Press (1995)
9. Shilnikov, A.L., Cymbalyuk, G.: Transition between tonic spiking and bursting in a neuron model via the blue-sky catastrophe. Phys. Rev. Lett. **94**(4), 048101 (2005)
10. Shilnikov, A.L., Calabrese, R.L., Cymbalyuk, G.: Mechanism of bistability: tonic spiking and bursting in a neuron model. Phys. Rev. E **71**, 056214 (2005)
11. Channell, P., Cymbalyuk, G., Shilnikov, A.: Origin of bursting through homoclinic spike adding in a neuron model. Phys. Rev. Lett. **98**(13), 134101 (2007a)
12. Channell, P., Cymbalyuk, G., Shilnikov, A.L.: Applications of the poincare mapping technique to analysis of neuronal dynamics. Neurocomputing **70**, 10–12 (2007b)
13. Shilnikov, A., Cymbalyuk, G.: PRL **94**, 048101 (2005)
14. Tikhonov, A.N.: Mat. Sb. **31** 575 (1952); N. Fenichel, J. Diff. Eq. **31**, 53 (1979)
15. Rinzel, J., Ermentrout, B, Koch, C., Segev, I.: Methods in Neuronal Modelling: From Synapses to Networks. MIT Press (1989)
16. Shilnikov, L.P., Shilnikov, A.L., Turaev, D.V., Chua, L.O.: Methods qualitative theory in nonlinear dynamics, Vols. I-II. World Sci. Publ. (1998, 2001); Shilnikov, A.L., Shilnikov, L.P., Turaev, D.V.: Moscow Math J. **5**(1), 205 (2005)
17. Shilnikov, A., Calabrese, R.L., Cymbalyuk, G.: Neurocomputing **65–66**, 869 (2005)
18. Cymbalyuk, G.S., Shilnikov, A.L.: J. Comp. Neuroscience **18**(3), 255 (2004); Regular & Chaotic Dynamics **9**(3), 281 (2004)
19. Wojcik, J., Shilnikov, A.: Voltage interval mappings for activity transitions in neuron models for elliptic bursters. Physica D **240**(14–15), 1164–1180 (2011)
20. Chay, T.R.: Chaos in a three-variable model of an excitable cell. Physica D **16**(2), 233–242 (1985)
21. Griffiths, R.E., Pernarowski, M.C.: Return map characterizations for a model of bursting with two slow variables. SIAM J. Appl. Math. **66**(6), 1917–1948 (2006)
22. Shilnikov, A.L., Rulkov, N.F.: Origin of chaos in a two-dimensional map modelling spiking-bursting neural activity. Int. J. Bifurcation Chaos **13**(11), 3325–3340 (2003)
23. Shilnikov, A.L., Rulkov, N.F.: Subthreshold oscillations in a map-based neuron model. Phys. Lett. A **328**(2–3), 177–184 (2004)
24. Shilnikov, A.L.: On bifurcations of the Lorenz attractor in the Shimizu-Morioka model. Physica D **62**(1–4), 338–346 (1993)
25. Channell, P., Fuwape, I., Neiman, A., Shilnikov, A.L.: Variability of bursting patterns in a neuronal model in the presence of noise. J. Computat. Neurosci. **27**(3), 527–42 (2009)
26. Marder, E., Calabrese, R.L.: Principles of rhythmic motor pattern generation. Physiol Rev. **76**(3), 687–717 (1996)
27. Kristan, W.B., Calabrese, R.L., Friesen, W.O.: Neuronal control of leech behavior. Prog. Neurobiol. **76**, 279 (2005)

28. Calin-Jageman, R.J., Tunstall, M.J., Mensh, B.D., Katz, P.S., Frost, W.N.: Parameter space analysis suggests multi-site plasticity contributes to motor pattern initiation in tritonia. J. Neurophysiol. **98**, 2382 (2007)
29. Newcomb, J.M., Sakurai, A., Lillvis, J.L., Gunaratne, C.A., Katz, P.S.: Homology and homoplasy of swimming behaviors and neural circuits in the nudipleura (mollusca, gastropoda, opistho-branchia). Proc. Natl. Acad. Sci. **109**(1), 10669–76 (2012)
30. Selverston, A. (ed.): Model Neural Networks and Behavior. Springer, Berlin (1985)
31. Bal, T., Nagy, F., Moulins, M.: The pyloric central pattern generator in crustacea: a set of conditional neural oscillators. J. Comparat. Physiol. A **163**(6), 715–727 (1988)
32. Katz, P.S., Hooper, S.L.: Invertebrate central pattern generators. In: North, G., Greenspan, R.R. (eds.) Invertebrate Neurobiology. Cold Spring Harbor Laboratory Press, NY, New York (2007)
33. Marder, E., Calabrese, R.L.: Principles of rhythmic motor pattern generation. Physiol Rev. **76**(3), 687–717 (1996). July
34. Kopell, N., Ermentrout, B.: Chemical and electrical synapses perform complementary roles in the synchronization of interneuronal networks. Proc. Natl. Acad Sci. **101**(43), 15482–15487 (2004)
35. Matsuoka, K.: Mechanisms of frequency and pattern control in the neural rhythms generators. Biol. Cybernetics **1**, 1 (1987)
36. Kopell, N.: Toward a theory of modelling central pattern generators. In: Cohen, A.H., Rossingol, S., Grillner, S. (eds.) Neural Control of Rhythmic Movements in Vertebrates. Wiley, New York (1988)
37. Canavier, C.C., Baxter, D.A., Clark, J.W., Byrne, J.H.: Multiple modes of activity in a model neuron suggest a novel mechanism for the effects of neuromodulators. J Neurophysiol **72**(2), 872–882 (1994). Aug
38. Skinner, F., Kopell, N., Marder, E.: Mechanisms for oscillation and frequency control in networks of mutually inhibitory relaxation oscillators. Comput. Neurosci. **1**, 69 (1994)
39. Dror, R.O., Canavier, C.C., Butera, R.J., Clark, J.W., Byrne, J.H.: A mathematical criterion based on phase response curves for stability in a ring of coupled oscillators. Biol Cybern **80**(1), 11–23 (1999). Jan
40. Prinz, A.A., Billimoria, C.P., Marder, E.: Alternative to hand-tuning conductance-based models: construction and analysis of databases of model neurons. J. Neurophysiol. **90**(6), 3998–4015 (2003). December
41. Belykh, I.V., Shilnikov, A.L.: When weak inhibition synchronizes strongly desynchronizing networks of bursting neurons. Phys. Rev. Lett. **101**(7), 078102 (2008)
42. Shilnikov, A.L., Gordon, R., Belykh, I.: Polyrhythmic synchronization in bursting networking motifs. Chaos **18**(3), 037120 (2008)
43. Sherwood, W.E., Harris-Warrick, R., Guckenheimer, J.M.: Synaptic patterning of left-right alternation in a computational model of the rodent hindlimb central pattern generator. J. Comput Neuroscience **30**(2), 323 (2010)
44. Koch, H., Garcia, A.J., Ramirez, J.-M.: Network reconfiguration and neuronal plasticity in rhythm-generating networks. Integrat. Comparat. Biol. **51**(6), 856–868 (2011)
45. Marder, E.: Neuromodulation of neuronal circuits: back to the future. Neuron **76**, 1 (2012)
46. Calabrese, R.L., Norris, B.J., Wenning, A., Wright, T.M.: Coping with variability in small neuronal networks. Integrat. Comparat. Biol. **51**(6), 845–855 (2011)
47. Kristan, W.B.: Neuronal decision-making circuits. Curr. Biol. **18**(19), R928–R932 (2008). Oct
48. Briggman, K.L., Kristan, W.B.: Multifunctional pattern-generating circuits. Annu. Rev. Neurosci. **31**, 271–294 (2008)
49. Wojcik, J., Clewley, R., Shilnikov, A.L.: Order parameter for bursting polyrhythms in multifunctional central pattern generators. Phys. Rev. E **83**, 056209–6 (2011)
50. Wojcik, J., Clewley, R., Schwabedal, J., Shilnikov, A.L.: Key bifurcations of bursting polyrhythms in 3-cell central pattern generators. PLoS ONE **9**(4) (2014)
51. Jalil, S., Belykh, I., Shilnikov, A.L.: pikes matter in phase-locking of inhibitory bursting networks. Phys. Rev. E. **85**, 36214 (2012)

52. Kopell, N., Somers, D.: Rapid synchronization through fast threshold modulation. Biol. Cybern. **68**, 5 (1993)
53. Marder, E.: Invertebrate neurobiology: polymorphic neural networks. Curr. Biol. **4**(8), 752–754 (1994)
54. Shilnikov, A.L.: Complete dynamical analysis of an interneuron model. J. Nonlinear Dyn. **68**(3), 305–328 (2012)
55. Jalil, S., Allen, D., Youker, J., Shilnikov, A.L.: Toward robust phase-locking in melibe swim central pattern generator models. J. Chaos **23**(4), 046105 (2013)
56. Knapper, D., Schwabedal, J., Shilnikov, A.L.: Qualitative and quantitative stability analysis of penta-rhythmic circuits. Nonlinearity **29**(12), 3647–3676 (2016)
57. Schwabedal, J., Pusuluri, K.: MotifToolBox *https://github.com/jusjusjus/Motiftoolbox* (2016)

Persistent Discrete-Time Dynamics on Measures

Horst R. Thieme

Abstract A discrete-time structured population model is formulated by a *population turnover map* F on the cone of finite nonnegative Borel measures that maps the structural population distribution of a given year to the one of the next year. F has a first order approximation at the zero measure (the extinction fixed point), which is a positive linear operator on the ordered vector space of real measures and can be interpreted as a *basic population turnover operator*. A spectral radius can be defined by the usual Gelfand formula and can be interpreted as *basic population turnover number*. We continue our investigation (Thieme, H.R.: Discrete-time population dynamics on the state space of measures, Math. Biosci. Engin. 17:1168–1217 (2020). doi: 10.3934/mbe.2020061) in how far the spectral radius serves as a threshold parameter between population extinction and population persistence. Emphasis is on conditions for various forms of uniform population persistence if the basic population turnover number exceeds 1.

Keywords Extinction · Basic reproduction number · Feller kernel · Eigenfunctions · Flat norm (also known as dual bounded lipschitz norm)

1 Introduction

Many animal and plant populations have yearly cycles with reproduction occurring once a year during a relatively short period. They also carry population structures which may be due to spatial distribution, age or rank structure, or degree of maturity.

It seems appropriate to describe such populations by discrete-time structured models in the form of difference equations,

H. R. Thieme (✉)
School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ 85287-1804, USA
e-mail: hthieme@asu.edu
URL: https://asu.digication.com/horst-thieme/welcome/published

S. Baigent et al. (eds.), *Progress on Difference Equations and Discrete Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 341,
https://doi.org/10.1007/978-3-030-60107-2_4

$$x_n = F(x_{n-1}), \qquad n \in \mathbb{N}, \quad x_0 \in X_+, \tag{1}$$

with the population structure being encoded in the closed subset $X_+ \ni 0$ of a normed vector space X over $\mathbb{R}$, $F(0) = 0$ [17, 20, 32]. The vector x_n describes the structural distribution of the population in year n while $F : X_+ \to X_+$ formulates the rule how the structural distribution in a given year follows from the structural distribution of the previous year. The norm $\|x_n\|$ is some measure of the population size in year n. F is called the *(yearly) population turnover operator*. The condition $F(0) = 0$ means that the population is closed: If there is no population this year, then there is no population next year. We use the notation

$$\dot{X}_+ = X_+ \setminus \{0\}. \tag{2}$$

Notice that (1) is solved by

$$x_n = F^n(x_0), \qquad n \in \mathbb{N}, \tag{3}$$

where F^n is the n-fold composition or iterate of the operator F and $\{F^n; n \in \mathbb{N}\}$ is the discrete semiflow on X_+ induced by the map F [26, Sect. 1.2].

Since this paper relies more heavily on dynamical systems theory than its prequel [32], we will rather use the iterates F^n than solutions of (1) to formulate our results.

A fundamental question is as to whether the population always dies out, $\|F^n(x_0)\| \to 0$ as $n \to \infty$ for all $x_0 \in X_+$, or whether it persists uniformly [26, 33]:

> There is some $\epsilon > 0$ such that for all $x_0 \in \dot{X}_+$ there is some $N \in \mathbb{N}$ such that $\|F(x_0)\| \geq \epsilon$ for all $n \geq N$ (with ϵ not depending on x_0).

In addressing this question, we assume that X_+ is *a (positively) homogeneous subset* of X:

> If $x \in X_+$ and $\alpha \in \mathbb{R}_+$, then $\alpha x \in X_+$.

We assume that F is directionally differentiable at $0 = F(0)$, i.e., that all directional differentials

$$B(x) = \partial F(0, x) = \lim_{\mathbb{R}_+ \ni b \to 0} \frac{1}{b} F(bx), \qquad x \in X_+, \tag{4}$$

exist. It is easy to see that the directional derivative $B : X_+ \to X_+$ at 0 is (positively) *homogeneous* (of degree one) [20, Theorem 3.1]:

> If $x \in X_+$ and $\alpha \in \mathbb{R}_+$, then $B(\alpha x) = \alpha B(x)$.

Since we rarely consider homogeneity in a different sense, B with this property is simply called *homogeneous*. B is a *first order approximation of F at 0* in a weak sense, and we will need B to be a first order approximation in a stronger sense [20, Sect. 3] with which we do not want to burden the reader quite yet. We call B the

basic population turnover operator because it approximates the turnover operator at low population densities.

The operator norm of a homogenous operator $B : X_+ \to X_+$ is defined as

$$\|B\| := \sup\{\|B(x)\|; x \in X_+, \|x\| \leq 1\}, \tag{5}$$

and B is called *bounded* if this supremum exists.

Lemma 1 *Assume that there are $\delta > 0$ and $c > 0$ such that $F : X_+ \to X_+$ satisfies $\|F(x)\| \leq c\|x\|$ for all $x \in X_+$ with $\|x\| \leq \delta$. Then the directional derivative B of F at 0 is bounded and $\|B\| \leq c$.*

1.1 The Spectral Radius of a Homogeneous Operator

Since B is homogeneous,

$$\|B(x)\| \leq \|B\| \, \|x\|, \qquad x \in X_+, \tag{6}$$

provided that B is bounded. This formula implies that the powers (iterates) $B^n : X_+ \to X_+$ of a homogeneous bounded B are bounded and $\|B^n\| \leq \|B\|^n$ for all $n \in \mathbb{N}$.

The *spectral radius* of a bounded homogeneous $B : X_+ \to X_+$ is defined by the *Gelfand* formula [13]

$$\mathbf{r}(B) = \inf_{n \in \mathbb{N}} \|B^n\|^{1/n} = \lim_{n \to \infty} \|B^n\|^{1/n}. \tag{7}$$

The last equality is shown in the same well-known way as for a bounded linear everywhere-defined map. See [20, 28, 31, 32] for more information.

For restrictions of bounded positive linear operators to a cone, the Gelfand formula for the spectral radius was used by Bonsall [4] under the name "partial spectral radius" and by Nussbaum [24] under the name "cone spectral radius." Mallet-Paret and Nussbaum [21, 22] used the Gelfand formula for homogeneous bounded operators on a cone under the name "Bonsall cone spectral radius". But since this formula also makes sense on homogeneous sets (which concept includes the vector space), we simply say "spectral radius".

If B has an interpretation as basic population turnover operator, then $\mathbf{r}(B)$ is called the *basic population turnover number* [17, 20, 32].

1.2 Preview of Extinction and Persistence Results

The following results, which highlight the role of the basic turnover number as threshold parameter between population extinction and persistence, hold under additional assumptions, all of which we do not mention here.

It will be not enough to assume that X_+ is a closed homogeneous subset of X; rather X_+ needs to be a closed cone.

A homogeneous subset X_+ of X is called a *cone* if it is also a convex subset of X and if $x = 0$ is the only vector in X such that x and $-x$ are both elements in X_+. A cone is called a closed cone if it is a closed subset of the normed vector space X.

Cones, wedges and ordered vector spaces are studied in this context in [20, 28, 32] to which we refer. Similarly, not much can be done without assuming that B is order-preserving i.e., for all $x, \tilde{x} \in X_+$,

$$x - \tilde{x} \in X_+ \implies B(x) - B(\tilde{x}) \in X_+. \tag{8}$$

- For the rest of this section, let X_+ be the closed cone of the normed vector space X.
- Further, let $B : X_+ \to X_+$ be homogeneous and order-preserving and let B be an appropriate first order approximation of F.

Theorem 1 *Let X_+ be a normal cone, $F, B : X_+ \to X_+$, $r = \mathbf{r}(B) < 1$. Then the extinction state 0 is locally asymptotically stable in the following sense:*

For each $\alpha \in (r, 1)$, there exist some $\delta_0 > 0$ and $M \geq 1$ such that $\|F^n(x)\| \leq M\alpha^n\|x\|$ for all $n \in \mathbb{N}$ and all $x \in X_+$ with $\|x\| \leq \delta_0$.

See [20, Theorem 4.2] for the precise formulation. A rigorously formulated application to a general population model in the state space of measures is given in Theorem 4.

Theorem 2 *Let $F, B : X_+ \to X_+$ and B be compact and continuous, $\mathbf{r}(B) > 1$.*

Then, under appropriate additional assumptions, the population persists uniformly weakly:

There exists some $\epsilon > 0$ such that for all $x \in \dot{X}_+$ and all $m \in \mathbb{N}$ there exists some $n \in \mathbb{N}$ with $n > m$ and that $\|F^n(x)\| \geq \epsilon$.

See [20, Theorem 5.2] for the precise formulation. A rigorously formulated application to a general population model in the state space of measures is given in Theorem 6 and to a more specific model for iteroparous populations in Theorem 25.

Theorem 3 *Let $F, B : X_+ \to X_+$ and B be compact and continuous, $\mathbf{r}(B) > 1$, and*

$$\limsup_{\|x\| \to \infty} \frac{\|F(x)\|}{\|x\|} < 1.$$

Then, under appropriate additional assumptions, the semiflow induced by F has a compact persistence attractor $\mathcal{A}_1 \subseteq X_+$:

(a) $\mathcal{A}_1$ *is a compact set,* $F(\mathcal{A}_1) = \mathcal{A}_1$, *and* $\inf_{x\in\mathcal{A}_1} \|x\| > 0$.

(b) $\mathcal{A}_1$ *attracts all compact subsets* K *of* X_+ *with* $\inf_{x\in K} \|x\| > 0$*:*
If K *is such a subset and* $\mathcal{U}$ *is an open set with* $\mathcal{A}_1 \subseteq \mathcal{U} \subseteq X_+$, *then there exists some* $N \in \mathbb{N}$ *such that* $F^n(\mathcal{K}) \subseteq \mathcal{U}$ *for all* $n \in \mathbb{N}$ *with* $n \geq N$.

Theorem 3 is a consequence of Theorem 2 and of the point-dissipativity Theorem 9 in Sect. 3 and is a special case of [26, Theorem 5.7] to which we refer for the precise assumptions. A rigorously formulated application to a general population model in the state space of measures is given in Theorem 22.

Corollary 1 *Let the assumptions of Theorem 3 be satisfied. Then there is some* $\epsilon_1 > 0$ *such that for any compact subset* $\mathcal{K}$ *of* X_+ *with* $\inf_{x\in\mathcal{K}} \|x\| > 0$ *there is some* $N \in \mathbb{N}$ *such that* $\|F^n(x)\| \geq \epsilon_1$ *for all* $x \in \mathcal{K}$ *and all* $n \in \mathbb{N}$ *with* $n \geq N$.

The theorems above are known if B can be extended to a bounded linear map on X and B is the Frechet derivative of F at 0 [7, 26, 33].

There are at least three motivations to consider the more general situation of a bounded homogenous order-preserving operator. The first is of mathematical nature, namely that the directional derivative is homogeneous but not necessarily linear and that homogenous operators are not Frechet differentiable at 0 unless they are linear [20, Sect. 3].

The second, biological, motivation are two-sex population models which often use homogeneous mating functions resulting in homogeneous first order approximations of the population turnover operator [18–20, 29–31].

The third motivation are structural population distributions which are best described by measures μ on a metric space S (see [1, 2, 32] and the references therein) which is the state space of individual characteristics [8]. A point in S gives an individual's characteristic, and the metric d describes how close the characteristics of two different individuals are to each other. If $\mu : \mathcal{B} \to \mathbb{R}_+$ is a measure on the σ-algebra $\mathcal{B}$ of Borel sets in S, $\mu(T)$ gives the number of individuals whose structural characteristic lies in the Borel subset T of S. This leads to choosing $X = \mathcal{M}(S)$ as population state space, the vector space of real finite Borel measures (or rather an appropriate closed subspace of it if S is not separable). Let $X_+ = \mathcal{M}_+(S)$ denote the cone of nonnegative measures and $\dot{X}_+ = \dot{\mathcal{M}}_+(S)$ be $\mathcal{M}_+(S)$ without the zero measure. The variation norm is too strong to provide the required compactness of the basic turnover operator B on X_+ in Theorem 2 even if B can be extended to a bounded linear operator on X. A suitable alternative is the flat norm aka dual bounded Lipschitz norm (see [14] and the references therein and Sect. 4). The flat norm has the trade off that important linear basic turnover operators defined on all of X are compact and continuous on X_+ but not bounded on X [32].

2 A General Framework for the State Space of Measures

In this paper, we will be guided by the third motivation, a population state space consisting of measures on a metric space S (Sect. 4).

2.1 Feller Kernels

Important building blocks for the turnover map F are Feller kernels $\kappa : \mathcal{B} \times S \to \mathbb{R}_+$ where $\mathcal{B}$ is the σ-algebra of Borel subsets of S ([32] and Sect. 5). In fact, the first order approximation of F at 0 mentioned before will be associated with a *Feller kernel*. As a first requirement,

- $\kappa(\cdot, s)$ is a nonnegative measure on $\mathcal{B}$ for each $s \in S$.

Then, for each $f \in C^b(S)$ (bounded continuous function), we can form the integrals

$$\int_S f(t)\kappa(dt, s) =: (A_* f)(s), \qquad s \in S. \tag{9}$$

As a second requirement,

- κ has the *Feller property*, i.e., definition (9) provides a continuous bounded function $A_* f$,

and a bounded linear map A_* on $C^b(S)$ is associated with κ.

Cf. [3, Sect. 19.3]. See Example 10.12 in [32].

$C^b(S)$, the vector space of bounded continuous real-valued functions, is a Banach space under the supremum norm and has $C^b_+(S)$, the subset of nonnegative functions in $C^b(S)$, as closed convex cone. $\dot{C}^b_+(S)$ denotes this cone without the zero function.

By [32, Proposition 6.3], if κ is a Feller kernel, $\kappa(U, \cdot)$ is a Borel measurable function on S for all open subsets U of S and thus for all Borel sets U in S. Consequently, A_* can be extended to $M^b(S)$ by (9), the Banach space of bounded Borel measurable functions with the supremum norm.

For each $\mu \in \mathcal{M}(S)$, we can define

$$\int_S \kappa(T, s)\mu(ds) = (A\mu)(T), \qquad T \in \mathcal{B}, \tag{10}$$

and obtain a measure $A\mu$ and a linear map on $\mathcal{M}(S)$ and the duality relation

$$\int_S (A_* f)\, d\mu = \int_S f\, d(A\mu), \qquad f \in M^b(S), \quad \mu \in \mathcal{M}(S). \tag{11}$$

The linear operator A on $\mathcal{M}(S)$ is bounded with respect to the variation norm,

$$\|A\| = \sup_{s\in S} \kappa(S, s) = \|A_*\|, \tag{12}$$

but not necessarily bounded with respect to the flat norm [32, Sect. 9, 10].

In some probabilistic applications, it is assumed that κ is also a Markov kernel, i.e., $\kappa(S, s) = 1$ for all $s \in S$. Then $\kappa(T, s)$ can be interpreted as the probability that an individual with characteristic $s \in S$ will have a characteristic within the set T after one year. This ignores that the individual may die during the year on the one hand or have offspring on the other hand.

So, we do not assume that κ is a Markov kernel, and $\kappa(T, s)$ is rather interpreted as follows: For an individual with characteristic feature $s \in S$, $\kappa(T, s)$ is the sum of the probability that, after one year, the individual is still alive and has its characteristic feature within the set T and of the amount of its surviving offspring that has also characteristic feature within the set T. For more on Feller kernels see Sect. 5.

Definition 1 A Feller kernel $\kappa : \mathcal{B} \times S \to \mathbb{R}_+$ is called a *uniform Feller kernel* if

$$\sup_{T\in\mathcal{B}} |\kappa(T, t) - \kappa(T, s)| \to 0, \qquad t \to s, \text{ for all } s \in S. \tag{13}$$

Equivalent characterizations of uniform Feller kernels are given in Proposition 9, in particular (13) implies the Feller property above. For more on uniform Feller kernels see Sect. 5.2.

2.1.1 Convolutions and Spectral Radius of Feller Kernels

The *convolution* of two Feller kernels $\kappa_j : \mathcal{B} \times S \to \mathbb{R}_+$, $j = 1, 2$, is defined by

$$(\kappa_1 \star \kappa_2)(T, s) = \int_S \kappa_1(T, t)\kappa_2(dt, s), \qquad T \in \mathcal{B}, s \in S. \tag{14}$$

$\kappa_1 \star \kappa_2$ is again a Feller kernel.

Definition 2 Let $\kappa : \mathcal{B} \times S \to \mathbb{R}_+$ be a Feller kernel. We inductively define the *multiple convolution kernels* $\kappa^{n\star}$ by $\kappa^{1\star} = \kappa$ and $\kappa^{(n+1)\star} = \kappa^{n\star} \star \kappa$.

The *spectral radius* of the Feller kernel κ is defined by

$$\mathbf{r}(\kappa) = \inf_{n\in\mathbb{N}} \left(\sup_{s\in S} \kappa^{n\star}(S, s) \right)^{1/n}. \tag{15}$$

If A_* is the map on $C^b(S)$ or on $M^b(S)$ induced by κ, then A_*^n is induced by $\kappa^{n\star}$. This implies that $\mathbf{r}(\kappa) = \mathbf{r}(A_*)$, and so, in (15), $\inf_{n\in\mathbb{N}}$ can be replaced by $\lim_{n\to\infty}$ because of (7). See [32, Sect. 9] for more details.

2.1.2 Irreducible Feller Kernels

Since a Feller kernel $\kappa : \mathcal{B} \times S \to \mathbb{R}_+$ induces the positive bounded linear map A_* on the Banach lattice $C^b(S)$ with the supremum norm, irreducibility of κ could be defined as irreducibility of A_* like in [25, III.8]. However, the following weaker irreducibility concept seems to be better tailored to a Feller kernel.

Definition 3 ([27]) A Feller kernel $\kappa : \mathcal{B} \times S \to \mathbb{R}_+$ is called *top-irreducible* (short for "topologically irreducible") if for any nonempty open subset U of S and for any $s \in S \setminus U$ there is some $n \in \mathbb{N}$ such that $\kappa^{n\star}(U, s) > 0$.

We will also use the following stronger concept.

Definition 4 A Feller kernel $\kappa : \mathcal{B} \times S \to \mathbb{R}_+$ is called *strongly top-irreducible* if for any nonempty open subset U of S and any nonempty compact subset K of S there exists some $n \in \mathbb{N}$ such that $\kappa^{n\star}(U, s) > 0$ for all $s \in K$.

For more on (strongly) irreducible Feller kernels see Sect. 5.3.

2.2 Turnover Maps on the State Space of Measures

We consider yearly turnover maps F of the following general form,

$$F(\mu)(T) = \int_S \kappa^\mu(T, s)\, \mu(ds), \qquad \mu \in \mathcal{M}_+(S), \qquad T \in \mathcal{B}, \tag{16}$$

where $\{\kappa^\mu;\ \mu \in \mathcal{M}_+(S)\}$ is a family of Feller kernels $\kappa^\mu : \mathcal{B} \times S \to \mathbb{R}_+$.

The interpretation of κ^μ is as before except that individual survival, development and reproduction play out in the environment being effected by the structural distribution μ of the population.

If μ is the zero measure, we use the notation κ^o. Often, the operator A associated with κ^o by (10) will turn out to be the first order approximation of F at the zero measure.

Finally, we emphasize that, while individual survival, development, and reproduction are modeled stochastically through the family of Feller kernels, the population model is completely deterministic.

A more specific model for a semelparous population can be found in [32, Sect. 2 and 12] and for an iteroparous population in Sect. 7.

Assumption 5 For each $\mu \in \mathcal{M}_+(S)$, κ^μ is a Feller kernel and $\left\{\kappa^\mu(S, t);\ \mu \in \mathcal{M}_+(S), t \in S\right\}$ is a bounded subset of $\mathbb{R}$.

Standard measure-theoretic arguments imply the following result.

Proposition 1 *Let the Assumption 5 be satisfied. Then F maps $\mathcal{M}_+(S)$ into itself.*

Definition 6 The kernel family $\{\kappa^\mu;\, \mu \in \mathcal{M}_+(S)\}$ is called *upper semicontinuous at the zero measure* if for any $\epsilon \in (0, 1)$ there is some $\delta > 0$ such that

$$\kappa^\mu(T, s) \le (1 + \epsilon)\kappa^o(T, s), \qquad T \in \mathcal{B}, s \in S,$$

for all $\mu \in \mathcal{M}_+(S)$ with $\mu(S) \le \delta$.

The kernel family $\{\kappa^\mu;\, \mu \in \mathcal{M}_+(S)\}$ is called *lower semicontinuous at the zero measure* if for any $\epsilon \in (0, 1)$ there is some $\delta > 0$ such that

$$\kappa^\mu(T, s) \ge (1 - \epsilon)\kappa^o(T, s), \qquad T \in \mathcal{B}, s \in S,$$

for all $\mu \in \mathcal{M}_+(S)$ with $\mu(S) \le \delta$.

The kernel family $\{\kappa^\mu;\, \mu \in \mathcal{M}_+(S)\}$ is called *continuous at the zero measure* if for any $\epsilon \in (0, 1)$ there is some $\delta > 0$ such that

$$(1 - \epsilon)\kappa^o(T, s) \le \kappa^\mu(T, s) \le (1 + \epsilon)\kappa^o(T, s), \qquad T \in \mathcal{B}, s \in S,$$

for all $\mu \in \mathcal{M}_+(S)$ with $\mu(S) \le \delta$.

In a preview of results, we will showcase the spectral radius of the basic turnover kernel κ^o as a crucial threshold parameter between local stability (in the subthreshold case $\mathbf{r}(\kappa^o) < 1$) and instability (in the superthreshold case $\mathbf{r}(\kappa^o) > 1$) of the extinction state represented by the zero measure; $\mathbf{r}(\kappa^o)$ is called the *basic population turnover number*. For a semelparous population, as it is considered in [32, Sect. 12], the *basic turnover number* coincides with the *basic reproduction number*.

2.3 Local (Global) Stability of the Zero Measure in the Subthreshold Case

For perspective, we cite the following result [32, Theorem 3.6].

Theorem 4 *Make Assumption 5 and let the kernel family* $\{\kappa^\mu;\, \mu \in \mathcal{M}_+(S)\}$ *be upper semicontinuous at the zero measure.*
(a) If $r = \mathbf{r}(\kappa^o) < 1$*, the zero measure (the extinction state) is locally asymptotically stable in the following sense:*

For each $\alpha \in (r, 1)$*, there exist some* $\delta_\alpha > 0$ *and* $M_\alpha \ge 1$ *such that,*

$$F^n(\mu)(S) \le \alpha^n M_\alpha\, \mu(S), \qquad n \in \mathbb{N},$$

if $\mu \in \mathcal{M}_+(S)$ *with* $\mu(S) \le \delta_\alpha$.

(b) If $r = \mathbf{r}(\kappa^o) < 1$ *and* $\kappa^\mu(T, s) \le \kappa^o(T, s)$ *for all* $T \in \mathcal{B}$, $s \in S$, *the zero measure is globally stable in the following sense:*

For each $\alpha \in (r, 1)$, there exists some $M_\alpha \geq 1$ such that

$$F^n(\mu)(S) \leq \alpha^n M_\alpha \, \mu(S), \qquad n \in \mathbb{N}, \quad \mu \in \mathcal{M}_+(S).$$

Recall that $F^n(\mu)(S)$ is the total population size in the nth year and $\mu(S)$ the population size at the beginning.

2.4 Instability of the Zero Measure in the Superthreshold Case

We consider the following concepts [10, 14, 16, 32].

Definition 7 Consider a subset $\mathcal{N}$ of $\mathcal{M}_+(S)$.

- $\mathcal{N}$ is called *tight* if for any $\epsilon > 0$ there exists a compact subset K of S such that $\mu(S \setminus K) < \epsilon$ for all $\mu \in \mathcal{N}$.
- A single measure $\mu \in \mathcal{M}_+(S)$ is called *tight*, and we write $\mu \in \mathcal{M}_+^t(S)$, if $\{\mu\}$ is tight.
- $\mathcal{N}$ is called *pre-tight* if for any $\epsilon > 0$ there exists a closed totally bounded subset T of S such that $\mu(S \setminus T) < \epsilon$ for all $\mu \in \mathcal{N}$.
- A single measure $\mu \in \mathcal{M}_+(S)$ is called *separable*, and we write $\mu \in \mathcal{M}_+^s(S)$, if there exists a countable subset T of S such that $\mu(S \setminus \bar{T}) = 0$.
- A single measure $\mu \in \mathcal{M}(S)$ is called *separable*, and we write $\mu \in \mathcal{M}^s(S)$, if its absolute value $|\mu|$ is separable.

By definition, a subset T of S is *totally bounded* if for any $\epsilon > 0$ there exists a finite subset K of T such that $T \subseteq \bigcup_{s \in K} U_\epsilon(s)$. Here $U_\epsilon(s) = \{t \in S; d(t, s) < \epsilon\}$ is the open neighborhood with center s and radius ϵ. $T \subseteq S$ is compact if and only if T is totally bounded and complete [3, Sect. 3.7].

If S is a compact metric space, $\mathcal{M}_+(S)$ is trivially tight. If S is a separable metric space, $\mathcal{M}_+(S) = \mathcal{M}_+^s(S)$.

Definition 8 A Feller kernel κ is called a *tight Feller kernel* if $\{\kappa(\cdot, s); \, s \in S\}$ is a tight set of measures.

A Feller kernel κ is called a *Feller kernel of separable measures* if all measures $\kappa(\cdot, s)$, $s \in S$, are separable.

The condition $\mathbf{r}(\kappa^o) < 1$ in Theorem 4 is almost sharp as seen from the next result ([32, Theorem 3.13] with switched roles of κ_1 and κ_2).

Theorem 5 *Make Assumption 5 and let the kernel family $\{\kappa^\mu; \, \mu \in \mathcal{M}_+(S)\}$ be lower semicontinuous at the zero measure.*

Assume that $\kappa^o = \kappa_1 + \kappa_2$ with two Feller kernels κ_j of separable measures and assume that κ_2 is a tight kernel and $r := \mathbf{r}(\kappa^o) > 1 \geq \mathbf{r}(\kappa_1)$.

Then there exists some eigenmeasure $\nu \in \mathcal{M}_+^s(S)$, $\nu(S) = 1$, *such that*

$$r\nu(T) = \int_S \kappa^o(T, s)\nu(ds), \qquad T \in \mathcal{B}.$$

Further, the zero measure is unstable: There is some $\delta_0 > 0$ *such for any* ν*-positive* $\mu \in \mathcal{M}_+(S)$ *there is some* $n \in \mathbb{Z}_+$ *with* $F^n(\mu)(S) \geq \delta_0$.

A measure $\mu \in \mathcal{M}_+(S)$ is called ν-positive if there exists some $\delta > 0$ such that $\mu(T) \geq \delta\nu(T)$ for all $T \in \mathcal{B}$.

In an iteroparous population, as we will consider it in Sect. 7, the Feller kernel κ_1 may be associated with adult survival and adult development and the Feller kernel κ_2 with reproduction and first year development. If $\mathbf{r}(\kappa_1) < 1$, $\kappa_1^\infty = \sum_{n=1}^\infty \kappa_1^{n\star}$ is a Feller kernel, and the Feller kernel

$$\kappa_2 + \kappa_2 \star \kappa_1^\infty = \kappa_2 + \sum_{n=1}^\infty \kappa_2 \star \kappa_1^{n\star}$$

can be interpreted as *next generation kernel* and its spectral radius as *basic* [9] (or *inherent net* [5, 6]) *reproduction number*. We again like to think of $\kappa^o = \kappa_1 + \kappa_2$ as *basic population turnover kernel* and its spectral radius as *basic turnover number*; this spectral radius has also been called *inherent population growth rate* [6].

Remark 1 Let $\mathbf{r}(\kappa_1) < 1$. The following trichotomy holds:

- $\mathbf{r}(\kappa_2 + \kappa_2 \star \kappa_1^\infty) > 1$ and $\mathbf{r}(\kappa_1 + \kappa_2) > 1$
  or
- $\mathbf{r}(\kappa_2 + \kappa_2 \star \kappa_1^\infty) = 1$ and $\mathbf{r}(\kappa_1 + \kappa_2) = 1$
  or
- $\mathbf{r}(\kappa_2 + \kappa_2 \star \kappa_1^\infty) < 1$ and $\mathbf{r}(\kappa_1 + \kappa_2) < 1$.

See [32, Remark 3.14, Theorem 7.16], but notice that the roles of κ_1 and κ_2 have been switched.

2.5 Persistence of the Population in the Superthreshold Case

We now give a preview of this paper's main results in the general framework for the population state space of measures. The proofs can be found in Sect. 6.

Assumption 9 For each $\mu \in \mathcal{M}_+^s(S)$, κ^μ is a Feller kernel of separable measures and $\left\{\kappa^\mu(S, t);\ \mu \in \mathcal{M}_+^s(S), t \in S\right\}$ is a bounded subset of $\mathbb{R}$.

Assumption 10 For any $\mu \in \dot{\mathcal{M}}_+^s(S)$, $\kappa^\mu(S, s) > 0$ for all $s \in S$.

Recall that $\dot{\mathcal{M}}_+^s(S)$ is the set of nonnegative separable measures without the zero measure.

Theorem 6 *Assume Assumptions 9 and 10. Let the kernel family $\{\kappa^\mu;\ \mu \in \mathcal{M}_+^s(S)\}$ be lower semicontinuous at the zero measure.*

Assume that κ^o is a top-irreducible Feller kernel and $\kappa^o = \kappa_1 + \kappa_2$ with two tight Feller kernels κ_j, where κ_2 is a uniform Feller kernel.

Finally, assume $r = \mathbf{r}(\kappa^o) > 1 \geq \mathbf{r}(\kappa_1)$.

Then the semiflow induced by F is uniformly weakly persistent: There exists some $\delta > 0$ such that $\limsup_{n\to\infty} F^n(\mu)(S) \geq \delta$ for all $\mu \in \dot{\mathcal{M}}_+^s(S)$.

The next assumption looks rather technical, but is often satisfied; the technicality is the prize we pay for the generality of the framework. We will derive from it that F is continuous on $\mathcal{M}_+^s(S)$ with respect to the flat norm.

Assumption 11 If $\mu \in \mathcal{M}_+^s(S)$ and (μ_n) is a sequence in $\mathcal{M}_+^s(S)$ such that $\int_S f d\mu_n \to \int_S f d\mu_n$ for all $f \in C_+^b(S)$, then

$$\int_S h(t)\, \kappa^{\mu_n}(dt, s) \overset{n\to\infty}{\longrightarrow} \int_S h(t)\, \kappa^{\mu}(dt, s) \tag{17}$$

uniformly for s in every closed totally bounded subset of S, for all $h \in \mathcal{L}$,

$$\mathcal{L} = \left\{h \in [0,1]^S;\ \forall t, \tilde{t} \in S:\ |h(t) - h(\tilde{t})| \leq d(t, \tilde{t})\right\}. \tag{18}$$

Assumption 12 If $\mathcal{N}$ is a bounded subset of $\mathcal{M}_+^s(S)$, then the set of measures $\{\kappa^\mu(\cdot, s);\ s \in S,\ \mu \in \mathcal{N}\}$ is tight and the set $\{\kappa^\mu(S, s);\ s \in S,\ \mu \in \mathcal{N}\}$ is bounded in $\mathbb{R}$.

This assumption will imply that F is compact on $\mathcal{M}_+^s(S)$ with respect to the flat norm.

Assumption 13 $\limsup_{\mu(S)\to\infty} \sup_{s\in S} \kappa^\mu(S, s) < 1$.

This assumption will allow us to use the abstract point-dissipativity result in the upcoming Sect. 3.

Theorem 7 *Make Assumptions 9, 10, 11, 12, 13 and let the kernel family $\{\kappa^\mu;\ \mu \in \mathcal{M}_+^s(S)\}$ be lower semicontinuous at the zero measure.*

Assume that κ^o is a top-irreducible Feller kernel and $\kappa^o = \kappa_1 + \kappa_2$ with two tight Feller kernels κ_j where κ_2 is a uniform Feller kernel.

Finally, assume $r = \mathbf{r}(\kappa^o) > 1 > \mathbf{r}(\kappa_1)$.

Then the semiflow induced by F is uniformly persistent: There exists some $\delta > 0$ such that $\liminf_{n\to\infty} F^n(\mu)(S) \geq \delta$ for all $\mu \in \dot{\mathcal{M}}_+^s(S)$.

To obtain uniform persistence in a stronger sense, we will assume the following.

Assumption 14 If $\mathcal{N}$ is a tight bounded subset of $\mathcal{M}_+(S)$, then there exists a strongly top-irreducible Feller kernel $\tilde{\kappa}$ such that

$$\kappa^{\mu}(T,s) \geq \tilde{\kappa}(T,s), \qquad T \in \mathcal{B}, s \in S, \mu \in \mathcal{N}.$$

Theorem 8 *Make Assumptions 9, 10, 11, 12, 13, 14 and let the kernel family* $\{\kappa^{\mu}; \mu \in \mathcal{M}_+^s(S)\}$ *be lower semicontinuous at the zero measure.*

Assume that κ^o *is a strongly top-irreducible Feller kernel and* $\kappa^o = \kappa_1 + \kappa_2$ *with two tight Feller kernels* κ_j*, where* κ_2 *is a uniform Feller kernel. Finally, assume* $r = \mathbf{r}(\kappa^o) > 1 \geq \mathbf{r}(\kappa_1)$.

Then the semiflow induced by F *is uniformly persistent in the following sense: For each* $f \in \dot{C}_+^b(S)$*, there exists some* $\epsilon_f > 0$ *with the following property:*

If $\mathcal{N}$ *is a bounded tight subset of* $\mathcal{M}_+^s(S)$ *with* $\inf_{\mu\in\mathcal{N}} \mu(S) > 0$*, there exists some* $N \in \mathbb{N}$ *such that*

$$\int_S f \, d\, F^n(\mu) \geq \epsilon_f \quad \text{for all } \mu \in \mathcal{N} \text{ and all } n \in \mathbb{N} \text{ with } n > N.$$

3 An Abstract Point-Dissipativity Result

The next abstract result will be used in proving Theorem 8.

Theorem 9 *Let* X_+ *be the closed cone of an ordered normed vector space* X*. Let* $F : X_+ \to X_+$ *map bounded subsets of* X_+ *into bounded subsets of* X_+*. Let* $\theta : X_+ \to \mathbb{R}_+$ *be homogeneous, subadditive, continuous and uniformly positive (there is some* $\epsilon > 0$ *such that* $\epsilon\|x\| \leq \theta(x)$ *for all* $x \in X_+$*). Assume that*

$$\limsup_{\|x\|\to\infty} \frac{\theta(F(x))}{\theta(x)} < 1. \tag{19}$$

Then, for any bounded subset B *of* X_+*, there exists a bounded convex subset* $\tilde{B}$ *of* X_+ *such that* $F^n(B) \subseteq \tilde{B}$ *for all* $n \in \mathbb{N}$*. Further, there exists a bounded convex subset* B_0 *of* X_+ *such that for each* $x \in X_+$ *there exists some* $m \in \mathbb{N}$ *such that* $F^n(x) \in B_0$ *for all* $n \geq m$*. If* F *is continuous and compact, the semiflow induced by* F *has a compact attractor of bounded sets [26, Sect. 2.2.3].*

Proof Cf. [26, L.7.1]. By (19) and the other properties of θ, there exists some $\xi \in (0, 1)$ and $R_1 > 0$ such that

$$\theta(F(x)) \leq \xi\theta(x), \qquad x \in X_+, \theta(x) \geq R_1. \tag{20}$$

We claim that there exists some $R_2 > 0$ such that, for all $x \in X_+$,

$$\theta(x) \leq R_2 \implies \theta(F(x)) \leq R_2. \tag{21}$$

If not, for any $n \in \mathbb{N}$, there exists some $x_n \in X_+$ such that $\theta(x_n) \le n < \theta(F(x_n))$. Since F maps bounded sets in X_+ into bounded sets of X_+ and θ is bounded, $\theta(x_n) \to \infty$ as $n \to \infty$. This leads to a contradiction for n large enough such that $\theta(x_n) \ge R_1$:

$$n < \theta(F(x_n)) \le \xi\theta(x_n) < n.$$

Let $R_3 = \max\{R_1, R_2\}$. Let $R \ge R_3$ and $B_R^+ = \{x \in X_+; \theta(x) \le R\}$. Since θ is convex and continuous, B_R^+ is convex and closed. Since θ is uniformly positive, B_R^+ is bounded. By (21), $F(B_R^+) \subseteq B_R^+$. Let B be a bounded subset of X_+. Then there exists some $R > R_3$ such that $B \subseteq B_R^+$ and $F^n(B) \subseteq B_R^+$ for all $n \in \mathbb{N}$. Let $x \in X_+$. If $\|x\| \le R_3$, $\limsup_{n\to\infty} \theta(F^n(x)) \le R_3$. If $\theta(x) > R_3$, by (20), $\theta(F^{n+1}(x)) \le \xi\theta(F^n(x))$ as long as $\theta(F^n(x)) \ge R_3$. So $\theta(F^n(x)) \le R_3$ for some $m \in \mathbb{N}$ and $\limsup_{n\to\infty} \theta(F^n(x)) \le R_3$ as well. Since θ is uniformly positive, there exists some $c > 0$ such that $\limsup_{n\to\infty} \|F^n(x)\| \le c$ for all $x \in X_+$.

In the language of [26, Definition 2.25], we have shown that the semiflow induced by F is point-dissipative and eventually bounded on every bounded set. If F is also continuous and the semiflow is asymptotically smooth (in particular if F is compact). then the semiflow has a compact attractor of bounded set by [26, Theorem 2.30]. □

4 The Ordered Vector Space of Real Measures

Let S be a nonempty set, $\mathcal{B}$ a σ-algebra on S, and $\mathcal{M}(S)$ denote the set of real measures on $\mathcal{B}$.

$\mathcal{M}(S)$ becomes a real vector space by the definitions $(\mu + \nu)(T) = \mu(T) + \nu(T)$ and $(\alpha\mu)(T) = \alpha\mu(T)$ where $T \in \mathcal{B}$ and $\alpha \in \mathbb{R}$ and $\mu, \nu \in \mathcal{M}(S)$.

$\mathcal{M}(S)$ contains the cone of all nonnegative measures, $\mathcal{M}_+(S)$ (a convex homogeneous set). $\mathcal{M}(S)$ is an order-complete vector lattice: Each subset $\mathcal{N}$ of $\mathcal{M}(S)$ which has an lower (upper) bound has an infimum (supremum).

The absolute value $|\mu|$ of a measure (in this context also called the variation of the measure) is given by

$$\begin{aligned} |\mu|(T) &= \sup\{\mu(U) - \mu(T \setminus U); \mathcal{B} \ni U \subseteq T\} \\ &= \sup\{|\mu(U)| + |\mu(T \setminus U)|; \mathcal{B} \ni U \subseteq T\} = \sup\Big\{\sum_{j=1}^{n} |\mu(T_j)|\Big\}, \end{aligned} \tag{22}$$

where the supremum is taken over all $n \in \mathbb{N}$ and subsets $\{T_1, \ldots, T_n\}$ of $\mathcal{B}$ such that T is its disjoint union [3, Corollary 10.54 and Theorem 10.56].

4.1 Measures Under the Variation Norm and the Flat Norm

The *variation norm* (also called *total variation*) on $\mathcal{M}(S)$ is defined by

$$\|\mu\|_\sharp = |\mu|(S), \qquad \mu \in \mathcal{M}(S), \tag{23}$$

where $|\mu|$ is the absolute value of μ defined by (22).

If $\mu \in \mathcal{M}_+(S)$, $\|\mu\|_\sharp = \mu(S)$. So the variation norm is additive and order-preserving on $\mathcal{M}_+(S)$, and $\mathcal{M}_+(S)$ is a normal cone. The variation norm makes $\mathcal{M}(S)$ a Banach lattice; in particular, $\mathcal{M}_+(S)$ is a non-flat generating cone: Every real-valued measure μ can be written as the difference of its positive and negative variation, $\mu = \mu_+ - \mu_-$, and $\|\mu_\pm\|_\sharp \le \|\mu\|_\sharp$.

The variation norm is equivalent to the supremum norm

$$\|\mu\|_\infty = \sup_{T \in \mathcal{B}} |\mu(T)|, \qquad \mu \in \mathcal{M}(S), \tag{24}$$

and the two norms are equal on $\mathcal{M}_+(S)$.

Let (S, d) be a metric space. $\mathcal{B}$ now denotes the *Borel* σ-algebra of S which is the smallest σ-algebra that contains all open and closed sets. The sets in the Borel σ-algebra are called *Borel sets*. In a metric space, the Borel σ-algebra is also the smallest σ-algebra for which all (bounded) continuous functions are continuous [11, Theorem 7.1.1]. This second σ-algebra is often [11] but not always [3] called the *Baire-σ-algebra*.

The following is a summary of results needed later. For more details, we refer to [14]. Many of the results can already been found in [10, 11]. See also [15, 16].

For perspective, we present the following result for the variation norm.

Theorem 10 *For all $\mu \in \mathcal{M}(S)$,*

$$\|\mu\|_\sharp = |\mu|(S) = \sup\left\{\left|\int_S f d\mu\right|;\ f \in C^b(S),\ \|f\|_\infty \le 1\right\}.$$

Proof By [12, IV.6.2], $\mu \mapsto \theta$ with $\theta(f) = \int_S f d\mu$, $f \in C^b(S)$, is an isometric isomorphism between the Banach space of regular additive set functions with the variation norm and the dual space of $C^b(S)$. The assertion now follows because every real measure on $\mathcal{B}$ is regular [3, Theorem 12.5]. □

We introduce the following functional on $\mathcal{M}(S)$,

$$\begin{aligned} \|\mu\|_\flat &= \sup_{f \in \mathcal{L}} \left|\int_S f d\mu\right|, \\ \mathcal{L} &= \left\{f \in [0,1]^S;\ \forall_{x,y \in S} |f(x) - f(y)| \le d(x,y)\right\}. \end{aligned} \tag{25}$$

Recall that M^S denotes the set of functions from S to a set M. $\|\cdot\|_\flat$ is a norm on $\mathcal{M}(S)$ [14], which we call the *flat norm*, and

$$\|\mu\|_\flat \leq \|\mu\|_\sharp, \qquad \mu \in \mathcal{M}(S). \tag{26}$$

In the literature, definitions different from (25) are used that lead to equivalent norms. For instance, $[0, 1]^S$ is replaced by $[-1, 1]^S$. Also different names are used for the flat norm or its equivalent definitions. For details see [14].

All the definitions have in common that

$$\|\mu\|_\flat = \mu(S) = \|\mu\|_\sharp, \qquad \mu \in \mathcal{M}_+(S). \tag{27}$$

This implies that the flat norm is additive and order-preserving on $\mathcal{M}_+(S)$.

In the following, all topological notions concerning $\mathcal{M}(S)$ and $\mathcal{M}_+(S)$ are meant with respect to the flat norm unless it is explicitly said otherwise.

Theorem 11 *$\mathcal{M}_+(S)$ is a generating, normal, closed cone.*

Lemma 2 *For $x \in S$, let δ_x denote the Dirac measure at x. Then $1 = \|\delta_x\|_\flat$ and, for $y, x \in S$,*

$$\|\delta_x - \delta_y\|_\flat = \min\{1, d(x, y)\}.$$

Corollary 2 ([16]) *If S is not uniformly discrete (i.e., its metric is not equivalent to the discrete metric), then the ordered normed vector space $\mathcal{M}(S)$ is not complete.*

4.1.1 Convergence in $\mathcal{M}_+(S)$

Definition 15 Let $\mathcal{F}$ be a set of functions $f : S \to \mathbb{R}$ and $s \in S$. $\mathcal{F}$ is called *equicontinuous* at s if for any $\epsilon > 0$ there exists some $\delta > 0$ such that $|f(t) - f(s)| < \epsilon$ for all $f \in \mathcal{F}$ and all $t \in S$ with $d(t, s) < \delta$. $\mathcal{F}$ is called equicontinuous on S if it is equicontinuous at all $s \in S$.

$\mathcal{F}$ is called *uniformly equicontinuous* on $\tilde{S} \subseteq S$ if for any $\epsilon > 0$ there is some $\delta > 0$ such that $|f(t) - f(s)| < \epsilon$ for all $f \in \mathcal{F}$ and all $s, t \in \tilde{S}$ with $d(t, s) < \delta$.

$\mathcal{F}$ is called *equibounded* if there exists some $c > 0$ such that $|f(s)| \leq c$ for all $s \in S$ and all $f \in \mathcal{F}$.

The following is proved in [32, Proposition 6.10].

Proposition 2 *Let $\mathcal{F}$ be an equicontinuous and equibounded family of functions $f : S \to \mathbb{R}_+$ and $\mu \in \mathcal{M}(S)$ and (μ_n) be a sequence in $\mathcal{M}_+(S)$ such that $\|\mu_n - \mu\|_\flat \to 0$ as $n \to \infty$. Then $\int_S f\, d\mu_n \to \int_S f\, d\mu$ as $n \to \infty$ uniformly for $f \in \mathcal{F}$.*

Recall the definition of a (pre-)tight set of measures (Definition 7).

To show that pre-tightness does not change under topologically equivalent metrics, we note the following.

Proposition 3 *$\mu \in \mathcal{M}_+(S)$ is separable if and only if it is pre-tight.*

Proposition 4 *The closure of a tight set of nonnegative measures is tight. The closure of a pre-tight set of nonnegative measures is pre-tight.*

The closure of a set of separable nonnegative measures consists of separable measures.

Proof The first two statements follow from [14, Theorem 4.10d]. The third statement holds because a countable union of countable sets is countable.

Corollary 3 *$\mathcal{M}_+^s(S)$ is a closed cone of $\mathcal{M}(S)$.*

Here is the characterization of convergence.

Theorem 12 *Let (μ_n) in $\mathcal{M}_+(S)$ and $\mu \in \mathcal{M}_+^s(S)$. Equivalent are*

(i) $\|\mu_n - \mu\|_\flat \to 0$,
(ii) $\int_S f\, d(\mu_n - \mu) \to 0$ *for all continuous functions* $f \in C^b(S)$,
(iii) $\int_S f\, d(\mu_n - \mu) \to 0$ *for all Lipschitz continuous functions* $f : S \to [0, 1]$.

4.1.2 Compactness and Completeness in $\mathcal{M}_+(S)$

Theorem 13 *Let (μ_n) be a tight sequence in $\mathcal{M}_+(S)$ such that $(\mu_n(S))$ is bounded. Then (μ_n) has a converging subsequence (with the limit measure being tight as well).*

Proposition 5 *Let $\mathcal{N} \subseteq \mathcal{M}_+^s(S)$ be a totally bounded set of pre-tight measures. Then $\mathcal{N}$ is pre-tight and, if S is complete, tight.*

Theorem 14 ([16, Theorem 3.8]) *$\mathcal{M}_+^s(S)$ is complete if and only if S is complete.*

5 More on Feller Kernels

Let S be metrizable topological space and $\mathcal{B}$ and the respective Borel σ-algebra.

Definition 16 A function $\kappa : \mathcal{B} \times S \to \mathbb{R}_+$ is called a *Feller kernel* if

$\kappa(\cdot, s) \in \mathcal{M}_+(S)$ for all $s \in S$ and if κ has the *Feller property*
$\int_S f(y)\kappa(dy, \cdot) \in C^b(S)$ for any $f \in C^b(S)$.

A Feller kernel κ is called a *Feller kernel of separable measures* if

$\kappa(\cdot, s) \in \mathcal{M}_+^s(S)$ for all $s \in \tilde{S}$.

Cf. [3, Sect. 19.3] and Sect. 2.1. For examples and details see [32]. Recall that every Feller kernel induces maps $A : \mathcal{M}(S) \to \mathcal{M}(S)$ and $A_* : M^b(S) \to M^b(S)$ with $M^b(S)$ denoting the Banach space of bounded measurable functions with the supremum norm. See (10) and (9). Since κ is a Feller kernel, A_* maps $C^b(S)$ to $C^b(S)$.

The next result is part of [32, Theorem 10.4].

Theorem 15 *Let $\kappa : \mathcal{B} \times S \to \mathbb{R}_+$ be a Feller kernel of separable measures. Then the following hold:*

(a) A maps $\mathcal{M}^s_+(S)$ into $\mathcal{M}^s_+(S)$, and $A : \mathcal{M}^s_+(S) \to \mathcal{M}^s_+(S)$ is continuous with respect to the flat norm.
(b) A maps $\mathcal{M}^s(S)$ into $\mathcal{M}^s(S)$.

Remark 2 Let κ be a Feller kernel of separable measures and A^s denote the restriction of A from $\mathcal{M}^s(S)$ to $\mathcal{M}^s(S)$ and A^s_+ the restriction of A from $\mathcal{M}^s_+(S)$ to $\mathcal{M}^s_+(S)$. Since the Dirac measures are separable, we still have for the operator norms that $\|A^s\| = \|A^s_+\| = \sup_{s\in S} \kappa(S, s)$, (see 12). By (15),

$$\mathbf{r}(A^s) = \mathbf{r}(A) = \mathbf{r}(A^s_+) = \mathbf{r}(\kappa).$$

Remark 3 The map A induced by a Feller kernel via (10) is continuous from $\mathcal{M}_+(S)$ to $\mathcal{M}_+(S)$ with respect to the variation norms even without the Feller type property. But it seems difficult to come up with conditions for A to be compact with respect to the variation norm.

5.1 Tight Feller Kernels

Definition 17 A Feller kernel $\kappa : \mathcal{B} \times S \to \mathbb{R}_+$ is called a *tight Feller kernel* if the set of measures $\{\kappa(\cdot, x); x \in S\}$ is tight.

A Feller kernel κ is called a *pre-tight Feller kernel* if set of measures $\{\kappa(\cdot, x); x \in S\}$ is pre-tight.

See [32, Sect. 10] for the proofs of the following and other results and for examples.

Proposition 6 *Let $\kappa : \mathcal{B} \times S \to \mathbb{R}_+$ be a tight Feller kernel. Then A is continuous and compact from $\mathcal{M}_+(S)$ to $\mathcal{M}_+(S)$ with respect to the flat norm and maps $\mathcal{M}_+(S)$ into $\mathcal{M}^t_+(S)$.*

Proposition 7 *Let $P : \mathcal{B} \times S \to \mathbb{R}_+$ be a tight Feller kernel and $g \in C^b_+(S \times S)$. Then $\tilde{\kappa} : \mathcal{B} \times S \to \mathbb{R}_+$,*

$$\tilde{\kappa}(T, s) = \int_T g(s, t) P(dt, s), \qquad s \in S, T \in \mathcal{B}, \tag{28}$$

is a tight Feller kernel. In particular, $\tilde{\kappa}(S, \cdot) \in C^b(S)$.

5.2 Uniform Feller Kernels

We start from the observation that tight Feller kernels are related to compactness in $C^b(S)$. Recall the concepts of equicontinuity and equiboundedness, Definition 15.

Proposition 8 *Let κ be a tight Feller kernel and Q be an equicontinuous bounded subset of $C^b(S)$. Let A_* be the map on $C^b(S)$ induced by κ via (9). Then $A_*(Q)$ has compact closure in in $C^b(S)$.*

Proof Let (g_n) be a sequence in Q. Since κ is tight, there exists a sequence (K_j) of compact subsets of S such that

$$\sup_{s\in S} \kappa(S \setminus K_j, s) \to 0, \qquad j \to \infty. \tag{29}$$

Set $\tilde{S} = \bigcup_{j\in\mathbb{N}} K_j$. Then $\tilde{S}$ is separable. By a version of the Arzela-Ascoli theorem [23, Theorem 8.5], there exists a subsequence (g_{n_i}) and some $g \in C^b(\tilde{S})$ such that $g_{n_i} \to g$ pointwise on $\tilde{S}$ and uniformly on each K_j. Set $h_n = A_* g_n$ and $h(s) = \int_{\tilde{S}} g(t)\kappa(dt, s)$, $s \in S$. Then $h_n \in C^b(S)$ and $h \in M^b(S)$. For each $s \in S$ and $j, i \in \mathbb{N}$,

$$|h_{n_i}(s) - h(s)| \le \int_{S\setminus K_j} |g_{n_i}(t)|\kappa(dt, s) + \int_{K_j} |g_{n_i}(t) - g(t)|\kappa(dt, s) + \int_{\tilde{S}\setminus K_j} |g(t)|\kappa(dt, s).$$

By our various assumptions, there is some $c > 0$ such that, for all $i, j \in \mathbb{N}$,

$$\|h_{n_i} - h\|_\infty \le 2c \sup_{s\in S} \kappa(S \setminus K_j, s) + c \sup_{t\in K_j} |g_{n_i}(t) - g(t)|.$$

For all $j \in \mathbb{N}$, since $g_{n_i} \to g$ as $i \to \infty$ uniformly on K_j,

$$\limsup_{i\to\infty} \|h_{n_i} - h\|_\infty \le 2c\kappa(S \setminus K_j).$$

By (29), we can take the limit as $i \to \infty$,

$$\limsup_{i\to\infty} \|h_{n_i} - h\|_\infty = 0.$$

This shows $A_*(Q)$ is a compact subset of $C^b(S)$. Since all h_n are continuous, h is continuous as well. □

The preceding result motivates us to look for Feller kernels that are related to equicontinuous sets of functions.

Proposition 9 *Let $\kappa : \mathcal{B} \times S \to \mathbb{R}_+$ be a Feller kernel and A the induced linear map on $\mathcal{M}(S)$ and A_* the induced linear map on $M^b(S)$ via (10) and (9), respectively.*

Then the following are equivalent:

(a) $\sup_{T\in\mathcal{B}} |\kappa(T, t) - \kappa(T, s)| \to 0, \qquad t \to s,$ *for all* $s \in S$.

(b) *If Q is a bounded subset of $M^b(S)$, then $A_*(Q)$ is an equicontinuous and equibounded subset of $C^b(S)$.*
(c) *If Q is a bounded subset of $C^b(S)$, then $A_*(Q)$ is an equicontinuous and equibounded subset of $C^b(S)$.*
(d) *A is continuous from $\mathcal{M}_+(S)$ with the flat norm to $\mathcal{M}_+(S)$ with the variation norm.*

Proof Assume (a). Let $f(t) = \sum_{i=1}^m \alpha_i \chi_{T_i}$ be a measurable function of finitely many values $\alpha_1, \ldots, \alpha_m$, where $T_i \in \mathcal{B}$ are pairwise disjoint. Then

$$\big|A_*f)(t) - A_*(f)(s)\big| \le \|f\|_\infty \sum_{i=1}^m \big|\kappa(T_i, t) - \kappa(T_i, s)\big|.$$

Since $\kappa(\cdot, t) - \kappa(\cdot, s)$ is a real-valued measure and the T_i are pairwise disjoint,

$$\big|A_*f)(t) - A_*(f)(s)\big| \le 2\|f\|_\infty \sup_{T\in\mathcal{B}} \big|\kappa(T, t) - \kappa(T, s)\big|. \tag{30}$$

If $f \in M^b(S)$, f is the uniform limit of a sequence of such finitely-valued measurable functions and (30) holds for $f \in M^b(S)$. This implies that A_* maps $M^b(S)$ into $C^b(S)$. Let Q be a bounded subset of $M^b(S)$. Then $A_*(Q)$ is a bounded subset of $C^b(S)$ and an equicontinuous subset by (a) and (30), and (b) follows.

Obviously, (b) implies (c).

Assume (c). Let (μ_n) be a sequence in $\mathcal{M}_+(S)$ and $\mu \in \mathcal{M}_+(S)$ such that $\|\mu_n - \mu\|_\flat \to 0$ as $n \to \infty$. By (c), $\{A_*f;\, f \in C^b(S), 0 \le f \le 1\}$ is a uniformly equicontinuous and equibounded family of functions from S to $\mathbb{R}_+$. By Proposition 2, $\int_S (A_*f)d\mu_n \to \int_S (A_*f)d\mu$ as $n \to \infty$ uniformly for $f \in C^b(S)$ with $0 \le f \le 1$. Let $f \in C^b(S)$ with $\|f\|_\infty \le 1$. Then $f = f_+ - f_-$ with $0 \le f_\pm \le 1$. So $\int_S (A_*f)d\mu_n \to \int_S (A_*f)d\mu$ uniformly for $f \in C^b(S)$ with $\|f\|_\infty \le 1$. By the duality between A_* and A, (11), $\int_S f d(A\mu_n) \to \int_S f d(A\mu)$ as $n \to \infty$ uniformly for $f \in C^b(S)$ with $\|f\|_\infty \le 1$. Assertion (d) now follows from Theorem 10.

Assume (d). As $t \to s$, $\|\delta_t - \delta_s\|_\flat \to 0$ by Lemma 2 and, by (d), $A\delta_t \to A\delta_s$ in variation norm and $\sup_{T\in\mathcal{B}} |\kappa(T, t) - \kappa(T, s)| \to 0$ by (10).

Definition 18 A Feller kernel $\kappa : \mathcal{B} \times S \to \mathbb{R}_+$ is called a *uniform Feller kernel* if it satisfies property (a) of Proposition 9.

Corollary 4 *If κ is a Feller kernel and the map $A_* : C^b(S) \to C^b(S)$ associated with κ is compact, then κ is a uniform Feller kernel.*

Corollary 5 *Let κ_1 be a Feller kernel on S and κ_2 a uniform Feller kernel on S. Then $\kappa_1 \star \kappa_2$ is a uniform Feller kernel on S.*

Proof Let A_i be the linear maps on $\mathcal{M}_+(S)$ induced by κ_i via (10). By Proposition 9, A_2 continuously maps $\mathcal{M}_+(S)$ with the flat norm into $\mathcal{M}_+(S)$ with the variation norm, while A_1 is a bounded liner map on $\mathcal{M}(S)$ with the variation norm. So A_1A_2

continuously maps $\mathcal{M}_+(S)$ with the flat norm into $\mathcal{M}_+(S)$ with the variation norm. Since A_1A_2 is induced by $\kappa_1 \star \kappa_2$ [32, L.9.2], $\kappa_1 \star \kappa_2$ is a uniform Feller kernel by Proposition 9.

Proposition 10 *Let κ be a uniform Feller kernel. Let $g : S \times S \to \mathbb{R}$ be bounded and $g(s, \cdot)$ be Borel measurable and $g(\cdot, s)$ be continuous on S for every $s \in S$.*

Let $\tilde{\kappa} : \mathcal{B} \times S \to \mathbb{R}$ be given by

$$\tilde{\kappa}(T, s) = \int_T g(s, t)\kappa(dt, s), \qquad s \in S, \quad T \in \mathcal{B}.$$

Then $\tilde{\kappa}$ is a uniform Feller kernel.

Proof Let $s \in S$. Since $g(s, \cdot)$ is Borel measurable and bounded and $\kappa(\cdot, s)$ is a finite non-negative measure, $\tilde{\kappa}(\cdot, s)$ is a finite non-negative measure.

Let (s_n) be a sequence in S and $s \in S$ and $s_n \to s$. Then

$$\begin{aligned}
&\Big| \int_T g(s_n, t)\kappa(dt, s_n) - \int_T g(s, t)\kappa(dt, s) \Big| \\
\le &\Big| \int_T g(s_n, t)\kappa(dt, s_n) - \int_T g(s_n, t)\kappa(dt, s) \Big| \\
&+ \Big| \int_T g(s_n, t)\kappa(dt, s) - \int_T g(s, t)\kappa(dt, s) \Big| \\
\le &\, 2 \sup |g| \sup_{\tilde{T} \in \mathcal{B}} |\kappa(\tilde{T}, s_n) - \kappa(\tilde{T}, s)| + \int_S |g(s_n, t) - g(s, t)| \kappa(dt, s).
\end{aligned}$$

The last integral converges to 0 as $n \to \infty$ by Lebesgue's dominated convergence theorem because $|g(s_n, t) - g(s, t)| \to 0$ as $n \to \infty$ pointwise in $t \in S$ and $|g(s_n, t) - g(s, t)| \le 2 \sup g(S \times S)$ for all $n \in \mathbb{N}$.

Notice that the last expression in the inequality converges to 0 as $n \to \infty$ uniformly for $T \in \mathcal{B}$. □

This trivially provides examples for uniform Feller kernels.

Example 1 Let $\nu \in \mathcal{M}_+(S)$ and $g : S \times S \to \mathbb{R}$ be bounded and $g(s, \cdot)$ be Borel measurable and $g(\cdot, s)$ continuous on S for every $s \in S$.

Let $\kappa : \mathcal{B} \times S \to \mathbb{R}$ be given by

$$\kappa(T, s) = \int_T g(s, t)\nu(dt), \qquad s \in S, T \in \mathcal{B}.$$

Then κ is a uniform Feller kernel.

The class of Feller kernels provided this way can be quite comprehensive.

Example 2 Let S be a separable metric space and $\kappa : \mathcal{B} \times S \to \mathbb{R}_+$ be a uniform Feller kernel.

Choose a countable dense subset $\{s_n; n \in \mathbb{N}\}$ in S.
Define $\nu \in \mathcal{M}_+(S)$.

$$\nu(T) = \sum_{n=1} 2^{-n} \kappa(T, s_n), \qquad T \in \mathcal{B}.$$

Let $T \in \mathcal{B}$ and $\nu(T) = 0$. Then $\kappa(T, s_n) = 0$ for all $n \in \mathbb{N}$. Since $\{s_n; n \in \mathbb{N}\}$ is dense in S and κ is a uniform Feller kernel, $\kappa(T, s) = 0$ for all $s \in S$. By the Radon-Nikodym theorem, for any $s \in S$, there exists a Borel measurable function $g(s, \cdot)$ such that

$$\kappa(T, s) = \int_T g(s, t)\nu(dt), \qquad s \in \mathcal{B}. \tag{31}$$

Since κ is a uniform Feller kernel,

$$\int_S |g(s, t) - g(\tilde{s}, t)|\nu(dt) \to 0, \qquad s \to \tilde{s}. \tag{32}$$

Conversely, any kernel of the form (31) satisfying (32) is a uniform Feller kernel.

Theorem 16 *Let κ_1 be a tight Feller kernel and κ_2 a uniform Feller kernel. Then*

$$(A_{1*}f)(s) = \int_S f(t)\kappa_1(dt, s), \qquad s \in S, f \in M^b(S),$$

defines a bounded positive linear map A_{1*} from $M^b(S)$ to $M^b(S)$ and from $C^b(S)$ to $C^b(S)$, and

$$(A_{2*}f)(s) = \int_S f(t)\kappa_2(dt, s), \qquad s \in S, f \in M^b(S),$$

is a bounded positive linear map A_{2*} from $M^b(S)$ to $C^b(S)$ such that $A_{1*}A_{2*}$ is compact from $M^b(S)$ to $C^b(S)$.

Proof Combine Propositions 8 and 9.

Theorem 17 *Let κ_2 be a uniform Feller kernel that is tight. Let κ_1 be a tight Feller kernel and $\kappa = \kappa_1 + \kappa_2$. Assume that $\mathbf{r}(\kappa) > \mathbf{r}(\kappa_1)$. Then there exists some $f \in \dot{C}^b_+(S)$ such that $\mathbf{r}(\kappa) f(s) = \int_S f(t)\kappa(dt, s)$ for all $s \in S$.*

Proof Let A_{*j} be the operators on $C^b(S)$ associated with κ_j. By Theorem 16, $A_{*1}A_{*2}$ and A^2_{*2} are compact on $C^b(S)$. The assertion now follows from [32, Theorem 7.17].

5.3 Irreducible and Colonization Kernels

Recall the definition of a (strongly) top-irreducible Feller kernel (Sect. 2.1.2).

Lemma 3 *Let $\kappa : \mathcal{B} \times S \to \mathbb{R}_+$ be a Feller kernel and A_* be the bounded linear map on $C^b(S)$ induced by (9). Then the following are equivalent:*

(a) κ is top-irreducible.
(b) For any nonempty open strict subset U of S there exists some $s \in S \setminus U$ such that $\kappa(U, s) > 0$.
(c) For any $f \in \dot{C}^b_+(S)$, $S = \bigcup_{n\in\mathbb{Z}_+} \{A^n_* f > 0\} =: U(f)$.
(d) For any Lipschitz continuous $f : S \to \mathbb{R}_+$ that is not identically equal to 0, $S = \bigcup_{n\in\mathbb{Z}_+} \{A^n_* f > 0\} =: U(f)$.

Here, $\{A^n_* f > 0\}$ is a shorthand for $\{s \in S; (A^n_* f)(s) > 0\}$.

Proof (a)⇒(b): Suppose that (b) does not hold: Then there exists some nonempty open strict subset U of S such that $\kappa(U, s) = 0$ for all $s \in S \setminus U$. Since $\kappa(S, \cdot)$ is bounded, there exists some $c > 0$ such that $\kappa(U, s) \le c\chi_U(s)$ for all $s \in S$. Then

$$\kappa^{*2}(U, s) = \int_S \kappa(U, t)\kappa(dt, s) \le \int_S c\chi_U(s)\kappa(dt, s) = c\kappa(U, s) \le c^2\chi_U(s).$$

By induction, $\kappa^{n*}(U, s) \le c^n \chi_U(s)$ for all $s \in S$ and all $n \in \mathbb{N}$. So (a) does not hold.

(b)⇒(c): Since κ is a Feller kernel, the functions $A^n_* f$ in part (c) are continuous and $U(f)$ is open as union of open sets. Since f is not the zero function and $A^0_* f = f$, $U(f)$ is nonempty. Suppose $U(f) \neq S$. By (b), there exists some $s \in S \setminus U(f)$ such that $\kappa(U(f), s) > 0$. Since the measure $\kappa(\cdot, s)$ is continuous from below, there is some $n \in \mathbb{N}$ such that $\kappa\big(\{A^n_* f > 0\}, s\big) > 0$. This implies that $(A^{n+1}_* f)(s) > 0$ and $s \in U(f)$, a contradiction.

(c)⇒(d): obvious.

(d)⇒(a): Let U be a nonempty open subset of S. Choose some $t_0 \in U$. Then there exists some Lipschitz continuous $f : S \to [0, 1]$ such that $f(t_0) = 1$, $f(t) \le \chi_U(t)$ for all $t \in S$ [14, L.2.1]. By (d), for any $s \in S$, there is some $n \in \mathbb{Z}_+$ such that $0 < (A^n_* f)(s)$. Let $s \in S \setminus U$. Then $(A^0_* f)(s) = f(s) \le \chi_U(s) = 0$ and $0 < (A^n_* f)(s)$ for some $n \in \mathbb{N}$. Since A^n_* is induced by κ^{n*},

$$0 < (A^n_* f)(s) \le \int_S \chi_U(t)\kappa^{n*}(dt, s) \le \kappa^{n*}(U, s).$$

So (a) holds.

Remark 4 Assume that S is not a singleton set. If $\kappa : \mathcal{B} \times S \to \mathbb{R}_+$ is a top-irreducible Feller kernel, then $\kappa(S \setminus \{s\}, s) > 0$ for all $s \in S$.

Proof Let $s \in S$ and $T = S \setminus \{s\}$. Since S is not a singleton set, T is a nonempty open subset of S. Since κ is top-irreducible, by Lemma 3(b), there exists some $\tilde{s} \in S \setminus T$ such that $\kappa(T, \tilde{s}) > 0$. Since $S \setminus T = \{s\}$, $\kappa(T, s) > 0$.

Theorem 18 *Let κ be a top-irreducible Feller kernel, A_* the associated linear bounded map on $C^b(S)$, $r > 0$. Let $f \in \dot{C}^b_+(S)$ be an eigenfunction $rf = A_* f$. Then $f(s) > 0$ for all $s \in S$.*

Proof For all $n \in \mathbb{N}$, $f = r^{-n} A_*^n f$ and so $f(s) > 0$ for all $s \in S$ by Lemma 3(c).

Theorem 19 *Let κ be a top-irreducible Feller kernel. Then, for any $\mu \in \dot{\mathcal{M}}_+(S)$ and $f \in \dot{C}^b_+(S)$, there is some $n \in \mathbb{Z}_+$ such that $\int_S f \, d(A^n \mu) = \int_S A_*^n f \, d\mu > 0$.*

Proof Let $\mu \in \dot{\mathcal{M}}_+(S)$ and $f \in \dot{C}^b_+(S)$. By Lemma 3(c),

$$S = \bigcup_{n \in \mathbb{Z}_+} S_n(f), \qquad S_n(f) = \big\{ A_*^n f > 0 \big\}.$$

The last is a shorthand for $\big\{ s \in S, (A_*^n f)(s) > 0 \big\}$. Analogous shorthands will be used in the following.

Since μ is continuous from below and $\mu(S) > 0$, there exists some $m \in \mathbb{N}$ such that $0 < \mu\big(\bigcup_{n=0}^m S_n(f) \big)$. Since $\bigcup_{n=0}^m S_n(f) = \big\{ \sum_{n=0}^m A_*^n f > 0 \big\}$, there is some $k \in \mathbb{N}$ such that $\mu\big(T_{mk}(f)\big) > 0$, $T_{mk}(f) = \big\{ \sum_{n=0}^m A_*^n f > 1/k \big\}$. Now

$$\sum_{n=0}^m \int_S f \, d\,(A^n \mu) = \int_S \sum_{n=0}^m (A_*^n f) \, d\mu$$
$$\geq \int_{T_{mk}(f)} \Big(\sum_{n=0}^m A_*^n f \Big) d\mu \geq (1/k) \mu\big(T_{mk}(f)\big) > 0.$$

So there is some $n \in \mathbb{Z}_+$ such that $\int_S f \, d(A^n \mu) = \int_S A_*^n f \, d\mu > 0$.

Corollary 6 *Let κ be a top-irreducible Feller kernel, A the associated linear map on $\mathcal{M}_+(S)$, $r > 0$. Let $\mu \in \dot{\mathcal{M}}_+(S)$ be an eigenmeasure $r\mu = A\mu$. Then $\int_S f d\mu > 0$ for any $f \in \dot{C}^b_+(S)$.*

Proof For all $n \in \mathbb{N}$, $\mu = r^{-n} A^n \mu$ and the assertion follows from Theorem 19.

Proposition 11 *Let κ be a top-irreducible Feller kernel and let $\mathcal{N}$ be a tight subset of $\mathcal{M}_+(S)$ with $\inf_{\mu \in \mathcal{N}} \mu(S) > 0$. Then, for any $f \in \dot{C}^b_+(S)$, there exist some $m \in \mathbb{N}$ and $\delta > 0$ such that*

$$\sum_{n=0}^m \int_S A_*^n f \, d\mu \geq \delta, \qquad \mu \in \mathcal{N}.$$

Proof Let $\eta = (1/2) \inf_{\mu \in \mathcal{N}} \mu(S)$. Then $\eta > 0$. Since $\mathcal{N}$ is tight, there exists some compact subset K of S such that $\mu(S \setminus K) \leq \eta$ for all $\mu \in \mathcal{N}$ and so

$$\mu(K) \geq \eta, \qquad \mu \in \mathcal{N}. \tag{33}$$

Let $f \in C_+^b(S)$, $f \neq 0$. Since κ is top-irreducible, $S = \bigcup_{n\in\mathbb{Z}_+} S_n(f)$ with open sets $S_n(f) = \{A_*^n f > 0\}$ by Lemma 3(c). Since K is compact, there exists some $m \in \mathbb{N}$ such that $K \subseteq \bigcup_{n=0}^m S_n(f)$. So there exists some $\tilde{\delta} > 0$ such that

$$\sum_{n=0}^{m} (A_*^n f)(s) \geq \tilde{\delta}, \qquad s \in K.$$

For all $\mu \in \mathcal{N}$, by (33),

$$\sum_{n=0}^{m} \int_S A_*^n f \, d\mu \geq \int_K \Big(\sum_{n=0}^{m} A_*^n f \Big) \, d\mu \geq \tilde{\delta}\mu(K) \geq \tilde{\delta}\eta > 0.$$

5.3.1 Strongly Top-Irreducible Feller Kernels.

Recall the definition of a strongly top-irreducible Feller kernel (Definition 4).

Lemma 4 *Let $\kappa : \mathcal{B} \times S \to \mathbb{R}_+$ be a Feller kernel and A_* be the associated bounded linear map on $C^b(S)$. Then the following are equivalent:*

(a) κ is strongly top-irreducible.
(b) For any $f \in \dot{C}_+^b(S)$ and any nonempty compact subset K of S there exists some $n \in \mathbb{Z}_+$ such that $(A_*^n f)(s) > 0$ for all $s \in K$.
(c) For any Lipschitz continuous $f : S \to \mathbb{R}_+$ that is not identically equal to 0 and any nonempty compact subset K of S, there exists some $n \in \mathbb{Z}_+$ such that $(A_*^n f)(s) > 0$ for all $s \in K$.

Proof (a) $\Rightarrow$ (b):

Let $f \in \dot{C}_+^b(S)$ and K be a compact subset of S. Then $U = \{t \in S; f(t) > \|f\|_\infty/2\}$ is a nonempty open subset of S. Since κ is strongly top-irreducible, there exists some $n \in \mathbb{N}$ such that, for all $s \in K$.

$$0 < \kappa^{n\star}(U, s) \leq \int_U \frac{2f(t)}{\|f\|_\infty} \kappa^{n\star}(dt, s) \leq \frac{2}{\|f\|_\infty} (A_*^n f)(s).$$

Obviously (b) implies (c).

(c) $\Rightarrow$ (a) follows similarly as in Lemma 3(d) $\Rightarrow$ (a).

Proposition 12 *Let $\kappa : \mathcal{B} \times S \to \mathbb{R}_+$ be a Feller kernel with the following property for any $f \in \dot{C}_+^b(S)$:*

For all $s \in S$ there exists some neighborhood $U_s \subseteq S$ of s and some $n_s \in \mathbb{N}$ such that $\int_S f(t)\kappa^{n\star}(dt, \tilde{s}) > 0$ for all $n \in \mathbb{N}$, $n \geq n_s$, and all $\tilde{s} \in U_s$.

Then κ is a strongly top-irreducible Feller kernel.

Proof The neighborhoods U_s can be chosen as open sets containing s. Let K be a nonempty compact subset of S. Then $K \subseteq \bigcup_{s\in S} U_s$ and there exists a finite subset $\tilde{S}$ of S such that $K \subseteq \bigcup_{s\in\tilde{S}} U_s$. Set $m = \max_{s\in\tilde{S}} n_s$. Then $m \in \mathbb{N}$ and

$$\int_S f(t)\kappa^{m\star}(dt, \tilde{s}) > 0, \qquad \tilde{s} \in K.$$

By Lemma 4, κ is strongly top-irreducible. □

A similar proof as for Proposition 11 yields the following result.

Proposition 13 *Let κ be a strongly top-irreducible Feller kernel and let $\mathcal{N}$ be a tight subset of $\mathcal{M}_+(S)$ with $\inf_{\mu\in\mathcal{N}} \mu(S) > 0$. Then, for any $f \in \dot{C}^b_+(S)$, there exist some $n \in \mathbb{N}$ and $\delta > 0$ such that*

$$\int_S A^n_* f \, d\mu \geq \delta, \qquad \mu \in \mathcal{N}.$$

Proposition 14 *Let $P : \mathcal{B} \times \tilde{S} \to \mathbb{R}_+$ be a Feller kernel, $g \in C^b_+(S \times S)$, and $\kappa : \mathcal{B} \times S \to \mathbb{R}_+$ be defined by*

$$\kappa(T, s) = \int_T g(s, t)P(dt, s), \qquad s \in S, T \in \mathcal{B}. \tag{34}$$

Assume that κ is also a Feller kernel and that $g(s, t) > 0$ for all $s, t \in S$.

(a) P is top-irreducible if and only if κ is top-irreducible.
(b) P is strongly top-irreducible if and only if κ is strongly top-irreducible.

Proof For $f_0 \in \dot{C}^b_+(S)$, set $U_n = \{f_n > 0\}$ and $V_n = \{h_n > 0\}$ where $f_{n+1} = \int_S f_n(t)P(dt, \cdot)$ and $h_{n+1} = \int_S h_n(t)\kappa(dt, \cdot)$ for all $n \in \mathbb{N}$. Let $U(f_0)$ and $V(f_0)$ be the respective unions over $n \in \mathbb{N}$.

For any $f \in \dot{C}^b_+(S)$ and $s \in S$, we have the equivalence of the following two statements:

(i) $\int_S f(t)P(dt, s) > 0$,
(ii) $P\big(\{f > 0\}, s\big) > 0$.

An analogous equivalence holds for κ replacing P.

Since g is strictly positive on S^2, statement (ii) for P is equivalent to the statement (ii) for κ replacing P.

With this observation, it follows by induction that $U_n = V_n$ for all $n \in \mathbb{N}$ such that $U(f) = V(f)$. So $S = U(f)$ if and only if $S = V(f)$.

The equivalence in (a) follows from Lemma 3(c).

The equivalence in (b) follows from Lemma 4(b).

In these lemmata, $U_n = \{A^n_* f > 0\}$ if A_* is induced by P and $V_n = \{A^n_* f > 0\}$ if A_* is induced by κ.

5.3.2 Colonization Kernels.

The following example of strongly top-irreducible kernels seems particularly suited for spatially structured populations, but less for populations with other structures.

Definition 19 Let $\kappa : \mathcal{B} \times S \to \mathbb{R}_+$ be a Feller kernel. κ is called a *colonization kernel* if for any $s \in S$ there is an open subset $U \ni s$ of S such that $\kappa(V, s) > 0$ for all nonempty open subsets V of U.

Proposition 15 *Let S be connected and κ be a colonization Feller kernel. Then, for any $f \in \dot{C}^b_+(S)$, $S = \bigcup_{n\in\mathbb{Z}_+} S_n(f)$ where $S_n(f) = \{A^n_* f > 0\}$ form an increasing sequence of open sets, and κ is strongly top-irreducible.*

Here A_* is the operator defined in (9).

Proof Let $f \in \dot{C}^b_+(S)$ and define $S_n(f)$ as above.

Since $A^n_* f$ is continuous, the sets $S_n(f)$ form a sequence of open subsets of S. We claim that this sequence is increasing with respect to the subset relation. It is sufficient to show that $S_0(f) \subseteq S_1(f)$ because $S_{n+1}(f) = S_1(A^n_*(f))$. Let $s \in S$ and $f(s) > 0$. Since κ is a colonization kernel, there is an open subset $U \ni s$ of S such that $\kappa(V, s) > 0$ for all nonempty open subsets V of U. Set $V = \{t \in U;\, f(t) > f(s)/2\}$. Then V an open subset of U and $s \in V$; so

$$(A_* f)(s) \geq \int_V f(t)\kappa(dt, s) \geq \frac{f(s)}{2}\kappa(V, s) > 0.$$

This implies $S_0(f) \subseteq S_1(f)$.

Set $S(f) = \bigcup_{n\in\mathbb{N}} S_n(f)$. $S(f)$ is open as union of open sets. To show that $S(f)$ is closed, let $s \in S$ be a limit point of $S(f)$. Since κ is a colonization kernel, there is an open subset $U \ni s$ of S such that $\kappa(V, s) > 0$ for all nonempty open subsets V of U. Since s is a limit point of $S(f)$, $U \cap S(f) \neq \emptyset$ and $U \cap S_n(f) \neq \emptyset$ for some $n \in \mathbb{Z}_+$. Since $S_n(f) = \bigcup_{m\in\mathbb{N}} \{A^n_* f > 1/m\}$, there exists a nonempty open subset V of U and some $m \in \mathbb{N}$ such that $(A^n_* f)(t) > 1/m$ for all $t \in V$. For all $x \in U$,

$$(A^{n+1}_* f)(s) \geq \int_V (A^n_* f)(t)\kappa(dt, s) \geq (1/m)\kappa(V, s) > 0.$$

So $s \in S_{n+1}(f) \subseteq S(f)$. Since $S(f)$ is open and closed in the connected set S, $S = S(f)$.

Let K be a compact subset of S. Then there exists some $n \in \mathbb{N}$ such that $K \subseteq \bigcup_{j=1}^n S_j(f)$. Since the $S_n(f)$ form an increasing sequence of sets, $K \subseteq S_n(f)$, i.e., $(A^n_* f)(s) > 0$ for all $s \in K$. So, κ is strongly top-irreducible by Lemma 4.

Lemma 5 *Let $\kappa : \mathcal{B} \times S \to \mathbb{R}_+$ be a tight colonization Feller kernel and $g : S \times S \to (0, \infty)$ be continuous and bounded. Then $\tilde{\kappa} : \mathcal{B} \times S \to \mathbb{R}_+$ defined by*

$$\tilde{\kappa}(T, s) = \int_T g(s, t)\kappa(dt, s), \qquad T \in \mathcal{B}, s \in S,$$

is also a tight colonization Feller kernel.

Proof By Proposition 7, $\tilde{\kappa}$ is a tight Feller kernel.

Let $s \in S$. Since κ is a colonization kernel, there is some open subset $U \ni s$ of S such that $\kappa(V, s) > 0$ for all nonempty open subsets V of U. Since g is strictly positive and continuous, $V = \bigcup_{n\in\mathbb{N}} V_n$ with open subsets $V_n = \{t \in V; g(s, t) > 1/n\}$ of V. For all $n \in \mathbb{N}$,

$$\tilde{\kappa}(V, s) \geq \int_{V_n} g(s, t)\kappa(dt, s) \geq (1/n)\kappa(V_n, s).$$

Since $\kappa(\cdot, s)$ is continuous from below, $\kappa(V_n, s) \to \kappa(V, s) > 0$ as $n \to \infty$ and $\tilde{\kappa}(V, s) > 0$.

6 Proofs for the General Framework for The state Space of Measures. Tight Bounded Persistence Attractors

Recall that we consider yearly turnover maps F of the following form,

$$F(\mu)(T) = \int_S \kappa^\mu(T, s)\ \mu(ds), \qquad \mu \in \mathcal{M}^s_+(S), \qquad T \in \mathcal{B},$$

where $\{\kappa^\mu; \mu \in \mathcal{M}^s_+(S)\}$ is a set of Feller kernels $\kappa^\mu : \mathcal{B} \times S \to \mathbb{R}_+$.

If μ is the zero measure, we use the notation κ^o.

Proposition 16 *Let the Assumption 9 be satisfied. Then F maps $\mathcal{M}^s_+(S)$ into itself.*

Proof Theorem 15(a).

Lemma 6 *Let $(\tilde{f}_n)$ be a bounded sequence in $C^b(S)$ and (μ_n) be a bounded pre-tight sequence in $\mathcal{M}_+(S)$. Then*

$$\int_S \tilde{f}_n d\mu_n \overset{n\to\infty}{\longrightarrow} 0 \qquad \text{if} \quad \tilde{f}_n \overset{n\to\infty}{\longrightarrow} 0$$

uniformly on every totally bounded subset of S.

Proof Let $\epsilon > 0$. Since $\{\mu_n; n \in \mathbb{N}\}$ is pre-tight, there exists a closed totally bounded subset T of S such that $\mu_n(S \setminus T) < \epsilon$ for all $n \in \mathbb{N}$. For all $n \in \mathbb{N}$,

$$\begin{aligned}\Big|\int_S \tilde{f}_n d\mu_n\Big| &\leq \int_T |\tilde{f}_n| d\mu_n + \int_{S\setminus T} |\tilde{f}_n| d\mu_n \\ &\leq \sup_T |\tilde{f}_n| \sup_{k\in\mathbb{N}} \mu_k(S) + \sup_{k\in\mathbb{N}} \sup_S |\tilde{f}_k|\ \mu_n(S \setminus T).\end{aligned}$$

Since $\tilde{f}_n \to 0$ uniformly on T, the last but one expression converges to 0 as $n \to \infty$ and

$$\limsup_{n\to\infty} \left| \int_S \tilde{f}_n d\mu_n \right| \le \sup_{k\in\mathbb{N}} \sup_S |\tilde{f}_k| \, \epsilon.$$

Since this holds for arbitrary $\epsilon > 0$, the limit superior is zero and we have proved the assertion.

Proposition 17 *Let the family of Feller kernels $\{\kappa^\mu); \mu \in \mathcal{M}_+^s(S)\}$ satisfy the Assumptions 9 and 11. Then $F : \mathcal{M}_+^s(S) \to \mathcal{M}_+^s(S)$ is continuous with respect to the flat norm.*

Proof Let $\mu \in \mathcal{M}_+^s(S)$ and (μ_n) be a sequence in $\mathcal{M}_+^s(S)$ such that $\|\mu_n - \mu\|_\flat \to 0$. By Theorem 12,

$$\int_S \tilde{f} d\mu_n \to \int_S \tilde{f} d\mu, \qquad \tilde{f} \in C_+^b(S). \tag{35}$$

Then $\{\mu_n; n \in \mathbb{N}\}$ is a compact subset of $\mathcal{M}_+(S)$ with respect to the flat norm and pre-tight by Proposition 5 and a bounded subset of $\mathcal{M}_+(S)$.

Let $f \in \mathcal{F}$. By (16),

$$\left| \int_S f \, dF(\mu_n) - \int_S f \, dF(\mu) \right| = \left| \int_S f_n d\mu_n - \int_S \tilde{f} d\mu \right| \tag{36}$$

with

$$f_n(s) = \int_S f(t) \kappa^{\mu_n}(dt, s), \qquad \tilde{f}(s) = \int_S f(t) \kappa^{\mu}(dt, s).$$

By Theorem 12, it is sufficient that the expression on the right hand side of (36) converges to 0 as $n \to \infty$.

By the triangle inequality and (36),

$$\left| \int_S f \, dF(\mu_n) - \int_S f \, dF(\mu) \right| \le \left| \int_S (f_n - \tilde{f}) d\mu_n \right| + \left| \int_S \tilde{f} d\mu_n - \int_S \tilde{f} d\mu \right|.$$

Since κ^μ is a Feller kernel, $\tilde{f} \in C_+^b(S)$ and the second term on the right hand side of the last inequality converges to 0 as $n \to \infty$ by (35). As for the first term, by Assumption 11, for any closed totally bounded subset T of S

$$f_n(s) - \tilde{f}(s) \to 0, \qquad n \to \infty, \text{ uniformly for } s \in T. \tag{37}$$

Further, by Assumption 9, $(f_n - \tilde{f})$ is a bounded sequence in $C^b(S)$. Now the first term of the last inequality converges to 0 by Lemma 6.

Proposition 18 *Under the Assumptions 9 and 12, the yearly population turnover map $F : \mathcal{M}_+^s(S) \to \mathcal{M}_+^s(S)$ is compact; for any bounded subset $\mathcal{N}$ of $\mathcal{M}_+^s(S)$, $F(\mathcal{N})$ is a tight bounded subset of $\mathcal{M}_+^s(S)$.*

Proof Let $\mathcal{N}$ be a bounded subset of $\mathcal{M}_+^s(S)$. For any set $T \in \mathcal{B}$ and $\mu \in \mathcal{N}$,

$$F(\mu)(S \setminus T) = \int_S \kappa^\mu(S \setminus T, s)\, \mu(ds) \le \sup_{s \in S} \kappa^\mu(S \setminus T, s)\, \mu(S). \tag{38}$$

For $T = \emptyset$, we obtain that $\{F(\mu)(S);\, \mu \in \mathcal{N}\}$ is bounded in $\mathbb{R}$ by Assumption 12.

Let $\epsilon > 0$. By Assumption 12, there exists some compact set T in S such that

$$\kappa^\mu(S \setminus T, s) \le \epsilon \Big(1 + \sup_{\mu \in \mathcal{N}} \mu(S)\Big)^{-1}, \qquad s \in S.$$

By (38), $F(\mu)(S \setminus T) \le \epsilon$ for all $\mu \in \mathcal{N}$. By Definition 7, $F(\mathcal{N})$ is a tight subset of $\mathcal{M}_+^s(S)$.

By Theorem 13, $F(\mathcal{N})$ has compact closure in $\mathcal{M}_+^s(S)$.

Proposition 19 *Let the Assumptions 9 and 13 be satisfied. Then*

$$\limsup_{\mu(S) \to \infty} \frac{F(\mu)(S)}{\mu(S)} < 1.$$

Proof For all $\mu \in \mathcal{M}_+^s(S)$,

$$\mathcal{F}(\mu)(S) = \int_S \kappa^\mu(S, s)\, \mu(ds) \le \sup_{s \in S} \kappa^\mu(S, s)\, \mu(S).$$

This implies the assertion.

Theorem 20 *Let the Assumptions 9, 11, 12, and 13 be satisfied.*
Then the semiflow induced by F has a compact attractor of bounded sets.

Proof We apply Theorem 9. By Assumption 13 and Proposition 19, inequality (19) is satisfied with $\theta(\mu) = \mu(S)$. F is continuous by Proposition 17 and compact and thus asymptotically smooth by Proposition 18. All assumptions of Theorem 9 are satisfied which implies that the semiflow induced by F has a compact attractor of bounded sets. □

Let us spell out what Theorem 20 means [26, Chap. 2].

Remark 5 Under the assumptions of Theorem 20, there exists a subset $\mathcal{K}$ of $\mathcal{M}_+^s(S)$ which is tight, compact with respect to the flat norm, and satisfies $F(\mathcal{K}) = \mathcal{K}$. Further, if $\mathcal{N}$ is a bounded subset of $\mathcal{M}_+^s(S)$ and $\mathcal{U}$ an open set in $\mathcal{M}_+^s(S)$ with respect to the flat norm with $\mathcal{K} \subseteq U$, there exists some $N \in \mathbb{N}$ such that $F^n(\mathcal{N}) \subseteq \mathcal{U}$ for all $n \in \mathbb{N}$ with $n \ge N$.

The tightness of $\mathcal{K}$ follows from Proposition 18 and $F(\mathcal{K}) = \mathcal{K}$.

Proposition 20 *Under the Assumptions 9 and 10, F maps $\dot{\mathcal{M}}_+^s(S)$ into itself.*

Proof Let $\mu \in \dot{\mathcal{M}}_+(S)$. Then $\mu(S) > 0$. By Assumption 10, $S = \bigcup_{j\in\mathbb{N}} T_j$ with

$$T_j = \{s \in S; \kappa^\mu(S, s) \geq 1/j\}.$$

Notice that $T_j \subseteq T_{j+1}$ for all $j \in \mathbb{N}$. Since μ is continuous from below, $0 < \mu(S) = \lim_{j\to\infty} \mu(T_j)$. So, for some $j \in \mathbb{N}$, $\mu(T_j) > 0$ and

$$F(\mu)(S) \geq \int_{T_j} \kappa^\mu(S, s)\mu(ds) \geq (1/j)\mu(T_j) > 0$$

and $F(\mu) \in \dot{\mathcal{M}}^s_+(S)$. □

The following result implies that the extinction state is unstable.

Theorem 21 *Make Assumptions 9 and 10 and let the kernel family* $\{\kappa^\mu; \mu \in \mathcal{M}^s_+(S)\}$ *be lower semicontinuous at the zero measure.*

Further assume that there exists some $r > 0$ *and* $f \in C^b_+(S)$ *such that* $f(s) > 0$ *for all* $s \in S$ *and*

$$\int_S f(t)\kappa^o(dt, s) \geq rf(s), \qquad s \in S.$$

Then the semiflow induced by F *is uniformly weakly persistent: There exists some* $\delta > 0$ *such that* $\limsup_{n\to\infty} F^n(\mu)(S) \geq \delta$ *for all* $\mu \in \dot{\mathcal{M}}^s_+(S)$.

Proof We apply [20, Theorem 5.2] with

$$(B\mu)(T) = \int_S \kappa^o(T, s)\mu(ds)$$

and

$$\theta(\mu) = \int_S f d\mu, \qquad \mu \in \mathcal{M}^s_+(S).$$

The assumptions (a) and (b) are satisfied by Assumption 10 and the strict positivity of f. Assumption (c) follows from the lower semicontinuity of the kernel family. □

Proof of Theorem 6. We apply Theorem 21. By Theorem 17, there is some $f \in \dot{C}^b_+(S)$ such that

$$\int_S f(t)\kappa^o(dt, s) = rf(s), \qquad s \in S,$$

$r = \mathbf{r}(\kappa^o)$. f is strictly positive by Corollary 18. □

Proof of Theorem 7. We combine [26, Theorem 4.5], Theorems 20 and 6.

6.1 Compact Persistence Attractor

Theorem 22 *Make Assumptions 9, 10, 11, 12, 13, 14 and let the kernel family* $\{\kappa^\mu;\, \mu \in \mathcal{M}^s_+(S)\}$ *be lower semicontinuous at the zero measure.*

Assume that κ^o *is a strongly top-irreducible Feller kernel and* $\kappa^o = \kappa_1 + \kappa_2$ *with two tight Feller kernels* κ_j*, where* κ_2 *is a uniform Feller kernel. Finally, assume* $r = \mathbf{r}(\kappa^o) > 1 \geq \mathbf{r}(\kappa_1)$.

Then the semiflow induced by F *has a compact connected persistence attractor* $\mathcal{A}_1$*:*

(a) $\mathcal{A}_1$ *is a compact set with respect to the flat norm,* $F(\mathcal{A}_1) = \mathcal{A}_1$*, and* $\mathcal{A}_1$ *is a tight set of measures.*

(b) $\mathcal{A}_1$ *attracts all subsets* $\mathcal{N}$ *of* $\mathcal{M}^s_+(S)$ *with* $\inf_{\mu\in\mathcal{N}} \mu(S) > 0$ *that are compact with respect to the flat norm or are bounded and tight: If* $\mathcal{N}$ *is such a subset and* $\mathcal{U}$ *is an open set in* $\mathcal{M}^s_+(S)$ *with respect to the flat norm,* $\mathcal{A}_1 \subseteq \mathcal{U}$*, then there exists some* $N \in \mathbb{N}$ *such that* $F^n(\mathcal{N}) \subseteq \mathcal{U}$ *for all* $n \in \mathbb{N}$ *with* $n \geq \mathbb{N}$.

(c) For any $f \in \dot{C}^b_+(S)$*, there exists some* $\epsilon_f > 0$ *such that* $\int_S f\,d\mu \geq \epsilon_f$ *for all* $\mu \in \mathcal{A}_1$.

(d) $\mathcal{A}_1$ *is connected with respect to the flat norm. In particular, for any* $f \in \dot{C}^b_+(S)$*,* $\left\{\int_S f\,d\mu;\, \mu \in \mathcal{A}_1\right\}$ *is a compact interval (possibly a singleton set) contained in* $(0, \infty)$.

Proof We apply [26, Sect. 5.2] with $X = \mathcal{M}^s_+(S)$ and $\rho(\mu) = \mu(S)$ for $\mu \in \mathcal{M}^s_+(S)$. Since $F(0) = 0$ and $F(X \setminus \{0\}) \subseteq X \setminus \{0\}$ by Proposition 20, the set $X_0 := \left\{\mu \in X;\, \forall n \in \mathbb{Z}_+ : F^n(\mu) = 0\right\} = \{0\}$.

By Theorem 6, the semiflow $\{F^n;\, n \in \mathbb{Z}_+\}$ is uniformly weakly ρ-persistent.

The statements (a) and (b) follow from [26, Theorem 5.7](b) as does

$$\delta := \inf_{\mu\in\mathcal{A}_1} \mu(S) > 0. \tag{39}$$

(c) By Assumption 14, there exists a strongly top-irreducible Feller kernel $\tilde{\kappa}$ such that

$$\kappa^\nu(T, s) \geq \tilde{\kappa}(T, s), \qquad T \in \mathcal{B}, s \in S, \nu \in \mathcal{A}_1.$$

Let $\tilde{A}_*$ be the map on $C^b(S)$ associated with $\tilde{\kappa}$. For any $f \in C^b_+(S)$, $\mu \in \mathcal{A}_1$,

$$\begin{aligned}\int_S f\ d\,F(\mu) &= \int_S \Big(\int_S f(t)\kappa^\mu(dt, s)\Big)\mu(ds)\\ &\geq \int_S \Big(\int_S f(t)\tilde{\kappa}(dt, s)\Big)\mu(ds) = \int_S (\tilde{A}_* f)\ d\,\mu.\end{aligned}$$

By induction, for any $f \in C^b_+(S)$,

$$\int_S f\ dF^k(\mu) \geq \int_S (\tilde{A}^k_* f)\ d\,\mu, \qquad k \in \mathbb{N}, \mu \in \mathcal{A}_1. \tag{40}$$

Now let $f \in \dot{C}^b_+(S)$. By Proposition 13 since $\tilde{\kappa}$ is a strongly top-irreducible Feller kernel, there exists some $n \in \mathbb{N}$ and $\epsilon_f > 0$ such that

$$\epsilon_f \leq \int_S (\tilde{A}^n_* f)\, d\,\mu \leq \int_S f\, dF^n(\mu), \qquad \mu \in \mathcal{A}_1.$$

Since $F^n(\mathcal{A}_1) = \mathcal{A}_1$, this implies that $\int_S f\, d\nu \geq \epsilon_f > 0$ for all $\nu \in \mathcal{A}_1$.

(d) Connectedness from $\mathcal{A}_1$ follows from [26, Proposition 5.9] because ρ with $\rho(\mu) = \mu(S)$ is concave, actually additive on $\mathcal{M}^s_+(S)$. By Theorem 12, for any $f \in C^b(S)$, the map $\phi_f : \mathcal{M}^s_+(S) \to [0, \infty)$, $\phi_f(\mu) = \int_S f d\mu$, is continuous under the flat norm. Since continuous images of compact (connected) sets are compact (connected), $\phi_f(\mathcal{A}_1)$ is compact and connected and, by (c), a subset of $(0, \infty)$ if $f \in \dot{C}^b_+(S)$. □

Proof of Theorem 8. Let $\mathcal{A}_1$ be the persistence attractor from Theorem 22 and $f \in \dot{C}^b_+(S)$. Then there exists some $\epsilon_f > 0$ such that $\int_S f d\mu > \epsilon_f$ for all $\mu \in \mathcal{A}_1$. Set $\mathcal{U} = \{\nu \in \mathcal{M}^s_+(S); \int_S f d\nu > \epsilon_f\}$. By Theorem 12, $\mathcal{U}$ is an open set in $\mathcal{M}^s_+(S)$ with respect to the flat norm and $\mathcal{A}_1 \subseteq \mathcal{U} \subseteq \mathcal{M}^s_+(S)$. The statement now follows from Theorem 22(b) and (c). □

7 A More Specific Model for an Iteroparous Population

We consider a structured population the dynamics of which are governed by the processes of birth, death, and structural development, with the last being spatial movement to be specific.

We assume that each year has one very short reproductive season. We count the years in such a way that the census period is just before the reproductive season. At the end of the year, juveniles born at the beginning of the year have matured enough that they are reproductive as well and are counted as adults. This means that each year, at the very beginning of the year, just before the reproductive season, all individuals are adults. Differently from the model for a semelparous population considered in [32], individuals can reproduce several times during their life-time.

Births and deaths can be affected by competition for resources. Consider a typical adult individual at location $t \in S$. Let $q_1(s, t)$ denote the competitive effect it has on an adult located at $s \in S$ and $q_2(s, t)$ denote the competitive effect it has on a neonate located at $s \in S$. Here $q_j : S^2 \to \mathbb{R}_+$. If $\mu \in \mathcal{M}_+(S)$ is the distribution of adult individuals at the beginning of the year and $s \in S$,

$$(Q_1\mu)(s) = \int_S q_1(s, t)\mu(dt) \tag{41}$$

is the competition level exerted by μ on an adult that has been at s at the beginning of the year. while

$$(Q_2\mu)(s) = \int_S q_2(s,t)\mu(dt) \tag{42}$$

is the competition level exerted by μ on a juvenile born at s.

Further, let $g_1(s,q)$ be the probability of an adult located at $s \in S$ at the beginning of the year to survive competition till the end of the year when the competition level at s is $q \in \mathbb{R}_+$, $g_1 : S \times \mathbb{R}_+ \to [0,1]$.

Let $g_2 : S \times \mathbb{R}_+ \to \mathbb{R}_+$ be the effective per capita birth function, i.e., $g_2(s,q)$ is the per capita amount of offspring that is produced at $s \in S$ by an adult located at s and that survives competition till the end of the year when the competition level at s is $q \in \mathbb{R}_+$.

We assume that the migration patterns of neonates and adults are possibly different.

Let $P_1(T,s)$ be the probability that an adult staying at $s \in S$ at the beginning of the year does not die from noncompetitive causes till the end of the year and is located at some point in the set T at the end of the year.

Similarly, let $P_2(T,s)$ be the probability that a neonate born at $s \in S$ at the beginning of the year does not die from noncompetitive causes till the end of the year and is located at some point in the set T at the end of the year.

If the measure ν represents the spatial distribution of neonates shortly after the reproductive season at the beginning of the year,

$$(A_2\nu)(T) = \int_S P_2(T,s)\nu(ds), \qquad T \in \mathcal{B}, \tag{43}$$

provides the resulting number of adults that, at the end of the year, have not died from noncompetitive causes and are located within the set T.

A similar formula holds for the relation between the spatial distribution of adults at the beginning of the year and the resulting distribution of survivors at the end of the year.

In combination, the turnover kernel for a population with spatial distribution $\mu \in \mathcal{M}_+(S)$ is

$$\left.\begin{aligned} \kappa^\mu(T,s) &= \kappa_1^\mu(T,s) + \kappa_2^\mu(T,s) \\ \kappa_j^\mu(T,s) &= P_j(T,s)\, g_j\big(s, (Q_j\mu)(s)\big) \end{aligned}\right\} \qquad T \in \mathcal{B}, s \in S, \tag{44}$$

and $Q_j\mu$ from (41) and (42). Notice that

$$\kappa_j^o(T,s) = P_j(T,s)\, g_j(s,0), \qquad T \in \mathcal{B}, s \in S, \tag{45}$$

Assumption 20 For the per capita survival and reproduction rate functions g_1 and g_2,

(g1) $g_j : S \times \mathbb{R}_+ \to \mathbb{R}_+$ is continuous and bounded, $j = 1,2$; $g_1(s,q) \le 1$ for all $s \in S, q \in \mathbb{R}_+$.

(g2) $g_j(s,0) > 0$ for all $s \in S$ and $\dfrac{g_j(s,u)}{g_j(s,0)} \to 1$ as $u \to 0$ uniformly for $s \in S$.

For the competitive influence functions q_1 and q_2,

(q1) $q_j : S^2 \to \mathbb{R}_+$ is continuous and bounded.

For the survival/migration kernels P_1 and P_2,

(P1) $P_j : \mathcal{B} \times S \to \mathbb{R}_+$ is a Feller kernel (Definition 16) of separable measures.
(P2) $0 \le P_j(S,s) \le 1$ for all $s \in S$.

Here S^2 and $S \times \mathbb{R}_+$ are equipped with the respective product topologies.

Lemma 7 *Let the Assumptions 20 be satisfied and $j = 1, 2$. Then, for any $\mu \in \mathcal{M}_+(S)$, $Q_j\mu$ is continuous on S and $g_j\big(s, (Q_j\mu)(s)\big)$ is a continuous function of $s \in S$.*

For each $\mu \in \mathcal{M}_+(S)$, the kernels κ_j^μ, $j = 1, 2$, and κ^μ are Feller kernels of separable measures, and the Assumptions 5 are satisfied for κ_j^μ and $\kappa^\mu = \kappa_1^\mu + \kappa_2^\mu$.

Further, the kernel families $\{\kappa_j^\mu;\, \mu \in \mathcal{M}_+(S)\}$, $j = 1, 2$, and $\{\kappa^\mu;\, \mu \in \mathcal{M}_+(S)\}$ are continuous at the zero measure.

Moreover, $\kappa_1^\mu(S,s) \le 1$ for all $\mu \in \mathcal{M}_+^s(S)$ and all $s \in S$ and $\mathbf{r}(\kappa_1^o) \le 1$.

Finally, if the kernel P_2 is tight, so is the kernel κ_2^o.

Proof Let $\mu \in \mathcal{M}_+(S)$. Then

$$(Q_j\mu)(s) = \int_S q_j(s,t)\mu(dt)$$

is a continuous function of s by Lebesgue's theorem of dominated convergence because q_j is continuous and bounded by Assumption 20. By the same assumption, $g_j\big(s, (Q_j\mu)(s)\big)$ is a continuous function of $s \in S$ as composition of continuous functions.

Let $f \in C_+^b(S)$. Then

$$\int_S f(t)\kappa_j^\mu(dt,s) = h_j(s)\, g_j\big(s, (Q_j\mu)(s)\big),$$

$$h_j(s) = \int_S f(t) P_j(dt,s),$$

$h_j \in C^b(S)$ because P_j is a Feller kernel. As product of continuous functions, $\int_S f(t)\kappa_j^\mu(dt,s)$ is a continuous function of $s \in S$.

This implies that κ_j^μ is a Feller kernel and so is κ^μ.

Further, $\kappa_j^\mu(S,s) \le P_j(S,s) \sup g_j(S \times \mathbb{R}_+) \le \sup g_j(S \times \mathbb{R}_+)$ is a bounded function of $s \in S$ and $\kappa_1^\mu(S,s) \le 1$ by Assumption 20.

The separability of $\kappa_j^\mu(\cdot, s)$ is inherited from the separability of $P_j(\cdot, s)$.

The continuity of the kernel families at the zero measure follows from Assumption 20 (g2) and (44) and (45).

κ_2 inherits tightness from P_2 via the boundedness of g_2. □

The subsequent stability result follows from Theorem 4 [32, Theorem 3.6].

Theorem 23 *Let the Assumptions 20 be satisfied and* $r = \mathbf{r}(\kappa^o) < 1$.

(a) The extinction state is locally asymptotically stable in the following sense: For each $\alpha \in (r, 1)$*, there exist some* $\delta_\alpha > 0$ *and* $M_\alpha \geq 1$ *such that,*

$$F^n(\mu)(S) \leq \alpha^n M_\alpha \, \mu(S), \qquad n \in \mathbb{N},$$

if $\mu \in \mathcal{M}_+(S)$ *with* $\mu(S) \leq \delta_\alpha$.

(b) If $g_j(s, q) \leq g_j(s, 0)$ *for all* $s \in S$, $q \in \mathbb{R}_+$, $j = 1, 2$, *the extinction state is globally stable in the following sense: For each* $\alpha \in (r, 1)$*, there exists some* $M_\alpha \geq 1$ *such that*

$$F^n(\mu)(S) \leq \alpha^n M_\alpha \, \mu(S), \qquad n \in \mathbb{N}, \quad \mu \in \mathcal{M}_+(S).$$

The subsequent instability result follows from Theorem 5 and from Lemma 7 and shows that the assumption $\mathbf{r}(\kappa^o) < 1$ in Theorem 23 is almost sharp.

Theorem 24 *Let the Assumptions 20 be satisfied and* P_2 *be a tight Feller kernel. Let* $r = \mathbf{r}(\kappa^o) > 1$.

Then there exists some eigenmeasure $\nu \in \mathcal{M}_+^s(S)$, $\nu(S) = 1$, *such that*

$$r\nu(T) = \int_S \kappa^o(T, s)\nu(ds), \qquad T \in \mathcal{B}.$$

Further, the zero measure is unstable: There is some $\delta_0 > 0$ *such that for any* ν*-positive* $\mu \in \mathcal{M}_+(S)$ *there is some* $n \in \mathbb{Z}_+$ *with* $F^n(\mu)(S) \geq \delta_0$.

Recall that $\mu \in \mathcal{M}_+(S)$ is ν-positive if there exists some $\delta > 0$ such that $\mu(T) \geq \delta\nu(T)$ for all $T \in \mathcal{B}$.

Proposition 21 *Let Assumption 20 be satisfied. Assume that* P_1 *and* P_2 *are tight Feller kernels. Then, for any* $\mu \in \mathcal{M}_+(S)$, κ_1^μ, κ_2^μ *and* κ^μ *are tight Feller kernels. Further, the sets of measures*

$$\{\kappa_j^\mu(\cdot, s); s \in S, \mu \in \mathcal{M}_+(S)\}, \quad j = 1, 2, \quad \text{and } \{\kappa^\mu(\cdot, s); s \in S, \mu \in \mathcal{M}_+(S)\}$$

are tight and the sets

$$\{\kappa_j^\mu(S, s); s \in S, \mu \in \mathcal{M}_+(S)\}, \quad j = 1, 2, \text{ and } \{\kappa^\mu(S, s); s \in S, \mu \in \mathcal{M}_+(S)\}$$

are bounded in $\mathbb{R}$. *In particular, Assumption 12 is satisfied.*

Proof κ_1^μ, κ_2^μ and κ^μ are tight Feller kernels by Proposition 7 and Lemma 7.

Since the functions g_j are bounded, there exists some $c > 0$ such that $g_j\big(s, (Q_j\mu)(s)\big) \le c$ for all $s \in S$ and $\mu \in \mathcal{M}_+(S)$, $j = 1, 2$. For any $T \in \mathcal{B}$,

$$\kappa_j^\mu(S \setminus T, s) \le \kappa^\mu(S \setminus T, s) \le c\big(P_1(S \setminus T, s) + P_2(S \setminus T, s)\big).$$

Since P_j are Feller kernels, the right hand side has a common upper bound for $s \in S$, $T \in \mathcal{B}$. This implies the boundedness of the various sets in $\mathbb{R}$ in the assertion of the Proposition. Let $\epsilon > 0$. For $j = 1, 2$, since the P_j are tight kernels, there exist compact sets $T_j \in \mathcal{B}$ such that $P_j(S \setminus T_j, s) \le \epsilon/(2c)$. Set $T = T_1 \cup T_2$. Then T is compact and, for all $\mu \in \mathcal{M}_+(S)$, $s \in S$,

$$\kappa_j^\mu(S \setminus T, s) \le \kappa^\mu(S \setminus T, s) \le c\big(P_1(S \setminus T_1, s) + P_2(S \setminus T_2, s)\big) \le \epsilon.$$

Proposition 22 *Let Assumption 20 be satisfied. If $P := P_1 + P_2$ is a (strongly) top-irreducible kernel, so is κ^o.*

Proof Set $h(s) = \min\{g_1(s, 0), g_2(s, 0)\}$, $s \in S$. Then $h \in C_+^b(S)$ and, by Assumption 20, $h(s) > 0$ for all $s \in S$.

By (44),

$$\kappa^o(T, s) \ge P(T, s)h(s) =: \tilde{\kappa}(T, s), \qquad T \in \mathcal{B}, s \in S.$$

Since P is a (strongly) top-irreducible kernel and h is strictly positive, $\tilde{\kappa}$ is a (strongly) top-irreducible kernel by Proposition 14 and so is κ^o as one sees from Definition 19. □

In view of these results, we collect the following set of assumptions.

Assumption 21 • P_1 and P_2 are tight Feller kernels.
• $g_j(s, q) > 0$ for $j = 1, 2$ and all $s \in S$ and $q \in \mathbb{R}_+$.

Lemma 8 *Assume that $g_j(s, q) > 0$ for all $s \in S$, $q \in \mathbb{R}_+$. Let $\mathcal{N}$ be a bounded subset of $\mathcal{M}_+(S)$ and $P = P_1 + P_2$ be a strongly top-irreducible kernel. Then there exists a strongly top-irreducible kernel $\tilde{\kappa}$ such that $\kappa^\mu(T, s) \ge \tilde{\kappa}(T, s)$ for all $T \in \mathcal{B}$, $s \in S$ and $\mu \in \mathcal{N}$.*

In particular, Assumption 14 is satisfied.

Proof Let $\mathcal{N}$ be a bounded subset of $\mathcal{M}_+(S)$. Since q_j is bounded, by (41) and (42) there exists some $c \in (0, \infty)$ such that $(Q_j\mu)(s) \le c$ for $j = 1, 2$, $s \in S$, and $\mu \in \mathcal{N}$. Set

$$h_j(s) = \inf_{0 \le q \le c} g_j(s, q), \quad s \in S, \qquad j = 1, 2.$$

Since g_j is continuous and $g_j(s, q) > 0$ for all $s \in S, q \in \mathbb{R}_+, h_j(s) > 0$ for all $s \in S$. To show that h_j is continuous, let $s \in S$ and (s_ℓ) be a sequence in S such that $s_\ell \to s$ as $\ell \to \infty$. Then $T = \{s_\ell; \ell \in \mathbb{N}\} \cup \{s\}$ is a compact subset of S and $T \times [0, c]$ is a

compact subset of $S \times \mathbb{R}$ and g_j is uniformly continuous on $T \times [0,c]$. This implies that $g_j(s_\ell, q) \to g_j(s,q)$ as $\ell \to \infty$ uniformly for $q \in [0,c]$ and so $h_j(s_\ell) \to h_j(s)$ for $\ell \to \infty$. Finally set $h(s) = \min\{h_1(s), h_2(s)\}$. Then $h \in C^b_+(S)$ and $h(s) > 0$ for all $s \in S$.

By (44),

$$\kappa^\mu(T,s) \geq P(T,s)h(s) =: \tilde{\kappa}(T,s).$$

Since P is a strongly top-irreducible kernel and h is strictly positive, $\tilde{\kappa}$ is a strongly top-irreducible kernel by Proposition 14. In particular, for each $\mu \in \mathcal{M}_+(S)$, $\kappa^\mu(S,s) \geq \tilde{\kappa}(S,s) > 0$ for all $s \in S$.

Theorem 25 *Let the Assumptions 20 and 21 be satisfied, P_2 be a uniform Feller kernel, $P_1 + P_2$ be top-irreducible and $r = \mathbf{r}(\kappa^o) > 1$.*

Then there exists some strictly positive eigenfunction $f \in \dot{C}^b_+(S)$ with $\int_S f(t)\kappa^o(dt,s) = rf(s)$ for all $s \in S$.

Further, the semiflow generated by F is uniformly weakly persistent: There exists some $\delta > 0$ such that $\limsup_{n\to\infty} F^n(\mu)(S) \geq \delta$ for all $\mu \in \dot{\mathcal{M}}^s_+(S)$.

Proof By Proposition 10, κ^o_2 is a uniform Feller kernel. By Proposition 7, κ^o_j is a tight Feller kernel, $j = 1, 2$. By Proposition 22, κ^o is a top-irreducible Feller kernel. By Lemma 7, the kernel family $\{\kappa^\mu; \mu \in \mathcal{M}^s_+(S)\}$ is lower semicontinuous at the zero measure.

Assumption 9 is satisfied by Proposition 21, and Assumption 10 is satisfied by Lemma 8. By Lemma 7, $\mathbf{r}(\kappa^o_1) \leq 1$. Apply Theorem 6. □

Recall Definition 15.

Assumption 22 (a) For any closed totally bounded subset T of S, $\{q_j(s,\cdot); s \in T\}$ is equicontinuous on S, $j = 1, 2$.

(b) For any closed totally bounded subset T of S, $\{g_j(s,\cdot); s \in T\}$ is uniformly equicontinuous on bounded subsets of $\mathbb{R}$, $j = 1, 2$.

Lemma 9 *Assumption 22 is satisfied if S be completely metrizable, and Assumption 20 holds.*

Proof Let T be a closed totally bounded subset of S and let S be completely metrizable. Then T is compact.

(b) Let $c > 0$. Then the set $T \times [0,c]$ is compact. Since g_j is continuous on $S \times \mathbb{R}_+$, g_j is uniformly continuous on $T \times [0,c]$. This implies (b).

(a) Suppose that Assumption 22(a) is false for $j = 1$ or $j = 2$. Then there is some $\tilde{s} \in S$ such that $\{q_j(s,\cdot); s \in T\}$ is not equicontinuous at $\tilde{s}$.

Then there exists some $\epsilon > 0$ and a sequence (s_n) in T and a sequence $(\tilde{s}_n)$ in S such that $\tilde{s}_n \to \tilde{s}$ as $n \to \infty$ and

$$|q_j(s_n, \tilde{s}_n) - q_j(s_n, \tilde{s})| > \epsilon, \qquad n \in \mathbb{N}.$$

Since $T \times \big(\{\tilde{s}_n; n \in \mathbb{N}\} \cup \{\tilde{s}\}\big)$ is a compact subset of S^2, q_j is uniformly continuous on this set, a contradiction.

Lemma 10 *Let the Assumptions 20 and 22(a) be satisfied. Further let* $\mu \in \mathcal{M}_+^s(S)$ *and* (μ_n) *be a sequence in* $\mathcal{M}_+^s(S)$, $\|\mu_n - \mu\|_\flat \to 0$ *as* $n \to \infty$.

Then, for $j = 1, 2$, $(Q_j\mu_n)(s) \to (Q_j\mu)(s)$ *as* $n \to \infty$ *uniformly for* s *in any closed totally bounded subset of* S. *Further* $Q_j\mu_n$ *and* $Q_j\mu$ *are bounded functions.*

Proof The convergence statement follows from Proposition 2 and (41) and (42). The boundedness statements are immediate.

Lemma 11 *Let the Assumptions 20 and 22(b) be satisfied. Let* T *be a closed totally bounded subset of* S *and* $f_n : T \to \mathbb{R}_+$, $n \in \mathbb{N}$, *and* $f : T \to \mathbb{R}_+$ *be bounded functions such that* $f_n \to f$ *uniformly on* T. *Then* $g_j\big(s, f_n(s)\big) \to g_j\big(s, f(s)\big)$ *as* $n \to \infty$ *uniformly for* $s \in T$.

Proof There exists some $c \in (0, \infty)$ such that $f_n(s), f(s) \le c$ for all $n \in \mathbb{N}, s \in T$. Since $\{g_j(s, \cdot); s \in T\}$ is uniformly equicontinuous on $[0, c]$, the assertion follows.

Proposition 23 *Let the Assumptions 20 and 22 be satisfied. Then Assumption 11 is satisfied for* κ_j^μ, $j = 1, 2$ *and* κ^μ.

Proof It is sufficient to show the claim for κ_1^μ. Let (μ_n) be a sequence in $\mathcal{M}_+^s(S)$ and $\mu \in \mathcal{M}_+^s(S)$ such that $\int_S f d\mu_n \to \int_S f d\mu$ as $n \to \infty$ for all $f \in C_+^b(S)$. Then $\|\mu_n - \mu\|_\flat \to 0$ as $n \to \infty$ by Theorem 12.

Let $h \in C_+^b(S)$. For $s \in S$,

$$\begin{aligned}&\int_S h(t)\kappa_1^{\mu_n}(dt, s) - \int_S h(t)\kappa_1^{\mu}(dt, s)\\ &= \int_S h(t)P_1(dt, s)\big[g_1\big(s, (Q_1\mu_n)(s)\big) - g_1\big(s, (Q_1\mu)(s)\big)\big].\end{aligned}$$

Since $\int_S h(t)P_1(dt, s) \le \sup h(S)$, it is sufficient to show that

$$g_1\big(s, (Q_1\mu_n)(s)\big) \to g_1\big(s, (Q_1\mu)(s)\big), \qquad n \to \infty$$

uniformly on every closed totally bounded subset T of S. But this follows by combining Lemmas 10 and 11.

Assumption 23 $\sup_{s\in S, q\ge 0} P_1(S, s)g_1(s, q) < 1$; $\inf_{s,t\in S} q_2(s, t) > 0$;

$$g_2(s, q) \to 0 \text{ as } q \to \infty, \text{ uniformly for } s \in S.$$

From the interpretation of g_1 as probability of surviving competition, it is suggestive that $0 \le g_1(s, q) \le 1$ (Assumption 20 g1). So, together with $P_1(S, s) \le 1$, the first of the assumptions is not really drastic. The second assumption means that competitive influence on somebody else's reproduction reaches everywhere in the habitat. The third assumption means that fertility drops very low if resources are very low due to large competition.

Proposition 24 *Under the Assumptions 23,*

$$\sup_{\mu\in\mathcal{M}_+^s(S)}\sup_{s\in S}\kappa_1^\mu(S,s)<1,\qquad \sup_{s\in S}\kappa_2^\mu(S,s)\to 0\ as\ \mu(S)\to\infty,$$

and Assumption 13 is satisfied. Further $\mathbf{r}(\kappa_1^o)<1$.

Proof Recall that

$$\kappa_1^\mu(S,s)=P_1(S,s)g_1\big(s,(Q_1\mu)(s)\big)\le P_1(S,s)\sup_{q\in\mathbb{R}_+}g_1(s,q),$$

which implies the first assertion. Further

$$(Q_2\mu)(s)\ge\inf_{s,t\in S}q_2(s,t)\,\mu(S)\overset{\mu(S)\to\infty}{\longrightarrow}\infty$$

uniformly for $s\in S$, and so

$$\kappa_2^\mu(S,s)=P_2(S,s)g_2\big(s,(Q_\mu)(s)\big)\to 0,\qquad \mu(S)\to\infty,$$

uniformly for $s\in S$. We combine,

$$\begin{aligned}\limsup_{\mu(S)\to\infty}\sup_{s\in S}\kappa^\mu(S,s)&\le\sup_{\mu\in\mathcal{M}_+^s(S),\,s\in S}\kappa_1^\mu(S,s)+\limsup_{\mu(S)\to\infty}\sup_{s\in S}\kappa_2^\mu(S,s)\\&=\sup_{\mu\in\mathcal{M}_+^s(S),\,s\in S}\kappa_1^\mu(S,s)<1.\end{aligned}$$

Theorem 26 *Let the Assumptions 20, 22 and 23 be satisfied. Assume that* P_1 *and* P_2 *are tight Feller kernels and* $\mathbf{r}(\kappa^o)>1$.

Then there exists a fixed point $F(\mu)=\mu\in\dot{\mathcal{M}}_+^s(S)$.

Proof We apply [32, Theorem 3.19]. Its assumptions are satisfied by Lemma 7, Propositions 21, 23 and 24.

Theorem 27 *Let the Assumptions 20, 21, 22 and 23 be satisfied. Assume that* P_2 *is a uniform Feller kernel,* P_1+P_2 *is top-irreducible and* $\mathbf{r}(\kappa^o)>1$.

Then the population is uniformly persistent in the following sense: There exists some $\epsilon_0>0$ *such that* $\liminf_{n\to\infty}F^n(\mu)(S)\ge\epsilon_0$ *for all* $\mu\in\dot{M}_+^s(S)$.

Proof We apply Theorem 7. Its assumptions are satisfied by Lemma 7, Propositions 21, 23, 24. □

Theorem 28 *Let the Assumptions 20, 21, 22 and 23 be satisfied. Assume that* P_2 *is a uniform Feller kernel,* P_1+P_2 *is strongly top-irreducible and* $\mathbf{r}(\kappa^o)>1$.

Then the semiflow induced by F *is uniformly persistent in the following sense: For each* $f\in\dot{C}_+^b(S)$, *there exists some* $\epsilon_f>0$ *with the following property:*

If $\mathcal{N}$ is a compact (or bounded tight) subset of $\mathcal{M}_+^s(S)$ with $\inf_{\mu\in\mathcal{N}} \mu(S) > 0$, there exists some $N \in \mathbb{N}$ such that

$$\int_S f \, dF^n(\mu) \geq \epsilon_f \quad \text{for all } \mu \in \mathcal{N} \text{ and all } n \in \mathbb{N} \text{ with } n > N.$$

Proof We apply Theorem 8. Its assumptions are satisfied by Lemma 7, Propositions 21, 23, 24, Lemma 8. □

Acknowledgements The author thanks Azmy Ackleh for useful hints and Odo Diekmann and Eugenia Franco for helpful suggestions.

References

1. Ackleh, A.S., Colombo, R.M., Goatin, P., Hille, S.C., Muntean, A. (guest editors): Mathematical modeling with measures. Math. Biosci. Eng. **17**(special issue) (2020)
2. Ackleh, A.S., Colombo, R.M., Hille, S.C., Muntean, A. (guest editors): Modeling with measures. Math. Biosci. Eng. **12**(special issue) (2015)
3. Aliprantis, C.D., Border, K.C.: Infinite Dimensional Analysis. A Hitchhiker's Guide, 3rd edn. Springer, Berlin (1999, 2006)
4. Bonsall, F.F.: Linear operators in complete positive cones. Proc. Lond. Math. Soc. **8**, 53–75 (1958)
5. Cushing, J.M., Zhou, Y.: The net reproductive value and stability in matrix population models. Nat. Res. Mod. **8**, 297–333 (1994)
6. Cushing, J.M.: On the relationship between r and R_0 and its role in the bifurcation of stable equilibria of Darwinian matrix models. J. Biol. Dyn. **5**, 277–297 (2011)
7. Desch, W., Schappacher, W.: Linearized stability for nonlinear semigroups. In: Favini, A., Obrecht, E. (eds.) Differential Equations in Banach Spaces. Lecture Notes in Mathematics, vol. 1223, pp. 61–67. Springer, Berlin (1986)
8. Diekmann, O., Gyllenberg, M., Metz, J.A.J., Thieme, H.R.: The 'cumulative' formulation of (physiologically) structured population models. In: Clément, Ph., Lumer, G. (eds.) Evolution Equations, Control Theory, and Biomathematics, pp. 145–154. Dekker, Marcel (1994)
9. Diekmann, O., Heesterbeek, J.A.P., Metz, J.A.J.: On the definition and the computation of the basic reproduction number R_0 in models for infectious diseases in heterogeneous populations. J. Math. Biol. **28**, 365–382 (1990)
10. Dudley, R.M.: Convergence of Baire measures. Stud. Math. **27**, 251–268 (1966). (Correction to "Convergence of Baire measures". Stud. Math. **51**, 275 (1974))
11. Dudley, R.M.: Real Analysis and Probability, 2nd edn. Cambridge University Press, Cambridge (2002)
12. Dunford, N., Schwartz, J.T.: Linear Operators. Part I. General Theory, John Wiley, Classics Library Edition, New York (1988)
13. Gel'fand, I.M.: Normierte Ringe. Mat Sbornik NS **9**, 3–24 (1941)
14. Gwiazda, P., Marciniak-Czochra, A., Thieme, H.R.: Measures under the flat norm as ordered normed vector space. Positivity **22**, 105–138 (2018). (Correction. Positivity **22**, 139–140 (2018))
15. Hille, S.C., Worm, D.T.H.: Continuity properties of Markov semigroups and their restrictions to invariant L^1 spaces. Semigroup Forum **79**, 575–600 (2009)
16. Hille, S.C., Worm, D.T.H.: Embedding of semigroups of Lipschitz maps into positive linear semigroups on ordered Banach spaces generated by measures. Integral Equ. Oper. Theory **63**, 351–371 (2009)

17. Jin, W., Smith, H.L., Thieme, H.R.: Persistence versus extinction for a class of discrete-time structured population models. J. Math. Biol. **72**, 821–850 (2016)
18. Jin, W., Smith, H.L., Thieme, H.R.: Persistence and critical domain size for diffusing populations with two sexes and short reproductive season. J. Dyn. Differ. Eqn. **28**, 689–705 (2016)
19. Jin, W., Thieme, H.R.: Persistence and extinction of diffusing populations with two sexes and short reproductive season. Discret. Contin. Dyn. Syst. B **19**, 3209–3218 (2014)
20. Jin, W., Thieme, H.R.: An extinction/persistence threshold for sexually reproducing populations: the cone spectral radius. Discret. Contin. Dyn. Syst. B **21**, 447–470 (2016)
21. Mallet-Paret, J., Nussbaum, R.D.: Eigenvalues for a class of homogeneous cone maps arising from max-plus operators. Disc. Cont. Dyn. Syst. A **8**, 519–562
22. Mallet-Paret, J., Nussbaum, R.D.: Generalizing the Krein-Rutman theorem, measures of noncompactness and the fixed point index. J. Fixed Point Theory Appl. **7**, 103–143 (2010)
23. McDonald, J.N., Weiss, N.A.: A Course in Real Analysis. Academic Press, San Diego (1999)
24. Nussbaum, R.D.: Eigenvectors of nonlinear positive operators and the linear Krein-Rutman theorem. In: Fadell, E., Fournier, G. (eds.) Fixed Point Theory, pp. 309–331. Springer, Berlin (1981)
25. Schaefer, H.H.: Banach Lattices and Positive Operators. Springer, Berlin (1974)
26. Smith, H.L., Thieme, H.R.: Dynamical Systems and Population Persistence. American Mathmetical Society, Providence (2011)
27. Shurenkov, V.M.: On the relationship between spectral radii and Perron Roots. Chalmers University of Technology and Göteborg University (preprint)
28. Thieme, H.R.: Spectral radii and Collatz-Wielandt numbers for homogeneous order-preserving maps and the monotone companion norm. In: de Jeu, M., de Pagter, B., van Gaans, O., Veraar, M. (eds.) Ordered Structures and Applications. Trends in Mathematics, pp. 415–467. Birkhäuser/Springer, Cham (2016)
29. Thieme, H.R.: Eigenfunctionals of homogeneous order-preserving maps with applications to sexually reproducing populations. J. Dyn. Differ. Eqn. **28**, 1115–1144 (2016)
30. Thieme, H.R.: Eigenvectors of homogeneous order-bounded order-preserving maps. Discret. Contin. Dyn. Syst. B **22**, 1073–1097 (2017)
31. Thieme, H.R.: From homogeneous eigenvalue problems to two-sex population dynamics. J. Math. Biol. **75**, 783–804 (2017)
32. Thieme, H.R.: Discrete-time population dynamics on the state space of measures. Math. Biosci. Eng. **17**, 1168–1217 (2020). https://doi.org/10.3934/mbe.2020061
33. Zhao, X.-Q.: Dynamical Systems in Population Biology. Springer, New York (2003)

Discrete Splines and Its Applications

Patricia J. Y. Wong

Abstract In this paper, we survey the contributions made to discrete splines in the literature and present some applications of discrete splines in the numerical treatment of boundary value problems.

Keywords Discrete spline interpolation · Error estimates · Numerical solution · Boundary value problems

1 Introduction

In the well familiar continuous spline interpolation, we not only interpolate the function of interest at each knot, but also interpolate a number of derivatives of the function at certain knots. Therefore, the function is required to be sufficiently smooth. However, in the real world situation, not only that it may be difficult to compute the derivatives of a function, the derivatives may not even exist at some points. In such a situation, the usual continuous spline interpolation will not be suitable. We therefore introduce 'discrete' spline interpolation schemes that involve only differences. Since no derivatives are involved, the interpolates can be constructed for a more general class of function and therefore this type of interpolation has a wider range of applications.

Discrete splines are piecewise polynomials where continuity of differences rather than derivatives are satisfied at the joining knots of the polynomial pieces. The difference operator used may be forward difference operator [18, 70] or central difference operator [20–23, 41, 42]. In contrast, the continuity conditions of the familiar *continuous splines* at the joining knots are in terms of *derivatives*.

Discrete splines were first introduced by Mangasarian and Schumaker [43] in 1971 as solutions to constrained minimization problems in real Euclidean space, which

P. J. Y. Wong (✉)
School of Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore
e-mail: ejywong@ntu.edu.sg

S. Baigent et al. (eds.), *Progress on Difference Equations and Discrete Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 341,
https://doi.org/10.1007/978-3-030-60107-2_5

are discrete analogs of minimization problems in Banach space whose solutions are generalized splines. The solutions to the constrained minimization problems in real Euclidean space exhibit a spline-like structure and therefore they are termed 'discrete splines' in [43].

In subsequent sections, we shall present discrete quintic spline interpolation, periodic discrete quintic spline interpolation, as well as their error analysis. The two-variable cases are also tackled.

Discrete splines have been used in the numerical treatment of boundary value problems and integral equations. There is an advantage of spline method over finite difference method—once the spline solution is obtained, any information between the mesh points becomes immediately available. Indeed, applications of discrete splines to Fredholm integral equations as well as to second order and fourth order boundary value problems have been investigated in [19, 21–23].

The outline of this paper is as follows. In Sect. 2, we present the discrete quintic spline interpolation, the periodic discrete quintic spline interpolation and their error analysis. In Sects. 3 and 4, we illustrate the application of discrete splines in the numerical treatment of second order boundary value problems. Finally, a brief conclusion is given in Sect. 5.

2 Discrete Spline Interpolation and Error Estimates

After the work of Mangasarian and Schumaker [43], subsequent investigations on discrete splines can be found in the work of Schumaker [59], Astor and Duris [13], Lyche [41, 42] and Wong et al. [18, 20, 70].

There are two basic approaches to developing splines—the *variational approach* wherein splines are defined as the solutions of certain constrained minimization problems, and the *constructive approach* wherein they are defined by piecing together classes of functions at certain knots. In the very first paper on discrete splines [43], the variational approach has been used and discrete splines are introduced as solutions to constrained minimization problems in real Euclidean space. These same discrete splines also play a fundamental role in certain best summation formulae for a finite sequence of real numbers [44]. On the other hand, the constructive approach has been employed in the work of [13, 41, 42, 59]. Both Schumaker [59] and Lyche [41, 42] deal with discrete polynomial splines. In [59], discrete B-splines, which are discrete analogs of the classical B-splines, are explored to give the general construction of discrete splines—here *forward differences* are involved. In comparison, the discrete cubic spline discussed in [41] involves *central differences*. Another work by Lyche [42] investigates several discrete spline approximation methods for fitting functions and data, the respective error analysis shows that some results in the continuous case can be obtained from the discrete analog. On a separate note, in [13] discrete L-splines are constructively defined so as to parallel the development of continuous L-splines.

Motivated by the earlier research on discrete splines, in [70] we have developed a *discrete cubic spline* via constructive approach, while in [18] a *discrete quintic spline* is developed also via constructive approach. Our definition of discrete spline involves *forward differences* and is in the spirit of that in [59]. However, the method of construction is different—we derive matrix equations which can be solved uniquely to obtain the discrete spline, which, unlike [42, 59], is *not* in terms of B-splines. Our approach is parallel to the technique in [41]. The main contributions in the work [18, 70] are (i) the development of a class of discrete cubic/quintic spline interpolation and the derivation of explicit error estimates between the function and its discrete spline interpolate; and (ii) the extension to two-variable discrete bi-cubic/bi-quintic spline interpolation and the related error estimates.

The work [18, 70] naturally complements the literature and especially the work of [41, 42, 59]. We remark that in generalizing the cubic case [70] to the quintic case [18], there is a quantum leap in the level of complexity. Moreover, the papers [18, 70] have extended the work of [1, 67, 68] on continuous spline to discrete case. It is also noted that [1, 67, 68] extends the work of Schultz [58], and other work on different types of continuous splines [1, 2, 27, 33, 38, 57, 60, 64, 69].

In [20], we have developed a class of *periodic discrete quintic spline* involving *central differences* and establish the related existence, uniqueness and error estimates. The two-variable case has also been considered. Our work naturally extends the literature and especially complements and/or extends the work of [28, 29, 40, 42, 52] on *one-variable* discrete cubic splines. We also extend the research of [1, 67] on continuous spline to discrete case, as well as complement the work of [18, 70] on discrete cubic/quintic splines involving *forward differences*.

2.1 Discrete Quintic Splines

This section illustrates the work of [18]. Let a, b, c, d $(b > a,\ d > c)$ be integers. We shall denote the discrete interval $N[a, b] = \{a, a + 1, \cdots, b\}$. Throughout, let

$$\rho : a = k_1 < k_2 < \cdots < k_m = b,\ \ k_i \in \mathbf{Z},\ 1 \le i \le m\ \ (m \ge 7)$$

and

$$\rho' : c = l_1 < l_2 < \cdots < l_n = d,\ \ l_j \in \mathbf{Z},\ 1 \le j \le n\ \ (n \ge 7)$$

be uniform partitions of $N[a, b]$ and $N[c, d]$ respectively with step sizes

$$h = k_{i+1} - k_i\ (\ge 6),\ 1 \le i \le m-1 \quad \text{and} \quad h' = l_{j+1} - l_j\ (\ge 6),\ 1 \le j \le n-1.$$

Further, we let $\tau = \rho \times \rho'$ be a rectangular partition of $N[a, b] \times N[c, d]$. The symbol Δ, as usual, denotes the forward difference operator with step size 1. For $x \in \mathbb{R}$ and k a nonnegative integer, the factorial expression $x^{(k)} = \prod_{i=0}^{k-1}(x - i)$, and we use the convention $0^{(0)} = 1$. For a function $f(x)$ defined on $N[a, b + 2]$,

we define the usual ℓ_∞ norm as $\|f\| = \max_{x\in N[a,b+2]} |f(x)|$. Similarly, for a function $f(x, y)$ defined on $N[a, b+2] \times N[c, d+2]$, the usual ℓ_∞ norm is defined as $\|f\| = \max_{(x,y)\in N[a,b+2]\times N[c,d+2]} |f(x, y)|$.

Definition 1 Let $j \in \mathbb{N} (< h)$ be fixed. Let $f_i(x)$ be defined on $N[k_i, k_{i+1} + j]$, $1 \le i \le m-2$, and $f_{m-1}(x)$ be defined on $N[k_{m-1}, b+2]$. Let $f(x) \equiv \cup_{1\le i\le m-1} f_i(x)$. We say that $f(x) \in D^{(j)}[a, b]$ if

$$\Delta^l f_{i-1}(k_i) = \Delta^l f_i(k_i), \quad 2 \le i \le m-1,\ 0 \le l \le j. \tag{1}$$

Note that (1) is also equivalent to

$$f_{i-1}(k_i + l) = f_i(k_i + l), \quad 2 \le i \le m-1,\ 0 \le l \le j. \tag{2}$$

Hence, the function $f(x) = \cup_{1\le i\le m-1} f_i(x)$ is well defined on $N[a, b+2]$. The set $D^{(p,q)}([a, b] \times [c, d])$ where $p, q \in \mathbb{N}$, $p < h$, $q < h'$, is analogously defined.

Define the set $H(\rho) = \{g(x) \in D^{(2)}[a, b] : g(x)$ is a quintic polynomial in each subinterval $N[k_i, k_{i+1}]$, $1 \le i \le m-2$ and $N[k_{m-1}, b+2]\}$. Clearly, $H(\rho)$ is of dimension $3m$. The next lemma gives a basis for $H(\rho)$.

Lemma 1 *[17] The functions $h_i(x)$, $\bar{h}_i(x)$ and $\bar{\bar{h}}_i(x)$, $1 \le i \le m$ form a basis of $H(\rho)$. Here, for $1 \le i, j \le m$,*

$$\begin{aligned} &h_i(k_j) = \delta_{ij}, \quad \Delta h_i(k_j) = \Delta^2 h_i(k_j) = 0,\\ &\Delta\bar{h}_i(k_j) = \delta_{ij}, \quad \bar{h}_i(k_j) = \Delta^2\bar{h}_i(k_j) = 0,\\ &\Delta^2\bar{\bar{h}}_i(k_j) = \delta_{ij}, \ \bar{\bar{h}}_i(k_j) = \Delta\bar{\bar{h}}_i(k_j) = 0. \end{aligned}$$

The explicit expressions of $h_i(x)$, $\bar{h}_i(x)$ and $\bar{\bar{h}}_i(x)$, $1 \le i \le m$ can be computed directly.

We are now ready to develop discrete spline interpolation. To begin, we define the set $S(\rho) = \{g(x) \in D^{(4)}[a, b] : g(x)$ is a quintic polynomial in each subinterval $N[k_i, k_{i+1}]$, $1 \le i \le m-2$ and $N[k_{m-1}, b+2]\}$. Clearly, $S(\rho)$ is of dimension $(m+4)$.

Definition 2 For a given function $f(x)$ defined on $N[a, b+2]$, we say $S_\rho f(x)$ is the *$S(\rho)$-interpolate of $f(x)$*, also known as the *discrete spline interpolate of $f(x)$*, if $S_\rho f(x) \in S(\rho)$ with $S_\rho f(k_i) = f(k_i)$, $1 \le i \le m$, and $\Delta^j S_\rho f(k_1) = \Delta^j f(k_1)$, $\Delta^j S_\rho f(k_m) = \Delta^j f(k_m)$, $j = 1, 2$.

Lemma 2 *For a given $g(x) \in H(\rho)$, we define $c_i = g(k_i)$, $\Delta c_i = \Delta g(k_i)$, $\Delta^2 c_i = \Delta^2 g(k_i)$, $1 \le i \le m$. Then, $g(x) \in S(\rho)$ if and only if the vectors $\Delta c = [\Delta c_i]_{i=2}^{m-1}$ and $\Delta^2 c = [\Delta^2 c_i]_{i=2}^{m-1}$ satisfy the matrix equations*

$$B^1(\Delta c) = w^1 \quad \text{and} \quad B^2(\Delta^2 c) = w^2, \tag{3}$$

where $B^1,\ B^2$ *are 5-band diagonal* $(m-2)\times(m-2)$ *matrices whose elements are in terms of* h; $w^1,\ w^2$ *are* $(m-2)\times 1$ *vectors whose elements are in terms of* $c_i,\ 1\le i\le m,\ \Delta c_1,\ \Delta c_m,\ \Delta^2 c_1$ *and* $\Delta^2 c_m$. *Moreover,* B^1 *and* B^2 *are invertible, hence from (3) the values of* $\Delta c_i,\ \Delta^2 c_i,\ 2\le i\le m-1$ *can be obtained uniquely in terms of* $c_i,\ 1\le i\le m,\ \Delta c_1,\ \Delta c_m,\ \Delta^2 c_1$ *and* $\Delta^2 c_m$.

The next result gives an explicit expression of the discrete spline interpolate $S_\rho f$.

Theorem 1 *Let* $f(x)$ *be defined on* $N[a,b+2]$. *If (3), with* $c_i=f(k_i),\ 1\le i\le m$ *and* $\Delta^j c_\ell=\Delta^j f(k_\ell),\ \ell=1,m,\ j=1,2$, *has unique solutions* Δc *and* $\Delta^2 c$, *then* $S_\rho f(x)$ *exists and is unique. Moreover,* $S_\rho f(x)$ *can be expressed as*

$$\begin{aligned} S_\rho f(x)=&\sum_{i=1}^{m} f(k_i)h_i(x)+\Delta f(k_1)\bar{h}_1(x)+\Delta f(k_m)\bar{h}_m(x)+\sum_{i=2}^{m-1}(\Delta c_i)\bar{h}_i(x)\\ &+\Delta^2 f(k_1)\bar{\bar{h}}_1(x)+\Delta^2 f(k_m)\bar{\bar{h}}_m(x)+\sum_{i=2}^{m-1}(\Delta^2 c_i)\bar{\bar{h}}_i(x). \end{aligned} \tag{4}$$

Remark 1 We can describe a basis for $S(\rho)$, namely the 'cardinal splines', $\{s_i(x)\}_{i=1}^{m+4}$, which are defined by the following interpolating conditions

$$\begin{aligned} &s_i(k_j)=\delta_{ij},\ \Delta s_i(a)=\Delta s_i(b)=0,\ \Delta^2 s_i(a)=\Delta^2 s_i(b)=0,\ \ 1\le i,j\le m\\ &s_{m+1}(k_i)=0,\ \Delta s_{m+1}(a)=1,\ \Delta s_{m+1}(b)=0,\ \Delta^2 s_{m+1}(a)=\Delta^2 s_{m+1}(b)=0,\\ &s_{m+2}(k_i)=0,\ \Delta s_{m+2}(a)=0,\ \Delta s_{m+2}(b)=1,\ \Delta^2 s_{m+2}(a)=\Delta^2 s_{m+2}(b)=0,\\ &s_{m+3}(k_i)=0,\ \Delta^2 s_{m+3}(a)=1,\ \Delta^2 s_{m+3}(b)=0,\ \Delta s_{m+3}(a)=\Delta s_{m+3}(b)=0,\\ &s_{m+4}(k_i)=0,\ \Delta^2 s_{m+4}(a)=0,\ \Delta^2 s_{m+4}(b)=1,\ \Delta s_{m+4}(a)=\Delta s_{m+4}(b)=0,\\ &\qquad\qquad 1\le i\le m. \end{aligned}$$

The discrete spline interpolate $S_\rho f(x)$ can also be explicitly expressed as

$$\begin{aligned} S_\rho f(x)=&\sum_{i=1}^{m} f(k_i)s_i(x)+\Delta f(a)s_{m+1}(x)+\Delta f(b)s_{m+2}(x)\\ &+\Delta^2 f(a)s_{m+3}(x)+\Delta^2 f(b)s_{m+4}(x). \end{aligned} \tag{5}$$

We are now ready to introduce two-variable discrete spline interpolation. For a given $\tau\ (=\rho\times\rho')$, we define $S(\tau)=S(\rho)\oplus S(\rho')$ (the tensor product) = Span $\left\{s_i(x)s_j(y)\right\}_{i=1\ \ j=1}^{m+4\ \ n+4}=\{g(x,y)\in D^{(4,4)}([a,b]\times[c,d]) : g(x,y)$ is a two-dimensional polynomial of degree 5 in each variable and in each subrectangle $N[k_i,k_{i+1}]\times N[l_j,l_{j+1}],\ N[k_{m-1},b+2]\times N[l_j,l_{j+1}],\ N[k_i,k_{i+1}]\times N[l_{n-1},d+2],\ 1\le i\le m-2,\ 1\le j\le n-2$ and $N[k_{m-1},b+2]\times N[l_{n-1},d+2]\}$. Since $S(\tau)$ is the tensor product of $S(\rho)$ and $S(\rho')$ which are of dimensions $(m+4)$ and $(n+4)$ respectively, $S(\tau)$ is of dimension $(m+4)(n+4)$.

Definition 3 For a given function $f(x,y)$ defined on $N[a,b+2]\times N[c,d+2]$, we shall denote $f_{i,j}^{\mu,\nu}=\Delta_x^\mu\Delta_y^\nu f(k_i,l_j),\ \mu,\nu=0,1,2,\ 1\le i\le m,\ 1\le j\le n$. We

say $S_\tau f(x, y)$ is the $S(\tau)$-*interpolate of* $f(x, y)$, also known as the *discrete spline interpolate of* $f(x, y)$, if $S_\tau f(x, y) \in S(\tau)$ with $\Delta_x^\mu \Delta_y^\nu \, S_\tau f(k_i, l_j) = f_{i,j}^{\mu,\nu}$ where $\mu,\ \nu,\ i$ and j satisfy

(i) if $\mu = \nu = 0$, then $1 \le i \le m,\ 1 \le j \le n$;
(ii) if $\mu = 1, 2,\ \nu = 0$, then $i = 1, m,\ 1 \le j \le n$;
(iii) if $\mu = 0,\ \nu = 1, 2$, then $1 \le i \le m,\ j = 1, n$; and
(iv) if $\mu = 1, 2,\ \nu = 1, 2$, then $(i, j) = (1, 1), (1, n), (m, 1), (m, n)$.

The next result gives an explicit expression of the two-variable discrete spline interpolate $S_\tau f$.

Theorem 2 *For any function* $f(x, y)$ *defined on* $N[a, b+2] \times N[c, d+2]$, $S_\tau f(x, y)$ *exists and is unique. Further,* $S_\tau f(x, y)$ *can be explicitly expressed in terms of cardinal splines as*

$$\begin{aligned} S_\tau f(x, y) &= \sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j}^{0,0} s_i(x) s_j(y) \\ &+ \sum_{i=1}^{m} \left[f_{i,1}^{0,1} s_{n+1}(y) + f_{i,n}^{0,1} s_{n+2}(y) + f_{i,1}^{0,2} s_{n+3}(y) + f_{i,n}^{0,2} s_{n+4}(y) \right] s_i(x) \\ &+ \sum_{i=1}^{n} \left[f_{1,i}^{1,0} s_{m+1}(x) + f_{m,i}^{1,0} s_{m+2}(x) + f_{1,i}^{2,0} s_{m+3}(x) + f_{m,i}^{2,0} s_{m+4}(x) \right] s_i(y) \\ &+ f_{1,1}^{1,1} s_{m+1}(x) s_{n+1}(y) + f_{1,n}^{1,1} s_{m+1}(x) s_{n+2}(y) + f_{m,1}^{1,1} s_{m+2}(x) s_{n+1}(y) \\ &+ f_{m,n}^{1,1} s_{m+2}(x) s_{n+2}(y) + f_{1,1}^{2,1} s_{m+3}(x) s_{n+1}(y) + f_{1,n}^{2,1} s_{m+3}(x) s_{n+2}(y) \\ &+ f_{m,1}^{2,1} s_{m+4}(x) s_{n+1}(y) + f_{m,n}^{2,1} s_{m+4}(x) s_{n+2}(y) + f_{1,1}^{1,2} s_{m+1}(x) s_{n+3}(y) \\ &+ f_{1,n}^{1,2} s_{m+1}(x) s_{n+4}(y) + f_{m,1}^{1,2} s_{m+2}(x) s_{n+3}(y) + f_{m,n}^{1,2} s_{m+2}(x) s_{n+4}(y) \\ &+ f_{1,1}^{2,2} s_{m+3}(x) s_{n+3}(y) + f_{1,n}^{2,2} s_{m+3}(x) s_{n+4}(y) + f_{m,1}^{2,2} s_{m+4}(x) s_{n+3}(y) \\ &+ f_{m,n}^{2,2} s_{m+4}(x) s_{n+4}(y). \end{aligned} \tag{6}$$

The next theorem gives the error estimates for the discrete spline interpolation in one variable and two variables. In the one-variable case, the idea is to use the inequality

$$\|f - S_\rho f\| \le \|f - H_\rho f\| + \|H_\rho f - S_\rho f\|$$

where $H_\rho f$ is the discrete Hermite interpolate of f [17]. In the two-variable case, the key inequality used is

$$\|f - S_\tau f\| \le \|f - S_\rho f\| + \|S_\rho (f - S_{\rho'} f)\| + \|f - S_{\rho'} f\|.$$

Theorem 3

(a) Let $f(x)$ *be defined on* $N[a, b+2]$. *Then*

$$\|f - S_\rho f\| \le d_j(h) \max_{x \in N[a, b+2-j]} |\Delta^j f(x)|,\ 2 \le j \le 6 \tag{7}$$

where $d_j(h),\ 2 \le j \le 6$ *are in terms of* h *and are explicitly known. Further, it is known that* $d_j(h) = O(h^j),\ 2 \le j \le 6$.

(b) Let $f(x, y)$ *be defined on* $N[a, b+2] \times N[c, d+2]$. *Then, we have the following error estimates*

$$\begin{aligned} \|f - S_\tau f\| \le\ & d_6(h)\|\Delta_x^6 f(x, y)\| + d_2(h)d_4(h')\|\Delta_x^2\Delta_y^4 f(x, y)\| \\ & + d_6(h')\|\Delta_y^6 f(x, y)\|, \end{aligned} \tag{8}$$

$$\begin{aligned} \|f - S_\tau f\| \le\ & d_6(h)\|\Delta_x^6 f(x, y)\| + d_3(h)d_3(h')\|\Delta_x^3\Delta_y^3 f(x, y)\| \\ & + d_6(h')\|\Delta_y^6 f(x, y)\|, \end{aligned} \tag{9}$$

$$\begin{aligned} \|f - S_\tau f\| \le\ & d_6(h)\|\Delta_x^6 f(x, y)\| + d_4(h)d_2(h')\|\Delta_x^4\Delta_y^2 f(x, y)\| \\ & + d_6(h')\|\Delta_y^6 f(x, y)\| \end{aligned} \tag{10}$$

where $\|\Delta_x^\mu\Delta_y^\nu f(x, y)\| = \max_{(x,y)\in N[a,b+2-\mu]\times N[c,d+2-\nu]} |\Delta_x^\mu\Delta_y^\nu f(x, y)|$. *Further, since* $d_j(h) = O(h^j),\ j = 2, 3, 4, 6$, *the error bounds (8)–(10) are of* $O(\hat{h}^6)$ *where* $\hat{h} = \max\{h, h'\}$.

We shall illustrate the sharpness of the error estimates obtained in Theorem 3 by two numerical examples. In each example, we take a function f and construct its discrete spline interpolate, then we calculate the actual error as well as the respective bound in Theorem 3. We remark that the functions considered in the examples are not differentiable at certain points and therefore cannot be approximated by continuous spline interpolation (which involves derivatives).

Example 1 Consider

$$f(t) = |t|(t^5 - 3t + 1)(t - 8)|t - 6| \ln(t + 1)/10^9$$

with $a = 0$ and $b = 60$.

The steps taken to construct $S_\rho f(t)$ and the related bound are as follows:

(i) For a function $f(t)$ defined on $N[a, b+2]$, fix the partition ρ and the step size h.

(ii) Obtain the values $f(k_i),\ 1 \le i \le m$ and $\Delta^j f(k_\ell),\ \ell = 1, m,\ j = 1, 2$. In (3), with $c_i = f(k_i),\ 1 \le i \le m$ and $\Delta^j c_\ell = \Delta^j f(k_\ell),\ \ell = 1, m,\ j = 1, 2$, solve for $\Delta c = [\Delta c_i]_{i=2}^{m-1}$ and $\Delta^2 c = [\Delta^2 c_i]_{i=2}^{m-1}$.

(iii) We construct $S_\rho f(t)$ in each subinterval $N[k_{i-1}, k_i],\ 2 \le i \le m$ as follows:

$$\begin{aligned} S_\rho f(t) = & f(k_i)h_i(t) + f(k_{i-1})h_{i-1}(t) + \Delta c_i \bar{h}_i(t) + \Delta c_{i-1}\bar{h}_{i-1}(t) \\ & + \Delta^2 c_i \bar{\bar{h}}_i(t) + \Delta^2 c_{i-1}\bar{\bar{h}}_{i-1}(t). \end{aligned}$$

(iv) Compute the actual error

$$\|f - S_\rho f\| = \max_{t\in N[a,b+2]} |f(t) - S_\rho f(t)|.$$

Table 1 (Example 1) Actual errors and error bounds

m	7 ($h = 10$)	11 ($h = 6$)
$\|f - S_\rho f\|$	$0.19143081e+01$	$0.13903024e+00$
Bound	$0.46240560e+02$	$0.54135637e+01$

(v) Obtain the bound in the right side of (7) for $j = 6$.

The results are presentated in Table 1.

Example 2 Consider

$$f(t,u) = \left(1 - e^{tu/400}\right)|t|/100$$

with $a = c = 0$ and $b = d = 48$. For a fixed partition τ, we shall obtain $S_\tau f(t,u)$, the biquintic spline interpolate of $f(t,u)$. Then, we calculate the actual error $\|f - S_\tau f\|$ as well as the bounds in (8)–(10).

To construct $S_\tau f(t,u)$, in view of (6) we need only to construct the cardinal splines $s_i(t),\ 1 \le i \le m+4$ and $s_j(u),\ 1 \le j \le n+4$. To compute a particular cardinal spline say $s_1(t)$, from Remark 1 we know exactly the values of $c_i = s_1(k_i),\ 1 \le i \le m$ and $\Delta^j c_\ell = \Delta^j s_1(k_\ell),\ \ell = 1, m,\ j = 1, 2$, substitute these into the two matrix equations in (3) and solve for the values Δc_i and $\Delta^2 c_i,\ 2 \le i \le m-1$. Then, noting (4) the cardinal spline s_1 has the expression

$$\begin{aligned} s_1(t) = \sum_{i=1}^{m} s_1(k_i)h_i(t) + \Delta s_1(k_1)\bar{h}_1(t) + \Delta s_1(k_m)\bar{h}_m(t) + \sum_{i=2}^{m-1} \Delta c_i \bar{h}_i(t) \\ + \Delta^2 s_1(k_1)\bar{\bar{h}}_1(t) + \Delta^2 s_1(k_m)\bar{\bar{h}}_m(t) + \sum_{i=2}^{m-1} \Delta^2 c_i \bar{\bar{h}}_i(t). \end{aligned}$$

Indeed, from the expressions of $h_i,\ \bar{h}_i$ and $\bar{\bar{h}}_i$, we see that in each subinterval $N[k_{i-1}, k_i],\ 2 \le i \le m$,

$$\begin{aligned} s_1(t) = s_1(k_i)h_i(t) + s_1(k_{i-1})h_{i-1}(t) + \Delta c_i \bar{h}_i(t) + \Delta c_{i-1}\bar{h}_{i-1}(t) \\ + \Delta^2 c_i \bar{\bar{h}}_i(t) + \Delta^2 c_{i-1}\bar{\bar{h}}_{i-1}(t). \end{aligned}$$

Then, we compute the actual error

$$\|f - S_\tau f\| = \max_{(t,u)\in N[a,b+2]\times N[c,d+2]} |f(t,u) - S_\tau f(t,u)|$$

as well as the bounds in (8)–(10). The results are presented in Table 2.

To illustrate graphically, in the following figures we shall plot the case $m = n = 9$. Figure 1 shows the original function and its spline interpolate, due to the close approximation the graphs are presented separately, otherwise they would just appear

Table 2 (Example 2) Actual errors and error bounds

$m\ (=n)$	7 $(h = h' = 8)$	9 $(h = h' = 6)$
$\|f - S_\tau f\|$	$0.66187125e-02$	$0.21483362e-02$
Bound (8)	$0.13637518e+02$	$0.73463482e+01$
Bound (9)	$0.27812870e+02$	$0.15152823e+02$
Bound (10)	$0.16876111e+02$	$0.89818595e+01$

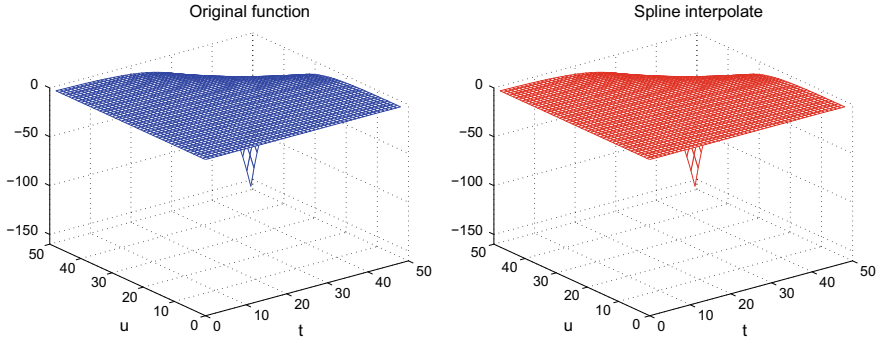

Fig. 1 (Example 2) f and $S_\tau f$ when $m = n = 9$ $(h = h' = 6)$

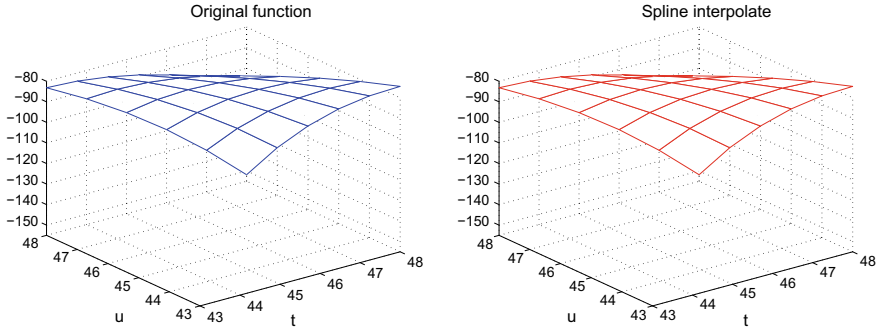

Fig. 2 (Example 2) Enlarged portion of Fig. 1 where the error $|f(t, u) - S_\tau f(t, u)|$ is large

as one graph. Figure 2 shows the portion where the error $|f(t, u) - S_\tau f(t, u)|$ is large, note that the maximum error occurs at $(t, u) = (44, 48)$.

2.2 Periodic Discrete Quintic Splines

Unlike the previous section where forward differences are involved in the discrete spline interpolation, we present another type of spline involving central differences

and it is particularly suitable for the interpolation of periodic function. This section illustrates the work of [20].

For a given $h > 0$, we recall the *central difference operator* D_h applying to a function F gives

$$D_h^{\{0\}}F(x) = F(x); \quad D_h^{\{1\}}F(x) = \frac{1}{2h}[F(x+h) - F(x-h)];$$
$$D_h^{\{2\}}F(x) = \frac{1}{h^2}[F(x+h) - 2F(x) + F(x-h)];$$
$$D_h^{\{3\}}F(x) = \frac{1}{2h^3}[F(x+2h) - 2F(x+h) + 2F(x-h) - F(x-2h)];$$
$$D_h^{\{4\}}F(x) = \frac{1}{h^4}[F(x+2h) - 4F(x+h) + 6F(x) - 4F(x-h) + F(x-2h)].$$

We also use the basic polynomials $x^{\{j\}}$ introduced by Lyche [42]

$$x^{\{j\}} = x^j, \quad j = 0, 1, 2$$

$$x^{\{3\}} = x(x^2 - h^2), \quad x^{\{4\}} = x^2(x^2 - h^2), \quad x^{\{5\}} = x(x^2 - h^2)(x^2 - 4h^2).$$

It is noted that $D_h^{\{1\}}x^{\{j\}} = jx^{\{j-1\}}$, $j = 0, 1, 2, 3, 5$ and $D_h^{\{1\}}x^{\{4\}} = 2x(2x^2 + h^2)$.

Let $a, b, c, d \in \mathbb{R}$ with $a < b$ and $c < d$. We let

$$\varphi : a = t_0 < t_1 \cdots < t_n = b \qquad \text{and} \qquad \varphi' : c = u_0 < u_1 \cdots < u_m = d$$

denote the unform partitions of $[a, b]$ and $[c, d]$ with step sizes $p = \frac{b-a}{n}$ and $p' = \frac{d-c}{m}$ respectively. Further, let $\phi = \varphi \times \varphi'$ be a rectangular partition of $[a, b] \times [c, d]$. Throughout, let $0 < h \leq \min\{p, p'\}$ be fixed and denote the discrete interval

$$[\alpha, \beta]_h = \{\alpha, \alpha + h, \alpha + 2h, \cdots\} \cap [\alpha, \beta].$$

We assume that p and p' are multiples of h. Then, it is clear that t_i's are in $[a, b]_h$ and u_i's are in $[c, d]_h$.

Definition 4 A function $S(t; \varphi, h)$ is called a *discrete quintic spline* if its restriction S_i on $[t_{i-1}, t_i]$ is a quintic polynomial for $i = 1, 2, \cdots, n$ and

$$D_h^{\{\mu\}}S_i(t_i) = D_h^{\{\mu\}}S_{i+1}(t_i), \quad 1 \leq i \leq n-1, \ \mu = 0, 1, 2, 3, 4. \tag{11}$$

For a positive number P_0, we say a function g is P_0-*periodic* if

$$g(t) = g(t + P_0).$$

We shall now introduce *periodic* discrete quintic spline. In the spirit of [28, 29] where periodic discrete *cubic* spline is studied, let

$$S_h(\varphi) = \Big\{ S(t;\varphi,h) : S(t;\varphi,h) \text{ is a discrete quintic spline and it is } (b-a)\text{-periodic} \Big\}.$$

Definition 5 For a $(b-a)$-periodic function f defined on $[a-2h, b+2h]_h$, we say $S_\varphi f$ is the $S_h(\varphi)$-*interpolate of* f, also known as the *periodic discrete spline interpolate of* f, if $S_\varphi f \in S_h(\varphi)$ with

$$S_\varphi f(t_i) = f(t_i) \equiv f_i, \quad 0 \le i \le n-1. \tag{12}$$

Remark 2 In Definition 5, it actually suffices to have the periodic function f defined on the uniform partition φ. However, for the error analysis in the next section, we require the periodic function f to be defined on $[a-2h, b+2h]_h$. To be consistent, we therefore impose throughout that the $(b-a)$-*periodic function* f *is defined on* $[a-2h, b+2h]_h$.

We shall give an explicit expression of $S_\varphi f$. For this, Let the functions g_i, $\bar{g}_i$ and $\bar{\bar{g}}_i$ satisfy the following for $0 \le i, j \le n-1$:

$$\begin{aligned} &g_i(t_j) = \delta_{ij}, \qquad D_h^{\{2\}} g_i(t_j) = D_h^{\{4\}} g_i(t_j) = 0, \\ &D_h^{\{2\}} \bar{g}_i(t_j) = \delta_{ij}, \ \bar{g}_i(t_j) = D_h^{\{4\}} \bar{g}_i(t_j) = 0, \\ &D_h^{\{4\}} \bar{\bar{g}}_i(t_j) = \delta_{ij}, \ \bar{\bar{g}}_i(t_j) = D_h^{\{2\}} \bar{\bar{g}}_i(t_j) = 0. \end{aligned}$$

The explicit expressions of g_i, $\bar{g}_i$ and $\bar{\bar{g}}_i$ can be obtained by direct computation.

Lemma 3 *Let* $M_i = D_h^{\{2\}} S_\varphi f(t_i)$ *and* $F_i = D_h^{\{4\}} S_\varphi f(t_i)$, $0 \le i \le n$. *Then,* $S_\varphi f$ *can be written as*

$$S_\varphi f(t) = \sum_{i=0}^{n-1} \big[f_i g_i(t) + M_i \bar{g}_i(t) + F_i \bar{\bar{g}}_i(t) \big], \quad t \in [a,b]. \tag{13}$$

In particular, for $t \in [t_{i-1}, t_i]$, $1 \le i \le n$, *the spline interpolate* $S_\varphi f$ *has the expression*

$$\begin{aligned} S_\varphi f(t) = (S_\varphi f)_i(t) = {} & f_{i-1} g_{i-1}(t) + f_i g_i(t) + M_{i-1} \bar{g}_{i-1}(t) + M_i \bar{g}_i(t) \\ & + F_{i-1} \bar{\bar{g}}_{i-1}(t) + F_i \bar{\bar{g}}_i(t), \quad t \in [t_{i-1}, t_i], \ 1 \le i \le n. \end{aligned} \tag{14}$$

Theorem 4 *Let* f *be a given* $(b-a)$-*periodic function defined on* $[a-2h, b+2h]_h$. *Then, there exists a unique periodic discrete spline interpolate* $S_\varphi f$. *Here,* M_i *and* F_i, $0 \le i \le n-1$ *in (13) are uniquely determined by the systems of equations*

$$\begin{aligned} &a_1 M_{i-2} + a_2 M_{i-1} + a_3 M_i + a_2 M_{i+1} + a_1 M_{i+2} \\ &= \frac{1}{6} [(p^2 - h^2) f_{i-2} + 2(2h^2 + p^2) f_{i-1} - 6(h^2 + p^2) f_i + 2(2h^2 + p^2) f_{i+1} \\ &\quad + (p^2 - h^2) f_{i+2}] \end{aligned} \tag{15}$$

and

$$a_1F_{i-2} + a_2F_{i-1} + a_3F_i + a_2F_{i+1} + a_1F_{i+2} = f_{i-2} - 4f_{i-1} + 6f_i - 4f_{i+1} + f_{i+2} \tag{16}$$

where $a_1 = \frac{1}{120}(p^2 - h^2)(p^2 - 4h^2)$, $a_2 = \frac{1}{60}(p^2 - h^2)(8h^2 + 13p^2)$ *and* $a_3 = \frac{1}{20}(4h^4 + 5h^2p^2 + 11p^4)$. *Note that* $a_3 > 2(|a_1| + a_2)$.

Remark 3 It is possible to describe a basis for $S_h(\varphi)$, namely the 'cardinal splines', $\{s_i\}_{i=0}^{n-1}$, defined by the following interpolating conditions

$$s_i(t_j) = \delta_{ij}^* = \begin{cases} 1, \text{ if } j = i + nk \ (k \in \mathbf{Z}) \\ 0, \text{ otherwise.} \end{cases}$$

Obviously $S_\varphi f$ can be expressed as

$$S_\varphi f(t) = \sum_{i=0}^{n-1} f_i s_i(t). \tag{17}$$

We shall now introduce the two-variable periodic discrete quintic spline interpolation. For any positive numbers P_1 and P_2, we say a two-variable function g is (P_1, P_2)-*periodic* if

$$g(t + P_1, u) = g(t, u), \quad g(t, u + P_2) = g(t, u) \quad \text{and} \quad g(t + P_1, u + P_2) = g(t, u).$$

For convenience, we shall denote $g^{\{\mu,\nu\}}(t, u) = D_{h,t}^{\{\mu\}} D_{h,u}^{\{\nu\}} g(t, u)$, and with respect to the partition $\phi = \varphi \times \varphi'$, denote $g_{i,j}^{\{\mu,\nu\}} = D_{h,t}^{\{\mu\}} D_{h,u}^{\{\nu\}} g(t_i, u_j)$.

Define $S_h(\phi) = S_h(\varphi) \oplus S_h(\varphi')$ (the tensor product) $= \text{Span } \{s_i s_j\}_{i=0,\ j=0}^{n-1,\ m-1} = \{S : S$ is a two-dimensional polynomial of degree 5 in each variable, its restriction S_{ij} on $[t_{i-1}, t_i] \times [u_{j-1}, u_j]$, $1 \le i \le n$, $1 \le j \le m$ is biquintic, $S_{ij}^{\{\mu,\nu\}}(t_i, u_j) = S_{i+1,j}^{\{\mu,\nu\}}(t_i, u_j) = S_{i,j+1}^{\{\mu,\nu\}}(t_i, u_j) = S_{i+1,j+1}^{\{\mu,\nu\}}(t_i, u_j)$, $1 \le i \le n-1$, $1 \le j \le m-1$, $\mu, \nu = 0, 1, 2, 3, 4$, and S is $(b-a, d-c)$-periodic$\}$.

Definition 6 For a $(b-a, d-c)$-periodic function f defined on $[a-2h, b+2h]_h \times [c-2h, d+2h]_h$, we say $S_\phi f$ is a $S_h(\phi)$-*interpolate of* f, also known as the *periodic discrete spline interpolate of* f, if $S_\phi f \in S_h(\phi)$ with

$$S_\phi f(t_i, u_j) = f(t_i, u_j) \equiv f_{ij}, \quad 0 \le i \le n-1, \ 0 \le j \le m-1. \tag{18}$$

Remark 4 In Definition 6, it actually suffices to have the periodic function f defined on the partition ϕ. However, the subsequent error analysis requires the periodic function f to be defined on $[a-2h, b+2h]_h \times [c-2h, d+2h]_h$. To be consistent, we therefore impose throughout that the $(b-a, d-c)$-*periodic function* f *is defined on* $[a-2h, b+2h]_h \times [c-2h, d+2h]_h$.

We shall give an explicit expression of $S_\phi f$ in the next result.

Lemma 4 *Let* $c_{i,j}^{\{\mu,\nu\}} = (S_\phi f)^{\{\mu,\nu\}}(t_i, u_j),\ 0 \le i \le n,\ 0 \le j \le m,\ \mu, \nu \in \{0, 2, 4\}$ *(note that* $c_{i,j}^{\{0,0\}} = f_{ij}$*). Then,* $S_\phi f$ *can be written as*

$$\begin{aligned} S_\phi f(t,u) = \sum_{i=0}^{n-1}\sum_{j=0}^{m-1}\Big[& f_{ij}g_i(t)g_j(u) + c_{i,j}^{\{0,2\}}g_i(t)\bar{g}_j(u) + c_{i,j}^{\{0,4\}}g_i(t)\bar{\bar{g}}_j(u) \\ & + c_{i,j}^{\{2,0\}}\bar{g}_i(t)g_j(u) + c_{i,j}^{\{2,2\}}\bar{g}_i(t)\bar{g}_j(u) + c_{i,j}^{\{2,4\}}\bar{g}_i(t)\bar{\bar{g}}_j(u) \\ & + c_{i,j}^{\{4,0\}}\bar{\bar{g}}_i(t)g_j(u) + c_{i,j}^{\{4,2\}}\bar{\bar{g}}_i(t)\bar{g}_j(u) + c_{i,j}^{\{4,4\}}\bar{\bar{g}}_i(t)\bar{\bar{g}}_j(u)\Big]. \end{aligned} \tag{19}$$

Theorem 5 *Let* f *be a given* $(b-a, d-c)$*-periodic function defined on* $[a-2h, b+2h]_h \times [c-2h, d+2h]_h$*. Then, there exists a unique periodic discrete spline interpolate* $S_\phi f$*. Here,* $c_{i,j}^{\{\mu,\nu\}}$*'s in (19) are uniquely determined by systems analogous to (15) and (16).*

Remark 5 In view of Remark 3, $S_\phi f$ can be expressed in terms of cardinal splines as

$$S_\phi f(t,u) = \sum_{i=0}^{n-1}\sum_{j=0}^{m-1} f_{ij}s_i(t)s_j(u). \tag{20}$$

We shall next establish the error estimates for the periodic discrete spline interpolation in one variable and two variables. For a function $g(t)$ defined on $[\bar{a}, \bar{b}]_h$, we introduce the *modulus of smoothness* and the *norm* as

$$w(g,r) = \max\left\{|g(t) - g(t')| : |t - t'| < r,\ t, t' \in [\bar{a}, \bar{b}]_h\right\}, \qquad \|g\| = \max_{t\in[\bar{a},\bar{b}]_h} |g(t)|.$$

For a function $g(t,u)$ defined on $[\bar{a}, \bar{b}]_h \times [\bar{c}, \bar{d}]_h$, the norm

$$\|g\| = \max_{(t,u)\in[\bar{a},\bar{b}]_h\times[\bar{c},\bar{d}]_h} |g(t,u)|.$$

To prove the error estimate result in the one-variable case, we require the following lemma from [40].

Lemma 5 *[40]*

(a) Let $\alpha,\ \beta$ *be given real numbers such that* $\alpha < \beta$ *and* $\beta \in \{\alpha, \alpha + h, \alpha + 2h, \cdots\}$ *for some* $h > 0$*. Let* $g : [\alpha - h, \beta + h]_h \to \mathbb{R}$ *be a given function, define the operators* L *and* U *by*

$$(Lg)(t) = \frac{t-\alpha}{\beta-\alpha}g(\beta) + \frac{\beta-t}{\beta-\alpha}g(\alpha), \qquad (Ug)(t) = g(t) - (Lg)(t).$$

Then, we have

$$\|Ug\| \le w(g, \beta-\alpha), \quad \|Ug\| \le \frac{(\beta-\alpha)^2}{8}\|g^{\{2\}}\|, \quad \|D_h^{\{1\}}Ug\| \le \frac{\beta-\alpha}{2}\|g^{\{2\}}\|.$$

(b) Let $\{a_i\}_{i=1}^I$ and $\{b_j\}_{j=1}^J$ be given sequences of nonnegative real numbers such that $\sum_{i=1}^I a_i = \sum_{j=1}^J b_j$. Then, for any real valued function g defined on a discrete interval $[\alpha, \beta]_h$, we have

$$\left|\sum_{i=1}^{I} a_i g[t_{i0}, t_{i1}, \cdots, t_{ik}] - \sum_{j=1}^{J} b_j g[u_{j0}, u_{j1}, \cdots, u_{jk}]\right| \leq \frac{1}{k!}\left(\sum_{i=1}^{I} a_i\right) w\left(g^{\{k\}}, |\beta - \alpha - kh|\right)$$

where $t_{i\ell}$'s and $u_{j\ell}$'s are in $[\alpha, \beta]_h$.

In the two-variable case, the key inequality used in deriving the error estimates is

$$\|f - S_\phi f\| \leq \|f - S_\varphi f\| + \|S_\varphi(f - S_{\varphi'} f) - (f - S_{\varphi'} f)\| + \|f - S_{\varphi'} f\|.$$

The next theorem gives the error estimates for the periodic discrete spline interpolation in one variable and two variables.

Theorem 6 *Define the constant $\gamma = \frac{40p^4}{(4h^2+p^2)(h^2+p^2)}$.*

(a) Let f be a $(b-a)$-periodic function defined on $[a-2h, b+2h]_h$ and $e = S_\varphi f - f$. Then, we have

$$\begin{aligned}
\|e^{\{4\}}\| &\leq (1+\gamma) w\left(f^{\{4\}}, p\right), \\
\|e^{\{3\}}\| &\leq \frac{p}{2}\|e^{\{4\}}\| + \frac{2}{p}\gamma w\left(f^{\{2\}}, p\right) \leq \frac{p}{2}(1+\gamma) w\left(f^{\{4\}}, p\right) + \frac{2}{p}\gamma w\left(f^{\{2\}}, p\right), \\
\|e^{\{2\}}\| &\leq \frac{p^2}{8}\|e^{\{4\}}\| + \gamma w\left(f^{\{2\}}, p\right) \leq \frac{p^2}{8}(1+\gamma) w\left(f^{\{4\}}, p\right) + \gamma w\left(f^{\{2\}}, p\right), \\
\|e^{\{1\}}\| &\leq \frac{p}{2}\|e^{\{2\}}\| \leq \frac{p^3}{16}(1+\gamma) w\left(f^{\{4\}}, p\right) + \frac{p}{2}\gamma w\left(f^{\{2\}}, p\right), \\
\|e\| &\leq \frac{p^2}{8}\|e^{\{2\}}\| \leq \frac{p^4}{64}(1+\gamma) w\left(f^{\{4\}}, p\right) + \frac{p^2}{8}\gamma w\left(f^{\{2\}}, p\right).
\end{aligned}$$

(b) Let f be a $(b-a, d-c)$-periodic function defined on $[a-2h, b+2h]_h \times [c-2h, d+2h]_h$. Then, we have

$$\begin{aligned}
\|f - S_\phi f\| \leq\ & \frac{p^4}{64}(1+\gamma) w_t\left(f^{\{4,0\}}, p\right) + \frac{p^2}{8}\gamma w_t\left(f^{\{2,0\}}, p\right) \\
& + \frac{(p')^4}{64}(1+\gamma') w_u\left(f^{\{0,4\}}, p'\right) + \frac{(p')^2}{8}\gamma' w_u\left(f^{\{0,2\}}, p'\right) \\
& + \frac{p^4}{32}(1+\gamma)\left[\frac{(p')^4}{64}(1+\gamma') w_u\left(f^{\{4,4\}}, p'\right) + \frac{(p')^2}{8}\gamma' w_u\left(f^{\{4,2\}}, p'\right)\right] \\
& + \frac{p^2}{4}\gamma\left[\frac{(p')^4}{64}(1+\gamma') w_u\left(f^{\{2,4\}}, p'\right) + \frac{(p')^2}{8}\gamma' w_u\left(f^{\{2,2\}}, p'\right)\right]
\end{aligned}$$

where γ' is the same as γ with p replaced by p'.

Table 3 (Example 3) Actual errors and error bounds

	$p = \frac{1}{10},\ h = p$	$p = \frac{1}{8},\ h = \frac{p}{4}$	$p = \frac{1}{10},\ h = \frac{p}{4}$	$p = \frac{1}{15},\ h = \frac{p}{4}$
$\|e\|$	0	$0.82039194e-08$	$0.20521179e-08$	$0.17547366e-09$
Bound	$0.96628467e-05$	$0.47809205e-03$	$0.24082627e-03$	$0.70158943e-04$
$\|e^{\{1\}}\|$	0	$0.13126271e-06$	$0.41042359e-07$	$0.52642099e-08$
Bound	$0.38651387e-03$	$0.15298946e-01$	$0.96330507e-02$	$0.42095366e-02$
$\|e^{\{2\}}\|$	0	$0.88913504e-05$	$0.35473364e-05$	$0.68230315e-06$
Bound	$0.77302774e-02$	$0.63964914e+00$	$0.19266101e+00$	$0.12628610e+00$
$\|e^{\{3\}}\|$	0	$0.25175638e-03$	$0.12575208e-3$	$0.36340402e-04$
Bound	$0.16331306e+00$	$0.42046640e+01$	$0.40397862e+01$	$0.38724574e+00$
$\|e^{\{4\}}\|$	0	$0.19214105e-01$	$0.12196109e-01$	$0.53762787e-02$
Bound	$0.34830056e+00$	$0.92202866e+01$	$0.74626360e+01$	$0.50324713e+01$

We shall now present some examples to illustrate the periodic discrete spline interpolation as well as the corresponding error bounds obtained in Theorem 6.

Example 3 Consider the function

$$f(t) = \frac{1}{100}\left[\sin^2(\pi t) + \frac{19}{20}\cos^2(\pi t)\right],\quad t \in [0, 1].$$

Here, we have $[a, b] = [0, 1]$ and the steps taken to obtain the periodic discrete spline interpolate $S_\varphi f$ and the errors are listed as follows.

(i) Fix the uniform partition φ (i.e., step size p) and choose a value for h.
(ii) Solve the systems (15) and (16) to get M_i's and F_i's respectively. Then, $S_\varphi f$ can be constructed in each subinterval $[t_{i-1}, t_i]$ following (14).
(iii) Compute the actual errors

$$\|e^{\{\mu\}}\| = \|f^{\{\mu\}} - (S_\varphi f)^{\{\mu\}}\| = \max_{t\in[0,1]_h} |f^{\{\mu\}}(t) - (S_\varphi f)^{\{\mu\}}(t)|,\ \mu = 0, 1, 2, 3, 4.$$

(iv) Compute the bounds given in Theorem 6.

The actual errors and the error bounds are presented in Table 3. To illustrate graphically, we have plotted $S_\varphi f$ and f in Fig. 3.

Example 4 Consider the function

$$f(t, u) = \frac{1}{100}\left[\sin^2(\pi t) + \frac{19}{20}\cos^2(\pi u)\right],\quad (t, u) \in [0, 1] \times [0, 1].$$

Here, we have $[a, b] = [c, d] = [0, 1]$. Fix $p = p'$ and take $h = \frac{p}{4}$.

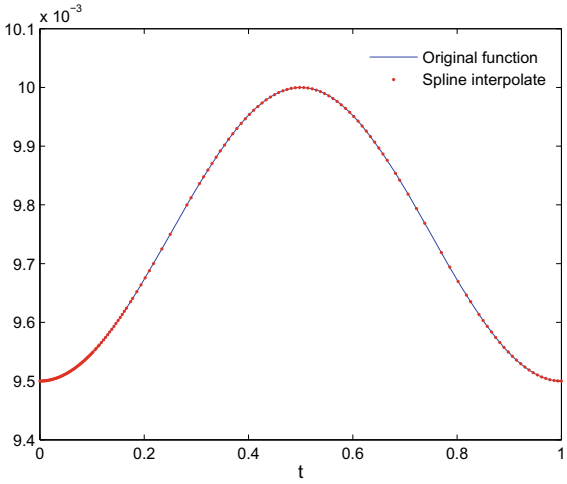

Fig. 3 (Example 3) f and $S_\varphi f$ when $h = p = \frac{1}{10}$

Table 4 (Example 4) Actual errors and error bounds

	$p = p' = \frac{1}{8}$	$p = p' = \frac{1}{10}$
$\|f - S_\phi f\|$	$0.56790328e-03$	$0.28049052e-03$
Bound	$0.18645590e-01$	$0.93922244e-02$

To construct $S_\phi f$, in view of Remark 5 we need only to construct the cardinal splines s_i, $t \in [0, 1]_h$, $0 \le i \le n-1$ and s_j, $u \in [0, 1]_h$, $0 \le j \le m-1$. To obtain a particular cardinal spline, we solve the systems (15) and (16) and then the cardinal spline can be written explicitly using (13) or (14).

We also compute the actual error

$$\|f - S_\phi f\| = \max_{(t,u)\in[0,1]_h\times[0,1]_h} |f(t,u) - S_\phi f(t,u)|$$

and the bound in Theorem 6. The results are presented in Table 4. To illustrate graphically, we have plotted $S_\phi f$ and f in Figs. 4 and 5.

3 Discrete Cubic Spline Method for Second Order Boundary Value Problem

In this section, we use discrete cubic spline to obtain approximate solution of a second order boundary value problem. We shall show that the method is of order 4 if a parameter takes a specific value, and it is of order 2 otherwise. Two numerical

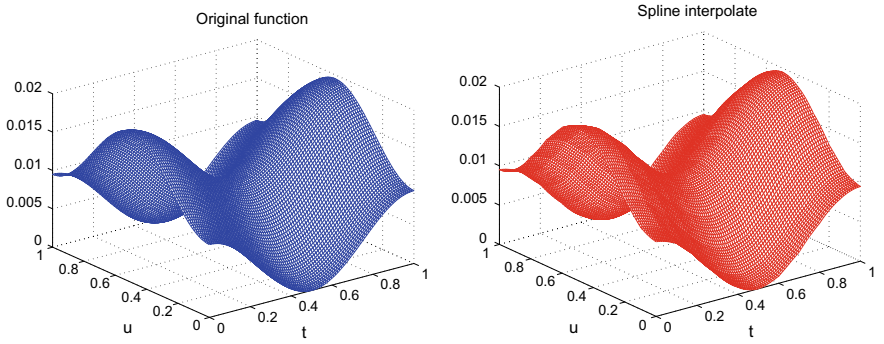

Fig. 4 (Example 4) f and $S_\phi f$ when $p = p' = 4h = \frac{1}{10}$

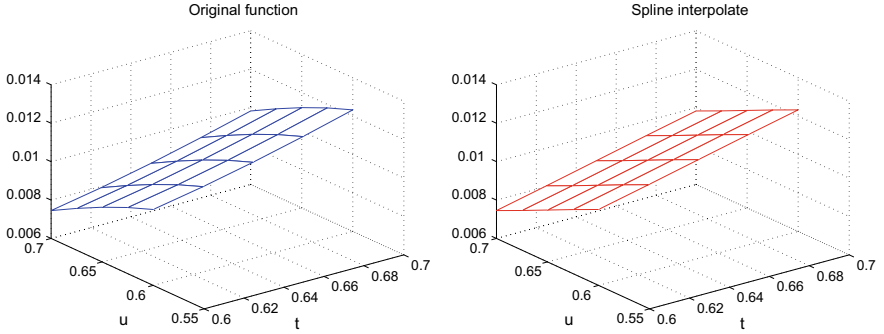

Fig. 5 (Example 4) Enlarged portion of Fig. 4 where the error $|f(t,u) - S_\phi f(t,u)|$ is large

examples are presented to illustrate the efficiency of our method as well as to compare the performance with other numerical methods proposed in the literature. This section refers to the work of [21].

We consider the second order boundary value problem

$$\begin{aligned} y''(x) &= f(x)y(x) + g(x), \quad a \le x \le b \\ y(a) &= \bar{\alpha}, \quad y(b) = \bar{\beta} \end{aligned} \tag{21}$$

where f and g are continuous functions on $[a, b]$. Such problems arise from many real world situations, for example in the theory which describes the deflection of plates and a variety of other scientific applications. In general it is difficult to obtain the analytical solution of (21) for arbitrary f and g and we usually resort to some numerical methods. In the literature, finite difference method has been commonly used for the numerical treatment of (21) and this method has been discussed by many authors, see, for example [12, 24, 34, 63]. On the other hand, Ahlberg et al. [3] have introduced (continuous) splines in solving initial as well as boundary value problems. Following this several authors [4, 5, 16, 32, 36] have investigated the use of cubic

splines in solving two-point boundary value problems. Other methods that involve quadratic splines as well as collocation methods with splines as basis functions have further been applied to solve various second order boundary value problems, see for example [3, 6, 11, 35, 48, 56, 65] and the references therein.

We note that in the literature those methods that solve (21) by *cubic splines* [4, 5, 16, 32, 36] are of order 2 in most cases, except that of Khan [36], which is of order 4 when certain parameters take specific values and is of order 2 otherwise. In comparison, our discrete cubic spline method is fourth order convergent if a parameter takes a specific value, else it is second order convergent—this convergence is 'on par' with the method of Khan [36] and better than those in [4, 5, 16, 32]. Moreover, computationally our method is much easier compared to [36]. Indeed, we shall show by two numerical examples that our method outperforms other collocation, finite-difference and spline methods for solving (21).

3.1 Discrete Cubic Spline Method

Let $P : a = x_0 < x_1 < \cdots < x_n = b$ be a uniform mesh of $[a, b]$ with $x_i - x_{i-1} = p,\ 1 \le i \le n$. For any function $F(x)$, we denote its k-th derivative at x_i as $F_i^{(k)}$.

Let $h \in (0, p]$ be a given constant used in the central difference operator D_h.

Definition 7 Let $S(x; h)$ be a piecewise continuous function defined over $[a, b]$ (with mesh P) and $S_i(x)$ be its restriction on $[x_{i-1}, x_i],\ 1 \le i \le n$ passing through the points (x_{i-1}, s_{i-1}) and (x_i, s_i). We say $S(x; h)$ is a *discrete cubic spline* if $S_i(x),\ 1 \le i \le n$ is a polynomial of degree 3 or less and

$$(S_{i+1} - S_i)(x_i + jh) = 0, \quad j = -1, 0, 1, \quad 1 \le i \le n - 1 \tag{22}$$

or equivalently

$$D_h^{\{j\}} S_i(x_i) = D_h^{\{j\}} S_{i+1}(x_i), \quad j = 0, 1, 2, \quad 1 \le i \le n - 1. \tag{23}$$

The above definition of discrete spline is based on *central differences*. Indeed, Lyche [42] has the same definition for discrete spline.

We shall approximate a solution $y(x)$ of (21) by the discrete cubic spline $S(x; h)$. Hence, for any $x \in [a, b]$ (x may be between mesh points), we propose the following approximation

$$y(x) \cong S(x; h), \qquad y'(x) \cong D_h^{\{1\}} S(x; h), \qquad y''(x) \cong f(x)S(x; h) + g(x). \tag{24}$$

In particular, at the mesh points we have

$$y_i \cong s_i \equiv S_i(x_i), \quad y_i' \cong s_i' \equiv D_h^{\{1\}} S_i(x_i), \quad y_i'' \cong s_i'' \equiv f_i S_i(x_i) + g_i, \quad 0 \le i \le n. \tag{25}$$

From the boundary conditions, we note that $s_0 = y_0 = \bar{\alpha}$ and $s_n = y_n = \bar{\beta}$, and s_i is an approximate of y_i, $1 \le i \le n-1$.

Our immediate task is to obtain an explicit expression of $S_i(x)$. Clearly, $S_i(x)$ should pass through the points (x_{i-1}, s_{i-1}) and (x_i, s_i). Let $c_i = D_h^{\{2\}} S(x_i; h)$, $0 \le i \le n$ denote the 'discrete moments'. Since $D_h^{\{2\}} S(x; h)$ is piecewise linear, we have for $x \in [x_{i-1}, x_i]$, $1 \le i \le n$,

$$D_h^{\{2\}} S(x; h) = D_h^{\{2\}} S_i(x) = \frac{x_i - x}{p} c_{i-1} + \frac{x - x_{i-1}}{p} c_i. \tag{26}$$

It follows that for $x \in [x_{i-1}, x_i]$, $1 \le i \le n$,

$$S_i(x) = \frac{(x_i - x)^{\{3\}}}{6p} c_{i-1} + \frac{(x - x_{i-1})^{\{3\}}}{6p} c_i + \frac{x_i - x}{p} u_i + \frac{x - x_{i-1}}{p} v_i \tag{27}$$

where u_i and v_i are arbitrary constants that can be determined by the interpolation conditions $S_i(x_{i-1}) = s_{i-1}$ and $S_i(x_i) = s_i$. It is found that

$$u_i = s_{i-1} - \frac{p^2 - h^2}{6} c_{i-1}, \quad v_i = s_i - \frac{p^2 - h^2}{6} c_i, \quad 1 \le i \le n. \tag{28}$$

Hence, upon substituting (28) into (27), we obtain an explicit expression of $S_i(x)$ in terms of s_{i-1}, s_i, c_{i-1} and c_i.

Taking central difference of (27) then gives for $x \in [x_{i-1}, x_i]$, $1 \le i \le n$,

$$D_h^{\{1\}} S_i(x) = -\frac{(x_i - x)^2}{2p} c_{i-1} - \frac{(p^2 - h^2)(c_i - c_{i-1})}{6p} + \frac{(x - x_{i-1})^2}{2p} c_i + \frac{s_i - s_{i-1}}{p}. \tag{29}$$

The 'continuity' requirement $D_h^{\{1\}} S_i(x_i) = D_h^{\{1\}} S_{i+1}(x_i)$ (see (23)) then leads to the system of equations

$$\frac{(p^2 - h^2)}{6} c_{i-1} + \frac{2(2p^2 + h^2)}{6} c_i + \frac{(p^2 - h^2)}{6} c_{i+1} = s_{i-1} - 2s_i + s_{i+1}, \quad 1 \le i \le n-1. \tag{30}$$

In view of the fact that we approximate $y(x)$ by $S(x; h)$ and (25), we set $c_i = s_i''$ or

$$c_i = f_i s_i + g_i, \quad 0 \le i \le n. \tag{31}$$

Upon substituting (31) into (30), the system (30) becomes

$$\begin{aligned} &\left[\frac{(p^2 - h^2)}{6} f_{i-1} - 1\right] s_{i-1} + \left[\frac{2(2p^2 + h^2)}{6} f_i + 2\right] s_i + \left[\frac{(p^2 - h^2)}{6} f_{i+1} - 1\right] s_{i+1} \\ &\quad = -\frac{(p^2 - h^2)}{6} (g_{i-1} + g_{i+1}) - \frac{2(2p^2 + h^2)}{6} g_i, \quad 1 \le i \le n-1. \end{aligned} \tag{32}$$

Together with the boundary conditions $s_0 = \bar{\alpha},\ s_n = \bar{\beta}$, we may solve (32) to get $s_i,\ 1 \le i \le n-1$. The unique solvability of the system (32) will be shown in the next section.

Finally, we list the steps of computing the discrete cubic spline solution of (21) as follows:

(i) Fix the mesh P (and hence the mesh size p) and choose a value for h ($\in (0, p]$).
(ii) Solve (32) to get $s_i,\ 1 \le i \le n-1$, which approximates y_i.
(iii) Calculate c_i by using $c_i = f_i s_i + g_i,\ 0 \le i \le n$ (see (31)). Noting (25), $s_i'' = c_i$ serves as an approximate to y_i''.
(iv) Compute $D_h^{\{1\}} S_i(x_i)$ from (29). Noting (25), $s_i' = D_h^{\{1\}} S_i(x_i)$ serves as an approximate to y_i'.
(v) The discrete cubic spline solution $S_i(x)$ over the subinterval $[x_{i-1}, x_i]$ can be obtained using (27). The first central difference $D_h^{\{1\}} S_i(x)$ can also be obtained using (29). These can be used to approximate $y(x)$ and $y'(x)$ for any $x \in [a, b]$.

3.2 Convergence Analysis

In this section, we shall establish the existence of a unique discrete cubic spline solution for (21) (i.e., (32) has a unique solution) and also conduct a convergence analysis for the method presented in the previous section. As usual, the norms of a column vector $t = [t_i]$ and a matrix $A = [a_{ij}]$ are given by

$$\|t\| = \max_i |t_i| \qquad \text{and} \qquad \|A\| = \max_i \sum_j |a_{ij}|.$$

Let $e_i = y_i - s_i,\ 1 \le i \le n-1$ be the errors. Let $y = [y_i],\ s = [s_i],\ r = [r_i],\ t = [t_i]$ and $e = [e_i]$ be $(n-1)$-dimensional column vectors. The system (32) can be written as

$$As = r \tag{33}$$

where

$$A = A_0 + Q, \qquad Q = BF, \qquad F = \text{diag}(f_i),\ \ i = 1, 2, \cdots, n-1, \tag{34}$$

$B = [b_{ij}]$ and $A_0 = [a_{ij}]$ are $(n-1) \times (n-1)$ tridiagonal matrices given by

$$b_{ij} = \begin{cases} \dfrac{2(2p^2 + h^2)}{6}, & i = j \\ \dfrac{(p^2 - h^2)}{6}, & |i - j| = 1 \\ 0, & \text{otherwise}, \end{cases} \qquad a_{ij} = \begin{cases} 2, & i = j \\ -1, & |i - j| = 1 \\ 0, & \text{otherwise} \end{cases} \tag{35}$$

and

$$r_i = \begin{cases} \bar{\alpha} - \frac{1}{6}[(p^2-h^2)f_0\bar{\alpha} + (p^2-h^2)g_0 + 2(2p^2+h^2)g_1 + (p^2-h^2)g_2], & i = 1 \\ -\frac{1}{6}[(p^2-h^2)g_{i-1} + 2(2p^2+h^2)g_i + (p^2-h^2)g_{i+1}], & 2 \le i \le n-2 \\ \bar{\beta} - \frac{1}{6}[(p^2-h^2)f_n\bar{\beta} + (p^2-h^2)g_{n-2} + 2(2p^2+h^2)g_{n-1} + (p^2-h^2)g_n], & \\ & i = n-1. \end{cases} \tag{36}$$

From (33), we have $A(y-e) = r$ or

$$Ay = r + t \tag{37}$$

where

$$t = Ae. \tag{38}$$

The i-th equation of the system (37) is

$$-y_{i-1} + 2y_i - y_{i+1} = -\frac{1}{6}[(p^2-h^2)y''_{i-1} + 2(2p^2+h^2)y''_i + (p^2-h^2)y''_{i+1}] + t_i$$

where t_i, $1 \le i \le n-1$ are the local truncation errors given by

$$t_i = \frac{p^2(p^2-2h^2)}{12}y_i^{(4)} + \frac{p^4(4p^2-5h^2)}{360}y_i^{(6)} + O(p^8). \tag{39}$$

Remark 6 When $h = \frac{p}{\sqrt{2}}$, it is clear from (39) that $t_i = \frac{1}{240}p^6 y_i^{(6)} + O(p^8)$. Thus,

$$\|t\| = \frac{1}{240}p^6 M_6 \tag{40}$$

where $M_6 = \max_x |y^{(6)}(x)|$.

Lemma 6 *[63] The inverse of A_0, namely $A_0^{-1} = [\eta_{ij}]$, is given by*

$$\eta_{ij} = \begin{cases} \frac{j(n-i)}{n}, & i \ge j \\ \frac{i(n-j)}{n}, & i \le j. \end{cases}$$

Note that $A_0^{-1} > 0$, i.e, all the entries of A_0^{-1} are positive. Moreover,

$$\|A_0^{-1}\| \le \frac{n^2}{8}. \tag{41}$$

Lemma 7 *[31] Let W be a square matrix such that $\|W\| < 1$. Then, $(I + W)$ is nonsingular and*

$$\|(I + W)^{-1}\| \leq \frac{1}{1 - \|W\|}.$$

Our first result guarantees the existence of a unique discrete cubic spline solution for (21).

Theorem 7

(a) The system (33) has a unique solution if

$$K\hat{f} < 1 \tag{42}$$

where $K = \frac{1}{8}(b-a)^2$ and $\hat{f} = \max_{1\leq i\leq n-1} |f_i|$.

(b) The system (33) has a unique solution if $f(x) \equiv f_0 > 0$.

Proof (a) If (33) has a unique solution, then it can be written as

$$s = A^{-1}r = (A_0 + Q)^{-1}r = [A_0(I + A_0^{-1}Q)]^{-1}r = (I + A_0^{-1}Q)^{-1}A_0^{-1}r. \tag{43}$$

From Lemma 6 the inverse A_0^{-1} exists, if we can show that $(I + A_0^{-1}Q)^{-1}$ also exists, then it is immediate that (33) has a unique solution given by (43).

It is clear that $\|B\| = p^2$. Since $Q = BF$, we find

$$\|Q\| \leq \|B\| \, \|F\| \leq p^2\hat{f}. \tag{44}$$

Then, applying (41), the fact $n = \frac{b-a}{p}$, (44) and (42) successively, we get

$$\|A_0^{-1}Q\| \leq \|A_0^{-1}\| \, \|Q\| \leq \frac{(b-a)^2}{8p^2}\left(p^2\hat{f}\right) = K\hat{f} < 1. \tag{45}$$

Since $\|A_0^{-1}Q\| < 1$, it follows from Lemma 7 that $(I + A_0^{-1}Q)$ is nonsingular. Hence, $(I + A_0^{-1}Q)^{-1}$ exists and (33) has a unique solution given by (43).

(b) If $f(x) \equiv f_0 > 0$, we can show that the coefficient matrix A in (33) is strictly diagonally dominant, then A^{-1} exists and the conclusion is immediate.

In fact, from (32) we see that the $(n-1) \times (n-1)$ matrix A is tridiagonal and is given by

$$A = [a_{ij}] = \begin{bmatrix} \frac{2(2p^2+h^2)}{6}f_0 + 2 & \frac{(p^2-h^2)}{6}f_0 - 1 & & & \\ \frac{(p^2-h^2)}{6}f_0 - 1 & \frac{2(2p^2+h^2)}{6}f_0 + 2 & \frac{(p^2-h^2)}{6}f_0 - 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{(p^2-h^2)}{6}f_0 - 1 & \frac{2(2p^2+h^2)}{6}f_0 + 2 & \frac{(p^2-h^2)}{6}f_0 - 1 \\ & & & \frac{(p^2-h^2)}{6}f_0 - 1 & \frac{2(2p^2+h^2)}{6}f_0 + 2 \end{bmatrix}.$$

It can be easily checked that for $2 \leq i \leq n-2$,

$$|a_{ii}| - \sum_{j \neq i} |a_{ij}| = \left\{ \begin{array}{ll} \frac{2(p^2+2h^2)}{6} f_0 + 4, & \frac{(p^2-h^2)}{6} f_0 - 1 \geq 0 \\ p^2 f_0, & \frac{(p^2-h^2)}{6} f_0 - 1 \leq 0 \end{array} \right\} > 0; \tag{46}$$

while for $i = 1,\ n-1$,

$$|a_{ii}| - \sum_{j \neq i} |a_{ij}| = \left\{ \begin{array}{ll} \frac{(p^2+h^2)}{2} f_0 + 3, & \frac{(p^2-h^2)}{6} f_0 - 1 \geq 0 \\ \frac{(5p^2+h^2)}{6} f_0 + 1, & \frac{(p^2-h^2)}{6} f_0 - 1 \leq 0 \end{array} \right\} > 0. \tag{47}$$

Hence, the matrix A is indeed strictly diagonally dominant, this completes the proof. □

The next result gives the convergence order of the discrete cubic spline method.

Theorem 8 *Suppose $K\hat{f} < 1$ or $f(x) \equiv f_0 > 0$. Then,*

$$\|e\| \cong O(p^4) \quad \text{if} \quad h = \frac{p}{\sqrt{2}}$$

and $\|e\| \cong O(p^2)$ for other values of $h \in (0, p]$. In conclusion, the discrete cubic spline method (33) is fourth order convergent if $h = \frac{p}{\sqrt{2}}$ and is second order convergent otherwise.

Proof First, suppose $K\hat{f} < 1$. We consider the special case when $h = \frac{p}{\sqrt{2}}$. From (38) we have

$$e = A^{-1}t = (A_0 + Q)^{-1}t = (I + A_0^{-1}Q)^{-1}A_0^{-1}t.$$

Noting (45), we apply Lemma 7, (41), (40) and the fact $n = \frac{b-a}{p}$, giving

$$\begin{aligned} \|e\| &\leq \|(I + A_0^{-1}Q)^{-1}\|\ \|A_0^{-1}\|\ \|t\| \\ &\leq \frac{\|A_0^{-1}\|\ \|t\|}{1 - \|A_0^{-1}Q\|} \\ &\leq \frac{(b-a)^2}{8p^2} \left(\frac{1}{240} p^6 M_6\right) \left(\frac{1}{1 - K\hat{f}}\right) \\ &= \frac{K M_6 p^4}{240(1 - K\hat{f})} \ \cong\ O(p^4). \end{aligned}$$

This shows that (33) is fourth order convergent when $h = \frac{p}{\sqrt{2}}$. For other values of $h \in (0, p]$, from (39) we have $\|t\| \cong O(p^4)$ and a similar argument then leads to (33) is second order convergent.

Next, suppose $f(x) \equiv f_0 > 0$. Here, the matrix A is strictly diagonally dominant. It is well known that for a strictly diagonally dominant matrix,

$$\|A^{-1}\| \leq \left[\min_i \left(|a_{ii}| - \sum_{j \neq i} |a_{ij}| \right) \right]^{-1}.$$

Comparing (46) and (47), it is easily checked that $\min_i \left[|a_{ii}| - \sum_{j \neq i} |a_{ij}| \right]$ occurs when $2 \leq i \leq n-2$. Hence, using (46), we find, if $\frac{(p^2-h^2)}{6} f_0 - 1 \geq 0$,

$$\|A^{-1}\| \leq \left[\frac{2(p^2+2h^2)}{6} f_0 + 4 \right]^{-1} \leq \frac{1}{\frac{p^2}{3} f_0 + 4} \leq \frac{3}{p^2 f_0}. \tag{48}$$

If $\frac{(p^2-h^2)}{6} f_0 - 1 \leq 0$, we get

$$\|A^{-1}\| \leq \frac{1}{p^2 f_0}. \tag{49}$$

A combination of (48) and (49) leads to

$$\|A^{-1}\| \leq \max \left\{ \frac{3}{p^2 f_0}, \frac{1}{p^2 f_0} \right\} = \frac{3}{p^2 f_0}. \tag{50}$$

Now for the special case $h = \frac{p}{\sqrt{2}}$, from (38), (40) and (50) we get

$$\|e\| \leq \|A^{-1}\| \, \|t\| \leq \frac{3}{p^2 f_0} \left(\frac{1}{240} p^6 M_6 \right) = \frac{M_6 p^4}{80 f_0} \cong O(p^4).$$

Hence, (33) is fourth order convergent when $h = \frac{p}{\sqrt{2}}$. For other values of $h \in (0, p]$, from (39) we have $\|t\| \cong O(p^4)$ and it follows that (33) is second order convergent. □

3.3 Examples

In this section, we present two numerical examples to demonstrate the discrete cubic spline method proposed in Sect. 3.1 as well as to illustrate the comparative performance with some well known numerical methods.

Example 5 Consider the boundary value problem

$$y'' = \frac{2}{x^2} y - \frac{1}{x}, \qquad y(2) = y(3) = 0. \tag{51}$$

The exact solution is given by $y(x) = \frac{1}{38} \left(-5x^2 + 19x - \frac{36}{x} \right)$.

Table 5 (Examples 5 and 6) Maximum absolute errors of discrete cubic spline method with $h = \frac{p}{\sqrt{2}}$

p	BVP (51)	BVP (52)
1/8	1.74×10^{-7}	1.74×10^{-3}
1/16	1.10×10^{-8}	1.13×10^{-4}
1/32	6.85×10^{-10}	7.28×10^{-6}

Example 6 Consider the boundary value problem

$$y'' = 100y, \qquad y(0) = y(1) = 1. \tag{52}$$

The exact solution is given by $y(x) = \frac{\cosh(10x-5)}{\cosh 5}$.

Clearly, both (51) and (52) satisfy the conditions of Theorem 7 and so each has a unique discrete cubic spline solution.

First, we choose $h = \frac{p}{\sqrt{2}}$. The maximum absolute errors $\max_i |y_i - s_i|$ for different mesh sizes p are given in Table 5. We note that if the mesh size p is reduced by a factor of $\frac{1}{2}$, then the maximum absolute errors are approximately reduced by $\left(\frac{1}{2}\right)^4 = \frac{1}{16}$. Thus, the numerical results confirm that our method is fourth order convergent when $h = \frac{p}{\sqrt{2}}$, which verifies Theorem 8. Moreover, the maximum absolute errors recorded in Table 5 *coincide* with those obtained by Khan [36] using the parametric cubic spline method with the parameters $(\alpha, \beta) = \left(\frac{1}{12}, \frac{5}{12}\right)$ in which case the method is also of order 4. We remark that the expression of the spline given by our method is much easier to obtain and the approximate values s_i are easy to compute, while in [36] only s_i can be computed but the expression of the parametric cubic spline cannot be obtained.

Next, we choose $h = \frac{3}{4}p$ in order to compare with other second order methods. The maximum absolute errors $\max_i |y_i - s_i|$ obtained by various methods for the boundary value problem (51) are given in Table 6. The numerical experiment confirms that our method is second order convergent when $h = \frac{3}{4}p$ (Theorem 8), and our results are notably *better* than others'.

Next, we shall compare the performance of the 'non-traditional' *continuous* cubic spline method of [5] (which is superior to traditional cubic spline method) with our *discrete* cubic spline method. The values of $\max_i |y_i^{(k)} - s_i^{(k)}|$, $k = 0, 1, 2$ obtained for the boundary value problem (51) by using the method in [5] and our method with $h = \frac{2}{3}p$ (second order convergent) are presented in Table 7. We observe that the actual error $\max_i |y_i - s_i|$ of our method is much smaller, whereas $\max_i |y_i' - s_i'|$ is slightly worse, but $\max_i |y_i'' - s_i''|$ is again much smaller—this indicates that our discrete cubic spline method gives better approximation of $y(x_i)$ and $y''(x_i)$ for the boundary value problem (51).

Finally, in Table 8 we present the maximum absolute errors $\max_i |y_i - s_i|$ for the boundary value problem (52) obtained by various second order methods. Once again we observe that our method is second order convergent and offers *better* results than

Table 6 (Example 5) Comparison with other *second order* methods (BVP (51))

Method	$p = 1/4$	$p = 1/8$	$p = 1/16$
Our method with $h = (3/4)p$	1.77×10^{-5}	5.00×10^{-6}	1.29×10^{-6}
Parametric cubic spline [36]			
$(\alpha, \beta) = (1/14, 3/7)$	2.05×10^{-5}	5.74×10^{-6}	1.47×10^{-6}
$(\alpha, \beta) = (1/10, 2/5)$	3.50×10^{-5}	8.46×10^{-6}	2.09×10^{-6}
$(\alpha, \beta) = (1/18, 4/9)$	5.14×10^{-5}	1.36×10^{-6}	3.46×10^{-6}
Cubic spline [5]	5.49×10^{-5}	1.87×10^{-5}	5.07×10^{-6}
Collocation—quadratic spline [56]	7.93×10^{-5}	2.06×10^{-5}	5.20×10^{-6}
Quadratic spline [6]	1.60×10^{-4}	2.66×10^{-5}	5.58×10^{-6}
Cubic spline [4]	1.65×10^{-4}	4.17×10^{-5}	1.04×10^{-5}
Second order centered-difference	2.79×10^{-4}	5.42×10^{-5}	1.19×10^{-5}

Table 7 (Example 5) Maximum absolute errors for BVP (51)

Method	p	$\max_i \lvert y_i - s_i \rvert$	$\max_i \lvert y_i' - s_i' \rvert$	$\max_i \lvert y_i'' - s_i'' \rvert$
Continuous cubic	1/10	1.247×10^{-5}	7.818×10^{-5}	8.734×10^{-4}
spline method [5]	1/20	3.286×10^{-6}	1.931×10^{-5}	2.211×10^{-4}
	1/40	8.466×10^{-7}	4.812×10^{-6}	5.546×10^{-5}
Discrete cubic	1/10	3.038×10^{-6}	2.797×10^{-4}	1.082×10^{-6}
spline method	1/20	7.461×10^{-7}	7.003×10^{-5}	2.655×10^{-7}
with $h = \frac{2}{3}p$	1/40	1.858×10^{-7}	1.751×10^{-5}	6.630×10^{-8}

other methods. While doing the numerical experiments with different $h \in (0, p]$, we observe that as $h \to 0$, the result reduces to that of the continuous cubic spline [4]; when $h \to \frac{p}{\sqrt{2}}$, either approaching from 0 or approaching from p, the maximum absolute errors become smaller, this is in agreement with our theoretical results given in Theorem 8.

In Fig. 6, we plot the graphs of the discrete cubic spline solutions and the exact solutions of boundary value problems (51) and (52) for comparison.

Table 8 (Example 6) Comparison with other *second order* methods (BVP (52))

Method	$p = 1/16$	$p = 1/32$	$p = 1/20$	$p = 1/40$
Our method with				
$h = (2/3)p$	7.64×10^{-4}	1.74×10^{-4}	4.75×10^{-4}	1.10×10^{-4}
$h = (3/4)p$	6.18×10^{-4}	1.80×10^{-4}	4.32×10^{-4}	1.17×10^{-4}
Parametric cubic spline [36]				
$(\alpha, \beta) = (1/10, 2/5)$	1.28×10^{-3}	3.07×10^{-4}	8.17×10^{-4}	1.95×10^{-4}
$(\alpha, \beta) = (1/14, 3/7)$	7.22×10^{-4}	2.06×10^{-4}	5.00×10^{-4}	1.34×10^{-4}
$(\alpha, \beta) = (1/18, 4/9)$	1.83×10^{-3}	4.91×10^{-4}	1.22×10^{-3}	3.16×10^{-4}
Cubic spline [5]	2.27×10^{-3}	6.84×10^{-4}	1.57×10^{-3}	4.53×10^{-4}
Collocation—quadratic spline [56]	3.06×10^{-3}	7.58×10^{-4}		
Cubic spline [4]	6.05×10^{-3}	1.51×10^{-3}	3.93×10^{-3}	9.66×10^{-4}
Collocation—quadratic spline [35]			1.8×10^{-3}	4.7×10^{-4}

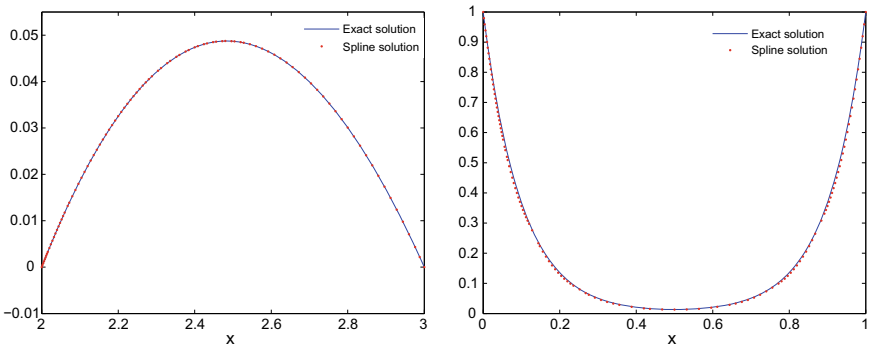

Fig. 6 (Example 6) BVP (51) (left) and BVP (52) (right) when $p = \frac{1}{8}$, $h = \frac{p}{\sqrt{2}}$

4 Deficient Discrete Cubic Spline Method for Second Order Boundary Value Problem

In this section, we use *deficient discrete cubic spline* to obtain approximate solution of a system of second order boundary value problems. It is shown that the method is of order 2 when a parameter takes a specific value. A well known numerical example

is presented to illustrate our method as well as to compare the performance with other numerical methods proposed in the literature. This section illustrates the work of [22].

The system of second order boundary value problems we consider is of the form

$$\begin{aligned} y'' &= \begin{cases} f(x), & a \le x \le c \\ g(x)y(x) + f(x) + r, & c \le x \le d \\ f(x), & d \le x \le b \end{cases} \\ y(a) &= \bar{\alpha}, \quad y(b) = \bar{\beta} \end{aligned} \tag{53}$$

with continuity conditions of y and y' at c and d. Here, f and g are continuous functions on $[a, b]$ and $[c, d]$ respectively, r, $\bar{\alpha}$ and $\bar{\beta}$ are real finite constants. This type of systems arises in the study of obstacle, unilateral, moving and free boundary value problems and has important applications in other branches of pure and applied sciences, see [15, 25, 26, 37, 46, 47, 55].

The literature on the numerical treatment of (53) is abundant. Noor and Khalifa [48] have solved (53) using a collocation method with *first-order* accuracy, adopting cubic B-splines as basis functions. Also, Noor and Tirmizi [51] have used finite difference schemes based on the central difference and the well known Numerov method to solve (53), all these give *first-order* accurate approximations to the solution of (53). In the paper of Al-Said et al. [11], the authors use spline and finite difference methods to obtain numerical solutions of (53) – it is shown that the numerical solutions derived using both spline and finite difference techniques are *first-order* accurate approximations regardless of the order of the methods used, and the authors illustrate this idea further with the second-order cubic spline method of Albasiny and Hoskins [4] and the fourth-order quintic spline method of Usmani and Warsi [66]. For methods of *second-order* accuracy, we note that a modified Numerov method has been discussed in [10]. Polynomial splines have also been employed in solving (53), for example in the papers of Al-Said [7–9], quadratic and cubic spline methods have been developed and analyzed, these methods are of *second order*. Further, quintic spline is used in [14] to solve (53), the method developed is *second-order* accurate. On the other hand, non-polynomial spline methods have been discussed in the papers [54, 61, 62], here the non-polynomial splines consist of polynomial and trigonometric parts such as span$\{1, x, \sin kx, \cos kx\}$ (cubic non-polynomial spline). So far the methods mentioned above are *non-iterative*. Some *iterative* numerical algorithms used to solve (53) include a modified decomposition method based on the Adomian decomposition method [45], as well as the variational iteration method [49]. Both of these methods do not require discretization, and the variational iteration method has the extra advantage of not having tedious computation of Adomian polynomials.

Motivated by all the above research especially the use of splines in solving (53), we shall employ a *deficient discrete cubic spline* to get a numerical solution of (53). In our proposed method, we shall relax the continuity of y' at c and d, and instead impose the continuity of the *first central difference* of y at c and d. The deficient discrete cubic spline is uniquely determined and it enables us to approximate y and

its derivatives at every point in the range of integration. Our proposed method is second-order convergent, and through a well know example on obstacle boundary value problem, we illustrate that our method outperforms other collocation, finite difference and spline methods for solving (53) in the literature [7–11, 14, 48, 51].

4.1 Deficient Discrete Cubic Spline Method

Let $P : a = x_0 < x_1 < \cdots < x_n = b$ be a uniform mesh of $[a, b]$ with step size $p = \frac{b-a}{n}$. Without loss of generality, we shall take

$$c = \frac{3a+b}{4} = x_{n/4} \quad \text{and} \quad d = \frac{a+3b}{4} = x_{3n/4},$$

and require the positive integer n (≥ 8) to be divisible by 4.

Let $h \in (0, p]$ be a given constant used in the central difference operator D_h.

Definition 8 Let $S(x; h)$ be a piecewise continuous function defined over $[a, b]$ (with mesh P) and $S_i(x)$ be its restriction on $[x_{i-1}, x_i]$, $1 \leq i \leq n$ passing through the points (x_{i-1}, s_{i-1}) and (x_i, s_i). We say $S(x; h)$ is a *deficient discrete cubic spline* if $S_i(x)$, $1 \leq i \leq n$ is a polynomial of degree 3 or less and

$$\begin{aligned} &(S_{i+1} - S_i)(x_i + jh) = 0, \quad j = -1, 0, 1, \quad i \in I \\ &(S_{i+1} - S_i)(x_i) = 0, \quad (S_{i+1} - S_i)(x_i + h) = (S_{i+1} - S_i)(x_i - h), \quad i = \tfrac{n}{4}, \tfrac{3n}{4} \end{aligned} \tag{54}$$

or equivalently

$$\begin{aligned} &D_h^{\{j\}} S_i(x_i) = D_h^{\{j\}} S_{i+1}(x_i), \quad j = 0, 1, \quad 1 \leq i \leq n-1 \\ &D_h^{\{2\}} S_i(x_i) = D_h^{\{2\}} S_{i+1}(x_i), \quad i \in I. \end{aligned} \tag{55}$$

where $I = \left\{ i \in \mathbf{Z} \;\middle|\; 1 \leq i \leq n-1,\ i \neq \tfrac{n}{4}, \tfrac{3n}{4} \right\}$.

The above definition of deficient discrete cubic spline coincides with that given in the paper of Rana and Dubey [53]. It has been observed [53] that deficient splines are more applicable than usual splines as they require less continuity requirement at the mesh points.

We shall approximate a solution $y(x)$ of (53) by the deficient discrete cubic spline $S(x; h)$, i.e., $y(x)$ will be approximated by $S_i(x)$ over the subinterval $[x_{i-1}, x_i]$, $1 \leq i \leq n$. Indeed, for any $x \in [a, b]$ (x may be between mesh points), we propose the following approximation

$$\begin{aligned} &y(x) \cong S(x; h), \quad y'(x) \cong D_h^{\{1\}} S(x; h), \quad x \in [a, b] \\ &y''(x) \cong \begin{cases} f(x), & x \in [a, c) \cup (d, b] \\ g(x)S(x; h) + f(x) + r, & x \in (c, d). \end{cases} \end{aligned} \tag{56}$$

In particular, at the mesh points we have

$$\begin{aligned} &y_i \cong s_i \equiv S_i(x_i), \quad y_i' \cong s_i' \equiv D_h^{\{1\}} S_i(x_i), \quad 0 \le i \le n \\ &y_i'' \cong \begin{cases} f_i, & 0 \le i \le \frac{n}{4} - 1 \text{ and } \frac{3n}{4} + 1 \le i \le n \\ g_i s_i + f_i + r, & \frac{n}{4} + 1 \le i \le \frac{3n}{4} - 1. \end{cases} \end{aligned} \tag{57}$$

Moreover, since the left and right second derivatives y_-'' and y_+'' exist at both c and d, we propose

$$\begin{aligned} &\text{when } i = \tfrac{n}{4}, \quad y_{i-}'' \cong f_i, \qquad\qquad y_{i+}'' \cong g_i s_i + f_i + r, \\ &\text{when } i = \tfrac{3n}{4}, \quad y_{i-}'' \cong g_i s_i + f_i + r, \quad y_{i+}'' \cong f_i. \end{aligned} \tag{58}$$

From the boundary conditions, we note that $s_0 = y_0 = \bar{\alpha}$ and $s_n = y_n = \bar{\beta}$, and s_i is an approximate of y_i, $1 \le i \le n-1$.

We shall now obtain an explicit expression of $S_i(x)$. Let $c_i = D_h^{\{2\}} S(x_i; h)$ denote the 'discrete moments'. Taking into account the fact that we approximate $y(x)$ by $S(x; h)$ as well as (57) and (58), we set

$$\begin{aligned} c_i = &\begin{cases} f_i, & 0 \le i \le \frac{n}{4} - 1 \text{ and } \frac{3n}{4} + 1 \le i \le n \\ g_i s_i + f_i + r, & \frac{n}{4} + 1 \le i \le \frac{3n}{4} - 1 \end{cases} \\ &\text{when } i = \tfrac{n}{4}, \quad c_{i-} = f_i, \qquad\qquad c_{i+} = g_i s_i + f_i + r, \\ &\text{when } i = \tfrac{3n}{4}, \quad c_{i-} = g_i s_i + f_i + r, \quad c_{i+} = f_i. \end{aligned} \tag{59}$$

Since $D_h^{\{2\}} S(x; h)$ is piecewise linear, using a similar argument as in Sect. 2, we obtain (26)–(29) for $x \in [x_{i-1}, x_i]$, $1 \le i \le n$. Here, when $i = \frac{n}{4}$, $\frac{3n}{4}$, we take $c_i = c_{i,-}$; when $i = \frac{n}{4} + 1$, $\frac{3n}{4} + 1$, we take $c_{i-1} = c_{i-1,+}$ (see (59) for the definitions).

For $i \in I$, the 'continuity' requirement $D_h^{\{1\}} S_i(x_i) = D_h^{\{1\}} S_{i+1}(x_i)$ leads to the system of equations

$$\frac{(p^2 - h^2)}{6} c_{i-1} + \frac{2(2p^2 + h^2)}{6} c_i + \frac{(p^2 - h^2)}{6} c_{i+1} = s_{i-1} - 2s_i + s_{i+1}, \quad i \in I. \tag{60}$$

Note that in (60), when $i = \frac{n}{4} - 1$, $\frac{3n}{4} - 1$, we take $c_{i+1} = c_{i+1,-}$; when $i = \frac{n}{4} + 1$, $\frac{3n}{4} + 1$, we take $c_{i-1} = c_{i-1,+}$ (see (59) for the definitions).

When $i = \frac{n}{4}$, $\frac{3n}{4}$, from (29) we have the following

$$D_h^{\{1\}} S_i(x_i) = \frac{p}{2} c_{i-} + \frac{s_i - s_{i-1}}{p} - \frac{(p^2 - h^2)}{6p} (c_{i-} - c_{i-1}),$$

$$D_h^{\{1\}} S_{i+1}(x_i) = -\frac{p}{2} c_{i+} + \frac{s_{i+1} - s_i}{p} - \frac{(p^2 - h^2)}{6p} (c_{i+1} - c_{i+}),$$

and $D_h^{\{1\}} S_i(x_i) = D_h^{\{1\}} S_{i+1}(x_i)$ yields

$$\frac{(p^2-h^2)}{6}c_{i-1}+\frac{(h^2+2p^2)}{6}(c_{i+}+c_{i-})+\frac{(p^2-h^2)}{6}c_{i+1}=s_{i-1}-2s_i+s_{i+1},\quad i=\tfrac{n}{4},\tfrac{3n}{4}. \tag{61}$$

Upon substituting (59) into (60) and (61), we obtain

$$-s_{i-1}+2s_i-s_{i+1}=-\tfrac{(p^2-h^2)}{6}f_{i-1}-\tfrac{2(2p^2+h^2)}{6}f_i-\tfrac{(p^2-h^2)}{6}f_{i+1},\quad 1\le i\le \tfrac{n}{4}-1 \text{ and } \tfrac{3n}{4}+1\le i\le n-1 \tag{62}$$

$$\left[-1+\tfrac{(p^2-h^2)}{6}g_{i-1}\right]s_{i-1}+\left[2+\tfrac{2(2p^2+h^2)}{6}g_i\right]s_i+\left[-1+\tfrac{(p^2-h^2)}{6}g_{i+1}\right]s_{i+1}$$
$$=-\tfrac{(p^2-h^2)}{6}f_{i-1}-\tfrac{2(2p^2+h^2)}{6}f_i-\tfrac{(p^2-h^2)}{6}f_{i+1}-p^2r,\quad \tfrac{n}{4}+1\le i\le \tfrac{3n}{4}-1 \tag{63}$$

and

$$-s_{i-1}+\left[2+\tfrac{(2p^2+h^2)}{6}g_i\right]s_i+\left[-1+\tfrac{(p^2-h^2)}{6}g_{i+1}\right]s_{i+1}$$
$$=-\tfrac{(p^2-h^2)}{6}f_{i-1}-\tfrac{2(2p^2+h^2)}{6}f_i-\tfrac{(p^2-h^2)}{6}f_{i+1}-\tfrac{p^2}{2}r,\quad i=\tfrac{n}{4},\tfrac{3n}{4}. \tag{64}$$

With $s_0=\bar{\alpha}$ and $s_n=\bar{\beta}$, we may solve (62)–(64) to get $s_i,\ 1\le i\le n-1$. The unique solvability of the system (62)–(64) will be proved in the next section.

For clarity, the steps of computing the deficient discrete cubic spline solution of (53) are listed as follows:

(i) Fix the mesh P (and hence the mesh size p) and choose a value for h ($\in (0, p]$).
(ii) Solve (62)–(64) to get $s_i,\ 1\le i\le n-1$, which approximates y_i.
(iii) Calculate c_i by using (59). Noting (57) and (58), c_i serves as an approximate to y_i''.
(iv) Compute $D_h^{\{1\}}S_i(x_i)$ from (29). Noting (57), $s_i'=D_h^{\{1\}}S_i(x_i)$ serves as an approximate to y_i'.
(v) The deficient discrete cubic spline solution $S_i(x)$ over the subinterval $[x_{i-1}, x_i]$ can be obtained using (27). The first central difference $D_h^{\{1\}}S_i(x)$ can also be obtained using (29). These can be used to approximate $y(x)$ and $y'(x)$ for any $x\in[a,b]$.

4.2 Convergence Analysis

In this section, we shall establish the existence of a unique deficient discrete cubic spline solution for (53) (i.e., (62)–(64) has a unique solution) and also conduct a convergence analysis for the method presented in the previous section.

Let $e_i=y_i-s_i,\ 1\le i\le n-1$ be the errors. Let $Y=[y_i],\ S=[s_i],\ W=[w_i],\ T=[t_i]$ and $E=[e_i]$ be $(n-1)$-dimensional column vectors. The system (62)–(64) can be written as

$$\bar{A}S = W \tag{65}$$

where

$$\bar{A} = A_0 + \bar{B}, \tag{66}$$

$A_0 = [a_{ij}]$ is given in (35) and $\bar{B} = [\bar{b}_{ij}]$ is a $(n-1) \times (n-1)$ matrix given by

$$\bar{b}_{ij} = \begin{cases} \frac{1}{6}2(2p^2+h^2)g_i, & i = j, \quad \frac{n}{4}+1 \le i \le \frac{3n}{4}-1 \\ \frac{1}{6}(p^2-h^2)g_{i-1}, & i-j = 1, \quad \frac{n}{4}+1 \le i \le \frac{3n}{4}-1 \\ \frac{1}{6}(p^2-h^2)g_{i+1}, & j-i = 1, \quad \frac{n}{4}+1 \le i \le \frac{3n}{4}-1 \\ \frac{1}{6}(2p^2+h^2)g_i, & i = j, \quad i = \frac{n}{4}, \frac{3n}{4} \\ \frac{1}{6}(p^2-h^2)g_{i+1}, & j-i = 1, \quad i = \frac{n}{4}, \frac{3n}{4} \\ 0, & \text{otherwise} \end{cases} \tag{67}$$

and

$$w_i = \begin{cases} \bar{\alpha} - \frac{1}{6}\left[(p^2-h^2)f_{i-1} + 2(2p^2+h^2)f_i + (p^2-h^2)f_{i+1}\right], & i = 1 \\ -\frac{1}{6}\left[(p^2-h^2)f_{i-1} + 2(2p^2+h^2)f_i + (p^2-h^2)f_{i+1}\right], & \\ \qquad 2 \le i \le \frac{n}{4}-1 \text{ and } \frac{3n}{4}+1 \le i \le n-2 & \\ -\frac{1}{6}\left[(p^2-h^2)f_{i-1} + 2(2p^2+h^2)f_i + (p^2-h^2)f_{i+1}\right] - \frac{p^2}{2}r, & i = \frac{n}{4}, \frac{3n}{4} \\ -\frac{1}{6}\left[(p^2-h^2)f_{i-1} + 2(2p^2+h^2)f_i + (p^2-h^2)f_{i+1} + 6p^2 r\right], & \\ \qquad \frac{n}{4}+1 \le i \le \frac{3n}{4}-1 & \\ \bar{\beta} - \frac{1}{6}\left[(p^2-h^2)f_{i-1} + 2(2p^2+h^2)f_i + (p^2-h^2)f_{i+1}\right], & i = n-1. \end{cases} \tag{68}$$

From (65) we have $\bar{A}(Y-E) = W$ or

$$\bar{A}Y = W + T \tag{69}$$

where

$$T = \bar{A}E. \tag{70}$$

For $i \in I$, the i-th equation of system (69) is

$$-y_{i-1} + 2y_i - y_{i+1} = -\frac{1}{6}[(p^2-h^2)y''_{i-1} + 2(2p^2+h^2)y''_i + (p^2-h^2)y''_{i+1}] + t_i$$

while for $i = \frac{n}{4}, \frac{3n}{4}$, we have

$$-y_{i-1} + 2y_i - y_{i+1} = -\frac{(p^2-h^2)}{6}(y''_{i-1} + y''_{i+1}) - \frac{(h^2+2p^2)}{6}(y''_{i+} + y''_{i-}) + t_i.$$

By Taylor series expansion, we obtain the truncation error t_i, $1 \le i \le n-1$ as (39). Hence, when $h = \frac{p}{\sqrt{2}}$, we get

$$\|T\| = \frac{1}{240} p^6 M_6 \tag{71}$$

where $M_6 = \max_x |y^{(6)}(x)|$.

The next result gives the convergence order of the deficient discrete cubic spline method.

Theorem 9 *Suppose* $L = \frac{(b-a)^2}{8}\hat{g} < 1$ *where* $\hat{g} = \max_{0\le i\le n} |g_i|$. *Then, the system (65) has a unique solution and*

$$\|E\| \cong O(p^2) \quad \text{if} \quad h = \frac{p}{\sqrt{2}},$$

i.e., the deficient discrete cubic spline method (65) is second-order convergent if $h = \frac{p}{\sqrt{2}}$.

Proof If (65) has a unique solution, then it can be written as

$$S = \bar{A}^{-1} W = (A_0 + \bar{B})^{-1} W = [A_0(I + A_0^{-1}\bar{B})]^{-1} W = (I + A_0^{-1}\bar{B})^{-1} A_0^{-1} W. \tag{72}$$

From Lemma 6 the inverse A_0^{-1} exists, hence it remains to show that $(I + A_0^{-1}\bar{B})$ is nonsingular.

enumerateFrom (67), we find $\|\bar{B}\| \le p^2\hat{g}$. Then, using (41) gives

$$\|A_0^{-1}\bar{B}\| \le \|A_0^{-1}\|\,\|\bar{B}\| \le \frac{(b-a)^2}{8p^2}\left(p^2\hat{g}\right) = L < 1. \tag{73}$$

Hence, we conclude from Lemma 7 that $(I + A_0^{-1}\bar{B})$ is nonsingular, and (65) has a unique solution given by (72).

Next, we consider the special case when $h = \frac{p}{\sqrt{2}}$. From (70), we have

$$E = \bar{A}^{-1} T = (A_0 + \bar{B})^{-1} T = (I + A_0^{-1}\bar{B})^{-1} A_0^{-1} T. \tag{74}$$

Applying (41), (71), (73) and Lemma 7, it follows from (74) that

$$\begin{aligned}
\|E\| &\le \|(I + A_0^{-1}\bar{B})^{-1}\|\,\|A_0^{-1}\|\,\|T\| \\
&\le \frac{\|A_0^{-1}\|\,\|T\|}{1 - \|A_0^{-1}\bar{B}\|} \\
&\le \frac{(b-a)^2}{8p^2}\left(\frac{1}{240}p^6 M_6\right)\left(\frac{1}{1-L}\right) \\
&= \frac{(b-a)^2 M_6 p^4}{1920(1-L)}.
\end{aligned}$$

This shows that (65) is a fourth-order convergence method when $h = \frac{p}{\sqrt{2}}$. However, as observed in [14] the solution exists continuously only up to the second derivative,

therefore the method is only second-order accurate over the whole interval. In fact, in the paper [62] it is also shown that $\|E\| \approx O(p^4)$ yet the method developed is second-order convergent, similar observations have also been noted in [10, 11]. □

Remark 7 Theorem 9 gives a *sufficient* condition for the existence and uniqueness of deficient discrete cubic spline solution and the order of convergence. Actually, the weakest condition is just to have the matrix $\bar{A}$ *invertible*. Then, the system (65) has a unique solution $S = \bar{A}^{-1}W$. Moreover, the deficient discrete cubic spline method (65) is convergent when $h = \frac{p}{\sqrt{2}}$, since from (70) we have $E = \bar{A}^{-1}T$ which in view of (71) leads to

$$\|E\| \leq \|\bar{A}^{-1}\| \, \|T\| \leq \frac{1}{240} p^6 M_6 \|\bar{A}^{-1}\| < \infty.$$

4.3 Obstacle Boundary Value Problem

To illustrate the application of the deficient discrete cubic spline method, we consider the obstacle boundary value problem

$$\begin{array}{ll} -y''(x) \geq f(x), & \text{on } \Omega = [0, \pi] \\ y(x) \geq \psi(x), & \text{on } \Omega = [0, \pi] \\ (y''(x) - f(x))(y(x) - \psi(x)) = 0, & \text{on } \Omega = [0, \pi] \\ y(0) = y(\pi) = 0, & \end{array} \tag{75}$$

where $f(x)$ is a given force acting on the string and $\psi(x)$ is the elastic obstacle.

The problem (75) has been considered by almost every paper on system of second order boundary value problems. As first discussed by Noor and Khalifa [48], by using the variational inequality approach, (75) is equivalent to the variational inequality problem (see [15, 25, 37, 50])

$$a(y, v - y) \geq \langle f, v - y \rangle, \quad \text{for all } v \in K \tag{76}$$

where $a(\cdot, \cdot)$ is a coercive continuous bilinear form and K is the closed convex set given by $K = \{v \in H_0^1(\Omega) \mid v \geq \psi \text{ on } \Omega\}$ ($H_0^1(\Omega)$ is a Sobolev space). Following the idea and technique of Lewy and Stampacchia [39], the variational inequality (76) can be written as

$$\begin{array}{l} y'' - [\mu(y - \psi)](y - \psi) = f, \quad 0 < x < \pi \\ y(0) = y(\pi) = 0 \end{array} \tag{77}$$

where $\mu(t)$, known as the penalty function, is the discontinuous function defined by

$$\mu(t) = \begin{cases} 1, & t \geq 0 \\ 0, & t < 0 \end{cases} \tag{78}$$

and ψ is the given obstacle function defined by

$$\psi(x) = \begin{cases} -1, & 0 \le x \le \frac{\pi}{4} \\ 1, & \frac{\pi}{4} \le x \le \frac{3\pi}{4} \\ -1, & \frac{3\pi}{4} \le x \le \pi. \end{cases} \tag{79}$$

Equation (77) describes the equilibrium configuration of an obstacle string pulled at the ends and lying over elastic step of constant height 1 and unit rigidity. Since the obstacle function ψ is known, it is possible to find the solution of the problem in the interval $[0, \pi]$.

From equations (77)–(79), one obtains the following system of boundary value problem

$$\begin{aligned} y'' &= \begin{cases} f, & 0 \le x \le \frac{\pi}{4} \text{ and } \frac{3\pi}{4} \le x \le \pi \\ y + f - 1, & \frac{\pi}{4} \le x \le \frac{3\pi}{4} \end{cases} \\ y(0) &= y(\pi) = 0 \end{aligned} \tag{80}$$

and the condition of the continuity of y and y' at $\frac{\pi}{4}$ and $\frac{3\pi}{4}$. We shall consider a special case of the system (80) in Example 7, this example is first discussed in [48] and subsequently considered in almost every paper on system of second order boundary value problems.

Example 7 [48] We consider the system (80) when $f = 0$, i.e.,

$$\begin{aligned} y'' &= \begin{cases} 0, & 0 \le x \le \frac{\pi}{4} \text{ and } \frac{3\pi}{4} \le x \le \pi \\ y - 1, & \frac{\pi}{4} \le x \le \frac{3\pi}{4} \end{cases} \\ y(0) &= y(\pi) = 0. \end{aligned} \tag{81}$$

The analytical solution for this problem is given by

$$y(x) = \begin{cases} \frac{4}{\gamma_1} x, & 0 \le x \le \frac{\pi}{4} \\ 1 - \frac{4}{\gamma_2} \cosh\left(\frac{\pi}{2} - x\right), & \frac{\pi}{4} \le x \le \frac{3\pi}{4} \\ \frac{4}{\gamma_1}(\pi - x), & \frac{3\pi}{4} \le x \le \pi \end{cases} \tag{82}$$

where $\gamma_1 = \pi + 4 \coth \frac{\pi}{4}$ and $\gamma_2 = \pi \sinh \frac{\pi}{4} + 4 \cosh \frac{\pi}{4}$.

We observe from the analytical solution that y and y' are continuous at $\frac{\pi}{4}$ and $\frac{3\pi}{4}$, but y'' is not continuous at these two points, so the overall accuracy of our method is only second order. This is also verified from the numerical evidence in Table 9.

In this example, we take $h = \frac{p}{\sqrt{2}}$. The system of linear equations (62)–(64) is explicitly given as

$$\begin{cases} -s_{i-1} + 2s_i - s_{i+1} = 0, \quad 1 \le i \le \frac{n}{4} - 1 \text{ and } \frac{3n}{4} + 1 \le i \le n - 1 \\ \left(-1 + \frac{p^2}{12}\right) s_{i-1} + \left(2 + \frac{5p^2}{6}\right) s_i + \left(-1 + \frac{p^2}{12}\right) s_{i+1} = p^2, \quad \frac{n}{4} + 1 \le i \le \frac{3n}{4} - 1 \\ -s_{i-1} + \left(2 + \frac{5p^2}{12}\right) s_i + \left(-1 + \frac{p^2}{12}\right) s_{i+1} = \frac{p^2}{2}, \quad i = \frac{n}{4}, \frac{3n}{4}. \end{cases}$$

Table 9 (Example 7) Maximum absolute errors $\max_i |y_i - s_i|$

Method	$p = \pi/20$	$p = \pi/40$	$p = \pi/80$
Deficient discrete cubic spline	1.19×10^{-3}	3.04×10^{-4}	7.68×10^{-5}
Cubic spline [9]	1.26×10^{-3}	3.29×10^{-4}	8.43×10^{-5}
Modified Numerov method [10]	1.65×10^{-3}	4.33×10^{-4}	1.11×10^{-4}
Cubic spline [8]	1.94×10^{-3}	4.99×10^{-4}	1.27×10^{-4}
Quadratic spline [7]	2.20×10^{-3}	5.87×10^{-4}	1.51×10^{-4}
Quintic spline [14]	2.57×10^{-3}	7.31×10^{-4}	1.94×10^{-4}
Collocation-cubic [48]	1.40×10^{-2}	7.71×10^{-3}	4.04×10^{-3}
Cubic spline [11]	1.80×10^{-2}	9.13×10^{-3}	4.60×10^{-3}
Quintic spline [11]	1.82×10^{-2}	9.17×10^{-3}	4.61×10^{-3}
Numerov [51]	2.32×10^{-2}	1.21×10^{-2}	6.17×10^{-3}
Scheme (4.9) [51]	2.50×10^{-2}	1.29×10^{-2}	6.58×10^{-3}
Cubic non-polynomial spline			
[62] $\alpha = \frac{1}{16}$, $\beta = \frac{7}{16}$	6.43×10^{-4}	1.83×10^{-4}	4.87×10^{-5}
[30] $\alpha = \frac{1}{16}$, $\beta = \frac{7}{16}$	5.01×10^{-4}	1.33×10^{-4}	3.40×10^{-5}
Quintic non-polynomial spline [54]			
$\alpha = \frac{1}{12}$, $\beta = \frac{5}{12}$	5.32×10^{-10}	2.90×10^{-12}	5.85×10^{-14}

For different values of p, we can solve the unknowns s_i, $1 \le i \le n-1$ from the above system. Then, we can get c_i using (59) and finally obtain the deficient discrete spline $S_i(x)$ as well as $D_h^{\{1\}} S_i(x)$ in (27) and (29) respectively. In Tables 9 and 10 respectively, we present the maximum absolute errors $\max_i |y_i - s_i|$ and $\max_i |y_i' - s_i'|$ obtained from our method as well as from other methods in the literature.

From Table 9, the numerical results confirm that our method is of second order. Moreover, our method *outperforms* other methods in [7–11, 14, 48, 51]. The non-polynomial spline methods [30, 54, 62] presented in Table 9 are of *second order* (cubic non-polynomial spline) and *sixth order* (quintic non-polynomial spline). Although non-polynomial spline methods offer excellent approximations to y_i's, but unlike our method, they may *not* be able to approximate y and its derivatives at every point in the range of integration, since the non-polynomial splines are *not* computable.

From Table 10, we see that our method gives better approximations for y' compared to the cubic and quintic spline methods in [9, 11].

To compare graphically, in Fig. 7 we plot the exact solution y and the deficient discrete cubic spline solution $S(x; h)$; in Fig. 8 we plot the exact y', $D_h^{\{1\}} S(x; h)$ and the first derivative of the cubic spline solution obtained in [8]; in Fig. 9 we plot the exact y'', $D_h^{\{2\}} S(x; h)$ and the second derivative of the cubic spline solution obtained in [8]. It is seen from the figures that our method gives better approximations for y, y' and y''.

Table 10 (Example 7) Maximum absolute errors $\max_i |y_i' - s_i'|$

Method	$p = \pi/20$	$p = \pi/40$	$p = \pi/80$
Deficient discrete cubic spline	7.58×10^{-4}	1.91×10^{-4}	4.79×10^{-5}
Cubic spline [9]	8.32×10^{-4}	2.09×10^{-4}	5.22×10^{-5}
Cubic spline [11]	2.75×10^{-2}	1.39×10^{-2}	7.02×10^{-3}
Quintic spline [11]	9.05×10^{-2}	4.70×10^{-2}	2.44×10^{-2}

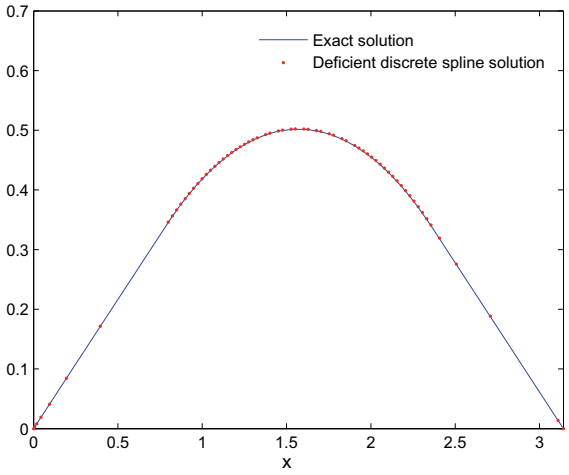

Fig. 7 (Example 7) Exact solution versus deficient discrete spline solution $\left(p = \frac{\pi}{20}\right)$

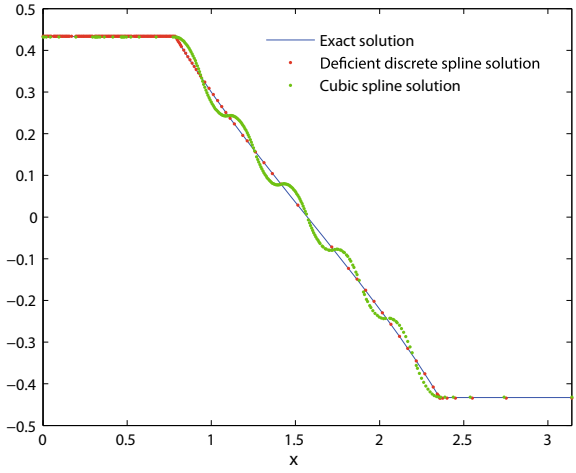

Fig. 8 (Example 7) The first derivative of exact solution versus the first central difference of deficient discrete cubic spline solution/first derivative of cubic spline solution [8] $\left(p = \frac{\pi}{20}\right)$

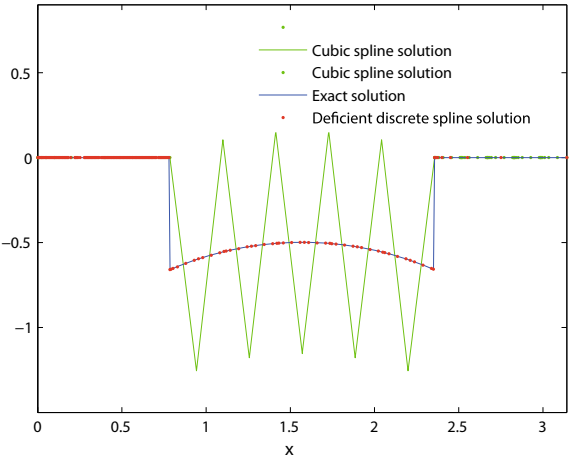

Fig. 9 (Example 7) The second derivative of exact solution versus the second central difference of deficient discrete cubic spline solution/second derivative of cubic spline solution [8] $\left(p = \frac{\pi}{20}\right)$

5 Conclusion

In this paper, we survey the contributions made to discrete splines in the literature and present some applications of discrete splines in the numerical treatment of boundary value problems. More specifically, we illustrate two types of discrete spline interpolation, namely the *discrete quintic spline interpolation* involving forward differences and the *periodic discrete quintic spline interpolation* involving central differences. In both cases, the explicit error estimates between the function and its discrete spline interpolate are obtained, and the interpolation of two-variable functions (including error estimates) is also developed. Further, to demonstrate the usefulness of discrete splines in numerical methods, we present a *discrete cubic spline method* for a second order boundary value problem that arises from plate deflection theory, and a *deficient discrete cubic spline method* for a system of second order boundary value problems that arises from obstacle, unilateral, moving and free boundary value problems. The convergence analysis as well as numerical examples are presented to illustrate the efficiency of the methods.

References

1. Agarwal, R.P., Wong, P.J.Y.: Error Inequalities in Polynomial Interpolation and Their Applications. Kluwer, Dordrecht (1993)
2. Agarwal, R.P., Wong, P.J.Y.: Explicit error bounds for the derivatives of spline interpolation in L_2 norm. Appl. Anal. **55**, 189–205 (1994)
3. Ahlberg, J.H., Nilson, E.N., Walsh, J.L.: The Theory of Splines and Applications. Academic Press, New York (1967)
4. Albasiny, E.L., Hoskins, W.D.: Cubic spline solutions to two-point boundary value problems. Comput. J. **12**, 151–153 (1969)

5. Al-Said, E.A.: Cubic spline method for solving two-point boundary-value problems. Korean J. Comput. Appl. Math. **5**, 669–680 (1998)
6. Al-Said, E.A.: Quadratic spline solution of two point boundary value problems. J. Nat. Geom. **12**, 125–134 (1997)
7. Al-Said, E.A.: Spline solutions for system of second-order boundary-value problems. Int. J. Comput. Math. **62**, 143–154 (1996)
8. Al-Said, E.A.: Spline methods for solving system of second-order boundary-value problems. Int. J. Comput. Math. **70**, 717–727 (1999)
9. Al-Said, E.A.: The use of cubic splines in the numerical solution of a system of second-order boundary value problems. Comput. Math. Appl. **42**, 861–869 (2001)
10. Al-Said, E.A., Noor, M.A.: Modified Numerov method for solving system of second-order boundary-value problems. Korean J. Comput. Appl. Math. **8**, 129–136 (2001)
11. Al-Said, E.A., Noor, M.A., Al-Shejari, A.A.: Numerical solutions for system of second order boundary value problems. Korean J. Comput. Appl. Math. **5**, 659–667 (1998)
12. Ascher, U.M., Mettheij, R.M., Russell, R.D.: Numerical Solution of Boundary Value Problems for Ordinary Differential Equations. SIAM, Philadelphia (1995)
13. Astor, P.H., Duris, C.S.: Discrete L-splines. Numer. Math. **22**, 393–402 (1974)
14. Aziz, T., Khan, A., Khan, I.: Quintic splines method for second-order boundary value problems. Int. J. Comput. Math. **85**, 735–743 (2008)
15. Baiocchi, C., Capelo, A.: Variational and Quasi-Variational Inequalities. Wiley, New York (1984)
16. Bickley, W.G.: Piecewise cubic interpolation and two-point boundary problems. Comput. J. **11**, 206–208 (1968)
17. Chen, F., Wong, P.J.Y.: Error inequalities for quintic and biquintic discrete Hermite interpolation. J. Comput. Appl. Math. **235**, 4589–4600 (2011)
18. Chen, F., Wong, P.J.Y.: Error estimates for discrete spline interpolation: quintic and biquintic splines. J. Comput. Appl. Math. **236**, 3835–3854 (2012)
19. Chen, F., Wong, P.J.Y.: Solutions of Fredholm integral equations via discrete biquintic splines. Math. Comput. Model. **57**, 551–563 (2013)
20. Chen, F., Wong, P.J.Y.: On periodic discrete spline interpolation: quintic and biquintic cases. J. Comput. Appl. Math. **255**, 282–296 (2014)
21. Chen, F., Wong, P.J.Y.: Discrete cubic spline method for second order boundary value problems. Int. J. Comput. Math. **91**, 1041–1053 (2014)
22. Chen, F., Wong, P.J.Y.: Deficient discrete cubic spline solution for a system of second order boundary value problems. Numer. Algorithms **66**, 793–809 (2014)
23. Chen, F., Wong, P.J.Y.: Numerical solutions of fourth order Lidstone boundary value problems using discrete quintic splines. J. Comput. Anal. Appl. **16**, 540–551 (2014)
24. Collatz, L.: The Numerical Treatment of Differential Equations. Springer, Berlin (1960)
25. Cottle, R.W., Giannessi, F., Lions, J.L.: Variational Inequalities and Complementarity Problems: Theory and Applications. Wiley, Chichester (1980)
26. Crank, J.: Free and Moving Boundary Problems. Clarendon Press, Oxford, UK (1984)
27. Danumjaya, P., Pani, A.K.: Orthogonal cubic spline collocation method for the extended Fisher-Kolmogorov equation. J. Comput. Appl. Math. **174**, 101–117 (2005)
28. Dikshit, H.P., Powar, P.L.: Discrete cubic spline interpolation. Numer. Math. **40**, 71–78 (1982)
29. Dikshit, H.P., Rana, S.S.: Discrete cubic spline interpolation over a nonuniform mesh. Rocky Mt. J. Math. **17**, 709–718 (1987)
30. Ding, Q., Wong, P.J.Y.: Mid-knot cubic non-polynomial spline for a system of second-order boundary value problems. Boundary Value Problems 2018, paper no. 156, 16 pp (2018)
31. Fröberg, C.: Numerical Mathematics. Theory and Computer Applications. Benjamin/Cummings, Reading, MA (1985)
32. Fyfe, D.J.: The use of cubic splines in the solution of two-point boundary value problems. Comput. J. **12**, 188–192 (1969)
33. Han, X.: A degree by degree recursive construction of Hermite spline interpolants. J. Comput. Appl. Math. **225**, 113–123 (2009)

34. Henrici, P.: Discrete Variable Methods in Ordinary Differential Equations. Wiley, New York (1961)
35. Khalifa, A.K., Eilbeck, J.C.: Collocation with quadratic and cubic splines. IMA J. Numer. Anal. **2**, 111–121 (1982)
36. Khan, A.: Parametric cubic spline solution of two point boundary value problems. Appl. Math. Comput. **154**, 175–182 (2004)
37. Kikuchi, N., Oden, J.T.: Contact Problems in Elasticity. SIAM Publishing Co., Philadelphia (1988)
38. Kouibia, A., Pasadas, M.: Approximation by interpolating variational splines. J. Comput. Appl. Math. **218**, 342–349 (2008)
39. Lewy, H., Stampacchia, G.: On the regularity of the solution of a variational inequality. Comm. Pure Appl. Math. **22**, 153–188 (1969)
40. Lyche, T.: Discrete cubic spline interpolation. Report RRI 5, University of Oslo (1975)
41. Lyche, T.: Discrete polynomial spline approximation methods. Spline Functions. Lecture Notes in Mathematics, vol. 501, pp. 144–176. Springer, Berlin (1976)
42. Lyche, T.: Discrete cubic spline interpolation. BIT **16**, 281–290 (1976)
43. Mangasarian, O.L., Schumaker, L.L.: Discrete splines via mathematical programming. SIAM J. Control **9**, 174–183 (1971)
44. Mangasarian, O.L., Schumaker, L.L.: Best summation formulae and discrete splines. SIAM J. Numer. Anal. **10**, 448–459 (1973)
45. Momani, S.: Solving a system of second order obstacle problems by a modified decomposition method. Appl. Math. E-Notes **6**, 141–147 (2006)
46. Noor, M.A.: Some recent advances in variational inequalities, Part I. Basic concepts. New Zealand J. Math. **26**, 53–80 (1997)
47. Noor, M.A.: Some recent advances in variational inequalities, Part II. Other concepts. New Zealand J. Math. **26**, 229–255 (1997)
48. Noor, M.A., Khalifa, A.K.: Cubic splines collocation methods for unilateral problems. Internat. J. Engrg. Sci. **25**, 1527–1530 (1987)
49. Noor, M.A., Noor, K.I., Rafiq, M., Al-Said, E.A.: Variational iteration method for solving a system of second-order boundary value problems. Int. J. Nonlinear Sci. Numer. Simul. **11**, 1109–1120 (2010)
50. Noor, M.A., Noor, K.I., Rassias, Th: Some aspects of variational inequalities. J. Comput. Appl. Math. **47**, 285–312 (1993)
51. Noor, M.A., Tirmizi, S.I.A.: Finite difference technique for solving obstacle problems. Appl. Math. Lett. **1**, 267–271 (1988)
52. Rana, S.S.: Local behaviour of the first difference of a discrete cubic spline interpolator. Approx. Theory Appl. **6**, 58–64 (1990)
53. Rana, S.S., Dubey, Y.P.: Local behaviour of the deficient discrete cubic spline interpolator. J. Approx. Theory **86**, 120–127 (1996)
54. Rashidinia, J., Jalilian, R., Mohammadi, R.: Non-polynomial spline methods for the solution of a system of obstacle problems. Appl. Math. Comput. **188**, 1984–1990 (2007)
55. Rodrigues, J.F.: Obstacle Problems in Mathematical Physics. North-Holland, Amsterdam (1987)
56. Sakai, M., Usmani, R.A.: Quadratic spline and two-point boundary value problem. Publ. Res. Inst. Math. Sci. **19**, 7–13 (1983)
57. Sakai, M., Usmani, R.A.: On recursion relations for splines. J. Approx. Theory **65**, 200–206 (1991)
58. Schultz, M.H.: Spline Analysis. Prentice-Hall, Englewood Cliffs (1973)
59. Schumaker, L.L.: Constructive aspects of discrete polynomial spline functions. In: Lorentz, G.G. (ed.) Approximation Theory, pp. 469–476. Academic Press, New York (1973)
60. Siewer, R.: A constructive approach to nodal splines. J. Comput. Appl. Math. **203**, 289–308 (2007)
61. Siraj-ul-Islam, Noor, M.A., Tirmizi, I.A., Khan, M.A.: Quadratic non-polynomial spline approach to the solution of a system of second-order boundary-value problems. Appl. Math. Comput. **179**, 153–160 (2006)

62. Siraj-ul-Islam, Tirmizi, I.A.: Nonpolynomial spline approach to the solution of a system of second-order boundary-value problems. Appl. Math. Comput. **173**, 1208–1218 (2006)
63. Usmani, R.A.: Bounds for the solution of a second order differential equation with mixed boundary conditions. J. Engrg. Math. **9**, 159–164 (1975)
64. Usmani, R.A.: On quadratic spline interpolation. BIT **27**, 615–622 (1987)
65. Usmani, R.A., Sakai, M.: A connection between quartic spline and Numerov solution of a boundary value problems. Int. J. Comput. Math. **26**, 263–273 (1989)
66. Usmani, R.A., Warsi, S.A.: Quintic spline solutions of boundary value problems. Comput. Math. Appl. **6**, 197–203 (1980)
67. Wong, P.J.Y., Agarwal, R.P.: Explicit error estimates for quintic and biquintic spline interpolation. Comput. Math. Appl. **18**, 701–722 (1989)
68. Wong, P.J.Y., Agarwal, R.P.: Quintic spline solutions of Fredholm integral equations of the second kind. Int. J. Comput. Math. **33**, 237–249 (1990)
69. Wong, P.J.Y., Agarwal, R.P.: Sharp error bounds for the derivatives of Lidstone-spline interpolation. Comput. Math. Appl. **28**, 23–53 (1994)
70. Wong, P.J.Y., Agarwal, R.P.: Error inequalities for discrete Hermite and spline interpolation. In: Milovanovic, G.V. (ed.) Recent Progress in Inequalities, pp. 397–422. Kluwer, Dordrecht (1998)

Contributed Papers

Persistence of a Discrete-Time Predator-Prey Model with Stage-Structure in the Predator

AZMY S. ACKLEH, Md. Istiaq Hossain, Amy Veprauskas, and Aijun Zhang

Abstract We propose and investigate a discrete-time predator-prey model with a structured predator population. We describe the predator population using two stages, juveniles and adults, and assume that only the adult stage consumes the prey species. For this model, we discuss conditions for the existence and global stability of the extinction and predator-free equilibria as well as conditions for the existence and uniqueness of an interior equilibrium. We show that when the predator-free equilibrium destabilizes, the interior equilibrium is stable in a neighborhood of the bifurcation. We also find the conditions for the persistence of both prey and predator populations. Finally, we use numerical simulations to demonstrate various dynamical scenarios. We find that introducing stage-structure into the predator population allows for complicated dynamics that are not possible when the predator is unstructured.

Keywords Discrete-time predator-prey models · Stability · Persistence · Global attractors · Stage-structure

1 Introduction

Predator-prey models play an important role in understanding the possible ecological outcomes of interacting species. The earliest predator-prey models were introduced and investigated independently by both Lotka and Volterra [28, 29, 40]. Ample extensions of these models, in both continuous-time and discrete-time, have since been developed to describe different ecological predator-prey scenarios. These include various types of functional responses [21, 24, 26, 41], developmental delays and stage structure [14, 15, 25], the co-evolution between predator and prey [1, 2, 42], and more complicated predator-prey interactions such as intraguild predation [7, 23, 36]. In certain situations, such as when the species have non-overlapping generations,

A. S. ACKLEH (✉) · Md. I. Hossain · A. Veprauskas · A. Zhang
Department of Mathematics, University of Louisiana at Lafayette, Lafayette, LA 70504, USA
e-mail: azmy.ackleh@louisiana.edu

S. Baigent et al. (eds.), *Progress on Difference Equations and Discrete Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 341,
https://doi.org/10.1007/978-3-030-60107-2_6

it has been suggested that discrete-time models governed by difference equations may be more appropriate than continuous-time modeling approaches [5, 20, 35].

To model a biological population mathematically, species can often be better described using stage-structure rather than treating all individuals as physiologically identical [6, 12, 13, 39]. In fact, various studies have found that unstructured population models are less efficient at predicting population abundance [8, 31, 34]. When modeling interacting species, stage structure can become particularly important when only specific developmental stages interact. For instance, in predator-prey interactions, it may be the case that juveniles and adults of the predator population consume different prey species. Such a situation may occur if different developmental stages inhabit different environments or size differences shift diet preferences due to changes in foraging ability. These ontogenetic niche shifts are prevalent in many aquatic organisms which undergo dramatic changes in body size throughout their lifetimes [34]. For example, newborn Eurasian perch feed on zooplankton, shift to benthic resources at intermediate sizes, then feed on fish at larger sizes [22]. Despite the documented importance of structure in species interactions, when compared to the extensive study of single species models with stage-structure, there are relatively few models for predator-prey interactions with stage structure [25, 33]. In large part, this is likely due to the mathematical intractability of such models [25].

The purpose of this work is to extend the discrete-time predator-prey model developed in [3] to include a stage-structured predator population in which the predator is classified according to two developmental stages: juveniles and adults. We model the prey population as unstructured, that is individuals are described by the same average biological vital rates. We assume that this population grows according to a monotonic nonlinearity, such as the Beverton-Holt function, in the absence of predators. For the predator population, we assume that only the adult predators are capable of attacking and consuming the prey population with prey consumption regulating predator reproduction. The transition probability of the juvenile predators, and the survival probabilities of both predator stages are assumed to be time and density independent. We thoroughly investigate the various dynamical behaviors of this discrete-time predator-prey model such as the existence and uniqueness of the extinction, predator-free, and interior equilibria as well as the local and global stability of the equilibria and the persistence of the system.

This paper is organized as follows. In Sect. 2, we introduce the discrete-time predator-prey model with stage-structure in the predator. We derive conditions for the existence of two boundary equilibria, namely, the extinction and predator-free equilibria. We also prove the global asymptotic stability of these two equilibria. Next, we derive conditions for the existence and uniqueness of the interior equilibrium and show that, when the predator-free equilibrium destabilizes, this equilibrium is stable in a neighborhood of the bifurcation. We also establish conditions for the persistence of both the prey and predator populations. In Sect. 3, we provide numerical examples showing various dynamical scenarios in support of the theoretical results. These numerical simulations also show the existence of rich dynamics that are not observed when the predator is unstructured [4]. Finally, we provide some concluding remarks in Sect. 4.

2 The Predator-Prey Model

In this section, we introduce the discrete-time predator-prey model for a single prey and a single predator population with stage-structure in the predator population. We consider two stages for the predator, namely juveniles and adults. Let n denote the prey density and p_1 and p_2 the densities of the juvenile and adult predator stages, respectively. We assume that after each time step a fraction $\gamma, 0 < \gamma \leq 1$, of surviving juvenile predators become adult predators. We also assume that only adult predators attack and consume the prey population. Specifically, the model is given by

$$\begin{cases} n(t+1) = \phi(n(t))\,(1 - f(p_2(t))p_2(t))\,n(t), \\ p_1(t+1) = (1-\gamma)s_1 p_1(t) + b(n(t))n(t)f(p_2(t))p_2(t), \\ p_2(t+1) = \gamma s_1 p_1(t) + s_2 p_2(t), \end{cases} \tag{1}$$

where the constants $0 < s_1, s_2 < 1$ represent the density-independent probabilities of a juvenile or adult predator, respectively, surviving a unit of time.

The nonlinearities ϕ, b, and f are assumed to be smooth functions in the set X defined by:

$$X := \{v \in C^1[0,\infty) |\, v'(x) < 0,\ (v(x)x)' > 0,\ \lim_{x\to\infty} v(x) = 0,\ \text{and}\ \lim_{x\to\infty} v(x)x < \infty\}.$$

This set includes functions of the Beverton-Holt type [9]. The quantity $f(p_2)$ is defined to be the fraction of prey consumed by a single adult predator individual per unit time when p_2 predators are present, $0 \leq f(p_2) \leq 1$. Thus, $0 \leq f(p_2)p_2 \leq 1$ gives the fraction of prey consumed by p_2 adult predators and $1 - f(p_2)p_2$ gives the fraction of prey that escape predation per unit time. This quantity modifies the predator-free prey growth given by $\phi(n)$. The reproductive output of an adult predator individual is assumed to be determined by the amount of prey consumed, where the conversion factor $b(n)n$ gives the number of new births that would result from the consumption of the entire prey population n. This term is defined so that predator reproduction is limited by biological factors and cannot grow unbounded as prey density gets large. The assumption that predator reproduction is regulated by prey consumption is appropriate for many species, such as snakes and marine birds, where reproductive output is determined by energy availability [10, 37].

2.1 Equilibria and Stability

The equilibrium equations of model (1) are given by the following system of equations:

$$\begin{cases} n = \phi(n)\,(1 - f(p_2)p_2)\,n, \\ p_1 = (1-\gamma)s_1 p_1 + b(n)nf(p_2)p_2, \\ p_2 = \gamma s_1 p_1 + s_2 p_2. \end{cases} \tag{2}$$

Solving the above system, we find that model (1) has three equilibria, namely the extinction equilibrium where both species die out, the predator-free equilibrium where the prey survives but predator goes extinct, and an interior equilibrium where both prey and predator densities are positive. In this section, we discuss the existence and stability of these equilibria. To determine the local asymptotic stability of the equilibria, we find when the eigenvalues of Jacobian matrix of system (1) evaluated at each of the equilibria have magnitude less than one. This Jacobian matrix is given by

$$J(n, p_1, p_2) = \begin{pmatrix} (\phi(n)n)'\,(1 - f(p_2)p_2) & 0 & -(\phi(n)n)\,(f(p_2)p_2)' \\ (b(n)n)'\,(f(p_2)p_2) & (1-\gamma)s_1 & (b(n)n)\,(f(p_2)p_2)' \\ 0 & \gamma s_1 & s_2 \end{pmatrix}. \tag{3}$$

2.1.1 Existence and Stability of Boundary Equilibria

In this section, we discuss the existence and stability of the two boundary equilibria of model (1), which are the extinction equilibrium and the predator-free equilibrium. First, in Theorem 1, we show that the extinction equilibrium is globally asymptotically stable when the inherent growth rate of the prey $\phi(0)$ is less than one.

Theorem 1 *The extinction equilibrium* $(0, 0, 0)$ *of model (1) is globally asymptotically stable if* $\phi(0) < 1$ *and unstable if* $\phi(0) > 1$.

Proof The Jacobian matrix (3) evaluated at the extinction equilibrium $(0, 0, 0)$ is given by

$$J(0, 0, 0) = \begin{pmatrix} \phi(0) & 0 & 0 \\ 0 & (1-\gamma)s_1 & 0 \\ 0 & \gamma s_1 & s_2 \end{pmatrix}.$$

The eigenvalues of this matrix are $\lambda_1 = \phi(0)$, $\lambda_2 = (1-\gamma)s_1$, and $\lambda_3 = s_2$. Since $0 < \gamma \leq 1$ and $0 < s_1, s_2 < 1$, $|\lambda_i| < 1$ for $i = 2, 3$. Thus the extinction equilibrium $(0, 0, 0)$ is locally asymptotically stable if $\phi(0) < 1$ and unstable if $\phi(0) > 1$. Since the function $\phi(n)$ is a function satisfying all the conditions in the set X, there exists a positive constant D such that $\phi(n)n \leq D$ for all $n \geq 0$. Since

$$n(t+1) = \phi(n(t))n(t)(1 - f(p_2(t))p_2(t)) \leq \phi(n(t))n(t) \leq D \quad \text{for all } n \geq 0,$$

we have that

$$\begin{aligned} p_1(t+2) &= (1-\gamma)s_1 p_1(t+1) + b(n(t+1))n(t+1)f(p_2(t+1))p_2(t+1), \\ &\leq (1-\gamma)s_1 p_1(t+1) + b(D)D. \end{aligned}$$

Consider the difference equation $q_1(t+1) = (1-\gamma)s_1 p_1(t) + b(D)D$, where $q_1(0) = p_1(1)$. Then we have that $\lim_{t\to\infty} q_1(t) = \dfrac{b(D)D}{1-(1-\gamma)s_1}$. Hence, $\limsup_{t\to\infty} p_1(t) \le \dfrac{b(D)D}{1-(1-\gamma)s_1}$. As a result,

$$p_2(t+3) \le \gamma s_1 \left(\frac{b(D)D}{1-(1-\gamma)s_1}\right) + s_2 p_2(t+2).$$

Letting $q_2(t+1) = \gamma s_1 \left(\frac{b(D)D}{1-(1-\gamma)s_1}\right) + s_2 q_2(t)$ with $q_2(0) = p_2(2)$, it follows that $\lim_{t\to\infty} q_2(t) = \left(\dfrac{\gamma s_1}{1-s_2}\right)\left(\dfrac{b(D)D}{1-(1-\gamma)s_1}\right)$. Hence, we have $\limsup_{t\to\infty} p_2(t) \le \left(\dfrac{\gamma s_1}{1-s_2}\right)\left(\dfrac{b(D)D}{1-(1-\gamma)s_1}\right)$. Therefore, the solutions of system (1) remain non-negative and bounded for all forward time. Since $\phi \in X$, $\phi(n) \le \phi(0)$ for any $n \ge 0$, and thus $n(t+1) \le \phi(n(t))n(t) \le \phi(0)n(t)$ for $t \ge 0$. From this we have $\lim_{t\to\infty} n(t) = 0$ whenever $\phi(0) < 1$. As a result, $\lim_{t\to\infty} p_1(t) = \lim_{t\to\infty} p_2(t) = 0$, and hence for $\phi(0) < 1$, all the solutions of (1) will converge to the extinction equilibrium $(0, 0, 0)$. Thus the extinction equilibrium $(0, 0, 0)$ is globally asymptotically stable when $\phi(0) < 1$, and unstable if $\phi(0) > 1$.

Next, in Theorem 2, we show that a predator-free equilibrium $(\bar{n}, 0, 0)$ exists when $\phi(0) > 1$. This predator-free equilibrium is globally asymptotically stable when the invasion net reproductive number $R_{\bar{n}}$ is less than one. This quantity is the inherent net reproductive number of the predator when the prey is at its predator-free state. Hence, when $R_{\bar{n}} < 1$, the predator is unable to invade the system.

To calculate the invasion net reproductive number $R_{\bar{n}}$, we first consider the predator subsystem given by

$$P(t+1) = A(n(t), p_1(t), p_2(t))P(t), \tag{4}$$

where

$$A(n, p_1, p_2) = \begin{pmatrix} (1-\gamma)s_1 & b(n)nf(p_2) \\ \gamma s_1 & s_2 \end{pmatrix} \text{ and } P(t) = \begin{pmatrix} p_1(t) \\ p_2(t) \end{pmatrix}.$$

Suppose that the prey is at its predator-free equilibrium $\bar{n}$. Then, the inherent projection matrix for the predator population when the prey is at its predator-free state is given by

$$A(\bar{n}, 0, 0) = \begin{pmatrix} (1-\gamma)s_1 & b(\bar{n})\bar{n}f(0) \\ \gamma s_1 & s_2 \end{pmatrix}.$$

This matrix $A(\bar{n}, 0, 0)$ can now be decomposed as $A = T + F$, where the transition matrix T and the fertility matrix F are given by

$$T = \begin{pmatrix} (1-\gamma)s_1 & 0 \\ \gamma s_1 & s_2 \end{pmatrix} \text{ and } F = \begin{pmatrix} 0 & b(\bar{n})\bar{n}f(0) \\ 0 & 0 \end{pmatrix}.$$

Then the next generation matrix N is given by

$$N = (I_2 - T)^{-1} = \begin{pmatrix} \frac{1}{1-(1-\gamma)s_1} & 0 \\ \frac{\gamma s_1}{(1-s_2)(1-(1-\gamma)s_1)} & \frac{1}{1-s_2} \end{pmatrix},$$

where I_2 is the identity matrix of size 2×2 and $N(i, j)$ is the expected number of visits over a lifetime to stage i for an individual starting in stage j [11]. The invasion net reproductive number for the predator, $R_{\bar{n}}$ when the prey is at its equilibrium state is given by the dominant eigenvalue of the matrix FN [19]. From this we obtain

$$R_{\bar{n}} = \frac{\gamma s_1 b(\bar{n})\bar{n}f(0)}{(1-s_2)(1-(1-\gamma)s_1)}. \tag{5}$$

Next, we show that a predator-free equilibrium exists when the inherent growth rate of the prey is greater than one and is globally asymptotically stable when $R_{\bar{n}} < 1$.

Theorem 2 *The predator-free equilibrium* $(\bar{n}, 0, 0)$ *of model* (1)*, where* $\bar{n} := \phi^{-1}(1)$*, exists when* $\phi(0) > 1$*. Moreover, if* $R_{\bar{n}} < 1$*, where* $R_{\bar{n}}$ *is defined as in* (5)*, then* $(\bar{n}, 0, 0)$ *is globally asymptotically stable. If* $R_{\bar{n}} > 1$*, then* $(\bar{n}, 0, 0)$ *is unstable.*

Proof First, it is easy to see that the predator-free equilibrium $(\bar{n}, 0, 0)$, where $\bar{n} := \phi^{-1}(1)$ exists when $\phi(0) > 1$. Applying the Jury conditions to the Jacobian matrix evaluated at $(\bar{n}, 0, 0)$, we find that this equilibrium is locally asymptotically stable when

$$s_2(1-(1-\gamma)s_1) + (1-\gamma)s_1 + \gamma s_1 b(\bar{n})\bar{n}f(0) < 1.$$

This is equivalent to $R_{\bar{n}} < 1$. It remains to show that $(\bar{n}, 0, 0)$ is globally asymptotically stable when $R_{\bar{n}} < 1$.

Since $\phi \in X, n(t+1) \le \phi(n(t))n(t) \le D$ for some $D > 0$ and $t \ge 0$. Moreover,

$$\limsup_{t\to\infty} n(t+1) \le \limsup_{t\to\infty} (\phi(n(t))n(t)) \le \phi(\limsup_{t\to\infty} n(t)) \limsup_{t\to\infty} n(t).$$

Thus $\phi(\limsup_{t\to\infty} n(t)) \ge 1$, which implies that $\limsup_{t\to\infty} n(t) \le \bar{n}$. Therefore for any $\epsilon > 0$, there exists $t_0 > 0$ such that for $t \ge t_0$, $n(t) \le \bar{n} + \epsilon$ and

$$s_2(1-(1-\gamma)s_1) + (1-\gamma)s_1 + \gamma s_1 b(\bar{n}+\epsilon)(\bar{n}+\epsilon)f(0) < 1. \tag{6}$$

On the other hand, with $f(0) \ge f(p_2)$ we have

$$\begin{aligned}
p_2(t+2) &= \gamma s_1 p_1(t+1) + s_2 p_2(t+1), \\
&\le \gamma s_1 \left((1-\gamma)s_1 p_1(t) + b(n)nf(0)p_2(t)\right) + s_2 p_2(t+1), \\
&= \gamma s_1(1-\gamma)s_1 \left(\frac{p_2(t+1) - s_2 p_2(t)}{\gamma s_1}\right) + \gamma s_1 b(n(t))n(t) f(0) p_2(t) + s_2 p_2(t+1), \\
&= (1-\gamma)s_1 p_2(t+1) - (1-\gamma)s_1 s_2 p_2(t) + \gamma s_1 b(n(t))n(t) f(0) p_2(t) + s_2 p_2(t+1), \\
&\le ((1-\gamma)s_1 + s_2) p_2(t+1) - ((1-\gamma)s_1 s_2 - \gamma s_1 b(\bar{n}+\epsilon)(\bar{n}+\epsilon) f(0)) p_2(t).
\end{aligned}$$

Thus, letting $c_1 = (1-\gamma)s_1 + s_2$ and $c_2 = (1-\gamma)s_1 s_2 - \gamma s_1 b(\bar{n}+\epsilon)(\bar{n}+\epsilon) f(0)$, we have

$$p_2(t+2) \le c_1 p_2(t+1) - c_2 p_2(t). \tag{7}$$

We claim that $\lim_{t\to\infty} p_2(t) = 0$. To this end, we solve $x^2 - c_1 x + c_2 = 0$ which has two roots $d_1 = \frac{c_1 - \sqrt{c_1^2 - 4c_2}}{2}$ and $d_2 = \frac{c_1 + \sqrt{c_1^2 - 4c_2}}{2}$. By (6),

$$c_1 - c_2 = ((1-\gamma)s_1 + s_2) - ((1-\gamma)s_1 s_2 - \gamma s_1 b(\bar{n}+\epsilon)(\bar{n}+\epsilon) f(0)) < 1.$$

As a result, we have $0 < d_i < 1$ for $i = 1, 2$. We can rewrite (7) as

$$p_2(t+2) - d_1 p_2(t+1) \le d_2 (p_2(t+1) - d_1 p_2(t)). \tag{8}$$

If $p_2(t+1) - d_1 p_2(t) \le 0$ for $t = \tau$, then by (8), $p_2(t+1) - d_1 p_2(t) \le 0$ for all $t > \tau$. Under this case, $d_1 < 1$ implies that $\lim_{t\to\infty} p_2(t) = 0$. Otherwise, suppose that $p_2(t+1) - d_1 p_2(t) > 0$ for all t. By (8), $d_2 < 1$ implies that $\lim_{t\to\infty} (p_2(t+1) - d_1 p_2(t)) = 0$ and thus $\lim_{t\to\infty} p_2(t) = 0$ since $d_1 < 1$. As a result of $\lim_{t\to\infty} p_2(t) = 0$, with the third equation in the model, we have $\lim_{t\to\infty} p_1(t) = 0$. Finally, with the first equation in the model, we must have $\lim_{t\to\infty} n(t) = \bar{n}$ if $\phi(0) > 1$.

2.1.2 Existence and Stability of the Interior Equilibrium

Next, we show that a unique interior equilibrium of (1) exists if and only if the invasion net reproductive number of the predator $R_{\bar{n}}$ is greater than one.

Theorem 3 *A unique interior equilibrium* (n^*, p_1^*, p_2^*) *of model (1) exists if and only if* $\phi(0) > 1$ *and* $R_{\bar{n}} > 1$, *where* $R_{\bar{n}}$ *is defined in* (5).

Proof The equilibrium equations of (1) can be written as

$$\begin{aligned}
\phi(n)(1 - f(p_2)p_2) &= 1, \\
(1-\gamma)s_1 p_1 + b(n)nf(p_2)p_2 &= p_1, \\
\gamma s_1 p_1 + s_2 p_2 &= p_2.
\end{aligned} \tag{9}$$

From the third equation of (9), we have $p_1 = \left(\frac{1-s_2}{\gamma s_1}\right) p_2$. Using this relation in the second equation above and simplifying, we obtain

$$\frac{(1-\gamma)(1-s_2)}{\gamma} + b(n)nf(p_2) = \frac{1-s_2}{\gamma s_1}, \tag{10}$$

which is equivalent to

$$\tilde{s} + b(n)nf(p_2) = 1, \tag{11}$$

with $\tilde{s} := 1 + \frac{(1-\gamma)(1-s_2)}{\gamma} - \frac{1-s_2}{\gamma s_1}$. This final equation together with the first equilibrium equation of (9) reduces the equilibrium equations to

$$\begin{aligned} \phi(n)(1 - f(p_2)p_2) &= 1, \\ \tilde{s} + b(n)nf(p_2) &= 1. \end{aligned} \tag{12}$$

This system of equilibrium equations is the same as the equilibrium equations for the non-evolutionary model discussed in [3]. Hence, this system has a unique equilibrium (n, p_2) (and hence (n, p_1, p_2)) if and only if $\phi(0) > 1$ and $\tilde{s} + b(\bar{n})\bar{n}f(0) > 1$. This latter inequality is equivalent to $R_{\bar{n}} > 1$. Thus a unique interior equilibrium of model (1) exists if and only if $\phi(0) > 1$ and $R_{\bar{n}} > 1$.

Remark 1 We note that, in the equilibrium equations for the non-evolutionary model studied in [3], we have $0 < \tilde{s} < 1$. Meanwhile, for model (1) we have $\tilde{s} < 1$ but not necessarily $\tilde{s} > 0$, as is the case in Example 1 shown below. However, all steps in the proof of Theorem 2.2(a) from [3] continue to hold when $\tilde{s} < 0$.

In Theorem 3, we showed that the predator-free equilibrium destabilizes as $R_{\bar{n}}$ increases past one. We further showed that a unique interior equilibrium exists for $R_{\bar{n}} > 1$. In Theorem 4, we apply perturbation arguments to establish the local stability of the interior equilibrium in a neighborhood of $R_{\bar{n}} \gtrapprox 1$. To do this, we introduce the bifurcation parameter $b_0 := b(0)$. In terms of this parameter, $R_{\bar{n}} \gtrapprox 1$ is equivalent to $b_0 \gtrapprox \hat{b}$ for

$$\hat{b} := \frac{(1-(1-\gamma)s_1)(1-s_2)}{\gamma s_1 \beta(\bar{n})\bar{n} f(0)}, \quad \beta(n) := b(n)/b_0. \tag{13}$$

The arguments used to prove Theorem 4 are analogous to those applied in [17, 18].

Theorem 4 *Define $b_0 := b(0)$ and assume $\phi(0) > 1$. There exists a branch of positive equilibria (n, p_1, p_2) of model (1) bifurcating from the predator-free equilibrium $(\bar{n}, 0, 0)$ at $b_0 = \hat{b}$ where $\hat{b}$ is defined in (13). The bifurcating equilibria are locally asymptotically stable for $b_0 \gtrapprox \hat{b}$.*

Proof Consider the equilibria equations given by (9). Solve the second equation in (9) for p_1 and substitute this solution into the third equation. We now have that an interior equilibrium must satisfy

$$g_1(n, p_1, p_2, b_0) = g_2(n, p_1, p_2, b_0) = g_3(n, p_1, p_2, b_0) = 0,$$

where

$$\begin{aligned} g_1(n, p_1, p_2, b_0) :=& 1 - \phi(n)(1 - f(p_2)p_2), \\ g_2(n, p_1, p_2, b_0) :=& p_1 - (1-\gamma)s_1 p_1 - b_0\beta(n)nf(p_2)p_2, \\ g_3(n, p_1, p_2, b_0) :=& 1 - \frac{\gamma s_1 b_0 \beta(n) n f(p_2)}{1-(1-\gamma)s_1} - s_2, \end{aligned} \tag{14}$$

and we have explicitly denoted dependence on the bifurcation parameter b_0 by defining $\beta(n) := b(n)/b_0$. Since $(\bar{n}, 0, 0)$ is a solution to (14) only for $b_0 = \hat{b}$, by the Implicit Function Theorem there exists a branch of equilibria bifurcating from $(\bar{n}, 0, 0)$ at $b_0 = \hat{b}$ provided that the determinant of the Jacobian of (14) evaluated at the known solution $(\bar{n}, 0, 0, \hat{b})$ is non-zero. This determinant is given by

$$\kappa := \frac{(1-(1-\gamma)s_1)(1-s_2)(f(0)^2(\beta(\bar{n}) + \bar{n}\beta'(\bar{n})) + \bar{n}\beta(\bar{n})f'(0)\phi'(\bar{n}))}{\bar{n}\beta(\bar{n})f(0)} > 0,$$

which is nonzero since $\phi, b, f \in X$. It follows that a branch of equilibria of the form $(n(\epsilon), p_1(\epsilon), p_2(\epsilon), b_0(\epsilon))$ with $\epsilon \approx 0, b_0 = \hat{b}(1+\epsilon)$, and $(n(0), p_1(0), p_2(0), b_0(0)) = (\bar{n}, 0, 0, \hat{b})$ bifurcates from the predator-free equilibrium at $b_0 = \hat{b}$.

To determine the parametrization of the bifurcating equilibria, we differentiate (14) with respect to ϵ and evaluate the derivatives at $\epsilon = 0$. We arrive at the following parameterizations

$$\begin{aligned} n(\epsilon) &= \bar{n} - \frac{f(0)(1-(1-\gamma)s_1)(1-s_2)}{\kappa}\epsilon + \mathcal{O}(\epsilon^2), \\ p_1(\epsilon) &= -\frac{(1-(1-\gamma)s_1)\phi'(\bar{n})}{\gamma s_1 \kappa}\epsilon + \mathcal{O}(\epsilon^2), \\ p_2(\epsilon) &= -\frac{(1-(1-\gamma)s_1)(1-s_2)\phi'(\bar{n})}{\kappa}\epsilon + \mathcal{O}(\epsilon^2). \end{aligned} \tag{15}$$

Since $\phi'(\bar{n}) < 0$, it follows that an interior equilibrium exists for $\epsilon \gtrapprox 0$, or equivalently, $b_0 \gtrapprox \hat{b}$. Notice that the parameterization for n says that, as the predator population becomes positive, the prey population decreases, as is to be expected.

Finally, to determine the stability of the branch of interior equilibria, we parameterize the eigenvalues in terms of ϵ,

$$\lambda_i(\epsilon) = \lambda_i(0) + \lambda_i'(0)\epsilon + \mathcal{O}(\epsilon^2).$$

The Jacobian matrix (3) evaluated at the positive equilibrium (15) has the expansion $J(n(\epsilon), p_1(\epsilon), p_2(\epsilon), b_0(\epsilon)) = J(0) + J'(0)\epsilon + \mathcal{O}(\epsilon^2)$ where $J(0)$ is the Jacobian matrix (3) evaluated at $(\bar{n}, 0, 0, \hat{b})$,

$$J(0) = \begin{pmatrix} \bar{n}\phi'(\bar{n}) + \phi(\bar{n}) & 0 & -\bar{n}\phi(\bar{n}) f(0) \\ 0 & (1-\gamma)s_1 & \hat{b}\bar{n}\beta(\bar{n}) f(0) \\ 0 & \gamma s_1 & s_2 \end{pmatrix},$$

and matrix $J'(0)$ may be found by substituting the positive equilibrium $(n(\epsilon), p_1(\epsilon), p_2(\epsilon), b_0(\epsilon))$ into (3), differentiating this Jacobian matrix with respect to ϵ, and evaluating the resulting matrix at $\epsilon = 0$. Matrix $J(0)$ has eigenvalues $\lambda_1(0) = 1$, $\lambda_2(0) = -1 + (1-\gamma)s_1 + s_2$, and $\lambda_3(0) = 1 + \bar{n}\phi'(\bar{n})$. Meanwhile, $\lambda_i'(0)$ may be found by first linearizing the eigenvalue equation $J(\epsilon)v(\epsilon) = \lambda(\epsilon)v(\epsilon)$ around $\epsilon = 0$ to obtain $(J(0) - \lambda(0)I)v'(0) = (\lambda'(0)I - J'(0))v(0)$. Next we may either apply inner product arguments or the Fredholm Alternative to show that this equation has a solution if and only if $w(0)(\lambda'(0)I - J'(0))v(0) = 0$ where $w(0)$ is the left eigenvector of $J(0)$ [18]. Thus we have $\lambda_i'(0) = \frac{w(0)J'(0)v(0)}{w(0)v(0)}$. From the eigenvalues of $J(0)$, it is clear that the stability of the interior equilibrium is determined by $\lambda_1(\epsilon)$, where $v(0) = \left(\frac{f(0)}{\phi'(\bar{n})}, \frac{1-s_2}{\gamma s_1}, 1\right)^{\mathsf{T}}$ and $w(0) = \left(0, \frac{\gamma s_1}{1-(1-\gamma)s_1}, 1\right)$ are the right and left eigenvectors corresponding to $\lambda_1(0)$. All together, we arrive at $\lambda_1'(0) = -\frac{(1-(1-\gamma)s_1)(1-s_2)}{2-(1-\gamma)s_1-s_2} < 0$. Thus, the interior equilibrium is locally asymptotically stable for $\epsilon \lessapprox 0$, or equivalently, $b_0 \gtrapprox \hat{b}$.

Remark 2 In Theorem 4, we have shown that the branch of interior equilibria bifurcates forward or supercritically (meaning the equilibria are positive for $b_0 \gtrapprox \hat{b}$) and is stable in a neighborhood of the bifurcation. This is a similar dynamic scenario as is described by the Fundamental Bifurcation Theorem [16]. However, that theorem applies to models with primitive inherent projection matrices whereas the projection matrix for model (1) is reducible.

Remark 3 After submission of this paper, it came to the authors' attention that a more general result, developed in [32], can be applied to establish Theorem 4. While both the proof of Theorem 4 and result from [32] use a similar Lyapunov-Schmidt expansion technique, which relies on an application of the Implicit Function Theorem and the Fredholm Alternative, they apply different forms for the expansion of the bifurcation parameter.

Next, we provide an example that shows the conditions for the existence and stability of the equilibria for a specific set of nonlinearities satisfying the conditions in X.

Example 1 We assume the following set of nonlinearities:

$$\phi(n) = \frac{r_0}{1+mn}, \quad b(n) = \frac{b_0}{1+\delta n}, \text{ and } f(p_2) = \frac{c}{1+cp_2}. \tag{16}$$

Note that $\{\phi, b, f\} \subset X$ with $\phi(0) = r_0$, $f(0) = c$, $\bar{n} = \phi^{-1}(1) = \frac{r_0-1}{m}$, and

$$R_{\bar{n}} = \frac{\gamma s_1 c b_0 (r_0 - 1)}{(m + \delta(r_0 - 1))(1 - s_2)(1 - (1-\gamma)s_1)}.$$

(i) The extinction equilibrium $(0, 0, 0)$ is globally asymptotically stable if $r_0 < 1$ and unstable otherwise.
(ii) The predator-free equilibrium $(\bar{n}, 0, 0)$ exists if $r_0 > 1$ and is globally asymptotically stable if $R_{\bar{n}} < 1$.
(iii) For $\tilde{s} := 1 + \frac{(1-\gamma)(1-s_2)}{\gamma} - \frac{1-s_2}{\gamma s_1}$ and $k := \frac{r_0(1-\tilde{s})}{b_0 c} > 0$, the interior equilibrium (n^*, p_1^*, p_2^*) with

$$n^* := \frac{(k\delta - 1) + \sqrt{(k\delta - 1)^2 + 4km}}{2m},$$

$$p_1^* := \left(\frac{1 - s_2}{\gamma s_1}\right)\left(\frac{r_0 - \left(1 + \frac{1}{2}\left((k\delta - 1) + \sqrt{(k\delta - 1)^2 + 4km}\right)\right)}{c + \frac{c}{2}\left((k\delta - 1) + \sqrt{(k\delta - 1)^2 + 4km}\right)}\right),$$

$$p_2^* := \frac{r_0 - \left(1 + \frac{1}{2}\left((k\delta - 1) + \sqrt{(k\delta - 1)^2 + 4km}\right)\right)}{c + \frac{c}{2}\left((k\delta - 1) + \sqrt{(k\delta - 1)^2 + 4km}\right)},$$

exists when $R_{\bar{n}} > 1$.

2.2 Persistence

In this section, we investigate the persistence of the prey and predator populations. Consider the difference equation system

$$x(t + 1) = F(x(t)), \tag{17}$$

where $x(t) = (x_1(t), x_2(t), ..., x_n(t))^T$, $F = (f_1, f_2, ..., f_n)^T$ is a smooth function from $\mathbb{R}^n$ to $\mathbb{R}^n$, and $f_i = f_i(x(t))$, $i = 1, 2, ..., n$ for a postive integer n. Define

$$Z := \{x \in \mathbb{R}^n | x \geq 0\}, \quad Z^+ := \{x \in \mathbb{R}^n | x_i > 0, \forall i = 1, 2, ..., n\}, \tag{18}$$

and let $\rho : Z \to [0, \infty)$ be a continuous function. Suppose $M_0 := \{z \in Z : \rho(z) > 0\}$ and $\partial M_0 := \{z \in Z : \rho(z) = 0\}$. We assume that the system is invariant in M_0. Then, the definition of ρ-uniform persistence is as follows.

Definition 1 *(ρ-uniform persistence* [30]*)* System (17) is said to be ρ-uniformly persistent if there exists $\epsilon > 0$ such that $\liminf_{n\to\infty} \rho(F^n(x)) \geq \epsilon, \forall x \in M_0$; system (17) is weakly ρ-uniformly persistent if there exists $\epsilon > 0$ such that $\limsup_{n\to\infty} \rho(F^n(x)) \geq \epsilon, \forall x \in M_0$.

To establish the persistence of system (1), we assume the system satisfies the following hypotheses:

$$\textbf{(H1)} \quad \phi(0) > 1 \text{ and } R_{\bar{n}} > 1 \text{ with } \bar{n} = \phi^{-1}(1).$$

Lemma 1 *Assume (H1). There exist positive constants N, P_1 and P_2 such that $\mathcal{A} = \{(n, p_1, p_2) | 0 \le n \le N \le \bar{n}, 0 \le p_1 \le P_1, 0 \le p_2 \le P_2\}$ is an attracting set under system (1).*

Proof This lemma follows from the arguments provided in the proof of Theorem 1 with $N = D$, $P_1 = \frac{b(D)D}{1-(1-\gamma)s_1}$, and $P_2 = \left(\frac{\gamma s_1}{1-s_2}\right)\left(\frac{b(D)D}{1-(1-\gamma)s_1}\right)$, where D is such that $\phi(n)n \le D$.

Next, in Theorem 5 we show that model (1) is persistent. For the predator population, we show that all stages are bounded away from zero, referred to as c-persistence in [27]. This condition provides the coexistence of the predator stages by ensuring that no orbits converge to the boundary of the positive cone.

Theorem 5 *Assume (H1). Then model (1) is persistent, that is there exists an $\epsilon > 0$ such that $\min\{\liminf_{t\to\infty} n(t), \liminf_{t\to\infty} p_1(t), \liminf_{t\to\infty} p_2(t)\} > \epsilon$ for any initial condition in Z^+.*

Proof We prove the theorem by showing the persistence of the prey, the juvenile predator, and the adult predator populations with the following four steps.

- Step 1: Show the prey is persistent (i.e. $\liminf_{t\to\infty} n(t) > \epsilon$ for all $z_0 \in Z^+$).
  Let $N_0 = \{(n, p_1, p_2) | n = 0\}$, on which the subsystem is given by

$$\begin{cases} p_1(t+1) = (1-\gamma)s_1 p_1(t), \\ p_2(t+1) = s_2 p_2(t). \end{cases} \tag{19}$$

  If the extinction equilibrium of subsystem (19) is globally stable, then the omega limit set of N_0 is $\Omega(N_0) = \{(0, 0, 0)\} =: \{\mathcal{P}_1\}$. Let $A(\mathcal{P}) := \phi(n)(1 - f(p_2)p_2)$ with $\mathcal{P} \in \Omega(N_0)$. Then if $A(\mathcal{P}_1) = \phi(0) > 1$, $A(\mathcal{P}_1)$ is primitive and the spectral radius of $A(\mathcal{P}_1)$ is greater than one. Then by Corollary 1 in [38], N_0 is a uniformly weak repeller. Finally, by Theorem 2.3 in [38], we have that the system is $n-$persistent i.e. $\liminf_{t\to\infty} n(t) > \epsilon$.
- Step 2: We prove that the system is $\rho-$ persistent with $\rho = p_1 + p_2$, that is, there exists $\epsilon > 0$ such that for all $z_0 \in Z^+$

$$\liminf_{t\to\infty} (p_1 + p_2) > \epsilon. \tag{20}$$

  Following [38], define $f : \mathbb{R}^1_1 \times \mathbb{R}^2_+ \to \mathbb{R}^1_1$ and $g : \mathbb{R}^1_1 \times \mathbb{R}^2_+ \to \mathbb{R}^2_+$ such that for all $z \in \mathbb{R}^1_1 \times \mathbb{R}^2_+$, $F(z) = (f(z), g(z))$. Consider the following dynamical system:

$$z(t+1) = F(z(t)), \text{ for all } z \in \mathbb{R}^1_1 \times \mathbb{R}^2_+. \tag{21}$$

  Equation (21) can be written as

$$\begin{aligned} x(t+1) &= f(z(t)), \\ y(t+1) &= A(z(t))y(t), \end{aligned} \tag{22}$$

where, $N_1 = \{z = (x, y) \in \mathbb{R}^1_+ \times \mathbb{R}^2_+ | y = 0\}$, and $A(z)$ is a continuous matrix function satisfying $A(x, 0) \geq 0$. Let $x = n$ and $y = (p_1, p_2)$. The set N_1 represents the positively invariant predator-free sub-space, in which the dynamics are given by

$$n(t+1) = \phi(n(t))n(t). \tag{23}$$

Let $n(t+1) = n(t) = \bar{n}$ be the predator-free equilibrium for the prey, that is $\bar{n} = \phi^{-1}(1)$. Note that $\bar{n}$ exists when $\phi(0) > 1$. Model (1) can now be put in the form (21) with

$$A(n, p_1, p_2) = \begin{pmatrix} (1-\gamma)s_1 & b(n)nf(p_2) \\ \gamma s_1 & s_2 \end{pmatrix}.$$

We note that the matrix $A(n, p_1, p_2)$ is nonnegative on the set $\mathcal{A}_1 = \{(n, p_1, p_2) | \epsilon \leq n \leq \bar{n}, 0 \leq p_1 \leq P_1, 0 \leq p_2 \leq P_2\}$ for some $\epsilon > 0$. Now the matrix A evaluated at $(\bar{n}, 0, 0)$,

$$A(\bar{n}, 0, 0) = \begin{pmatrix} (1-\gamma)s_1 & b(\bar{n})\bar{n}f(0) \\ \gamma s_1 & s_2 \end{pmatrix},$$

is primitive, since all the entries are non-negative provided that $\phi(0) > 1$. Also, $A(z)\eta \neq 0$ for all $(z, \eta) \in \mathcal{A}_1 \times U_+$, where η is a unit vector in $\mathbb{R}^2_+$ and U_+ is the set of such unit vectors in $\mathbb{R}^2_+$. We note that both the sets Z and $Z \setminus N_1$ are non-empty and positively invariant. Suppose $M_1 = \mathcal{A}_1 \cap N_1$. Then M_1 is a non-empty, compact, positively invariant set that is bounded away from zero and attracts all non-zero points of N_1. Moreover, $\Omega(M_1) = \{(\bar{n}, 0, 0)\}$. Note that, $\text{tr}(A) = (1-\gamma)s_1 + s_2$, and $\det(A) = (1-\gamma)s_1s_2 - \gamma s_1 b(\bar{n})\bar{n}f(0)$. Thus the eigenvalues of A are given by

$$\begin{aligned} \lambda_{1,2} &= \frac{\text{tr}(A) \pm \sqrt{\text{tr}^2(A) - 4\det(A)}}{2}, \\ &= \frac{(1-\gamma)s_1 + s_2 \pm \sqrt{((1-\gamma)s_1 + s_2)^2 + 4((1-\gamma)s_1s_2 - \gamma s_1 b(\bar{n})\bar{n}f(0))}}{2}. \end{aligned}$$

If $R_{\bar{n}} > 1$, then it is easy to verify that the spectral radius of $A(\bar{n}, 0, 0)$, given by λ_1, is greater than one. Applying Corollary 1 along with Theorem 2.3 in [38], M_1 is a uniformly weak repeller and thus the system is $\rho-$ persistent with $\rho = p_1 + p_2$. That is, there exists an $\epsilon > 0$ such that for all $z_0 \in Z^+$

$$\liminf_{t \to \infty} (p_1 + p_2) > \epsilon.$$

- Step 3: In this step we show that

$$\liminf_{t\to\infty}(p_1+p_2)>\epsilon \implies \liminf_{t\to\infty} p_1>\epsilon_1 \text{ or } \liminf_{t\to\infty} p_2>\epsilon_2, \tag{24}$$

i.e. either the juvenile or the adult population persists. Indeed, we have that

$$\liminf_{t\to\infty}(p_1+p_2)>\epsilon \implies \limsup_{t\to\infty} p_1>\epsilon_1 \text{ or } \limsup_{t\to\infty} p_2>\epsilon_2, \tag{25}$$

However, by Proposition 3.2 in [30], for $k=1,2$, we have

$$\limsup_{t\to\infty} p_k>\epsilon_k \implies \liminf_{t\to\infty} p_k>\epsilon_k.$$

- Step 4: We show that

$$\liminf_{t\to\infty} p_1>\epsilon_1 \iff \liminf_{t\to\infty} p_2>\epsilon_2.$$

Suppose first that p_1 is strongly uniform persistent, i.e. $\liminf_{t\to\infty} p_1>\epsilon_1$. We claim that p_2 is weakly uniform persistent, i.e. $\limsup_{t\to\infty} p_2>\epsilon_2$ for some $\epsilon_2>0$. We prove the claim by way of contradiction. Suppose the claim to be false, that is p_2 is not weakly uniform persistent. Then $\limsup_{t\to\infty} p_2<\epsilon_2$ for any $\epsilon_2>0$. Then we have that $\lim_{t\to\infty} p_2=0$ and from the third equation of (1), $\lim_{t\to\infty} p_1=0$. This contradicts the fact that p_1 is strongly uniform persistent. As a result, we have that $\limsup_{t\to\infty} p_2>\epsilon_2$. By Proposition 3.2 in [30], $\limsup_{t\to\infty} p_2>\epsilon_2$ implies $\liminf_{t\to\infty} p_2>\epsilon_2$. Thus, $\liminf_{t\to\infty} p_1>\epsilon_1 \implies \liminf_{t\to\infty} p_2>\epsilon_2$. In a similar manner, we can show that $\liminf_{t\to\infty} p_2>\epsilon_2 \implies \liminf_{t\to\infty} p_1>\epsilon_1$. This completes the proof.

3 Numerical Studies

In this section, we demonstrate the possible dynamics of model (1). In Example 2, we illustrate the results of the theorems presented in the previous section. Meanwhile, in Example 3 we show that model (1) may exhibit rich dynamics. This is in stark contrast to the unstructured predator-prey model developed in [3]. For this model, only stable equilibria were observed when the nonlinearities are given by the Beverton-Holt functions (16) [4].

Example 2 In Fig. 1 we illustrate the dynamics of the predator-prey model (1) under the conditions stated in Theorems 1 and 2. The time-series graphs were obtained using the functions given in (16) and varying the inherent prey growth rate r_0, while keeping all other parameters fixed at the following values: $m=1$, $c=1$, $s_1=0.5$, $s_2=0.6$, $\gamma=0.5$, $\delta=1$, and $b_0=2$. For the simulation presented in Fig. 1a we let

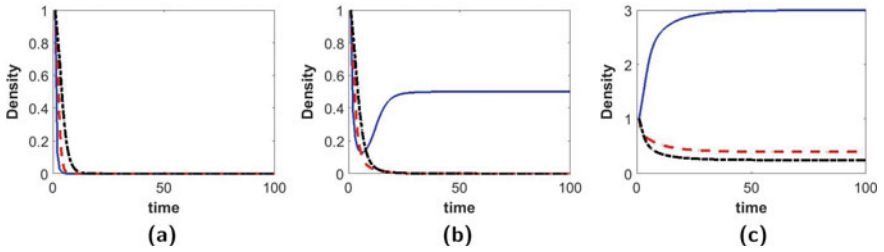

Fig. 1 Shown are the time series dynamics for the predator-prey model (1) obtained using the nonlinearities given in (16) and various values of r_0. Here the solid blue lines are the prey, the dashed red lines are the juvenile predator, and the dashed-dotted black lines are the adult predator. For all graphs we use the parameter values $m = 1,\ c = 1,\ s_1 = 0.5,\ s_2 = 0.6,\ \gamma = 0.5,\ \delta = 1,$ and $b_0 = 2$. **a** For $\phi(0) = r_0 = 0.5$, both the prey and predator go extinct since $\phi(0) < 1$; **b** For $\phi(0) = r_0 = 1.5$, the prey survives while the predator goes extinct since $\phi(0) > 1$ and $R_{\bar{n}} = 0.5556 < 1$; **c** For $\phi(0) = r_0 = 5$, both the prey and the predator persist since $\phi(0) > 1$ and $R_{\bar{n}} = 1.3333 > 1$

$r_0 = 0.5$. For this value of r_0, we have $\phi(0) = r_0 < 1$ and, hence, both the predator and prey populations go extinct, as discussed in Theorem 1. For the simulation presented in Fig. 1b we let $\phi(0) = r_0 = 1.5$. Here we observe that the prey survives but the predator population goes extinct, as stated in Theorem 2. Finally, in Fig. 1c we use $\phi(0) = r_0 = 5$ resulting in both the prey and predator populations persisting, as concluded in Theorem 5. Here we observe a stable interior equilibrium.

Example 3 *(Rich dynamics resulting from predator structure)*

We generate bifurcation diagrams for model (1) using the nonlinearities given in (16) along with the following four sets of parameter values:

(i) $r_0 = 5, m = 1.1,\ c = 1,\ s_1 = 0.95,\ \gamma = 0.5,\ \delta = 1.1,$ and $b_0 = 2$,
(ii) $r_0 = 5, m = 0.1,\ c = 1,\ s_1 = 0.5,\ \gamma = 0.5,\ \delta = 0.1,$ and $b_0 = 2$,
(iii) $r_0 = 5, m = 0.1,\ c = 1,\ s_1 = 0.95,\ \gamma = 0.5,\ \delta = 0.1,$ and $b_0 = 2$,
(iv) $r_0 = 5, m = 0.1,\ c = 1,\ s_1 = 0.95,\ \gamma = 1,\ \delta = 0.1,$ and $b_0 = 2$.

In Fig. 2, we give bifurcation diagrams for model (1) with respect to the parameter value s_2. Each row was obtained using the corresponding set of parameters listed above. To generate these graphs, for each value of s_2 running from $s_2 = 0$ to $s_2 = 1$ with a step-size of 0.001, we ran the model for 10000 iterations and plotted the last 200 data points for each of n, p_1, and p_2 against the corresponding s_2 values.

For the parameter values in (i), we observe stable equilibria for all values of s_2 (shown in Fig. 2a–c). On the other hand, for the parameter values in (ii) and (iii), the system shows chaotic behavior (as was verified through calculation of the Lyapunov exponent, not shown). We note that the only parameter that differs in these two parameter sets is s_1, with $s_1 = 0.5$ for (ii) (shown in Fig. 2d–f) and $s_1 = 0.95$ for (iii) (shown in Fig. 2d–f). These graphs show that increasing the juvenile predator survival may cause the chaotic region to shift left, resulting in chaotic behavior for

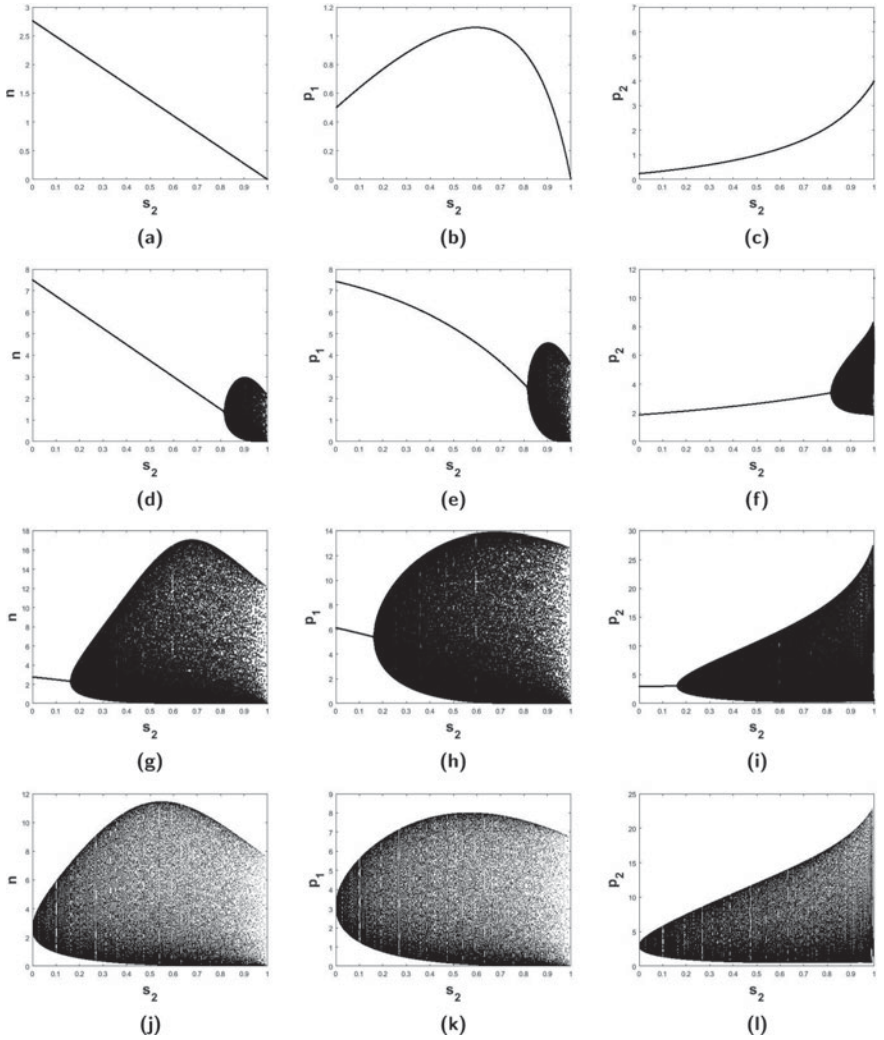

Fig. 2 Shown are the bifurcation diagrams for the prey (first column), the juvenile predator (second column), and the adult predator (third column) using the sets of parameter values listed in Example 3. Figures **a–c** use the parameters listed in (i); Figures **d–f** use the parameters listed in (ii); Figures **g–i** use the parameters listed in (iii); and Figures **j–l** use the parameters listed in (iv)

smaller values of s_2. Finally, the system exhibits rich dynamics even when $s_2 = 0$ if $\gamma = 1$. This corresponds to the case where juvenile predators mature after one time step (shown in Fig. 2j–l).

4 Conclusion

We investigated the dynamics of a discrete-time predator-prey model with stage-structure in the predator. In this model, we assumed that the predator population consists of a juvenile and an adult stage, with only the adult stage consuming the prey population. We established the existence and global stability of the boundary equilibria of this model in Theorems 1 and 2. The conditions for the existence of a unique interior equilibrium are given in Theorem 3. These conditions are also the same conditions for the persistence of the prey and predator populations, which is shown in Theorem 5. The conditions in these theorems depend on two quantities: the inherent growth rate of the prey $\phi(0)$ and the invasion net reproductive number of the predator when the prey is at its predator-free equilibrium density $R_{\bar{n}}$. These quantities are defined in terms of general non-linear functions, as given by the set X.

While we proved that the interior equilibrium is locally asymptotically stable when $R_{\bar{n}} \gtrapprox 1$, it remains an open problem to determine the extent of the stability of the interior equilibrium. Though it was shown for the case of an unstructured predator population that, given a certain restriction on the nonlinearity f, the interior equilibrium is always stable when it exists [3], numerical simulations show that the interior equilibrium of model (1) may be unstable in some parameter ranges. In fact, numerical simulations of this simple discrete-time model have revealed rich dynamics. In Example 3, we show that model (1) may have stable equilibria or may exhibit chaotic dynamics depending on the choice of parameter values. For our particular examples, these different scenarios were obtained by varying the parameters s_1 and γ. In particular, increased ranges of chaotic dynamics were observed when either of these values were increased, which corresponds to increasing $R_{\bar{n}}$.

The investigations in this paper contribute to the understanding of how stage-structure may influence predator-prey interactions. In particular, we observe that stage-structure may introduce complicated dynamics that are not observed for unstructured predator and prey populations. However, this model has a number of simplifying assumptions. In particular, we assumed that prey consumption is only dependent on the adult predator and we did not explicitly consider juvenile consumption of resources. Natural extensions of this model would be to assume that both juvenile and adult predators consume the prey or that adults and juveniles have two distinct prey populations. It is also of interest to consider how prey evolution, as considered in [3], may impact model dynamics. In future work, we will address these issues in order to gain a better understanding of the intriguing nature of interacting species.

Acknowledgements This research is part of the Littoral Acoustic Demonstration Center-Gulf Ecological Monitoring and Modeling (LADC-GEMM) consortium project supported by Gulf of Mexico Research Initiative Year 5-7 Consortia Grants (RFP-IV). Data are publicly available through the Gulf of Mexico Research Initiative Information & Data Cooperative (GRIIDC) at https://data.gulfresearchinitiative.org (doi:10.7266/4D5V2P5D).

References

1. Abrams, P.A.: Prey evolution as a cause of predator-prey cycles. Evolution **51**, 1740–1748 (1997)
2. Abrams, P.A.: Modelling the adaptive dynamics of traits involved in inter- and intraspecific interactions: an assessment of three methods. Ecol. Lett. **4**(2), 166–175 (2001). https://doi.org/10.1046/j.1461-0248.2001.00199.x
3. Ackleh, A.S., Hossain, M.I., Veprauskas, A., Zhang, A.: Persistence and stability analysis of discrete-time predator-prey models: a study of population and evolutionary dynamics. J. Differ. Equ. Appl. **25**(11), 1568–1603 (2019). https://doi.org/10.1080/10236198.2019.1669579
4. Ackleh, A.S., Hossain, M.I., Veprauskas, A., Zhang, A.: Long-term dynamics of discrete-time predator-prey models: Stability of equilibria, cycles, and chaos. J. Differ. Equ. Appl. (2020). https://doi.org/10.1080/10236198.2020.1786818
5. Agarwal, R.P.: Difference Equations and Inequalities: Theory, Methods, and Applications. CRC Press (2000)
6. Al-Omari, J.F.M.: The effect of state dependent delay and harvesting on a stage-structured predator-prey model. Appl. Math. Comput. **271**, 142–153 (2015)
7. Amarasekare, P.: Coexistence of intraguild predators and prey in resource-rich environments. Ecology **89**(10), 2786–2797 (2008)
8. Arditi, R., Ginzburg, L.R.: Coupling in predator-prey dynamics: ratio-dependence. J. Theoret. Biol. **139**(3), 311–326 (1989)
9. Beverton, R.J.H., Holt, S.J.: On the Dynamics of Exploited Fish Populations, vol. 11. Springer Science & Business Media (2012)
10. Bonnet, X., Naulleau, G., Shine, R., Lourdais, O.: Short-term versus long-term effects of food intake on reproductive output in a viviparous snake, vipera aspis. Oikos **92**(2), 297–308 (2001)
11. Caswell, H.: Matrix population models. Encyclopedia of Environmetrics **3** (2006)
12. Chen, F., Chen, W., Wu, Y., Ma, Z.: Permanence of a stage-structured predator-prey system. Appl. Math. Comput. **219**(17), 8856–8862 (2013)
13. Chen, F., Xie, X., Li, Z.: Partial survival and extinction of a delayed predator-prey model with stage structure. Appl. Math. Comput. **219**(8), 4157–4162 (2012)
14. Chen, X.W., Fu, X.L., Jing, Z.J.: Dynamics in a discrete-time predator-prey system with allee effect. Acta Mathematicae Applicatae Sinica, English Series **29**(1), 143–164 (2013)
15. Choudhury, S.R.: On bifurcations and chaos in predator-prey models with delay. Chaos Solitons Fractals **2**(4), 393–409 (1992)
16. Cushing, J.M.: An introduction to structured population dynamics, vol. 71. SIAM (1998)
17. Cushing, J.M.: Nonlinear semelparous leslie models. Math. Biosci. Eng. **3**(1), 17 (2006)
18. Cushing, J.M., Henson, S.M.: Stable bifurcations in semelparous leslie models. J. Biolog. Dyn. **6**(sup2), 80–102 (2012)
19. Cushing, J.M., Yicang, Z.: The net reproductive value and stability in matrix population models. Natural Resou. Model. **8**(4), 297–333 (1994)
20. Freedman, H.I.: Deterministic mathematical models in population ecology, vol. 57. Marcel Dekker Incorporated (1980)
21. Hirsch, M.W.: Systems of differential equations that are competitive or cooperative ii: Convergence almost everywhere. SIAM J. Math. Anal. **16**(3), 423–439 (1985)
22. Hjelm, J., Persson, L., Christensen, B.: Growth, morphological variation and ontogenetic niche shifts in perch (perca fluviatilis) in relation to resource availability. Oecologia **122**(2), 190–199 (2000)
23. Holt, R.D., Huxel, G.R.: Alternative prey and the dynamics of intraguild predation: theoretical perspectives. Ecology **88**(11), 2706–2712 (2007)
24. Huffaker, C.: Experimental studies on predation: dispersion factors and predator-prey oscillations. Hilgardia **27**(14), 343–383 (1958)
25. Jang, S.R.J., Ackleh, A.S.: Discrete-time, discrete stage-structured predator-prey models. J. Differ. Equ. Appl. **11**(4–5), 399–413 (2005)

26. Kazarinoff, N., Van Den Driessche, P.: A model predator-prey system with functional response. Math. Biosci. **39**(1–2), 125–134 (1978)
27. Kon, R.: Nonexistence of synchronous orbits and class coexistence in matrix population models. SIAM J. Appl. Math. **66**(2), 616–626 (2005)
28. Lotka, A.J.: Contribution to the theory of periodic reactions. J. Phys. Chem. **14**(3), 271–274 (1909). https://doi.org/10.1021/j150111a004
29. Lotka, A.J.: Elements of physical biology. Williams & Wilkins Company, Baltimore (1925). http://library.wur.nl/WebQuery/clc/529141
30. Magal, P., Zhao, X.Q.: Global attractors and steady states for uniformly persistent dynamical systems. SIAM J. Math. Anal. **37**(1), 251–275 (2005). https://doi.org/10.1137/S0036141003439173
31. McCauley, E., Murdoch, W.W., Watson, S.: Simple models and variation in plankton densities among lakes. The American Naturalist **132**(3), 383–403 (1988)
32. Meissen, E.P.: Invading a structured population: a bifurcation approach (2017). http://repository.arizona.edu/handle/10150/625610
33. Miller, T.E., Rudolf, V.H.: Thinking inside the box: community-level consequences of stage-structured populations. Trends Ecol. Evolut. **26**(9), 457–466 (2011)
34. Mittelbach, G., Osenberg, C., Leibold, M.: Trophic relations and ontogenetic niche shifts in aquatic ecosystems. In: Size-structured populations, pp. 219–235. Springer (1988)
35. Murray, J.D.: Mathematical biology, vol. 19 of biomathematics (1989)
36. Daugherty, M.P., Harmon, J.P., Briggs, C.J.: Trophic supplements to intraguild predation. Oikos **116**(4), 662–677 (2007)
37. Rindorf, A., Wanless, S., Harris, M.: Effects of changes in sandeel availability on the reproductive output of seabirds. Marine Ecol. Progress Series **202**, 241–252 (2000)
38. Salceanu, P., Smith, H.: Lyapunov exponents and persistence in discrete dynamical systems. Discrete Contin. Dyn. Syst. Ser. B **12**(1), 187–203 (2009). https://doi.org/10.3934/dcdsb.2009.12.187
39. Sun, L., Fu, S., Ma, W.: Pattern formation in a predator-prey diffusion model with stage structure for the predator. Comput. Math. Appl. **70**(12), 2988–3000 (2015)
40. Volterra, V.: Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. C. Ferrari (1927)
41. Wang, Q., Fan, M., Wang, K.: Dynamics of a class of nonautonomous semi-ratio-dependent predator-prey systems with functional responses. J. Math. Anal. Appl. **278**(2), 443–471 (2003)
42. Yoshida, T., Jones, L.E., Ellner, S.P., Fussmann, G.F., Hairston Jr., N.G.: Rapid evolution drives ecological dynamics in a predator-prey system. Nature **424**(6946), 303 (2003)

Techniques on Solving Systems of Nonlinear Difference Equations

JERICO B. BACANI and Julius Fergy T. Rabago

Abstract This paper provides alternative techniques on solving some systems of difference equations. These techniques are analytical and much explanatory in nature as compared to methods used in existing literatures. We applied these methods particularly to the systems studied by Touafek in his paper Touafek (Iran J Math Sci Info 9(2): 303–305, 2014, [33]). We found out that these strategies can be used also in solving other systems that are closely related to our work. Interestingly, some of the systems are found to posses closed-form solutions that consist of intriguing integer sequences, such as those found in nature and polyenoids.

Keywords Difference equations · Systems of difference equations · Closed-form solutions

1 Introduction

1.1 Background and Motivation

Let I be a subset of the set of all real numbers $\mathbb{R}$ and $f : I^{k+1} \to I$ be a continuously differentiable function. Any equation of the form given by

$$x_{n+1} = f(x_n, x_{n-1}, \ldots, x_{n-k}), \quad n \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}, \tag{1}$$

is called a *difference equation of order* $k + 1$—a specific type of *recurrence relation*. It is known that for every set of initial conditions $\{x_n\}_{n=-k}^{0} \subset I$, the difference

J. B. BACANI (✉)
University of the Philippines Baguio, 2600 Baguio, Philippines
e-mail: jbbacani@up.edu.ph
URL: http://upb.edu.ph

J. F. T. Rabago
Department of Complex Systems Science, Graduate School of Informatics, Nagoya University, A4-2 (780) Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan

S. Baigent et al. (eds.), *Progress on Difference Equations and Discrete Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 341,
https://doi.org/10.1007/978-3-030-60107-2_7

Eq. (1) has a unique solution $\{x_n\}_{n=-k}^{\infty} := \{x_n\}_{-k}^{\infty}$ (cf. [15]). Some well-known difference equations such as the Fibonacci sequences [16, 37] were originally discovered to model population dynamics. In present times, difference equations are fundamentally important in various fields of mathematics and related sciences such as physics, probability theory, biology, ecology, epidemiology, etc. They are used extensively in both theoretical and empirical economics [29]. These equations are actually discrete analogues of differential equations and are used in solving their continuous 'counterparts' numerically. Difference equations also have great importance in analysis of algorithms [30] and have valuable applications in digital signal processing [22]. For more applications of difference equations in physical, life and natural sciences, we refer the readers to the monograph of Jagerman [14], the text of Kulenović and Ladas [17], and books of Mickens [20] and Sharkovsky [31]. For a good introduction about theory on difference equations, we recommend a book by Elaydi [6].

For the past few years, difference equations have attracted the attention of many researchers. We have witnessed a rapid growth in the number of papers published dealing with these types of equations. One of the hot topics that gains much interest on this field is the problem of finding the closed-form solutions of some solvable systems of nonlinear difference equations. This is advantageous on our part as researchers for if we know the solution form, we can examine easily and predict the dynamical behavior of such systems. We can also easily understand the concepts of boundedness, asymptoticity and periodicity of the solutions.

In terms of solving linear difference equations, various methods can readily be found in existing literatures. In [1], for instance, several methods have been presented in solving special linear recurrence equations related to Fibonacci, Pell, Jacobsthal and Balancing number sequences, but there are still no general methods available for solving nonlinear types of difference equations. Also, as far as we know, not so much effort has been done to provide readers analytical methods in solving systems of difference equations. Nevertheless, the method of mathematical induction is usually used by most experts to establish the solution forms of these solvable systems of difference equations (cf. [7–9, 33–36] and the references cited therein). We emphasize, however, that this method has some disadvantages. For instance, it does not give much detail on how could one derive analytically the solution form of a particular system of difference equations. Meanwhile, Rabago [23, 27, 28] was able to study intensively several systems of nonlinear difference equations by reducing them to linear types through appropriate transformations. In fact, various techniques were also devised to solve several nonlinear difference equations whose solutions were expressible in closed-forms (see, e.g., [2, 11, 24, 28]), providing clear explanations on some existing results that were first justified only through the induction principle (see [25, 26] for instance).

In [3], Brand was able to find the solution form of the Riccati difference equation

$$x_{n+1} = \frac{a + bx_n}{c + dx_n}, \qquad n \in \mathbb{N}_0,$$

through the transformation $x_n = y_n - d/c$. He examined completely its limiting properties and obtained an interesting result related to continued fractions by using the transformation $y_n = z_{n+1}/z_n$. The same problem appeared and was considered by Stevic [32]. In [13], Iricanin and Liu described in a simpler and more elegant way the behavior of positive solutions of the higher-order difference equations

$$x_n = \frac{cx_{n-p}x_{n-p-q}}{x_{n-q}}, \qquad n \in \mathbb{N}_0, \tag{2}$$

where $p, q \in \mathbb{N}$ and $c > 0$. The method employed by Iricanin and Liu to investigate equation (2) uses some elementary properties of logarithms. This approach was discovered independently by Rabago and was applied in [23, 28] in examining similar problems. Moreover, this approach is found to be effective in dealing with such types of problems, especially in determining the periodicity of solutions of some systems of difference equations (cf. [23]).

In this work we revisit some systems of nonlinear difference equations which were previously studied by Touafek [33]. This time, we use alternative methods in deriving the solution forms of these systems of equations. One technique that we find very powerful in addressing this problem is the so-called *method of differences*, also known as *the method of telescoping sums*. This method simplifies the sum $\sum_{n=1}^{N}\{a_n - a_{n-1}\}$, where $\{a_n\}_1^N$ is some number sequence. More precisely, given a number sequence $\{a_n\}_1^N$, we get the identity

$$\sum_{n=1}^{N}\{a_n - a_{n-1}\} = a_N - a_0.$$

through 'telescoping'. As we shall see in our discussion, this method works perfectly in determining the solution form of the following system of difference equations

$$x_{n+1} = \frac{y_{n-3}y_nx_{n-2}}{y_{n-3}x_{n-2} \pm y_{n-3}y_n \pm y_nx_{n-2}}, \quad y_{n+1} = \frac{y_{n-2}x_{n-1}}{2y_{n-2} \pm x_{n-1}}, \quad n \in \mathbb{N}_0, \tag{S}$$

when reduced to linear types via appropriate transformations.

Now, in relation to our set objective, that is, to find the closed-form solution of system (S), the following results and notations are needed.

Definition 1 (*Periodicity*) A sequence $\{x_n\}_{-k}^{\infty}$ is said to be periodic with period p if $x_{n+p} = x_n$ for every $n \geq -k$.

Definition 2 ([10]) A solution $\{x_n\}_{-k}^{\infty}$ of (1) is called *eventually periodic with period* p if there exists an integer $N \geq -k$ such that $\{x_n\}_N^{\infty}$ is periodic with period p; that is, $x_{n+p} = x_n$, for all $n \geq N$.

The rest of the paper is structured as follows. In Sect. 2, we present some preliminaries which are requisites to our main results presented in Sect. 3. Results are accompanied by illustrations. Then, in Sect. 4, we end our paper with a summary of discussion, plus suggested works for future investigations.

2 Preliminary Results

2.1 Review of Integer Sequences

In the following discussion, we present two integer sequences that are essential to our results.

Sequence No. A000045. The widely-studied Fibonacci sequence $\{f_n\}_0^\infty$ satisfies the second-order linear recurrence equation $f_{n+1} = f_n + f_{n-1}$, with initial values $f_0 = 0$ and $f_1 = 1$. Its first few terms, starting with $n = 0$, are 0, 1, 1, 2, 3, 5, 8, 13, 21, 34, ... (cf. Sequence No. A000045 in O.E.I.S [21]). This sequence can also be extended into negative indexes. More precisely, one can generate the sequence $\{f_{-n}\}_1^\infty$—the Fibonacci numbers with negative indexes—using the relation $f_{-n} = (-1)^{n+1} f_n$. Hence, one easily finds that $f_{-1} = 1$, $f_{-2} = -1$, $f_{-3} = 2$, $f_{-4} = -3$ and so on. This number sequence is also known to posses many exciting properties (see, e.g., [5, 16, 37, 38]) and has been generalized in various ways (see, e.g., Larcombe's survey paper [18] about Horadam sequences—a second-order linear recurrence sequence named after Horadam [12] for his extensive study of these numbers; and the interesting paper of Lucas [19]). It's Binet formula is given by

$$f_n = \frac{\phi^n - (1-\phi)^n}{\sqrt{5}}, \qquad n \in \mathbb{N}_0,$$

where ϕ denotes the well-known golden ratio [5, 37]. In Sect. 3, we shall see how the solution form of a particular case of system (S) can be expressed in terms of the Fibonacci numbers. In this regard, the following lemma will be useful.

Lemma 1 *Let $\{f_n\}_0^\infty$ denote the Fibonacci sequence and $\{g_n\}_0^\infty$ be an integer sequence generating the number series*

$$0,\ 1,\ 4,\ 17,\ 72,\ 305,\ 1292,\ 5473,\ \ldots,\ g_{n+1} = 4g_n + g_{n-1},\ \ldots.$$

Then, for all $n \in \mathbb{N}_0$, the following identities hold.

(i) $g_{n+1} - g_n = f_{3n+1}$,

(ii) $g_{n+1} + g_n = \frac{1}{2}\{3g_{n+1} + 2g_n + g_{n-1}\} = f_{3n+2}$,

(iii) $2g_{n+1} = \frac{1}{2}\{g_{n+2} - g_n\} = f_{3n+3}$.

Proof Consider the recurrence equation given by $g_{n+1} = 4g_n + g_{n-1}$, with initial values $g_0 = 0$ and $g_1 = 1$. We can find the Binet's form for g_n as follows: Using the ansatz $g_n = \lambda^n$ ($n \in \mathbb{N}_0$), we obtain the quadratic equation $P(\lambda) = \lambda^2 - 4\lambda - 1 = 0$. Since $P(\lambda) = 0$ has two distinct roots $\lambda_{1,2} = 2 \pm \sqrt{5}$, we can express g_n as

$$g_n = c_1\lambda_1^n + c_2\lambda_2^n, \qquad n \in \mathbb{N}_0,$$

where c_1 and c_2 can be computed by solving the system of equations given by

$$c_1 + c_1 = 0 \qquad \text{and} \qquad \lambda_1 c_1 + \lambda_2 c_2 = 1.$$

Hence, we have

$$\forall n \in \mathbb{N}_0 : \qquad g_n = \frac{(2+\sqrt{5})^n - (2-\sqrt{5})^n}{2\sqrt{5}}.$$

One can check that, indeed,

$$\{g_n\}_0^\infty = \left\{ \frac{(2+\sqrt{5})^n - (2-\sqrt{5})^n}{2\sqrt{5}} \right\}_0^\infty = \{0, 1, 4, 17, 72, 305, 1292, 5473, \ldots\}.$$

Identity (i). For all $n \in \mathbb{N}_0$, we have, using the relations $\phi^3 = 2 + \sqrt{5}$ and $(1 - \phi)^3 = 2 - \sqrt{5}$,

$$\begin{aligned} \frac{1}{2}\{g_{n+1} - g_n\} &= \frac{1}{2}\left\{ \frac{(2+\sqrt{5})^{n+1} - (2-\sqrt{5})^{n+1}}{2\sqrt{5}} - \frac{(2+\sqrt{5})^n - (2-\sqrt{5})^n}{2\sqrt{5}} \right\} \\ &= \frac{\phi^{3n}\phi - (1-\phi)^{3n}(1-\phi)}{\sqrt{5}} = f_{3n+1}. \end{aligned}$$

Identity (ii). Furthermore, with the identities $\phi^2 = 3 + \sqrt{5}$ and $(1 - \phi)^2 = 3 - \sqrt{5}$, we have

$$\begin{aligned} g_{n+1} + g_n &= \frac{(2+\sqrt{5})^{n+1} - (2-\sqrt{5})^{n+1}}{2\sqrt{5}} + \frac{(2+\sqrt{5})^n - (2-\sqrt{5})^n}{2\sqrt{5}} \\ &= \frac{\phi^{3n}\phi^2 - (1-\phi)^{3n}(1-\phi)^2}{\sqrt{5}} = f_{3n+2}. \end{aligned}$$

Identity (iii). Noting that $\phi^3 = 2 + \sqrt{5}$ and $(1 - \phi)^3 = 2 - \sqrt{5}$,

$$2g_{n+1} = 2\left\{ \frac{(2+\sqrt{5})^{n+1} - (2-\sqrt{5})^{n+1}}{2\sqrt{5}} \right\} = \frac{\phi^{3n} - (1-\phi)^{3n}}{\sqrt{5}} = f_{3n}.$$

This proves the lemma.

Remark 1 The number sequence $\{g_n\}_0^\infty$ is, in fact, the number sequence numbered as Sequence No. A001076 in O.E.I.S. [21].

Sequence No. A000912. Consider the number sequence $\{u_n(u_0, u_1)\}_0^\infty$ defined by the recurrence relation, given its real initial values u_0 and u_1 (not simultaneously zero),

$$u_{n+1} = 2u_n + 3u_{n-1}, \quad n \geq 1. \tag{3}$$

Its first few terms, with u_0 and u_1 set to 0 and 2, respectively, are given by, starting with $n = 2$,

$$u_2 = 4,\ u_3 = 14,\ u_4 = 40,\ u_5 = 132,\ u_6 = 424,\ u_7 = 1430, u_8 = 4848,\ \ldots$$

With the usual approach in solving linear recurrence equations, the corresponding Binet's formula of u_n is easily established as follows: Using the ansatz $u_n = \lambda^n, n \in \mathbb{N}_0$, in Eq. (3) we get the quadratic equation $P(\lambda) = \lambda^2 - 2\lambda - 3 = 0$, whose roots are given by $\lambda_1 = 3$ and $\lambda_2 = -1$. Since $\lambda_1 \neq \lambda_2$, then u_n can be written in the form $u_n = c_1\lambda_1^n + c_2\lambda_2^2$, where c_1 and c_2 are computable constants. Indeed, $c_1 + c_2 = u_0$ and $3c_1 - c_2 = u_1$. Computing for the unknowns c_1 and c_2, we get $c_1 = \frac{1}{4}(u_0 + u_1)$ and $c_2 = \frac{1}{4}(3u_0 - u_1)$, respectively. Thus, the n-th term of the sequence $\{u_n(0, 2)\}_0^\infty$ can be found explicitly using the formula

$$u_n = \frac{1}{2}\left\{3^n - (-1)^n\right\}, \qquad n \in \mathbb{N}_0. \tag{4}$$

Interestingly, the sequence $\{u_n(0, 2)\}_0^\infty = \{0, 2, 4, 14, 40, 132, 424, 1430, 4848, \ldots\}$ (with 1 and 0 replaced by 0 and 2, respectively, as its first two values) appears to be Sequence No. A000912 in O.E.I.S. Apparently, this sequence, whose n-th term is actually given by the formula

$$n\text{-th term of Seq. A00912} = \begin{cases} C(n), & \text{if } n \text{ is even,} \\ C(n) - C(\frac{n-1}{2}), & \text{if } n \text{ is odd,} \end{cases}$$

where

$$C(n) = \frac{1}{n+1}\binom{2n}{n},$$

denoted as the n-th Catalan number (cf. Sequence No. A000108 in O.E.I.S.), seems to have some sorts of connection with the number of bond-rooted polyenoids with $2n - 1$ edges (cf. [4]). Consequently, with the above relations, we are able to describe a new formula for the sequence A000912 in O.E.I.S. More precisely, we have the following proposition.

Proposition 1 *For all $n \in \mathbb{N}_0$, we have*

$$n\text{-th term of Seq. A00912} = \begin{cases} 1, & \text{if } n = 0, \\ \frac{1}{2}\left\{3^{n-1} - (-1)^{n-1}\right\}, & \text{otherwise.} \end{cases}$$

3 Main Results

Let $n \in \mathbb{N}_0$ and consider the following systems of nonlinear difference equations:

$$x_{n+1} = \frac{y_{n-3}y_n x_{n-2}}{y_{n-3}x_{n-2} + y_{n-3}y_n + y_n x_{n-2}}, \qquad y_{n+1} = \frac{y_{n-2}x_{n-1}}{2y_{n-2} + x_{n-1}}, \tag{S.1}$$

$$x_{n+1} = \frac{y_{n-3}y_n x_{n-2}}{y_{n-3}x_{n-2} - y_{n-3}y_n + y_n x_{n-2}}, \qquad y_{n+1} = \frac{y_{n-2}x_{n-1}}{2y_{n-2} + x_{n-1}}, \tag{S.2}$$

$$x_{n+1} = \frac{y_{n-3}y_n x_{n-2}}{y_{n-3}x_{n-2} + y_{n-3}y_n - y_n x_{n-2}}, \qquad y_{n+1} = \frac{y_{n-2}x_{n-1}}{2y_{n-2} + x_{n-1}}, \tag{S.3}$$

$$x_{n+1} = \frac{y_{n-3}y_n x_{n-2}}{y_{n-3}x_{n-2} + y_{n-3}y_n + y_n x_{n-2}}, \qquad y_{n+1} = \frac{y_{n-2}x_{n-1}}{2y_{n-2} - x_{n-1}}, \tag{S.4}$$

$$x_{n+1} = \frac{y_{n-3}y_n x_{n-2}}{y_{n-3}x_{n-2} - y_{n-3}y_n - y_n x_{n-2}}, \qquad y_{n+1} = \frac{y_{n-2}x_{n-1}}{2y_{n-2} + x_{n-1}}, \tag{S.5}$$

$$x_{n+1} = \frac{y_{n-3}y_n x_{n-2}}{y_{n-3}x_{n-2} + y_{n-3}y_n - y_n x_{n-2}}, \qquad y_{n+1} = \frac{y_{n-2}x_{n-1}}{2y_{n-2} - x_{n-1}}, \tag{S.6}$$

$$x_{n+1} = \frac{y_{n-3}y_n x_{n-2}}{y_{n-3}x_{n-2} - y_{n-3}y_n + y_n x_{n-2}}, \qquad y_{n+1} = \frac{y_{n-2}x_{n-1}}{2y_{n-2} - x_{n-1}}, \tag{S.7}$$

$$x_{n+1} = \frac{y_{n-3}y_n x_{n-2}}{y_{n-3}x_{n-2} - y_{n-3}y_n - y_n x_{n-2}}, \qquad y_{n+1} = \frac{y_{n-2}x_{n-1}}{2y_{n-2} - x_{n-1}}, \tag{S.8}$$

with real nonzero initial values $x_{-2}, x_{-1}, x_0, y_{-3}, y_{-2}, y_{-1}$ and y_0.

In this study, we analyze the forms and behaviors of the well-defined solutions of the above systems by taking into account the following substitutions on the phase variables:

$$w_n = \frac{1}{x_n} \quad \text{and} \quad z_n = \frac{1}{y_n}, \quad \text{for all } n \in \mathbb{N}_0. \tag{5}$$

Remark 2 By well-defined solutions of systems (S.1)–(S.8), we mean a solution generated by the set of initial points $\{x_n\}_{-2}^{0}$ and $\{y_n\}_{-3}^{0}$ taken outside the systems' respective singularity sets. An initial set of points $\{x_n\}_{-2}^{0}$ and $\{y_n\}_{-3}^{0}$ that generates a solution $\{(x_n, y_n)\}_1^{\infty}$ of equation (S) with at least one of the denominators equal to

zero for some least index n leads to an undefined value for x_{n+1} and/or y_{n+1}. We call this set of all such initial points the *singularity set* of equations (S). The singularity set is also called the "forbidden set" in the literature (cf. [10, 17]).

Remark 3 We mention that four out of the eight systems above have already been studied by Touafek in [33]. Particularly, Touafek established the solution forms of systems (S.1), (S.4), (S.5) and (S.8) through induction principle. In this paper, as alluded in the Introduction, we will use a different approach in establishing the solution forms of these systems.

3.1 Solution Form of System (S.1)

Consider the system

$$x_{n+1} = \frac{y_{n-3}y_n x_{n-2}}{y_{n-3}x_{n-2} + y_{n-3}y_n + y_n x_{n-2}}, \qquad y_{n+1} = \frac{y_{n-2}x_{n-1}}{2y_{n-2} + x_{n-1}}, \quad n \in \mathbb{N}_0.$$

Using the substitution defined in (5), and after some transpositions, we get the following transformed system of equations:

$$\forall n \in \mathbb{N}_0 : \begin{cases} w_{n+1} = z_n + w_{n-2} + z_{n-3} \iff z_n + z_{n-3} = w_{n+1} - w_{n-2}, \\ z_{n+1} = 2w_{n-1} + z_{n-2} \qquad \iff \qquad w_{n-1} = \dfrac{1}{2}\{z_{n+1} - z_{n-2}\}. \end{cases}$$

Eliminating the first variable yields the following one-dimensional difference equation $z_{n+3} = 4z_n + z_{n-3}$, for all $n \in \mathbb{N}_0$. Replacing n by $3n - i$, where $i = 0, 1, 2$, and then iterating the right hand side of the resulting equation, we get

$$\begin{aligned} \forall n \in \mathbb{N}_0 : \quad z_{3(n+1)-i} &= 4z_{3n-i} + z_{3(n-1)-i} \\ &= 4\left\{4z_{3(n-1)-i} + z_{3(n-2)-i}\right\} + z_{3(n-1)-i} \\ &= 17z_{3(n-1)+i} + 4z_{3(n-2)-i} \\ &= 17\left\{4z_{3(n-2)-i} + z_{3(n-3)-i}\right\} + 4z_{3(n-2)-i} \\ &= 72z_{3n-i} + 17z_{3(n-1)-i} \\ &\vdots \\ &= g_{n+1}z_{3-i} + g_n z_{-i}, \end{aligned}$$

where $\{g_n\}_0^\infty = \{0, 1, 4, 17, 72, 305, 1292, 5473, \ldots\}$ is Sequence No. A001076 in O.E.I.S. [21]). Now, referring to equation $z_{3(n+1)-i} = g_{n+1}z_{3-i} + g_n z_{-i}$, we have, in view of the substitution defined in (5) and the expression for y_{3-i} computed using the original system (S.1),

$$\forall n \in \mathbb{N}_0: \quad y_{3(n+1)-i} = \left\{ \frac{g_{n+1}}{y_{3-i}} + \frac{g_n}{y_{-i}} \right\}^{-1} = \left\{ \frac{g_{n+1}(2y_{-i} + x_{1-i})}{y_{-i}x_{1-i}} + \frac{g_n}{y_{-i}} \right\}^{-1}$$
$$= \frac{y_{-i}x_{1-i}}{2g_{n+1}y_{-i} + \{g_{n+1} + g_n\}x_{1-i}}, \qquad i = 0, 1, 2.$$

For the first few terms of the solution $\{y_n\}_1^\infty$, one can check that the formula works well for $i = 1, 2$. However, for $i = 0$, we find that the solution form for y_{3n} has an x_1 term. Hence, the exact form for y_{3n} must be given by, in view of the expression for x_1 computed using the original system,

$$\forall n \in \mathbb{N}_0: \quad y_{3(n+1)} = \left\{ \frac{2g_{n+1}}{x_1} + \frac{g_{n+1} + g_n}{y_0} \right\}^{-1}$$
$$= \left\{ \frac{2g_{n+1}(y_{-3}x_{-2} + y_{-3}y_0 + y_0x_{-2})}{y_{-3}y_0x_{-2}} + \frac{g_{n+1} + g_n}{y_0} \right\}^{-1}$$
$$= \frac{y_{-3}y_0x_{-2}}{\{3g_{n+1} + g_n\}y_{-3}x_{-2} + 2g_{n+1}y_{-3}y_0 + 2g_{n+1}y_0x_{-2}}.$$

Now, with the formula for y_n at hand, we can compute for the solution form of x_n through the equation $w_{n-1} = \frac{1}{2}\{z_{n+1} - z_{n-2}\}$. To do this, we replace n by $3n + 2 - i$ where $i = 0, 1, 2$ and then use the substitution defined in (5) to obtain

$$\forall n \in \mathbb{N}_0: \quad x_{3n+1-i} = 2\left\{ \frac{1}{y_{3(n+1)-i}} - \frac{1}{y_{3n-i}} \right\}^{-1}.$$

For $i = 0$, we have

$$\forall n \in \mathbb{N}_0: \; x_{3n+1} = 2\left[\frac{\{3g_{n+1} + g_n\}y_{-3}x_{-2} + 2g_{n+1}y_{-3}y_0 + 2g_{n+1}y_0x_{-2}}{y_{-3}y_0x_{-2}} \right.$$
$$\left. - \frac{\{3g_n + g_{n-1}\}y_{-3}x_{-2} + 2g_ny_{-3}y_0 + 2g_ny_0x_{-2}}{y_{-3}y_0x_{-2}} \right]^{-1}$$
$$= 2\left[\frac{\{3g_{n+1} - 2g_n - g_{n-1}\}y_{-3}x_{-2}}{y_{-3}y_0x_{-2}} \right]^{-1}$$
$$+ 2\left[\frac{2\{g_{n+1} - g_n\}y_{-3}y_0 + 2\{g_{n+1} - g_n\}y_0x_{-2}}{y_{-3}y_0x_{-2}} \right]^{-1}.$$

Meanwhile, for $i = 1, 2$, we have

$$\begin{aligned}\forall n \in \mathbb{N}_0 : \; x_{3n+1-i} &= 2\left[\frac{1}{y_{3(n+1)-i}} - \frac{1}{y_{3n-i}}\right]^{-1} \\ &= 2\left[\frac{2g_{n+1}y_{-i} + \{g_{n+1} + g_n\}x_{1-i}}{y_{-i}x_{1-i}}\right. \\ &\qquad \left. - \frac{2g_n y_{-i} + \{g_n + g_{n-1}\}x_{1-i}}{y_{-i}x_{1-i}}\right]^{-1} \\ &= \frac{y_{-i}x_{1-i}}{\{g_{n+1} - g_n\}y_{-i} + \frac{1}{2}\{g_{n+1} - g_{n-1}\}x_{1-i}}.\end{aligned}$$

Now, in reference to Lemma 1, we finally have the following theorem.

Theorem 1 *Let* $\{x_n\}_{-2}^{0}$ *and* $\{y_n\}_{-3}^{0}$ *be non-zero real numbers such that* $y_{-1}/x_0,\ y_{-2}/y_{-1} \notin \{-f_{3n}/f_{3n+1}\}_1^{\infty}$ *and* $(y_{-3}y_0 + y_0x_{-2})/y_{-3}x_{-2} \notin \{-f_{3n+2}/f_{3n+1}\}_0^{\infty}$. *Then, every solution* $\{(x_n, y_n)\}_1^{\infty}$ *of system (S.1) takes the form*

$$x_{3n+1-i} = \begin{cases} \dfrac{y_{-3}y_0x_{-2}}{f_{3n+2}y_{-3}x_{-2} + f_{3n+1}y_{-3}y_0 + f_{3n+1}y_0x_{-2}}, & \text{for } i = 0, \\ \dfrac{y_{-i}x_{1-i}}{f_{3n+1}y_{-i} + f_{3n}x_{1-i}}, & \text{for } i = 1, 2, \end{cases}$$

and

$$y_{3(n+1)-i} = \begin{cases} \dfrac{y_{-3}y_0x_{-2}}{f_{3n+4}y_{-3}x_{-2} + f_{3n+3}y_{-3}y_0 + f_{3n+3}y_0x_{-2}}, & \text{for } i = 0, \\ \dfrac{y_{-i}x_{1-i}}{f_{3n+3}y_{-i} + f_{3n+2}x_{1-i}}, & \text{for } i = 1, 2, \end{cases}$$

where $\{f_n\}_0^{\infty}$ *is the Fibonacci sequence.*

Note that the Fibonacci sequence grows to infinity as n increases without bound. As a consequence and in view of the previous theorem, we see that every solution of system (S.1) converges to zero as n goes to infinity. The following example illustrates this observation.

Example 1 Figure 1 illustrates the long term dynamics of system (S.1) with random initial values taken from the unit interval [0,1].

3.2 Solution Form of System (S.2)

Consider the system

$$x_{n+1} = \frac{y_{n-3}y_nx_{n-2}}{y_{n-3}x_{n-2} - y_{n-3}y_n + y_nx_{n-2}}, \qquad y_{n+1} = \frac{y_{n-2}x_{n-1}}{2y_{n-2} + x_{n-1}}, \quad n \in \mathbb{N}_0.$$

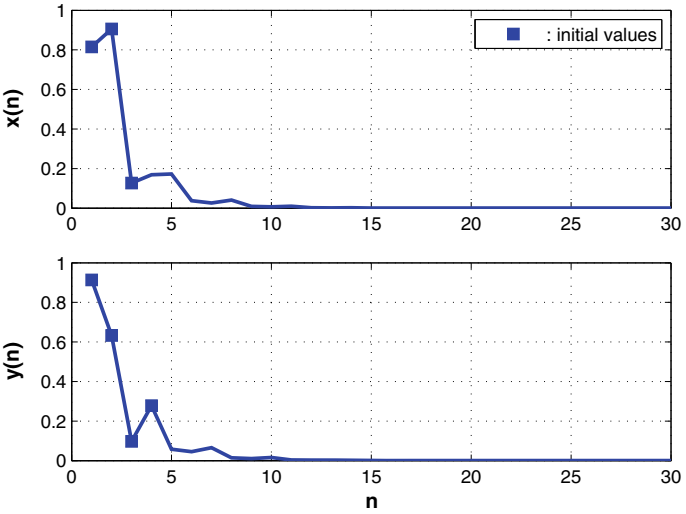

Fig. 1 Behavior of a particular solution of system (S.1)

Using the substitution defined in (5), we can transform the above system into

$$\forall n \in \mathbb{N}_0 : \begin{cases} w_{n+1} = z_n - w_{n-2} + z_{n-3} \iff z_n + z_{n-3} = w_{n+1} + w_{n-2}, \\ z_{n+1} = 2w_{n-1} + z_{n-2} \qquad \iff \qquad w_{n-1} = \dfrac{1}{2}\{z_{n+1} - z_{n-2}\}. \end{cases}$$

Hence, $z_n + z_{n-3} = w_{n+1} + w_{n-2} = \frac{1}{2}\{z_{n+3} - z_n\} + \frac{1}{2}\{z_n - z_{n-3}\} = \frac{1}{2}\{z_{n+3} - z_{n-3}\}$. Therefore, $2z_n + 2z_{n-3} = z_{n+3} - z_{n-3}$ or equivalently, $z_{n+3} + z_n = 3\{z_n + z_{n-3}\}$. Putting $v_n := z_n + z_{n-3}$, we get $v_{n+3} = 3v_n$. Replacing n by $3n - i$, where $i = 0, 1, 2$, and then iterating the right hand side (RHS) of the resulting equation leads to

$$\forall n \in \mathbb{N}_0 : \quad v_{3(n+1)-i} = 3^n v_{3-i}, \quad i = 0, 1, 2.$$

This in turn will give us the equation $v_{3(n+1)-i} = z_{3(n+1)-i} + z_{3n-i}$ which can be rewritten as, after replacing n by j, $z_{3(j+1)-i} + z_{3j-i} = 3^j v_{3-i}$. By multiplying both sides of the latter equation by $(-1)^j$ and then summing it up from 0 to $n - 1$, we get

$$\begin{aligned} (-1)^{n-1} z_{3n-i} + z_{-i} &= \sum_{j=0}^{n-1} (-1)^j \{z_{3(j+1)-i} + z_{3j-i}\} = \sum_{j=0}^{n-1} (-3)^j v_{3-i} \\ &= v_{3-i} \left\{ \frac{(-3)^n - 1}{-4} \right\} = \frac{(-1)^{n-1}}{2} \left\{ \frac{3^n - (-1)^n}{2} \right\} v_{3-i}. \end{aligned}$$

The above relation can be rearranged to get

$$\forall n \in \mathbb{N}_0 : \quad z_{3n-i} = (-1)^n z_{-i} + \frac{1}{2} u_n v_{3-i}, \qquad i = 0, 1, 2.$$

Hence, in reference to Eq. (5) together with the definition of v_n, we have

$$\forall n \in \mathbb{N}_0 : \quad y_{3n-i} = \frac{2y_{3-i}\, y_{-i}}{2(-1)^n y_{3-i} + u_n (y_{3-i} + y_{-i})}, \qquad i = 0, 1, 2.$$

Since the RHS of the above equation has the term y_{3-i}, we further simplify it as follows. Noting that $y_{3-i} = y_{-i} x_{1-i}/(2y_{-i} + x_{1-i})$ for all $i = 0, 1, 2$, we have

$$\begin{aligned} \forall n \in \mathbb{N}_0 : \quad y_{3n-i} &= \frac{2y_{-i}}{\{2(-1)^n + u_n\} + y_{-i} u_n \left\{ \dfrac{2y_{-i} + x_{1-i}}{y_{-i} x_{1-i}} \right\}} \\ &= \frac{y_{-i} x_{1-i}}{u_n y_{-i} + \{u_n + (-1)^n\} x_{1-i}}, \qquad i = 0, 1, 2. \end{aligned}$$

Notice that the RHS of the above equation depends on its initial values except when $i = 0$. In this case, however, we can express y_{3n} as follows:

$$\begin{aligned} \forall n \in \mathbb{N}_0 : \quad y_{3n} &= \frac{y_0}{\{u_n + (-1)^n\} + u_n y_0 / x_1} \\ &= \frac{y_0}{\{u_n + (-1)^n\} + u_n y_0 \left\{ \dfrac{y_{-3} x_{-2} - y_{-3} y_0 + y_0 x_{-2}}{y_{-3} y_0 x_{-2}} \right\}} \\ &= \frac{y_{-3} y_0 x_{-2}}{\{u_n + (-1)^n\} y_{-3} x_{-2} + u_n \{y_{-3} x_{-2} - y_{-3} y_0 + y_0 x_{-2}\}} \\ &= \frac{y_{-3} y_0 x_{-2}}{\{2u_n + (-1)^n\} y_{-3} x_{-2} - u_n y_{-3} y_0 + u_n y_0 x_{-2}}, \qquad i = 0, 1, 2. \end{aligned}$$

Now, on the other hand, by applying the same approach to $w_{n+1} + w_{n-2} = z_n + z_{n-3}$, we get

$$\begin{aligned} (-1)^{n-1} w_{3n+1-i} + w_{1-i} &= \sum_{j=0}^{n-1} (-1)^j \{w_{3(j+1)+1-i} + w_{3j+1-i}\} \\ &= \sum_{j=0}^{n-1} (-1)^j \{z_{3(j+1)-i} + z_{3j-i}\} \\ &= (-1)^{n-1} z_{3n-i} + z_{-i}. \end{aligned}$$

Using our result for the phase variable w_n, we get

$$\forall n \in \mathbb{N}_0 : \quad w_{3n+1-i} = (-1)^n w_{1-i} + \frac{1}{2} u_n v_{3-i}, \qquad i = 0, 1, 2.$$

In view of the substitution (5), we obtain

$$\forall n \in \mathbb{N}_0 : \quad x_{3n+1-i} = \frac{2y_{3-i}y_{-i}x_{1-i}}{2(-1)^n y_{3-i}y_{-i} + u_n(y_{3-i} + y_{-i})x_{1-i}}, \qquad i = 0, 1, 2.$$

The above expression can be further simplified, as we did in our previous result, in the following manner

$$\begin{aligned}\forall n \in \mathbb{N}_0 : \quad x_{3n+1-i} &= \frac{2y_{-i}x_{1-i}}{\{2(-1)^n y_{-i} + u_n x_{1-i}\} + u_n y_{-i}x_{1-i}/y_{3-i}}, \\ &= \frac{2y_{-i}x_{1-i}}{\{2(-1)^n y_{-i} + u_n x_{1-i}\} + u_n y_{-i}x_{1-i}\left\{\dfrac{2y_{-i} + x_{1-i}}{y_{-i}x_{1-i}}\right\}}, \\ &= \frac{y_{-i}x_{1-i}}{\{u_n + (-1)^n\}\, y_{-i} + u_n x_{1-i}}, \qquad i = 0, 1, 2,\end{aligned}$$

where at $i = 0$, we have, for all $n \in \mathbb{N}_0$

$$\begin{aligned}x_{3n+1} &= \frac{y_0}{u_n + \{u_n + (-1)^n\}\, y_0/x_1} \\ &= \frac{y_{-3}y_0x_{-2}}{\{2u_n + (-1)^n\}\, y_{-3}x_{-2} - \{u_n + (-1)^n\}\, y_{-3}y_0 + \{u_n + (-1)^n\}\, y_0x_{-2}}.\end{aligned}$$

The following theorem summarizes our previous discussion.

Theorem 2 *Let $\{x_n\}_{-2}^0$ and $\{y_n\}_{-3}^0$ be non-zero real numbers such that they satisfy the conditions that*

$$\frac{y_{-1}}{x_0}, \ \frac{y_{-2}}{y_{-1}} \notin \left(\left\{-\frac{u_n}{u_n + (-1)^n}\right\}_1^\infty \cup \left\{-\frac{u_n + (-1)^n}{u_n}\right\}_1^\infty\right)$$

and

$$\frac{-y_{-3}y_0 + y_0x_{-2}}{y_{-3}x_{-2}} \notin \left(\left\{-\frac{3^n}{u_n + (-1)^n}\right\}_0^\infty \cup \left\{-\frac{3^n}{u_n}\right\}_1^\infty\right).$$

Then, every solution $\{(x_n, y_n)\}_1^\infty$ of system (S.2) takes the form

$$x_{3n+1-i} = \begin{cases} \dfrac{y_{-3}y_0x_{-2}}{3^n y_{-3}x_{-2} - \{u_n + (-1)^n\}\, y_{-3}y_0 + \{u_n + (-1)^n\}\, y_0x_{-2}}, & \text{for } i = 0 \\ \dfrac{y_{-i}x_{1-i}}{\{u_n + (-1)^n\}\, y_{-i} + u_n x_{1-i}} & \text{for } i = 1, 2, \end{cases}$$

and

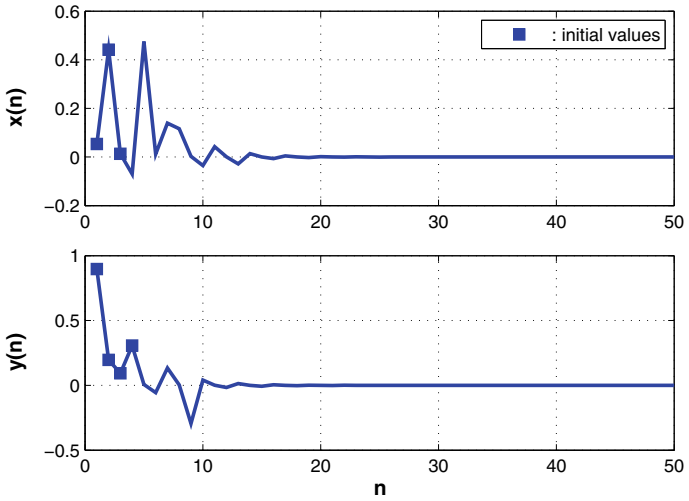

Fig. 2 Behavior of a particular solution of system (S.2)

$$y_{3n-i} = \begin{cases} \dfrac{y_{-3}y_0x_{-2}}{3^n y_{-3}x_{-2} - u_n y_{-3}y_0 + u_n y_0 x_{-2}}, & \text{for } i = 0 \\ \dfrac{y_{-i}x_{1-i}}{u_n y_{-i} + \{u_n + (-1)^n\}x_{1-i}} & \text{for } i = 1, 2, \end{cases}$$

where $\{u_n\}_0^\infty$ *is the sequence defined in Proposition 1.*

Note that the sequence $\{u_n\}_0^\infty$ grows indefinitely as n increases without bound. Hence, it follows immediately from the above theorem that every solution (x_n, y_n) of system (S.3) goes to (0, 0) as $n \to \infty$.

For confirming the above statement we provide the following example.

Example 2 Figure 2 illustrates the long term dynamics of system (S.2) with random initial values taken from the unit interval [0,1].

3.3 Solution Form of System (S.3)

Consider the system

$$x_{n+1} = \frac{y_{n-3}y_n x_{n-2}}{y_{n-3}x_{n-2} + y_{n-3}y_n - y_n x_{n-2}}, \qquad y_{n+1} = \frac{y_{n-2}x_{n-1}}{2y_{n-2} + x_{n-1}}, \quad n \in \mathbb{N}_0.$$

With the substitution (5), the above system can be written equivalently as follows

$$\forall n \in \mathbb{N}_0 : \begin{cases} w_{n+1} = z_n + w_{n-2} - z_{n-3} \Longleftrightarrow z_n - z_{n-3} = w_{n+1} - w_{n-2}, \\ z_{n+1} = 2w_{n-1} + z_{n-2} \quad \Longleftrightarrow \quad w_{n-1} = \dfrac{1}{2}\{z_{n+1} - z_{n-2}\}. \end{cases}$$

Eliminating the phase variable z_n yields $2w_{n-2} = w_{n+1} - w_{n-2}$, which is equivalent to $w_{n+1} = 3w_{n-2}$. Replacing n by $3n - 1 - i$ where $i = 0, 1, 2$ and then iterating the RHS of this equation, we obtain

$$\forall n \in \mathbb{N}_0 : \quad w_{3n-i} = 3^n w_{-i}, \quad i = 0, 1, 2,$$

or equivalently, with reference to Eq. (5),

$$\forall n \in \mathbb{N}_0 : \quad x_{3n-i} = \left(\frac{1}{3}\right)^n x_{-i}, \quad i = 0, 1, 2.$$

Notice that, at $n = 1$ and $i = 2$, we have $x_1 = \frac{1}{3}x_{-2}$ which obviously differs from the expression

$$x_1 = \frac{y_{-3}y_0x_{-2}}{y_{-3}x_{-2} + y_{-3}y_0 - y_0x_{-2}}$$

obtained from the original system (S.3). These two expressions are, in fact, equivalent. To see this, we use the second equation in system (S.3). That is, we have, at $n = -1$,

$$\begin{aligned} y_0 = \frac{y_{-3}x_{-2}}{2y_{-3} + x_{-2}} &\Longleftrightarrow 2y_{-3}y_0 = y_{-3}x_{-2} - y_0x_{-2} \\ &\Longleftrightarrow 3y_{-3}y_0 = y_{-3}x_{-2} + y_{-3}y_0 - y_0x_{-2} \\ &\Longleftrightarrow \frac{y_{-3}y_0}{y_{-3}x_{-2} + y_{-3}y_0 - y_0x_{-2}} = \frac{1}{3} \\ &\Longleftrightarrow \frac{1}{3}x_{-2} = \frac{y_{-3}y_0x_{-2}}{y_{-3}x_{-2} + y_{-3}y_0 - y_0x_{-2}}. \end{aligned} \tag{6}$$

Furthermore, we find that the exact solution form for x_n is given by

$$\forall n \in \mathbb{N}_0 : \quad x_{3n-i} = \begin{cases} \left(\dfrac{1}{3}\right)^n x_{-i}, & \text{for } i = 0, 1, \\ \left(\dfrac{1}{3}\right)^{n-1} \left\{\dfrac{y_{-3}y_0x_{-2}}{y_{-3}x_{-2} + y_{-3}y_0 - y_0x_{-2}}\right\}, & \text{for } i = 2. \end{cases}$$

Now, on the other hand, since $z_{n+1} = 2w_{n-1} + z_{n-2}$, then, after replacing n by $3n + 1 - i$, we have

$$\forall n \in \mathbb{N}_0 : \quad z_{3(n+1)-i-1} = 2 \cdot 3^n w_{-i} + z_{3n-i-1}, \quad i = 0, 1.$$

Again, iterating the RHS of the above equation, we get

$$\begin{aligned}\forall n \in \mathbb{N}_0: \quad z_{3(n+1)-1-i} &= 2\cdot 3^n w_{-i} + z_{3n-i-1}\\ &= 2\cdot 3^n w_{-i} + 2\cdot 3^{n-1} w_{-i} + z_{3(n-1)-i-1}\\ &= 2\cdot 3^n w_{-i} + 2\cdot 3^{n-1} w_{-i} + 2\cdot 3^{n-2} w_{-i} + z_{3(n-2)-i-1}\\ &\vdots\\ &= 2\cdot 3^n w_{-i} + 2\cdot 3^{n-1} w_{-i} + \cdots + 2w_{-i} + z_{-i-1}\\ &= z_{-i-1} + 2w_{-i}\left\{\frac{3^{n+1}-1}{2}\right\}, \quad i = 0, 1.\end{aligned}$$

Hence, using the substitution defined in (5) and upon replacing $n+1$ by n, the above relation can be written equivalently in the form

$$\forall n \in \mathbb{N}_0: \quad y_{3n-1-i} = \frac{y_{-1-i}x_{-i}}{2\left\{\frac{1}{2}(3^n-1)\right\}y_{-1-i} + x_{-i}}, \qquad i = 0, 1.$$

Meanwhile, to obtain for the form of solution for y_{3n}, we proceed as follows. Note that, in a similar argument as above, we have

$$\forall n \in \mathbb{N}_0: \quad z_{3(n+1)} = 2\cdot 3^n w_1 + z_{3n} = z_0 + (3^n - 1)w_1.$$

Hence, in view of the substitution defined in (5) and the expression for x_1, we have

$$\begin{aligned}\forall n \in \mathbb{N}_0: \quad y_{3n} &= \left\{\frac{1}{y_0} + \frac{(3^n-1)}{x_1}\right\}^{-1}\\ &= \left\{\frac{1}{y_0} + \frac{(3^n-1)(y_{-3}x_{-2} + y_{-3}y_0 - y_0x_{-2})}{y_{-3}y_0x_{-2}}\right\}^{-1}\\ &= \frac{y_{-3}y_0x_{-2}}{3^n y_{-3}x_{-2} + (3^n-1)y_{-3}y_0 - (3^n-1)y_0x_{-2}}.\end{aligned}$$

Combining all of our results exhibited above, we arrive at the following theorem.

Theorem 3 *Let $\{x_n\}_{-2}^{0}$ and $\{y_n\}_{-3}^{0}$ be non-zero real numbers such that they satisfy the conditions that $y_{-3}x_{-2} + y_{-3}y_0 - y_0x_{-2} \neq 0$, $y_{-2}, x_{-1}, y_{-1}/x_0 \notin \{-1/2t_n\}_1^\infty$ and $(y_{-3}y_0 - y_0x_{-2})/y_{-3}x_{-2} \notin \{-(2t_n+1)/2t_n\}_1^\infty$. Then, every solution $\{(x_n, y_n)\}_1^\infty$ of system (S.3) takes the form*

$$\forall n \in \mathbb{N}_0: \quad x_{3n-i} = \begin{cases}\left(\frac{1}{3}\right)^n x_{-i}, & \text{for } i = 0, 1,\\ \left(\frac{1}{3}\right)^{n-1}\left\{\frac{y_{-3}y_0x_{-2}}{y_{-3}x_{-2} + y_{-3}y_0 - y_0x_{-2}}\right\}, & \text{for } i = 2.\end{cases}$$

and

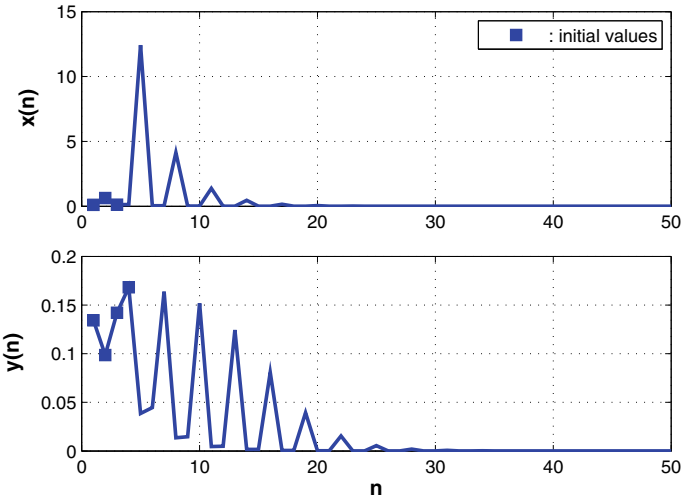

Fig. 3 Behavior of a particular solution of system (S.3)

$$\forall n \in \mathbb{N}_0 : \quad y_{3n-i} = \begin{cases} \dfrac{y_{-i}x_{1-i}}{2t_n y_{-i} + x_{1-i}}, & \text{for } i = 2, 1, \\ \dfrac{y_{-3}y_0x_{-2}}{\{2t_n+1\}y_{-3}x_{-2} + 2t_n y_{-3}y_0 - 2t_n y_0 x_{-2}}, & \text{for } i = 0, \end{cases},$$

where $\{t_n\}_0^\infty := \{\frac{1}{2}(3^n - 1)\}_0^\infty = \{0, 1, 4, 13, 40, 121, 364, 1093, 3280, \ldots\}$.

Remark 4 We emphasize that the same approach employed in the previous section can be used to formulate the solution form for the phase variable y_n. However, we intended to use a different method here to give the readers a different way to derive the solution form.

It is evident from Theorem 3 that $(x_n, y_n) \to (0, 0)$ as $n \to \infty$ (see example below).

Example 3 Figure 3 illustrates the long term dynamics of system (S.3) with random initial values from [0, 1].

3.4 Solution Form of System (S.4)

Consider the system

$$x_{n+1} = \frac{y_{n-3}y_n x_{n-2}}{y_{n-3}x_{n-2} + y_{n-3}y_n + y_n x_{n-2}}, \qquad y_{n+1} = \frac{y_{n-2}x_{n-1}}{2y_{n-2} - x_{n-1}}, \quad n \in \mathbb{N}_0.$$

Using the substitution (5), the above equations are transformed into

$$\forall n \in \mathbb{N}_0 : \begin{cases} w_{n+1} = z_n + w_{n-2} + z_{n-3} \iff z_n + z_{n-3} = w_{n+1} - w_{n-2}, \\ z_{n+1} = 2w_{n-1} - z_{n-2} \quad \iff \quad w_{n-1} = \dfrac{1}{2}\{z_{n+1} + z_{n-2}\}. \end{cases}$$

These relations imply that $w_{n+1} = 3w_{n-2}$. In reference to the previous case, we readily have $w_{3n-i} = 3^n w_{-i}$, for each $i = 0, 1, 2$, for all $n \in \mathbb{N}_0$. Therefore, we have

$$\forall n \in \mathbb{N}_0 : \quad x_{3n-i} = \left(\frac{1}{3}\right)^n x_{-i}, \quad i = 0, 1, 2.$$

Obviously, the above formula works well for $i = 0, 1$. However, for $i = 2$, the RHS of the equation seems to be different from the form of solution, for instance of x_1, obtained from the original system (S.4). To fix this problem, we must compute for the exact form of x_1 from the original system. That is, we have

$$x_1 = \frac{y_{-3}y_0x_{-2}}{y_{-3}x_{-2} + y_{-3}y_0 + y_0x_{-2}}.$$

Hence, the exact solution form for x_n is given by

$$\forall n \in \mathbb{N}_0 : \quad x_{3n-i} = \begin{cases} \left(\dfrac{1}{3}\right)^n \left\{\dfrac{y_{-3}y_0x_{-2}}{y_{-3}x_{-2} + y_{-3}y_0 + y_0x_{-2}}\right\}, & \text{for } i = 2, \\ \left(\dfrac{1}{3}\right)^n x_{-i}, & \text{for } i = 1, 0. \end{cases}$$

Now, consider the equation $z_{n+1} + z_{n-2} = 2w_{n-1}$. Note that the homogeneous equation $z_{n+1} = -z_{n-2}$ has the solution $z_{3n-i} = (-1)^n z_{-i}$, where $i = 0, 1, 2$, for all $n \in \mathbb{N}_0$. To see this, we can simply iterate the RHS of the equation $z_{3n-i} = -z_{3(n-1)-i}$ as follows: $z_{3n-i} = -z_{3(n-1)-i} = (-1)^2 z_{3(n-2)-i} = (-1)^3 z_{3(n-3)-i} = \cdots = (-1)^n z_{-i}$. Meanwhile, the non-homogeneous case $z_{n+1} + z_{n-2} = 2w_{n-1}$ can be transformed, upon replacing n by $3n - i + 1$ as in the previous case, as follows:

$$z_{3(n+1)-1-i} = 2w_{3n-i} - z_{3n-i-1} = \cdots = -z_{-i-1} + 2w_{-i} \sum_{j=0}^{n} (-3)^j, \quad i = 0, 1, 2.$$

This implies that, upon replacing $n + 1$ by n together with the substitution (5) and Proposition 1, we have

$$\begin{aligned} \forall n \in \mathbb{N}_0 : \quad y_{3n-i-1} &= \left[-\frac{1}{y_{-i-1}} + \frac{(-1)^{n-1}}{x_{-i}} \left\{\frac{3^n - (-1)^n}{2}\right\}\right]^{-1} \\ &= \frac{y_{-1-i}x_{-i}}{y_{-i-1}u_n + (-1)^n x_{-i}}, \qquad i = 0, 1, 2. \end{aligned}$$

It can be verified that the formula above gives exact values for $i = 1, 2$; however, a different form will be obtained for y_{3n}. Even so, the solution form for y_{3n} can be established in a similar fashion as above. Note that

$$\begin{aligned} z_{3n} &= 2w_{3n-2} - z_{3n-3} = \cdots = -z_0 + 2w_1 \sum_{j=0}^{n-1} (-3)^j \\ &= -z_0 + (-1)^{n-1} w_1 \left\{ \frac{3^n - (-1)^n}{2} \right\}, \end{aligned}$$

which, in turn, can be rewritten as, upon using the substitution defined in (5) and in reference to Proposition 1,

$$y_{3n} = \left\{ \frac{u_n}{x_1} + \frac{(-1)^n}{y_0} \right\}^{-1}.$$

Thus, in view of the form for x_1, we finally have

$$\begin{aligned} \forall n \in \mathbb{N}_0 : \quad y_{3n} &= \left\{ \frac{u_n}{x_1} + \frac{(-1)^n}{y_0} \right\}^{-1} \\ &= \left\{ \frac{u_n (y_{-3}x_{-2} + y_{-3}y_0 + y_0 x_{-2})}{y_{-3}y_0 x_{-2}} + \frac{(-1)^n}{y_0} \right\}^{-1} \\ &= \frac{y_{-3}y_0 x_{-2}}{\{u_n + (-1)^n\} y_{-3}x_{-2} + u_n y_{-3}y_0 + u_n y_0 x_{-2}}. \end{aligned}$$

In conclusion, we have the following theorem.

Theorem 4 *Let $\{x_n\}_{-2}^0$ and $\{y_n\}_{-3}^0$ be non-zero real numbers such that they satisfy the conditions that $y_{-3}x_{-2} + y_{-3}y_0 + y_0 x_{-2} \neq 0$, y_{-2}, x_{-1}, $y_{-1}/x_0 \notin \{-(-1)^n/u_n\}_1^\infty$ and $(y_{-3}y_0 + y_0 x_{-2})/y_{-3}x_{-2} \notin \{-[u_n + (-1)^n]/u_n\}_1^\infty$. Then, every solution $\{(x_n, y_n)\}_1^\infty$ of system (S.4) takes the form*

$$\forall n \in \mathbb{N}_0 : \quad x_{3n-i} = \begin{cases} \left(\frac{1}{3}\right)^n \left\{ \frac{y_{-3}y_0 x_{-2}}{y_{-3}x_{-2} + y_{-3}y_0 + y_0 x_{-2}} \right\}, & \text{for } i = 2, \\ \left(\frac{1}{3}\right)^n x_{-i}, & \text{for } i = 1, 0, \end{cases}$$

and

$$\forall n \in \mathbb{N}_0 : \quad y_{3n-i} = \begin{cases} \frac{y_{-i}x_{1-i}}{y_{-i}u_n + (-1)^n x_{1-i}}, & \text{for } i = 2, 1, \\ \frac{y_{-3}y_0 x_{-2}}{\{u_n + (-1)^n\} y_{-3}x_{-2} + u_n y_{-3}y_0 + u_n y_0 x_{-2}}, & \text{for } i = 0, \end{cases}$$

where $\{u_n\}_0^\infty$ is the sequence defined in Proposition 1.

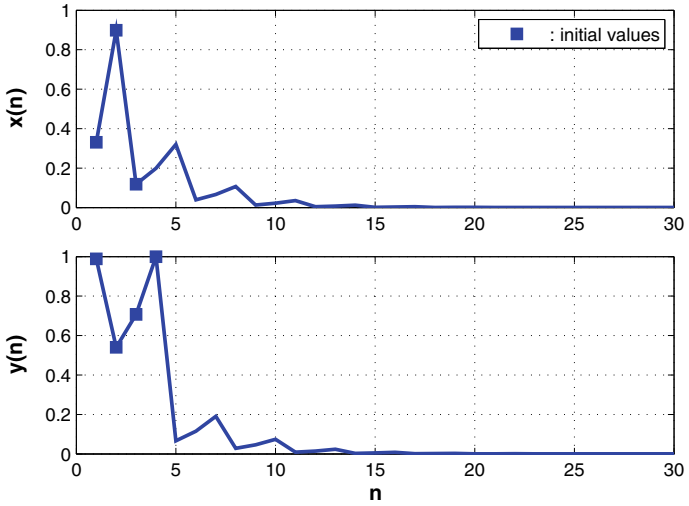

Fig. 4 The long term behavior of a particular solution of system (S.4)

Note that the sequence $\{(1/3)^n\}_0^\infty$ converges to zero as $n \to \infty$ while $\{u_n\}_0^\infty$ grows indefinitely. Hence, $\max\{x_n, y_n\} \to 0$ as $n \to \infty$. As an illustration of this observation, we provide the following example.

Example 4 Figure 4 illustrates a particular plot for the long term dynamics of system (S.4) with random initial values from [0, 1].

3.5 Solution Form of System (S.5)

Consider the system

$$x_{n+1} = \frac{y_{n-3}y_n x_{n-2}}{y_{n-3}x_{n-2} - y_{n-3}y_n - y_n x_{n-2}}, \qquad y_{n+1} = \frac{y_{n-2}x_{n-1}}{2y_{n-2} + x_{n-1}}, \quad n \in \mathbb{N}_0.$$

Using the substitution (5), the above equations are transformed into

$$\forall n \in \mathbb{N}_0 : \begin{cases} w_{n+1} = z_n - w_{n-2} - z_{n-3} \iff z_n - z_{n-3} = w_{n+1} + w_{n-2}, \\ z_{n+1} = 2w_{n-1} + z_{n-2} \qquad \iff \qquad w_{n-1} = \dfrac{1}{2}\{z_{n+1} - z_{n-2}\}. \end{cases}$$

From above equations, it follows that $2w_{n-1} = w_{n+1} + w_{n-2}$ or equivalently, $w_{n+1} = w_{n-2}$. Replacing n by $3n - 1 - i$ together with Eq. (5), where $i = 0, 1, 2$, we get

$$\forall n \in \mathbb{N}_0 : \quad x_{3n-i} = x_{3(n-1)-i} \iff x_{3n-i} = x_{-i}, \qquad i = 0, 1, 2.$$

Hence, $\{x_n\}_1^\infty$ is periodic with period 3. The formula we have obtained above works well for $i = 0, 1$. However, for $i = 2$, we will get the relation $x_{3n-2} = x_{-2}$. At $n = 1$, this equation will give us $x_1 = x_{-2}$, which is inconsistent with the form of x_1 obtained from the original system (S.5). This problem, however, can be fixed easily by computing for x_1. Thus, the exact solution form for x_n is given as follows:

$$\forall n \in \mathbb{N}_0 : \quad x_{3n-i} = \begin{cases} \dfrac{y_{-3}y_0x_{-2}}{y_{-3}x_{-2} - y_{-3}y_0 - y_0x_{-2}}, & \text{for } i = 2, \\ x_{-i}, & \text{for } i = 1, 0. \end{cases}$$

On the other hand, going back to the transformed equations, we have $z_{n+1} = 2w_{n-1} + z_{n-3}$. Replacing n by $3n - i - 1$ in this equation and then using the relation $w_{3n-i} = w_{-i}$, where $i = 0, 1$, we obtain $z_{3(n+1)-i-1} = 2w_{-i} + z_{3n-i-1}$ or equivalently, $z_{3(n+1)-i-1} - z_{3n-i-1} = 2w_{-i}$. Once again, replacing n by j and then summing up each side of the equation from 0 to $n - 1$, we get

$$\begin{aligned} z_{3n-i-1} - z_{-i-1} &= \sum_{j=0}^{n-1} \left\{ z_{3(j+1)-i-1} - z_{3j-i-1} \right\} \\ &= \sum_{j=0}^{n-1} 2w_{-i} = 2nw_{-i}, \qquad i = 0, 1. \end{aligned}$$

Hence, $z_{3n-i-1} = 2nw_{-i} + z_{-i-1}$ for each $i = 0, 1$, for all $n \in \mathbb{N}_0$. Thus, using Eq. (5), we get

$$\forall n \in \mathbb{N}_0 : \quad y_{3n-i-1} = \frac{y_{-i-1}x_{-i}}{2ny_{-i-1} + x_{-i}}, \qquad i = 0, 1.$$

Meanwhile, to find for the solution form for y_{3n}, we replace n by $3j - 1$ in the equation $2w_{n-1} = z_{n+1} - z_{n-2}$ and then sum up the resulting equation from 1 to n in order to get

$$z_{3n} - z_3 = \sum_{j=1}^{n} \{z_{3j} - z_{3(j-1)}\} = 2 \sum_{j=1}^{n} w_{3j-2}.$$

Using the substitution defined in (5), together with the computed form for x_{3n-2}, we can rewrite the above equation as follows:

$$\begin{aligned} \forall n \in \mathbb{N}_0 : \quad \frac{1}{y_{3n}} - \frac{1}{y_0} &= 2 \sum_{j=1}^{n} \frac{1}{x_{3j-2}} \\ &= \frac{2n(y_{-3}x_{-2} - y_{-3}y_0 - y_0x_{-2})}{y_{-3}y_0x_{-2}} \end{aligned}$$

which, when rearranged, becomes

$$\forall n \in \mathbb{N}_0: \quad y_{3n} = \left\{ \frac{2n(y_{-3}x_{-2} - y_{-3}y_0 - y_0x_{-2})}{y_{-3}y_0x_{-2}} + \frac{1}{y_0} \right\}^{-1} = \frac{y_{-3}y_0x_{-2}}{(2n+1)y_{-3}x_{-2} - 2ny_{-3}y_0 - 2ny_0x_{-2}}.$$

Our previous discussion proves the following theorem.

Theorem 5 *Let $\{x_n\}_{-2}^{0}$ and $\{y_n\}_{-3}^{0}$ be non-zero real numbers such that they satisfy the conditions that $y_{-3}x_{-2} - y_{-3}y_0 - y_0x_{-2} \neq 0$, $y_{-2}, x_{-1}, \ y_{-1}/x_0 \notin \{-1/2n\}_1^\infty$ and $(y_{-3}y_0 + y_0x_{-2})/y_{-3}x_{-2} \notin \{(2n+1)/2n\}_1^\infty$. Then, every solution $\{(x_n, y_n)\}_1^\infty$ of system (S.5) takes the form*

$$\forall n \in \mathbb{N}_0: \quad x_{3n-i} = \begin{cases} \dfrac{y_{-3}y_0x_{-2}}{y_{-3}x_{-2} - y_{-3}y_0 - y_0x_{-2}}, & \text{for } i = 2, \\ x_{-i}, & \text{for } i = 1, 0, \end{cases}$$

and

$$\forall n \in \mathbb{N}_0: \quad y_{3n-i} = \begin{cases} \dfrac{y_{-i}x_{1-i}}{2ny_{-i} + x_{1-i}}, & \text{for } i = 2, 1, \\ \dfrac{y_{-3}y_0x_{-2}}{(2n+1)y_{-3}x_{-2} - 2ny_{-3}y_0 - 2ny_0x_{-2}}, & \text{for } i = 0. \end{cases}$$

The following result is immediate from the above theorem.

Corollary 1 *Let $\{(x_n, y_n)\}_1^\infty$ be a solution of system (S.5). Then, the sequence $\{x_n\}_1^\infty$ is periodic with period 3 while the sequence $\{y_n\}_1^\infty$ converges to zero as n increases without bound.*

The following example illustrates the virtue of the previous corollary. Notice that the solution $\{x_n\}_1^\infty$ is periodic of period 3 while $\{y_n\}_1^\infty$ converges to zero as n goes to infinity.

Example 5 Figure 5 illustrates a particular plot for the long term dynamics of system (S.5) with random initial values from [0, 1].

3.6 Solution Form of System (S.6)

Consider the system

$$x_{n+1} = \frac{y_{n-3}y_nx_{n-2}}{y_{n-3}x_{n-2} + y_{n-3}y_n - y_nx_{n-2}}, \qquad y_{n+1} = \frac{y_{n-2}x_{n-1}}{2y_{n-2} - x_{n-1}}, \quad n \in \mathbb{N}_0.$$

In view of the substitution (5), the above equations are, therefore, equivalent to the following transformations:

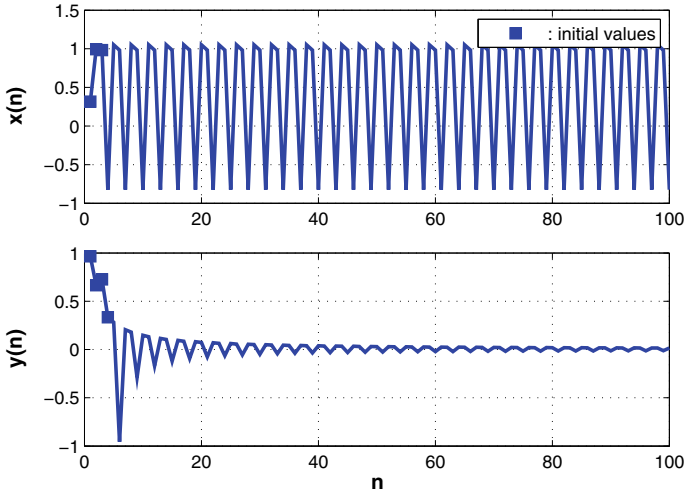

Fig. 5 The long term behavior of a particular solution of system (S.5)

$$\forall n \in \mathbb{N}_0 : \begin{cases} w_{n+1} = z_n + w_{n-2} - z_{n-3} \Longleftrightarrow z_n - z_{n-3} = w_{n+1} - w_{n-2}, \\ z_{n+1} = 2w_{n-1} - z_{n-2} \quad \Longleftrightarrow \quad w_{n-1} = \dfrac{1}{2}\{z_{n+1} + z_{n-2}\}. \end{cases}$$

Clearly, the above equations imply that, upon eliminating the phase variable w_n, $2z_n - 2z_{n-3} = z_{n+3} - z_{n-3}$, which can equivalently be written as $z_{n+3} - z_n = z_n - z_{n-3}$. Replacing n by $3j + i$ and then summing up the resulting equation from 0 to n, with $i = 1, 2$, we get

$$\begin{aligned} \forall n \in \mathbb{N}_0 : \quad z_{3(n+1)+i} - z_{-i} &= \sum_{j=0}^{n} \{z_{3(j+1)+i} - z_{3j+i}\} \\ &= \sum_{j=0}^{n} \{z_{3j+i} - z_{3(j-1)+i}\} \\ &= z_{3n+i} - z_{-3+i}, \qquad i = 1, 2. \end{aligned}$$

Rearranging the resulting equation obtained above, we get the relation $z_{3(n+1)+i} - z_{3n+i} = z_{-i} - z_{-3+i}$. Now, replacing n by j and then summing up each side of this equation from 0 to n, we obtain

$$\begin{aligned} \forall n \in \mathbb{N}_0 : \quad z_{3(n+1)+i} - z_i &= \sum_{j=0}^{n} \{z_{3(j+1)+i} - z_{3j+i}\} = \sum_{j=0}^{n} \{z_{-i} - z_{-3+i}\} \\ &= (n+1)[z_{-i} - z_{-3+i}], \qquad i = 1, 2. \end{aligned}$$

Once again, after replacing $n+1$ by n, the above equation can be rearranged to obtain

$$\forall n \in \mathbb{N}_0: \quad z_{3n+i} = (n+1)z_i - nz_{-3+i}, \qquad i = 1, 2.$$

Therefore, in view of Eq. (5), we have

$$\forall n \in \mathbb{N}_0: \quad y_{3n+i} = \frac{y_i y_{-3+i}}{(n+1)y_{-3+i} - ny_i}, \qquad i = 1, 2. \tag{7}$$

Observe from the above formula that the RHS is dependent from y_1 and y_2 for $i = 1, 2$, respectively. So, in order to establish the exact expression for the n-th term solution y_n, we need to compute for y_1 and y_2. Thus, in view of the original system (S.6), we have, for $i = 1, 2$,

$$y_i = \frac{y_{i-3}x_{i-2}}{2y_{i-3} - x_{i-2}}.$$

Hence, we now have

$$\begin{aligned}
\forall n \in \mathbb{N}_0: \quad y_{3n+i} &= \frac{y_i y_{-3+i}}{(n+1)y_{-3+i} - ny_i} = \frac{1}{\dfrac{n+1}{y_i} - \dfrac{n}{y_{-3+i}}} \\
&= \frac{1}{\dfrac{(n+1)(2y_{i-3} - x_{i-2})}{y_{i-3}x_{i-2}} - \dfrac{n}{y_{i-3}}} \\
&= \frac{y_{i-3}x_{i-2}}{2(n+1)y_{i-3} - (2n+1)x_{i-2}}, \qquad i = 1, 2.
\end{aligned}$$

Now, for the terms of the form y_{3n}, we replace n by $3j$ on both sides of the equation $z_{n+3} - z_n = z_n - z_{n-3}$ and then sum up the resulting equation from 1 to n so that we will have

$$\begin{aligned}
\forall n \in \mathbb{N}_0: \quad z_{3(n+1)} - z_3 &= \sum_{j=1}^{n} \{z_{3(j+1)} - z_{3j}\} = \sum_{j=1}^{n} \{z_{3j} - z_{3(j-1)}\} \\
&= z_{3n} - z_0.
\end{aligned}$$

Again, the above equation can be rearranged to obtain $z_{3(n+1)} - z_{3n} = z_3 - z_0$, for all $n \in \mathbb{N}_0$. Replacing n by j and then summing up the resulting equation from 0 to n, we get

$$\begin{aligned}
\forall n \in \mathbb{N}_0: \quad z_{3(n+1)} - z_0 &= \sum_{j=0}^{n} \left\{z_{3(j+1)} - z_{3j}\right\} = \sum_{j=0}^{n} \{z_3 - z_0\} \\
&= (n+1)(z_3 - z_0).
\end{aligned}$$

In reference to the substitution defined in (5), we can rewrite the equation to obtain

$$\forall n \in \mathbb{N}_0: \quad y_{3(n+1)} = \left\{ \frac{n+1}{y_3} - \frac{n}{y_0} \right\}^{-1}.$$

Now, from the original system of equations (S.6), we can find $1/y_3$ as follows:

$$\begin{aligned} \frac{1}{y_3} &= \frac{2y_0 - x_1}{y_0 x_1} = \frac{2}{x_1} - \frac{1}{y_0} \\ &= \frac{2(y_{-3}x_{-2} + y_{-3}y_0 - y_0 x_{-2})}{y_{-3}y_0 x_{-2}} - \frac{1}{y_0} \\ &= \frac{y_{-3}x_{-2} + 2y_{-3}y_0 - 2y_0 x_{-2}}{y_{-3}y_0 x_{-2}}. \end{aligned}$$

Hence, we now have

$$\begin{aligned} \forall n \in \mathbb{N}_0: \quad y_{3(n+1)} &= \left\{ \frac{(n+1)(y_{-3}x_{-2} + 2y_{-3}y_0 - 2y_0 x_{-2})}{y_{-3}y_0 x_{-2}} - \frac{n}{y_0} \right\}^{-1} \\ &= \frac{y_{-3}y_0 x_{-2}}{y_{-3}x_{-2} + 2(n+1)y_{-3}y_0 - 2(n+1)y_0 x_{-2}}. \end{aligned}$$

On the other hand, upon replacing n by $3n + i - 1$, with $i = 1, 2$, in $w_{n-1} = \frac{1}{2}\{z_{n+1} + z_{n-2}\}$, we have

$$\begin{aligned} w_{3n-2+i} &= \frac{1}{2}\{z_{3n-i} + z_{3(n-1)-i}\} \\ &= \frac{1}{2}\{\left[(n+1)z_i - nz_{-3+i}\right] + \left[nz_i - (n-1)z_{-3+i}\right]\} \\ &= \frac{1}{2}\{(2n+1)z_i - (2n-1)z_{-3+i}\}. \end{aligned}$$

Using the substitution defined in (5), we have

$$\begin{aligned} \forall n \in \mathbb{N}_0: \quad \frac{1}{x_{3n-2+i}} &= \frac{1}{2}\left\{ \frac{(2n+1)}{y_i} - \frac{(2n-1)}{y_{-3+i}} \right\} \\ &= \frac{1}{2}\left\{ \frac{(2n+1)y_{-3+i} - (2n-1)y_i}{y_{-3+i}y_i} \right\}, \qquad i = 1, 2, \end{aligned}$$

or equivalently,

$$\forall n \in \mathbb{N}_0: \quad x_{3n-2+i} = \frac{2y_{-3+i}y_i}{(2n+1)y_{-3+i} - (2n-1)y_i}, \qquad i = 1, 2.$$

Again, we observe that the RHS of the above equation is dependent on y_1 and y_2 for $i = 1, 2$, respectively. So, in view of the form for y_1 and y_2, we have

$$\begin{aligned}\forall n \in \mathbb{N}_0: \quad x_{3n-2+i} &= \frac{2}{\dfrac{2n+1}{y_i} - \dfrac{2n-1}{y_{-3+i}}} = \frac{2}{\dfrac{(2n+1)(2y_{i-3} - x_{i-2})}{y_{i-3}x_{i-2}} - \dfrac{2n-1}{y_{i-3}}} \\ &= \frac{y_{i-3}x_{i-2}}{(2n+1)y_{i-3} - 2nx_{i-2}}, \qquad i = 1, 2.\end{aligned}$$

Now, to compute for the form of the terms x_{3n+1}, we replace n by $3j$ in the equation $w_{n+1} - w_{n-2} = z_n - z_{n-3}$ and then sum up the resulting equation from 1 to n. Hence, we have

$$\begin{aligned}\forall n \in \mathbb{N}_0: \quad w_{3n+1} - w_1 &= \sum_{j=1}^{n} \{w_{3j+1} - w_{3(j-1)+1}\} = \sum_{j=1}^{n} \{z_{3j} - z_{3(j-1)}\} \\ &= z_{3n} - z_0,\end{aligned}$$

or equivalently, in view of the substitution (5),

$$\forall n \in \mathbb{N}_0: \quad x_{3n+1} = \left\{ \frac{1}{y_{3n}} + \frac{1}{x_1} - \frac{1}{y_0} \right\}^{-1}.$$

With the expression for x_1 computed using the original system together with the form of solution for y_{3n}, we get

$$\begin{aligned}\forall n \in \mathbb{N}_0: \; x_{3n+1} &= \left\{ \frac{y_{-3}x_{-2} + 2ny_{-3}y_0 - 2ny_0x_{-2}}{y_{-3}y_0x_{-2}} \right. \\ &\qquad \left. + \frac{y_{-3}x_{-2} + y_{-3}y_0 - y_0x_{-2}}{y_{-3}y_0x_{-2}} - \frac{1}{y_0} \right\}^{-1} \\ &= \left\{ \frac{y_{-3}x_{-2} + 2ny_{-3}y_0 - 2ny_0x_{-2}}{y_{-3}y_0x_{-2}} + \frac{y_{-3}y_0 - y_0x_{-2}}{y_{-3}y_0x_{-2}} \right\}^{-1} \\ &= \frac{y_{-3}y_0x_{-2}}{y_{-3}x_{-2} + (2n+1)y_{-3}y_0 - (2n+1)y_0x_{-2}}.\end{aligned}$$

Combining the results we have exhibited above, we arrive at the following theorem.

Theorem 6 *Let $\{x_n\}_{-2}^{0}$ and $\{y_n\}_{-3}^{0}$ be non-zero real numbers such that $y_{-1}/x_0,\ y_{-2}/y_{-1} \notin \left(\{2n/(2n+1)\}_0^\infty \cup \{(2n+1)/(2n+2)\}_0^\infty\right)$ and $(y_{-3}y_0 - y_0x_{-2})/y_{-3}x_{-2} \notin \{-1/n\}_1^\infty$. Then, every solution $\{(x_n, y_n)\}_1^\infty$ of system (S.6) takes the form*

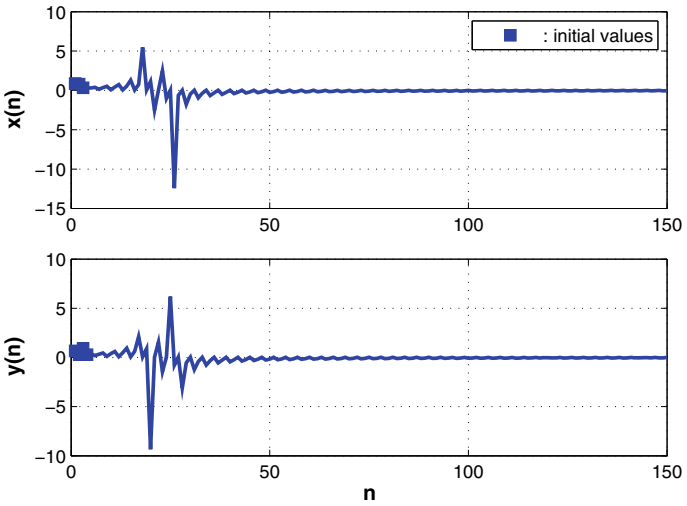

Fig. 6 The long term behavior of a particular solution of system (S.6)

$$\forall n \in \mathbb{N}_0 : \quad x_{3n-2+i} = \begin{cases} \dfrac{y_{-3}y_0x_{-2}}{y_{-3}x_{-2} + (2n+1)y_{-3}y_0 - (2n+1)y_0x_{-2}}, & \text{for } i = 3, \\ \dfrac{y_{i-3}x_{i-2}}{(2n+1)y_{i-3} - 2nx_{i-2}}, & \text{for } i = 2, 1, \end{cases}$$

and

$$\forall n \in \mathbb{N}_0 : \quad y_{3n+i} = \begin{cases} \dfrac{y_{i-3}x_{i-2}}{2(n+1)y_{i-3} - (2n+1)x_{i-2}}, & \text{for } i = 1, 2, \\ \dfrac{y_{-3}y_0x_{-2}}{y_{-3}x_{-2} + 2(n+1)y_{-3}y_0 - 2(n+1)y_0x_{-2}}, & \text{for } i = 3. \end{cases}$$

In the following example, we provide a numerical illustration describing the long term behavior of system (S.6) for some arbitrary initial values taken randomly from the unit interval [0, 1]. Notice that, in the illustrated plots, the solution converges to zero as n goes to infinity.

Example 6 Figures 6, 7 and 8 illustrate several plots for the long term dynamics of system (S.6) with random initial values from [0, 1].

In contrast to the first two plots shown previously, the illustrated behavior of solution of system (S.6) shown in Fig. 8 is diverging.

In view of the illustrations, it is natural to ask what particular set of initial values will give a converging, diverging or periodic solution (if there is) to system (S.6). The answer to this question shall be the subject of further investigation elsewhere.

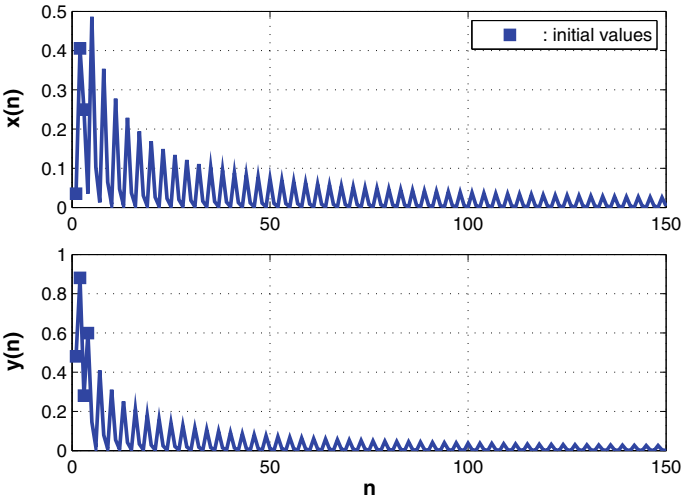

Fig. 7 Another possible long term behavior of a particular solution of system (S.6)

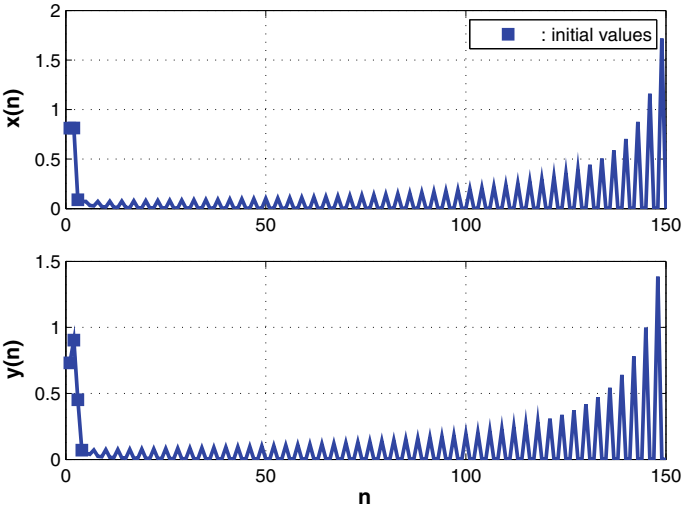

Fig. 8 Another possible long term behavior of a particular solution of system (S.6)

3.7 Solution Form of System (S.7)

Consider the system

$$x_{n+1} = \frac{y_{n-3}y_n x_{n-2}}{y_{n-3}x_{n-2} - y_{n-3}y_n + y_n x_{n-2}}, \qquad y_{n+1} = \frac{y_{n-2}x_{n-1}}{2y_{n-2} - x_{n-1}}, \quad n \in \mathbb{N}_0.$$

In view of the substitution (5), the above equations are therefore equivalent to the following transformations

$$\forall n \in \mathbb{N}_0 : \begin{cases} w_{n+1} = z_n - w_{n-2} + z_{n-3} \iff z_n + z_{n-3} = w_{n+1} + w_{n-2}, \\ z_{n+1} = 2w_{n-1} - z_{n-2} \qquad \iff \qquad w_{n-1} = \dfrac{1}{2}\{z_{n+1} + z_{n-2}\}. \end{cases}$$

With the above equations at hand, we easily obtained the equation $2w_{n-2} = w_{n+1} + w_{n-2}$ or equivalently, $w_{n+1} = w_{n-2}$. Following our result for system (S.5), we have

$$\forall n \in \mathbb{N}_0 : \quad x_{3n-i} = x_{3(n-1)-i} \quad \iff \quad x_{3n-i} = x_{-i}, \qquad i = 0, 1, 2,$$

which in turn implies that $\{x_n\}_1^\infty$ is periodic with period 3. Obviously, at $i = 2$, the equation $x_{3n-2} = x_{-2}$ does not hold for $n = 0$, but, this can be fixed by solving for x_1 directly from the original system (S.7). Therefore, the correct solution form for x_n is given by

$$\forall n \in \mathbb{N}_0 : \quad x_{3n-i} = \begin{cases} \dfrac{y_{-3}y_0x_{-2}}{y_{-3}x_{-2} - y_{-3}y_0 + y_0x_{-2}}, & \text{for } i = 2, \\ x_{-i}, & \text{for } i = 1, 0. \end{cases}$$

Now, from the transformation of the original system obtained through the substitution (5), we have $2z_n + 2z_{n-3} = \{z_{n+3} + z_n\} + \{z_n + z_{n-3}\}$ or equivalently, $z_{n+3} = z_{n-3}$. Hence, $y_{n+3} = y_{n-3}$ for all $n \in \mathbb{N}_0$ which, in turn, implies that y_n is periodic of period 6. Using this formula, at $n = 0$, we'll get $y_3 = y_{-3}$. This value for y_3, however, does not agree with

$$y_3 = \frac{y_0x_1}{2y_0 - x_1} = \frac{1}{\dfrac{2}{x_1} - \dfrac{1}{y_0}} = \frac{1}{\dfrac{2y_{-3}x_{-2} - 2y_{-3}y_0 + 2y_0x_{-2}}{y_{-3}y_0x_{-2}} - \dfrac{1}{y_0}}$$
$$= \frac{y_{-3}y_0x_{-2}}{y_{-3}x_{-2} - 2y_{-3}y_0 + 2y_0x_{-2}},$$

which is obtained from the original system (S.7). So, in view of this result, the solution for y_n must take the form $\{y_n\}_0^\infty = \{\overline{y_{-2}, y_{-1}, y_0, y_1, y_2, y_3}\}$ from which it suggests that y_1 and y_2 must be computed from the original system. These results now deliver the following theorem.

Theorem 7 *Let* $\{x_n\}_{-2}^{0}$ *and* $\{y_n\}_{-3}^{0}$ *be non-zero real numbers such that the following inequalities are satisfied:*

$$(y_{-3}x_{-2} - y_{-3}y_0 + y_0x_{-2}) \neq 0,$$

and

$$(2y_{-2} - x_{-1})(2y_{-1} - x_0)(y_{-3}x_{-2} - 2y_{-3}y_0 + 2y_0x_{-2}) \neq 0.$$

Then, every solution $\{(x_n, y_n)\}_1^\infty$ *of system (S.7) takes the form*

$$\forall n \in \mathbb{N}_0 : \quad x_{3n-i} = \begin{cases} \dfrac{y_{-3}y_0x_{-2}}{y_{-3}x_{-2} - y_{-3}y_0 + y_0x_{-2}}, & \text{for } i = 2, \\ x_{-i}, & \text{for } i = 1, 0, \end{cases}$$

and

$$\forall n \in \mathbb{N}_0 : \quad y_{6n+i} = \begin{cases} \dfrac{y_{i-3}x_{i-2}}{2y_{i-3} - x_{i-2}}, & \text{for } i = 1, 2 \\ \dfrac{y_{-3}y_0x_{-2}}{y_{-3}x_{-2} - 2y_{-3}y_0 + 2y_0x_{-2}}, & \text{for } i = 3, \\ y_{i-6}, & \text{for } i = 4, 5, 6. \end{cases}$$

Corollary 2 *Let* $\{(x_n, y_n)\}_1^\infty$ *be a solution of system (S.7). Then,* $\{x_n\}_1^\infty$ *is periodic of period* 3 *and* $\{y_n\}_1^\infty$ *is periodic of period* 6.

For confirming the virtue of the above corollary, we provide the following example.

Example 7 Figure 9 illustrates the long term dynamics of system (S.7) with random initial values taken from the unit interval [0, 1].

3.8 Solution Form of System (S.8)

Consider the system

$$x_{n+1} = \frac{y_{n-3}y_nx_{n-2}}{y_{n-3}x_{n-2} - y_{n-3}y_n - y_nx_{n-2}}, \qquad y_{n+1} = \frac{y_{n-2}x_{n-1}}{2y_{n-2} - x_{n-1}}, \quad n \in \mathbb{N}_0.$$

In view of the substitution (5), the above equations are therefore equivalent to the following transformations

$$\forall n \in \mathbb{N}_0 : \begin{cases} w_{n+1} = z_n - w_{n-2} - z_{n-3} \iff z_n - z_{n-3} = w_{n+1} + w_{n-2}, \\ z_{n+1} = 2w_{n-1} - z_{n-2} \qquad \iff \qquad w_{n-1} = \dfrac{1}{2}\{z_{n+1} + z_{n-2}\}. \end{cases}$$

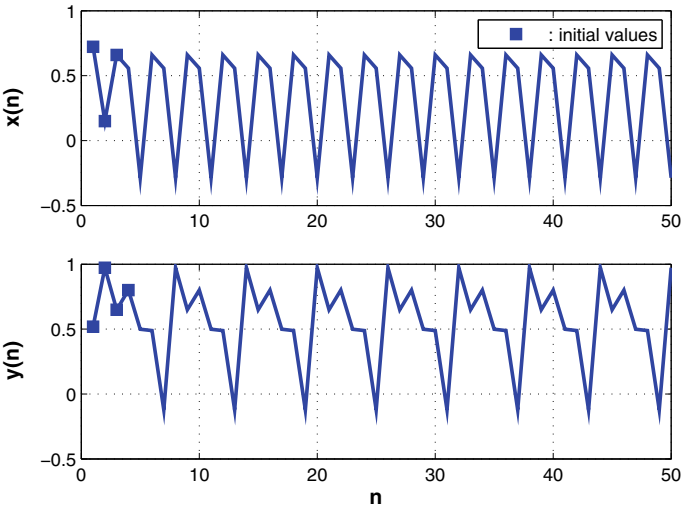

Fig. 9 The periodicity of solution for system (S.7) can be observed easily in above plots. The upper plot clearly shows that the solution $\{x_n\}_1^\infty$ is periodic with period 3 while it shows (lower plot) that the solution $\{y_n\}_1^\infty$ is of period 6. These behaviors agree with Corollary 2

The above equations imply, upon eliminating the phase variable w_n, $z_{n+3} = -3z_{n-3}$ or equivalently, $z_{n+6} = -3z_n$. Hence, replacing n by $6n - i$, where $i = -3, -2, -1, 0, 1, 2$, we get $z_{6(n+1)-i} = -3z_{6n-i}$. Iterating the RHS of this equation n times yields $z_{6(n+1)-i} = (-3)^{n+1} z_{-i}$. Then, using the substitution (5), we obtain

$$\forall n \in \mathbb{N}_0 : \quad y_{6(n+1)-i} = \left(-\frac{1}{3}\right)^{n+1} y_{-i}, \qquad i = -3, -2, -1, 0, 1, 2.$$

In view of the above formula, we see that $\{y_n\}_1^\infty$ is of period 6. Moreover, at $n = 0$, we will have the equation $y_{6-i} = -\frac{1}{3} y_{-i}$, $i = -3, -2, -1, 0, 1, 2$. For indices $i = -1, -2, -3$, the LHS of this equation depends on the terms y_1, y_2 and y_3, respectively. In this regard, we need to compute for the values of y_1, y_2 and y_3 in order to establish completely the closed form solution for $\{y_n\}_1^\infty$. These expressions, however, are easily obtained as follows:

$$y_1 = \frac{y_{-2}x_{-1}}{2y_{-2} - x_{-1}}, \qquad y_2 = \frac{y_{-1}x_0}{2y_{-1} - x_0}, \tag{8}$$

and

$$y_3 = \frac{y_0 x_1}{2y_0 - x_1} = \frac{y_{-3}y_0 x_{-2}}{x_{-2}y_{-3} - 2y_0 y_{-3} - 2y_0 x_{-2}}. \tag{9}$$

Now to find the solution form for x_n, we go directly to the original system (S.8). This part needs a little more work. From the original equation

$$y_{n+1} = \frac{y_{n-2}x_{n-1}}{2y_{n-2} - x_{n-1}},$$

we could find x_{n-1}, which is given by

$$x_{n-1} = \frac{2y_{n+1}y_{n-2}}{y_{n+1} + y_{n-2}}. \tag{10}$$

Replacing n by $6n + 7 - i$, where $i = 0, 1, 2, 3, 4, 5$, we get

$$x_{6n+6-i} = \frac{2y_{6(n+1)+2-i}\,y_{6(n+1)-1-i}}{y_{6(n+1)+2-i} + y_{6(n+1)-1-i}}.$$

In view of the formula for y_n we obtained earlier, the above equation can be transformed into

$$x_{6n+6-i} = \frac{2}{\dfrac{1}{y_{6(n+1)-1-i}} + \dfrac{1}{y_{6(n+1)+2-i}}} = \left(-\frac{1}{3}\right)^{n+1} \frac{2y_{2-i}\,y_{-1-i}}{y_{2-i} + y_{-1-i}}$$

for $i = 0, 1$. Now, using Eq. (8), we have

$$\forall n \in \mathbb{N}_0 : \quad x_{6n+6-i} = \left(-\frac{1}{3}\right)^{n+1} x_{-i}, \qquad i = 0, 1.$$

For $i = 2$, in reference to Eq. (9) together with the formula for y_n, we have

$$\begin{aligned}
\forall n \in \mathbb{N}_0 : \quad x_{6n+4} &= \frac{2}{\dfrac{1}{y_{6n+3}} + \dfrac{1}{y_{6(n+1)}}} = \left(-\frac{1}{3}\right)^{n} \left\{ \frac{2}{\dfrac{1}{y_3} - \dfrac{3}{y_0}} \right\} \\
&= \left(-\frac{1}{3}\right)^{n} \left\{ \frac{2}{\dfrac{x_{-2}y_{-3} - 2y_0y_{-3} - 2y_0x_{-2}}{y_{-3}y_0x_{-2}} - \dfrac{3}{y_0}} \right\} \\
&= -\left(-\frac{1}{3}\right)^{n} \left\{ \frac{y_{-3}y_0x_{-2}}{x_{-2}y_{-3} + y_0y_{-3} + y_0x_{-2}} \right\}.
\end{aligned}$$

Meanwhile, using Eq. (8) and the formula for y_n, we have

$$\forall n \in \mathbb{N}_0: \quad x_{6n+6-i} = \frac{2}{\frac{1}{y_{6(n+1)-1-i}} + \frac{1}{y_{6(n+1)+2-i}}} = \left(-\frac{1}{3}\right)^n \left\{ \frac{2}{\frac{1}{y_{5-i}} - \frac{3}{y_{2-i}}} \right\}$$
$$= \left(-\frac{1}{3}\right)^n \left\{ \frac{2}{\frac{2y_{2-i} - x_{3-i}}{y_{2-i}x_{3-i}} - \frac{3}{y_{2-i}}} \right\}$$
$$= \left(-\frac{1}{3}\right)^n \left\{ \frac{y_{2-i}x_{3-i}}{y_{2-i} - 2x_{3-i}} \right\}, \qquad i = 3, 4.$$

Finally, at $i = 5$, we have, with reference to Eq. (9) and the formula for y_n,

$$\forall n \in \mathbb{N}_0: \quad x_{6n+1} = \frac{2}{\frac{1}{y_{6n}} + \frac{1}{y_{6n+3}}} = \left(-\frac{1}{3}\right)^n \left\{ \frac{2}{\frac{1}{y_0} + \frac{x_{-2}y_{-3} - 2y_0y_{-3} - 2y_0x_{-2}}{y_{-3}y_0x_{-2}}} \right\}$$
$$= \left(-\frac{1}{3}\right)^n \left\{ \frac{y_{-3}y_0x_{-2}}{x_{-2}y_{-3} - y_0y_{-3} - y_0x_{-2}} \right\}.$$

In summary, we have the following theorem.

Theorem 8 *Let $\{x_n\}_{-2}^0$ and $\{y_n\}_{-3}^0$ be non-zero real numbers such that the following inequalities are satisfied:*

$$(x_{-2}y_{-3} - y_0y_{-3} - y_0x_{-2})(y_{-2} - 2x_{-1})(y_{-1} - 2x_0)(x_{-2}y_{-3} + y_0y_{-3} + y_0x_{-2}) \neq 0,$$

and

$$(x_{-2}y_{-3} - y_0y_{-3} - y_0x_{-2})(2y_{-2} - x_{-1})(2y_{-1} - x_0)(x_{-2}y_{-3} + y_0y_{-3} + y_0x_{-2}) \neq 0.$$

Then, every solution $\{(x_n, y_n)\}_1^\infty$ of system (S.8) takes the form

$$\forall n \in \mathbb{N}_0: \quad x_{6(n+1)-i} = \begin{cases} \left(-\frac{1}{3}\right)^n \left\{ \frac{y_{-3}y_0x_{-2}}{x_{-2}y_{-3} - y_0y_{-3} - y_0x_{-2}} \right\}, & \text{for } i = 5, \\ \left(-\frac{1}{3}\right)^n \left\{ \frac{y_{2-i}x_{3-i}}{y_{2-i} - 2x_{3-i}} \right\}, & \text{for } i = 4, 3, \\ -\left(-\frac{1}{3}\right)^n \left\{ \frac{y_{-3}y_0x_{-2}}{x_{-2}y_{-3} + y_0y_{-3} + y_0x_{-2}} \right\}, & \text{for } i = 2, \\ \left(-\frac{1}{3}\right)^{n+1} x_{-i}, & \text{for } i = 1, 0. \end{cases}$$

and

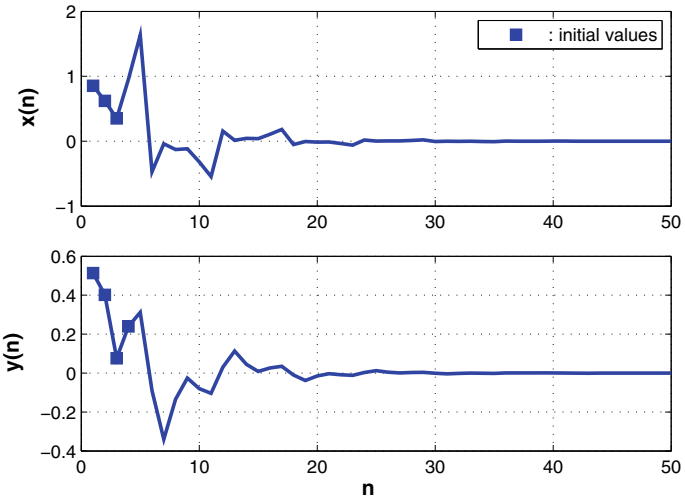

Fig. 10 Behavior of a particular solution of system (S.8)

$$\forall n \in \mathbb{N}_0 : \quad y_{6n-i} = \begin{cases} \left(-\frac{1}{3}\right)^n \left\{ \frac{y_{-(3+i)} x_{-(2+i)}}{2y_{-(3+i)} - x_{-(2+i)}} \right\}, & \text{for } i = -1, -2, \\ \left(-\frac{1}{3}\right)^n \left\{ \frac{y_{-3} y_0 x_{-2}}{x_{-2} y_{-3} - 2y_0 y_{-3} - 2y_0 x_{-2}} \right\}, & \text{for } i = -3, \\ \left(-\frac{1}{3}\right)^n y_{-i}, & \text{for } i = 2, 1, 0. \end{cases}$$

In view of the above theorem, it is evident that $\max\{x_n, y_n\} \to 0$ as $n \to \infty$ since $\lim_{n\to\infty}(-1/3)^n = \lim_{n\to\infty}(-1/3)^{n+1} = 0$ (refer to Fig. 10 for an illustration).

Example 8 Figure 10 illustrates the long term dynamics of system (S.8) with random initial values from [0, 1]. The computed solution $\{(x_n, y_n)\}_1^\infty$ to the system (S.8) clearly converges to (0, 0) as n approaches infinity.

4 Summary and a Statement of Future Work

We have considered in this work several systems of nonlinear difference equations. Some of these systems were already studied in [33]. The main goal of the paper was achieved by providing the readers new techniques—more explanatory and efficient—in determining the solution forms of these systems of equations. It has been shown that the method of differences or telescoping sums works perfectly in deriving the closed-form solutions. Furthermore, well-known integer sequences are seen in the solution forms.

Obviously, the techniques presented here can be employed in handling other systems of equations related to our work. Thus, our next agenda is to continue solving other forms of systems of difference equations via the method of telescoping sums.

Acknowledgements This work was completed in 2016 when JFTR was still at the Department of Mathematics and Computer Science, College of Science, University of the Philippines Baguio.

References

1. Bacani, J.B., Rabago, J.F.T.: On linear recursive sequences with coefficients in arithmetic-geometric progressions. Appl. Math. Sci. (Ruse) **9**(52), 2595–2607 (2015)
2. Bacani, J.B., Rabago, J.F.T.: An analytical approach in solving a system of nonlinear difference equations. Nat. Res. Counc. Philippines Res. J. **17**(3), 37–51 (2018)
3. Brand, L.: A sequence defined by a difference equation. Am. Math. Mon. **62**, 489–492 (1955)
4. Cyvin, S.J., Brunvoll, J., Brendsdal, E., Cyvin, B.N., Lloyd, E.K.: Enumeration of polyene hydrocarbons: a complete mathematical solution. J. Chem. Inf. Comput. Sci. **35**, 743–751 (1995)
5. Dunlap, R.A.: The Golden Ratio and Fibonacci Numbers. World Scientific, Singapore (1997)
6. Elaydi, S.: An Introduction to Difference Equations, 2nd edn. Springer, New York (1999)
7. Elsayed, E.M., Ibrahim, T.F.: Periodicity and solutions for some systems of nonlinear rational difference equations. Hacet. J. Math. Stat. **44**(6), 1361–1390 (2015)
8. Elsayed, E.M.: Solution for systems of difference equations of rational form of order two. Comput. Appl. Math. **33**(3), 751–765 (2014)
9. Elsayed, E.M., El-Metwwally, H.: On the solutions of some nonlinear systems of difference equations. Adv. Diff. Equ. **2013**, 161 (2013)
10. Grove, E.A., Ladas, G.: Advances in Discrete Mathematics and Applications. Chapman & Hall/CRC, Boca Raton (2005)
11. Halim, Y., Rabago, J.F.T.: On the solutions of a second-order difference equation in terms of generalized Padovan sequences. Math. Slovaca **68**(1), 625–638 (2018)
12. Horadam, A.F.: Basic properties of a certain generalized sequence of numbers. Fib. Quart. **3**, 161–176 (1965)
13. Iričanin, B.D., Liu, W.: On a higher-order difference equation. Disc. Dyn. Nat. Soc. **2010**, Article ID 891564, 6 pages (2010)
14. Jagerman, D.L.: Difference Equations with Applications to Queues. Chapman & Hall/CRC (2000)
15. Kocič, V.L., Ladas, G.: Mathematics and its Applications. Kluwer Academic Publishers, Dordrecht, The Netherlands (1993)
16. Koshy, T.: Fibonacci and Lucas Numbers with Applications. Pure and Applied Mathematics, Wiley-Interscience, New York (2001)
17. Kulenović, M.R.S., Ladas, G.: Dynamics of Second Order Rational Difference Equations: With Open Problems and Conjectures. Chapman & Hall/CRC, Boca Raton (2002)
18. Larcombe, P.J., Bagdasar, O.D., Fennessey, E.J.: Horadam sequences: a survey. Bull. I.C.A. **67**, 49–72 (2013)
19. Lucas, E.: Théorie des Fonctions Numériques Simplement Périodiques. Am. J. Math. **1**, 184–240, 289–321 (1878); reprinted as "The Theory of Simply Periodic Numerical Functions", Santa Clara, CA: The Fibonacci Association (1969)
20. Mickens, R.E.: Difference Equations: Theory, Applications and Advanced Topics, 3rd edn. Chapman and Hall/CRC (2015)
21. OEIS Foundation Inc.: The On-Line Encyclopedia of Integer Sequences (2011). http://oeis.org

22. Proakis, J.G., Manolakis, D.G.: Digital Signal Processing: Principles, Algorithms, and Applications, 4th edn. (2007)
23. Rabago, J.F.T.: Effective methods on determining the periodicity and form of solutions of some systems of nonlinear difference equations. Int. J. Dyn. Syst. Differ. Equ. **7**(2), 112–135 (2017)
24. Rabago, J.F.T.: An intriguing application of telescoping sums. In: Proceedings of the 2016 Asian Mathematical Conference, IOP Conference Series: Journal of Physics: Conference Series, vol. 893, p. 012005 (2017)
25. Rabago, J.F.T., Halim, Y.: Supplement to the paper of Halim, Touafek and Elsayed: Part I. Dyn. Contin. Discrete Impuls. Syst. Ser. A Math. Anal. **24**:2, 121–131 (2017)
26. Rabago, J.F.T., Halim, Y.: Supplement to the paper of Halim, Touafek and Elsayed: Part II. Dyn. Contin. Discret. Impuls. Syst. Ser. A Math. Anal. **24**:5, 333–345 (2017)
27. Rabago, J.F.T.: On the closed-form solution of a nonlinear difference equation and another proof to Sroysang's conjecture. Iran. J. Math. Sci. Inform. **13**(1), 139–151 (2018)
28. Rabago, J.F.T.: On an open question concerning product-type difference equations. Iran. J. Sci. Technol. Trans. Sci. **42**, 1499–1503 (2018)
29. Sargent, T.J.: Dynamic Macroeconomic Theory. Harvard University Press (1987)
30. Sedgewick, R., Flajolet, F.: An Introduction to the Analysis of Algorithms. Addison-Wesley (2013)
31. Sharkovsky, A.N., Maistrenko, Y.L., Yu Romanenko, E.: Difference Equations and Their Applications, Mathematics and Its Applications, Vol. 250. Springer, Netherlands (1993)
32. Stévic, S.: Representation of solutions of bilinear equations in terms of generalized Fibonacci sequences. Electron. J. Qual. Theory Differ. Equ. **67**, 1–15 (2014)
33. Touafek, N.: On some fractional systems of difference equations. Iran. J. Math. Sci. Info. **9**(2), 303–305 (2014)
34. Touafek, N.: On a second order rational difference equation. Hacet. J. Math. Stat. **41**, 867–874 (2012)
35. Touafek, N., Elsayed, E.M.: On the solutions of systems of rational difference equations. Math. Comput. Model. **55**, 1987–1997 (2012)
36. Touafek, N., Elsayed, E.M.: on the periodicity of some systems of nonlinear difference equations. Bull. Math. Soc. Sci. Math. Roum. Nouv. Sr. **55**(103), 217–224 (2012)
37. Vajda, S.A.: Fibonacci & Lucas Numbers and The Golden Section: Theory and Applications. Ellis Horwood Ltd., Chishester (1989)
38. Vorobév, N.N.: Fibonacci Numbers. Birkhäuser, Basel (2002)

A Note on q-partial Differential Equations for Generalized q-2D Hermite Polynomials

JIAN CAO, Tianxin Cai, and Li-Ping Cai

Abstract In this short paper, we generalize Ismail–Zhang's q-2D Hermite polynomials (Trans Am Math Soc 369:6779–6821 (2017), [14]) with an extra parameter and prove that if an analytic function in several variables satisfies a set of partial differential equations of second order, then it can be expanded in terms of the product of the generalized q-2D Hermite polynomials. In addition, we give some generating functions as applications.

Keywords q-partial differential equation · Generating function · Generalized q-2d hermite polynomials

1 Introduction

In this paper, we follow the notations and terminology in [9] and suppose that $0 < q < 1$. The compact factorials of q-shifted factorials are defined respectively by

$$(a;q)_0 = 1, \quad [a]_q := \frac{1-q^a}{1-q}, \quad (a;q)_n = \prod_{k=0}^{n-1}(1-aq^k), \quad (a;q)_\infty = \prod_{k=0}^{\infty}(1-aq^k) \tag{1}$$

and $(a_1, a_2, \ldots c, a_m; q)_n = (a_1;q)_n(a_2;q)_n \cdots (a_m;q)_n$, where $m \in \mathbb{N} := \{1, 2, 3, \ldots\}$ and $n \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$.

J. CAO (✉) · L.-P. Cai
Department of Mathematics, Hangzhou Normal University, Hangzhou 311121, Zhejiang, China
e-mail: 21caojian@hznu.edu.cn
URL: http://lxy.hznu.edu.cn/c/2012-05-16/255464.shtml

T. Cai
Department of Mathematics, Zhejiang University, Hangzhou 310027, China
e-mail: txcai@zju.edu.cn

S. Baigent et al. (eds.), *Progress on Difference Equations and Discrete Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 341,
https://doi.org/10.1007/978-3-030-60107-2_8

The complex Hermite polynomials $\{H_{m,n}(z_1, z_2)\}_{m,n=0}^{\infty}$ were defined first by Itô [15]

$$H_{m,n}(z_1, z_2) = \sum_{k=0}^{m\wedge n}(-1)^k k!\binom{m}{k}\binom{n}{k}z_1^{m-k}z_2^{n-k} \tag{2}$$

in his study of complex multiple Wiener integrals and applied in normal stochastic processes. In recent years, several mathematical physicists studied complex Hermite polynomials from mathematical and physical points of view, applying them to Landau levels and coherent states [1], quantum optics and quasi-probabilities [21, 22]. (See details in [1, 8, 10, 15, 21, 22], respectively.)

Just recently, Ismail and Zhang [12] defined the following two q-analogue complex Hermite polynomials:

Definition 1 ([12, *Eqs.* (3.1) *and* (4.2)]) For $m, n \in \mathbb{N}_0$ and $m \wedge n = min\{m, n\}$, the q-2D Hermite polynomials are

$$H_{m,n}(x, y|q) = \sum_{k=0}^{m\wedge n}\begin{bmatrix}m\\k\end{bmatrix}_q\begin{bmatrix}n\\k\end{bmatrix}_q(-1)^k q^{\binom{k}{2}}(q;q)_k x^{m-k}y^{n-k}, \tag{3}$$

$$G_{m,n}(x, y|q) = \sum_{k=0}^{m\wedge n}\begin{bmatrix}m\\k\end{bmatrix}_q\begin{bmatrix}n\\k\end{bmatrix}_q q^{(m-k)(n-k)}(-1)^k(q;q)_k x^{m-k}y^{n-k}, \tag{4}$$

where $\begin{bmatrix}m\\k\end{bmatrix}_q := \begin{cases}\frac{(q;q)_n}{(q;q)_k(q;q)_{n-k}}, & if \quad 0 \le k \le n,\\ 0, & otherwise.\end{cases}$

The q-2D Hermite polynomials $H_{m,n}(x, y|q)$ and $G_{m,n}(x, y|q)$ transform into each other under $q \to 1/q$. Ismail and Zhang produced several orthogonality measures for both families of q-2D Hermite polynomials, and found raising and lowering operators for both families of q-2D Hermite polynomials together with the Sturm–Liouville equations which they satisfy, see details in [12–14].

Recently, Liu [17] introduced the concept of the ternary classical 2D Hermite polynomials and then proved that if an analytic function in several variables satisfies a set of partial differential equations of second order, then it can be expanded in terms of the product of the ternary classical Hermite polynomials (see details in [17, 18]).

Motivated by Ismail–Zhang's results for q-2D Hermite polynomials and Liu's method for classical 2D Hermite polynomials, it's natural to ask if we can add an extra parameter on Ismail–Zhang's q-2D Hermite polynomials (generalized q-2D Hermite polynomials) and prove that if an analytic function in several variables satisfies a set of partial differential equations of second order, then it can be expanded in terms of the product of the generalized q-2D Hermite polynomials. With this method, we give a new proof and further generalize Ismail–Zhang's results.

Definition 2 For $m, n \in \mathbb{N}_0$, we define

$$H_{m,n}(x, y, z|q) = \sum_{k=0}^{m\wedge n} \begin{bmatrix} m \\ k \end{bmatrix}_q \begin{bmatrix} n \\ k \end{bmatrix}_q (-1)^k q^{\binom{k}{2}} (q;q)_k x^{m-k} y^{n-k} z^k, \tag{5}$$

$$G_{m,n}(x, y, z|q) = \sum_{k=0}^{m\wedge n} \begin{bmatrix} m \\ k \end{bmatrix}_q \begin{bmatrix} n \\ k \end{bmatrix}_q (q;q)_k q^{(m-k)(n-k)} x^{m-k} y^{n-k} z^k. \tag{6}$$

We will prove:

Theorem 3 *If $f(x, y, z)$ is a 3-variable analytic function at $(0, 0, 0) \in \mathbb{C}^3$, then*

(I) f can be expanded in an absolutely and uniformly convergent polynomial $H_{m,n}(x, y, z|q)$, if and only if f satisfies the partial differential equation

$$\frac{\partial_{q^{-1}}}{\partial_{q^{-1}} z} f = -\frac{\partial_q^2}{\partial_q x \partial_q y} f, \tag{7}$$

where $f = f(x, y, z)$ and

$$\frac{\partial_q}{\partial_q z} f = \frac{f(x) - f(qx)}{x}, \quad \frac{\partial_q^2}{\partial_q x \partial_q y} f = \frac{\partial_q}{\partial_q y} \left\{ \frac{\partial_q}{\partial_q x} f \right\}. \tag{8}$$

(II) f can be expanded in an absolutely and uniformly convergent polynomial $G_{m,n}(x, y, z|q)$, if and only if f satisfies the partial differential equation

$$\frac{\partial_q}{\partial_q z} \eta_z f = \frac{\partial_{q^{-1}}^2}{\partial_{q^{-1}} x \partial_{q^{-1}} y} f \tag{9}$$

where $\eta_z f = f(x, y, zq)$.

These expansion theorems allow us to easily deduce identities for generalized q-2D Hermite polynomials.

The rest of the paper is organized as follows. In Sect. 2, we give the proof of Theorem 12. In Sect. 3, using Theorem 12, we obtain the Srivastava–Agarwal type generating functions for the generalized q-2D Hermite polynomials. In Sect. 4, we gain the mixed generating functions for the generalized q-2D Hermite polynomials.

2 The Proof of Theorem 3

In order to prove Theorem 3, the following lemmas are necessary.

Lemma 4 ([12, Eqs. (3.3) and (4.6)]) *We have*

$$\sum_{m,n=0}^{\infty} H_{m,n}(x,y,z|q)\frac{s^m t^n}{(q;q)_m(q;q)_n} = \frac{(stz;q)_\infty}{(sx,ty;q)_\infty}, \quad \max\{|sx|,|ty|\}<1, \tag{10}$$

$$\sum_{m,n=0}^{\infty} G_{m,n}(x,y,z|q)\frac{q^{(m-n)^2/2}s^m t^n}{(q;q)_m(q;q)_n} = \frac{\left(-sxq^{1/2},-tyq^{1/2};q\right)_\infty}{(stz;q)_\infty}, \quad |stz|<1. \tag{11}$$

Proof (Proof of Lemma 4) Direct summation in Eqs. (3) and (4) yields (10) and (11) respectively.

Remark 5 Taking $z=1$ in Lemma 4, Eqs. (10) and (11) reduce to [12, Eqs. (3.3) and (4.6)] respectively.

Lemma 6 *For $m,n\in\mathbb{N}_0$, polynomials $H_{m,n}(x,y,z|q)$ and $G_{m,n}(x,y,z|q)$ satisfy the following partial differential equations respectively*

$$\begin{aligned}
\frac{\partial_{q^{-1}}}{\partial_{q^{-1}}z}\{H_{m,n}(x,y,z|q)\} &= -\frac{\partial_q^2}{\partial_q x\partial_q y}\{H_{m,n}(x,y,z|q)\},\\
\frac{\partial_q}{\partial_q z}\eta_z\{G_{m,n}(x,y,z|q)\} &= \frac{\partial_{q^{-1}}^2}{\partial_{q^{-1}}x\partial_{q^{-1}}y}\{G_{m,n}(x,y,z|q)\}.
\end{aligned} \tag{12}$$

Proof (Proof of Lemma 6) Applying q-partial differential operator $\frac{\partial_q^2}{\partial_q x\partial_q y}$ to both sides of the Eq. (10), we deduce

$$\sum_{m,n=0}^{\infty}\frac{\partial_q^2 H_{m,n}}{\partial_q x\partial_q y}\frac{s^m t^n}{(q;q)_m(q;q)_n} = \frac{st(stz;q)_\infty}{(sx,ty;q)_\infty}. \tag{13}$$

Similarly, taking the q-partial differential operator on both sides of the Eq. (10) with z yields

$$\sum_{m,n=0}^{\infty}\frac{\partial_{q^{-1}} H_{m,n}}{\partial_{q^{-1}}z}\frac{s^m t^n}{(q;q)_m(q;q)_n} = -\frac{st(stz;q)_\infty}{(sx,ty;q)_\infty}. \tag{14}$$

Comparing the above two equation, we have

$$\sum_{m,n=0}^{\infty}\frac{\partial_q^2 H_{m,n}}{\partial_q x\partial_q y}\frac{s^m t^n}{(q;q)_m(q;q)_n} = -\sum_{m,n=0}^{\infty}\frac{\partial_{q^{-1}} H_{m,n}}{\partial_{q^{-1}}z}\frac{s^m t^n}{(q;q)_m(q;q)_n}. \tag{15}$$

Equating coefficients of s and t on both sides of the above Eq. (15), we obtain the first formula in Eq. (12). Similarly, we deduce the second formula in Eq. (12). The proof of Lemma 6 is complete.

Lemma 7 *For $m, n \in \mathbb{N}_0$, $H_{m,n}(x, y, z|q)$ and $G_{m,n}(x, y, z|q)$ have the following operational representation*

$$\begin{aligned} H_{m,n}(x, y, z|q) &= (z\partial_q x\partial_q y; q)_\infty\{x^m y^n\}, \\ G_{m,n}(x, y, z|q) &= \frac{q^{mn}}{(q^{-1}z\partial_{q^{-1}} x\partial_{q^{-1}} y; q)_\infty}\{x^m y^n\}. \end{aligned} \tag{16}$$

Proof (Proof of Lemma 7) Direct computation gives

$$\left(\frac{\partial_q^2}{\partial_q x\partial_q y}\right)^k \{x^m y^n\} = \begin{cases} \begin{bmatrix} m \\ k \end{bmatrix}_q \begin{bmatrix} n \\ k \end{bmatrix}_q (q; q)_k^2 x^{m-k} y^{n-k}, & k \le min\{m, n\}, \\ 0, & k > min\{m, n\}. \end{cases}$$

Using the q-binomial theorem, we get

$$(z\partial_q x\partial_q y; q)_\infty\{x^m y^n\} = \sum_{k=0}^{m\wedge n} \frac{(-1)^k q^{\binom{k}{2}} z^k}{(q; q)_k} \left(\frac{\partial^2}{\partial_q x\partial_q y}\right)^k \{x^m y^n\} = H_{m,n}(x, y, z|q).$$

In the same way, we deduce the second formula in Eq. (16). The proof of Lemma 7 is complete.

We usually use the following Hartogs's theorem to determine whether a function is an analytic function in several complex variables. For more information, please refer to [11, 19, 20, 24].

Proposition 8 (Hartogs's theorem [11]) *If a complex-valued function is holomorphic (analytic) in each variable separately in an open domain $D \subseteq \mathbb{C}^n$, then it is holomorphic (analytic) in D.*

Proposition 9 ([20]) *If $f(x_1, x_2, \ldots, x_k)$ is analytic at the origin $(0, 0, \ldots, 0) \in \mathbb{C}^k$, then, f can be expanded in an absolutely and uniformly convergent power series,*

$$f(x_1, x_2, \ldots, x_k) = \sum_{n_1, n_2, \ldots, n_k = 0}^{\infty} \lambda_{n_1, n_2, \ldots, n_k} x_1^{n_1} x_2^{n_2} \ldots x_k^{n_k}.$$

Proof (Proof of Theorem 3) The proof of theorem can be deduced by induction. Since f is analytic at the origin $(0, 0, 0)$, we know that f can be expanded in an absolutely and uniformly convergent power series in a neighborhood of $(0, 0, 0)$. Thus there exists a sequence $\lambda_{m,m,p}$ independent of x_1, y_1, z_1, such that

$$f(x_1, y_1, z_1) = \sum_{m,n,p=0}^{\infty} \lambda_{m,n,p} \cdot x_1^m y_1^n z_1^p. \tag{17}$$

Substituting the Eq. (17) into the q-partial differential equation (7) yields

$$\sum_{m,n,p=0}^{\infty} (1-q^p) q^{1-p} \lambda_{m,n,p} \cdot x_1^m y_1^n z_1^{p-1} = -\frac{\partial_q^2}{\partial_q x_1 \partial_q y_1} \left\{ \sum_{m,n,p=0}^{\infty} \lambda_{m,n,p} \cdot x_1^m y_1^n z_1^p \right\}. \tag{18}$$

Equating the coefficients of z_1^{p-1} on both sides of the Eq. (18), we obtain

$$(1-q^p) q^{1-p} \sum_{m,n=0}^{\infty} \lambda_{m,n,p} \cdot x_1^m y_1^n = -\frac{\partial_q^2}{\partial_q x_1 \partial_q y_1} \left\{ \sum_{m,n=0}^{\infty} \lambda_{m,n,p-1} \cdot x_1^m y_1^n \right\}.$$

Iterating this relation $p-1$ times and interchanging the order of the differentiation and summation, we deduce

$$\begin{aligned} \sum_{m,n=0}^{\infty} \lambda_{m,n,p} \cdot x_1^m y_1^n &= \frac{1}{(q;q)_p} \frac{\partial_q^{2p}}{\partial_q x_1^p \partial_q y_1^p} \left\{ \sum_{m,n=0}^{\infty} \lambda_{m,n,0} \cdot x_1^m y_1^n \right\} \\ &= \frac{(-1)^p q^{\binom{p}{2}}}{(q;q)_p} \sum_{m,n=0}^{\infty} \lambda_{m,n,0} \cdot \frac{\partial_q^{2p}}{\partial_q x_1^p \partial_q y_1^p} \left\{ x_1^m y_1^n \right\}. \end{aligned} \tag{19}$$

Substituting the above Eq. (19) into (17) and interchanging the order of the summation, we gain

$$\begin{aligned} f(x_1, y_1, z_1) &= \sum_{p=0}^{\infty} z_1^p \sum_{m,n=0}^{\infty} \lambda_{m,n,p} \cdot x_1^m y_1^n \\ &= \sum_{p=0}^{\infty} \frac{(-1)^p q^{\binom{p}{2}} z_1^p}{(q;q)_p} \sum_{m,n=0}^{\infty} \lambda_{m,n,0} \cdot \frac{\partial_q^{2p}}{\partial_q x_1^p \partial_q y_1^p} \left\{ x_1^m y_1^n \right\} \\ &= \sum_{m,n=0}^{\infty} \lambda_{m,n,0} \cdot (z_1 \partial_q x_1 \partial_q y_1; q)_\infty \{ x_1^m y_1^n \} \\ &= \sum_{m,n=0}^{\infty} \lambda_{m,n,0} \cdot H_{m,n}(x_1, y_1, z_1 | q). \end{aligned}$$

Conversely, if $f(x_1, y_1, z_1)$ can be expanded in terms of $H_{m,n}(x_1, y_1, z_1|q)$, then using Lemma 6, we can obtain Eq. (7). Similarly, we deduce the Eq. (9). This completes the proof of Theorem 3.

3 Srivastava–Agarwal Type Generating Functions for the Generalized q-2D Hermite Polynomials

In this section, we obtain Srivastava–Agarwal type generating functions by q-partial differential equations.

Theorem 10 *For* $max\{|sx|, |ty|, |stz|\} < 1$, *we have*

$$\sum_{m,n=0}^{\infty} H_{m,n}(x, y, z|q) \frac{s^m (u/s; q)_m t^n (v/t; q)_n}{(q; q)_m (q; q)_n} = \frac{(ux, vy; q)_\infty}{(sx, ty; q)_\infty} {}_2\phi_2 \left[\begin{matrix} u/s, v/t \\ ux, vy \end{matrix} ; q, stz \right]. \tag{20}$$

For $max\{|sx|, |ty|, |stz|\} < 1$, *we have*

$$\sum_{m,n=0}^{\infty} G_{m,n}(x, y, z|q) \frac{(-1)^{m+n} q^{(m+n-2mn)/2} u^m (s/u; q)_m v^n (t/v; q)_n}{(q; q)_m (q; q)_n}$$
$$= \frac{(-sxq^{1/2}, -tyq^{1/2}; q)_\infty}{(-uxq^{1/2}, -vyq^{1/2}; q)_\infty} {}_3\phi_2 \left[\begin{matrix} s/u, t/v, 0 \\ -q^{1/2}/(ux), -q^{1/2}/(vy) \end{matrix} ; q, \frac{zq}{xy} \right]. \tag{21}$$

Remark 11 Letting $u = v = 0$ in Theorem 10, Eqs. (20) and (21) reduce to Eqs. (10) and (11) respectively.

Proof (Proof of Theorem 10) We denote the right-hand side of Eq. (20) by

$$f(x, y, z) = \frac{(ux, vy; q)_\infty}{(sx, ty; q)_\infty} {}_2\phi_2 \left[\begin{matrix} u/s, v/t \\ ux, vy \end{matrix} ; q, stz \right].$$

It is easily seen that $f(x, y, z)$ is an analytic function of x, y, z, for any x, y and $\max\{|sx|, |ty|, |stz|\} < 1$. Hence $f(x, y, z)$ is analytic at $(x, y, z) = (0, 0, 0)$. Hence $f(x, y, z)$ satisfies Theorem 3: there exists a sequence $\lambda_{m,n}$ independent of x, y, z such that

$$f(x, y, z) = \sum_{m,n=0}^{\infty} \lambda_{m,n} \cdot H_{m,n}(x, y, z|q). \tag{22}$$

Setting $z = 0$, and using $H_{m,n}(x, y, 0|q) = x^m y^n$ in the resulting equation, we obtain

$$f(x, y, 0) = \frac{(ux, vy; q)_\infty}{(sx, ty; q)_\infty} = \sum_{m,n=0}^{\infty} \frac{(sx)^m (u/s; q)_m (ty)^n (v/t; q)_n}{(q; q)_m (q; q)_n} = \sum_{m,n=0}^{\infty} \lambda_{m,n} \cdot x^m y^n.$$

Comparing the coefficients of $x^m y^n$, we gain

$$\lambda_{m,n} = \frac{s^m (u/s;q)_m t^n (v/t;q)_n}{(q;q)_m (q;q)_n}.$$

Substituting the above equation into Eq. (22), we deduce $f(x, y, z)$ equals the left-hand side of Eq. (20). Similarly, we can deduce the Eq. (21). The proof is complete.

4 Mixed Generating Functions for the Generalized q-2D Hermite Polynomials

The Rogers–Szegö polynomials

$$h_n(a|q) = \sum_{k=0}^{n} \begin{bmatrix} n \\ k \end{bmatrix} a^k, \quad g_n(a|q) = \sum_{k=0}^{n} \begin{bmatrix} n \\ k \end{bmatrix} q^{k(k-n)} a^k \tag{23}$$

and their corresponding generating functions

$$\sum_{m=0}^{\infty} h_m(a|q) \frac{s^m}{(q;q)_m} = \frac{1}{(as, s; q)_\infty}, \quad \sum_{m=0}^{\infty} g_m(a|q) \frac{(-1)^m q^{\binom{m}{2}} s^m}{(q;q)_m} = (as, s; q)_\infty \tag{24}$$

play an important role in the theory of orthogonal polynomials, see details in [2, 5–7, 23].

In this section, we deduce the following mixed generating functions by q-partial differential equations.

Theorem 12 *For* $\max\{|asx|, |bty|, |ty|, |sx|\} < 1$, *we have*

$$\sum_{m,n=0}^{\infty} H_{m,n}(x, y, z|q) h_m(a|q) h_n(b|q) \frac{s^m t^n}{(q;q)_m (q;q)_n}$$
$$= \frac{(abstz;q)_\infty}{(asx, bty;q)_\infty} \sum_{j=0}^{\infty} \frac{(asz/y;q)_j (ty)^j}{(q;q)_j (abstz;q)_j} \sum_{k=0}^{\infty} \frac{(btzq^j/x, aszq^j/y;q)_k (sx)^k}{(q, abstzq^j, asz/y;q)_k} {}_2\phi_1 \left[\begin{matrix} q^{-j}, 0 \\ aszq^k/y \end{matrix} ; q, szq^j/y \right]. \tag{25}$$

For $|abstz| < 1$, *we have*

$$\sum_{m,n=0}^{\infty} G_{m,n}(x, y, z|q) g_m(a|q) g_n(b|q) \frac{q^{(m-n)^2/2} s^m t^n}{(q;q)_m (q;q)_n} = \frac{(-asxq^{1/2}, -btyq^{1/2};q)_\infty}{(abstz;q)_\infty}$$
$$\times \sum_{j=0}^{\infty} \frac{(-yq^{1/2}/(asz);q)_j (-1)^j q^{\binom{j+1}{2}} (1/b)^j}{(q, q/(abstz);q)_j}$$
$$\times \sum_{k=0}^{\infty} \frac{(-xq^{1/2+j}/(btz), -yq^{1/2+j}/(asz);q)_k (-q^{-j}/a)^k}{(q, q^{1+j}/(abstz), -yq^{1/2}/(asz);q)_k} \sum_{n=0}^{\infty} \frac{(q^{-j};q)_n q^{(n+k+1)(n+k)/2}}{(q; -yq^{1/2+k}/(asz);q)_n} \left(-\frac{1}{a}\right)^n. \tag{26}$$

Remark 13 For $a = b = 0$ in Theorem 12, Eqs. (25) and (26) reduce to Eqs. (10) and (11) respectively. For $z = 0$ in Theorem 12, Eqs. (25) and (26) reduce to results in Eqs. (24) respectively.

Proof (Proof of Theorem 12) We denote the right-hand side of Eq. (25) by

$$f(x, y, z) = \frac{(abstz; q)_\infty}{(asx, bty; q)_\infty} \sum_{j=0}^{\infty} \frac{(asz/y; q)_j (ty)^j}{(q; q)_j (absyz; q)_j} \times \sum_{k=0}^{\infty} \frac{(btzq^j/x, aszq^j/y; q)_k (sx)^k}{(q, abstzq^j, asz/y; q)_k} {}_2\phi_1 \left[\begin{matrix} q^{-j}, 0 \\ aszq^k/y \end{matrix} ; q, szq^j/y \right].$$

Since $f(x, y, z)$ is analytic, Theorem 3 shows that there exists a sequence $\lambda_{m,n}$ independent of x, y, z such that

$$f(x, y, z) = \sum_{m,n=0}^{\infty} \lambda_{m,n} \cdot H_{m,n}(x, y, z|q). \tag{27}$$

Setting $z = 0$ and using that $H_{m,n}(x, y, 0|q) = x^m y^n$ in the above Eq. (27), we have

$$f(x, y, 0) = \frac{1}{(sx, ty, asx, bty; q)_\infty} = \sum_{m,n=0}^{\infty} h_m(a|q) h_n(b|q) \frac{(sx)^m (ty)^n}{(q; q)_m (q; q)_n} = \sum_{m,n=0}^{\infty} \lambda_{m,n} \cdot x^m y^n. \tag{28}$$

Comparing the coefficients of $x^m y^n$ in the above Eq. (28), we obtain

$$\lambda_{m,n} = h_m(a|q) h_n(b|q) \frac{s^m t^n}{(q; q)_m (q; q)_n}. \tag{29}$$

Substituting the above Eq. (29) into Eq. (27) yields

$$f(x, y, z) = \sum_{m,n=0}^{\infty} h_m(a|q) h_n(b|q) H_{m,n}(x, y, z|q) \frac{s^m t^n}{(q; q)_m (q; q)_n},$$

which is the left-hand side of the Eq. (25). Similarly, we deduce the Eq. (26). The proof of Theorem 12 is complete.

5 Concluding Remarks

q-Partial Differential Equations are powerful methods, see details in [3, 4, 16, 25]. Compared to traditional combinatorial transformation methods used for analysis of q-2D Hermite polynomials, we build the relations between q-partial differential equations and q-2D Hermite polynomials. That is, if an analytic function in several variables satisfies a kind of q-partial differential equation of second order, then it can be expanded in terms of the product of the generalized q-2D Hermite polynomials. This method is a useful tool for proving formulas involving the generalized q-2D Hermite polynomials, which allow us to develop a systematic way to derive identities involving others q-2D polynomials.

Acknowledgments The first author thanks the professor Steve Baigent of University College London for the hospitality during the ICDEA 2019 when the major part of this work was carried out. This work was supported by the Zhejiang Provincial Natural Science Foundation of China (No. LY21A010019) and the National Natural Science Foundation of China (No. 12071421).

References

1. Ali, S.T., Bagarello, F., Honnouvo, G.: Modular structures on trace class operators and applications to Landau levels. J. Phys. A **43**, 105–202 (2010)
2. Al-Salam, W.A.: Some orthogonal q-polynomials. Math. Nachr. **30**, 47–61 (1965)
3. Cao, J.: New proofs of generating functions for Rogers-Szegö polynomials. Appl. Math. Comput. **207**, 486–492 (2009)
4. Cao, J.: Homogeneous q-partial difference equations and some applications. Adv. Appl. Math. **84**, 47–72 (2017)
5. Cao, J., Srivastava, H.M., Liu, Z.-G.: Some iterated fractional q-integrals and their applications. Fract. Calc. Appl. Anal. **21**, 672–695 (2018)
6. Carlitz, L.: Some polynomials related to theta functions. Ann. Mat. Pur. Appl. **41**, 359–373 (1956)
7. Chen, W.Y.C., Saad, H.L., Sun, L.H.: The bivariate Rogers-Szegö polynomials. J. Phys. A: Math. Theor. **40**, 6071–6084 (2007)
8. Cotfas, N., Gazeau, J.P., Górska, K.: Complex and real Hermite polynomials and related quantizations. J. Phys. A **43**, 305304, 14 (2010)
9. Gasper, G., Rahman, M.: Basic Hypergeometric Series. Camberdge University Press (2004)
10. Ghanmi, A.: Operational formulae for the complex Hermite polynomials $H_{p,q}(z, \bar{z})$. Integral Trans. Spec. Funct. **24**, 884–895 (2013)
11. Gunning, R.: Introduction to holomorphic functions of several variables, vol. I. Function Theory, Wadsworth and Brooks/Cole, Belmont, California (1990)
12. Ismail, M.E.H., Simeonov, P.: Complex Hermite polynomials: their combinatorics and integral operators. Proc. Am. Math. Soc. **143**, 1397–1410 (2015)
13. Ismail, M.E.H., Zhang, R.: Kibble-Slepian formula and generating functions for 2D polynomials. Adv. Appl. Math. **80**, 70–92 (2016)
14. Ismail, M.E.H., Zhang, R.: On some 2D orthogonal q-polynomials. Trans. Am. Math. Soc. **369**, 6779–6821 (2017)
15. Itô, K.: Complex multiple Wiener integral. Jpn. J. Math. **22**, 63–86 (1952)
16. Liu, Z.-G.: On the q-partial differential equations and q-series. In: The Legacy of Srinivasa Ramanujan 213–250, Ramanujan Mathematical Society Lecture Notes Series, vol. 20. Ramanujan Mathematical Society, Mysore (2013)

17. Liu, Z.-G.: On the complex Hermite polynomials and partial differential equations. arXiv:1707.08708
18. Liu, Y.-K.: The linear q-difference equation $y(x) = ay(qx) + f(x)$. Appl. Math. Lett. **8**, 15–18 (1995)
19. Liu, Z.-G.: On a system of partial differential equations and the bivariate Hermite polynomials. J. Math. Anal. Appl. **454**, 1–17 (2017)
20. Malgrange, B.: Lectures on the Theory of Functions of Several Complex Variables. Springer, Berlin (1984)
21. Szegö, G.: Ein Beitrag zur Theorie der Thetafunktionen, Sitz. Preuss. Akad. Wiss. Phys. Math. Kiasse. **19**(1926), 242–252
22. Taylor, J.: Several complex variables with connections to algebraic geometry and lie groups, Graduate Studies in Mathematics, American Mathematical Society. Providence **46**, (2002)
23. Wang, G.-T.: Twin iterative positive solutions of fractional q-difference Schrödinger equations. Appl. Math. Lett. **76**, 103–109 (2018)
24. Wünsche, A.: Laguerre 2D-functions and their application in quantum optics. J. Phys. A **31**, 8267–8287 (1998)
25. Wünsche, A.: Transformations of Laguerre 2D-polynomials and their applications to quasiprobabilities. J. Phys. A **32**, 3179–3199 (1999)

Stability of a Spring-Mass System with Generalized Piecewise Constant Argument

DUYGU ARUĞASLAN ÇINÇIN and Nur Cengiz

Abstract In this paper, we address a damped spring-mass system and develop it with piecewise constant argument of generalized type (PCAG). We investigate existence and uniqueness of the solutions of the proposed mechanical system. Then, we give sufficient conditions guaranteeing the uniform asymptotic stability of the trivial solution. While doing the stability examination, we use Lyapunov-Razumikhin method developed by Akhmet and Aruğaslan (Discrete and continuous dynamical systems. Series A, vol 25(2), pp 457–466, 2009, [1]) for differential equations with PCAG (EPCAG). Additionally, we present several examples with simulations.

Keywords Stability · Differential equations with piecewise constant argument of generalized type · Lyapunov-Razumikhin method · Spring-mass system · Simulations

1 Introduction and Preliminaries

Differential equations with piecewise constant argument (EPCA) are in the class of differential equations with deviating arguments [11, 17, 18, 22, 23, 25, 27–29]. In these type differential equations, the greatest integer function is considered as deviating argument. By taking any piecewise function instead of the greatest integer function, EPCAG are introduced in the papers [2–4] and developed in the papers [4–8]. Recently, research of EPCAG has attracted the attention of great number of researchers [9, 10, 12–15, 30]. Using the theory of EPCAG, we aim to study a spring-mass system which is one of the most remarkable models of real life problems and plays an important role in many fields such as physics, mathematics, biomechanics,

D. ARUĞASLAN ÇINÇIN (✉)
Department of Mathematics, Süleyman Demirel University, 32260 Isparta, Turkey
e-mail: duyguarugaslan@sdu.edu.tr

N. Cengiz
Graduate School of Natural and Applied Sciences, Süleyman Demirel University, 32260 Isparta, Turkey

S. Baigent et al. (eds.), *Progress on Difference Equations and Discrete Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 341,
https://doi.org/10.1007/978-3-030-60107-2_9

biology and engineering. In the present paper, we model the spring mass system using PCAG as follows

$$my''(t) + cy'(t) + ky(t) = Ay(\beta(t)). \tag{1}$$

This model can be considered as a damped harmonic oscillator. Let $\mathbb{R}$, $\mathbb{N}$, $\mathbb{N}_0$ and $\mathbb{R}^+$ be the sets of all real numbers, positive integers, non-negative integers and non-negative real numbers, respectively, i.e., $\mathbb{R} = (-\infty, \infty)$, $\mathbb{N} = \{1, 2, 3, ...\}$, $\mathbb{N}_0 = \{0, 1, 2, ...\}$ and $\mathbb{R}^+ = [0, \infty)$. Denote the $n-$dimensional real space by $\mathbb{R}^n$, $n \in \mathbb{N}$, and the Euclidean norm in $\mathbb{R}^n$ by $\|.\|$. Here, the positive constants m, c, k denote the mass, the coefficient of damping, the spring constant, respectively. A and y correspond to the magnitude of the generalized piecewise constant force and the displacement of the mass, respectively. Fix a real valued sequence $\theta = \{\theta_i\}, i \in \mathbb{N}_0$, such that $0 \leq \theta_i < \theta_{i+1}$ for all $i \in \mathbb{N}_0$ and $\theta_i \to \infty$ as $i \to \infty$. Let us assume without loss of generality that $\theta_i < t_0 \leq \theta_{i+1}$ for some $i \in \mathbb{N}_0$, where $t_0 \in \mathbb{R}^+$ is an initial moment. We denote $\mathbb{D} = [t_0, \infty)$. In (1), $y \in \mathbb{R}$, $t \geq t_0$, and $\beta(t) = \theta_i$ if $t \in [\theta_i, \theta_{i+1})$, $i \in \mathbb{N}_0$. There are numerous studies on the spring-mass systems with piecewise constant argument in the literature [13, 14, 19].

With $x_1(t) = y(t)$, $x_2(t) = y'(t)$, the damped spring-mass system (1) can be reduced to a first-order differential equation as

$$\begin{cases} x_1'(t) = x_2(t) \\ x_2'(t) = -\dfrac{c}{m}x_2(t) - \dfrac{k}{m}x_1(t) + \dfrac{A}{m}x_1(\beta(t)). \end{cases} \tag{2}$$

Let $x(t) = (x_1(t), x_2(t))$, and $x_1(t)$, $x_2(t)$ be abbreviated as x_1, x_2, respectively, throughout the paper. The definition of a solution of (2) on $[t_0, \infty)$ is defined below.

Definition 1 ([1]) A function $x(t)$ is a solution of (2) on $\mathbb{D}$ if:

(i) $x(t)$ is continuous on $\mathbb{D}$;
(ii) the derivative $x'(t)$ exists for $t \in \mathbb{D}$ with the possible exception of the points θ_i, $i \in \mathbb{N}_0$, where one-sided derivatives exist;
(iii) equation (2) is satisfied by $x(t)$ on each interval (θ_i, θ_{i+1}), $i \in \mathbb{N}_0$, and it holds for the right derivative of $x(t)$ at the points θ_i, $i \in \mathbb{N}_0$.

The main purpose of the paper is to give sufficient conditions for uniform asymptotic stability of the trivial solution of the mechanical system (2) ((1)) with PCAG, by means of Lyapunov-Razumikhin method, without finding the exact solution of the system and without transforming the system into a discrete one. Accordingly, the paper is organized as follows: In Sect. 2, we give a remarkable inequality which shows the relation between the value of the solution at the deviating argument $\beta(t)$ and the value of the solution at any t. In Sect. 3, we examine the stability of the trivial solution of (2). While investigating the stability, we consider Lyapunov-Razumikhin method developed by Akhmet and Aruğaslan [1] for EPCAG. Then, in Sect. 4, our theoretical results are exemplified and simulations are presented. Finally, in Sect. 5, we present conclusions and discuss what can be done in future studies.

2 An Auxiliary Result and Existence-Uniqueness of the Solutions

In this section, a crucial auxiliary result which has an importance in the proofs of stability theorems in the sense of the Lyapunov-Razumikhin method is given by Lemma 1. Moreover, sufficient conditions guaranteeing the existence and uniqueness of the solutions are presented in Lemma 2 and Theorem 1. The following assumption will be needed throughout the paper:

(C1) There exists a positive number $\overline{\theta}$ such that $\theta_{i+1} - \theta_i \leq \overline{\theta}$, $i \in \mathbb{N}_0$.

Additionally, introducing the following notations,

$$\Omega = e^{\frac{\overline{\theta}}{m}\left[c+\overline{\theta}k\right]},$$

$$\kappa_1 = \frac{\overline{\theta}}{m}(k+|A|)\left(\Omega\overline{\theta}\frac{1}{m}\left[c+\overline{\theta}k\right]+1\right),$$

$$\kappa = \max\left\{\left\{1-\overline{\theta}\Omega\frac{1}{m}(c+\overline{\theta}k)\right\}^{-2}(1+\kappa_1), \left\{1-\frac{\overline{\theta}^2}{m}(k+|A|)\,\Omega\right\}^{-2}(1+\overline{\theta}\Omega)\right\},$$

$$\mu = \max\left\{\kappa_1, \overline{\theta}\Omega\right\},$$

$$M = \max\left\{A^2+k^2+2\,|A|\,k+c\,|A|+ck, m^2+c^2+c\,|A|+ck\right\},$$

we will assume that the conditions below are valid:

(C2) $\overline{\theta}\Omega\frac{1}{m}(c+\overline{\theta}k) < 1$, $\frac{\overline{\theta}^2}{m}(k+|A|)\,\Omega < 1$, $\mu\kappa < 1$.

(C3) $2\frac{1}{m}\overline{\theta}M^{1/2} < 1$;

(C4) $\overline{\theta}e^{\frac{1}{m}\overline{\theta}M^{1/2}}\left(2\sqrt{A^2+k\,|A|+c\,|A|}+M^{1/2}\right) < m$.

Note that κ is positive by the condition (C2).

Lemma 1 *Let (C1) and (C2) be satisfied. Then the following inequality*

$$\|x(\beta(t))\| \leq \left\{\frac{1}{\kappa}-\mu\right\}^{-1/2}\|x(t)\| \tag{3}$$

holds for all $t \in \mathbb{D},\ x_1,\ x_2 \in \mathbb{R}$.

Proof For $t \in [\theta_i, \theta_{i+1})$, the solution (x_1, x_2) of system (2) can be written as follows

$$x_1 = x_1(\theta_i) + \int_{\theta_i}^{t} x_2(s)ds, \tag{4}$$

$$x_2 = x_2(\theta_i) + \int_{\theta_i}^{t} \left(-\frac{c}{m}x_2(s) - \frac{k}{m}x_1(s) + \frac{A}{m}x_1(\theta_i) \right) ds. \tag{5}$$

Then,

$$|x_2| \leq |x_2(\theta_i)| + \frac{\overline{\theta}}{m}(k + |A|)\,|x_1(\theta_i)| + \frac{c}{m}\int_{\theta_i}^{t} |x_2(s)|\,ds + \frac{k}{m}\int_{\theta_i}^{t}\int_{\theta_i}^{s} |x_2(u)|\,duds.$$

Using a Gronwall type inequality, stated by Bykov and Salpagarov [16, 20, 21] for integral equations including integral and double integral, for $t \geq s \geq \theta_i$, $i \in \mathbb{Z}$, we can write

$$\begin{aligned} |x_2| &\leq \left\{ |x_2(\theta_i)| + \frac{\overline{\theta}}{m}(k + |A|)\,|x_1(\theta_i)| \right\} e^{\int_{\theta_i}^{t} \frac{c}{m}ds + \int_{\theta_i}^{t}\int_{\theta_i}^{s} \frac{k}{m}duds} \\ &\leq \left\{ |x_2(\theta_i)| + \frac{\overline{\theta}}{m}(k + |A|)\,|x_1(\theta_i)| \right\} \Omega. \end{aligned} \tag{6}$$

For (4), it is seen that

$$\begin{aligned} |x_1| &\leq |x_1(\theta_i)| + \int_{\theta_i}^{t} \left\{ |x_2(\theta_i)| + \frac{\overline{\theta}}{m}(k + |A|)\,|x_1(\theta_i)| \right\} \Omega ds \\ &\leq |x_1(\theta_i)| \left[1 + \frac{\overline{\theta}^2}{m}(k + |A|)\,\Omega \right] + |x_2(\theta_i)|\,\overline{\theta}\Omega. \end{aligned} \tag{7}$$

Moreover, for $t \in [\theta_i, \theta_{i+1})$, we can write that

$$x_1(\theta_i) = x_1 - \int_{\theta_i}^{t} x_2(s)ds, \tag{8}$$

$$x_2(\theta_i) = x_2 - \int_{\theta_i}^{t} \left(-\frac{c}{m} x_2(s) - \frac{k}{m} x_1(s) + \frac{A}{m} x_1(\theta_i) \right) ds. \tag{9}$$

Then, considering (6) and (7), we reach the following inequality:

$$|x_2(\theta_i)| \leq \left\{ 1 - \overline{\theta}\Omega \frac{1}{m}(c + \overline{\theta}k) \right\}^{-1} (|x_2| + |x_1(\theta_i)|\,\kappa_1)\,. \tag{10}$$

Additionally, it follows from (8) and then (6) that

$$|x_1(\theta_i)| \leq |x_1| + \overline{\theta} \left\{ |x_2(\theta_i)| + \frac{\overline{\theta}}{m} (k + |A|)\, |x_1(\theta_i)| \right\} \Omega\,.$$

From the last inequality, we obtain that

$$|x_1(\theta_i)| \leq \left\{ 1 - \frac{\overline{\theta}^2}{m} (k + |A|)\, \Omega \right\}^{-1} \left(|x_1| + |x_2(\theta_i)|\, \overline{\theta}\Omega \right). \tag{11}$$

Using (10) and (11) together with the fact that $2\,|uv| \leq u^2 + v^2$, it is obtained that

$$x_2^2(\theta_i) \leq \left\{ 1 - \overline{\theta}\Omega \frac{1}{m}(c + \overline{\theta}k) \right\}^{-2} (1 + \kappa_1) \left(x_2^2 + x_1^2(\theta_i)\kappa_1 \right) \tag{12}$$

and

$$x_1^2(\theta_i) \leq \left\{ 1 - \frac{\overline{\theta}^2}{m} (k + |A|)\, \Omega \right\}^{-2} (1 + \overline{\theta}\Omega) \left(x_1^2 + x_2^2(\theta_i)\overline{\theta}\Omega \right). \tag{13}$$

Then, we get

$$x_1^2(\theta_i) + x_2^2(\theta_i) \leq \kappa \left(x_1^2 + x_2^2 \right) + \kappa\mu(x_1^2(\theta_i) + x_2^2(\theta_i)),$$

and so

$$x_1^2(\theta_i) + x_2^2(\theta_i) \leq \{1 - \mu\kappa\}^{-1} \kappa \left(x_1^2 + x_2^2 \right) = \left\{ \frac{1}{\kappa} - \mu \right\}^{-1} \left(x_1^2 + x_2^2 \right). \tag{14}$$

It follows from condition (C2) that $\|x(\theta_i)\| \leq \left\{ \frac{1}{\kappa} - \mu \right\}^{-1/2} \|x(t)\|$ for $t \in [\theta_i, \theta_{i+1})$. Thus, (3) holds for all $t \geq t_0$. □

Now, for arbitrary initial moment ξ, sufficient conditions for the existence and uniqueness of the solution of (2) on $\left[\theta_i, \theta_{i+1}\right]$ can be seen with help of the following lemma.

Lemma 2 *Let (C1), (C3) and (C4) be satisfied and $i \in \mathbb{N}_0$ be fixed. Then for every $(\xi, x_0) \in \left[\theta_i, \theta_{i+1}\right] \times \mathbb{R}^2$, there exists a unique solution $x(t) = x(t, \xi, x_0)$ of (2) on $\left[\theta_i, \theta_{i+1}\right]$.*

Proof *Existence*: Fix $i \in \mathbb{N}_0$ and assume without loss of generality that $\theta_i \leq \xi \leq \theta_{i+1}$. Define a norm $\|x(t)\|_0 = \max\limits_{[\theta_i, \xi]} \|x(t)\|$. Take $x^0(t) = x^0$ and a sequence

$$x^{p+1}(t) = x^0 + \int_{\xi}^{t} \begin{bmatrix} x_2^p(s) \\ -\dfrac{c}{m} x_2^p(s) - \dfrac{k}{m} x_1^p(s) + \dfrac{A}{m} x_1^p(\theta_i) \end{bmatrix} ds, \;\; p \geq 0, \;\; t \in [\theta_i, \theta_{i+1}).$$

First, for $p = 0$, we have

$$\begin{aligned} \left\|x^1(t) - x^0(t)\right\| &\leq \max_{[\theta_i, \xi]} \left\|x^1(t) - x^0(t)\right\| \\ &= \left\|x^1(t) - x^0(t)\right\|_0 \\ &\leq \max_{[\theta_i, \xi]} \int_{\xi}^{t} \left\{ \frac{(A-k)^2}{m^2} (x_1^0)^2 + \left(1 + \frac{c^2}{m^2}\right) (x_2^0)^2 + \right. \\ &\qquad \left. + 2 \frac{c\,|A-k|}{m^2} \left|x_1^0\right| \left|x_2^0\right| \right\}^{1/2} ds. \end{aligned}$$

Using $2\,|uv| \leq u^2 + v^2$, the last inequality takes the following form

$$\left\|x^1(t) - x^0(t)\right\|_0 \leq \frac{1}{m} \max_{[\theta_i, \xi]} \int_{\xi}^{t} \sqrt{M \left\{(x_1^0)^2 + (x_2^0)^2\right\}} ds.$$

So, we obtain

$$\left\|x^1(t) - x^0(t)\right\|_0 \leq \frac{1}{m} \max_{[\theta_i, \xi]} \int_{\xi}^{t} M^{1/2} \left\|x^0\right\| ds \leq \frac{1}{m} \overline{\theta} M^{1/2} \left\|x^0\right\|. \tag{15}$$

Similarly, for $p = 0$ and $p = 1$, we get

$$\left\|x^2(t)-x^1(t)\right\|_0=\max_{[\theta_i,\xi]}\int_{\xi}^{t}\left\{\left(1+\frac{c^2}{m^2}\right)\left(x_2^1(s)-x_2^0(s)\right)^2+\frac{k^2}{m^2}\left(x_1^1(s)-x_1^0(s)\right)^2+\right.$$
$$+2\frac{kc}{m^2}\left|x_2^1(s)-x_2^0(s)\right|\left|x_1^1(s)-x_1^0(s)\right|+$$
$$+\frac{A^2}{m^2}\left(x_1^1(\theta_i)-x_1^0(\theta_i)\right)^2+$$
$$+2\frac{|A|}{m^2}\left|x_1^1(\theta_i)-x_1^0(\theta_i)\right|k\left|x_1^1(s)-x_1^0(s)\right|+$$
$$\left.+2\frac{|A|}{m^2}\left|x_1^1(\theta_i)-x_1^0(\theta_i)\right|c\left|x_2^1(s)-x_2^0(s)\right|\right\}^{1/2}ds.$$

Using $2|uv|\le u^2+v^2$, we have

$$\left\|x^2(t)-x^1(t)\right\|_0\le\frac{1}{m}\max_{[\theta_i,\xi]}\int_{\xi}^{t}\left\{\left[k^2+k|A|+ck\right](x_1^1(s)-x_1^0(s))^2\right.$$
$$+\left[m^2+c^2+ck+c|A|\right](x_2^1(s)-x_2^0(s))^2$$
$$\left.+(|A|+c+k)|A|(x_1^1(\theta_i)-x_1^0(\theta_i))^2\right\}^{1/2}ds$$
$$\le\frac{1}{m}\overline{\theta}M^{1/2}\left\|x^1(\theta_i)-x^0(\theta_i)\right\|+$$
$$+\frac{1}{m}\max_{[\theta_i,\xi]}\int_{\xi}^{t}M^{1/2}\left\|x^1(s)-x^0(s)\right\|ds.$$

Thus, by (15), we obtain

$$\left\|x^2(t)-x^1(t)\right\|_0\le\frac{1}{m^2}\overline{\theta}^2M\left\|x^0\right\|+\frac{1}{m^2}\overline{\theta}^2M\left\|x^0\right\|\le2\frac{1}{m^2}\overline{\theta}^2M\left\|x^0\right\|.$$

By induction, it can be easily seen that we reach the conclusion

$$\left\|x^{p+1}(t)-x^p(t)\right\|_0\le2^p\frac{1}{m^{p+1}}\overline{\theta}^{p+1}M^{(p+1)/2}\left\|x^0\right\|=\frac{1}{2}\left\{2\frac{1}{m}\overline{\theta}M^{1/2}\right\}^{(p+1)}\left\|x^0\right\|.$$

Then, the condition (C3) implies that the sequence $x^p(t)$ is convergent and its limit $x(t)$ satisfies

$$x(t)=x^0+\int_{\xi}^{t}\begin{bmatrix}x_2(s)\\-\frac{c}{m}x_2(s)-\frac{k}{m}x_1(s)+\frac{A}{m}x_1(\theta_i)\end{bmatrix}ds$$

on $[\theta_i,\xi]$. The existence is proved.

Uniqueness: Let $x^j(t) = x\left(t, \xi, (x^0)^j\right)$, $x^j(\xi) = (x^0)^j$, $j = 1, 2$, denote the solutions of

$$x'(t) = \begin{bmatrix} x_2(t) \\ -\dfrac{c}{m}x_2(t) - \dfrac{k}{m}x_1(t) + \dfrac{A}{m}x_1(\theta_i) \end{bmatrix}$$

where $\theta_i \leq \xi \leq \theta_{i+1}$. Now, we shall show that $(x^0)^1 \neq (x^0)^2$ implies $x^1(t) \neq x^2(t)$ for every $t \in [\theta_i, \theta_{i+1}]$. The solutions $x^1(t)$ and $x^2(t)$ satisfy, respectively,

$$x^1(t) = (x^0)^1 + \int_\xi^t \begin{bmatrix} x_2^1(s) \\ -\dfrac{c}{m}x_2^1(s) - \dfrac{k}{m}x_1^1(s) + \dfrac{A}{m}x_1^1(\theta_i) \end{bmatrix} ds$$

and

$$x^2(t) = (x^0)^2 + \int_\xi^t \begin{bmatrix} x_2^2(s) \\ -\dfrac{c}{m}x_2^2(s) - \dfrac{k}{m}x_1^2(s) + \dfrac{A}{m}x_1^2(\theta_i) \end{bmatrix} ds,$$

for all $t \in [\theta_i, \theta_{i+1}]$. We can write

$$\begin{aligned} \left\|x^1(t) - x^2(t)\right\| &\leq \left\|(x^0)^1 - (x^0)^2\right\| + \\ &\quad + \frac{1}{m}\int_\xi^t \Big\{(k^2 + ck + k\,|A|)\left(x_1^1(s) - x_1^2(s)\right)^2 + \\ &\qquad + (m^2 + c^2 + ck + c\,|A|)\left(x_2^1(s) - x_2^2(s)\right)^2 + \\ &\qquad + (A^2 + k\,|A| + c\,|A|)\left(x_1^1(\theta_i) - x_1^2(\theta_i)\right)^2\Big\}^{1/2} ds \\ &\leq \left\|(x^0)^1 - (x^0)^2\right\| + \frac{1}{m}\overline{\theta}\sqrt{A^2 + k\,|A| + c\,|A|}\left\|x^1(\theta_i) - x^2(\theta_i)\right\| + \\ &\quad + \frac{1}{m}\int_\xi^t M^{1/2}\left\|x^1(s) - x^2(s)\right\| ds. \end{aligned}$$

By Gronwall-Bellman inequality, we have

$$\begin{aligned} \left\|x^1(t) - x^2(t)\right\| &\leq \left\|(x^0)^1 - (x^0)^2\right\| e^{\frac{1}{m}\overline{\theta}M^{1/2}} + \\ &\quad + \frac{1}{m}\overline{\theta}\sqrt{A^2 + k\,|A| + c\,|A|}\left\|x^1(\theta_i) - x^2(\theta_i)\right\| e^{\frac{1}{m}\overline{\theta}M^{1/2}}. \quad (16) \end{aligned}$$

For $t = \theta_i$, it is true that

$$\left\|x^1(\theta_i) - x^2(\theta_i)\right\| \leq \left\|(x^0)^1 - (x^0)^2\right\| e^{\frac{1}{m}\overline{\theta}M^{1/2}} + \\ + \frac{1}{m}\overline{\theta}\sqrt{A^2 + k\,|A| + c\,|A|}\left\|x^1(\theta_i) - x^2(\theta_i)\right\| e^{\frac{1}{m}\overline{\theta}M^{1/2}}.$$

Hence,

$$\left\|x^1(\theta_i) - x^2(\theta_i)\right\| \leq \frac{e^{\frac{1}{m}\overline{\theta}M^{1/2}} m}{m - \overline{\theta}\sqrt{A^2 + k\,|A| + c\,|A|}e^{\frac{1}{m}\overline{\theta}M^{1/2}}}\left\|(x^0)^1 - (x^0)^2\right\|.$$

Substituting the last inequality in (16) and by rearranging the terms, we obtain

$$\left\|x^1(t) - x^2(t)\right\| \leq \frac{e^{\frac{1}{m}\overline{\theta}M^{1/2}} m}{m - \overline{\theta}\sqrt{A^2 + k\,|A| + c\,|A|}e^{\frac{1}{m}\overline{\theta}M^{1/2}}}\left\|(x^0)^1 - (x^0)^2\right\|. \quad (17)$$

If it is assumed on the contrary that there exists a $t^* \in [\theta_i, \theta_{i+1}]$ such that $x^1(t^*) = x^2(t^*)$, then we get

$$(x^0)^1 - (x^0)^2 = \int_{\xi}^{t^*} \begin{bmatrix} x_2^2(s) - x_2^1(s) \\ \left(-\frac{c}{m}\left(x_2^2(s) - x_2^1(s)\right) - \frac{k}{m}\left(x_1^2(s) - x_1^1(s)\right) + \right. \\ \left. + \frac{A}{m}\left(x_1^2(\theta_i) - x_1^1(\theta_i)\right)\right) \end{bmatrix} ds.$$

By the last equality, we reach

$$\begin{aligned} \left\|(x^0)^1 - (x^0)^2\right\| &\leq \frac{1}{m}\int_{\xi}^{t^*} \left\{\left(x_2^2(s) - x_2^1(s)\right)^2\left[m^2 + c^2 + ck + c\,|A|\right] + \right. \\ &\quad + \left(x_1^2(s) - x_1^1(s)\right)^2\left[k^2 + ck + k\,|A|\right] + \\ &\quad \left. + \left(x_1^2(\theta_i) - x_1^1(\theta_i)\right)^2\left[A^2 + c\,|A| + k\,|A|\right]\right\}^{1/2} ds \\ &\leq \frac{\overline{\theta}\left(\sqrt{A^2 + k\,|A| + c\,|A|} + M^{1/2}\right) e^{\frac{1}{m}\overline{\theta}M^{1/2}}}{m - \overline{\theta}\sqrt{A^2 + k\,|A| + c\,|A|}e^{\frac{1}{m}\overline{\theta}M^{1/2}}}\left\|(x^0)^1 - (x^0)^2\right\| \\ &< \left\|(x^0)^1 - (x^0)^2\right\|, \end{aligned}$$

which is a contradiction due to the condition (C4). □

Then, the following theorem guarantees sufficient conditions for the existence and the uniqueness of the solution of (2) on $\mathbb{D}$, which can be proved by the mathematical induction and in view of Lemma 2.

Theorem 1 *Assume that the conditions (C1), (C3) and (C4) hold true. Then, for every* $(t_0, x_0) \in \mathbb{R}^+ \times \mathbb{R}^2$*, there exists a unique solution* $x(t) = x(t, t_0, x_0)$ *of (2) on* $\mathbb{D}$ *in the sense of Definition 1 such that* $x(t_0) = x_0$.

3 Stability Analysis in the Sense of Lyapunov-Razumikhin Method

Now, we shall consider the Lyapunov-Razumikhin method developed by Akhmet and Aruğaslan [1] for differential equations with PCAG in the following form

$$x'(t) = f\left(t, x(t), x(\beta(t))\right). \tag{18}$$

In the investigation of the system (18) [3–5], a new approach based on the construction of an equivalent integral equation has been used. Definitions of stability for EPCAG coincide with the definitions used for ordinary differential equations [1, 7, 12, 24].

Let us describe special sets as follows:

$$\mathcal{A} = \left\{a \in C\left(\mathbb{R}^+, \mathbb{R}^+\right) : \text{ strictly increasing and } a(0) = 0\right\},$$

and

$$\mathcal{B} = \left\{b \in C\left(\mathbb{R}^+, \mathbb{R}^+\right) : b(0) = 0, b(s) > 0 \text{ for } s > 0\right\}.$$

The technique developed in [1, 30] enables stability analysis by constructing a positive definite Lyapunov function $V(t, x)$ which

(i) is continuous on $\mathbb{D} \times \mathbb{R}$ and $V(t, 0) \equiv 0$ for all $t \in \mathbb{D}$,
(ii) is continuously differentiable on $(\theta_i, \theta_{i+1}) \times \mathbb{R}$ and for each $x \in \mathbb{R}$, the right derivative exists at $t = \theta_i$, $i \in \mathbb{N}_0$;

and by finding conditions giving a negative definite derivative of $V(t, x)$ along the trajectories of (18) whenever there exists a relation between the values of this Lyapunov function at the deviation argument $\beta(t)$ and any time t according to Theorem 3.1 in [1] and Theorem 5 in [30]. Here, the derivative of $V(t, x)$ with respect to system (18) is defined by

$$V'_{(18)}(t, x, y) = \frac{\partial V(t, x)}{\partial t} + \langle \nabla V(t, x), f(t, x, y) \rangle,$$

for all $t \neq \theta_i$ in $\mathbb{D}$ and $x, y \in \mathbb{R}$, where ∇V denotes the gradient vector of V with respect to x [1, 30].

By this method, we investigate the uniform asymptotic stability of the spring-mass system (2) ((1)) with PCAG. The following further assumptions will be needed for the stability analysis in sense of Lyapunov-Razumikhin method:

(C5) $2k - |A| \geq 0$ and $(k + m)(2c - |A|) - 2mk \geq 0$;

$$\text{(C6)}\ \min\left\{\frac{(2k-|A|)k(k+c)}{|A|\,(k+c+2m)(2k+m)},\frac{[(k+m)(2c-|A|)-2mk]\,k(k+c)}{(m+2)^2m\,|A|}\right\}>1.$$

3.1 Theoretical Results Together with the Construction of the Lyapunov Function

Based on the Lyapunov-Razumikhin method developed in paper [1], the next theorem gives sufficient conditions for uniform asymptotic stability of the trivial solution of (2).

Theorem 2 *Assume that the conditions (C1)–(C6) are satisfied. Then, the trivial solution of (2) is uniformly asymptotically stable.*

Proof Consider the following Lyapunov function

$$V(x)=(1+a)\,x_1^2+(1+b)x_2^2+2x_1x_2, \tag{19}$$

where $x=(x_1,x_2)$ and $a=\dfrac{k+c}{m}$, $b=\dfrac{m}{k}$. It is obvious that the Lyapunov function (19) is positive definite:

$$V(x)=ax_1^2+bx_2^2+(x_1+x_2)^2\geq ax_1^2+bx_2^2.$$

Therefore, we can find a constant $\delta_1=\delta_1(m,c,k)>0$ such that $V\geq\delta_1\left(x_1^2+x_2^2\right)$, and thus we can find a function $u\in\mathcal{A}$ which satisfies the inequalities $u(\|x\|)\geq 0$ and $u(\|x\|)\leq V(t,x)$. Besides, a function $v\in\mathcal{A}$ with the property $v(\|x\|)\geq V(t,x)$ can be found:

$$V(x)\leq(1+a)\,x_1^2+(1+b)x_2^2+x_1^2+x_2^2=(2+a)\,x_1^2+(2+b)x_2^2,$$

and so, we can find a constant $\delta_2=\delta_2(m,c,k)>0$ such that $V\leq\delta_2\left\{x_1^2+x_2^2\right\}$. Define $\psi(s)=\sigma s$ which is a continuous nondecreasing function for $s>0$ and define a function $w(u)$ which is given by $w(u)=\dfrac{\aleph}{m}u^2\in\mathcal{B}$. Let us take a constant σ such that

$$1<\sigma<\min\left\{\frac{(2k-|A|)k(k+c)}{|A|\,(k+c+2m)(2k+m)},\frac{[(k+m)(2c-|A|)-2mk]\,k(k+c)}{(m+2)^2m\,|A|}\right\}.$$

Assume that $\aleph$ is a positive constant defined by

$$\aleph=\min\left\{2k-|A|-\sigma\,|A|\left(2+\frac{m}{k}\right)\left(\frac{1}{m}+\frac{2}{k+c}\right)m,\right.$$
$$\left.,\left(1+\frac{m}{k}\right)(2c-|A|)-2m-\sigma m\left(2+\frac{m}{k}\right)^2\frac{|A|}{k+c}\right\}.$$

Let us evaluate the derivative of the Lyapunov function (19) with respect to t, for $t \neq \theta_i,\ i \in \mathbb{N}_0$:

$$\begin{aligned} V'_{(2)}(x, x(\beta(t))) \leq & -x_1^2 \left\{ \frac{2k}{m} \right\} - x_2^2 \left\{ 2(1+b)\frac{c}{m} - 2 \right\} + \\ & + 2x_1 x_2 \left\{ 1 + a - \frac{k}{m}(1+b) - \frac{c}{m} \right\} + \\ & + 2\frac{|A|}{m} |x_1| \, |x_1(\beta(t))| + 2(1+b)\frac{|A|}{m} |x_2| \, |x_1(\beta(t))| . \end{aligned}$$

Here, it can be seen that the coefficient of $x_1 x_2$ is equal to zero. By the inequality $2|u||v| \leq u^2 + v^2$, the last inequality takes the following form

$$\begin{aligned} V'_{(2)}(x, x(\beta(t))) \leq & -x_1^2 \left\{ \frac{2k}{m} \right\} - x_2^2 \left\{ 2(1+b)\frac{c}{m} - 2 \right\} + \\ & + x_1^2 \frac{|A|}{m} + x_2^2 (1+b)\frac{|A|}{m} + (2+b)\frac{|A|}{k+c} a x_1^2(\beta(t)) \\ = & -x_1^2 \left\{ \frac{2k}{m} - \frac{|A|}{m} \right\} - x_2^2 \left\{ (1+\frac{m}{k})\frac{2c-|A|}{m} - 2 \right\} + \\ & + \left(2 + \frac{m}{k} \right) \frac{|A|}{k+c} a x_1^2(\beta(t)) \\ \leq & -x_1^2 \left\{ \frac{2k}{m} - \frac{|A|}{m} \right\} - x_2^2 \left\{ (1+\frac{m}{k})\frac{2c-|A|}{m} - 2 \right\} + \\ & + \left(2 + \frac{m}{k} \right) \frac{|A|}{k+c} \left(a x_1^2(\beta(t)) + b x_2^2(\beta(t)) \right) + \\ & + \left(2 + \frac{m}{k} \right) \frac{|A|}{k+c} \left(x_1(\beta(t)) + x_2(\beta(t)) \right)^2 . \end{aligned}$$

Now, we can complete the proof based on the Theorem 3.2.3 in [1]. Therefore, we have

$$\begin{aligned} V'_{(2)}(x, x(\beta(t))) \leq & -x_1^2 \left\{ \frac{2k-|A|}{m} - \sigma |A| \left(2 + \frac{m}{k} \right) \left(\frac{1}{m} + \frac{2}{k+c} \right) \right\} \\ & - x_2^2 \left\{ (1+\frac{m}{k})\frac{2c-|A|}{m} - 2 - \sigma \left(2 + \frac{m}{k} \right)^2 \frac{|A|}{k+c} \right\} \\ \leq & -\frac{\aleph}{m} \left(x_1^2 + x_2^2 \right) \\ = & -w(\|x\|) \end{aligned}$$

whenever $V(x(\beta(t))) \leq \psi(V(x(t)))$ according to Theorem 3.2.3 in [1]. In other words, the trivial solution of (2) ((1)) is uniformly asymptotically stable whenever $a x_1^2(\beta(t)) + b x_2^2(\beta(t)) + (x_1(\beta(t)) + x_2(\beta(t)))^2 \leq \sigma a x_1^2 + \sigma b x_2^2 + \sigma (x_1 + x_2)^2$ by (3). This completes the proof. □

4 Illustrative Examples

In this section, taking the obtained theoretical results into account, we give some examples together with simulations using the MATLAB package program.

Example 1 Let $m = 3$, $c = 15$, $k = 3$, $A = 0.03$ in (1) and let $\theta_i = \frac{i}{40} + (-1)^i \frac{1}{120}$, $i \in \mathbb{N}_0$. So consider the following spring-mass system

$$\begin{cases} x_1'(t) = x_2(t) \\ x_2'(t) = -5x_2(t) - x_1(t) - 0.01x_1(\beta(t)) \end{cases} \tag{20}$$

with PCAG. Assume that the solutions $x_i(t)$ start at initial points $x_i(0.01) = 0.1$, $i = 1, 2$. By simple calculation, we obtain $\overline{\theta} = \frac{5}{120}, \overline{\theta}\Omega\frac{1}{m}(c + \overline{\theta}k) = 0.25917606 < 1$, $\frac{\overline{\theta}^2}{m}(k + |A|)\Omega = 0.0021633704 < 1$, $\mu\kappa = 0.10166956 < 1$; $2\frac{1}{m}\overline{\theta}M^{1/2} = 0.46435439 < 1$; $\overline{\theta}e^{\frac{1}{m}\overline{\theta}M^{1/2}}\left(2\sqrt{A^2 + k|A| + c|A|} + M^{1/2}\right) = 0.95587091 < m = 3$. So, the conditions (C1)–(C4) hold true, and it follows from Lemma 2 and Theorem 1 that there exists a unique solution with the initial value $(x_1(0.01), x_2(0.01)) = (0.1, 0.1)$. Additionally, we have $2k - |A| = 5.97 \geq 0$ and $(k + m)(2c - |A|) - 2mk = 161.82 \geq 0$;

$$\min\left\{\frac{(2k - |A|)k(k + c)}{|A|(k + c + 2m)(2k + m)}, \frac{[(k + m)(2c - |A|) - 2mk]k(k + c)}{(m + 2)^2 m|A|}\right\} = 49.75 > 1.$$

Thus, (C5)–(C6) are satisfied. We can choose $\sigma = 49$ which satisfies the inequality $1 < \sigma < 49.75$. Hence, the conditions of Theorem 2 are satisfied, and Lyapunov-Razumikhin technique says that the trivial solution of (20) is uniformly asymptotically stable as seen by Fig. 1.

Example 2 Let $m = 3$, $c = 1.6$, $k = 3$, $A = 0.1$ in (1) and let $\theta_i = \frac{i}{40} + (-1)^i \frac{1}{120}$, $i \in \mathbb{N}_0$. So, consider the following spring-mass system

$$\begin{cases} x_1'(t) = x_2(t) \\ x_2'(t) = -\frac{1.6}{3}x_2(t) - x_1(t) - \frac{0.1}{3}x_1(\beta(t)). \end{cases} \tag{21}$$

By simple calculation, we obtain $\overline{\theta} = \frac{5}{120}$, $\overline{\theta}\Omega\frac{1}{m}(c + \overline{\theta}k) = 0.0245393 < 1$, $\frac{\overline{\theta}^2}{m}(k + |A|)\Omega = 0.0018375 < 1$, $\mu\kappa = 0.0484045 < 1$; $2\frac{1}{m}\overline{\theta}M^{1/2} = 0.1129022$

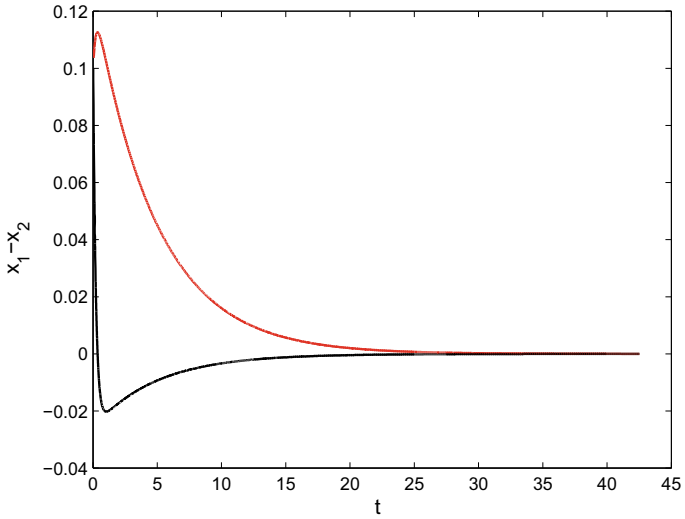

Fig. 1 Time response of state variables $x_1(t)$ (the red one) and $x_2(t)$ (the black one) of (20) for $t \in [0.01, 42.50833333]$ while $m = 3$, $c = 15$, $k = 3$, 0.03, $\theta_i = \dfrac{i}{40} + (-1)^i \dfrac{1}{120}$, $i \in \mathbb{N}_0$, at $(x_1(0.01), x_2(0.01)) = (0.1, 0.1)$

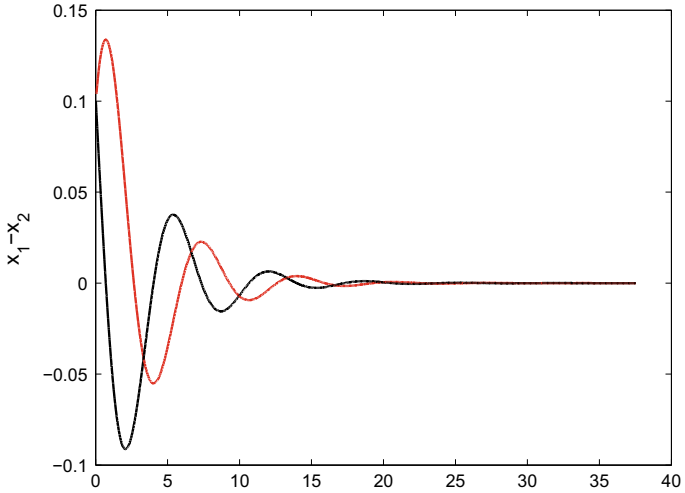

Fig. 2 Time response of state variables $x_1(t)$ (the red one) and $x_2(t)$ (the black one) of (21) for $t \in [0.01, 37.50833333]$ while $m = 3$, $c = 1.6$, $k = 3$, $A = 0.1$, $\theta_i = \dfrac{i}{40} + (-1)^i \dfrac{1}{120}$, $i \in \mathbb{N}_0$, at $(x_1(0.01), x_2(0.01)) = (0.1, 0.1)$

$< 1; \overline{\theta} e^{\frac{1}{m}\overline{\theta} M^{1/2}} \left(2\sqrt{A^2 + k\,|A| + c\,|A|} + M^{1/2}\right) = 0.2396368 < m = 3$. So, the conditions (C1)–(C4) are fulfilled, and it follows from Lemma 2 and Theorem 1 that there exists a unique solution with the initial value $(x_1(0.01), x_2(0.01)) = (0.1, 0.1)$. We have $2k - |A| = 5.9 \geq 0$ and $(k+m)(2c - |A|) - 2mk = 0.6 \geq 0$; $\min\left\{\dfrac{(2k-|A|)k(k+c)}{|A|\,(k+c+2m)(2k+m)}, \dfrac{[(k+m)(2c-|A|) - 2mk]\,k(k+c)}{(m+2)^2 m\,|A|}\right\} = 1.104 > 1$. Thus, (C5)–(C6) are satisfied. We can choose $\sigma = 1.1$ which satisfies the inequality $1 < \sigma < 1.104$. Hence, the conditions of Theorem 2 are satisfied, and Lyapunov-Razumikhin technique says that the trivial solution of (21) is uniformly asymptotically stable. The simulation showing the uniform asymptotic stability of the trivial solution of (21) is given in Fig. 2.

5 Conclusion

In the present work, we address a damped spring-mass system which is modeled by a piecewise function, i.e. by PCAG, as a deviating argument. Since PCAG is a more general argument than the greatest integer function, we get a more general system and thus obtain more general results for the mechanical system mathematically. Thus, when examining stability of the system's behavior, it is possible to obtain more appropriate results in terms of the reality of the model. Although this argument develops the systems, it makes difficult to find the exact solution of such systems. Therefore, it is very important to be able to analyze the system without finding its solution. In this case, Lyapunov methods are very useful [1, 26]. In our study, Lyapunov-Razumikhin method [1], which was developed for analysis of applications where there is a PCAG, is considered. This method is more applicable than other methods available in the literature for similar analyses. With the help of this method, i.e. the Lyapunov-Razumikhin method for EPCAG, results can be achieved more efficiently. In order to observe this, studies conducted, for instance, by Lyapunov-Krasovskii method, in the same direction can be examined. Moreover, in the present paper, it is of great importance that the achieved results are supported by examples and simulations. In future studies, the mechanical model considered in this study can be developed by the argument $\gamma(t)$ which is introduced by Akhmet [8]. The argument function $\gamma(t)$ is of the alternate type which means that it is both advanced type and delayed type. In our paper, the argument function $\beta(t)$ is of delayed type only.

Acknowledgements This work is supported by TÜBİTAK (The Scientific and Technological Research Council of Turkey) under project no 118F185 and supported partially by TÜBİTAK under project no 118F161. The authors want to express their sincere gratitude to the referee for the valuable remarks that improve the paper.

References

1. Akhmet, M.U., Aruğaslan, D.: Lyapunov-Razumikhin method for differential equations with piecewise constant argument. Discrete and Continuous Dynamical Systems. Series A, vol. 25(2), pp. 457–466 (2009)
2. Akhmet, M.U.: On the integral manifolds of the differential equations with piecewise constant argument of generalized type. In: Agarval, R.P., Perera, K. (eds.), Proceedings of the Conference on Differential and Difference Equations at the Florida Institute of Technology, pp. 11–20. Hindawi Publishing Corporation, Melbourne, Florida (2005)
3. Akhmet, M.U.: Integral manifolds of differential equations with piecewise constant argument of generalized type. Nonlinear Anal.: Theory Methods Appl. **66**, 367–383 (2007)
4. Akhmet, M.U.: On the reduction principle for differential equations with piecewise constant argument of generalized type. J. Math. Anal. Appl. **336**, 646–663 (2007)
5. Akhmet, M.U.: Stability of differential equations with piecewise constant arguments of generalized type. Nonlinear Anal. **68**, 794–803 (2008)
6. Akhmet, M.U.: Principles of Discontinuous Dynamical Systems. Springer, New York (2010)
7. Akhmet, M.U.: Nonlinear Hybrid Continuous Discrete-Time Models. Atlantis Press, Amsterdam-Paris (2011)
8. Akhmet, M.U.: Quasilinear retarded differential equations with functional dependence on piecewise constant argument. Commun. Pure Appl. Anal. **13**(2), 929–947 (2014)
9. Akhmet, M.U., Buyukadali, C.: Differential equations with a state-dependent piecewise constant argument. Nonlinear Anal.: TMA **72**, 4200–4210 (2010)
10. Akhmet, M.U., Arugaslan, D., Yilmaz, E.: Method of Lyapunov functions for differential equations with piecewise constant delay. J. Comput. Appl. Math. **235**, 4554–4560 (2011)
11. Alonso, A., Hong, J., Obaya, R.: Almost periodic type solutions of differential equations with piecewise constant argument via almost periodic type sequences. Appl. Math. Lett. **13**, 131–137 (2000)
12. Aruğaslan, D.: Differential Equations with Discontinuities and Population Dynamics. Ph.D. dissertation, Middle East Technical University (2009)
13. Aruğaslan, D., Cengiz, N.: Green's function and periodic solutions of a spring-mass system in which the forces are functionally dependent on piecewise constant argument. Süleyman Demirel Univ. J. Nat. Appl. Sci. **21**(1), 266–278 (2017)
14. Aruğaslan, D., Cengiz, N.: Existence of periodic solutions for a mechanical system with piecewise constant forces. Hacet. J. Math. Stat. **47**(3), 521–538 (2018)
15. Aruğaslan, D., Özer, A.: Stability analysis of a predator-prey model with piecewise constant argument of generalized type using Lyapunov functions. Nonlinear Oscillations **16**(4), 452–459 (2013)
16. Bykov, J.A.V., Salpagarov, H.M.K.: Theorii Integrodifferential'nyh Uravnenijam v Kirgizii. Izd-vo AN Kirg. SSR (2) (1962)
17. Cooke, K.L., Wiener, J.: Retarded differential equations with piecewise constant delays. J. Math. Anal. Appl. **99**, 265–297 (1984)
18. Dai, L.: Nonlinear Dynamics of Piecewise Constant Systems and Implementation of Piecewise Constant Arguments. World Scientific, Hackensack, NJ (2008)
19. Dai, L., Singh, M.C.: On oscillatory motion of spring-mass systems subjected to piecewise constant forces. J. Sound Vib. **173**(2), 217–231 (1994)
20. Dragomir, S.S.: Some Gronwall Type Inequalities and Applications. Nova Science Publishers Inc., Hauppauge, NY (Reviewer: B. G. Pachpatte) (2003)
21. Filatov, A., Šarova, L.: Integral'nye Neravenstva i Teorija Nelineinyh Kolebanii. Moskva (1976)
22. Hale, J.: Functional Differential Equations. Springer, New-York (1971)
23. Hale, J.K.: Theory of Functional Differential Equations. Springer, New York, Heidelberg, Berlin (1997)
24. Hartman, P.: Ordinary Differential Equations. Society for Industrial and Applied Mathematics, Philadelphia (2002)

25. Pinto, M.: Asymptotic equivalence of nonlinear and quasi linear differential equations with piecewise constant arguments. Math. Comput. Model. **49**, 1750–1758 (2009)
26. Razumikhin, B.S.: Stability of delay systems. Prikl. Mat. Mekh. **20**, 500–512 (1956)
27. Shah, S.M., Wiener, J.: Advanced differential equations with piecewise constant argument deviations. Int. J. Math. Math. Sci. **6**, 671–703 (1983)
28. Wang, Y., Yan, J.: A necessary and sufficient condition for the oscillation of a delay equation with continuous and piecewise constant arguments. Acta Math. Hungar. **79**, 229–235 (1998)
29. Wiener, J., Cooke, K.L.: Oscillations in systems of differential equations with piecewise constant argument. J. Math. Anal. Appl. **137**, 221–239 (1989)
30. Xi, Q.: Razumikhin-type theorems for impulsive differential equations with piecewise constant argument of generalized type. Adv. Differ. Eq. **2018**, 267 (2018)

A Darwinian Ricker Equation

Jim M. Cushing

Abstract The classic Ricker equation $x_{t+1} = bx_t \exp(-cx_t)$ has positive equilibria for $b > 1$ that destabilize when $b > e^2$ after which its asymptotic dynamics are oscillatory and complex. We study an evolutionary version of the Ricker equation in which coefficients depend on a phenotypic trait subject to Darwinian evolution. We are interested in the question of whether evolution will select against or will promote complex dynamics. Toward this end, we study the existence and stability of its positive equilibria and focus on equilibrium destabilization as an indicator of the onset of complex dynamics. We find that the answer relies crucially on the speed of evolution and on how the intra-specific competition coefficient c depends on the evolving trait. In the case of a hierarchical dependence, equilibrium destabilization generally occurs after e^2 when the speed of evolution is sufficiently slow (in which case we say evolution selects against complex dynamics). When evolution proceeds at a faster pace, destabilization can occur before e^2 (in which case we say evolution promotes complex dynamics) provided the competition coefficient is highly sensitive to changes in the trait v. We also show that destabilization does not always result in a period doubling bifurcation, as in the non-evolutionary Ricker equation, but under certain circumstances can result in a Neimark-Sacker bifurcation.

Keywords Ricker equation · Darwinian Ricker equation · Chaos · Evolutionary game theory

1 Introduction

It is well known that difference equations can predict complex asymptotic dynamics in the form of non-equilibrium attractors. The exponential or Ricker equation

J. M. Cushing (✉)
University of Arizona, Tucson, AZ 85721, USA
e-mail: cushing@math.arizona.edu
URL: https://www.math.arizona.edu/cushing/

S. Baigent et al. (eds.), *Progress on Difference Equations and Discrete Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 341,
https://doi.org/10.1007/978-3-030-60107-2_10

$$x_{(t+1)} = bx_{(t)} \exp(-cx_{(t)}) \tag{1}$$

is the iconic example of a period doubling route to chaos which, as $b > 1$ increases, initiates after $b = e^2$ where the positive equilibrium $x = c^{-1} \ln b$ destabilizes. Despite the ubiquity of this phenomenon in difference equations used as population dynamic models, unequivocal evidence of its occurrence in biological populations is sparse and is, for the most part, limited to populations manipulated in laboratory settings [10]. Several explanations for this can be found in the literature. One is that population time series data tends to be too short to be able to identify complex dynamics and data is usually "noisy" and, as a result, it is difficult to tell the difference between stochastic fluctuations and deterministic fluctuations (such as chaos) [5, 8]. Another explanation is that most populations in the natural world are subject to interactions with other species that can serve to dampen complex dynamics [7]. Yet another explanation is that biological populations are subject to evolutionary change by Darwinian principles and that evolution might select to reduce dynamic complexity, i.e. non-equilibrium dynamics such as periodic oscillations or chaos [4]. In this paper we briefly consider the latter possibility by subjecting the parameters in the Ricker equation (1) to evolutionary changes according to a methodology called evolutionary game theory (or Darwinian dynamics) [9]. This derivation results in a system of difference equations that we refer to as a Darwinian Ricker model. In this short note, we do not strive to carry out a study of the non-equilibrium dynamics that are possible in Darwinian Ricker equations, but instead focus simply on whether or not positive equilibria destabilize for b greater than some critical value and, if they do, whether the critical value is greater or less than e^2. If it is greater than e^2, then we say that evolution selects against non-equilibrium and complex dynamics in the sense that the de-stabilization of the equilibrium occurs for larger values of b than it does when evolution is absent. If the critical value of b occurs before e^2, then we say that evolution promotes non-equilibrium and complex dynamics.

Darwinian Ricker model equations are derived by evolutionary game theoretic methods in Sect. 2. The existence and stability (by linearization) of equilibria of this system of two nonlinear difference equations are studied in Sect. 3. Conclusions obtained from this analysis with regard to the effect of evolution on non-equilibrium dynamics are discussed in Sect. 4.

2 A Darwinian Ricker Equation

In the Ricker equation (1) x represents the total size or density of a population consisting of individual biological organisms. We interpret b as the inherent (i.e. density free) per capita fertility rate. The coefficient c is a measure of the effect that increased population density has on the per capita fertility rate, as might be due to competition with con-specifics for resources (food, space, mates, etc.). We refer to c as the competition coefficient. We assume that both b and c, as coefficients relating to an individual's inherent fertility and susceptibility to intra-specific competition

respectively, are functions of a phenotypic trait of the individual, denoted by v, that is subject to evolutionary change over time. Under the axioms of Darwinian evolution (trait variability, heritability, and differential trait dependent fitness), the method of evolutionary game theory [9] provides a dynamic model for the population density and the population's mean phenotypic trait, under the assumption that the trait has a Gaussian distribution with fixed variance throughout the population at all times. Thus, the distribution of the trait v in the population at any point in time is determined by the population mean trait, which we denote by u.

In the Ricker equation we assume the fertility rate b is a function of v alone, since it is the density free fertility rate of an individual with trait v (i.e. not subject to the presence of other individuals and hence to the population mean u). The competition coefficient c, on the other hand, we assume is dependent on the individual's trait v and that of other individuals with whom it competes, as represented by the mean trait u. Thus we assume

$$b = b(v), \quad c = c(v, u).$$

The density dependent fertility rate is then

$$r(x, v, u) = b(v) \exp(-c(v, u)\, x). \tag{2}$$

The Darwinian equations governing both population and mean trait dynamics are

$$x_{t+1} = r(x_t, v, u_t)|_{v=u_t}\, x_t \tag{3}$$

$$u_{t+1} = u_t + \sigma^2 \left.\frac{\partial \ln r(x_t, v, u_t)}{\partial v}\right|_{v=u_t} \tag{4}$$

where $\sigma^2 \geq 0$ is called the speed of evolution (it is proportional to the constant variance of v) [2, 9]. The trait equation (4) says that the change in mean trait is proportional to the fitness gradient, with fitness taken to be $\ln r$ (the equation is often called Lande's or Fisher's equation or the canonical equation of evolution).

To further specify the model, we will place assumptions on $b(v)$ and $c(v, u)$. In this paper we assume that there is a trait at which inherent fertility has a maximum, denoted by b_0, and we choose that trait to be the reference point for v. We also assume that fertility $b(v)$ is distributed in a Gaussian fashion around its maximum b_0 $v = 0$ and, without loss in generality, we scale the trait v so that the variance of $b(v)$ equals 1:

$$b(v) = b_0 \exp\left(-\frac{v^2}{2}\right). \tag{5}$$

With (2) and this choice for $b(v)$, the Darwinian equations (3)–(4) become

$$x_{t+1} = b_0 \left(\exp\left(-\frac{v}{2}\right) \exp\left(-c\left(v, u_t\right) x_t\right)\right)\Big|_{v=u_t} x_t \tag{6}$$

$$u_{t+1} = u_t + \sigma^2 \left(-u_t - \frac{\partial c\left(v, u_t\right)}{\partial v}\Big|_{v=u_t} x_t \right). \tag{7}$$

A common assumption that is made concerning trait dependency of competition coefficients in Darwinian models is that they are functions of the difference $v - u$. In other words, the competition that an individual experiences depends on how different its trait v is from the typical individual in the population, as represented by the mean trait u. We make this assumption here and write $c = c\left(v - u\right)$ where the function $c\left(z\right)$ is continuously differentiable for all values of its argument z. Under this assumption equations (6)–(7) become

$$x_{t+1} = b_0 \left(\exp\left(-\frac{u_t^2}{2}\right) \exp\left(-c\left(0\right) x_t\right)\right) x_t$$

$$u_{t+1} = u_t + \sigma^2 \left(-u_t - \frac{dc\left(z\right)}{dz}\Big|_{z=0} x_t \right).$$

As a final scaling, we assume population units for x are chosen so that $c\left(0\right) = 1$ and obtain the model equations

$$x_{t+1} = b_0 \exp\left(-\frac{u_t^2}{2}\right) \exp\left(-x_t\right) x_t \tag{8}$$

$$u_{t+1} = -c_1 \sigma^2 x_t + \left(1 - \sigma^2\right) u_t \tag{9}$$

where

$$c_1 := \frac{dc\left(z\right)}{dz}\Big|_{z=0}.$$

There are three coefficients in the Eqs. (8)–(9). The coefficient b_0 is the *maximal possible fertility rate*, as a function of the trait v, and the coefficient σ^2 is the *speed of evolution*. The coefficient c_1 is the sensitivity of the competition competition $c\left(z\right)$ to changes in the difference $z = v - u$ at when $v = u$. If $c_1 \neq 0$ then c_1 measures the difference between the competition intensities experienced by individuals that have the population mean trait and those whose traits are slightly different from the mean. For example, if $c_1 > 0$ then an individual that inherits a trait slightly larger (smaller) than the mean u will experience increased (decreased) intraspecific competition. These interpretations can also hold, of course, if $c_1 = 0$ unless $c\left(z\right)$ has an extrema at $z = 0$. In fact, a common modeling assumption is that maximum competition is experience by individuals with the population mean trait, in which case $c\left(z\right)$ has a maximum at $z = 0$ and $c_1 = 0$. A commonly used model for $c\left(z\right)$ assumes it has a Gaussian type distribution

$$c(z) = \exp\left(-\frac{z^2}{2\omega^2}\right) \tag{10}$$

(with variance ω^2). In contrast, if for example

$$c(z) = \exp(c_1 z) \tag{11}$$

then competition intensity either decreases as v decreases or increases from the mean u, depending on the sign of c_1. We refer to this type of competition coefficient $c(z)$, i.e. one for which $c_1 \neq 0$ as heirarchical.

3 Equilibria of the Darwinian Ricker

Our goal is the study the existence and stability properties of equilibria of the Darwinian Ricker equations (8)–(9) using b_0 as a bifurcation parameter. We are interested in equilibria (x, u) with a positive x-component, which we define to be a *positive equilibrium* pair. The equations for a positive equilibrium pair are

$$\begin{aligned} 1 &= b_0 \exp\left(-\frac{u^2}{2}\right)\exp(-x) \\ 0 &= -c_1 x - u. \end{aligned}$$

If $b_0 < 1$, then one sees from the first equation that there is no positive equilibrium (x, u). However, if $b_0 > 1$ then there exists a unique positive equilibrium obtained from the equations

$$1 = b_0 \exp\left(-\frac{c_1^2 x^2}{2}\right)\exp(-x), \qquad u = -c_1 x \tag{12}$$

The positive root of

$$\frac{c_1^2 x^2}{2} + x = \ln b_0 \tag{13}$$

yields the formulas for positive equilibria:

$$(x(b_0), u(b_0)) = \begin{cases} (\ln b_0, 0) & \text{if } c_1 = 0 \\ \left(\frac{-1+\sqrt{1+2c_1^2 \ln b_0}}{c_1^2}, \frac{1-\sqrt{1+2c_1^2 \ln b_0}}{c_1}\right) & \text{if } c_1 \neq 0 \end{cases}. \tag{14}$$

The Jacobian of Eqs. (8)–(9)

$$J(x, u) = \begin{pmatrix} b_0 e^{-\frac{1}{2}u^2} e^{-x}(1-x) & -u x b_0 e^{-\frac{1}{2}u^2} e^{-x} \\ -c_1\sigma^2 & 1-\sigma^2 \end{pmatrix}$$

evaluated at the positive equilibrium becomes, when Eq. (12) are utilized, is

$$J\left(x\left(b_0\right), u\left(b_0\right)\right) = \begin{pmatrix} 1 - x\left(b_0\right) & c_1 x^2\left(b_0\right) \\ -c_1\sigma^2 & 1-\sigma^2 \end{pmatrix}$$

which by (13), further simplifies to

$$J\left(x\left(b_0\right), u\left(b_0\right)\right) = \begin{pmatrix} 1 - x\left(b_0\right) & \frac{2}{c_1}\left(\ln b_0 - x\left(b_0\right)\right) \\ -c_1\sigma^2 & 1-\sigma^2 \end{pmatrix} \tag{15}$$

Motivated by the question posed in Sect. 1 we are interested in the case when the positive equilibria are stable for $b_0 > 1$, but near 1 and destabilize at some value of $b_0 > 1$. For b_0 near 1 the eigenvalues of the Jacobian $J\left(x\left(b_0\right), u\left(b_0\right)\right)$ are

$$\begin{aligned} \lambda_1\left(b_0\right) &= 1 - \left(b_0 - 1\right) + O\left(\left(b_0 - 1\right)^2\right) \\ \lambda_2\left(b_0\right) &= \sigma^2 - 1 + O\left(\left(b_0 - 1\right)^2\right). \end{aligned}$$

It follows by the Linearization Principle that for b_0 greater than, but near 1, the equilibria $\left(x\left(b_0\right), u\left(b_0\right)\right)$ are stable if $\sigma^2 < 2$ and unstable if $\sigma^2 > 2$. Therefore, *we will assume that* $\sigma^2 < 2$.

For the case $c_1 = 0$ the eigenvalues of this Jacobian are

$$\lambda_1 = 1 - \ln b_0 \quad \text{and} \quad \lambda_2 = 1 - \sigma^2$$

and the destabilization of the positive equilibrium occurs at the same critical value as does the classic Ricker equation (1).

Theorem 1 *Assume* $c_1 = 0$ *and* $\sigma^2 < 2$ *in the Darwinian Ricker equations (8)–(9). There exists positive equilibrium for and only for* $b_0 > 1$. *They are locally asymptotically stable if* $1 < b_0 < e^2$ *and unstable if* $b_0 > e^2$. *When* $b_0 = e^2$ *the Jacobian has eigenvalue value* -1.

In general when $c_1 = 0$, the trait equation (9) decouples from the population equation (8) and $\lim_{t\to+\infty} u_t = 0$ under the assumption $\sigma^2 < 2$. In this case, the population equation (8) is asymptotically autonomous and the classic Ricker (1) is its limiting equation. This fact allows for further analysis of the dynamics of the Darwinian Ricker model [1, 6], but we will not pursue further analysis here . Note that Theorem 1 applies when the competition coefficient has the Gaussian form (10).

Consider now the case $c_1 \neq 0$. To study the eigenvalues of the Jacobian we employ the trace and determinant criteria which imply both eigenvalues have magnitude <1 if and only if the three inequalities

$$tr\,J\left(x, u\right) < 1 + \det J\left(x, u\right) \tag{16}$$

$$-1 - \det J\left(x, u\right) < tr\,J\left(x, u\right) \tag{17}$$

$$\det J\,(x,u) < 1 \tag{18}$$

all hold [3]. If inequality (16) or (17) become equalities, then the Jacobian has an eigenvalue equal to $+1$ or -1 respectively. If inequality (18) becomes an equality, then the Jacobian has a complex eigenvalue whose absolute value equals 1.

For (15) we have

$$tr\,J\,(x\,(b_0)\,,u\,(b_0)) = 2 - x\,(b_0) - \sigma^2 \tag{19}$$

$$\det\,J\,(x\,(b_0)\,,u\,(b_0)) = (1 - x\,(b_0))\left(1-\sigma^2\right) + 2\sigma^2\,(\ln b_0 - x\,(b_0))\,. \tag{20}$$

Lemma 1 *Assume $c_1 \neq 0$ in the Darwinian Ricker equations (8)–(9). Inequality (16) holds for all σ^2 and $b_0 > 1$.*

Proof Using (19) and (20), it is easy to show that inequality (16) reduces to $x\,(b_0) < 2\ln b_0$. From the Formula (14) and $c_1 \neq 0$, this inequality is

$$\frac{-1+\sqrt{1+2c_1^2\ln b_0}}{c_1^2} < 2\ln b_0$$

or $\sqrt{1+2c_1^2\ln b_0} < 1 + 2c_1^2\ln b_0$, which is clearly true and completes the proof.

Next we turn attention to inequality (17).

Lemma 2 *Assume $c_1 \neq 0$ and $\sigma^2 < 2$ in the Darwinian Ricker equations (8)–(9).*

(a) If

$$\sigma^2 < \frac{2}{1+8c_1^2} \tag{21}$$

then there exist a real $b_2 > e^2$ such that inequality (17) holds for b_0 satisfying $1 < b_0 < b_2$. Inequality (17) is reversed if b_0 is greater than but near b_2. The Jacobian $J\,(x\,(b_n)\,,u\,(b_n))$ has eigenvalue -1.

(b) If

$$\sigma^2 > \frac{2}{1+8c_1^2} \tag{22}$$

then inequality (17) holds for all $b_0 > 1$.

Proof Using (19) and (20) together with the equilibrium formulas (14), one can re-arrange inequality (17) to the inequality

$$\left(2+\sigma^2\right)\sqrt{2c_1^2z+1} < 2 + \sigma^2 + 2c_1^2\left(2-\sigma^2\right) + 2\sigma^2c_1^2z$$

where we have defined

$$z = \ln b_0 > 0.$$

Since both sides are positive, we can retain the inequality by squaring both sides, after which we re-arrange the result into an equivalent inequality $0 < q_1(z)$ where $q_1(z)$ is the quadratic polynomial

$$q_1(z) := 2c_1^2\left(2-\sigma^2\right)\left(\sigma^2+2+\left(2-\sigma^2\right)c_1^2\right) - c_1^2\left(2-\sigma^2\right)\left(2+\left(1-4c_1^2\right)\sigma^2\right)z + 2\sigma^4c_1^4z^2.$$

The quadratic $q_1(z)$ has a global minimum

$$q_1(z_c) = \frac{1}{8\sigma^4}\left(2-\sigma^2\right)\left(\sigma^2+2\right)^2\left(1+8c_1^2\right)\left(\sigma^2-\frac{2}{1+8c_1^2}\right)$$

attained at the critical point

$$z_c = \frac{2-\sigma^2}{4\sigma^4c_1^2}\left(2+\sigma^2\left(1-4c_1^2\right)\right).$$

(a) Inequality (21) implies $q_1(z_c) < 0$ and hence the existence of two real roots of $q_1(z)$. Since $q_1(0) = 2c_1^2\left(2-\sigma^2\right)\left(2+\sigma^2+\left(2-\sigma^2\right)c_1^2\right) > 0$, it follows that the two roots are both negative or both positive, depending on whether $z_c < 0$ or $z_c > 0$ respectively. Clearly $z_c > 0$ if $1-4c_1^2 \geq 0$. Suppose, on the other hand, that $1-4c_1^2 < 0$. Then $z_c > 0$ if and only if $\sigma^2 < 2\left(4c_1^2-1\right)^{-1}$ which holds by (21) since $\left(4c_1^2-1\right)^{-1} > \left(1+8c_1^2\right)^{-1}$. Thus, in this case, $q_1(z)$ has two positive roots. If we denote the smaller by z_2 then $0 < q_1(z)$ for $0 < z < z_2$ and $q_1(z)$ changes sign as z increases through z_2. Since $q_1(z_2) = 0$ inequality (17) becomes an equality which means the Jacobian has an eigenvalue of -1. Finally we need to show that $z_2 > 2$. One way to do this is to show $q_1(2) > 0$ and $q_1'(2) < 0$. Calculations in fact show $q_1(2) = 2c_1^4\left(\sigma^2+2\right)^2 > 0$ and, using (21),

$$q_1'(2) = c_1^2\left(\sigma^2+2\right)\left(\sigma^2+4\sigma^2c_1^2-2\right) < -\frac{1}{2}c_1^2\left(\sigma^2+2\right)\left(2-\sigma^2\right) < 0.$$

(b) Inequality (22) implies $q_1(z_c) > 0$ and hence $q_1(z) > 0$ for all z. This completes the proof.

Finally we consider inequality (18).

Lemma 3 *Assume $c_1 \neq 0$ and $\sigma^2 < 2$ in the Darwinian Ricker equations (8)–(9). There exists a real $b_n > \exp(1/2)$ such that inequality (18) holds for $1 < b_0 < b_n$. The Jacobian $J(x(b_n), u(b_n))$ has a complex eigenvalue of absolute value 1. The inequality (18) is reversed for $b_0 > b_n$.*

Proof Inequality (18) can be re-arranged as

$$\sigma^2(2\ln b_0 - 1) < x(b_0)\left(\sigma^2+1\right)$$

which is true for $1 < b_0 < \exp(1/2)$. For $b_0 > \exp(1/2)$ we use the Formula (14) for $x(b_0)$ and re-arrange the inequality as

$$1 + (2z - 1)\frac{\sigma^2}{(\sigma^2 + 1)}c_1^2 < \sqrt{1 + 2c_1^2 z}$$

where $z = \ln b_0 > 1/2$. Since both sides are positive, we can square them and re-arrange the inequality to obtain an equivalent inequality

$$0 < q_2(z) := \sigma^2\left(2\sigma^2 - \sigma^2 c_1^2 + 2\right) + 2\left(-\sigma^4 + 2\sigma^4 c_1^2 + 1\right)z - 4\sigma^4 c_1^2 z^2$$

Since $q_2(1/2) = \left(\sigma^2 + 1\right)^2 > 0$, this quadratic polynomial has a unique positive root $z_n > 1/2$ and $q_2(z) > 0$ for $1/2 < z < z_n$. Since $q_2(z_n) = 0$, inequality (18) becomes an equality, which implies the Jacobian has a complex eigenvalue of absolute value 1. This completes the proof.

In Lemma 2(a), the real b_2 is equal to $\exp(z_2)$ where z_2 is the smaller of the positive roots of $q_1(z)$. When (21) holds.a formula for $z_2 > 2$ is

$$z_2 = \frac{\left(2 - \sigma^2\right)\left(2 + \sigma^2 - 4\sigma^2 c_1^2\right) - \left(\sigma^2 + 2\right)\sqrt{\left(2 - \sigma^2\right)\left(2 - \sigma^2 - 8\sigma^2 c_1^2\right)}}{4\sigma^4 c_1^2} \tag{23}$$

When (21) holds define

$$b_2 := \exp(z_2) > e^2. \tag{24}$$

In Lemma 2, the real b_n is equal to $\exp(z_n)$ where z_n is the unique positive root $>1/2$ of $q_2(z)$. A formula for z_n is

$$z_n = \frac{1 - \sigma^4 + 2\sigma^4 c_1^2 + \left(\sigma^2 + 1\right)\sqrt{\left(\sigma^2 - 1\right)^2 + 4\sigma^4 c_1^2}}{4\sigma^4 c_1^2} > \frac{1}{2}. \tag{25}$$

Define

$$b_n := \exp(z_n) > e^{1/2}. \tag{26}$$

The three Trace-Determinant stability inequalities (16)–(18) for local stability, together with the three Lemmas 1, 2, and 3, yield the following theorem.

Theorem 2 *Assume $c_1 \neq 0$ and $\sigma^2 < 2$ in the Darwinian Ricker equations (8)–(9) and let b_2 and b_n be defined by (24) and (26).*

(a) Assume

$$\sigma^2 < \frac{2}{1 + 8c_1^2}$$

and define $b_m = \min\{b_2, b_n\}$. *The positive equilibrium (14) is locally asymptotically stable for* $1 < b_0 < b_m$ *and is unstable for* b_0 *greater than but near* b_m. *If* $b_m = b_2$ *then the Jacobian has an eigenvalue* -1 *when* $b_0 = b_m$. *If* $b_m = b_n$ *then the Jacobian has a complex eigenvalue of absolute value* 1 *when* $b_0 = b_m$.

(b) *If*

$$\sigma^2 > \frac{2}{1 + 8c_1^2}$$

then the positive equilibrium (14) is locally asymptotically stable for $1 < b_0 < b_n$. *and unstable for* b_0 *greater than, but near* b_n. *The Jacobian has a complex eigenvalue of absolute value* 1 *when* $b_0 = b_n$.

Note that the denominators in the Formulas (23) and (25) for z_2 and z_n are identical and the numerator of z_2 vanishes while that of z_n equals 2 when $\sigma^2 = 0$. Thus, for σ^2 small it follows that $b_2 < b_n$. Theorem 2(a) implies the following corollary.

Corollary 1 *Assume* $c_1 \neq 0$ *in the Darwinian Ricker equations (8)–(9). For* σ^2 *sufficiently small, the destabilization of the positive equilibria occurs at* $b_2 > e^2$.

For a fixed value of $c_1 \neq 0$, sufficiently large values of σ^2 (but less than 2) can result in destabilization at b_n, which can be either greater than or less than e^2. Examples are provided in the next section.

4 Concluding Remarks

It is not our purpose in this paper to rigorously study the nature of the bifurcations in the Darwinian Ricker equations that occur when the positive equilibrium destabilizes (i.e. to formally prove that they do result in new invariant sets, what the direction of bifurcation is, their stability properties, etc.). We focus only on the occurrence of the destabilization an indicator of the onset of non-equilibrium and complex dynamics. At the point of bifurcation, the equilibrium is nonhyperbolic and, as a result, the linearization principle does not hold. This is irrelevant for our purposes here because it is no concern to us what the stability properties of the equilibrium are at the point of bifurcation; we are interested only in the fact that there is a change from equilibrium stability to instability before and after the bifurcation occurs. With regard to the type of bifurcation that occurs, i.e. what kind of stable invariant sets replace the destabilized equilibrium, we do point out in Theorems 1 and 2 what the Jacobian eigenvalues are at the bifurcation point, specifically where on the complex unit circle an eigenvalue lies. The reason for this is that this information tells us what kind of bifurcation we expect to occur. If at destabilization -1 is an eigenvalue of the Jacobian, then one expects a period doubling bifurcation. If the Jacobian has a complex eigenvalue of absolute value 1, then one expects a Neimark-Sacker bifurcation to an invariant loop [3].

Theorem 1 implies that when $c_1 = 0$ in the Darwinian Ricker equations (8)–(9) the positive equilibria destabilize at $b_0 = e^2$, which is no different from the non-

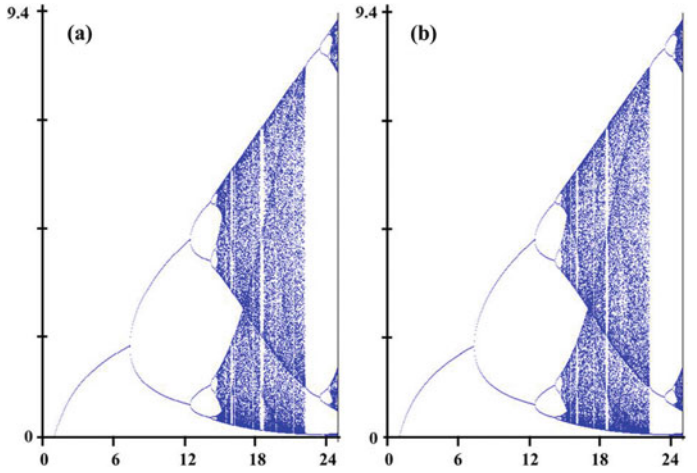

Fig. 1 **a** The familiar bifurcation diagram for the Ricker equation (1) with $c = 1$. **b** The bifurcation diagram showing the x component of the Darwinian Ricker equations (8)–(9) with $c_1 = 0$ and $\sigma^2 = 1$

evolutionary Ricker equation (1). The destabilization occurs because an eigenvalue of the Jacobian increases through -1 as b_0 increases through e^2, which is indicative of a period doubling bifurcation. This is also no different from the non-evolutionary Ricker equation. A sample bifurcation diagram appears in Fig. 1b that illustrates this bifurcation and what is apparently a period doubling route to chaos for the Darwinian Ricker equations that is identical with the non-evolutionary Ricker equation (Fig. 1a).

On the other hand, if $c_1 \neq 0$ then Theorem 2 shows that while destabilization does indeed occur at a critical value of b_0 in Darwinian Ricker equations, it does not necessarily indicate a period doubling bifurcation nor that it occurs at e^2, as in the non-evolutionary Ricker equation. The critical bifurcation point is either $b_2 > e^2$ (which is indicative of a period doubling bifurcation) or $b_n > e^{1/2}$ (which is indicative of a Neimark-Sacker bifurcation [3]). As stated in Corollary 1 equilibrium destabilization occurs at b_2 when the speed of evolution is not too fast. In fact, b_2 can be significantly larger than e^2 and the onset of complexity significantly delayed. Example bifurcation diagrams appear in Fig. 2.

Another difference between the evolutionary and non-evolutionary Ricker models is that destabilization does not necessarily result in period doubling. This occurs (for larger values of σ^2 and c_1^2) when $b_m = b_n$, which is indicative of a Neimark-Sacker bifurcation. Sample bifurcation diagrams appear in Fig. 3. One example (Fig. 3a) is when non-equilibrium dynamics are delayed, i.e. $b_n > e^2$ and the other (Fig. 3(b)) is when they are advanced, i.e. $b_n < e^2$. In the latter case, one could say evolution has promoted non-equilibrium and complexity dynamics.

For the Darwinian versions of the Ricker equation considered here, we arrive at several general conclusions. If $c_1 = 0$ in the trait dependent density coefficient $c(v - u)$, then there is no change in the destabilization point for the fertility rate b_0.

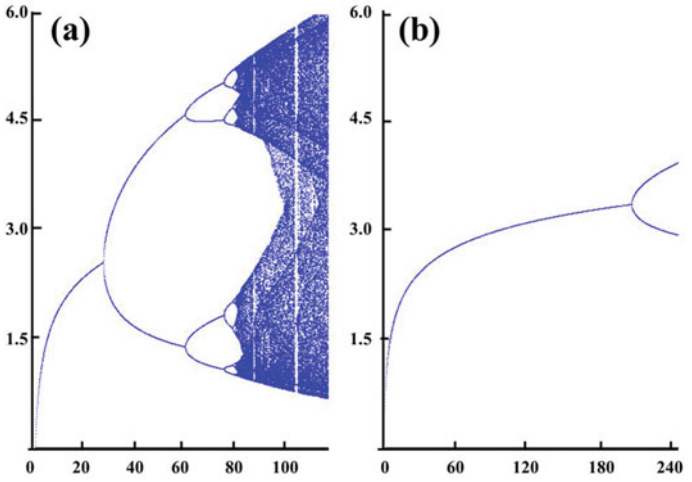

Fig. 2 The bifurcation diagram showing the x component of the Darwinian Ricker equations (8)–(9) with **a** $c_1 = 0.5$ and $\sigma^2 = 0.5$ and **b** $c_1 = 0.6$ and $\sigma^2 = 0.5$. The Formulas (24) and (26) for b_2 and b_n in these two cases yield **a** $b_2 \approx 28.121 < b_n \approx 2304.5$ and **b** $b_2 \approx 207.13 < b_n \approx 342.96$

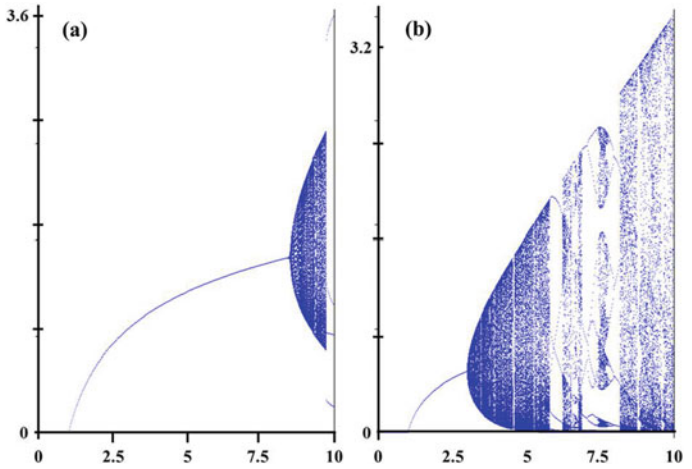

Fig. 3 The bifurcation diagram showing the x component of the Darwinian Ricker equations (8)–(9) with **a** $c_1 = 0.8$ and $\sigma^2 = 0.8$ and **b** $c_1 = 2$ and $\sigma^2 = 0.8$. For these cases, b_2 does not exist and **a** $b_n \approx 8.5253 > e^2 \approx 7.3891$ and **b** $b_n \approx 3.0004 < e^2$

Both models destabilize in period doubling bifurcations at the same critical value e^2. In this sense, we conclude that evolution has no effect on the onset of non-equilibrium and complex dynamics. The opposite is true in the case of hierarchical trait dependent competition coefficients, i.e. when $c_1 \neq 0$. In this case the onset of non-equilibrium and complex dynamics is delayed to a larger critical value of b_0 when evolution procedes slowly (i.e. σ^2 is small). In this case, we say that slow evolution selects against non-equilibrium and complex dynamics. If, on the other hand, evolution procedes at a faster speed, then there are two differences with the non-evolutionary Ricker equation, depending the magitude of the density effects, i.e. the size of c_1. First, the onset of non-equilibrium and complex dynamics can lead not to a period doubling bifurcation, but to a Neimark-Sacker bifurcation. Secondly, in the latter case, the bifurcation point can be either later or earlier than e^2. In the latter case (and only in this case), which occurs for larger σ^2 and c_1 values, we can conclude that evolution promotes non-equilibrium and complex dynamics.

These conclusions are drawn, of course, on the basis of the specific Darwinian Ricker equation considered here. To what extent they remain valid for other Darwinian equations with complex dynamics awaits further study.

References

1. Cushing, J.M.: A strong ergodic theorem for some nonlinear matrix models for structured population growth. Nat. Resourc. Model. **3**(3), 331–357 (1989)
2. Cushing, J.M.: Difference equations as models of evolutionary population dynamics. J. Biol. Dyn. **13**, 103–127 (2019)
3. Elaydi, S.N.: Discrete Chaos, with Applications in Science and Engineering, 2nd edn. Chapman & Hall/CRC, New York (2008)
4. Ferriere, R., Fox, G.A.: Chaos and evolution. Trends Ecol. Evol. **10**, 480–483 (1995)
5. May, R.M.: Simple mathematical models with very complicated dynamics. Nature **261**, 459–467 (1976)
6. Mokni, K., Elaydi, S., CH-Chaoui, M., Eladdadi, A.: Discrete evolutionary population models: a new approach. To appear in the J Biol Dyn (2020)
7. Neutel, A.M., Heesterbeek, J.A.P., de Ruiter, P.C.: Stability in real food webs: weak links in long loops. Science **296**, 1120–1123 (2002)
8. Sugihara, G., May, R.M.: Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. Nature **344**, 734–741 (1990)
9. Vincent, T., Brown, J.: Evolutionary Game Theory, Natural Selection, and Darwinian Dynamics. Cambridge University Press, Cambridge, UK (2005)
10. Zimmer, C.: Life after chaos. Science **284**, 83–86 (1999)

Difference Equations Related to Number Theory

BERNHARD HEIM and Markus Neuhauser

Abstract In this paper we investigate difference equations related to number theory. We apply a criterion to reduce these hereditary difference equations to finite length. The solutions are polynomials in one variable. We analyze the solutions with respect to convergence, periodicity, and boundedness. As an example we obtain and study Chebyshev polynomials of the second kind. We also apply Poincaré's theorem to transform a non-autonomous difference equation to an autonomous version.

Keywords Arithmetic functions · Chebyshev polynomials · Dedekind's η-function · Polynomials · Recurrence relations

1 Introduction

Powers of Dedekind's η-function are functions that are classically studied in number theory [8]. Dedekind's η-function is defined as a function on the upper complex half plane $\mathbb{H} = \{\tau \in \mathbb{C} : \operatorname{Im}(\tau) > 0\}$ by

$$\eta(\tau) = q^{1/24} \prod_{n=1}^{\infty} \left(1 - q^n\right)$$

employing $q = \mathrm{e}^{2\pi \mathrm{i} \tau}$. For powers thereof it can be shown that there are polynomials $P_n(x)$ such that

B. HEIM (✉)
Faculty of Mathematics, Computer Science, and Natural Sciences, RWTH Aachen University, 52056 Aachen, Germany
e-mail: bernhard.heim@rwth-aachen.de

M. Neuhauser
Kutaisi International University (KIU), Youth Avenue, Turn 5/7 Kutaisi, 4600, Georgia
e-mail: markus.neuhauser@kiu.edu.ge

S. Baigent et al. (eds.), *Progress on Difference Equations and Discrete Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 341,
https://doi.org/10.1007/978-3-030-60107-2_11

$$q^{x/24}\left(\eta\left(\tau\right)\right)^{-x}=\prod_{n=1}^{\infty}\left(1-q^{n}\right)^{-x}=1+\sum_{n=1}^{\infty}P_{n}\left(x\right)q^{n}$$

for $x\in\mathbb{C}$. The degree of the polynomial $P_n\left(x\right)$ is n. They satisfy $P_1\left(x\right)=x$ and the recurrence relation

$$P_{n}\left(x\right)=\frac{x}{n}\left(\sigma\left(n\right)+\sum_{k=1}^{n-1}\sigma\left(k\right)P_{n-k}\left(x\right)\right), \tag{1}$$

where $\sigma\left(k\right)=\sum_{d|k}d$ is the sum of divisors of k. In number theory the values $P_n\left(-24\right)$ play a special role. The Lehmer conjecture [7] on their non-vanishing is still open. Note that $P_n\left(1\right)$ are the so-called partition numbers p_n [1, 8]. For example $p_1=1$, $p_2=2$, $p_3=3$, $p_4=5$, …, $p_9=30$, …, $p_{50}=204226$.

To understand the dependence of these polynomials on the sum of divisors function σ we employ the following more general approach. We allow arbitrary functions instead of σ and additionally we also generalize the factor $\frac{1}{n}$ to $\frac{1}{h(n)}$, where h is an arithmetic function. To summarise we make the following definition (compare also [6]).

Definition 1 Let $g:\mathbb{N}\to\mathbb{C}$ be an arithmetic function normalised to $g\left(1\right)=1$ and $h:\mathbb{N}\to\mathbb{R}$, increasing, with $h\left(1\right)=1$. We define $P_n^{g,h}\left(x\right)$ by $P_1^{g,h}\left(x\right)=x$ and the recurrence relation

$$P_{n}^{g,h}\left(x\right)=\frac{x}{h\left(n\right)}\left(g\left(n\right)+\sum_{k=1}^{n-1}g\left(k\right)P_{n-k}^{g,h}\left(x\right)\right) \tag{2}$$

for $n\geq 2$. We abbreviate $P_n^{g,h}\left(x\right)$ by $P_n\left(x\right)=P_n^{g}\left(x\right)$ if $h=\mathrm{id}$ the identity on $\mathbb{N}$ and by $Q_n\left(x\right)=Q_n^{g}\left(x\right)$ if $h\left(n\right)=1$ for all $n\in\mathbb{N}$.

Note that for $g=\sigma$ and $h=\mathrm{id}$ we obtain that $P_n^{\sigma,\mathrm{id}}\left(x\right)=P_n^{\sigma}\left(x\right)=P_n\left(x\right)$ is exactly the nth coefficient of the $-x$th power of Dedekind's η-function.

In some cases, as in Example 2, it is possible to reduce the recurrence relation (2) to one of bounded length see [6, Theorem 2.1] (and Theorem 1). We recall this in Sect. 2.

Example 1 (Toy problems). Let $g\left(n\right)=1$ for all n. Then

$$P_{n}^{g,h}\left(x\right)=\frac{x}{h\left(n\right)}\prod_{m=1}^{n-1}\left(\frac{x}{h\left(m\right)}+1\right)$$

for $n\geq 1$ with (obvious) zeros $x\in\{0,-h\left(1\right),-h\left(2\right),\ldots,-h\left(n-1\right)\}$.

Example 2 (Chebyshev polynomials of the second kind). We obtain for $g=\mathrm{id}$ a relation to classical orthogonal polynomials [2]. In [6, Corollary 2.5] (see also Corollary 1 of Theorem 1) we obtained the reduced recurrence relation

$$Q_n(x) = (x+2)\,Q_{n-1}(x) - Q_{n-2}(x) \qquad (n \geq 3) \tag{3}$$

and the initial values $Q_1(x) = x$ and $Q_2(x) = (x+2)\,x$. As a solution of this recurrence relation in [6, part 1 of Remark 2.8] we obtained $Q_n(x) = xU_{n-1}(x/2+1)$ for $n \geq 1$ where $U_n(x)$ are the Chebyshev polynomials of the second kind. By the well-known relation $U_{n-1}(\cos t) = \frac{\sin(nt)}{\sin t}$ for $0 < t < \pi$ the zeros of $Q_n(x)$ are $x = 0$ and $x = \cos(k\pi/n)$ for $1 \leq k \leq n-1$ and $n \geq 2$.

Also the $P_n^g(x) = \sum_{k=1}^{n} \binom{n-1}{n-k} \frac{x^k}{k!} = \frac{x}{n} L_{n-1}^{(1)}(-x)$ are related to the associated Laguerre polynomials [2], which have the form $L_n^{(\alpha)}(x) = \sum_{k=0}^{n} \binom{n+\alpha}{n-k} \frac{(-x)^k}{k!}$ for $\alpha > -1$.

We assume throughout this work that the power series

$$G(z) = \sum_{k=1}^{\infty} g(k+1)\, z^k$$

has a positive radius of convergence R. Note in the case $h(n) = 1$ for all $n \in \mathbb{N}$ that (2) is actually a Volterra difference equation of convolution type [3, Chap. 6].

We are interested in the sequence $\left(P_n^{g,h}(x)\right)_{n\in\mathbb{N}}$ depending on x. Note that the sequence $P_n^{g,h}(0) = 0$. Let $x > 0$. Then the sequence $\left(P_n^{g,h}(x)\right)_{n\in\mathbb{N}}$ diverges if g is bounded from below (Proposition 1).

In Sect. 3 we are going to study the limiting behaviour (convergence, periodicity, or boundedness) of the sequences $\left(Q_n^g(x)\right)_{n\in\mathbb{N}}$ for $g(k) = 1$ depending on x.

In Sect. 4 we are going to identify the behaviour of the sequences depending on x for the more complicated case of $Q_n^{\mathrm{id}}(x)$ related to the Chebyshev polynomials of the second kind (Theorem 4).

Finally in Sect. 5 we are going to indicate that there is a close relation between the sequences $\left(P_n^g(x)\right)_{n\in\mathbb{N}}$ and $\left(Q_n^g(x)\right)_{n\in\mathbb{N}}$ provided by Poincaré's theorem for non-autonomous difference equations [3]. This is illustrated by the case $g(k) = 1$. The $Q_n^g(x)$ are usually simpler to study because in this case we actually have a genuine Volterra difference equation.

2 Previous Results

In this section we recall some results from our work [6], which we employ in the following.

Among the many obstacles that prevent the application of standard methods to solve explicitly recurrence relation (1) is that to our knowledge it is not possible to reduce it to a k-order recurrence relation, a recurrence relation of bounded length.

For other functions than $g = \sigma$, in [4] for polynomial g and in [6, Theorem 2.1] for more general g, we stated a method to reduce them to bounded length. Let us recall the result. Let T be the shift operator that is $Tg(k) = g(k+1)$.

Theorem 1 *Suppose that there are $M \in \mathbb{N}$ and $\alpha_0, \ldots, \alpha_M \in \mathbb{C}$ with $\alpha_M = 1$ such that*

$$\sum_{m=0}^{M} \alpha_m T^m g = 0. \tag{4}$$

Then a solution of (2) fulfills

$$\sum_{m=0}^{M} \left(\alpha_m \frac{h(n+m)}{h(n+M)} - \frac{x}{h(n+M)} \sum_{k=1}^{M-m} \alpha_{m+k} g(k) \right) P_{n+m}^{g,h}(x) = 0 \tag{5}$$

for $n \geq 1$ which is a recurrence relation of bounded length $\leq M+1$.

Remark 1 Note that the reduced equation (5) can only be applied when $n \geq M+1$. Thus instead of just one initial value $P_1^{g,h}(x) = x$ we use M initial values generated from this by the first M convolution equations.

The following states Theorem 1 in the case of $h(k) = 1$ for all $k \in \mathbb{N}$. This is a simplified version of [6, Corollary 2.5].

Corollary 1 *Let x be fixed. Let $h(n) = 1$ for $n \geq 1$. In this case a solution of (2) fulfills the difference equation*

$$\sum_{m=0}^{M} \left(\alpha_m - x \sum_{k=1}^{M-m} \alpha_{m+k} g(k) \right) Q_{n+m}^{g}(x) = 0 \tag{6}$$

for $n \geq 1$ with characteristic equation

$$\sum_{m=0}^{M} \alpha_m \lambda^m - x \sum_{m=0}^{M-1} \left(\sum_{k=1}^{M-m} \alpha_{m+k} g(k) \right) \lambda^m = 0.$$

In Sect. 4 we are going to consider the case related to the Chebyshev polynomials of the second kind (see Example 2). This is the case when $g = \mathrm{id}$. For the $Q_n^g(x)$ holds the following (see [6, Lemma 4.5]).

Lemma 1 *Let $x \in \mathbb{C} \setminus \{-4, 0\}$ be fixed and $D = D(x) = x^2 + 4x$. Define $\lambda_\pm = \lambda_\pm(x) = \left(x + 2 \pm \sqrt{D}\right)/2$, $b_- = b_-(x) = \frac{\lambda_+ - x - 2}{\lambda_+ - \lambda_-}$, and $b_+ = b_+(x) = \frac{x+2-\lambda_-}{\lambda_+ - \lambda_-}$. Then for all $n \geq 1$*

$$Q_n^g(x) = \left(b_- \lambda_-^{n-1} + b_+ \lambda_+^{n-1} \right) x.$$

Later in this work we are going to employ the following [6, Theorem 7.1] to show the divergence of certain sequences (see also [5] for a version where specifically $h = \mathrm{id}$ is considered).

Theorem 2 *Let $R > 0$ be the radius of convergence of*

$$G(q) := \sum_{k=1}^{\infty} g(k+1) q^k.$$

Let $\kappa > 0$ be such that $G\left(\frac{2}{\kappa}\right) \leq \frac{1}{2}$ (and $\frac{2}{\kappa} < R$). Then

$$\left|P_n^{g,h}(x)\right| > \frac{|x|}{2h(n)} \left|P_{n-1}^{g,h}(x)\right|$$

if $|x| > \kappa h(n-1)$, $n \geq 1$.

3 The Behaviour of Some of These Polynomial Sequences in the Limit

Note that for increasing g, there is no neighbourhood of 0 such that all sequences for x in this neighbourhood converge to 0 since $Q_n(x) \geq g(n) x \geq g(1) x$ for positive x. In case $\lim_{n\to\infty} g(n) = \infty$ this sequence is actually divergent.

Example 3 (Toy problem). Let $g(n) = 1$ for all $n \in \mathbb{N}$. Then $Q_n(x) = (x+1)^{n-1} x$ is the solution to the recurrence relation. This can be observed from

$$x + x \sum_{k=1}^{n-1} (x+1)^{k-1} x = x + \frac{(x+1)^{n-1} - 1}{x} x^2.$$

Actually the recurrence relation reduces to $Q_n(x) = (x+1) Q_{n-1}(x)$ for $n \geq 2$. This explicit formula immediately implies the following properties:

1. $x = 0$ is fixed.
2. If $|x+1| < 1$ then $(Q_n(x))_{n\in\mathbb{N}}$ converges to $x = 0$.
3. If $|x+1| > 1$ then $(Q_n(x))_{n\in\mathbb{N}}$ diverges to ∞.
4. If $|x+1| = 1$ and $x \neq 0$ then $(Q_n(x))_{n\in\mathbb{N}}$ moves around on the circle $\{z \in \mathbb{C} : |z| = |x|\}$. If $x = -1 + e^{\pi i t}$ with rational t the orbit is finite and otherwise infinite and dense in this circle.

We prove the last item 4 in the previous example in the following.

Proof (Item 4). Let $|x+1| = 1$. Then $x = -1 + e^{\pi i t}$ for some $t \in \mathbb{R}$.

If $t \in \mathbb{Q}$ then there is an $m \in \mathbb{N}$ such that $(x+1)^m = 1$. Thus the sequence $Q_n(x) = (x+1)^n x$ assumes only finitely many values.

For $t \in \mathbb{R}\backslash\mathbb{Q}$ it is well-known that $\{nt + \mathbb{Z} : n \in \mathbb{Z}\}$ is dense in $\mathbb{R}/\mathbb{Z}$ (see [10]). Since exp is continuous this implies that $\left\{e^{2\pi i n t} : n \in \mathbb{Z}\right\}$ is dense in the circle of radius 1 around 0 in the complex plane.

If g is bounded from below we have the following general result on the sequences $(Q_n(x))_{n\in\mathbb{N}}$.

Proposition 1 *Assume $g(n) \geq c$ for all $n \geq 2$ for some $c > 0$. Then $(Q_n(x))_{n\in\mathbb{N}}$ diverges for $x > 0$.*

Proof We have

$$\begin{aligned} Q_n(x) &= x\left(g(n) + \sum_{k=1}^{n-1} g(k)\, Q_{n-k}(x)\right) \\ &= xg(n) + x\sum_{k=1}^{n-1} g(k)\left(xg(n-k) + \sum_{j=1}^{n-k-1} g(j)\, Q_{n-k-j}(x)\right) \\ &\geq xg(n) + x^2\sum_{k=1}^{n-1} g(k)\, g(n-k) \\ &> x^2\left(2c + (n-3)\, c^2\right) \to \infty \end{aligned}$$

for $n \to \infty$.

We can also prove the following result for arbitrary g, in which we have a positive radius of convergence of $G(z) = \sum_{k=1}^{\infty} g(k+1)\, z^k$.

Theorem 3 *Let κ be such that $G(2/\kappa) \leq 1/2$. Suppose $|x| > \max\{2, \kappa\}$. Then $\left(P_n^{g,h}(h(n)\, x)\right)_{n\in\mathbb{N}}$ diverges.*

Proof By Theorem 2 [6, Theorem 7.1] for $2 \leq m \leq n$ we obtain

$$\left|P_m^{g,h}(h(n)\, x)\right| > \frac{|h(n)\, x|}{2h(m)}\left|P_{m-1}^{g,h}(h(n)\, x)\right| \geq \frac{|x|}{2}\left|P_{m-1}^{g,h}(h(n)\, x)\right|.$$

Thus,

$$\left|P_n^{g,h}(h(n)\, x)\right| > \left(\frac{|x|}{2}\right)^n \to \infty$$

follows by assumption on x.

Since $h(n) = 1$ for all $n \in \mathbb{N}$ in the case of the polynomials $Q_n^g(x)$ we obtain immediately the following.

Corollary 2 *Let κ be such that $G(2/\kappa) \leq 1/2$. Suppose $|x| > \max\{2, \kappa\}$. Then $\left(Q_n^g(x)\right)_{n\in\mathbb{N}}$ diverges.*

4 Chebyshev Polynomials of the Second Kind

In this section $g = \mathrm{id}$. Then $Q_n(x) = xU_{n-1}(x/2+1)$ where $U_n(x)$ are the Chebyshev polynomials of the second kind (see Example 2). We prove now for $x \in \mathbb{C}$ the following behaviours of $(Q_n(x))_{n\in\mathbb{N}}$.

Theorem 4 *1.* $Q_n(0) = 0$ *for all* n.
2. *If* $x = 2\cos(k\pi/n) - 2$ *for* $1 \le k \le n-1$ *then* $(Q_n(x))_{n\in\mathbb{N}}$ *has period* d *where* $d \mid 2n$.
3. *If* $-4 < x < 0$ *and* x *is not of the form* $2\cos(k\pi/n) - 2$ *for some* $n \in \mathbb{N}$ *and some* $1 \le k \le n-1$ *then* $(Q_n(x))_{n\in\mathbb{N}}$ *generates a bounded sequence dense in the interval* $\left[\frac{x}{\sqrt{-x-x^2/4}}, -\frac{x}{\sqrt{-x-x^2/4}}\right]$.
4. *If* $x \in \mathbb{C}\setminus(-4, 0]$ *then* $(Q_n(x))_{n\in\mathbb{N}}$ *diverges.*

In the following we prove this theorem by several auxiliary results.

Lemma 2 *Let* $n \ge 2$ *and* $1 \le k \le n-1$. *Then* $\cos(k\pi/n)$ *is a zero of* $Q_{\ell n}(x)$ *for all* $\ell \in \mathbb{N}$.

Proof We have $x = \cos\left(\frac{k}{n}\pi\right) = \cos\left(\frac{\ell k}{\ell n}\pi\right)$. Thus x is also a zero of $Q_{\ell n}(x)$ for all $\ell \in \mathbb{N}$.

By the following lemma we show that for all of these values $x = \cos(k\pi/n)$ that $Q_m(x)$ is $2n$-periodic for $m \ge 1$. Thus, the proof of part 2 of Theorem 4.

Lemma 3 *Any zero* $x = \cos(k\pi/n)$ *for some* $n \in \mathbb{N}$ *and* $1 \le k \le n-1$ *generates a* $2n$*-periodic sequence* $Q_m(x)$ *for* $m \ge 1$.

Proof Then $Q_{\ell n}(x) = 0$. We show $Q_{\ell n+k}(x) = -Q_{\ell n-k}(x)$. Obviously this holds for $k = 0$. Suppose now $1 \le k \le \ell n - 2$ and that for $0 \le j \le k-1$ holds $Q_{\ell n+j}(x) = -Q_{\ell n-j}(x)$. Then

$$\begin{aligned} Q_{\ell n+k}(x) &= (x+2)\,Q_{\ell n+k-1}(x) - Q_{\ell n+k-2}(x) \\ &= -(x+2)\,Q_{\ell n-k+1}(x) + Q_{\ell n-k+2}(x) \\ &= -(x+2)\,Q_{\ell n-k+1}(x) + (x+2)\,Q_{\ell n-k+1}(x) - Q_{\ell n-k}(x)\,. \end{aligned}$$

For $k \le 2(\ell-1)n - 2$ this implies

$$\begin{aligned} Q_{2\ell n+k}(x) &= -Q_{2\ell n-k}(x) = -Q_{(2\ell-1)n+n-k}(x) = Q_{(2\ell-1)n-n+k}(x) \\ &= Q_{2(\ell-1)n+k}(x)\,. \end{aligned}$$

Hence $Q_m(x)$ is $2n$-periodic.

The following proves part 3 of Theorem 4.

Lemma 4 *If* $-4 < x < 0$ *the orbit generated by* $(Q_n(x))_{n\in\mathbb{N}}$ *is bounded. If* $t = \arccos(x/2+1) \notin \pi\mathbb{Q}$ *then the sequence is dense in*

$$\left[\frac{x}{\sqrt{-x-x^2/4}}, -\frac{x}{\sqrt{-x-x^2/4}}\right].$$

Proof Let $t = \arccos(x/2+1)$. Thus $|Q_n(x)| = \left|x\frac{\sin(nt)}{\sin t}\right| \leq \left|\frac{x}{\sqrt{-x-x^2/4}}\right|$. If $t \notin \pi\mathbb{Q}$ then it is well-known that $\{nt + 2\pi\mathbb{Z} : n \in \mathbb{Z}\}$ is dense in $\mathbb{R}/(2\pi\mathbb{Z})$. Since sin is continuous $\{\sin(nt) : n \in \mathbb{Z}\}$ is dense in $[-1, 1]$.

Finally the following proves part 4 of Theorem 4. Here standard methods (see e.g. [3]) can be used to solve the recurrence relation via the characteristic equation (see Lemma 1 [6, Lemma 4.5]).

Proposition 2 *If* $x \in \mathbb{C}\setminus(-4, 0]$ *then the sequence generated by* $Q_n(x)$ *diverges.*

Proof For $x = -4$ this is easy to observe (see also [6, part 3 of Remark 2.8]).

The characteristic equation of (3) is

$$\lambda^2 - (x+2)\lambda + 1 = 0. \tag{7}$$

If $\lambda_\pm = \lambda_\pm(x)$ are the two solutions of (7) then $\lambda_+\lambda_- = 1$ and $\lambda_+ + \lambda_- = x+2$. Thus $\lambda_- = \lambda_+^{-1}$. For $x < -4$ or $x > 0$ we obtain $\Delta = (x+2)^2 - 4 = (x+4)x > 0$. Thus both $\lambda_\pm$ are real and $\lambda_+ \neq \lambda_-$ since $x \neq 0, -4$.

Now let $x = a + ib \in \mathbb{C}\setminus\mathbb{R}$ where $a = \mathrm{Re}(x)$, $b = \mathrm{Im}(x) \in \mathbb{R}$. Let $\lambda_+ = \alpha + i\beta$ with $\alpha = \mathrm{Re}(\lambda_+)$, $\beta = \mathrm{Im}(\lambda_+) \in \mathbb{R}$. As we have already observed $\lambda_+\lambda_- = 1$. Thus $\frac{\alpha - i\beta}{\alpha^2+\beta^2} = \lambda_+^{-1} = \lambda_- = x + 2 - \lambda_+ = a + 2 - \alpha + (b-\beta)i$. From the imaginary part we obtain $-\frac{\beta}{\alpha^2+\beta^2} = b - \beta$. Hence $\alpha^2 + \beta^2 = \frac{\beta}{\beta - b} \neq 1$ since by assumption $b = \mathrm{Im}(x) \neq 0$. Thus one of $|\lambda_+|$ or $|\lambda_-|$ is > 1.

In all cases i.e. $x < -4$ or $x > 0$ or $x \in \mathbb{C}\setminus\mathbb{R}$ we obtained in [6, Lemma 4.5] that $Q_n(x) = \left(b_+\lambda_+^{n-1} + b_-\lambda_-^{n-1}\right)x$ for $n \geq 1$ with $b_+ = \frac{\lambda_+ - x - 2}{\lambda_+ - \lambda_-}$ and $b_- = \frac{x+2-\lambda_-}{\lambda_+ - \lambda_-}$. From $\lambda_+ + \lambda_- = x + 2$ and $\lambda_\pm \neq 0$ it follows immediately that $b_\pm \neq 0$. Hence $(Q_n(x))_{n\in\mathbb{N}}$ diverges as one of the absolute values of $\lambda_\pm$ is strictly larger and one strictly less than 1.

5 Relations Between Two Sequences of Polynomials

In this section we link the recurrence relations of $P_n^g(x)$ and $Q_n^g(x)$ for $g(n) = 1$ by the following Theorem of Poincaré [9] adapted to the case considered in the present work (for its general form see [3, Theorem 8.9]). Recall that for $g(n) = 1$ we obtain $P_n^g(x) = \frac{x+n-1}{n}P_{n-1}^g(x) = (-1)^n\binom{-x}{n}$ for $n \geq 1$.

Theorem 5 *(H. Poincaré [9]) Assume that $P_n^g(x)$ satisfies a recurrence relation of finite length. If the limit for each of the coefficients in the recurrence relation exists and the roots of the characteristic equations with theses limits have distinct moduli then either the solution to the recurrence relation is eventually* 0 *or* $\lim_{n\to\infty} \frac{P_n^g(x)}{P_{n-1}^g(x)}$ *equals a root of the characteristic equation with the limits of the coefficients.*

For the coefficient we obtain

$$\lim_{n\to\infty} \frac{x+n-1}{n} = 1.$$

In the recurrence relation for $Q_n(x)$ we have the term $x+1$. But if we considered nx instead of x then $\lim_{n\to\infty} \frac{nx+n-1}{n} = x+1$. Unfortunately $\tilde{P}_n(x) = P_n(nx)$ does not fulfill the same recurrence relation as $P_n(x)$ but only $\tilde{P}_n(x) = \left(x+1-\frac{1}{n}\right) P_{n-1}(nx) = \left(x+1-\frac{1}{n}\right) \frac{P_{n-1}(nx)}{P_{n-1}((n-1)x)} \tilde{P}_{n-1}(x)$.

Theorem 6 *Let* $x < -1$ *then* $\lim_{n\to\infty} \frac{P_{n-1}(nx)}{P_{n-1}((n-1)x)} = \left(\frac{-x}{-x-1}\right)^{-x}$.

Proof We obtain

$$\begin{aligned}\ln\left(\frac{P_{n-1}(nx)}{P_{n-1}((n-1)x)}\right) &= \ln\left(P_{n-1}(nx)\right) - \ln\left(P_{n-1}((n-1)x)\right)\\ &= \sum_{k=0}^{n-1} \ln(-nx-k) - \ln(-(n-1)x-k).\end{aligned}$$

The terms in the sum can be written as an integral

$$\ln(-nx-k) - \ln(-(n-1)x-k) = \int_0^{-x} \frac{1}{t-(n-1)x-k}\,\mathrm{d}t.$$

As Riemannian sum for $n\to\infty$ we obtain

$$\begin{aligned}\sum_{k=0}^{n-1} \frac{1}{t-(n-1)x-k} &= \frac{1}{n}\sum_{k=0}^{n-1} \frac{1}{\frac{t}{n}-x-\frac{k-x}{n}}\\ &\to \int_0^1 \frac{1}{-x-s}\,\mathrm{d}s\\ &= -\ln(-x-1)+\ln(-x) = \ln\left(\frac{-x}{-x-1}\right).\end{aligned}$$

This implies

$$\ln\left(\frac{P_{n-1}(nx)}{P_{n-1}((n-1)x)}\right) \to \int_0^{-x} \ln\left(\frac{-x}{-x-1}\right)\mathrm{d}t = -x\ln\left(\frac{-x}{-x-1}\right).$$

Thus, $\frac{P_{n-1}(nx)}{P_{n-1}((n-1)x)} \to \left(\frac{-x}{-x-1}\right)^{-x}$.

Thus, for $x < -1$ in the limit the terms $\left(x+1-\frac{1}{n}\right)\frac{P_{n-1}(nx)}{P_{n-1}((n-1)x)}$ converge to $(x+1)\left(\frac{-x}{-x-1}\right)^{-x}$. Thus, we have a difference equation of Poincaré type with $\lambda = (x+1)\left(\frac{-x}{-x-1}\right)^{-x}$ for $x < -1$. By Poincaré's Theorem [9] (see also Theorem 5) we obtain that

$$\lim_{n\to\infty} \frac{\tilde{P}_n(x)}{\tilde{P}_{n-1}(x)} = (x+1)\left(\frac{-x}{-x-1}\right)^{-x}.$$

On the other hand if we use that $\lim_{n\to\infty} \frac{x+n-1}{n} = 1$ then $\lim_{n\to\infty} \frac{P_n(x)}{P_{n-1}(x)} = 1$. In the case considered the difference equation can be solved explicitly as $P_n(x) = (-1)^n \binom{-x}{n}$. We obtain the second limit as

$$\lim_{n\to\infty} \frac{P_n(x)}{P_{n-1}(x)} = \lim_{n\to\infty} \frac{-\binom{-x}{n}}{\binom{-x}{n-1}} = \lim_{n\to\infty} \frac{x+n-1}{n} = 1.$$

It would be interesing to understand the relation of $P_n(x)$ and $Q_n(x)$ in arbitrary cases. For example for $h = \mathrm{id}$ and $g = \mathrm{id}$ the reduced recurrence relation is

$$\begin{aligned} P_n^g(x) &= \frac{1}{n}\left((2n-2+x)\,P_{n-1}^g(x) - (n-2)\,P_{n-2}^g(x)\right) \\ &= \left(2+\frac{x-2}{n}\right)P_{n-1}^g(x) - \left(1-\frac{2}{n}\right)P_{n-2}^g(x) \end{aligned}$$

for $n \geq 3$ with initial values $P_1^g(x) = x$ and $P_2^g(x) = \frac{x}{2}(x+2)$.

For $P_n(x)$ we obtain in the limit the characteristic equation $0 = \lambda^2 - 2\lambda + 1 = (\lambda-1)^2$. In this case we cannot directly apply Poincaré's Theorem.

Acknowledgments We thank the organisers of the conference, especially Steve Baigent, for their invitation and their excellent job and the reviewer for several useful comments.

References

1. Andrews, G.E.: The Theory of Partitions. Cambridge Mathematical Library. Cambridge University Press, Cambridge (1998). Reprint of the 1976 original
2. Doman, B.G.S.: The Classical Orthogonal Polynomials. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ (2016)
3. Elaydi, S.: An Introduction to Difference Equations. Undergraduate Texts in Mathematics, 3rd edn. Springer, New York (2005)
4. Heim, B., Luca, F., Neuhauser, M.: Recurrence relations for polynomials obtained by arithmetic functions. Int. J. Numb. Theory **15**(6), 1291–1303 (2019). https://doi.org/10.1142/S1793042119500726
5. Heim, B., Neuhauser, M.: The Dedekind eta function and D'Arcais-type polynomials. Res. Math. Sci. **7**(1), Paper No. 3, 8 (2020). https://doi.org/10.1007/s40687-019-0201-5

6. Heim, B., Neuhauser, M., Tröger, R.: Zeros of recursively defined polynomials. J. Difference Equ. Appl. **26**(4), 510–531 (2020). https://doi.org/10.1080/10236198.2020.1748022
7. Lehmer, D.H.: The vanishing of Ramanujan's function $\tau(n)$. Duke Math. J. **14**, 429–433 (1947). http://projecteuclid.org/euclid.dmj/1077474140
8. Ono, K.: The web of modularity: arithmetic of the coefficients of modular forms and q-series. In: CBMS Regional Conference Series in Mathematics, vol. 102. Published for the Conference Board of the Mathematical Sciences, Washington, DC. American Mathematical Society, Providence, RI (2004)
9. Poincaré, H.: Sur les equations linéaires aux différentielles ordinaires et aux différences finies. Amer. J. Math. **7**(3), 203–258 (1885). https://doi.org/10.2307/2369270
10. Weyl, H.: Über die Gleichverteilung von Zahlen mod. Eins. Math. Ann. **77**(3), 313–352 (1916). https://doi.org/10.1007/BF01475864

A Note on Non-hyperbolic Fixed Points of One-Dimensional Maps

Sinan Kapçak

Abstract This paper deals with the local asymptotic stability of non-hyperbolic fixed points of one-dimensional maps. There are, basically, two stability conditions introduced in this study. One of them is for the stability of fixed points of non-oscillatory maps. The second one is a sufficient condition for the stability for oscillatory maps. Some properties and applications are also presented.

Keywords Non-hyperbolic fixed points · One-dimensional maps · Test of stability · Difference equations · Discrete dynamical systems

1 Introduction

Although the complete theory of the stability of hyperbolic and non-hyperbolic fixed points of one-dimensional maps was already studied [1, 5], the conditions given in these studies are based on the higher order derivatives evaluated at the fixed point, which usually requires lengthy calculations. For the oscillatory case, in [1], the composition function $g = f \circ f$ is used in order to determine the stability. In the same paper, Faà di Bruno's formula was proposed in order to take the higher order derivatives of the composition function. Using this idea, a generalization of Schwarzian derivatives is obtained in [6]. In [5], author uses only the higher order derivatives of f evaluated at the fixed point in order to determine the stability of both non-oscillatory and oscillatory fixed points.

This paper only deals with local stability, and proposes two conditions on the non-hyperbolic fixed points. Firstly, for a fixed point x^* of one-dimensional map $x_{n+1} = f(x_n)$, where $f'(x^*) = 1$, we introduce a new local stability condition. Secondly, we give a sufficient condition for local stability of non-hyperbolic fixed points of the maps with $f'(x^*) = -1$. Our main approach here for the stability conditions is that we will not focus on the fixed point itself but the vicinity of it. By this way, the

S. Kapçak (✉)
American University of the Middle East, Egaila, Kuwait
e-mail: sinan.kapcak@aum.edu.kw

S. Baigent et al. (eds.), *Progress on Difference Equations and Discrete Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 341,
https://doi.org/10.1007/978-3-030-60107-2_12

stability/instability for particular types of maps can be determined straightforwardly, which is the one of the strengths of this study. Moreover, the theorem which gives a sufficient condition for the stability (we call it Test of Stability) of non-hyperbolic fixed points of oscillatory maps is easy to apply and helps us generalize the stability conditions for some family of difference equations.

The next section is devoted to the main results of this study. Examples and applications will be given in Sect. 3. Some related lemmas and theorems can be found in Appendix.

2 Main Results

2.1 The Case $f'(x^*) = 1$

Let $\varepsilon > 0$ be an infinitesimal quantity and $a \in \mathbb{R}$. We will use the following notations for the types of neighborhood sets of point a:

$$B_\varepsilon(a) = (a-\varepsilon, a+\varepsilon) - \{a\}, \quad B_\varepsilon^-(a) = (a-\varepsilon, a), \quad B_\varepsilon^+(a) = (a, a+\varepsilon).$$

Throughout this paper, we assume that f is an analytic function. Clearly, the fixed point $x^* = 0$ is locally stable (not asymptotically) when $f(x) \equiv x$ on $B_\varepsilon(0)$. This is the case with $f'(x) = 1$ on $B_\varepsilon(0)$.

Without loss of generality, we may assume that the fixed point is at the origin. Now we present the following theorem which gives a local stability condition for the non-hyperbolic fixed points of non-oscillatory maps.

Theorem 1 *Consider the difference equation $x_{n+1} = f(x_n)$, where $f(0) = 0$ and $f'(0) = 1$. The fixed point $x^* = 0$ is*

1. *locally asymptotically stable if $f'(x) < 1$ on $B_\varepsilon(0)$.*
2. *unstable if $f'(x) > 1$ on $B_\varepsilon^+(0)$ or $B_\varepsilon^-(0)$.*

Proof We will prove the parts of the theorem separately.

1. Assume that $f'(x) < 1$ on $B_\varepsilon(0)$. Pick an initial point $x_0 \in B_\varepsilon(0)$.
   (a) If $x_0 \in B_\varepsilon^+(0)$: By Lemma 1, we have $0 < f(x) < x$. By Lemma 3, $x_n \to 0$ as $n \to \infty$.
   (b) If $x_0 \in B_\varepsilon^-(0)$: By Lemma 1, we have $x < f(x) < 0$. By Lemma 3, $x_n \to 0$ as $n \to \infty$.

   By Theorem 7, $x^* = 0$ is a locally asymptotically stable fixed point.
2. Assume that $f'(x) > 1$ on $B_\varepsilon(0)$. Pick an initial point $x_0 \in B_\varepsilon(0)$.
   (a) If $x_0 \in B_\varepsilon^+(0)$: By Lemma 1, we have $f(x) > x$. Lemma 4 completes the proof.

(b) If $x_0 \in B_\varepsilon^-(0)$: By Lemma 1, we have $f(x) < x$. Lemma 4 completes the proof.

Example 1 Let us discuss the difference equation

$$x_{n+1} = x_n - \sin x_n^5.$$

We have $f(0) = 0$ and $f'(x) = 1 - 5x^4 \cos x^5$. Hence, $f'(0) = 1$ and obviously, $f'(x) < 1$ on $B_\varepsilon(0)$. Therefore, the origin is a locally asymptotically stable fixed point. This map, with the methods given in [1, 5], requires the first five derivatives to conclude this result.

More generally, for the difference equation $x_{n+1} = x_n - \sin x_n^k$, we may, similarly, determine the stability and conclude that the fixed point is asymptotically stable (resp. unstable) when k is odd (resp. even).

2.1.1 The General Case

The stability condition of a fixed point $x = x^*$ for a general autonomous difference equation $x_{n+1} = f(x_n)$ for the non-oscillatory case $f'(x^*) = 1$ is already presented in [1, 5]. That condition can also be obtained by Theorem 1. Taylor series expansion of the derivative function $f'(x)$ evaluated at the fixed point tells us the stability character of the fixed point. Taking the fixed point $x^* = 0$ gives

$$f'(x) = 1 + \frac{f^{(m)}(0)}{(m-1)!}x^{m-1} + \frac{f^{(m+1)}(0)}{m!}x^m + \frac{f^{(m+2)}(0)}{(m+1)!}x^{m+1} + O(x^{m+2}) \quad (1)$$

where $m \in \{2, 3, 4, \ldots\}$ is the smallest number such that $f^{(m)}(0) \neq 0$. Thus, clearly, if $\frac{f^{(m)}(0)}{(m-1)!}x^{m-1} < 0$ on $B_\varepsilon(0)$, then the origin is locally asymptotically stable. Similarly, if $\frac{f^{(m)}(0)}{(m-1)!}x^{m-1} > 0$ on $B_\varepsilon^+(0)$ or $B_\varepsilon^-(0)$, then the origin is unstable. Here, the stability depends on whether m is even or odd. If m is even, then $\frac{f^{(m)}(0)}{(m-1)!}x^{m-1} > 0$ on $B_\varepsilon^+(0)$ or $B_\varepsilon^-(0)$, which yields instability. If m is odd, then the sign of $f^{(m)}(0)$ on $B_\varepsilon(0)$ will determine the stability. Clearly, if $f^{(m)}(0) > 0$ (resp. < 0), the fixed point is unstable (resp. asymptotically stable). Hence, we obtain the following theorem, which is already given in [1, 5].

Theorem 2 *Assume that x^* is a fixed point of the difference equation $x_{n+1} = f(x_n)$ and $f'(x^*) = 1$. Let $m \in \{2, 3, 4, \ldots\}$ be the smallest number such that $f^{(m)}(x^*) \neq 0$.*

1. *If m is even, then x^* is unstable (semi-stable).*
2. *If m is odd and $f^{(m)}(x^*) > 0$, then x^* is unstable.*
3. *If m is odd and $f^{(m)}(x^*) < 0$, then x^* is locally asymptotically stable.*

2.2 The Case $f'(x^*) = -1$

Consider, for example, the difference equation $x_{n+1} = -x_n e^{-x_n^3}$. With $f(x) = -xe^{-x^3}$, we have $f(0) = 0$, $f'(0) = -1$. For this case, one of the known methods to determine the stability of the fixed point $x^* = 0$ is evaluating the Schwarzian derivatives (up to third order for the given f), which includes higher order derivatives (up to seventh order for the given f). One of the other methods uses $g = f \circ f$ instead of f and focuses on g for the stability analysis. One can find that $g'(0) = 1$, $g''(0) = g^{(3)}(0) = \ldots = g^{(6)}(0) = 0$, and $g^{(7)}(0) = -15120$ which yields the local stability of the origin. Another method was introduced in [5], which is based on the higher order derivatives (for this case, again, seventh order derivative is required). Obviously, all of these methods require too lengthy calculations.

In this section, we will discuss the stability for oscillatory maps. For some cases, first order derivative will be sufficient for determining the stability. We will also give a theorem with a sufficient condition which may help us determine the stability of non-hyperbolic maps.

Clearly, there exists a stable (not asymptotically) period-2 orbit when $(f \circ f)(x) \equiv x$, that is when $f(x) = f^{-1}(x)$, on $B_\varepsilon(0)$.

Theorem 3 *Consider the map $x_{n+1} = f(x_n)$. Let $f(0) = 0$ and $f'(0) = -1$. Then the fixed point $x^* = 0$ is*

1. *locally asymptotically stable if $(f \circ f)'(x) < 1$ on $B_\varepsilon(0)$.*
2. *unstable if $(f \circ f)'(x) > 1$ on $B_\varepsilon^+(0)$ or $B_\varepsilon^-(0)$.*

Proof The result is straightforward by Lemma 5 and Theorem 1.

Example 2 For the difference equation $x_{n+1} = -x + ax^2$, where $a \in \mathbb{R}$, since $(f \circ f)'(x) = f'(x)f'(f(x)) = 1 - 6a^2x^2 < 1$ at the vicinity of the origin, $x = 0$ is an asymptotically stable fixed point. Here, Schwarzian derivative is $Sf(x) = -f'''(x) - \frac{3}{2}(f''(x))^2 = -6a^2 < 0$, which also yields asymptotic stability. Note that a generalization of Schwarzian derivatives can be obtained by finding the coefficients of the first nonzero term of the Taylor series expansion of the function $f'(x)f'(f(x)) - 1$.

Theorem 4 *Consider the discrete dynamical system $x_{n+1} = f(x_n)$, where $f(0) = 0$, and $f'(0) = -1$. Then the fixed point $x^* = 0$ is*

1. *locally asymptotically stable if $|f(x)| < |f^{-1}(x)|$ on $B_\varepsilon(0)$.*
2. *unstable if $|f(x)| > |f^{-1}(x)|$ on $B_\varepsilon(0)$.*

Proof By Lemma 5, we know that stability character of the fixed point $x^* = 0$ under f and $g = f \circ f$ are the same. Since $f(0) = 0$ and $f'(0) = -1$, we have $f(x) < 0$ on $B_\varepsilon^+(0)$ and $f(x) > 0$ on $B_\varepsilon^-(0)$. It is true that $f^{-1}(x) < 0$ on $B_\varepsilon^+(0)$ and $f^{-1}(x) > 0$ on $B_\varepsilon^-(0)$. Now, we will prove the first part of the theorem. Second part can be done similarly.

The assumption $|f(x)| < |f^{-1}(x)|$ in the first part of the theorem can be written as $f(x) > f^{-1}(x)$ on $B_\varepsilon^+(0)$ and $f(x) < f^{-1}(x)$ on $B_\varepsilon^-(0)$. Since f is a

decreasing continuous function, applying f to the both sides of the inequalities, we obtain $(f \circ f)(x) < x$ on $B_\varepsilon^+(0)$ and $(f \circ f)(x) > x$ on $B_\varepsilon^-(0)$. Therefore, we have $(f \circ f)'(x) < 1$ on $B_\varepsilon(0)$ and thus, by Theorem 3, the origin is locally asymptotically stable.

2.2.1 A Sufficient Condition

Now, we present one of the main results of this study, namely, a sufficient condition for the stability of non-hyperbolic fixed points for oscillatory maps.

Theorem 5 *(Test of Stability). Consider the map $x_{n+1} = f(x_n)$. Let $f(0) = 0$ and $f'(0) = -1$. Then the fixed point $x^* = 0$ is locally asymptotically stable if*

$$f'(x)f'(-x) < 1$$

for $x \in B_\varepsilon^+(0)$.

Proof Equivalently, we will prove the following statement: If the fixed point $x^* = 0$ is not locally asymptotically stable, then $f'(x)f'(-x) \geq 1$ on $B_\varepsilon^+(0)$.

Assume that $x^* = 0$ is not locally asymptotically stable fixed point. Hence, by Theorem 4, $|f(x)| \geq |f^{-1}(x)|$ on $B_\varepsilon(0)$. Therefore, $f(x) \leq f^{-1}(x)$ on $B_\varepsilon^+(0)$, and thus, clearly we have $f'(x) \leq (f^{-1})'(x)$ or

$$\frac{f'(x)}{(f^{-1})'(x)} \geq 1$$

on $B_\varepsilon^+(0)$. Now let us split the proof into following four cases.

1. The case $f(x) \leq f^{-1}(x) < -x$: Both $f^{-1}(x)$ and $f(x)$ are tangent to the line $y = -x$ at the origin and they are below the line $y = -x$ on $B_\varepsilon(0)$. Thus, f is concave down, that is $f''(x) < 0$, on $B_\varepsilon(0)$. Hence, f' is decreasing. Since f' is also continuous, applying f' to each parts of the inequality $f(x) \leq f^{-1}(x) < -x$, we obtain
$$f'(f(x)) \geq f'(f^{-1}(x)) > f'(-x).$$
Now, we multiply each part of the inequality by $f'(x) < 0$ and obtain
$$f'(x)f'(-x) > f'(x)f'(f^{-1}(x)) \geq (f \circ f)'(x)).$$
By the inequality on the left, we obtain
$$f'(x)f'(-x) > \frac{f'(x)}{(f^{-1})'(x)} > 1,$$
which completes the proof for this case.

2. The case $f(x) < -x < f^{-1}(x)$: When $x \in B_\varepsilon^+(0)$, each of the components $f(x)$, $-x$, and $f^{-1}(x)$ is negative. Since $y = f(x)$ is tangent to the line $y = -x$ at the origin and it is above the line $y = -x$ on $B_\varepsilon^-(0)$, we have $f''(x) > 0$ on $B_\varepsilon^-(0)$. Hence, f' is an increasing function on $B_\varepsilon^-(0)$. Now, applying f' to the inequality $f(x) < -x < f^{-1}(x)$ from the left, we obtain

$$f'(f(x)) < f'(-x) < f'(f^{-1}(x)).$$

Now, we will multiply each side of the above inequality by $f'(x)$, which is a negative quantity, to get

$$(f \circ f)'(x) > f'(x)f'(-x) > \frac{f'(x)}{(f^{-1})'(x)} > 1,$$

which yields the desired result.

3. The case $-x < f(x) \leq f^{-1}(x)$: This case is also similar to the previous cases. Knowing that $f''(x) > 0$, which means $f'(x)$ is increasing, we apply the composition with f' from the left. Since $f'(x) < 0$ on $B_\varepsilon(0)$, we obtain the inequality after multiplying by $f'(x)$:

$$f'(x)f'(-x) > (f \circ f)'(x) \geq \frac{f'(x)}{(f^{-1})'(x))} > 1.$$

4. The case $-x = f(x) = f^{-1}(x)$: It is straightforward that $f'(x)f'(-x) \geq 1$, which is the desired result.

Remark 1 Note that, Theorem 5 (Test of Stability) is only a sufficient condition, and hence there might be asymptotically stable fixed points which does not satisfy the condition in the theorem. For example, for the difference equation $x_{n+1} = -x_n + 3x_n^2 - 8x_n^3$, the usual method of taking the composition $g = f \circ f$ gives $g'(0) = 1$, $g''(0) = 0$, and $g'''(0) = -12 < 0$, which yields the asymptotic stability of the origin. However, we have $f'(x)f'(-x) > 1$ on $B_\varepsilon^+(0)$.

3 Applications

3.1 An Example

We will now show that the fixed point $x^* = 0$ is locally asymptotically stable for the difference equation

$$x_{n+1} = -x_n e^{-x_n^k}$$

for any positive integer k. Clearly, taking composition or using Schwarzian derivative will give us complicated expressions. However, Test of Stability (Theorem 5) gives us

very straightforward result. The first derivative will be sufficient for our discussion. We have $f(x) = -xe^{-x^k}$, hence $f'(x) = (kx^k - 1)e^{-x^k}$ and $f'(0) = -1$.

1. If k is odd, then $f'(-x)f'(x) = 1 - k^2x^{2k} < 1$ at the vicinity of $x^* = 0$.
2. If k is even, then $f'(-x)f'(x) = \frac{(1-kx^k)^2}{e^{2x^k}} < 1$ at the vicinity of $x^* = 0$.

Therefore, $x^* = 0$ is an asymptotically stable fixed point for any positive integer k.

Remark 2 Note that, when the fixed point x^* is not at the origin, it is easy to shift the fixed point to the origin by the transformation $y_n = x_n - x^*$ and apply Theorem 5. Another way to apply the theorem for a nonzero fixed point is, clearly, using the condition

$$f'(x^* - x)f'(x^* + x) < 1.$$

3.2 Population Models

In population dynamics, two of the well-known single-species population models are Logistic Map and the Ricker Map.

1. The Logistic Map $x_{n+1} = \mu x_n(1 - x_n)$, when $\mu = 3$, has two fixed points, one of which is $x^* = \frac{2}{3}$. We have $f'(x) = 3 - 6x$ and $f'\left(\frac{2}{3}\right) = -1$. Hence, we obtain $f'\left(\frac{2}{3} - x\right) f'\left(\frac{2}{3} + x\right) = 1 - 36x^2 < 1$ on $B_\varepsilon(0)$. Therefore, the fixed point $x^* = \frac{2}{3}$ of the logistic map is asymptotically stable.
2. Similarly, for the Ricker Map $x_{n+1} = x_n \exp(r - x_n)$, when $r = 2$, one of the fixed points is $x^* = 2$. For this case, $f'(2) = -1$ and $f'(2 - x)\, f'(2 + x) = 1 - x^2 < 1$ on $B_\varepsilon(0)$. Thus, the fixed point $x^* = 2$ of the Ricker map is asymptotically stable.

3.3 One-Dimensional Maps with Even or Odd Functions

The following theorem gives the stability/instability of some maps with even or odd functions, and other similar rules can be derived. These are direct results of our main theorems, and we will give the proof of only the second part of the theorem. The other parts can be done similarly. The non-oscillatory case is most of the time straightforward. One can easily apply Theorems 1 or 2 in order to investigate stability. Since the oscillatory case is usually more challenging, we mostly focus on that case by applying Test of Stability (Theorem 5).

Theorem 6 *Let $E(x)$ and $F(x)$ be even and odd functions, respectively. Then, following statements hold true.*

1. *For the map $x_{n+1} = x_n + E(x_n)$, where $E(0) = 0$, the fixed point $x^* = 0$ is unstable.*

2. *For the map* $x_{n+1} = -x_n + E(x_n)$, *where* $E(0) = 0$, *the fixed point* $x^* = 0$ *is asymptotically stable.*
3. *For the map* $x_{n+1} = F(x_n)$, *where* $F'(0) = -1$, *the fixed point* $x^* = 0$ *is asymptotically stable if* $|F'(x)| < 1$ *in a neighbourhood of the origin.*
4. *The fixed point* $x^* = 0$ *is asymptotically stable for the map*

$$x_{n+1} = -x_n e^{F(x_n)}.$$

Proof *(Part 2).* Since derivative of an even function is an odd function, taking $f(x) = -x + E(x)$, we obtain $f'(x)f'(-x) = 1 - (E'(x))^2 < 1$ at the vicinity of $x^* = 0$. Therefore, by Theorem 5, the origin is asymptotically stable.

Example 3 For the map $x_{n+1} = -x_n + x_n \sin^3 x_n + x_n^8 \cos x_n^5$, the origin is asymptotically stable by Part (2) of Theorem 6. Similarly, the origin is an asymptotically stable fixed point for the maps $x_{n+1} = -x_n + \ln(1 + x_n^2)$, $x_{n+1} = -x_n + 2x_n^2 - 3x_n^6$, and $x_{n+1} = -x_n + x_n^2 e^{-x_n^2}$.

Example 4 Consider the map

$$x_{n+1} = -x_n \cos x_n^k,$$

where k is a positive integer. Since the function $f(x) = -x\cos x^k$ is an odd function, and $f'(0) = -1$, we clearly have $|f'(x)| = |\cos x^k - kx^k \sin x^k| < 1$ at the vicinity of $x^* = 0$, and by Part (3) of Theorem 6, the origin is asymptotically stable for any positive integer k.

Note that, by any usual method, first $2k + 1$ derivatives at the fixed point must be evaluated. However, with the Test of Stability, we need only the first derivative at the vicinity of the fixed point.

4 Conclusions

The stability of non-hyperbolic fixed points of one-dimensional maps was investigated. We gave a stability condition which contains only the first order derivative and focuses on the vicinity of the fixed point. A sufficient condition for the stability (Test of Stability) for oscillatory maps was introduced. The condition is easy to apply and we may obtain the stability for complicated maps as well. For the oscillatory case, by the methods given in [1, 5], the derivative of the map must be taken at least three times. Since we have to evaluate the first derivative in any case, surely it is easier to firstly use Test of Stability by evaluating $f'(x)f'(-x)$, and if it does not work, then try one of the usual methods.

We used the Test of Stability for the well-known population models such as Logistic and Ricker Maps to determine the stability of the existence (positive) fixed point. Although the Test of Stability is a sufficient condition, it is still very powerful. It

allows us to construct some general rules for stability for some families of functions, for example maps with even or odd functions.

The results here can also be applied in center manifold theory when one obtains a one dimensional non-hyperbolic map on the center manifold. It is also possible to apply these results in the area of zero-diagonal planar maps [4] when converted to a one dimensional map with a non-hyperbolic fixed point.

A Related Lemmas/Theorems

Lemma 1 *Let f be an analytic function, $f(0) = 0$, and $f'(0) = 1$.*

1. *If $f'(x) < 1$ on $B_\varepsilon^+(0)$, then $0 < f(x) < x$ on $B_\varepsilon^+(0)$.*
2. *If $f'(x) < 1$ on $B_\varepsilon^-(0)$, then $x < f(x) < 0$ on $B_\varepsilon^-(0)$.*
3. *If $f'(x) > 1$ on $B_\varepsilon^+(0)$, then $f(x) > x$ on $B_\varepsilon^+(0)$.*
4. *If $f'(x) > 1$ on $B_\varepsilon^-(0)$, then $f(x) < x$ on $B_\varepsilon^-(0)$.*

Proof 1. Let $f'(x) < 1$ on $B_\varepsilon^+(0)$. Firstly, by contradiction, we will show that $f(x) < x$. Assume that $f(x) \geq x$ for some $x = a \in B_\varepsilon^+(0)$. Hence, by Mean Value Theorem (MVT), there exists a number $a_0 \in B_\varepsilon^+(0)$ such that $f'(a_0) = \frac{f(a)}{a} \geq 1$, which contradicts the assumption that $f'(x) < 1$ for all $x \in B_\varepsilon^+(0)$. Therefore, if $f'(x) < 1$, then $f(x) < x$ for $x \in B_\varepsilon^+(0)$.
Similarly, we can show that $0 < f(x)$. Since $f'(0) = 1$ is positive, we know by the continuity of f' that, f' is positive for some neighborhood of the origin. Let us assume that $f(x) \leq 0$ for some $x = a \in B_\varepsilon^+(0)$. Thus, by MVT, there exists a number $a_1 \in B_\varepsilon^+(0)$ such that $f'(a_1) = \frac{f(a)}{a} \leq 0$, which contradicts the fact that $f'(x) > 0$ for some neighborhood of the origin. Therefore, if $f'(x) < 1$, then $f(x) > 0$ for $x \in B_\varepsilon^+(0)$.
Combining the results, we complete the proof: if $f'(x) < 1$, then $0 < f(x) < x$ on $x \in B_\varepsilon^+(0)$.

2. Let $f'(x) < 1$ for all $x \in B_\varepsilon^-(0) = (-\varepsilon, 0)$. We will show that $x < f(x)$. We use the similar approach: Assume that $x \geq f(x)$ for some $x = a$ in the interval $B_\varepsilon^-(0)$. By MVT, there exists a number $a_0 \in B_\varepsilon^-(0)$ such that $f'(a_0) = \frac{f(a)}{a} \geq 1$, which contradicts the assumption that $f'(x) < 1$ for all $x \in B_\varepsilon^-(0)$. Therefore, if $f'(x) < 1$, then $x < f(x)$ for $x \in B_\varepsilon^-(0)$.
Similarly, we can show that $f(x) < 0$. Since $f'(0) = 1$ is positive, f' is positive for some neighborhood of the origin. Let us assume that $f(x) \geq 0$ for some $x = a \in B_\varepsilon^-(0)$. Thus, there exists a number $a_1 \in B_\varepsilon^-(0)$ such that $f'(a_1) = \frac{f(a)}{a} \leq 0$, which contradicts the fact that $f'(x) > 0$ for some neighborhood of the origin. Therefore, if $f'(x) < 1$, then $f(x) < 0$ for $x \in B_\varepsilon^-(0)$.

3. Let $f'(x) > 1$ on $B_\varepsilon^+(0)$ and assume that $f(x) \leq x$ for some $x = a \in B_\varepsilon^+(0)$. Hence, similarly, by MVT, there exists a number $a_0 \in B_varepsilon^+(0)$ such that $f'(a_0) = \frac{f(a)}{a} \leq 1$, which contradicts the assumption that $f'(x) > 1$ for all $x \in B_\varepsilon^+(0)$. Therefore, if $f'(x) > 1$, then $f(x) > x$ for $x \in B_\varepsilon^+(0)$.

4. Let $f'(x) > 1$ on $B_\varepsilon^-(0)$ and assume that $f(x) \geq x$ for some $x = a \in B_\varepsilon^-(0)$. Hence, by MVT, there exists a number $a_0 \in B_\varepsilon^-(0)$ such that $f'(a_0) = \frac{f(a)}{a} \leq 1$, which contradicts the assumption that $f'(x) > 1$ for all $x \in B_\varepsilon^-(0)$. Therefore, if $f'(x) > 1$, then $f(x) < x$ for $x \in B_\varepsilon^-(0)$.

Lemma 2 *Let f be an analytic function, $f(0) = 0$, and $f'(0) = -1$. The followings hold:*

1. *If $(f \circ f)'(x) < 1$ on $B_\varepsilon^+(0)$, then $f^{-1}(x) < f(x) < 0$ on $B_\varepsilon^+(0)$.*
2. *If $(f \circ f)'(x) < 1$ on $B_\varepsilon^-(0)$, then $0 < f(x) < f^{-1}(x) < 0$ on $B_\varepsilon^-(0)$.*
3. *If $(f \circ f)'(x) > 1$ on $B_\varepsilon^+(0)$, then $f^{-1}(x) < f(x)$ on $B_\varepsilon^+(0)$.*
4. *If $(f \circ f)'(x) > 1$ on $B_\varepsilon^-(0)$, then $f(x) < f^{-1}(x)$ on $B_\varepsilon^-(0)$.*

Proof Setting $g = f \circ f$, using Lemma 1 with function g, and applying the inverse function f^{-1} to the inequalities from the left, one can obtain the desired result. Note that f^{-1} is a decreasing function on $B_\varepsilon^-(0)$, which changes the direction of the inequalities.

Lemma 3 *Let f be an analytic function. Consider the discrete dynamical system $x_{n+1} = f(x_n)$, where $f(0) = 0$, and $f'(0) = 1$. Assume that the following condition is satisfied for some $\varepsilon > 0$.*

$$\begin{cases} 0 < f(x) < x, & \text{if } x \in B_\varepsilon^+(0), \\ x < f(x) < 0, & \text{if } x \in B_\varepsilon^-(0). \end{cases}$$

If $x_0 \in B_\varepsilon(0)$, then $x_n \to 0$ as $n \to \infty$.

Proof Since f is an analytic function and $f'(0) = 1 > 0$, for sufficiently small $\varepsilon > 0$, f is increasing on $B_\varepsilon(0)$. Now, let us focus on the case $x > 0$. Take an initial point $x_0 \in (0, \varepsilon)$. We have the inequality $0 < f(x_0) < x_0$. Since f is increasing on $B_\varepsilon(0)$, applying f to both sides of this inequality from the left, we obtain $0 = f(0) < x_2 = f(f(x_0)) < f(x_0) = x_1$. Similarly, we apply f to the obtained inequality over and over to get

$$0 < \ldots < x_k < \ldots < x_3 < x_2 < x_1 < x_0.$$

By Monotone Convergence Theorem, the limit of the sequence x_n exists. Let the limit be L and we have $L = \lim f^{n+1}(x_0) = f(\lim f^n(x_0)) = f(L)$. Hence, the limit must be 0, which is the only fixed point of the difference equation.

The case when $x < 0$ can be done similarly. Therefore, $\lim x_n \to 0$ for any $x_0 \in B_\varepsilon(0)$.

Lemma 4 *Let f be an analytic function, $f(0) = 0$, and $f'(0) = 1$.*

1. *If $f(x) > x$ on $B_\varepsilon^+(0)$, then 0 is unstable.*
2. *If $f(x) < x$ on $B_\varepsilon^-(0)$, then 0 is unstable.*

Proof 1. Take $\varepsilon' < \varepsilon$ and let $x_0 \in B^+_{\varepsilon'}(0)$. Hence, we have that $f(x) > x$ on $(0, \varepsilon']$. Assume that $0 < f^n(x_0) < \varepsilon'$ for all positive integer n. Then, $x_1 = f(x_0) > x_0$ and applying f from the left over and over, we obtain

$$0 < \ldots < x_0 < x_1 < x_2 < x_3 < \ldots < x_n < \ldots < \varepsilon'.$$

By Monotone Convergence Theorem, the limit of the sequence x_n exists. Let the limit be L and we have $L = \lim f^{n+1}(x_0) = f(\lim f^n(x_0)) = f(L)$. Hence, the limit must be a fixed point. However, since $f(x) > x$ on $(0, \varepsilon']$, there is no fixed point on this interval. This contradiction completes the proof.

2. Proof for this case can be done similarly.

Theorem 7 *Let z be an attracting fixed point of a continuous map $f : I \to \mathbb{R}$, where I is an interval. Then z is stable.*

Proof of Theorem 7 can be found in [3], on p. 239.

Lemma 5 *Consider the difference equations*

$$x_{n+1} = f(x_n) \tag{2}$$

and

$$x_{n+1} = (f \circ f)(x_n), \tag{3}$$

where $f(x^*) = x^*$ and $f'(x^*) = -1$.

x^* is asymptotically stable under equation (2) if and only if x^* is asymptotically stable under equation (3).

Proof See [2, 5].

References

1. Dannan, F.M., Elaydi, S.N., Ponomarenko, V.: Stability of hyperbolic and nonhyperbolic fixed points of one-dimensional maps. J. Diff. Equ. Appl. **9**(5), 449–457 (2003)
2. Elaydi, S.: An Introduction to Difference Equations. Springer (2000)
3. Elaydi, S.N.: Discrete Chaos: With Applications in Science and Engineering. Chapman and Hall/CRC (2008)
4. Kapçak, S.: On planar zero-diagonal and zero-trace iterated maps. J. Diff. Equ. Appl. **24**(9), 1521–1539 (2018)
5. Murakami, K.: Stability for non-hyperbolic fixed points of scalar difference equations. J. Math. Anal. Appl. **310**(2), 492–505 (2005)
6. Ponomarenko, V.: Faà di Bruno's formula and nonhyperbolic fixed points of one-dimensional maps. Int. J. Math. Math. Sci. **2004**(29), 1543–1549 (2004)

Impulse Effect on a Population Model with Piecewise Constant Argument

Fatma Karakoç

Abstract We consider a population model with piecewise constant argument under impulse effect. First, we deal with the model with impulses. Sufficient conditions for the oscillation of the solutions are obtained. We also investigate asymptotic behavior of the non-oscillatory solutions. Then we obtain similar results for the same model without impulse effect. Finally, we compare the results with non-impulsive case and we give some examples to illustrate our results.

Keywords Population model · Piecewise constant argument · Impulse · Difference equation · Linearized oscillation · Non-oscillation.

1 Introduction

In this paper we investigate asymptotic behavior of the positive solutions of the following population model

$$N'(t) = -\gamma N(t) + \frac{\beta N(t)}{r + N^m([t-k])},\ t \geq 0,\ t \neq n,\ n = 1, 2, ..., \tag{1}$$

$$N(n^+) = N(n^-)\left(\frac{N^*}{N(n-l)}\right)^b,\ n = 1, 2, ..., \tag{2}$$

where $\beta, \gamma, m \in (0, \infty), r \in [0, \infty),\ N^* = \left(\frac{\beta}{\gamma} - r\right)^{1/m}$ are constants, $k \in Z^+ = \{1, 2, 3, ...\}$ and $l \in \{2, 3, ...\}$ are fixed numbers, $b \geq 0$ is a constant, [.] denotes the greatest integer function, $N(n^+) = \lim_{t \to n^+} N(t)$ and $N(n^-) = \lim_{t \to n^-} N(t)$.

The following population model related to control of a single population of cells was presented by Nazerenko [1].

F. Karakoç (✉)
Ankara University, Faculty of Sciences, Department of Mathematics, Ankara 06100, Turkey
e-mail: fkarakoc@ankara.edu.tr

S. Baigent et al. (eds.), *Progress on Difference Equations and Discrete Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 341,
https://doi.org/10.1007/978-3-030-60107-2_13

$$x'(t) + px(t) - q\frac{x(t)}{r + x^n(t-\tau)} = 0$$

Then stability and oscillation of the solutions of the above differential equation was dealt with in [2].

Studies on delay differential equations with piecewise constant arguments instead of continuous arguments have been started in 1980's. In the beginning, stability, oscillation and existence of periodic solutions of the linear differential equations were investigated [3–9] and the references cited therein. To the best of our knowledge, there is only a few paper on the asymptotic behavior of biological models with piecewise constant arguments. One of the logistic model with piecewise constant arguments

$$\frac{dN}{dt} = rN(t)\left\{1 - \sum_{j=0}^{m} p_j N([t-j])\right\}.$$

was investigated in [10] and a necessary and sufficient condition for the oscillation of the positive solutions was established. Recently, asymptotic behavior of the solutions of non-impulsive differential Equation (1) in the case of $k = 1$ has been studied in [11]. In real world problems, it is known that the solutions of the mathematical models may be discontinuous as well as an exterior effect may change the asymptotic behavior of the solutions. Because of this reality, studies on the impulsive differential equations with piecewise constant arguments have been started with the works [12–15]. So, the aim of the present paper is to show how can an exterior effect change the asymptotic behavior of the population model (1). For this purpose we consider Eq. (1) with the impulse conditions (2). The main tool of our technique is linearized oscillation of difference equations. So, Sect. 2 is devoted to some fundamental definitions and results of linearized oscillation theory. In Sect. 3, we prove the main results. Finally, we consider some examples to compare the results of impulsive differential equations models with non-impulsive differential equations models.

2 Preliminaries

Define $k_0 = \max\{k, l\}$.

Definition 1 It is said that a function $N(t)$ defined on the set $\{-k_0, 1-k_0, \ldots -1\} \cup [0, \infty)$ is a solution of Eqs. (1)–(2) if it satisfies the following conditions:

(i) $N(t)$ is continuous on R^+ with the possible exception of the points $[t] \in [0, \infty)$,
(ii) $N(t)$ is right continuous and has left-hand limit at the points $[t] \in [0, \infty)$,
(iii) $N(t)$ is differentiable and satisfies Eq. (1) for any $t \in R^+$, with the possible exception of the points $[t] \in [0, \infty)$ where one-sided derivatives exist,
(iv) $N(t)$ satisfies impulse conditions (2) for $n \in Z^+$.

Definition 2 A function $x(t)$ defined on $[0, \infty)$ is called oscillatory about zero if there exist two real valued sequences $\{t_n\}_{n\geq 0}$, $\{t'_n\}_{n\geq 0} \subset [0, \infty)$ such that $t_n \to +\infty$, $t'_n \to +\infty$ as $n \to +\infty$ and $x(t_n) \leq 0 \leq x\left(t'_n\right)$ for $n \geq N_1$ where N_1 is sufficiently large. Otherwise, the function $x(t)$ is called non-oscillatory.

Remark 1 According to Definition 2, a piecewise continuous function $x : [0, \infty) \to R$ can be oscillatory even if $x(t) \neq 0$ for all $t \in [0, \infty)$.

Definition 3 ([16]) A function $x(t)$ is called oscillatory about K^* if the function $(x(t) - K^*)$ is oscillatory about zero.

Difference equations are main tool for the investigation of differential equations with piecewise constant arguments. So, we recall the following definition and theorems which will be used in the proofs of main results.

Definition 4 ([16]) The sequence $\{y_n\}$ is said oscillatory if it is neither eventually positive nor eventually negative. Otherwise, it is called non-oscillatory.

Theorem 1 *(*[16], Corollary 7.4.1). Assume that $\lim_{u\to 0} \frac{f_i(u)}{u} = 1$ for $i = 1, 2, ..., m$ and there exists a positive constant δ such that

$$\begin{cases} either\ f_i(u) \leq u\ for\ 0 \leq u \leq \delta\ and\ i = 1, 2, ..., m \\ or\ f_i(u) \geq u\ for\ -\delta \leq u \leq 0\ and\ i = 1, 2, ..., m. \end{cases}$$

Then every solution of equation

$$a_{n+1} - a_n + \sum_{i=1}^{m} p_i f_i(a_{n-k_i}) = 0,\ n = 0, 1, 2, ... \tag{3}$$

oscillates if and only if every solution of its linearized equation

$$b_{n+1} - b_n + \sum_{i=1}^{m} p_i b_{n-k_i} = 0,\ n = 0, 1, 2, ... \tag{4}$$

oscillates, where $p_i \in (0, \infty)$ and $k_i \in \{0, 1, 2...\}$ for $i = 1, 2, ..., m$, with $\sum_{i=1}^{m}(p_i + k_i) \neq 1$, $f_i \in C(R, R)$ and $uf_i(u) > 0$ for $u \neq 0$.

The following theorem gives a sufficient condition for the existence of oscillatory solutions for the linear equation (4).

Theorem 2 *(*[16], Theorem 7.3.1). Suppose that

$$\sum_{i=1}^{m} p_i \frac{(k_i + 1)^{k_i+1}}{k_i^{k_i}} > 1. \tag{5}$$

Then every solution of Eq. (4) oscillates.

By the biological meaning we consider differential Equation (1)–(2) with the positive initial conditions

$$N(-k_0) > 0,\ N(1-k_0) > 0, ..., N(-1) > 0,\ N(0) > 0. \tag{6}$$

3 Main Results

The purpose of this section is to investigate the asymptotic behavior of the positive solutions of Eqs. (1)–(2). Throughout this section we assume that $\frac{\beta}{\gamma} > r$. Then, it is easy to see that every solution of the Eqs. (1)–(2) with the positive initial conditions (6) is positive and $N^* = \left(\frac{\beta}{\gamma} - r\right)^{1/m}$ is the positive equilibrium point of the Eqs. (1)–(2).
Using substitution $N(t) = N^* e^{x(t)}$, Eqs. (1)–(2) reduces to following differential equation

$$x'(t) = -\gamma + \frac{\beta}{r + (N^*)^m e^{mx([t-k])}},\ t \neq n,\ n = 1, 2, ... \tag{7}$$

$$x(n^+) - x(n^-) = -bx(n-l),\ n = 1, 2, ... \tag{8}$$

So, we shall investigate the properties of the Eqs. (7)–(8). We consider Eqs. (7)–(8) with the initial conditions

$$x(-k_0) = \ln \frac{N_{-k_0}}{K} = x_{-k_o}, ...,\ x(-1) = \ln \frac{N_{-1}}{K} = x_{-1},\ x(0) = \ln \frac{N_0}{K} = x_0 \tag{9}$$

In the following theorem we obtain the unique solution of the initial value problem (7)–(9).

Theorem 3 *The unique solution $x(t)$ defined on $\{-k_0, 1-k_0, ..., -1\} \cup [0, \infty)$ of the initial value problem (7)–(9) has the following representation*

$$x\,(t) = y(n) + \left(-\gamma + \frac{\beta}{r + (N^*)^m e^{my(n-k)}}\right)(t-n), n \leq t < n+1,\ n \in N, \tag{10}$$

where the sequence $y(n)$ is the unique solution of the difference equation

$$y(n+1) - y(n) + \gamma - \frac{\beta}{r + (N^*)^m e^{my(n-k)}} + by(n-l+1) = 0 \tag{11}$$

with the initial conditions

$$y(-k_0) = x_{-k_0}, ..., y(-1) = x_{-1},\ y(0) = x_0. \tag{12}$$

Proof Let $x_n(t) \equiv x(t)$ be a solution of (7)–(8) on $n \leq t < n+1$. Equation (7) is rewritten in the form

$$x'(t) = -\gamma + \frac{\beta}{r + (N^*)^m e^{mx(n-k)}}, n < t < n+1. \tag{13}$$

Integrating both sides of Eqs. (13) from n to t we obtain that

$$x_n(t) = x\left(n^+\right) + \left(-\gamma + \frac{\beta}{r + (N^*)^m e^{mx(n-k)}}\right)(t-n),\ n < t < n+1. \tag{14}$$

On the other hand, if $x_{n-1}(t)$ is a solution of Eqs. (7)–(8) on $n-1 \leq t < n$, then we get

$$x_{n-1}(t) = x\left((n-1)^+\right) + \left(-\gamma + \frac{\beta}{r + (N^*)^m e^{mx(n-1-k)}}\right)(t-n+1),\ n-1 < t < n. \tag{15}$$

Using the impulse conditions (8), from (14) and (15) we obtain that

$$x\left(n^+\right) - x\left((n-1)^+\right) - \left(-\gamma + \frac{\beta}{r + (N^*)^m e^{mx(n-1-k)}}\right) = -bx(n-l).$$

Since x is right continuous at the points $t = n,\ n = 1, 2, ...$, above equation gives the difference equation (11). Considering the initial conditions (12), the solution of Eq. (11) can be obtained uniquely. Thus, the solution of (7)–(8) with (9) is obtained as (10).

Following theorem presents a sufficient condition for the oscillation of the difference Eq. (11).

Theorem 4 *Let assume that* $b > 0$, $\frac{\beta}{\gamma} > 2r$ *and*

$$\frac{m\beta(N^*)^m}{(r + (N^*)^m)^2}\frac{(k+1)^{k+1}}{k^k} + b\frac{l^l}{(l-1)^{l-1}} > 1. \tag{16}$$

Then every solution of Eq. (11) is oscillatory.

Proof We use linearized oscillation for difference equations to prove the result. Equation (11) can be rewritten as

$$y(n+1) - y(n) + p_1 f_1(y(n-k)) + p_2 f_2(y(n-l+1)) = 0,$$

where $p_1 = \frac{m\beta(N^*)^m}{(r+(N^*)^m)^2} > 0$, $f_1(u) = \frac{(r+(N^*)^m)(e^{mu}-1)}{m(r+(N^*)^m e^{mu})} \in C(R, R)$, $p_2 = b > 0$, $f_2(u) = u \in C(R, R)$.

It is clear that $\sum_{i=1}^{2}(p_i + k_i) = \frac{m\beta(N^*)^m}{(r+(N^*)^m)^2} + k + b + l - 1 \neq 1$. Moreover, it is satisfied that $uf_i(u) > 0$ for $u \neq 0$ and $\lim_{u\to\infty} \frac{f_i(u)}{u} = 1,\ i = 1, 2$. Moreover, since $\frac{\beta}{\gamma} > 2r$, it is shown that

$$\frac{df_1}{du} = \frac{e^{mu}\,(r + (N^*)^m)^2}{(r + (N^*)^m e^{mu})^2} < 1 \; for \; u > 0.$$

Hence,

$$\frac{d}{du}(f_1(u) - u) < 0 \; for \; u > 0.$$

So, we get $f_1(u) < u$ for $u > 0$. Moreover, it is clear that $f_2(u) = u$ is also satisfied the last condition of Theorem 1. Therefore, by Theorem 1, every solution of Eq. (11) is oscillatory if and only if every solution of linearized equation

$$y(n+1) - y(n) + \frac{m\beta(N^*)^m}{(r + (N^*)^m)^2} y(n-k) + by(n-l+1) = 0 \qquad (17)$$

is oscillatory. Note that, by Theorem 2, under the condition (16), every solution of Eq. (17) is oscillatory. So, the proof is completed.

Now we get the following result for the solutions of Eqs. (1)–(2).

Corollary 1 *Let assume that* $b > 0$, $\frac{\beta}{\gamma} > 2r$. *Every solution of Eqs. (1)–(2) oscillates about* N^* *if the condition (16) is satisfied.*

Theorem 5 *Let assume that* $b > 0$. *If a solution* $N(t)$ *of Eqs. (1)–(2) is non-oscillatory about* N^*, *then* $\lim_{t\to\infty} N(t) = N^*$.

Proof It is sufficient to prove that for every nonoscillatory solution $x(t)$ of the Eqs. (7)–(8) $\lim_{t\to\infty} x(t) = 0$. Let $x(t)$ be an eventually positive solution of Eqs. (7)–(8). From Eq. (7) for $n < t < n + 1$, we get

$$x'(t) = -pf\,(x([t-k])) < 0,$$

where, $p = \frac{m\beta(N^*)^m}{(r+(N^*)^m)^2} > 0$, $f(u) = \frac{(r+(N^*)^m)(e^{mu}-1)}{m(r+(N^*)^m e^{mu})} > 0$ for $u > 0$. On the other hand, from the impulse conditions (8), we have

$$x(n^+) < x(n^-).$$

So, $\lim_{t\to\infty} x(t) = l \geq 0$ exists. Since $x(t) = y(n)$ for $t = n$, $\lim_{t\to\infty} y(n) = l$. We claim that $l = 0$. Otherwise, taking the limit of both sides of Eq. (11) as $n \to \infty$, we obtain that

$$0 = l - l = \left(-\gamma + \frac{\beta}{r + (N^*)^m e^{ml}}\right) - bl < 0$$

which is a contradiction. So, $l = 0$. If $x(t)$ is an eventually negative solution of Eqs. (7)–(8), then we obtain same result.

Now, let us consider the non-impulsive differential equation

$$N'(t) = -\gamma N(t) + \frac{\beta N(t)}{r + N^m([t-k])}, \ t \geq 0. \tag{18}$$

In [11] Eq. (18) with $k = 1$ is studied. The following results are generalizations of Corollary 1 and Theorem 5 in [11], respectively.

Corollary 2 *If* $\frac{\beta}{\gamma} > 2r$ *and*

$$\frac{m\beta(N^*)^m}{(r + (N^*)^m)^2} \frac{(k+1)^{k+1}}{k^k} > 1, \tag{19}$$

then every solution of Eq. (18) oscillates about N^*.

Corollary 3 *If a solution* $N(t)$ *of Eq. (18) is nonoscillatory about* N^*, *then* $\lim_{t\to\infty} N(t) = N^*$.

4 Numerical Examples

In this section we give some examples to illustrate our results.

Example 1 Let us consider the following differential equation

$$N'(t) = -N(t) + \frac{5N(t)}{1 + N^2([t-1])}, \ t \geq 0, \ \ t \neq n, \ n = 1, 2, ... \tag{20}$$

$$N(n^+) = N(n^-)\left(\frac{2}{N(n-3)}\right)^b, \ n = 1, 2, ... \tag{21}$$

where $b \geq 0$ is a constant. It can be seen that the $N^* = 2$ is the positive equilibrium point of the Eqs. (20)–(21). Moreover it is shown that $\frac{\beta}{\gamma} > 2r$ and the condition (16) is satisfied for all $b \geq 0$. So, from Corollary 1 and Corollary 2 all solutions of impulsive differential Equation (20)–(21) as well as all solutions of non-impulsive differential Equation (20) are oscillate about 2. The solutions $N(t)$ of the non-impulsive differential Equation (20) and impulsive differential Eqs. (20)–(21) for $b = 1/5$ with the initial conditions $N(-2) = N(-1) = N(0) = 1$ are demonstrated in Fig. 1 and Fig. 2, respectively.

Example 2 **(i)** Let us consider the following non-impulsive population model

$$N'(t) = -\frac{1}{8}N(t) + \frac{N(t)}{N^{1/2}([t-2])}, \ t \geq 0. \tag{22}$$

It is clear that Eq. (22) is a special case of (18) with $\gamma = \frac{1}{8}$, $\beta = 1$, $r = 0$, $m = 1/2$, $k = 2$. It is easy to see that $N^* = 64$ is the positive equilibrium point of the Eq. (22) and $\frac{\beta}{\gamma} > 2r$. But,

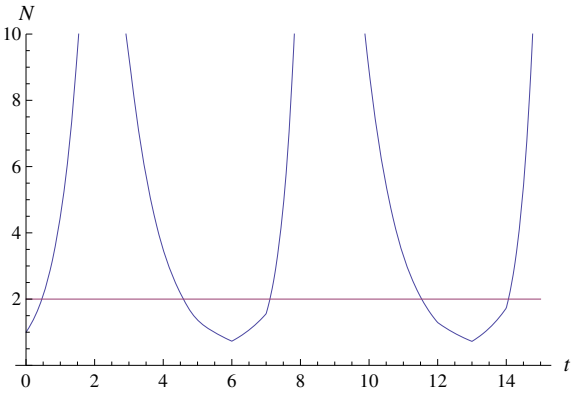

Fig. 1 The solution $N(t)$ of Eq. (20) with the initial conditions $N(-2) = N(-1) = 1$

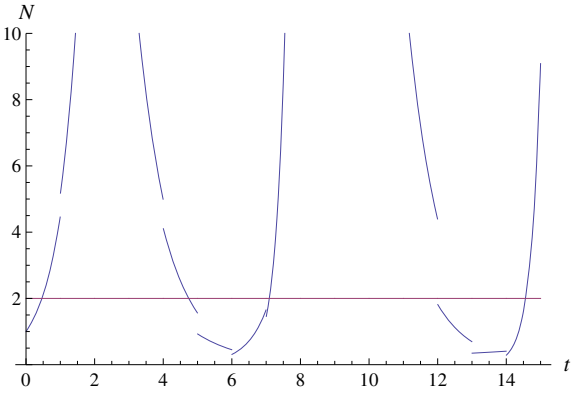

Fig. 2 The solution $N(t)$ of Eqs. (20)–(21) with the initial conditions $N(-2) = N(-1) = N(0) = 1$

$$\frac{m\beta(N^*)^m}{(r+(N^*)^m)^2}\frac{(k+1)^{k+1}}{k^k} < 1.$$

So, we can not apply Corollary 2. But, from Corollary 3, if a solution $N(t)$ of Eq. (22) is nonoscillatory about the positive equilibrium point 64, then $\lim_{t\to\infty} N(t) = 64$. The solution $N(t)$ of the Eq. (22) with the initial conditions $N(-2) = N(-1) = 0.5,\ N(0) = 1$ is demonstrated in Fig. 3.

(ii) Now let us consider the same population model under impulse effect

$$N'(t) = -\frac{1}{8}N(t) + \frac{N(t)}{N^{1/2}([t-2])},\ t \geq 0,\ t \neq n,\ n = 1, 2, ..., \tag{23}$$

$$N(n^+) = N(n^-)\left(\frac{64}{N(n-2)}\right)^{1/2},\ n = 1, 2, ... \tag{24}$$

It is clear that

$$\frac{m\beta(N^*)^m}{(r+(N^*)^m)^2}\frac{(k+1)^{k+1}}{k^k} + b\frac{l^l}{(l-1)^{l-1}} > 1.$$

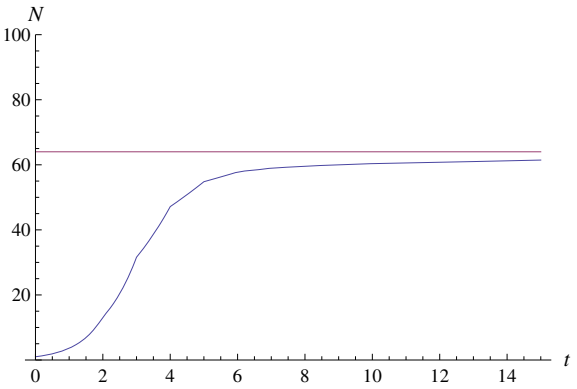

Fig. 3 The solution $N(t)$ of Eq. (22) with the initial conditions $N(-2) = N(-1) = 0.5,\ N(0) = 1$

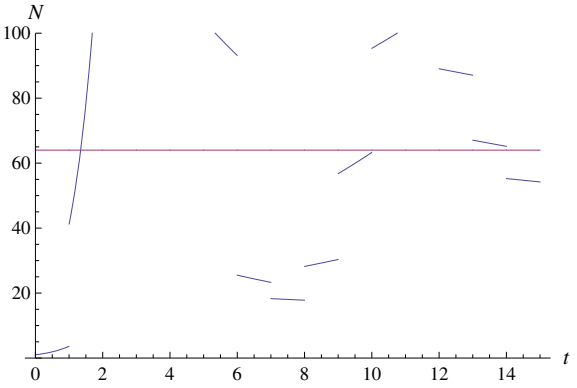

Fig. 4 The solution $N(t)$ of Eqs. (23)–(24) with the initial conditions $N(-2) = N(-1) = 0.5,\ N(0) = 1$

So, from Corollary 1, every solution of Eqs. (23)–(24) oscillates about the positive equilibrium point 64. The solution $N(t)$ of the Eqs. (23)–(24) with the initial conditions $N(-2) = N(-1) = 0.5,\ N(0) = 1$ is demonstrated in Fig. 4.

References

1. Nazarenko, V.G.: Influence of delay on auto-oscillation in cell populations. Biofisika **21**, 352–356 (1976)
2. Kubiaczyk, I., Saker, S.H.: Oscillation and stability in nonlinear delay differential equations of population dynamics. Math. Comput. Modelling **35**, 295–301 (2002)
3. Cooke, K.L., Wiener, J.: Retarded differential equations with piecewise constant delays. J. Math. Anal. Appl. **99**, 265–297 (1984)
4. Aftabizadeh, A.R., Wiener, J.: Oscillatory properties of first order linear functional differential equations. Appl. Anal. **20**, 165–187 (1985)
5. Aftabizadeh, A.R., Wiener, J., Ming Xu, J.: Oscillatory and periodic solutions of delay differential equations with piecewise constant argument. Proc. Am. Math. Soc. **99**, 673–679 (1987)

6. Gopalsamy, K., Gyori, I., Ladas, G.: Oscillations of a class of delay equations with continuous and piecewise constant arguments. Funkcialaj Ekvacioj **32**, 395–406 (1989)
7. Wiener, J.: Generalized Solutions of Functional Differential Equations. World Scientific, Singapore (1994)
8. Shen, J.H., Stavroulakis, I.P.: Oscillatory and nonoscillatory delay equations with piecewise constant argument. J. Math. Anal. Appl. **248**, 385–401 (2000)
9. Chiu, K.-S., Pinto, M.: Oscillatory and periodic solutions in alternately advanced and delayed differential equations. Carpathian J. Math. **29**(2), 149–158 (2013)
10. Gopalsamy, K., Kulenovic, M.R.S., Ladas, G.: On logistic equation with piecewise constant arguments. Differ. Integr. Equat. **4**, 215–223 (1990)
11. Karakoç, F.: Asymptotic behaviour of a population model with piecewise constant argument. Appl. Math. Letters **70**, 7–13 (2017)
12. Karakoç, F., Bereketoglu, H., Seyhan, G.: Oscillatory and periodic solutions of impulsive differential equations with piesewise constant argument. Acta Appl. Math. **110**, 499–510 (2010)
13. Bereketoglu, H., Seyhan, G., Ogun, A.: Advanced impulsive differential equations with piecewise constant arguments. Math. Model. Anal. **15**, 175–187 (2010)
14. Karakoç, F., Ogun Unal, A., Bereketoglu, H.: Oscillation of nonlinear impulsive differential equations with constant arguments. E. J. Qual. Theory Diff. Equ. No. **49**, 1–12 (2013)
15. Chiu, Kuo-Shou: On generalized impulsive piecewise constant delay differential equations. Sci. China Math. **58**(9), 1981–2002 (2015)
16. Gyori, I., Ladas, G.: Oscillation Theory of Delay Differential Equations with Applications. Oxford Science Publications, pp. xii+368. The Clarendon Press, Oxford University Press, New York (1991). ISBN: 0-19-853582-1

On a Second-Order Rational Difference Equation with Quadratic Terms, Part II

YEVGENIY KOSTROV and Zachary Kudlak

Abstract We give the character of solutions of the following second-order rational difference equation with quadratic denominator

$$x_{n+1} = \frac{\alpha + \beta x_n}{Bx_n + Dx_n x_{n-1} + x_{n-1}},$$

where the coefficients are positive numbers, and the initial conditions x_{-1} and x_0 are nonnegative such that the denominator is nonzero. In particular, we show that the unique positive equilibrium is locally asymptotically stable, and we give conditions on the coefficients for which the unique positive equilibrium is globally stable.

Keywords Local stability · Global stability · Rational difference equation · Rational recurrence relation

1 Introduction

In this paper, we will investigate the behavior of solutions of a second-order rational recurrence relation with a quadratic term.

Namely, we will consider the equation

$$x_{n+1} = \frac{\alpha + \beta x_n}{Bx_n + Dx_n x_{n-1} + x_{n-1}}, \text{ for } n = 0, 1, \ldots, \tag{1}$$

Y. KOSTROV (✉)
Manhattanville College, Purchase, NY 10577, USA
e-mail: yevgeniy.kostrov@mville.edu; kostrovy@mville.edu

Z. Kudlak
US Coast Guard Academy, New London, CT 06320, USA
e-mail: zachary.a.kudlak@uscga.edu

S. Baigent et al. (eds.), *Progress on Difference Equations and Discrete Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 341,
https://doi.org/10.1007/978-3-030-60107-2_14

where the coefficients are positive real numbers and the initial conditions are nonnegative real numbers such that the denominator is positive.

The difference equation in (1) is a special case of the more general difference equation

$$x_{n+1} = \frac{ax_n^2 + bx_nx_{n-1} + cx_{n-1}^2 + dx_n + ex_{n-1} + f}{Ax_n^2 + Bx_nx_{n-1} + Cx_{n-1}^2 + Dx_n + Ex_{n-1} + F}, n = 0, 1, \ldots, \quad (2)$$

with nonnegative coefficients and nonnegative initial conditions such that the denominator is positive. Several authors have investigated difference equations contained in (2), for several examples see [1, 3, 5, 7, 11].

Interestingly, upon investigation of (1), it is easy to see that the monotonicity with respect to x_n changes as the values of the parameters change. There are several recent papers where the authors investigate equations with varying monotone character, for reference, see [6, 10].

We will now state the main result of this paper.

Theorem 1 *Let $\{x_n\}_{n=-1}^{\infty}$ be a solution of (1). Then, there is a unique positive equilibrium $\bar{x}$ of (1). If any of the following are true*

1. $\beta \leq \alpha D + \sqrt{\alpha}B$; *or*
2. $\beta > \alpha D + \sqrt{\alpha}B$ *and* $B \geq 1$; *or*
3. $\beta > \alpha D + \sqrt{\alpha}B$, $B < 1$, *and* $\beta \leq \dfrac{\alpha D}{1 - B}$;

then the unique positive equilibrium is ***globally asymptotically stable***.

We will prove Theorem 1 in multiple steps. In Sect. 2 we provide some previously known results for reference. Next, in Sect. 3 we will state and prove several auxiliary results about (1). Section 4 will show that solutions of (1) will eventually enter an invariant interval. In Sect. 5, we prove that the unique positive equilibrium is a global attractor in the regions specified in Theorem 1.

2 Preliminaries

In this section, we state several well-known results which will be useful in this paper. We call the following two theorems the "m&M Theorems," see [4] for more details.

Theorem 2 *Let $g\colon [a, b] \times [a, b] \to [a, b]$ be a continuous function, where a and b are real numbers with $a < b$, and consider the difference equation*

$$x_{n+1} = g(x_n, x_{n-1}), \quad \text{for } n = 0, 1, \ldots . \quad (3)$$

Suppose that g satisfies the following two conditions:

1. *$g(x, y)$ is non-increasing in $x \in [a, b]$ for each fixed $y \in [a, b]$, and $g(x, y)$ is non-increasing in $y \in [a, b]$ for each fixed $x \in [a, b]$;*
2. *if (m, M) is a solution of the system*

$$m = g(M, M), \quad M = g(m, m), \tag{4}$$

then $m = M$.

Then, there exists exactly one equilibrium of (3), namely $\bar{x}$. Furthermore, every solution of (3) converges to $\bar{x}$.

Theorem 3 *Let $g\colon [a, b] \times [a, b] \to [a, b]$ be a continuous function, where a and b are real numbers with $a < b$, and consider the difference equation*

$$x_{n+1} = g(x_n, x_{n-1}), \quad \text{for } n = 0, 1, \ldots . \tag{5}$$

Suppose that g satisfies the following two conditions:

1. *$g(x, y)$ is non-decreasing in $x \in [a, b]$ for each fixed $y \in [a, b]$, and $g(x, y)$ is non-increasing in $y \in [a, b]$ for each fixed $x \in [a, b]$;*
2. *if (m, M) is a solution of the system*

$$m = g(m, M), \quad M = g(M, m), \tag{6}$$

then $m = M$.

Then, there exists exactly one equilibrium of (5), namely $\bar{x}$. Furthermore, every solution of (5) converges to $\bar{x}$.

Theorem 4 ***(Drymonis and Ladas, [2]).*** *Let*

$$x_{n+1} = f(x_n, x_{n-1}), \quad \text{for } n = 0, 1, \ldots \tag{7}$$

with

1. *$f \in C[(0, \infty) \times (0, \infty), (0, \infty)]$;*
2. *$f(u, v)$ is non-increasing in u and v respectively;*
3. *$xf(x, x)$ is non-decreasing in x;*
4. *Equation (7) has a unique positive equilibrium $\bar{x}$.*

Then, every positive solution of (7) which is bounded from above and from below by positive constants converges to $\bar{x}$.

We follow the terminology given in [4] for stability of an equilibrium. Here we restate some of their results for convenience.

Let I be an interval of real numbers. Suppose that $f : I \times I \to I$, is a continuous function, which defines the difference equation

$$x_{n+1} = f(x_n, x_{n-1}), \quad \text{for } n = 0, 1 \ldots, \tag{8}$$

and let $\bar{x} \in I$ be an equilibrium of (8). Further, suppose $f(u, v)$ is continuously differentiable in some neighborhood of the equilibrium $\bar{x}$. We define the following constants

$$a_1 = -\frac{\partial f}{\partial u}(\bar{x}, \bar{x}), \tag{9}$$

$$a_0 = -\frac{\partial f}{\partial v}(\bar{x}, \bar{x}). \tag{10}$$

The equation

$$\lambda^2 + a_1\lambda + a_0 = 0 \tag{11}$$

is called the characteristic equation of (8).

Theorem 5 ***(Linearized Stability Theorem, [4]).*** *Let $\bar{x}$ be an equilibrium of (8), and suppose that f is a continuously differentiable function defined on some open neighborhood of $\bar{x}$. If the roots of (11) have absolute value less than one, then the equilibrium $\bar{x}$ is locally asymptotically stable.*

Theorem 6 ***(Theorem 1.3 of [4]).*** *Consider the second-degree polynomial equation*

$$\lambda^2 + a_1\lambda + a_0 = 0, \tag{12}$$

where a_0 and a_1 are real numbers.

A necessary and sufficient condition for the roots of (12) to lie within the unit disc $|\lambda| < 1$ is

$$|a_1| < 1 + a_0 < 2. \tag{13}$$

We will use the method of full limiting sequences, as developed by Karakostas, see [8, 9], and we use the following result.

Theorem 7 ***(Theorem 1.8 of [4]).*** *Consider the difference equation*

$$x_{n+1} = f(x_n, x_{n-1}, \ldots, x_{n-k}), \tag{14}$$

where $f \in C[J^{k+1}, J]$ for some interval J of real numbers and some nonnegative integer k. Let $\{x_n\}_{n=-k}^{\infty}$ be a solution of (14). Set $I = \liminf_{n\to\infty} x_n$ and $S = \limsup_{n\to\infty} x_n$, and suppose that $I, S \in J$. Let $\mathcal{L}_0$ be a limit point of the sequence $\{x_n\}_{n=-k}^{\infty}$. Then the following statements are true.

1. *There exists a solution $\{L_n\}_{n=-\infty}^{\infty}$ of (14), called a full limiting sequence of $\{x_n\}_{n=-k}^{\infty}$, such that $L_0 = \mathcal{L}_0$, and such that for every $N \in \mathbb{Z}$, L_N is a limit point of $\{x_n\}_{n=-k}^{\infty}$. In particular,*

$$I \le L_n \le S \text{ for all } N \in \mathbb{Z}. \tag{15}$$

2. *For every $i_0 \in \mathbb{Z}$, there exists a subsequence $\{x_{r_i}\}_{i=0}^{\infty}$ of the solution $\{x_n\}_{n=-k}^{\infty}$ such that*

$$L_N = \lim_{i\to\infty} x_{r_i+N} \text{ for every } N \geq i_0. \tag{16}$$

The following inequality will be useful in the sequel, and can be found as an exercise in [12].

$$\min\left\{\frac{\alpha_1}{B_1}, \frac{\alpha_2}{B_2}, \ldots, \frac{\alpha_n}{B_n}\right\} \leq \frac{\sum_{k=1}^{n}\alpha_k}{\sum_{k=1}^{n}B_k} \leq \max\left\{\frac{\alpha_1}{B_1}, \frac{\alpha_2}{B_2}, \ldots, \frac{\alpha_n}{B_n}\right\}, \tag{17}$$

where $\alpha_1, \ldots, \alpha_n$ are nonnegative real numbers and $B_1, B_2, \ldots, B_n$ are positive real numbers.

3 Several Auxiliary Results on Equation (1)

In this section we will prove several auxiliary results concerning (1). These results will be useful in proving Theorem 1. We begin, by showing that every positive solution of (1) is bounded.

Theorem 8 *Every positive solution of (1) is bounded from above and from below by positive constants.*

Proof For the sake of contradiction, suppose that $\{x_n\}_{n=0}^{\infty}$ is an unbounded solution of equation (1). Then there exists an increasing sub-sequence of $\{x_n\}_{n=0}^{\infty}$, which we will denote $\{x_{n_j}\}$ such that

$$\lim x_{n_j} = \infty.$$

By considering the recursive definition of x_{n_j}, this implies that the subsequences $\{x_{n_j-1}\}$ and $\{x_{n_j-2}\}$ both converge to zero. Further, if $\lim x_{n_j-1} = 0$, then $\lim x_{n_j-3} = \infty$. Likewise, if $\lim x_{n_j-2} = 0$, then $\lim x_{n_j-4} = \infty$.

This is a contradiction since on one hand, we see that $\lim x_{n_j-4} = \infty$, but on the other, $\lim x_{n_j-4} = 0$. Therefore, there exists some $U > 0$ such that $x_n \leq U$ for all $n \geq 1$. Now consider for $n \geq 1$,

$$x_{n+1} = \frac{\alpha + \beta x_n}{Bx_n + Dx_n x_{n-1} + x_{n-1}} \geq \frac{\alpha}{BU + DU^2 + U}.$$

We define $L = \frac{\alpha}{BU+DU^2+U}$, and thus we have

$$U \leq x_n \leq L, \qquad \text{for } n \geq 1.$$

We will use the following cubic polynomial, which will aid us in the proofs of several results. We define the cubic polynomial h, by

$$h(x) = Dx^3 + (B+1)x^2 - \beta x - \alpha. \tag{18}$$

Theorem 9 *Equation (1) has a unique positive equilibrium.*

Proof Consider the equation

$$h(x) = 0 \tag{19}$$

where the function h is defined in (18). By the well-known Decartes' Rule of Signs, the cubic equation (19) has a unique positive root.

We define $\bar{x}$ as the unique positive solution of (19). Thus,

$$D\bar{x}^3 + (B+1)\bar{x}^2 - \beta\bar{x} - \alpha = 0 \tag{20}$$

$$(B+1)\bar{x}^2 + D\bar{x}^3 = \alpha + \beta\bar{x} \tag{21}$$

$$\bar{x} = \frac{\alpha + \beta\bar{x}}{B\bar{x} + D\bar{x}^2 + \bar{x}} \tag{22}$$

Hence, $\bar{x}$ is the positive equilibrium of (1), and $\bar{x}$ is unique.

In the remainder of this section, we will provide several results concerning the local stability of (1). We show that the unique equilibrium is locally asymptotically stable. Before we state and prove Theorem 10, we will give some helpful lemmas, the proofs of which follow by direct computation.

Lemma 1 *Let* $f(u,v) = \dfrac{\alpha + \beta u}{Bu + Duv + v}$, *then*

$$\frac{\partial f}{\partial u} = \frac{\beta v - \alpha B - \alpha D v}{(Bu + Duv + v)^2} \quad \text{and} \quad \frac{\partial f}{\partial v} = -\frac{(\alpha + \beta u)(Du + 1)}{(Bu + Duv + v)^2}. \tag{23}$$

Lemma 2 *The characteristic equation of (1) reduces to*

$$\lambda^2 + \left(\frac{-\beta\bar{x} + \alpha B + \alpha D\bar{x}}{((B+1)\bar{x} + D\bar{x}^2)^2}\right)\lambda + \left(\frac{\bar{x}(D\bar{x}+1)}{(B+1)\bar{x} + D\bar{x}^2}\right) = 0. \tag{24}$$

Proof Let constants a_1 and a_0 be as defined in (9) and (10). Then,

$$a_1 = \frac{-\beta\bar{x} + \alpha B + \alpha D\bar{x}}{((B+1)\bar{x} + D\bar{x}^2)^2}. \tag{25}$$

By using the definition of $\bar{x}$,

$$a_0 = \frac{(\alpha + \beta\bar{x})(D\bar{x}+1)}{((B+1)\bar{x} + D\bar{x}^2)^2} = \frac{\bar{x}(D\bar{x}+1)}{(B+1)\bar{x} + D\bar{x}^2}. \tag{26}$$

Theorem 10 *Let* $\bar{x}$ *be the unique positive equilibrium of (1), then,* $\bar{x}$ *is* locally asymptotically stable.

Proof According to Theorems 5 and 6, we need to show

$$\left|\frac{-\beta\bar{x}+\alpha B+\alpha D\bar{x}}{((B+1)\bar{x}+D\bar{x}^2)^2}\right| < 1+\frac{\bar{x}(D\bar{x}+1)}{(B+1)\bar{x}+D\bar{x}^2} < 2. \tag{27}$$

We will start by proving the right side of inequality (27) first. Consider

$$1+\frac{\bar{x}(D\bar{x}+1)}{(B+1)\bar{x}+D\bar{x}^2} = 1+\frac{D\bar{x}^2+\bar{x}}{B\bar{x}+\bar{x}+D\bar{x}^2} < 1+1=2.$$

Now we will prove the left side of (27). We will start by showing

$$\frac{-\beta\bar{x}+\alpha B+\alpha D\bar{x}}{((B+1)\bar{x}+D\bar{x}^2)^2} < 1+\frac{\bar{x}(D\bar{x}+1)}{(B+1)\bar{x}+D\bar{x}^2}.$$

Suppose that it is indeed the case, then,

$$\begin{aligned} -\beta\bar{x}+\alpha B+\alpha D\bar{x} &< ((B+1)\bar{x}+D\bar{x}^2)^2+((B+1)\bar{x}+D\bar{x}^2)\bar{x}(D\bar{x}+1)\\ \alpha\,(B+D\bar{x})-\beta\bar{x} &< ((B+1)\bar{x}+D\bar{x}^2)^2+((B+1)\bar{x}+D\bar{x}^2)\bar{x}(D\bar{x}+1). \end{aligned}$$

Now, we use the expression for α that is obtained from (19) to simplify.

$$\begin{aligned} \left((B+1)\bar{x}^2+D\bar{x}^3-\beta\bar{x}\right)(B+D\bar{x})-\beta\bar{x} &< (B+1)^2\,\bar{x}^2+2\,(B+1)\,D\bar{x}^3+D^2\bar{x}^4+\\ &\quad+\left((B+1)\bar{x}^2+D\bar{x}^3\right)(D\bar{x}+1)\\ \left(B\bar{x}^2+\bar{x}^2+D\bar{x}^3-\beta\bar{x}\right)(B+D\bar{x})-\beta\bar{x} &< \left(B^2+2B+1\right)\bar{x}^2+2BD\bar{x}^3+2D\bar{x}^3+\\ &\quad+D^2\bar{x}^4+\left(B\bar{x}^2+\bar{x}^2+D\bar{x}^3\right)(D\bar{x}+1) \end{aligned}$$

Now, by canceling out like terms on both sides of this inequality, we see that the left side is negative, while the right side is positive, and hence always true.

Now we will show that

$$-\left(1+\frac{\bar{x}(D\bar{x}+1)}{(B+1)\bar{x}+D\bar{x}^2}\right) < \frac{-\beta\bar{x}+\alpha B+\alpha D\bar{x}}{((B+1)\bar{x}+D\bar{x}^2)^2}.$$

Suppose that it is indeed true, then

$$-\left(((B+1)\bar{x}+D\bar{x}^2)^2+((B+1)\bar{x}+D\bar{x}^2)\bar{x}\,(D\bar{x}+1)\right) < -\beta\bar{x}+\alpha B+\alpha D\bar{x}.$$

By rearranging terms, we obtain,

$$\beta\bar{x} < \alpha B+\alpha D\bar{x}+((B+1)\bar{x}+D\bar{x}^2)^2+\left((B+1)\bar{x}^2+D\bar{x}^3\right)D\bar{x}+(B+1)\bar{x}^2+D\bar{x}^3.$$

We use the fact that $\beta\bar{x}=D\bar{x}^3+(B+1)\bar{x}^2-\alpha$, obtained from (19), and cancel out like terms to see that

$$-\alpha < \alpha B + \alpha D\bar{x} + \left((B+1)\bar{x} + D\bar{x}^2 + D\bar{x}^2\right)^2 + \left((B+1)\bar{x}^2 + D\bar{x}^3\right) D\bar{x}.$$

Thus, the left side is negative while the right side is positive. Hence, we have shown that the conditions of Theorem 5 are satisfied, and the unique positive equilibrium is locally asymptotically stable.

4 Existence of an Invariant Interval

In this section we will show that all solutions of (1) will eventually enter an invariant interval. We begin by studying the following quadratic function in variable β,

$$R(\beta) = (\beta - \alpha D)^2 - \alpha B^2. \tag{28}$$

It is clear that the roots of (28) are

$$\beta^- = -B\sqrt{\alpha} + \alpha D \quad \text{and} \quad \beta^+ = B\sqrt{\alpha} + \alpha D.$$

Furthermore, $R(\beta) < 0$ on the interval (β^-, β^+), and $R(\beta) > 0$ on $(-\infty, \beta^-) \cup (\beta^+, \infty)$. We will use the following technical lemma in the proofs of the lemmas to follow.

Lemma 3 *The following statements are true.*

1. *If* $\alpha D < \beta < \beta^+$ *then*

$$\max\left\{\frac{\alpha B}{\beta - \alpha D}, \frac{\beta B}{B^2 + D(\beta - \alpha D)}\right\} = \frac{\alpha B}{\beta - \alpha D}.$$

2. *If* $\beta > \beta^+$ *then*

$$\min\left\{\frac{\alpha B}{\beta - \alpha D}, \frac{\beta B}{B^2 + D(\beta - \alpha D)}\right\} = \frac{\alpha B}{\beta - \alpha D}.$$

Proof We prove part (1), the proof of part (2) is similar and will be omitted. We want to show that $\frac{\alpha B}{\beta - \alpha D} \geq \frac{\beta B}{B^2 + D(\beta - \alpha D)}$. Consider

$$\begin{aligned} \frac{\alpha B}{\beta - \alpha D} &\geq \frac{\beta B}{B^2 + D(\beta - \alpha D)} \\ \alpha B^2 + \alpha D(\beta - \alpha D) &\geq \beta(\beta - \alpha D) \\ 0 &\geq (\beta - \alpha D)^2 - \alpha B^2 \end{aligned}$$

which is true from the shape of the quadratic function (28).

Lemma 4 *Let $\alpha D < \beta < \beta^+$, then the following is true*

1. $\frac{\alpha B}{\beta-\alpha D} > \frac{\beta-\alpha D}{B}$
2. *If* $x_{n-1} \leq \frac{\alpha B}{\beta-\alpha D}$, *then* $x_{n+1} \geq \frac{\beta-\alpha D}{B}$
3. *If* $x_{n-1} \geq \frac{\beta-\alpha D}{B}$, *then* $x_{n+1} \leq \frac{\alpha B}{\beta-\alpha D}$

Proof The proof of statement (1) follows from the fact that the quadratic function defined in (28) is negative for $\alpha D < \beta < \beta^+$.

To prove (2), let's assume that $x_{n-1} \leq \frac{\alpha B}{\beta-\alpha D}$ and using that the function $f(u, v)$ is decreasing in v, we get

$$\begin{aligned} x_{n+1} &= \frac{\alpha+\beta x_n}{Bx_n + Dx_n x_{n-1} + x_{n-1}} \geq \frac{\alpha+\beta x_n}{Bx_n + Dx_n\left(\frac{\alpha B}{\beta-\alpha D}\right) + \left(\frac{\alpha B}{\beta-\alpha D}\right)} \\ &= \frac{\alpha+\beta x_n}{\left(\frac{\alpha B}{\beta-\alpha D}\right) + \left(B + \frac{D\alpha B}{\beta-\alpha D}\right) x_n} \geq \min\left\{\frac{\alpha}{\frac{\alpha B}{\beta-\alpha D}}, \frac{\beta}{\frac{B\beta}{\beta-\alpha D}}\right\} \\ &= \min\left\{\frac{\beta-\alpha D}{B}, \frac{\beta-\alpha D}{B}\right\} = \frac{\beta-\alpha D}{B} \end{aligned}$$

Now we will prove statement (3). Assume $x_{n-1} \geq \frac{\beta-\alpha D}{B}$, and by using the fact that the function $f(u, v)$ is decreasing in v along with Lemma 3 Part (1), we obtain

$$\begin{aligned} x_{n+1} &= \frac{\alpha+\beta x_n}{Bx_n + Dx_n x_{n-1} + x_n - 1} \leq \frac{\alpha+\beta x_n}{Bx_n + Dx_n\left(\frac{\beta-\alpha D}{B}\right) + \left(\frac{\beta-\alpha D}{B}\right)} \\ &= \frac{\alpha+\beta x_n}{\left(\frac{\beta-\alpha D}{B}\right) + \left(B + \frac{D(\beta-\alpha D)}{B}\right) x_n} \leq \max\left\{\frac{\alpha B}{\beta-\alpha D}, \frac{\beta B}{B^2 + D(\beta-\alpha D)}\right\} \\ &= \frac{\alpha B}{\beta-\alpha D}. \end{aligned}$$

Lemma 5 *Let $\beta > \beta^+$, then the following is true*

1. $\frac{\alpha B}{\beta-\alpha D} < \frac{\beta-\alpha D}{B}$
2. *If* $x_{n-1} \geq \frac{\alpha B}{\beta-\alpha D}$, *then* $x_{n+1} \leq \frac{\beta-\alpha D}{B}$
3. *If* $x_{n-1} \leq \frac{\beta-\alpha D}{B}$, *then* $x_{n+1} \geq \frac{\alpha B}{\beta-\alpha D}$

Proof The proof of statement (1) follows from the fact that the quadratic function defined in (28) is positive for $\beta > \beta^+$.

To prove (2), we assume that $x_{n-1} \geq \frac{\alpha B}{\beta-\alpha D}$ and using that the function $f(u, v)$ is decreasing in v, we get

$$x_{n+1} = \frac{\alpha + \beta x_n}{Bx_n + Dx_n x_{n-1} + x_{n-1}} \leq \frac{\alpha + \beta x_n}{Bx_n + Dx_n\left(\frac{\alpha B}{\beta - \alpha D}\right) + \left(\frac{\alpha B}{\beta - \alpha D}\right)}$$
$$= \frac{\alpha + \beta x_n}{\left(\frac{\alpha B}{\beta - \alpha D}\right) + \left(B + \frac{D\alpha B}{\beta - \alpha D}\right) x_n} \leq \max\left\{\frac{\alpha}{\frac{\alpha B}{\beta - \alpha D}}, \frac{\beta}{\frac{B\beta - B\alpha D + D\alpha B}{\beta - \alpha D}}\right\}$$
$$= \min\left\{\frac{\beta - \alpha D}{B}, \frac{\beta - \alpha D}{B}\right\} = \frac{\beta - \alpha D}{B}$$

Now let's prove (3). Assume $x_{n-1} \leq \frac{\beta - \alpha D}{B}$ and using that the function $f(u, v)$ is decreasing in v along with Lemma 3 Part (2), we get

$$x_{n+1} = \frac{\alpha + \beta x_n}{Bx_n + Dx_n x_{n-1} + x_n - 1} \geq \frac{\alpha + \beta x_n}{Bx_n + Dx_n\left(\frac{\beta - \alpha D}{B}\right) + \left(\frac{\beta - \alpha D}{B}\right)}$$
$$= \frac{\alpha + \beta x_n}{\left(\frac{\beta - \alpha D}{B}\right) + \left(B + \frac{D(\beta - \alpha D)}{B}\right) x_n} \geq \min\left\{\frac{\alpha B}{\beta - \alpha D}, \frac{\beta B}{B^2 + D(\beta - \alpha D)}\right\}$$
$$= \frac{\alpha B}{\beta - \alpha D}.$$

Lemma 6 *When $\beta = \beta^+ = \alpha D + B\sqrt{\alpha}$, the unique equilibrium of equation is $\bar{x} = \frac{\alpha B}{\beta - \alpha D} = \frac{\beta - \alpha D}{B}$.*

Proof It is clear that $\frac{\alpha B}{\beta - \alpha D} = \frac{\beta - \alpha D}{B}$ when $\beta = \beta^+ = \alpha D + B\sqrt{\alpha}$ from Eq. (28). Now we show that $\bar{x} = \frac{\alpha B}{\beta - \alpha D}$. We will evaluate $h\left(\frac{\alpha B}{\beta - \alpha D}\right)$, where h was the cubic equation defined in (18). Clearly,

$$h\left(\frac{\alpha B}{\beta - \alpha D}\right) = D\left(\frac{\alpha B}{\beta - \alpha D}\right)^3 + (B + 1)\left(\frac{\alpha B}{\beta - \alpha D}\right)^2 - \beta\left(\frac{\alpha B}{\beta - \alpha D}\right) - \alpha.$$

Simplifying this, we see that,

$$h\left(\frac{\alpha B}{\beta - \alpha D}\right) = D\left(\frac{\alpha B}{\beta - \alpha D}\right)^3 + (B + 1)\left(\frac{\alpha B}{\beta - \alpha D}\right)^2 - \beta\left(\frac{\alpha B}{\beta - \alpha D}\right) - \alpha$$
$$= \frac{D\alpha^3 B^3 + (B + 1)\alpha^2 B^2(\beta - \alpha D) - \beta\alpha B\,(\beta - \alpha D)^2 - \alpha\,(\beta - \alpha D)^3}{(\beta - \alpha D)^3}$$
$$= \frac{D\alpha^2 B(\alpha B^2) + (B + 1)\alpha(\alpha B^2)(\beta - \alpha D) - \beta\alpha B\,(\beta - \alpha D)^2 - \alpha\,(\beta - \alpha D)^3}{(\beta - \alpha D)^3}.$$

Since $\frac{\alpha B}{\beta - \alpha D} = \frac{\beta - \alpha D}{B}$ implies that $\alpha B^2 = (\beta - \alpha D)^2$, we get

$$h\left(\frac{\alpha B}{\beta-\alpha D}\right) = \frac{D\alpha^2 B\,(\beta-\alpha D)^2 + (B+1)\alpha\,(\beta-\alpha D)^3 - \beta\alpha B\,(\beta-\alpha D)^2 - \alpha\,(\beta-\alpha D)^3}{(\beta-\alpha D)^3}$$
$$= \frac{\left(D\alpha^2 B + (B\alpha+\alpha)(\beta-\alpha D) - \beta\alpha B - \alpha\beta + \alpha^2 D\right)(\beta-\alpha D)^2}{(\beta-\alpha D)^3}$$
$$= 0.$$

This shows that $\bar{x} = \frac{\alpha B}{\beta-\alpha D}$ when $\beta = \beta^+$.

Let's define the interval K,

$$K = \begin{cases} \left[\frac{\beta-\alpha D}{B}, \frac{\alpha B}{\beta-\alpha D}\right], & \text{if} \quad \alpha D < \beta \le \beta^+ \\ \\ \left[\frac{\alpha B}{\beta-\alpha D}, \frac{\beta-\alpha D}{B}\right], & \text{if} \quad \beta > \beta^+, \end{cases} \tag{29}$$

with the convention that when $\beta = \beta^+$, the "interval" K is a single point and $\bar{x} = K$, as was previously established in Lemma 4.

The next lemma establishes that the interval K is invariant, in the sense that if two consecutive terms of the solution are in K then the solution will remain in K for ever. The proof follows from Lemmas 4 and 5, and will be omitted.

Lemma 7 *Suppose that $\alpha D < \beta$ and there exists $N \in \mathbb{Z}^+$ such that $x_N, x_{N-1} \in K$, then $x_n \in K$ for all $n \ge N$.*

Amazingly, we will now show that K is also attracting. That is, solutions will always eventually enter K, and we state it formally in the following lemma.

Lemma 8 *If $\alpha D < \beta$ then K is an attracting interval. In other words, there exists $N \in \mathbb{Z}^+$ such that $x_n \in K$ for all $n \ge N$.*

Proof We will give the proof for $\alpha D < \beta \le \beta^+$. The proof for the other case follows similarly and will be omitted.

Let $I = \liminf_{n\to\infty} x_n$ and $S = \limsup_{n\to\infty} x_n$. Then, if both $I \in K$ and $S \in K$, then we are done. For the sake of contradiction, assume that $S \notin K$. It follows from Lemma 4 that $S > \frac{\alpha B}{\beta-\alpha D}$. Thus, there is an open neighborhood O containing S such that $O \cap K = \emptyset$. By Theorem 7, let S_{n+1} be a full-limiting sequence such that $\lim_{n\to\infty} S_{n+1} = S$. Thus, there exists a positive integer N, such that $S_n \in O$ for $n \ge N$. According to Lemma 4, if $S_n > \frac{\alpha B}{\beta-\alpha D} \ge \frac{\beta-\alpha D}{B}$, then $S_{n+1} < \frac{\alpha B}{\beta-\alpha D}$, which is a contradiction. Thus, it must be the case that S is in the interval K. The other case, when $I \notin K$ is proved the same way. Thus, it must be the case that both I and S are in the interval K, which completes the proof.

5 Global Attractivity of $\bar{x}$

Our proof of the main result, Theorem 1, will be based on cases which partition the values of the coefficient β.

Fig. 1 $\beta \leq \beta^+$

5.1 Case $\beta \leq \beta^+$

Suppose that $\beta \leq \beta^+$. Theorems 11 and 12 will show that the unique equilibrium of (1) is a global attractor (Fig. 1).

Theorem 11 *If $\beta = \frac{\alpha D}{B+1}$, then the unique positive equilibrium of (1), $\bar{x}$, is a global attractor.*

Proof We will apply Theorem 4, for which it only remains to show that $xf(x, x)$ is non-decreasing in x. Consider

$$xf(x,x) = \frac{x(\alpha + \beta x)}{Bx + Dx^2 + x} = \frac{\alpha + \beta x}{(B+1) + Dx}.$$

Then, we see that

$$\frac{d}{dx} xf(x,x) = \frac{\beta((B+1) + Dx) - (\alpha + \beta x) D}{((B+1) + Dx)^2} = 0.$$

Now all the conditions of the Theorem 4 are satisfied and so every solution converges to $\bar{x}$.

Theorem 12 *If $\beta \leq \beta^+$ and $\beta \neq \frac{\alpha D}{B+1}$, then the unique positive equilibrium of (1), $\bar{x}$, is a global attractor.*

Proof Suppose that $\beta \neq \frac{\alpha D}{B+1}$, and $\beta \leq \beta^+$. Then, consider $\frac{\partial f}{\partial u}$. From Lemma 1, and if $\beta \leq \alpha D$, we obtain

$$\frac{\partial f}{\partial u} = \frac{\beta v - \alpha B - \alpha D v}{Bu + Duv + v)^2} = \frac{(\beta - \alpha D)v - \alpha B}{(Bu + Duv + v)^2} \leq 0. \tag{30}$$

If $\alpha D < \beta < \beta^+ = \alpha D + B\sqrt{\alpha}$ then by Lemmas 7 and 8, and by the definition of interval K in (29), every solution $\{x_n\}$ will eventually enter the attracting interval $K = \left[\frac{\beta - \alpha D}{B}, \frac{\alpha B}{\beta - \alpha D}\right]$ and remain there. Hence, by Lemma 8, there exists a positive integer N such that for all integers $n \geq N$, $x_n \leq \frac{\alpha B}{\beta - \alpha D}$. Therefore, we again see that for $n \geq N$, $\frac{\partial f}{\partial u} \leq 0$ (Fig. 2).

Therefore, when $\beta \leq \beta^+$ and $n \geq N$, the function $f(u, v)$ is non-increasing in both u and v. We will use the "M&m" Theorem, Theorem 2. We define m, M as follows,

$$m = \frac{\alpha + \beta M}{(B+1)M + DM^2} \quad \text{and} \quad M = \frac{\alpha + \beta m}{(B+1)m + Dm^2}.$$

We clear the denominators to obtain

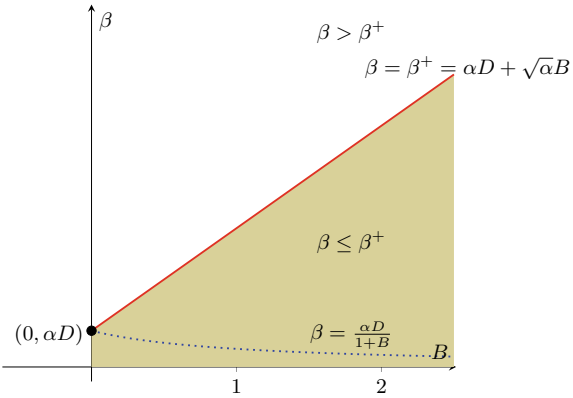

Fig. 2 When $\beta \leq \beta^+$ the unique equilibrium is a global attractor

$$(B+1)mM + DmM^2 = \alpha + \beta M \tag{31}$$

$$(B+1)Mm + DMm^2 = \alpha + \beta m. \tag{32}$$

We subtract (31) and (32) to get

$$DmM(M-m) = \beta(M-m).$$

We have a solution $M = m$. Assume $M \neq m$, then we get that $DmM = \beta$ which implies that $m = \frac{\beta}{DM}$ and we substitute this expression into (31) to obtain,

$$\begin{aligned}(B+1)\left(\frac{\beta}{DM}\right)M + D\left(\frac{\beta}{DM}\right)M^2 &= \alpha + \beta M\\ \frac{(B+1)\beta}{D} + \beta M &= \alpha + \beta M\\ (B+1)\,\beta &= \alpha D.\end{aligned}$$

Which, we assumed is not the case. Hence, we have proved convergence to $\bar{x}$ when $\beta \leq \alpha D$ and $(B+1)\,\beta \neq \alpha D$ or $\beta > \beta^+ = \alpha D + B\sqrt{\alpha}$.

Corollary 1 *If $\beta \leq \beta^+$ then the unique positive equilibrium of (1), $\bar{x}$ is a global attractor.*

5.2 Case $\beta > \beta^+$

Suppose that $\beta > \beta^+$. By Lemma 5, and the definition of the interval K in (29) we know that the interval $K = \left[\frac{\alpha B}{\beta - \alpha D}, \frac{\beta - \alpha D}{B}\right]$ is attracting and invariant, and hence,

Fig. 3 $\beta > \beta^+$

there exists a positive integer N such that for all integers $n \geq N$, $x_n \in K$. Hence, $f(u, v)$ is non-decreasing in u and non-increasing in v (Fig. 3).

We proceed by Theorem 3, setting up the system of equations

$$\begin{cases} m = \frac{\alpha+\beta m}{Bm+DmM+M} \\ M = \frac{\alpha+\beta M}{BM+DmM+m} \end{cases}. \tag{33}$$

By, clearing the denominators in (33), and subtracting the second from the first, we obtain

$$B(m - M)(m + M) + DmM(m - M) = \beta(m - M). \tag{34}$$

Clearly $m = M$ is a solution. Suppose that $m \neq M$ and divide both sides of (34) by $m - M$ to obtain,

$$B(m + M) + DmM = \beta. \tag{35}$$

We can see that by the symmetry, any solution to (33) for m must also be a solution for M. We solve for m in (35) to obtain,

$$m = \frac{\beta - BM}{B + DM}, \tag{36}$$

and substitute this into the second equation in system (33).

$$BM^2 + DmM^2 + mM = \alpha + \beta M \tag{37}$$

$$BM^2 + D\left(\frac{\beta - BM}{B + DM}\right)M^2 + \left(\frac{\beta - BM}{B + DM}\right)M = \alpha + \beta M \tag{38}$$

$$BM^2(B + DM) + DM^2(\beta - BM) + M(\beta - BM) = (\alpha + \beta M)(B + DM) \tag{39}$$

$$B^2M^2 + BDM^3 + D\beta M^2 - BDM^3 + \beta M - BM^2 = \alpha B + \alpha DM + \beta BM + \beta DM^2 \tag{40}$$

$$B^2M^2 + \beta M - BM^2 = \alpha B + \alpha DM + \beta BM \tag{41}$$

$$B(B - 1)M^2 + (\beta - \alpha D - \beta B)M - \alpha B = 0. \tag{42}$$

We will break this case up into three subcases, depending on whether $B = 1$, $B > 1$ or $B < 1$. For the cases when $B = 1$ or $B > 1$ we continue by using Theorem 3.

5.2.1 Case $B = 1$

If $B = 1$ then (42) reduces to

$$-\alpha DM - \alpha B = 0$$

which has no positive solutions. Hence $m = M$ was the unique solution to (33), and by Theorem (3) we would conclude that the unique positive equilibrium is a global attractor.

5.2.2 Case $B > 1$

If $B > 1$ then $B(B-1) > 0$ and $\beta(1-B) - \alpha D < 0$, then we see that M_1, M_2 are solutions to (42), where

$$M_1 = M_- = \frac{\beta(B-1) + \alpha D - \sqrt{(\beta(B-1) + \alpha D)^2 + 4\alpha B^2(B-1)}}{2B(B-1)} < 0,$$
$$M_2 = M_+ = \frac{\beta(B-1) + \alpha D + \sqrt{(\beta(B-1) + \alpha D)^2 + 4\alpha B^2(B-1)}}{2B(B-1)} > 0.$$

By (36), we find that

$$m_1 = \frac{\beta - BM_1}{B + DM_1} = \frac{\beta(B-1) + \alpha D + \sqrt{(\beta(B-1) + \alpha D)^2 + 4\alpha B^2(B-1)}}{2B(B-1)} = M_+,$$
$$m_2 = \frac{\beta - BM_2}{B + DM_2} = \frac{\beta(B-1) + \alpha D - \sqrt{(\beta(B-1) + \alpha D)^2 + 4\alpha B^2(B-1)}}{2B(B-1)} = M_-.$$

This gives us the following symmetric solutions,

$$(M_1, m_1) = (M_-, M_+),$$
$$(M_2, m_2) = (M_+, M_-),$$

where one of the components is negative in each solution. Hence, the unique solution of (33) is $M = m$, and so by Theorem 3 we have shown that the $\bar{x}$ is a global attractor.

5.2.3 Case $B < 1$

So, what remains to be shown, is global convergence for the case when $\beta > \beta+$ and $B < 1$ (Fig. 4).

When $B < 1$ we know that $B(B-1) < 0$. Looking at the coefficient of M in (42), we see that if $\beta - \alpha D - \beta B \leq 0$ then (42) will have no solutions, since all

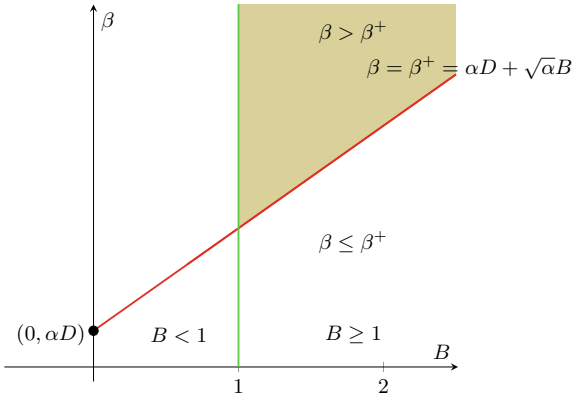

Fig. 4 When $\beta > \beta^+$ and $B \geq 1$ the unique equilibrium is a global attractor

coefficients in (42) will be negative. Hence,

$$\beta - \alpha D - \beta B \leq 0 \tag{43}$$

$$\beta(1 - B) \leq \alpha D \tag{44}$$

$$\beta \leq \frac{\alpha D}{1 - B}. \tag{45}$$

So for $\beta \leq \frac{\alpha D}{1-B}$ we have satisfied the requirements of Theorem 3, and every solution will converge to the unique positive equilibrium $\bar{x}$.

The global behavior of solutions is still open in the region for $\beta > \beta^+$, $\beta > \frac{\alpha D}{1-B}$ and $B < 1$.

The following theorem follows from the justification given above:

Theorem 13 *If $\beta > \beta^+$ and if $B \geq 1$, or if $B < 1$ and $\beta \leq \frac{\alpha D}{1-B}$, then the unique positive equilibrium of (1) is a global attractor.*

6 Conclusion

It has been shown that Eq. (1) has a unique positive equilibrium which is locally asymptotically stable and a global attractor when the values of the parameters satisfy the conditions of Theorem 1, and for all positive initial conditions x_0, x_{-1}. Hence, the equilibrium of (1) is globally asymptotically stable under the conditions of Theorem 1 (Fig. 5).

These authors believe that alternative techniques must be used to investigate the remaining region, namely when

$$\beta > \alpha D + \sqrt{\alpha} B, \quad \beta > \frac{\alpha D}{1 - B}, \quad \text{and} \quad B < 1.$$

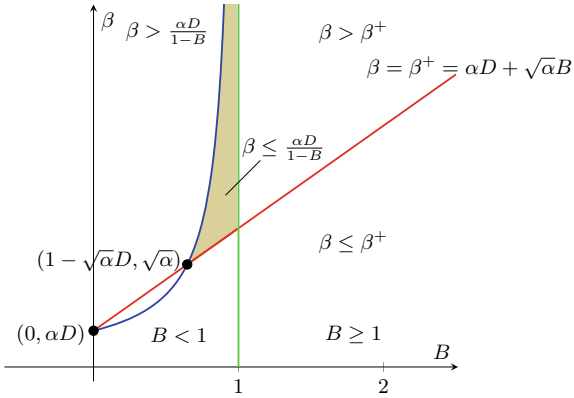

Fig. 5 When $\beta > \beta^+$, $B < 1$ and $\beta \leq \frac{\alpha D}{1-B}$ the unique equilibrium is a global attractor

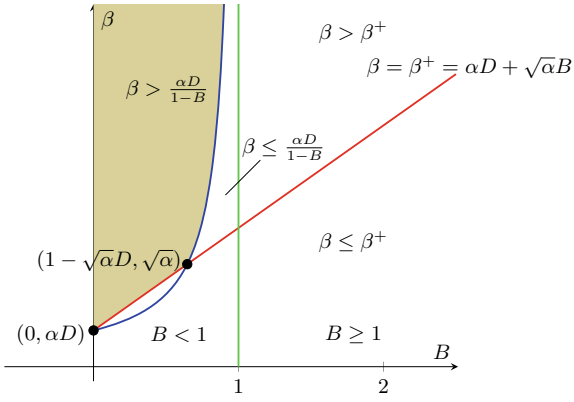

Fig. 6 Region for B and β for which global attractivity remains open

Theorem 3 does not apply with the given invariant interval of K defined in (29) because it can be shown that there are other solutions to the system in Theorem 3 for which $m \neq M$. Hence, one must find a smaller invariant interval if one is to apply this theorem, or use other techniques such as semi-cycle analysis, Lyapunov functions, etc. (Fig. 6).

Conjecture 1 The unique positive equilibrium of (1) is a global asymptotically stable for all positive values of the parameters.

We conclude the paper with the following open question, which would extend the results to a non-autonomous equation.

Question 1 Determine the behavior of solutions of (1) when the coefficients are periodic, or more generally positive sequences of real numbers bounded from above and from below by positive constants. In particular, determine if all positive solutions are bounded.

Acknowledgements The authors wish to thank the anonymous referee for his or her helpful comments for revising this paper.

References

1. DiPippo, M., Janowski, E., Kulenović, M.R.S.: Global asymptotic stability for quadratic fractional difference equation. Adv. Differ. Equ. 2015(179):13 (2015). https://doi.org/10.1186/s13662-015-0525-4. http://dx.doi.org/10.1186/s13662-015-0525-4
2. Drymonis, E., Ladas, G.: On the global character of the rational system $\frac{\alpha_1}{A_1+B_1x_n+y_n}$ and $\frac{\alpha_2+\beta_2x_n}{A_2+B_2x_n+C_2y_n}$. Sarajevo J. Math. **8**(21), 293–309 (2012). https://doi.org/10.5644/SJM.08.2.10
3. Garić-Demirović, M., Kulenović, M.R.S., Nurkanović, M.: Basins of attraction of certain homogeneous second order quadratic fractional difference equation. J. Concr. Appl. Math. **13**(1–2), 35–50 (2015)
4. Grove, E.A., Ladas, G.: Periodicities in nonlinear difference equations. In: Advances in Discrete Mathematics and Applications, vol. 4. Chapman & Hall/CRC, Boca Raton, FL (2005)
5. Kalabušić, S., Kulenović, M.R.S., Mehuljić, M.: Global period-doubling bifurcation of quadratic fractional second order difference equation. Discrete Dyn. Nat. Soc. Art. ID 920,410:13 (2014). https://doi.org/10.1155/2014/920410. http://dx.doi.org/10.1155/2014/920410
6. Kalabušić, S., Nurkanović, M., Nurkanović, Z.: Global dynamics of certain mix monotone difference equation. Mathematics **6**(1) (2018). https://doi.org/10.3390/math6010010. https://www.mdpi.com/2227-7390/6/1/10
7. Kalabušić, S., Kulenović, M.R.S., Mehuljić, M.: Global dynamics and bifurcations of two quadratic fractional second order difference equations. J. Comput. Anal. Appl. **21**(1), 132–143 (2016)
8. Karakostas, G.: Convergence of a difference equation via the full limiting sequences method. Differ. Equ. Dyn. Syst. **1**, 289–294 (1993)
9. Karakostas, G.: Asymptotic 2-periodic difference equations with diagonally self-invertible responses. J. Differ. Equ. Appl. **6**, 329–335 (2000)
10. Kostrov, Y., Kudlak, Z.: On a second-order rational difference equation with a quadratic term. Int. J. Differ. Equ. **11**(2), 179–202 (2016)
11. Kulenović, M.R.S., Pilav, E., Silić, E.: Local dynamics and global attractivity of a certain second-order quadratic fractional difference equation. Adv. Differ. Equ. **2014**(68) (2014)
12. Mitrinović, D.: Elementary Inequalities. P. Noordhoff LTD (1964)

Population Motivated Discrete-Time Disease Models

YE LI and Jiawei Xu

Abstract Infectious diseases are now widely analyzed by compartmental models. This paper introduces a SIR model coupled with a social mobility model (SMM). After discretization by a forward Euler Method, and a mixed type Euler method (structured with both forward and backward Euler elements), we obtained a difference equations model for our social mobility model. We calculate the basic reproduction number R_0 using the next-generation matrix method. When $R_0 < 1$, there will be a disease-free equilibrium (DFE), and $R_0 < 1$ implies DFE will be locally asymptotically stable, while $R_0 > 1$ implies DFE is unstable. When $R_0 = 1$, DFE may stable or unstable. Then we obtain a hyperbolic forward Kolmogorov equation corresponding to the SIR epidemic model. We also generate the hyperbolic forward Kolmogorov equations for the SIR model with SMM between 2 locations.

Keywords SIR epidemic model · Discrete-time model · Social mobility model · Forward Kolmogorov equation

1 Introduction

Infectious diseases are now widely analyzed by compartmental models, such as SEIR, SIR, SI, etc. [1–3]. Sattenspiel and Dietz considered a structured epidemic model by incorporating geographic mobility among regions [4]. Skufca and ben-Avraham considered a situation that accounts for the different dynamics arising from individuals on short trips and returning to home locations, based on a Gravity Model [5] and using the SMM (social mobility model) [6]. All of the above studies are formulated as a Markov Process. In Sect. 2, we review the SMM model in the form of

Y. LI (✉)
Department of Mathematics and Statistics, Texas Tech University, Lubbock, TX, USA
e-mail: ye.li@ttu.edu

J. Xu
Center of Health and Bioinformatics, Newcastle University, Newcastle, UK
e-mail: jxulincoln@gmail.com

S. Baigent et al. (eds.), *Progress on Difference Equations and Discrete Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 341,
https://doi.org/10.1007/978-3-030-60107-2_15

the continuous-time model. In Sect. 3, we formulate the SMM model in the form of the discrete-time model by the forward Euler method. In Sect. 4, we also define the Disease-free equilibrium (DFE) and R_0 (basic reproductive number) and analyze the forward Euler method discrete-time model. In Sect. 5, we give a PDE (Kolmogorov Equation) view of the SIR epidemic model and we construct the Kolmogorov Equations of the SIR epidemic model under population movement between 2 locations. Section 6 is the Numerical Simulation part, which shows the computed results on a simulative example based on the two discrete-time models. Finally, we discuss two discrete-time models and the difference equation SIR epidemic models under the Kolmogorov Equations representation in Sect. 7.

2 The Continuous-Time Model

2.1 Basic SIR Model Integrated with the SMM

According to [6], the SMM allows two types of motions: relocation and short trips. Let

$$(i, j) = (current\ location, home\ location)$$

Define the time scales $T \gg \theta \gg \tau, T \sim years, \theta \sim months, \tau \sim days$ [6]. The motion of people falls into two broad categories: (1) movement to relocate from one home to another, and (2) motion related to taking a trip with planned returns. Definition of forms in the SMM model is included in Table 1 and the transition paths of the model is shown in Fig. 1.

Table 1 Meaning of coefficients

τ	Time between travelling
θ	Time between trips
T	Rime between relocation
ω_{ij}	Likelihood relocate from i to j
ν_{ij}	Preference to travel from i to j
r	Probability

From the definition above, Skufca and ben-Avraham studied the following population motion process by using the parameters of Table 1 and the rates of Fig. 2 [6]. The resultant master equations are given by

$$\dot{\pi}_{ii} = \frac{r}{\tau}\sum_{j\neq i}\pi_{ji} - (\frac{1}{\theta}\sum_{j}\nu_{ij} + \frac{1}{T}\sum_{j}\omega_{ij})\pi_{ii} + \underbrace{\frac{1}{T}\sum_{j}\omega_{ji}\pi_{ij}}_{optional} \tag{1}$$

$$\dot{\pi}_{ij} = \frac{1}{\theta}\nu_{ji}\pi_{jj} + \frac{1}{T}\omega_{ij}\pi_{ii} + \frac{1-r}{\tau}\sum_{k\neq j}\nu_{ki}\pi_{kj} - (\frac{r}{\nu} + \frac{1-r}{\tau}\sum_{k\neq j}\nu_{ik})\pi_{ij} + \underbrace{\frac{1}{T}\sum_{k}(\omega_{kj}\pi_{ik} - \omega_{jk}\pi_{ij})}_{optional} \tag{2}$$

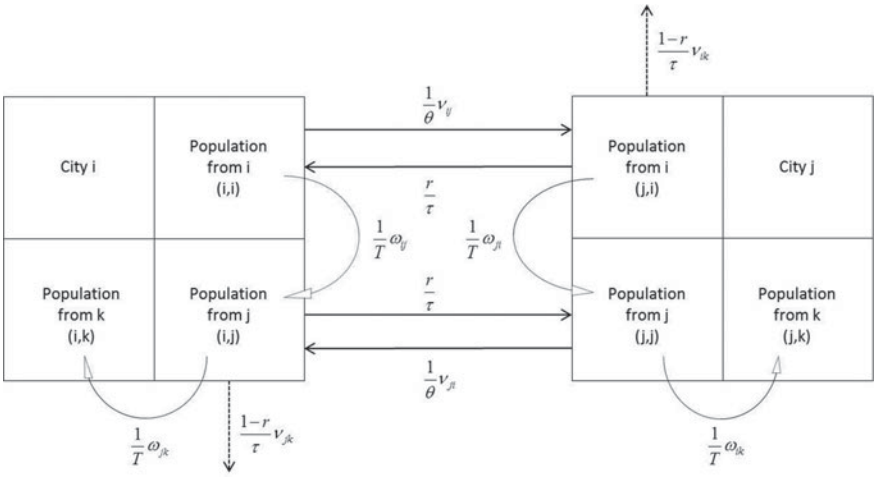

Fig. 1 Travel path of the model

Fig. 2 Transition rates for the SMM. In all cases, $i \neq j \neq k$

Transition	Rate	Description
	when at home	
$(i,i) \to (j,i)$	$\frac{1}{\theta}\nu_{ij}$	travel
$(i,i) \to (i,j)$	$\frac{1}{T}\omega_{ij}$	relocate
	when away from home	
$(i,j) \to (j,j)$	$\frac{r}{\tau}$	return
$(i,j) \to (k,j)$	$\frac{1-r}{\tau}\nu_{ik}$	continue trip
$(i,j) \to (i,k)$	$\frac{1}{T}\omega_{jk}$	relocate (optional)

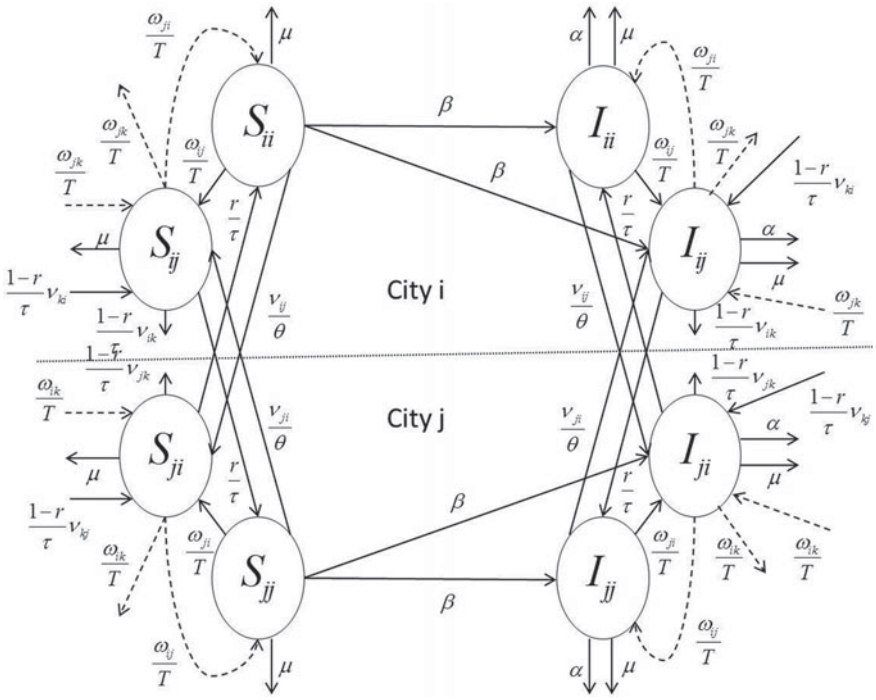

Fig. 3 State transition diagram for the model. Note that health condition is not affected by motion, and all paths from $S_{xy} \rightarrow I_{xy}$ represent disease propagation

We consider a basic SIR model with births and deaths occurring at equal rate, μ, disease recovery rate, α, disease transmission rate, β, and the total population, N, remaining fixed:

$$\dot{S} = \mu N - \mu S - \beta SI \tag{3}$$

$$\dot{I} = \beta SI - \alpha I - \mu I \tag{4}$$

$$\dot{R} = \alpha I - \mu R \tag{5}$$

Here S means susceptible, I means infectious, R means removed with $S + I + R = N$.

Let S_{ij} means susceptible move from j to i, I_{ij} means infectious move from j to i, R_{ij} means removed from j to i. Figure 3 illustrates the transition paths available in the model. From [6], the SMM allows two types of motions: relocation and short trip. given a two location example, we can get a general SMM model

$$\dot{S}_{ii} = \mu N_i - \mu S_{ii} - \beta S_{ii} \sum_j I_{ii} + \frac{r}{\tau} \sum_{j \neq i} S_{ji} - (\frac{1}{\theta} \sum_j \nu_{ij} + \frac{1}{T} \sum_j \omega_{ij}) S_{ii} \quad (6)$$

$$\begin{aligned} \dot{S}_{ij} = & - \mu S_{ij} - \beta S_{ij} \sum_j I_{ij} + \frac{1}{\theta} \nu_{ji} S_{ii} + \frac{1}{T} \omega_{ij} S_{ii} + \frac{1-r}{\tau} \sum_{k \neq j} \nu_{ki} S_{kj} \\ & - (\frac{r}{\tau} + \frac{1-r}{\tau} \sum_{k \neq j} \nu_{ik}) S_{ij} \end{aligned} \quad (7)$$

$$\dot{I}_{ii} = \beta S_{ii} \sum_j I_{ij} - (\alpha + \mu) I_{ii} + \frac{r}{\tau} \sum_{j \neq i} I_{ji} - (\frac{1}{\theta} \sum_j \nu_{ij} + \frac{1}{T} \sum_j \omega_{ij}) I_{ii} \quad (8)$$

$$\begin{aligned} \dot{I}_{ij} = & \beta S_{ij} \sum_j I_{ij} - (\alpha + \mu) I_{ij} + \frac{1}{\theta} \nu_{ji} I_{jj} + \frac{1}{T} \omega_{ij} Iii + \frac{1-r}{\tau} \sum_{k \neq j} \nu_{ki} I_{kj} \\ & - (\frac{r}{\tau} + \frac{1-r}{\tau} \sum_{k \neq j} \nu_{ij}) I_{ij} \end{aligned} \quad (9)$$

2.2 The Population Mobility Model with SMM

Set $N_{ij} = S_{ij} + I_{ij} + R_{ij}$. Define N_i be the residents from i, $N_i = \sum_{j=1}^{n} N_{ji}$. Applying the SMM model from [6], we can get a population mobility model as follows:

$$\dot{N}_{ii} = \mu(N_i - N_{ii}) + \frac{r}{\tau} \sum_{j \neq i} N_{ji} - (\frac{1}{\theta} \sum_j \nu_{ij} + \frac{1}{T} \sum_j \omega_{ij}) N_{ii} \quad (10)$$

$$\dot{N}_{ij} = -\mu N_{ij} + \frac{1}{\theta} \nu_{ji} N_{jj} + \frac{1}{T} \omega_{ij} N_{ii} + \frac{1-r}{\tau} \sum_{k \neq j} \nu_{ki} N_{kj} - (\frac{r}{\tau} + \frac{1-r}{\tau} \sum_{k \neq j} \nu_{ik}) N_{ij} \quad (11)$$

3 The Discrete-Time Model

In this section, we use the forward Euler method [7–10]. Define $\phi(t)$ as the time step. We obtain the discrete-time model as follows:

$$S_{ii}(t+1) - S_{ii}(t) = \phi(t)[\mu N_i - \mu S_{ii}(t) - \beta S_{ii}(t) \sum_j I_{ij}(t) + \frac{r}{\tau} \sum_{j \neq i} S_{ji}(t) - (\frac{1}{\theta} \sum_j \nu_{ij} + \frac{1}{T} \omega_{ij}) S_{ii}(t)] \tag{12}$$

$$S_{ij}(t+1) - S_{ij}(t) = \phi(t)[-\mu S_{ij}(t) - \beta S_{ij}(t) \sum_j I_{ij}(t) + \frac{1}{\theta} \nu_{ji} S_{jj}(t) + \frac{1}{T} \omega_{ij} S_{ii}(t) + \frac{1-r}{\tau} \sum_{k \neq j} \nu_{ki} S_{kj}(t) - (\frac{r}{\tau} + \frac{1-r}{\tau} \sum_{k \neq j} \nu_{ik}) S_{ij}(t)] \tag{13}$$

$$I_{ii}(t+1) - I_{ii}(t) = \phi(t)[\beta S_{ii}(t) \sum_j I_{ij}(t) - (\alpha + \mu) I_{ii}(t) + \frac{r}{\tau} \sum_{j \neq i} I_{ji}(t) - (\frac{1}{\theta} \sum_j \nu_{ij} + \frac{1}{T} \omega_{ij}) I_{ii}(t)] \tag{14}$$

$$I_{ij}(t+1) - I_{ij}(t) = \phi(t)[\beta S_{ij}(t) \sum_j I_{ij}(t) - (\alpha + \mu) I_{ij}(t) + \frac{1}{\theta} \nu_{ji} I_{jj}(t) + \frac{1}{T} \omega_{ij} I_{ii}(t) + \frac{1-r}{\tau} \sum_{k \neq j} \nu_{ki} I_{kj}(t) - (\frac{r}{\tau} + \frac{1-r}{\tau} \sum_{k \neq j} \nu_{ik}) I_{ij}(t)] \tag{15}$$

Similarly, we obtain the discrete population mobility model as follows:

$$N_{ii}(t) = \phi(t)[\mu(N_i - N_{ii}) + \frac{r}{\tau} \sum_{j \neq i} N_{ji} - (\frac{1}{\theta} \sum_j \nu_{ij} + \frac{1}{T} \sum_j \omega_{ij}) N_{ii}] \tag{16}$$

$$N_{ij}(t+1) - N_{ij}(t) = \phi(t)[-\mu N_{ij} + \frac{1}{\theta} \nu_{ji} N_{jj} + \frac{1}{T} \omega_{ij} N_{ii} + \frac{1-r}{\tau} \sum_{k \neq j} \nu_{ki} N_{kj} - (\frac{r}{\tau} + \frac{1-r}{\tau} \sum_{k \neq j} \nu_{ik}) N_{ij}] \tag{17}$$

If $I_{ij} = 0$ for all (i, j), the population will be 'disease free'. The Eqs. (12)–(15) will reduce to

$$S_{ii}(t+1) - S_{ii}(t) = \phi(t)[\mu N_i - \mu S_{ii}(t) + \frac{r}{\tau} \sum_{j \neq i} S_{ji}(t) - (\frac{1}{\theta} \sum_j \nu_{ij} + \frac{1}{T} \omega_{ij}) S_{ii}(t)] \tag{18}$$

$$S_{ij}(t+1) - S_{ij}(t) = \phi(t)[-\mu S_{ij}(t) + \frac{1}{\theta} \nu_{ji} S_{jj}(t) + \frac{1}{T} \omega_{ij} S_{ii}(t) + \frac{1-r}{\tau} \sum_{k \neq j} \nu_{ki} S_{kj}(t) - (\frac{r}{\tau} + \frac{1-r}{\tau} \sum_{k \neq j} \nu_{ik}) S_{ij}(t)] \tag{19}$$

Which is equivalent to model (16) and (17). If $I_{ij} = 0$ then $R_{ij} = 0$, and so $N_{ij} = S_{ij}$ for all time steps.

Following the method in [11] we consider discrete dynamical system

$$x(t+1) = f(x(t)),\ t = 0, 1, 2, ... \tag{20}$$

which has an equilibrium point x^* if $f(x^*) = x^*$. Writting (16) and (17) in the form of (20), we obtain equilibrium points where

$$\mu(N_i - N_{ii}) + \frac{r}{\tau}\sum_{j\neq i} N_{ji} - (\frac{1}{\theta}\sum_{j}\nu_{ij} + \frac{1}{T}\sum_{j}\omega_{ij})N_{ii} = 0 \tag{21}$$

$$-\ \mu N_{ij} + \frac{1}{\theta}\nu_{ji}N_{jj} + \frac{1}{T}\omega_{ij}N_{ii} + \frac{1-r}{\tau}\sum_{k\neq j}\nu_{ki}N_{kj} - (\frac{r}{\tau} + \frac{1-r}{\tau}\sum_{k\neq j}\nu_{ik})N_{ij} = 0 \tag{22}$$

The solution of (21) and (22) defines the equilibrium of model (16) and (17).

The solution of (21) and (22) orders as follows:

$$N^*_{11}, N^*_{12}, ..., N^*_{1n}, N^*_{21}, ..., N^*_{2n}, ..., N^*_{n1}, ..., N^*_{nn}$$

Which gives population mobility equilibrium $n^2 \times n^2$ diagonal matrix N^*.

4 The Basic Reproduction Number R_0 and Disease-Free Equilibrium (DFE)

Ordering the infectious variable

$$I_{11}, I_{12}, ..., I_{1n}, I_{21}, ..., I_{2n}, ..., I_{n1}, ..., I_{nn}.$$

Define matrix V as follow:

$$V=\phi(t)\begin{pmatrix}
a_{11} & 0 & 0 & \dots & 0 & -\frac{r}{\tau} & 0 & 0 & \dots & 0 & -\frac{r}{\tau} & 0 & 0 & \dots & 0\\
-\frac{\omega_{12}}{I} & a_{12} & 0 & \dots & 0 & 0 & -\frac{\nu_{21}}{\theta} & 0 & \dots & 0 & \dots\ 0 & \frac{r-1}{\tau}\nu_{n1} & 0 & \dots & 0\\
-\frac{\omega_{13}}{I} & 0 & a_{13} & \dots & 0 & 0 & 0 & \frac{r-1}{\tau}\nu_{21} & \dots & 0 & \dots\ 0 & 0 & \frac{r-1}{\tau}\nu_{n1} & \dots & 0\\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \\
-\frac{\omega_{1n}}{I} & 0 & 0 & \dots & a_{nn} & 0 & 0 & 0 & \dots & \frac{r-1}{\tau}\nu_{21} & 0 & 0 & 0 & \dots & -\frac{\nu_{n1}}{\theta}\\
-\frac{\nu_{12}}{\theta} & 0 & 0 & \dots & 0 & a_{21} & -\frac{\omega_{21}}{I} & 0 & \dots & 0 & \frac{r-1}{\tau}\nu_{n2} & 0 & 0 & \dots & 0\\
0 & -\frac{r}{\tau} & 0 & \dots & 0 & 0 & a_{22} & 0 & \dots & 0 & 0 & -\frac{r}{\tau} & 0 & \dots & 0\\
0 & 0 & \frac{r-1}{\tau}\nu_{12} & \dots & 0 & 0 & -\frac{\omega_{23}}{I} & a_{23} & \dots & 0 & \dots\ 0 & 0 & \frac{r-1}{\tau}\nu_{n2} & \dots & 0\\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots\\
0 & 0 & 0 & \dots & \frac{r-1}{\tau}\nu_{12} & 0 & -\frac{\omega_{2n}}{I} & 0 & \dots & a_{2n} & 0 & 0 & 0 & \dots & -\frac{\nu_{n2}}{\theta}\\
 & & \vdots & & & & & \vdots & & & \ddots & & \vdots & & \\
-\frac{\nu_{1n}}{\theta} & 0 & 0 & \dots & 0 & \frac{r-1}{\tau}\nu_{2n} & 0 & 0 & \dots & 0 & a_{n1} & 0 & 0 & \dots & -\frac{\omega_{n1}}{I}\\
0 & \frac{r-1}{\tau}\nu_{1n} & 0 & \dots & 0 & 0 & -\frac{\nu_{2n}}{\theta} & 0 & \dots & 0 & 0 & a_{n2} & 0 & \dots & -\frac{\omega_{n2}}{I}\\
0 & 0 & \frac{r-1}{\tau}\nu_{1n} & \dots & 0 & 0 & 0 & \frac{r-1}{\tau}\nu_{2n} & \dots & 0 & \dots\ 0 & 0 & a_{n3} & \dots & -\frac{\omega_{n3}}{I}\\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots\\
0 & 0 & 0 & \dots & -\frac{r}{\tau} & 0 & 0 & 0 & \dots & -\frac{r}{\tau} & 0 & 0 & 0 & \dots & a_{nn}
\end{pmatrix},$$

where, $a_{nn} = \alpha + \mu + \frac{1}{\theta}\sum_j \nu_{nj} + \frac{1}{I}\sum_j \omega_{nj}$, $a_{ni} = \alpha + \mu - \frac{1-r}{\mu}\nu_{nn} + (\frac{r}{\tau} + \frac{1-r}{\tau}\sum_{k\neq i}\nu_{nk})$. According to [12], the basic reproduction number of model (12)–(15) is:

$$R_0 = \rho(FV^{-1}),$$

where ρ is spectral radius.

Using the method of [7, 12] we construct the diagonal matrix F as follow:

$$F=\phi(t)\cdot\beta\cdot N^*=\phi(t)\cdot\beta\begin{pmatrix}
N_{11}^* & & & & & & & & & & & & & & \\
 & N_{12}^* & & & & & & & & & & & & & \\
 & & \ddots & & & & & & & & & & & & \\
 & & & N_{1n}^* & & & & & & & & & & & \\
 & & & & N_{21}^* & & & & & & & & & & \\
 & & & & & N_{22}^* & & & & & & & & & \\
 & & & & & & \ddots & & & & & & & & \\
 & & & & & & & N_{2n}^* & & & & & & & \\
 & & & & & & & & \ddots & & & & & & \\
 & & & & & & & & & N_{n1}^* & & & & & \\
 & & & & & & & & & & N_{n2}^* & & & & \\
 & & & & & & & & & & & \ddots & & & \\
 & & & & & & & & & & & & N_{nn}^* & &
\end{pmatrix},$$

Let E_0 be the DFE of model (12)–(15). At the DFE, Eqs. (18) and (19) are equivalent to (16) and (17), so

$$E_0 = [N_{11}^*, N_{12}^*, ..., N_{1n}^*, N_{21}^*, ..., N_{2n}^*, ..., N_{n1}^*, ..., N_{nn}^*, \underbrace{0, \ldots, 0}_{n\times n}]$$

in the ordering of

$$S_{11}, \ldots, S_{1n}, S_{21}, \ldots, S_{2n}, \ldots, S_{n1}, \ldots, S_{nn},$$
$$I_{11}, \ldots, I_{1n}, \ldots, I_{n1}, \ldots, I_{nn}.$$

Because the DFE considered as $I = 0$, which cause $R = 0$, E_0 not includes the removed part. There is a convenient way to analyze the stability of DFE according to Theorem 2 in [12].

Lemma 1 *$R_0 < 1$ implies DFE be locally asymptotically stable; $R_0 > 1$ implies DFE unstable.*

Proof Here the matrices F and V follow the method [12]. We easily verify that matrix F and V satisfies the (A1)–(A5) in [12] and that V is also a non-singular $M - matrix$, F is also a non-negative matrix. Let $C = F - V$, and $s(C)$ be the maximum real part of the eigenvalues of matrix C (spectual abscissa). According to the Theorem 1 and proved in [12] and Lemma 4.1 in [9], $R_0 > 1 \iff s(C) > 0$, $R_0 < 1 \iff s(C) < 0$. So we can get $R_0 < 1$ implies DFE be locally asymptotically stable; $R_0 > 1$ implies DFE unstable.

Theorem 1 *DFE is globally asymptotically stable when $R_0 < 1$; when $R_0 > 1$, the DFE is unstable.*

5 Example

In this section, we validate our algorithm for vectorization of the infected SIR-SMM model. We construct three examples of locations. These two examples show how coefficients affect the basic reproduction number (Figs. 4 and 5).

6 The PDE View of the SIR Model with Social Mobility

6.1 Basic SIR Model

Firstly, we consider the Moran process according to the method in [13, 14] in the basic SIR model (6)–(8). Define N be the total constant population. $P_{(N,\Delta t)}(t, n, m)$ be the probability there are n susceptible, m infectious and $N - m - n$ removed at time t. Define Δt to be the time step. Then we have

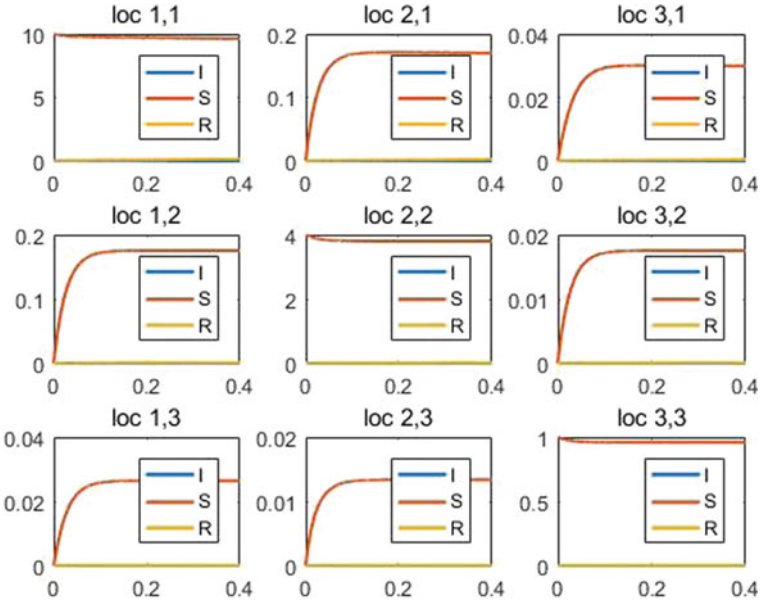

Fig. 4 $N_1 = 10, N_2 = 4, N_3 = 1, S = 0.99N, I = 0.01N, \phi(t) = 0.0002, \mu = 0.005, \beta = 4.8, \alpha = 48, \tau = \frac{5}{365}, \theta = 1, T = 10, r = 0.5$. The S_{ij} and I_{ij} as functions of time t. Here $R_0 < 1$ shows the disease will die out. $\rho(A) < 1, s(C) < 0$ shows the DFE is locally asymptotical stable

$$
\begin{aligned}
&P_{(N,\Delta t)}(t+\Delta t,n,m) = (\mu+\beta\frac{m-1}{N-1})\frac{n+1}{N}P_{(N,\Delta t)}(t,n+1,m-1)\\
&+(\alpha+\mu)\frac{m+1}{N}P_{(N,\Delta t)}(t,n,m+1)+\mu\frac{N-n-m+1}{N}P_{(N,\Delta t)}(t,n-1,m)\\
&+[\frac{n}{N}(1-(\mu+\beta\frac{m}{N-1}))+\frac{m}{N}(1-\alpha-\mu)+\frac{N-n-m}{N}(1-\mu)]P_{(N,\Delta t)}(t,n,m)
\end{aligned}
\tag{23}
$$

Let $x = \frac{n}{N}$, $y = \frac{m}{N}$, $p(t,x,y) = NP_{(N,\Delta t)}(t,xN,yN) = p$. Keeping terms of order $\frac{1}{N}$

$$
\begin{aligned}
&p_{(N,\Delta t)}(t+\Delta t,x,y) = (\mu+\beta\frac{Ny-1}{N-1})\frac{xN+1}{N}p(t,x+\frac{1}{N},y-\frac{1}{N})\\
&+(\alpha+\mu)(y+\frac{1}{N})p(t,x,y+\frac{1}{N})+\mu(1-x-y+\frac{1}{N})p(t,x-\frac{1}{N},y)\\
&+[x(1-(\mu+\beta\frac{y}{1-\frac{1}{N}}))+y(1-\alpha-\mu)+(1-x-y)(1-\mu)]p(t,x,y)\\
&\approx p+\frac{1}{N}[(3\mu+\alpha-\beta x+\beta y)p+(\beta xy+2\mu x+\mu y-\mu)p_x+(\alpha y+\mu y-\beta xy-\mu)p_y]\\
&\approx p+\frac{1}{N}[\partial_x((\beta xy+\mu(2x+y-1))p)+\partial_y(((\alpha-\beta x)y+\mu(y-1))p)]
\end{aligned}
\tag{24}
$$

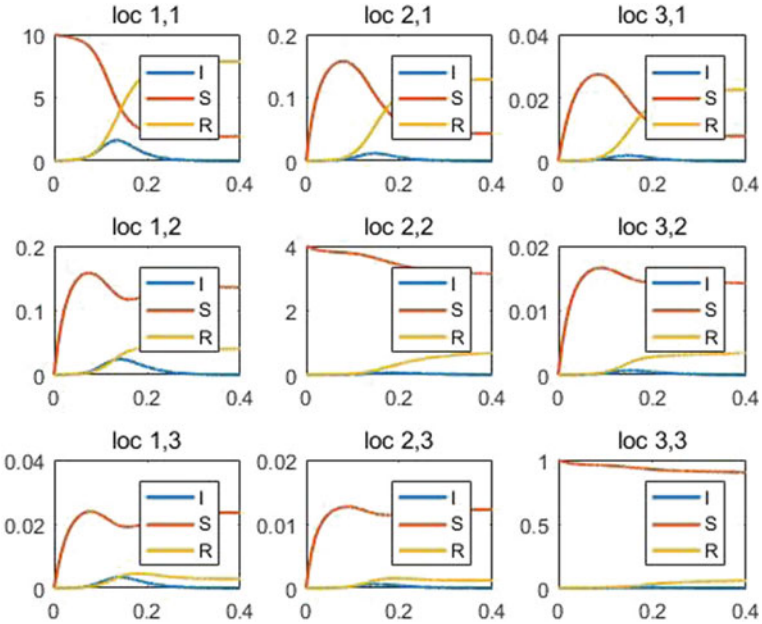

Fig. 5 The S_{ij} and I_{ij} as functions of time t. Here we suppose there are 3 locations and $N_1 = 10$, $N_2 = 4$, $N_3 = 1$, $S = 0.99N$, $I = 0.01N$, $\phi(t) = 0.0002$, $\mu = 0.005$, $\beta = 10$, $\alpha = 48$, $\tau = \frac{5}{365}$, $\theta = 1$, $T = 10$, $r = 0.5$, with $R_0 = 2.0036$

Let a, b, c be positive constants and defined by

$$\lim_{N\to\infty_{\Delta t\to 0}} \frac{\beta}{N\Delta t} = b$$

$$\lim_{N\to\infty_{\Delta t\to 0}} \frac{\alpha}{N\Delta t} = a$$

$$\lim_{N\to\infty_{\Delta t\to 0}} \frac{\mu}{N\Delta t} = c$$

where

$$ab \neq 0.$$

Then we have

$$\partial_t p = \partial_x(((bxy + c(2x + y - 1)))p) + \partial_y(((a - bx)y + c(y - 1))p) \quad (25)$$

6.2 2 Location Condition

According to the method in [13], define N_i be the total population in city i, n_{ij} be the susceptible, m_{ij} be the infectious and $N_i - n_{ij} - m_{ij}$ be the recovery from city j to city i. Define $P_{(N_i,\Delta t)}(t, n_{ij}, m_{ij})$ be the probability that there are n_{ij} susceptible, m_{ij} infections and $N_i - n_{ij} - m_{ij}$ removed at t steps.

Without loss of generality, in city 1, $x_1 = \frac{n_{11}}{N_1}$, $y_1 = \frac{m_{11}}{N_1}$, $y_2 = \frac{m_{12}}{N_1}$. For population from city 1 to city 1, let $p_1 = p(t, x_1, y_1) = N_1 P_{(N_1,\Delta t)}(t, x_1 N_1, y_1 N_1)$ For population from city 2 to city 1, let $p_2 = p(t, x_2, y_2) = N_1 P_{(N_1,\Delta t)}(t, x_2 N_1, y_2 N_1)$ For susceptible moves from city 1 to city 1, infectious moving from city 2 to city 1, let $p_3 = p(t, x_1, y_2) = N_1 P_{(N_1,\Delta t)}(t, x_1 N_1, y_2 N_1)$

Because S_{ij} cannot become an I_{ii}, there is no term $P_{(N_1,\Delta t)}(t, n_{12}, m_{11})$. Let

$$\lim_{N\to\infty_{\Delta t\to 0}} \frac{\frac{1}{\theta}}{N\Delta t} = d$$

$$\lim_{N\to\infty_{\Delta t\to 0}} \frac{\frac{1}{T}}{N\Delta t} = l$$

$$\lim_{N\to\infty_{\Delta t\to 0}} \frac{\frac{r}{\tau}}{N\Delta t} = h$$

Then we can obtain

$$\begin{aligned}\partial_t p_1 =& \partial_{x_1}[(2(b(y_1+y_2)+c+d\nu_{12}+l\omega_{12})x_1+(c+d\nu_{12}+l\omega_{12})(y_1-1))p_1]\\ &+\partial_{y_1}[((a-bx_1+c+d\nu_{12}+l\omega_{12})y_1-(by_2+c+d\nu_{12}+l\omega_{12})x_1)p_1]\end{aligned} \tag{26}$$

$$\begin{aligned}\partial_t p_2 =& \partial_{x_2}[((b(y_1+y_2)+2c+2h)x_2+(c+h)(y_2-1))p_2]\\ &+\partial_{y_2}[((a-bx_2+c+h)y_2-(by_1+c+h)x_2)p_2]\end{aligned} \tag{27}$$

$$\begin{aligned}\partial_t p_3 =& \partial_{x_1}[((b(y_1+y_2)+2c+d\nu_{12}+l\omega_{12}+h)x_1+(c+h)(y_2-1))p_3]\\ &+\partial_{y_2}[((a-bx_1+c+h)y_2-(by_1+c+d\nu_{12}+l\omega_{12})x_1)p_3]\end{aligned} \tag{28}$$

Equations (26)–(28) constitute a new stochastic model SIR and SMM based upon the constitution of Sect. 3. The analysis of (26)–(28) and their generalization to n locations will be the subject of future work.

7 Discussion

In this paper, we consider discrete methods for the SIR model with SMM based on the [15, 16]. Because the existence criteria of the steady states in the continuous-time and discrete-time models are the same, both continuous and discrete-time models have the same equilibrium [17]. According to the method introduced in [13, 14], we get the

hyperbolic forward Kolmogorov equation for the SIR model (6)–(8) corresponding to the Moran process. We also construct the Kolmogorov equations for the SIR-SMM model between 2 locations.

Our further work is to prove the uniqueness of the solution in (6) and (26)–(28), study how the solution changes with time, and generalize the Kolmogorov equations of SIR-SMM model (26)–(28) to n locations. How to find a suitable initial condition for (26)–(28) is also something we will investigate in our future work.

Acknowledgments I would like to thank Professor Joseph D. Skufca for his support and encouragement. He kindly read my paper and offered invaluable detailed advice on grammar, organization, and the theme of the paper. I sincerely thank my parents and friends, who provide pieces of advice and financial support. The product of this research paper would not be possible without all of them.

References

1. Murray, J.D.: Mathematical Biology, 3rd edn. Springer, Berlin, Heidelberg (2002)
2. Anderson, Roy M., May, Robert M.: Infectious Diseases of Humans Daynamics and Control, 1st edn. Oxford University Press, New York (1991)
3. Hethecote, H.W.: The mathematics of infectious diseases. SIAM Rev. **42**(4), 599–653
4. Sattenspiel, L., Dietz, K.: A structured epidemic model incorporating geographic mobility among regions. Math. Biosci. **128**(71–91) (1995)
5. Bharti, N., Xia, Y., Bjornstad, O.N., Grenfell, B.T.: Measles on the edge: coastal heterogeneities and infection dynamics. PLoS One **3**(4), e1941 (2008)
6. Skufca, J.D., Ben Avraham, D.: A model of human population motion. [physics.soc-ph], (arXiv:1006.1301v2), 7 Oct 2014
7. Arino, J., van den Driessche, P.: A multi-city epidemic model. Math. Popul. Stud. **10**, 175–193 (2003)
8. Arino, J., van den Driessche, P.: The basic reproduction number in a multi-city compartmental epidemic model. In: Positive Systems. LNCIS, vol. 294, pp. 135–142 (2003)
9. Ruan, S., Wang, W., Levin, S.A.: The effect of global travel on the spread of SARS. Math. Bopsci. Eng. **3**(1) (2006)
10. Ramani, A., Carstea, A.S., Willox, R., Grammaticos, B.: Oscillating epidemics: a discrete-time model. Phys. A **333**, 278–292) (2004)
11. Feng, Z., Velasco-Hernandez, J., Tapia-Santos, B., Leite, M.C.A.: A model for coupling within-host and between-host dynamics in an infectious disease. Nonlinear Dyn. (2012). https://doi.org/10.1007/s11071-011-0291-0(68:401411)
12. van den Driessche, P., Watmough, J.: Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. Math. Biosci. **180**, 29–8 (2002)
13. Chalub, F.A.C.C., Souza, M.O.: The SIR epidemic model from a PDE point of view. Math. Comput. Model. **53**, 1568–1574 (2011)
14. Chalub, F.A.C.C., Souza, M.O.: From discrete to continuous evolution models: a unifying approach to drift-diffusion and replicator dynamics. Theor. Popul. Biol. **76**, 268–277 (2009)
15. Allen, L.J.S.: Some discrete-time SI, SIR, and SIS epidemic models. Math. Biosci. **124**(1), 83–105 (1994)
16. Allen, L.J.S., van den Driessche, P.: The basic reproduction number in some discrete-time epidemic models. J. Diff. Equ. Appl. **14**(10–11) (2008)
17. Jang, S., Elaydi, S.: Difference equations from discretization of a continuous epidemic model with immigration of infectives. Can. Appl. Math. Q. **11**, 93–105 (2003)

Uniqueness Criterion and Cramer's Rule for Implicit Higher Order Linear Difference Equations Over Z

V. V. MARTSENIUK, Sergey L. Gefter, and A. L. Piven'

Abstract We obtain uniqueness criterion for integer solutions of implicit higher order difference equations which extends the result of Berestovskii and Nikonorov. This criterion is specified for an implicit third order difference equation. We also obtain existence and uniqueness theorems for a solution of an implicit higher order nonhomogeneous difference equation over the ring of p-adic integers and over the ring $\mathbf{Z}$. The possibility to obtain this solution by using the Cramer's rule is established. We also give the explicit form for this solution. These results generalize corresponding results for implicit first and second order nonhomogeneous difference equations.

Keywords Cramer's rule · Implicit linear difference equation · Integer solution · p-adic topology

MSC: 39A06 · 65Q10

V. V. MARTSENIUK (✉)
Yaroslav Mudryi National Law University, 77 Pushkinskaya Street, Kharkiv 61024, Ukraine
e-mail: martsenyukvv@gmail.com
URL: https://martsenyukvv.wixsite.com/martseniukvv

V. V. MARTSENIUK · S. L. Gefter · A. L. Piven'
School of Mathematics and Computer Science, V. N. Karazin Kharkiv National University, 4 Svobody Sq., Kharkiv 61022, Ukraine
e-mail: gefter@karazin.ua
URL: http://puremath.univer.kharkov.ua/~gefter

A. L. Piven'
e-mail: aleksei.piven@karazin.ua
URL: http://appmath.univer.kharkov.ua/Piven.htm

S. Baigent et al. (eds.), *Progress on Difference Equations and Discrete Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 341,
https://doi.org/10.1007/978-3-030-60107-2_16

1 Introduction

We consider the following higher order linear difference equation

$$c_N x_{n+N} = c_{N-1} x_{n+N-1} + \cdots + c_1 x_{n+1} + c_0 x_n - f_n, \quad n = 0, 1, 2, ..., \qquad (1)$$

where $N \geq 2$, $c_0, ..., c_N$ and f_n are integers ($n = 0, 1, 2, ...$), $c_0 \neq 0$, $c_N \neq 0$. If $c_N = \pm 1$ then we have an *explicit* equation. It is clear that any explicit equation has a unique integer solution for any initial data $x_0, x_1, ..., x_{N-1} \in \mathbf{Z}$ (see, for example, [1]). In what follows we assume that $c_N \neq \pm 1$ and at least one of $c_0, ..., c_{N-1}$ is not divisible by c_N. In this situation Eq. (1) is said to be *implicit* over the ring $\mathbf{Z}$.

Let us note that an implicit difference equation even of first order does not necessarily have an integer solution. For example, the general solution of the equation $3x_{n+1} = x_n + 1$ over $\mathbf{Q}$ has the form $x_n = \frac{a}{3^n} + \frac{1}{2}$, where $a \in \mathbf{Q}, n = 0, 1, 2, \ldots$. It is obvious that for any value of a constant a we cannot obtain an integer solution (see Example 2.1 in [2]). An implicit difference equation may have an integer solution. For example, the equation $3x_{n+1} = 4x_n + 5$ has the integer solution $x_n = -5, n = 0, 1, 2, \ldots$. Such unexpected integrality is sometimes called the "Laurent phenomena" (see [3]).

Equation (1) can be regarded as an infinite system of linear equations with coefficients in $\mathbf{Z}$. If Eq. (1) is explicit, then the corresponding infinite system has infinitely many solutions over $\mathbf{Z}$. For the implicit equation the situation is already different: the uniqueness of an integer solution takes place in many cases [4–6]. In Sect. 2 of the present paper we obtain the general uniqueness criterion of integer solution which extends the V.N. Berestovskii and Yu.G. Nikonorov's result [7, Theorem 6]. In Theorem 2 this criterion is specified for an implicit third order difference equation. The case of an implicit second order difference equation was considered in [6, Theorem 1].

As shown in [8, 9], if an integer solution to an implicit first or second order linear difference equation is unique, then this solution can be found by some analogue of Cramer's rule (see also Remark 4 of Sect. 4 in this paper). The application of the p-adic topology on the ring $\mathbf{Z}$ was essential for this [2]. For other applications of p-adic numbers to difference equations see [10]. In [11, Remark 3.4] Cramer's rule was considered for first order implicit linear difference equations in Fréchet spaces and other locally convex spaces. In this paper we obtain an analogue of Cramer's rule for some implicit higher order linear difference equations. Unlike the first and second order equations, this problem is much more complicated because the explicit expression for the solution of Eq. (1) is too cumbersome and the process of computing the corresponding finite order determinants becomes much more complicated (see the proofs of Lemma 1 and Theorem 3). The difference equation (1) does not necessarily have an integer solution. Therefore we begin by studying Eq. (1) over the ring $\mathbf{Z}_p$ of p-adic integers [12, Chap. 1, Sect. 3], and show that, under some additional assumptions, Eq. (1) has a unique solution over $\mathbf{Z}_p$. Moreover, this solution can be found by an analogue of the Cramer's rule (see Theorem 3). Under additional

conditions on the coefficients $c_0, ..., c_N$ of Eq. (1) we prove the main result of this paper concerning the possibility to find the unique integer solution by using of an analogue of Cramer's rule (see Theorem 4).

Some results of this paper were presented at the 25th International Conference on Difference Equations and Applications at UCL on 24th June–8th June 2019.

2 Uniqueness Criterion of an Integer Solution for a Higher Order Implicit Difference Equation

We prove the following criterion for the uniqueness of an integer solution of Eq. (1) in terms of the characteristic polynomial $\chi(\lambda) = c_N\lambda^N - \sum_{j=0}^{N-1} c_j\lambda^j$.

Theorem 1 *The implicit homogeneous equation*

$$c_N x_{n+N} = c_{N-1}x_{n+N-1} + \cdots + c_1 x_{n+1} + c_0 x_n, \quad n = 0, 1, 2, \ldots \tag{2}$$

has only the trivial solution in integer numbers if and only if the factorization of the characteristic polynomial $\chi(\lambda)$ on primitive irreducible over $\mathbf{Z}$ polynomials has no polynomials with leading coefficients equal to ± 1.

Proof By Gauss's theorem (see, for example, [13, Chap. IV, Sect. 2, Theorem 2.3]) the characteristic polynomial of Eq. (2) admits the factorization

$$\chi(\lambda) = cp_1(\lambda) \cdot \ldots \cdot p_m(\lambda), \tag{3}$$

where $c \in \mathbf{Z}$ and $p_1(\lambda), \ldots, p_m(\lambda)$ are primitive non-constant irreducible over $\mathbf{Z}$ polynomials. This factorization is unique up to order of factors. Define the shift operator $S : \mathbf{Z}^{\mathbf{N}\cup\{0\}} \to \mathbf{Z}^{\mathbf{N}\cup\{0\}}$ as follows

$$S(\{x_n\}_{n=0}^{\infty}) = \{x_{n+1}\}_{n=0}^{\infty}.$$

Equation (2) can be rewritten in the operator form

$$\chi(S)(\{x_n\}_{n=0}^{\infty}) = 0, \quad n = 0, 1, 2, \ldots,$$

or by the decomposition (3) in the form

$$p_1(S) \cdot \ldots \cdot p_m(S)(\{x_n\}_{n=0}^{\infty}) = 0 \tag{4}$$

(see [1, Kelly-Peterson, Sect. 3.3, Theorem 3.7]). To prove the sufficiency of the assertion of Theorem 1 we assume that leading coefficients of $p_1(\lambda), ..., p_m(\lambda)$ are not equal to ± 1. The polynomial $p_1(\lambda)$ is irreducible over $\mathbf{Q}$ and its leading coefficient is not a common divisor of the other coefficients. It follows from Theorem 6 [7] and Eq. (4) that the sequence $\{x_n\}_{n=0}^{\infty}$ is an integer solution of the equation

$$p_2(S) \cdot \ldots \cdot p_m(S)(\{x_n\}_{n=0}^{\infty}) = 0$$

Using similar arguments for polynomials $p_2(\lambda), \cdot \ldots \cdot, p_m(\lambda)$, we obtain $x_n = 0, \quad n = 0, 1, 2, \ldots$.

Now we prove the necessity of the assertion of Theorem 1. Let Eq. (2) have only the trivial solution in integer numbers. Assume to the contrary that the leading coefficient at least one of polynomials $p_1(\lambda), \ldots, p_m(\lambda)$ in the factorization (3) is equal to ± 1. Without loss of generality, suppose that

$$p_m(\lambda) = \lambda^k + \sum_{j=0}^{k-1} a_j \lambda^j, \quad a_j \in \mathbf{Z}, \ j = 0, 1, 2, \ldots.$$

Consider a non-trivial integer solution $\{x_n\}_{n=0}^{\infty}$ of the explicit difference equation

$$x_{n+k} + \sum_{j=0}^{k-1} a_j x_{n+j} = 0, \quad n = 0, 1, 2, \ldots.$$

Then $p_m(S)(\{x_n\}_{n=0}^{\infty}) = 0$ and

$$p(S)(\{x_n\}_{n=0}^{\infty}) = p_1(S) \cdot \ldots \cdot p_m(S)(\{x_n\}_{n=0}^{\infty}) = 0.$$

Consequently, $\{x_n\}_{n=0}^{\infty}$ is a non-trivial solution of Eq. (2). This contradicts assumption of the theorem. The proof is complete.

Corollary 1 *Let $f_n = f \in \mathbf{Z}$ for all $n = 0, 1, 2, \ldots$. If the expansion of the characteristic polynomial $\chi(\lambda)$ on primitive irreducible over $\mathbf{Z}$ polynomials has no polynomials with leading coefficients equal to ± 1, then $c_N - \sum_{j=0}^{N-1} c_j \neq 0$ and the nonhomogeneous equation*

$$c_N x_{n+N} = c_{N-1} x_{n+N-1} + \cdots + c_1 x_{n+1} + c_0 x_n - f, \quad n = 0, 1, 2, \ldots \tag{5}$$

has an integer solution if and only if $c_N - \sum_{j=0}^{N-1} c_j$ is a divisor of f. If an integer solution exists, then it is unique, it is constant and has the form

$$x_n = \frac{f}{\sum_{j=0}^{N-1} c_j - c_N}, \ n = 0, 1, 2, \ldots.$$

Proof By Theorem 1 Eq. (2) has only the trivial solution in integer numbers. Since the constant sequence $z_n = 1$ is not a solution of Eq. (2), we obtain $c_N - \sum_{j=0}^{N-1} c_j \neq 0$. Let $\{x_n\}_{n=0}^{\infty}$ be an integer solution of Eq. (5). Set $y_n = x_{n+1}$. Then the sequence $\{y_n\}_{n=0}^{\infty}$ is an integer solution of Eq. (5) as well. Then $y_n = x_n$, i.e. $x_n = x_0$ for all n

and $\left(c_N - \sum_{j=0}^{N-1} c_j\right) x_0 = -f$. Conversely, if $c_N - \sum_{j=0}^{N-1} c_j$ is a divisor of f, then the constant sequence $x_n = \dfrac{f}{\sum_{j=0}^{N-1} c_j - c_N}$ is an integer solution of (5).

Example 1 Consider the following third order implicit linear homogeneous difference equation:

$$4x_{n+3} = 4x_{n+2} - 3x_{n+1} + x_n, \quad n = 0, 1, 2, \ldots \tag{6}$$

The characteristic polynomial $\chi(\lambda) = 4\lambda^3 - 4\lambda^2 + 3\lambda - 1$ admits the following factorization on primitive irreducible over $\mathbf{Z}$ polynomials:

$$\chi(\lambda) = (2\lambda - 1)(2\lambda^2 - \lambda + 1)$$

By Theorem 1 Eq. (6) has only the trivial integer solution. Now let $f \in \mathbf{Z}$. By Corollary 1 the nonhomogeneous equation

$$4x_{n+3} = 4x_{n+2} - 3x_{n+1} + x_n - f, \quad n = 0, 1, 2, \ldots$$

has an integer solution if and only if f is even. This solution is unique and has the form $x_n = -\frac{f}{2}$.

Theorem 1 at once implies the following simple uniqueness criterion of an integer solution in the particular case of a second order implicit difference equation (a more complicated proof of this assertion may be found in [6, Theorem 1]).

Corollary 2 *The implicit second order homogeneous equation*

$$c_2 x_{n+2} = c_1 x_{n+1} + c_0 x_n, \quad n = 0, 1, 2, \ldots$$

has only the trivial solution in integer numbers if and only if the characteristic polynomial $\chi(\lambda) = c_2\lambda^2 - c_1\lambda - c_0$ has no integer roots.

The following example shows that Corollary 2 can fail for an implicit difference equation of an order greater then two.

Example 2 Consider the following third order implicit linear homogeneous difference equation:

$$3x_{n+3} = 4x_{n+2} + 2x_{n+1} - x_n, \quad n = 0, 1, 2, \ldots \tag{7}$$

The characteristic polynomial of this equation

$$3\lambda^3 - 4\lambda^2 - 2\lambda + 1 = (3\lambda - 1)(\lambda^2 - \lambda - 1)$$

has no integer roots. But the leading coefficient of the factor $\lambda^2 - \lambda - 1$ is equal to 1 and the Fibonacci sequence is an integer solution of Eq. (7). Then Eq. (7) has infinitely many integer solutions.

Now we prove the following uniqueness criterion for an implicit third order difference equation.

Theorem 2 *The implicit third order difference equation*

$$c_3 x_{n+3} = c_2 x_{n+2} + c_1 x_{n+1} + c_0 x_n, \quad n = 0, 1, 2, ... \tag{8}$$

with co-prime coefficients c_3, c_2, c_1, c_0 has only the trivial solution in integer numbers if and only if the one of the following two conditions holds:

1. *The characteristic polynomial $\chi(\lambda) = c_3\lambda^3 - c_2\lambda^2 - c_1\lambda - c_0$ has no rational roots.*
2. *The characteristic polynomial $\chi(\lambda) = c_3\lambda^3 - c_2\lambda^2 - c_1\lambda - c_0$ has no integer roots, but it has a rational root $\lambda = \frac{p}{q} (q \neq \pm c_3)$, where p and q are co-prime.*

Proof *Necessity.* Let Eq. (8) has only the trivial integer solution and the characteristic polynomial $\chi(\lambda)$ has a rational root $\lambda = \frac{p}{q}$, where p and q are co-prime. Assume $\lambda \in \mathbf{Z}$. Then Eq. (8) has the non-trivial integer solution $x_n = \lambda^n$ $(n = 0, 1, 2, ...)$. This contradicts with the assumption about the uniqueness of an integer solution for Eq. (8). Therefore $\lambda = \frac{p}{q} \in \mathbf{Q} \setminus \mathbf{Z}$ and $q \neq \pm 1$.

Let us prove that $q \neq \pm c_3$. Then the primitive polynomial $q\lambda - p$ is a divisor of the polynomial $\chi(\lambda)$. Then by Gauss's theorem [13, Chap. IV, Sect. 2, Theorem 2.3] all the coefficients of the polynomial $a_2\lambda^2 + a_1\lambda + a_0 = \frac{\chi(\lambda)}{q\lambda - p}$ are integers and $qa_2 = c_3$. Moreover, since the polynomial $\chi(\lambda)$ is primitive, the polynomial $a_2\lambda^2 + a_1\lambda + a_0$ is primitive too. Then we have the following decomposition of $\chi(\lambda)$ on primitive irreducible over $\mathbf{Z}$ polynomials

$$\chi(\lambda) = (a_2\lambda^2 + a_1\lambda + a_0) \cdot (q\lambda - p). \tag{9}$$

By Theorem 1 $a_2 \neq \pm 1$ and hence $q \neq \pm c_3$.

Sufficiency. Let Condition 1 is fulfilled, i.e. the polynomial $\chi(\lambda)$ has no rational roots. Then it is irreducible and by Theorem 6 [4] Eq. (8) has only the trivial integer solution. Now suppose Condition 2 holds, i.e. the characteristic polynomial $\chi(\lambda)$ has no integer roots, but it has a rational root $\lambda = \frac{p}{q} (q \neq \pm c_3)$, where p and q are co-prime. We again have the decomposition (9) on primitive irreducible over $\mathbf{Z}$ polynomials. Therefore $qa_2 = c_3$. Since $q \neq \pm c_3$, we obtain $a_2 \neq \pm 1$. By Theorem 1 Eq. (8) has only the trivial integer solution. The proof is complete.

3 Existence Theorem over the Ring of p-adic Integers

Let p be a prime. Consider Eq. (1) over the ring of p-adic integers $\mathbf{Z}_p$.

Lemma 1 *Suppose p is a common prime divisor of $c_1, \ldots c_N$, but c_0 be not divided by p. Then Eq. (1) has a unique solution over $\mathbf{Z}_p$ and this solution may be found as*

$$x_n = \sum_{k=0}^{\infty} y_{k+N-1} \frac{f_{n+k}}{c_0}, \quad n = 0, 1, 2, \ldots, \tag{10}$$

where the sequence $\{y_n\}_{n=0}^{\infty}$ belongs to $\mathbf{Q} \cap \mathbf{Z}_p$ and uniquely solves the initial problem

$$c_N y_n = \sum_{j=1}^{N} c_{N-j} y_{n+j}, \quad n = 0, 1, 2, \ldots, \tag{11}$$

$$y_0 = 0, \ldots, y_{N-2} = 0, \; y_{N-1} = 1. \tag{12}$$

All terms of the series (10) belong to the ring $\mathbf{Z}_p$ and the series (10) converges in the topology of this ring. Thus, a unique solution of (1) over $\mathbf{Z}_p$ is some convolution of the sequence $\{f_n\}_{n=0}^{\infty}$ and a "fundamental solution" $\{y_n\}_{n=0}^{\infty}$ of Eq. (11). Herewith Eq. (11) is some dual equation for Eq. (1).

Proof Since p is not a divisor of c_0, we have $\frac{1}{c_0} \in \mathbf{Z}_p$ (see [12, Chap. 1, Sect. 3, Theorem 4]). Therefore, the difference equation (11) can be written as the explicit equation over $\mathbf{Q} \cap \mathbf{Z}_p$

$$y_{n+N} = \frac{1}{c_0} \left(c_N y_n - \sum_{j=1}^{N-1} c_{N-j} y_{n+j} \right), \quad n = 0, 1, 2, \ldots . \tag{13}$$

Consequently, the solution of the initial problem (13), (12) belongs to $\mathbf{Q} \cap \mathbf{Z}_p$. We show that the series in the right-hand side of (10) converges in the topology of $\mathbf{Z}_p$. For this purpose it suffices to show that y_{k+N} is divided by $p^{[\frac{k+N}{N}]}$ in the ring $\mathbf{Z}_p$ for any $k = 0, 1, 2, \ldots$, i.e.

$$y_{k+N} = p^{[\frac{k+N}{N}]} z_{k+N}, \tag{14}$$

where $\{z_k\}_{k=1}^{\infty}$ is a sequence of elements of $\mathbf{Z}_p$ (see [12, Chap. 1, Sect. 3, Theorem 8]). We prove Formula (14) by induction on k. Let $k = 0$. Since c_1 is divided by p, from (12),(13) it follows that $y_N = c_1 \left(-\frac{1}{c_0} \right) \in p\mathbf{Z}_p$. Assume that the representation (14) holds for $k = 0, \ldots, m-1$, where $m \geq 2$. We show that it is also valid for $k = m$. Since $c_1, \ldots, c_N$ are divided by p, we have $c_j = pb_j, b_j \in \mathbf{Z}, \; j = 1, \ldots, N$. By Eq. (13) and the induction assumption,

$$y_{m+N} = \frac{1}{c_0}\left(c_N y_m - \sum_{j=1}^{N-1} c_{N-j} y_{m+j}\right) = \frac{1}{c_0}\left(pb_N y_m - \sum_{j=1}^{N-1} pb_{N-j} y_{m+j}\right) =$$

$$= \frac{1}{c_0}\left(p^{1+\left[\frac{m}{N}\right]} b_N z_m - \sum_{j=1}^{N-1} p^{1+\left[\frac{m+j}{N}\right]} b_{N-j} z_{m+j}\right) =$$

$$= p^{1+\left[\frac{m}{N}\right]} \frac{1}{c_0}\left(b_N z_m - \sum_{j=1}^{N-1} p^{\left[\frac{m+j}{N}\right]-\left[\frac{m}{N}\right]} b_{N-j} z_{m+j}\right) = p^{\left[\frac{N+m}{N}\right]} z_{m+N},$$

where

$$z_{m+N} = \frac{1}{c_0}\left(b_N z_m - \sum_{j=1}^{N-1} p^{\left[\frac{m+j}{N}\right]-\left[\frac{m}{N}\right]} b_{N-j} z_{m+j}\right) \in \mathbf{Z}_p.$$

Thus, the representation (14) holds and the series in the right-hand side of (10) converges in the topology of $\mathbf{Z}_p$.

Substituting (10) into Eq. (1), we obtain from (11)–(13) that the sequence x_n is a solution of Eq. (1) over $\mathbf{Z}_p$:

$$\sum_{j=0}^{N-1} c_j x_{j+n} - c_N x_{n+N} = \frac{1}{c_0}\sum_{j=0}^{N-1}\sum_{k=0}^{\infty} c_j f_{j+n+k} y_{k+N-1} - \frac{c_N}{c_0}\sum_{k=0}^{\infty} f_{n+k+N} y_{k+N-1} =$$

$$= \frac{1}{c_0}\sum_{j=0}^{N-1}\sum_{k=j}^{\infty} c_j f_{n+k} y_{N+k-j-1} - \frac{c_N}{c_0}\sum_{k=N}^{\infty} f_{n+k} y_{k-1} =$$

$$= \frac{1}{c_0}\sum_{j=0}^{N-1}\sum_{k=j}^{\infty} c_j f_{n+k} y_{N+k-j-1} - \frac{1}{c_0}\sum_{k=N}^{\infty}\sum_{j=1}^{N} c_{N-j} y_{j+k-1} f_{n+k} =$$

$$= \frac{1}{c_0}\sum_{j=0}^{N-1}\sum_{k=j}^{\infty} c_j f_{n+k} y_{N+k-j-1} - \frac{1}{c_0}\sum_{k=N}^{\infty}\sum_{j=0}^{N-1} c_j y_{k+N-1-j} f_{n+k} =$$

$$= \frac{1}{c_0}\sum_{j=0}^{N-1}\sum_{k=j}^{N-1} c_j f_{n+k} y_{N+k-j-1} = \frac{1}{c_0}\sum_{k=0}^{N-1}\sum_{j=0}^{k} c_j y_{N+k-j-1} f_{n+k} =$$

$$= \frac{1}{c_0}\sum_{k=0}^{N-1}\sum_{j=0}^{k} c_{k-j} y_{N+j-1} f_{n+k} = y_{N-1} f_n + \frac{1}{c_0}\sum_{k=1}^{N-1}\sum_{j=0}^{k} c_{k-j} y_{N+j-1} f_{n+k} =$$

$$= y_{N-1} f_n + \frac{1}{c_0} \sum_{k=1}^{N-1} \left(c_N y_{k-1} - \sum_{j=1}^{N-k-1} c_{N-j} y_{k-1+j} \right) f_{n+k} = f_n, \ n = 0, 1, 2, ...$$

Now we prove the uniqueness of this solution. For this purpose it suffices to prove that the homogeneous equation (2) has only the trivial solution $x_n = 0$ over $\mathbf{Z}_p$. It follows from (2) that

$$x_n = \frac{1}{c_0} \left(c_N x_{n+N} - \sum_{j=1}^{N-1} c_j x_{n+j} \right) =$$

$$= \frac{p}{c_0} \left(b_N x_{n+N} - \sum_{j=1}^{N-1} b_j x_{n+j} \right), \ n = 0, 1, 2, ...,$$

where $b_j = \frac{c_j}{p} \in \mathbf{Z}$, $j = 1, ..., N$. Hence, for any $m \in \mathbf{N}$, $n = 0, 1, 2, ...$ there exists a number $z_{mn} \in \mathbf{Z}_p$ such that $x_n = p^m z_{mn}$. Therefore, $x_n = 0$, $n = 0, 1, 2, ...$ The proof is complete.

Remark 1 In the case of the second order difference equation

$$c_2 x_{n+2} = c_1 x_{n+1} + c_0 x_n - f_n, \quad n = 0, 1, 2, ... \tag{15}$$

Formula (10) was obtained in [6, Formula (10)] by the more complicated method as the corollary of an explicit formula for a solution of Eq. (1)

$$x_n = \sum_{k=0}^{\infty} \left(\frac{\lambda_1^{k+1} - \lambda_2^{k+1}}{\lambda_1 - \lambda_2} \right) \frac{(-1)^k c_2^k}{c_0^{k+1}} f_{n+k}, \quad n = 0, 1, 2, ...,$$

where λ_1, λ_2 are the different roots of the characteristic polynomial $c_2\lambda^2 - c_1\lambda - c_0$. If this characteristic polynomial has a multiple root $\lambda_1 = \lambda_2$ (for example, $c_2 = 9, c_1 = 6, c_0 = -1, p = 3$). then Formula (10) is a corollary of the following explicit formula for the unique solution of Eq. (15) over $\mathbf{Z}_p$ [6, Formula 25]:

$$x_n = \sum_{k=0}^{\infty} \frac{(-1)^k (k+1)}{c_0^{k+1}} \left(\frac{c_1}{2} \right)^k f_{n+k}.$$

Note that in this case $\frac{c_1}{2} \in \mathbf{Z}_p$ for $p \neq 2$. If $p = 2$, then the number 2 is a common divisor of c_2 and c_1, and thus $\frac{c_1}{2} \in \mathbf{Z}$. Moreover, $c_1^2 + 4c_0c_1 = 0$ because we have a multiple root. Hence $\left(\frac{c_1}{2}\right)^2 = -c_0c_1$, i.e. c_1 is divisible by 4.

4 Cramer's Rule

We regard Eq. (1) as an infinite system of linear equations over $\mathbf{Z}_p$. Let the assumptions of Lemma 1 hold. Since c_0 is not divided by p, after dividing Eq. (1) by c_0 we obtain the equivalent system over $\mathbf{Z}_p$

$$\frac{c_N}{c_0}x_{n+N} = \frac{c_{N-1}}{c_0}x_{n+N-1} + \cdots + \frac{c_1}{c_0}x_{n+1} + x_n - \frac{f_n}{c_0}, \quad n = 0, 1, 2, ..., \tag{16}$$

By Lemma 1, this system has a unique solution over $\mathbf{Z}_p$. We consider elements of the set $S(\mathbf{Z}_p)$ of all sequences $x = \{x_n\}_{n=0}^{\infty}$ from $\mathbf{Z}_p$ as column vectors. We write (16) in the operator-vector form

$$\mathcal{A}x = \frac{f}{c_0}, \quad \mathcal{A} = \begin{pmatrix} 1 & c_1c_0^{-1} & c_2c_0^{-1} & \dots & c_{N-1}c_0^{-1} & -c_Nc_0^{-1} & 0 \\ 0 & 1 & c_1c_0^{-1} & \dots & c_{N-2}c_0^{-1} & c_{N-1}c_0^{-1} & -c_Nc_0^{-1} \\ 0 & 0 & 1 & \dots & c_{N-3}c_0^{-1} & c_{N-2}c_0^{-1} & c_{N-1}c_0^{-1} \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \ddots \end{pmatrix}, \ x \in S(\mathbf{Z}_p), \tag{17}$$

where elements of the column vector $f = \{f_n\}_{n=0}^{\infty}$ are integers. Let $\mathcal{A}_n$ be the matrix obtained from the matrix $\mathcal{A}$ by replacing the n-th column with the vector $\frac{f}{c_0}$ ($n = 0, 1, 2, ...$), i.e.

$$\mathcal{A}_0 = \begin{pmatrix} f_0c_0^{-1} & c_1c_0^{-1} & c_2c_0^{-1} & \dots & c_{N-1}c_0^{-1} & -c_Nc_0^{-1} & 0 \\ f_1c_0^{-1} & 1 & c_1c_0^{-1} & \dots & c_{N-2}c_0^{-1} & c_{N-1}c_0^{-1} & -c_Nc_0^{-1} \\ f_2c_0^{-1} & 0 & 1 & \dots & c_{N-3}c_0^{-1} & c_{N-2}c_0^{-1} & c_{N-1}c_0^{-1} \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \ddots \end{pmatrix},$$

$$\mathcal{A}_1 = \begin{pmatrix} 1 & f_0c_0^{-1} & c_1c_0^{-1} & c_2c_0^{-1} & \dots & c_{N-1}c_0^{-1} & -c_Nc_0^{-1} & 0 & \dots \\ 0 & f_1c_0^{-1} & 1 & c_1c_0^{-1} & \dots & c_{N-2}c_0^{-1} & c_{N-1}c_0^{-1} & -c_Nc_0^{-1} & \dots \\ 0 & f_2c_0^{-1} & 0 & 1 & \dots & c_{N-3}c_0^{-1} & c_{N-2}c_0^{-1} & c_{N-1}c_0^{-1} & \dots \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \dots .$$

We denote by Δ_m (respectively $\Delta_{n,m}$) the $(m+1)$th order leading principal minor of the matrix $\mathcal{A}$ (respectively $\mathcal{A}_n$), $m, n = 0, 1, 2, ...$ The following assertion shows that a unique solution over $\mathbf{Z}_p$ can be found by an analogue of the Cramer's rule.

Theorem 3 *Let the assumptions of Lemma 1 hold. Then Eq. (1) has a unique solution over $\mathbf{Z}_p$. This solution may be found by the following Cramer's rule:*

$$x_n = \frac{\det \mathcal{A}_n}{\det \mathcal{A}}, \quad n = 0, 1, 2, ..., \tag{18}$$

where the determinants of $\mathcal{A}$, $\mathcal{A}_n$ can be defined as limits in $\mathbf{Z}_p$ of the sequence of leading principal minors of these matrices, i.e.

$$\det \mathcal{A} = \lim_{m\to\infty} \Delta_m, \quad \det \mathcal{A}_n = \lim_{m\to\infty} \Delta_{n,m}. \tag{19}$$

Proof By Lemma 1 Eq. (1) has a unique solution over $\mathbf{Z}_p$. Without loss of generality we can assume that $c_0 = 1$. We show that Formula (10) for finding a solution of (1) can be regarded as a collection of Cramer's formulas for solving the infinite system of linear equations (17). We note that $\Delta_m = 1$. Therefore $\det \mathcal{A} = 1$. Consider the sequences of the leading principal minors of $\mathcal{A}_0$:

$$\Delta_{0,0} = f_0,\ \Delta_{0,k} = \begin{vmatrix} f_0 & c_1 & c_2 & \dots & c_k \\ f_1 & 1 & c_1 & \dots & c_{k-1} \\ f_2 & 0 & 1 & \dots & c_{k-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_k & 0 & 0 & \dots & 1 \end{vmatrix}, \quad k = 1, ..., N-1,$$

$$\Delta_{0,N} = \begin{vmatrix} f_0 & c_1 & c_2 & \dots & c_{N-1} & -c_N \\ f_1 & 1 & c_1 & \dots & c_{N-2} & c_{N-1} \\ f_2 & 0 & 1 & \dots & c_{N-3} & c_{N-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ f_N & 0 & 0 & \dots & 0 & 1 \end{vmatrix}, ...,$$

$$\Delta_{0,k} = \begin{vmatrix} f_0 & c_1 & c_2 & c_3 & \dots & c_{N-1} & -c_N & 0 & \dots & 0 \\ f_1 & 1 & c_1 & c_2 & \dots & c_{N-2} & c_{N-1} & -c_N & \dots & 0 \\ f_2 & 0 & 1 & c_1 & \dots & c_{N-3} & c_{N-2} & c_{N-1} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ f_k & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{vmatrix}, \quad k = N+1, N+2, ...$$

Fix $m = N, N+1, ...$ We show that

$$\Delta_{0,m} = \sum_{j=0}^{m} y_{j+N-1} f_j. \tag{20}$$

Denote $g_k = f_{m-k}, \quad k = 0, ..., m$ and consider the determinants

$$B_0 = g_0, \quad B_k = \begin{vmatrix} g_k & c_1 & c_2 & \dots & -c_k \\ g_{k-1} & 1 & c_1 & \dots & c_{k-1} \\ g_{k-2} & 0 & 1 & \dots & c_{k-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ g_0 & 0 & 0 & \dots & 1 \end{vmatrix}, \quad k = 1, ..., N-1,$$

$$B_N = \begin{vmatrix} g_N & c_1 & c_2 & \dots & c_{N-1} & -c_N \\ g_{N-1} & 1 & c_1 & \dots & c_{N-2} & c_{N-1} \\ g_{N-2} & 0 & 1 & \dots & c_{N-3} & c_{N-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ g_0 & 0 & 0 & \dots & 0 & 1 \end{vmatrix} \quad \text{and}$$

$$B_k = \begin{vmatrix} g_k & c_1 & c_2 & c_3 & \dots & c_{N-1} & -c_N & 0 & \dots & 0 \\ g_{k-1} & 1 & c_1 & c_2 & \dots & c_{N-2} & c_{N-1} & -c_N & \dots & 0 \\ g_{k-2} & 0 & 1 & c_1 & \dots & c_{N-3} & c_{N-2} & c_{N-1} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ g_0 & 0 & 0 & 0 & 0 & \dots 0 & 0 & 0 & \dots & 1 \end{vmatrix}, \quad k = N+1, N+2, ..., m.$$

Decompose B_k $(k = 0, ..., N-1)$ according to elements of the first row. Then the finite sequence $B_0, B_1, B_2, ..., B_{N-1}$ satisfies the following recurrence relation over $\mathbf{Z}_p$:

$$B_k + c_1 B_{k-1} + \cdots + c_k B_0 = g_k, \quad k = 0, 1, ..., N-1. \tag{21}$$

Now decompose B_k $(k = N, N+1..., m)$ relative to the first row. Then the finite sequence $B_0, B_1, , ..., B_m$ is a solution over $\mathbf{Z}_p$ of the following finite difference equation

$$B_k + c_1 B_{k-1} + \cdots + c_{N-1} B_{k-N+1} - c_N B_{k-N} = g_k, \quad k = N, N+1, ..., m. \tag{22}$$

The initial data $B_0, ..., B_{N-1}$ for the difference equation (22) are defined uniquely from the recurrence relations (21). Then the initial problem (22) with these initial conditions has a unique solution and $B_m = \Delta_{0,m}$. Let us prove that

$$s_k = \sum_{j=0}^{k} y_{j+N-1} g_{k-j}, \quad k = 0, ..., m \tag{23}$$

is a solution of Eq. (22) with initial data $B_0, ..., B_{N-1}$. We have

$$s_k = \sum_{j=0}^{k} g_j y_{k-j+N-1}, \quad k = 0, 1, \dots, m.$$

Then substituting s_k into (22) and taking into account (11) and (12), we find

$$\sum_{l=0}^{N-1} c_l s_{k-l} = \sum_{l=0}^{N-1} c_l \sum_{j=0}^{k-l} g_j y_{k-l-j+N-1} =$$

$$= \sum_{l=0}^{N-1} c_l \sum_{j=0}^{k-N} g_j y_{k-l-j+N-1} + \sum_{l=0}^{N-1} c_l \sum_{j=k-(N-1)}^{k-l} g_j y_{k-l-j+N-1} =$$

$$= \sum_{j=0}^{k-N} g_j \sum_{l=0}^{N-1} c_l y_{k-l-j+N-1} + \sum_{j=k-(N-1)}^{k} g_j \sum_{l=0}^{k-j} c_l y_{k-l-j+N-1} =$$

$$= c_N \sum_{j=0}^{k-N} g_j y_{k-j-1} + \sum_{j=k-(N-1)}^{k-1} g_j (c_N y_{k-j-1} - \sum_{l=k-j+1}^{N-1} c_l y_{k-l-j+N-1}) +$$

$$+ c_0 y_{N-1} g_k = g_k + c_N \sum_{j=0}^{k-N} g_j y_{k-j-1} = g_k + s_{k-N} c_N, \quad k = N, N+1, \ldots, m.$$

Substituting s_k into (21), we have $s_0 = g_0 = B_0$ and

$$\sum_{l=0}^{k} c_l s_{k-l} = \sum_{l=0}^{k} c_l \sum_{j=0}^{k-l} g_j y_{k-l-j+N-1} = \sum_{j=0}^{k} g_j \sum_{l=0}^{k-j} c_l y_{k-l-j+N-1} =$$

$$= g_k + \sum_{j=0}^{k-1} g_j \sum_{l=0}^{k-j} c_l y_{k-l-j+N-1} =$$

$$= g_k + \sum_{j=0}^{k-1} g_j (c_N y_{k-j-1} - \sum_{l=k-j+1}^{N-1} c_l y_{k-l-j+N-1}) = g_k, \quad k = 1, \ldots, N-1.$$

Consequently,

$$\Delta_{0,m} = s_m = \sum_{j=0}^{m} y_{j+N-1} f_j, \quad m = N, N+1, N+2, \ldots \tag{24}$$

and the relation (20) holds. By Lemma 1 $\lim_{m\to\infty} s_m = \lim_{m\to\infty} \Delta_{0,m}$ exists in the topology of $\mathbf{Z}_p$. Then $\det \mathcal{A}_0$ is well defined and Formula (10) for x_0 can be written as the Cramer's formula (18) with $n = 0$. Arguing in a similar way, we also find Formula (18) for the remaining components of the solution x_n, $n = 1, 2, \ldots$. The proof is complete.

Remark 2 Under the conditions of Theorem 3 the inverse operator to $\mathcal{A}$ can be found by Formula (10):

$$\mathcal{A}^{-1} = \begin{pmatrix} 1 & y_N & y_{N+1} & y_{N+2} & y_{N+3} & \cdots \\ 0 & 1 & y_N & y_{N+1} & y_{N+2} & \cdots \\ 0 & 0 & 1 & y_N & y_{N+1} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix}.$$

Remark 3 As shown in [8, Example 3.1], if the assumptions of Theorem 3 fail, then even for second order equation $\lim_{m\to\infty} \Delta_{0,m}$ may not exist in the topology of $\mathbf{Z}_p$ for all primes. Furthermore, Eq. (1) can have a unique integer solution and this solution cannot be found by the Cramer's rule (18), (19).

The following theorem yields the sufficient conditions for the existence and the uniqueness of an integer solution to Eq. (1) and the possibility to apply the Cramer's rule for finding this solution.

Theorem 4 *Let c_N and c_0 are co-prime integers and let any prime divisor of c_N divide $c_1, ..., c_{N-1}$. Assume that there exist numbers $x_0,...,x_{N-1} \in \mathbf{Z}$ such that (18) and (19) hold for any prime divisor p of c_N for $n = 0, 1, \ldots, N-1$. Then the implicit equation (1) has a unique integer solution $\{x_n\}_{n=0}^{\infty}$. This solution can be found by the Cramer's rule (18), (19).*

Proof We consider the prime decomposition $c_N = p_1^{l_1} \cdot p_2^{l_2} \cdot \ldots \cdot p_m^{l_m}$. By Theorem 3, Eq. (1) has a unique integer solution $\left\{x_n^{(p_j)}\right\}_{n=0}^{\infty}$ over $\mathbf{Z}_{p_j}$, $j = 1, \ldots, m$. Moreover, $x_n^{(p_j)} = x_n$ for $n = 0, \ldots, N-1, \quad j = 1, 2, \ldots, m$. We show $x_N^{(p_1)} = \cdots = x_N^{(p_m)} \in \mathbf{Z}$. By Eq. (1) we have $c_N x_N^{(p_j)} = c_{N-1}x_{N-1} + \cdots + c_1 x_1 + c_0 x_0 - f_0 \in \mathbf{Z}$. Consequently, $x_N^{(p_1)} = \cdots = x_N^{(p_m)}$ in $\mathbf{Q}$. Let $||z||_p$ be the p-adic norm of $z \in \mathbf{Z}_p$. Since $x_N^{(p_j)} \in \mathbf{Z}_{p_j}$, we have $||x_N^{(p_j)}||_{p_j} \le 1$ and $||c_N x_N^{(p_j)}||_{p_j} = ||c_N||_{p_j} \cdot ||x_N^{(p_j)}||_{p_j} \le p_j^{-l_j}$, $j = 1, \ldots, m$. Hence, an integer number $c_N x_N^{(p_j)}$ is divided by $p_j^{l_j}$ for all $j = 1, \ldots, m$. Thus, this number is divided by c_N. Therefore, $x_N^{(p_j)} \in \mathbf{Z}$, $j = 1, ..., m$. Repeating the above argument, we find $x_n^{(p_1)} = \cdots = x_n^{(p_m)} \in \mathbf{Z}$, where $n = N+1, N+2, \ldots$. The proof is complete.

Remark 4 We give an analogue of Theorem 4 in the case of the first order equation ($N = 1$):

$$c_1 x_{n+1} = c_0 x_n - f_n. \quad n = 0, 1, 2, ... \tag{25}$$

Assume that c_1 and c_0 are co-prime integers and there exists an integer $x_0 \in \mathbf{Z}$ such that the following equality holds

$$x_0 = \sum_{k=0}^{\infty} \frac{c_1^k}{c_0^{k+1}} f_k$$

for any prime divisor p of c_1 in the ring $\mathbf{Z}_p$. Then Eq. (25) has a unique integer solution $\{x_n\}_{n=0}^{\infty}$ and this solution can be found by using the Cramer's rule. This assertion may be proved by same arguments as in the case $N \geq 2$. The case $c_0 = 1$ has been considered in [2, 9].

References

1. Kelley, W.G., Peterson, A.C.: Difference Equation: An Introduction with Applications, 2nd edn., p. 404. Academic Press (2001)
2. Gerasimov, V.A., Gefter, S.L., Goncharuk, A.B.: Application of the p-adic topology on $\mathbf{Z}$ to the problem of finding solutions in integers of an implicit linear difference equation. J. Math. Sci. **235**, 256–261 (2018). https://doi.org/10.1007/s10958-018-4072-x
3. Fomin, S., Zelevinsky, A.: The laurent phenomenon. Adv. Appl. Math. **28**, 119–144 (2002). https://doi.org/10.1006/aama.2001.0770
4. Gefter, S.L., Goncharuk, A.B., Piven', A.L.: Integer solutions for a vector implicit linear difference equation in $\mathbf{Z}^{\mathbf{N}}$. Dopov. Nac. Akad. Nauk Ukr. **11**, 11–18 (in Ukrainian) (2018). https://doi.org/10.15407/dopovidi2018.11.011
5. Martseniuk, V., Gefter, S., Piven', A.: Integer solutions of implicit linear difference equations. In: Voronois Impact on Modern Science, Proceedings of the Sixth International Conference on Anal. Number Theory Spat. Tessellations, vol. 1, pp. 87–95. National Pedagogical Dragomanov University Publication (2018)
6. Hefter, S.L., Martseniuk, V.V., Piven, O.L.: Integer solutions of a second order implicit linear difference equation. Bukovinian Math. J. **6**(3–4), 40–46 (in Ukrainian) (2018). https://doi.org/10.31861/bmj2018.03.040
7. Berestovskii, V.N., Nikonorov, Y.G.: Continued fractions, the group $GL(2, \mathbf{Z})$ and Pisot numbers. Sib. Adv. Math. **17**(4), 268–290 (2007). https://doi.org/10.3103/S1055134407040025
8. Gefter, S.L., Martseniuk, V.V., Goncharuk, A.B., Piven' A.L.: Analogue of the Cramer rule for an implicit linear second order difference equation over the ring of integers. J. Math. Sci. **244**, 601–607 (2020). https://doi.org/10.1007/s10958-019-04635-w
9. Gefter, S., Goncharuk, A.: Generalized backward shift operators on the ring $\mathbf{Z}[[x]]$, Cramer's rule for infinite linear systems, and p-adic integers. Oper. Theory Adv. Appl. **268**, 247–259 (2018). https://doi.org/10.1007/978-3-319-75996-8_13
10. Everest, G., van der Poorten, A., Shparlinski, I., Ward, T.: Recurrence Sequences. American Mathematical Society (2003)
11. Gefter, S.L., Piven, A.L.: Implicit linear nonhomogeneous difference equation in banach and locally convex spaces. J. Math. Phys. Anal. Geom. **15**(3), 336–353 (2019). https://doi.org/10.15407/mag15.03.336
12. Borevich, Z.I., Shafarevich, I.R.: Number Theory. Academic Press, Boston (1966)
13. Lang, S.: Algebra, Rev. 3rd edn. Springer, New York Inc. (2002)

On the Neumann Boundary Optimal Control of a Frictional Quasistatic Contact Problem with Dry Friction

NICOLAE POP, Luige Vladareanu, and Victor Vladareanu

Abstract This paper deals with boundary optimal control problem of a frictional quasi-static contact problem with dry friction, described by a nonlocal version of Coulomb's law. We prove the existence of a boundary optimal control for the regularized problem obtained from a quasi-static contact problem with dry friction. For getting the necessary optimality conditions, we use some regularization techniques leading us to a control problem of a variational equality. The minimizing of cost function is a compromise between energy consumption and the finding of a traction force on the Neumann boundary condition, so that the actual displacement field is as close as possible to the desired displacement field, while the density of body force remain constant and small enough.

Keywords Boundary optimal control problem · Quasi-static contact problem with friction · Regularized state problem

1 Introduction

The quasi-static model of the contact problems with friction, without the inertia effects, was proposed by [9] and consists of the formulation obtained through the approximation with the finite differences of the variational inequality. The proof of the existence and uniqueness is based on the hypothesis that the displacements satisfy some conditions of regularity and the friction coefficient is small enough, see [7, 9]. The static contact problem with friction cannot describe the evolutive state of the contact conditions. For of this reason, the quasi-static formulation, of the

N. POP (✉) · L. Vladareanu · V. Vladareanu
Institute of Solid Mechanics of Romanian Academy, Bucharest, Romania
e-mail: nicolae.pop@imsar.ro

L. Vladareanu
e-mail: luige.vladareanu@yimsar.ro

V. Vladareanu
e-mail: victor.vladareanu@imsar.ro

S. Baigent et al. (eds.), *Progress on Difference Equations and Discrete Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 341,
https://doi.org/10.1007/978-3-030-60107-2_17

contact problem with friction is preferred, which contains a dynamic formulation of the contact conditions and the inertial term is no longer used. Through the temporal discretization of the quasi-static contact problem, the so called incremental problem is obtained, equivalent with a sequence of the static contact problems. Therefore, the quasi-static problem is solved step by step, at each time small deformations and displacements are calculated and are added at those calculated previously, as a result of a few small modifications of the applied forces, of the contact zone and of the contact conditions. Although, at each increment the dependence of the load-way is neglected, this hypothesis takes into account the way the applied forces change (modify themselves). From a mathematical point view, the problem obtained at each step is similar with a static problem. We will describe two methods for solving our contact problem, the first is the primal variational formulation problem, and the second the dual mixed variational formulation problem. The main results where it was demonstrated the existence of a boundary optimal control for the regularized problem obtained from a quasi-static contact problem with dry friction, we recall, [1–3, 8]). For getting the necessary optimality conditions, we use some regularization techniques leading us to a control problem of a variational equality. After describing the classical and variational form of the problem, we will first define the notion of boundary control, optimal pair, optimal control and regularized optimal control, after which we will present the existence of optimal boundary control.

2 Classical and Variational Formulation

Let $\Omega \subset \mathbb{R}^d, d = 2$ or 3, the domain occupied by a linear elastic body with a Lipschitz boundary Γ. Let Γ_1, Γ_2 and Γ_C be three open disjoint parts of Γ such that $\Gamma = \overline{\Gamma}_1 \cup \overline{\Gamma}_2 \cup \overline{\Gamma}_C$, $\overline{\Gamma}_1 \cap \overline{\Gamma}_C = \emptyset$ and mes $(\Gamma_1) > 0$. We assume that the body is subjected to volume forces of density $\mathbf{f} \in (L^2(\Omega))^d$, to surface traction of density $\mathbf{h} \in (L^2(\Gamma_2))^{d-1}d$ and is held fixed on Γ_1. The Γ_C denotes a contact part of boundary where unilateral contact and Coulomb friction condition between Ω and perfectly rigid foundation are considered. We denote by $\mathbf{u} = (u_1, \ldots, u_d)$ the displacement field, $\varepsilon = (\varepsilon_{ij}(\mathbf{u})) = \left(\frac{1}{2}(u_{i,j} + u_{j,i})\right)$ the strain tensor and $\sigma = (\sigma_{ij}(\mathbf{u})) = (a_{ijkl}\varepsilon_{kl}(\mathbf{u}))$ the stress tensor with the usual summation convention, where $i, j, k, l = 1, \ldots, d$. For the normal and tangential components of the displacement vector and stress vector, we use the following notation: $\mathbf{u}_N = u_i \cdot n_i$, $\mathbf{u}_T = \mathbf{u} - \mathbf{u}_N \cdot \mathbf{n}, \sigma_N = \sigma_{ij}u_i n_j, (\sigma_T)_i = \sigma_{ij}n_j - \sigma_N \cdot n_i$, where $\mathbf{n} = (n_i)$ is the outward unit normal vector to Γ. We denote by $g \in C(\bar{\Gamma}_C), g \geq 0$ the initial gap between the body and the rigid foundation and lets us denote by $\mathbf{f}$ and $\mathbf{h}$ the density of body and traction forces, respectively. We assume that $a_{ijkl} \in L^\infty(\Omega)$, $1 \leq i, j, k, l \leq d$, with usual condition of symmetry and elasticity, that is

$$a_{ijkl} = a_{jikl} = a_{klij}, \quad 1 \le i, j, k, l \le d,$$

$$\text{and} \quad \exists\, m_0 > 0,\ \forall\, \xi = (\xi_{ij}) \in \mathbb{R}^{d^2},\ \xi_{ij} = \xi_{ji},\ 1 \le i,\ j \le d,$$

$$a_{ijkl}\, \xi_{ij}\, \xi_{kl} \ge m_0 |\xi|^2 .$$

In this conditions, the fourth-order tensor $\mathbf{a} = (a_{ijkl})$ is invertible a.e., on Ω and if we denote its inverse by $\mathbf{b} = (b_{ijkl})$, we have $\varepsilon_{ij}(\mathbf{u})) = (b_{ijkl}\sigma_{kl}(\mathbf{u}))$, $i, j, k, l = 1, \ldots, d$.

The classical contact problem with dry friction in elasticity, in the particular case, is with the normal stress $\sigma_N(\mathbf{u})$ and Γ_C is assumed known and considered as obeying the normal compliance law, is the following.

Find $\mathbf{u} = \mathbf{u}(x, t)$ such that $\mathbf{u}(0, \cdot) = \mathbf{u}^0(\cdot)$ in Ω and for all $t \in [0, T]$,

$$-\operatorname{div} \sigma(\mathbf{u}) = \mathbf{f}, \quad \text{in } \Omega \tag{1}$$

$$\sigma_{ij}(\mathbf{u}) = a_{ijkl} \cdot \varepsilon_{kl}(\mathbf{u}), \quad \text{in } \Omega \tag{2}$$

$$\mathbf{u} = 0 \quad \text{on } \Gamma_1 \tag{3}$$

$$\sigma \cdot \mathbf{n} = \mathbf{h} \quad \text{on } \Gamma_2, \tag{4}$$

the contact condition:

$$\mathbf{u}_N \le g,\ \sigma_N(\mathbf{u}) \le 0,\ (\mathbf{u}_N - \mathbf{g})\sigma_N(\mathbf{u}) = 0 \quad \text{on } \Gamma_C \tag{5}$$

and Coulomb friction on Γ_C:

$$\|\sigma_T(\mathbf{u})\| \le \mu_F |\sigma_N(\mathbf{u})|, \tag{6}$$

such that:

$-$if $\quad \|\sigma_T(\mathbf{u})\| < \mu_F |\sigma_N(\mathbf{u})| \Rightarrow \mathbf{u}_T = 0$

$-$if $\quad \|\sigma_T(\mathbf{u})\| = \mu_F |\sigma_N(\mathbf{u})| \Rightarrow\ \exists \lambda \ge 0,$ such that $\dot{\mathbf{u}}_T = -\lambda \sigma_T$

where $\mathbf{u}^0$ denotes the initial displacement of the body. Supposing that a positive coefficient $\mu_F \in L^\infty(\Gamma_C)$, $\mu_F \ge \mu_0$ a.e. on Γ_C of Coulomb friction is given, we introduce the space of virtual displacements

$$V = \left\{ \mathbf{v} \in (H^1(\Omega))^d | \mathbf{v} = 0 \text{ on } \Gamma_1 \right\}$$

and its convex subset of kinematically admissible displacements

$$K = \{ v_N \in V | v_N \equiv v \cdot n \le g \text{ on } \Gamma_C \}.$$

We assume that the normal force on Γ_C is known (as normal compliance) so that one can evaluate the non-negative slip bound $p \in L^\infty(\Gamma_C)$ as a product of the friction

coefficient and the normal stress, i.e. $p = \mu_F \lambda_1$, when λ_1 is the normal stress. We assume that normal interface response (the normal compliance law) is:

$$\sigma_N(\mathbf{u}) = -c_N(\mathbf{u}_N - \mathbf{g})^{m_N}$$

where c_N and m_N are material constant depending on interface properties. **Problem** $(\mathbf{P_1})$ *Find* $\mathbf{u} \in K$ such that $J(\mathbf{u}) = \min_{\mathbf{v} \in K} J(\mathbf{v})$.

The minimized functional representing the total potential energy of the body has the form:

$$J(\mathbf{v}) = \frac{1}{2} a(\mathbf{v}, \mathbf{v}) - L(\mathbf{v}) + \overline{j}(\mathbf{v})$$

where:

- the bilinear form a is given by

$$a(\mathbf{v}, \mathbf{w}) = \int_\Omega a_{ijkl} \varepsilon_{ij}(\mathbf{v}) \varepsilon_{kl}(\mathbf{w}) dx$$

- linear functional L is given by:

$$L(\mathbf{v}) = \int_\Omega \mathbf{f}\mathbf{v} dx + \int_{\Gamma_2} \mathbf{h}\mathbf{v} ds;$$

- the sublinear functional $\overline{j}$ is given by:

$$\overline{j}(\mathbf{u}, \mathbf{v}) = \int_{\Gamma_C} p\, |\mathbf{v}_T|\, ds + \int_{\Gamma_C} c_N (\mathbf{u} - \mathbf{g})^{m_n} \mathbf{v}_N ds$$

where $\mathbf{v}_T \in (L^\infty(\Gamma_C))^{d-1}$ denotes the tangent vector to boundary Γ.

It is known that the problem $(\mathbf{P_1})$ is non-differentiable due to the sublinear term $\overline{j}$, and has a unique solution [6].

The variational formulation, in the quasi-static case, is equivalent to the quasi-variational inequality:

Problem $(\mathbf{P_2})$ *Find* $\mathbf{u}(x,t) \in K \times [0,T]$ s.t. $a(\mathbf{u}, \mathbf{v} - \dot{\mathbf{u}}) + \overline{j}(\mathbf{v} - \dot{\mathbf{u}}) \geq (L, \mathbf{u} - \dot{\mathbf{v}})$, $\forall \mathbf{v} \in K, \forall t \in [0,T], T > 0$, *with initial conditions* $\mathbf{u}(x,0) = u_0, \dot{\mathbf{u}}(x,0) = u_1$.

The existence and uniqueness of the solution of this quasi-variational inequality are proven under the assumption that μ_F is sufficiently small and $mes(\Gamma_0) > 0$ [4].

The Lagrangian formulation of the problem $(\mathbf{P_1})$ is given by introducing $L : V \times \Lambda_1 \times \Lambda_2 \to \mathbb{R}$, with

$$L(v, \mu_1, \mu_2) = \frac{1}{2} a(v, v) - L(v) + \langle \mu_1, v_N - g \rangle + \int_{\Gamma_C} \mu_2 v_T ds$$

where $\Lambda_1 = \{\mu_1 \in H^{-\frac{1}{2}}(\Gamma_C) | \mu_1 \geq 0\}$, $\Lambda_2 = \{\mu_2 \in L^\infty(\Gamma_C) |\ |\mu_2| \leq p \text{ on } \Gamma_C\}$.

The space $H^{-\frac{1}{2}}(\Gamma_C)$ is the dual of

$$H^{\frac{1}{2}}(\Gamma_C) = \{\gamma \in L^2(\Gamma_C) | \exists v \in V \ s.t. \ \gamma = v_N \text{ on } \Gamma_C\}$$

and the ordering $\mu_1 \geq 0$ means, in the variational form, that $\langle \mu_1, v_N - g \rangle \leq 0, \ \forall v \in K$, where $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $H^{-\frac{1}{2}}(\Gamma_C)$ and $H^{\frac{1}{2}}(\Gamma_C)$. Since $L^2(\Gamma_C)$ is dense in $H^{-\frac{1}{2}}(\Gamma_C)$, the duality pairing $\langle \cdot, \cdot \rangle$ is represented by a scalar product in $L^2(\Gamma_C)$.

Another approach is a mixed formulation. Mixed formulation is required by modern numerical solving techniques to solve contact problems. To this purpose, we will approach the formulation of the mixed contact issue with help the saddle point problem, using Lagrange multipliers.

The Lagrange multipliers $\mu_1, \ \mu_2$ are considered as functionals on the contact part of the boundary Γ. It is important that the Lagrange multipliers do have mechanical significance: while the first one is related to the non-penetration conditions and represents the normal stress, the second one removes the non-differentiability of the sublinear functional

$$j_2(v) = \sup_{\mu_2 \in \Lambda_2} \int_{\Gamma_C} \mu_2 v_T ds$$

and represents the tangential stress.

The equivalence between the problem $(\mathbf{P_1})$ and the lagrangian formulation is given by:

$$\inf_{v \in K} J(v) = \inf_{v \in V} \sup_{\mu_1 \in \Lambda_1, \mu_2 \in \Lambda_2} L(v, \mu_1, \mu_2).$$

By the mixed variational formulation of the problem $(\mathbf{P_1})$ we mean a saddle point problem:

Problem $(\mathbf{P_3})$. *Find*

$$(w, \lambda_1, \lambda_2) \in V \times \Lambda_1 \times \Lambda_2 \ \textit{such that}$$

$$L(w, \mu_1, \mu_2) \leq L(w, \lambda_1, \lambda_2) \leq L(v, \lambda_1, \lambda_2), \quad \forall (v, \mu_1, \mu_2) \in V \times \Lambda_1 \times \Lambda_2.$$

It is known that $(\mathbf{P_3})$ has a unique solution [4] and its first component $w = u \in K$ solves $(\mathbf{P_1})$ and the Lagrange multipliers $\lambda_1, \ \lambda_2$ represent the normal and tangential contact stress on the contact part of the boundary, respectively.

Remarks.

1^0. For the contact problem with Coulomb friction, we use the formula $p \equiv \mu_F \lambda_1$, for the slip bound on the contact boundary Γ_C, where $\lambda_1 \equiv \ \lambda_1(p)$ is the normal stress on Γ_C and μ_F is the coefficient of friction. Unfortunately this problem cannot be solved as a convex quadratic programming problem because p is an a priori parameter in $(\mathbf{P_3})$, while λ_1 is an a posteriori one.

2^0. Because we can consider the mapping $\Psi : \Lambda_1 \to \Lambda_1,\ \Psi : p \to \lambda_1 \equiv \lambda_1(p)$ defined by the second component of the solution for the contact problem with given friction ($\mathbf{P_3}$), the solution of the contact problem with Coulomb friction will be defined as a fixed point of this mapping in Λ_1. Results concerning the existence of fixed points for sufficiently small friction coefficients may be found in [6].

3 The Time Discretization of the Contact Problems with Coulomb Friction

Let us consider a partition $(t^0, t^1, \ldots, t^N)$ of time interval $[0, T]$ and also the incremental formulation obtained by using the backward finite difference approximation of the time derivative of $\mathbf{u}$.

If we use $u_h^k = u_h(x, t^k),\ \Delta u_h^k = u_h^{k+1} - u_h^k,\ \Delta t^k = t^{k+1} - t^k,\ \dot{u}_h(t^{k+1}) = \Delta u_h^k / \Delta t$, $f_h^k = f_h(k\Delta t)$, for $k = 0, 1, \ldots, N-1$ where $\Delta t = \frac{T}{N}$, we obtain, at each time t^k, the following quasi-variational inequality

$$\begin{aligned} &find\ \ \Delta u_h^k \in V_h \text{ s.t.} \\ &a(\Delta u_h^k, v_h - \Delta u_h^k) + \overline{j}(u_h^k + \Delta u_h^k, v_h - \Delta u_h^k) \geq \\ &\geq \Delta L^k(v_h - \Delta u_h^k) - F(u_h^k, v_h - \Delta u_h^k),\ \forall\, v_h \in K_h \end{aligned} \tag{7}$$

where $F(u_h^k, v_h - \Delta u_h^k) = a(u_h^k, v_h - \Delta u_h^k) - L^k(v_h - \Delta u_h^k)$.

The time discretization of the problem ($\mathbf{P_2}$) follows. For a given load history the quasi-static problem is approximated by a sequence of incremental problems (7); although every problem (7) is a static one, it requires appropriate updating of the displacements, so loads for each increment and so we obtain the following sequence:

Problem($\mathbf{P_2^{ht}}$). *Find* $\mathbf{u} \in \mathbf{K_h}$, *for each time* t^k *such that* $\mathbf{J}(\mathbf{u}) = \min_{\mathbf{v} \in \mathbf{K_h}} \mathbf{J}(\mathbf{v})$,

where $\mathbf{u} \equiv \Delta u_h^k$, $\mathbf{v} \equiv v_h$, $J(\mathbf{v}) = \frac{1}{2}\mathbf{v}^T\mathbf{K}\mathbf{v} - \mathbf{v}^T\mathbf{f} + \mathbf{p}^T|\mathbf{T}\mathbf{v}|$ and $K_h = \{\mathbf{v} \in \mathbb{R}^n | \mathbf{N}\mathbf{v} \leq \mathbf{g}\}$. Here, we by denote $\mathbf{K} \in \mathbb{R}^{n \times n}$ the positive definite stiffness matrix, $\mathbf{f} \in \mathbb{R}^n$ is the load vector, $\mathbf{p} \in \mathbb{R}^m$ is the nodal slip bounds vector for contact nodes. The matrices $\mathbf{N}, \mathbf{T} \in \mathbb{R}^{m \times n}$ contain the rows of the normal and tangential vectors in the contact nodes, respectively, and $\mathbf{g} \in \mathbb{R}^m$ is the vector of distances between the contact nodes and the rigid foundation.

The *matrix form of the Lagrangian* for the problem (P_2^{ht}), at each time t^k is:

$$L(\mathbf{v}, \mu_1, \mu_2) = \frac{1}{2}\mathbf{v}^T\mathbf{K}\mathbf{v} - \mathbf{f}^T\mathbf{v} + \mu_2^T\mathbf{T}\mathbf{v} + \mu_1^T(\mathbf{N}\mathbf{v} - \mathbf{g})$$

where $\mu_1 \in \Lambda_1$, $\mu_2 \in \Lambda_2$ are the Lagrange multipliers and

$$\Lambda_1 = \{\mu_1 \in \mathbb{R}^m | \mu_1 \geq \mathbf{0}\},\ \Lambda_2 = \{\mu_2 \in \mathbb{R}^m || \mu_2| \leq \mathbf{p}\}.$$

The algebraic mixed formulation of (P_2^{ht}) is:
Find $(\mathbf{v}, \mu_1, \mu_2) \in \mathbb{R}^n \times \Lambda_1 \times \Lambda_2$ such that

$$\mathbf{Ku} = \mathbf{f} - \mathbf{N}^T\lambda_1 - \mathbf{T}^T\lambda_2 \tag{8}$$

$$(\mathbf{Nu} - \mathbf{g})^T(\lambda_1 - \mu_1) + \mathbf{u}^T\mathbf{T}^T(\lambda_2 - \mu_2) \geq 0, \ (\mu_1, \mu_2) \in \Lambda_1 \times \Lambda_2. \tag{9}$$

After computing $\mathbf{u}$ from (8) and substituting u into (9), we obtain the *algebraic dual formulation*, for each time t^k, i.e.,

$$\min\left\{\frac{1}{2}\lambda^T\mathbf{A}\lambda - \lambda^T\mathbf{B}\right\} \ s.t. \ \lambda_1 \geq 0, \ \ |\lambda_1| \leq \mathbf{g}, \ \lambda = (\lambda_1^T, \lambda_2^T)^T, \tag{10}$$

where

$$\mathbf{A} = \begin{pmatrix} \mathbf{NK^{-1}N}^T & \mathbf{NK}^{-1}\mathbf{T}^T \\ \mathbf{TK}^{-1}\mathbf{N}^T & \mathbf{TK}^{-1}\mathbf{T}^T \end{pmatrix} \text{ and } \mathbf{B} = \begin{pmatrix} \mathbf{NK}^{-1}\mathbf{f} - \mathbf{g} \\ \mathbf{TK}^{-1}\mathbf{f} \end{pmatrix}.$$

The problem (10) is a *quadratic programming* problem that can be solved by several efficient algorithms.

4 Boundary Optimal Control Problem Formulation

For our contact problem, boundary optimal control problem consists: Let a fixed function $\mathbf{f} \in L^2(\Omega)^d$, we present the following *state problem*:
Problem $(\mathbf{SP_1})$. *Let a given function* $\mathbf{h} \in L^2(\Gamma_2)^{d-1}$, *called control.*
Find $\mathbf{v} \in V$, *such that*: $a(\mathbf{u}, \mathbf{v} - \mathbf{u}) + \overline{j}(\mathbf{u}, \mathbf{v}) - \overline{j}(\mathbf{u}, \mathbf{u}) \geq (L, \mathbf{u} - \mathbf{v}) \ \ \forall \mathbf{v} \in V$

Using the result from problem $(\mathbf{P_1})$, for all $\mathbf{h} \in L^2(\Gamma_2)^{d-1}$, the state problem $(\mathbf{SP_1})$ has a unique solution $\forall \mathbf{v} \in V$, $\mathbf{u} = \mathbf{u}(\mathbf{h})$. Now we will define the following functional: $\overline{J} : L^2(\Gamma_2)^{d-1} \times V \to \mathbb{R}$, with

$$\overline{J}(\mathbf{h}, \mathbf{u}) = \frac{\alpha}{2}\| \mathbf{u} - \mathbf{u}_d\|_V + \frac{\beta}{2}\| \mathbf{h}\|_{L^2(\Gamma_2)^{d-1}} \tag{11}$$

We denote:

$$V_{ad} = \left\{[\mathbf{u}, \mathbf{h}] | [\mathbf{u}, \mathbf{h}] \in V \times L^2(\Gamma_2)^{d-1}, s.t. \ (SP1) \ is \ verified\right\},$$

where α and β are two positive constants, and u_d is the desired target function, taking into account that we are studying a control that acts on the boundary Γ_2, so that the resulting stress σ is as close as possible to the desired target $\sigma_d = (\sigma_{ij}(\mathbf{u}_d)) = (a_{ijkl}\varepsilon_{kl}(\mathbf{u}_d))$.

The optimal control problem is:
Problem $(\mathbf{OC_1})$. *Find* $[\overline{\mathbf{u}}, \overline{\mathbf{h}}] \in V_{ad}$ *s.t.*

$$\overline{J}(\overline{\mathbf{h}}, \overline{\mathbf{u}}) = \min_{[\mathbf{u},\mathbf{h}]\in V_{ad}} \overline{J}(\mathbf{h}, \mathbf{u}),$$

with this notation, a solution of the problem $(\mathbf{OC_1})$ is called an optimal paire, and the second component is called an *optimal control*.

In these hypotheses and see [2], the following is the next theorem:

Theorem 1 *Problem* $(\mathbf{OC_1})$ *has at least one solution* $[\overline{\mathbf{u}}, \overline{\mathbf{h}}]$.

5 The Regularized State Problem for a Boundary Optimal Control Problem

The first step for obtaining the optimal control algorithm is the regularization of the nondifferentiable friction functional $\overline{j}$. For this purpose, we will estimate the functional $\overline{j}$ by a family of regularized functional j_r, which are convex and differentiable in the second argument

$$j_r : V \times V \to \mathbb{R},\ \ j_r(\mathbf{u}, \mathbf{v}) = \int_{\Gamma_C} p\psi_r(\mathbf{v}_T)ds + \int_{\Gamma_C} c_N(\mathbf{u} - \mathbf{g})^{m_n}\mathbf{v}_N ds$$

where the function $\psi_r : (L^2(\Gamma_C))^{d-1} \to L^2(\Gamma_C)$ represents an approximation of the modulus function, $|\cdot| : (L^2(\Gamma_C))^{d-1} \to L^2(\Gamma_C)$ and it can be defined in many other ways, for example, for $r > 0$, $\xi \in (L^2(\Gamma_C))^{d-1}$ and $x \in \Gamma_C$

$$\psi_r(\xi) = \begin{cases} r\left|\frac{\xi}{r}\right|^2\left(1 - \frac{1}{3}\left|\frac{\xi}{r}\right|\right), & \text{if } |\xi(x)| \le r \\ \varepsilon\left(\left|\frac{\xi}{r}\right| - \frac{1}{3}\right), & \text{if } |\xi(x)| > r. \end{cases}$$

The most common example for a regularization function is
$\psi_r(\xi) = \sqrt{\|\xi\|^2 + r}$, *or* $\psi_r(\xi) = \sqrt{\|\xi\|^2 + r^2} - r$.

Now we can replace the state problem $(\mathbf{SP_1})$ and see [7], we have the following regularized state problem:
Problem $(\mathbf{RSP_1})$. *Let a given function* $\mathbf{h} \in L^2(\Gamma_2)^{d-1}$, *called regularized control. Find* $\mathbf{v} \in V$, *such that*: $a(\mathbf{u}, \mathbf{v} - \mathbf{u}) + j_r(\mathbf{u}, \mathbf{v}) - j_r(\mathbf{u}, \mathbf{u}) \ge (L, \mathbf{u} - \mathbf{v})\ \ \forall \mathbf{v} \in V$.

The regularized state problem $(\mathbf{RSP_1})$ has a unique solution $\mathbf{u}_r \in V$ that depends of the Lipschitz continuously on the linear functional L, see [7].

From the presented assumtions result: for $\forall v \in L^2(\Gamma_2)^{d-1}$, the problem $(\mathbf{RSP_1})$ has a unique solution $\forall \mathbf{v} \in V$, $\mathbf{u} = \mathbf{u}(\mathbf{h})$.

If we denote

$$V_{ad}^r = \left\{[\mathbf{u}, \mathbf{h}] | [\mathbf{u}, \mathbf{h}] \in V \times L^2(\Gamma_2)^{d-1},\ s.t.\ (RSP1)\ \ is\ \ verified\right\},$$

using the regularized functional j_r, we can introduce the regularized optimal control problem:

Problem (**RSP$_2$**). *Find* $[\mathbf{u}^*, \mathbf{h}^*] \in V_{ad}^r$ *s. t.*

$$\overline{J}_r(\mathbf{u}^*, \mathbf{h}^*) = \min_{[\mathbf{u},\mathbf{h}] \in V_{ad}^r} \overline{J}_r(\mathbf{h}, \mathbf{u}).$$

With these we can affirm:

Theorem 2 *Problem* (RSP_2) *has at least one solution* $[\mathbf{u}^*, \mathbf{h}^*]$.

The solution of the problem (**RSP$_2$**) is called a regularized optimal pair and the second component $\mathbf{h}^*$ is called a *regularized optimal control*.

6 Conclusions

A classical control problem consists in finding a *control function* $\mathbf{h} \in (L^2(\Gamma_2))^{d-1}$ which minimizes the *cost functional* (3.1). The function $\mathbf{f}$ is the given density of body force, and $\mathbf{h}$ is traction force, which, in the Neumann boundary optimal control is a *control variable*. The second term of the *cost functional* is proportional to the consumed energy. The minimizing of $\overline{J}$ is a compromise between energy consumption and the finding of a traction force on the Neumann boundary condition $\mathbf{h}$, so that the actual displacement field $\mathbf{u}$ is as close as possible to the desired displacement field $\mathbf{u}_d$, while the stresses inside the body remain constant, small enough.

We prove the existence of a boundary optimal control for the regularized problem obtained from a quasi-static contact problem with dry friction. For getting the necessary optimality conditions, we use some regularization techniques leading us to a control problem of a variational equality.

One of the applications of this problem is the analysis of the dynamic systems with friction, which model and control the movement with friction of mobile robots. The structural components of these robots are considered deformable, not rigid, so frictional contact can be modeled much more correctly. The optimal control problem will minimize the effort made by the traction force on the Neumann boundary condition and the difference between the desired displacement field and the current displacement field of the structural components of the robot.

References

1. Amassad, A., Chenais, D., Fabre, C.: Optimal control of an elastic problem involving Tresca friction law. Nonlinear Anal. **48**, 1107–1135 (2002)
2. Capatina, A., Timofte, C.: Boundary optimal control for quasistatic bilateral frictional contact problems. Nonlinear Anal. Theory Method. Appl. **94**, 84–99 (2014)
3. Glowinski, R., Lions, J., Tremolieres, R.: Numerical Analysis of Variational Inequalities. Amsterdam, New York, Oxford, North Holland (1981)
4. Ju, J.W., Taylor, R.L.: A perturbed lagrangian formulation for the finite element solution of nonlinear frictional contact problem. J. de Mécanique Theoretique et Appliquée, Special Issue, suppl. **7**, 1–14 (1998)
5. Klarbring, A., Mikelic, A., Shillor, M.: Global existence result for the quasistatic frictional contact problem with normal compliance. In: Unilateral Problems in Structural Analysis IV (Capri. Birkhäuser, vol. 1991, pp. 85–111 (1989)
6. Matei, A., Micu, S.: Boundary optimal control for nonlinear antiplane problems. Non-linear Anal. Theory Method. Appl. **7**4(5), 16411652 (2011). ISSN 0362-546X. https://doi.org/10.1016/j.na.2010.10.034
7. Pop, N.: On the convergence of the solution of the quasi-static contact problems with friction using the Uzawa type alorithm, Studia Univ. "Babes-Bolyai", Mathematica, vol. XLVIII, no. 3, pp. 125–132 (2004)
8. Rocca, R., Cocou, M.: Existence and approximation of a solution to quasi-static Signorini problem with local friction. Int. J. Eng. Sci. **39**, 1253–1258 (2001)
9. Rocca, R., Cocou, M.: Numerical analysis of quasi-static unilateral contact problems with local friction. Siam J. Numer. Anal. **39**(4), 1324–1342 (2001)
10. Sofonea, M., Matei, A.: Variational inequalities with applications. a study of antiplane frictional contact problems. Adv. Mech. Math. **18** (2009)
11. Wriggers, P., Simo, J.C.: A note on tangent stiffness for fully nonlinear contact problems. Comm. in App. Num. Math. **1**, 199–203 (1985)

Recent Results on Summations and Volterra Difference Equations via Lyapunov Functionals

Youssef Raffoul

Abstract In this research we utilize Lyapunov functionals to obtain boundedness on all solutions, exponential stability and l_p-stability on the zero solution of summation equations and Volterra difference equations.

Keywords Volterra · Summations · Difference equations · Lyapunof functionals · Boundedness · Exponential stability

1 Introduction

In this chapter $\mathbb{R}, \mathbb{Z}$, and $\mathbb{Z}^+$ represents the sets of real numbers, all integers and all nonnegative integers, respectively and $\mathbb{Z}_{[-1,\infty)} = \mathbb{Z} \cap [-1, \infty)$. Throughout this paper the symbol $\triangle$ stands for $\triangle l(n) = l(n+1) - l(n)$, where l is any sequence $l : \mathbb{Z} \to \mathbb{R}$. In addition we adhere to the notation that $\sum_{n=a}^{b} l(n) = 0$ for $b < a$. In the introduction of [10], the author elaborated on the role that Volterra summation equations play in the qualitative analysis of neutral difference equations of the form

$$\triangle\big(H(n, x_n)\big) = f(n, x_n), \quad n \in \mathbb{Z}^+ \tag{1}$$

where H is some difference operator. For more on neutral difference equations, we refer to [2, 11].

In this study we consider the scalar Volterra summation equation

$$x(t) = a(t) - \sum_{s=0}^{t-1} C(t, s)x(s), \ t \in \mathbb{Z}^+ \tag{2}$$

Y. Raffoul (✉)
Department of Mathematics, University of Dayton, Dayton, OH 45469-2316, USA
e-mail: yraffoul1@udayton.edu

S. Baigent et al. (eds.), *Progress on Difference Equations and Discrete Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 341,
https://doi.org/10.1007/978-3-030-60107-2_18

and the scalar perturbed Volterra difference equation

$$x(t+1) = \mu(t)x(t) + \sum_{s=0}^{t-1} h(t,s)x(s) + f(t), \tag{3}$$

where $x, a, \mu, f : \mathbb{Z}^+ \to \mathbb{R}$, while $C : \mathbb{Z}^+ \times \mathbb{Z}_{[-1,\infty)} \to \mathbb{R}$ and $h : \mathbb{Z}^+ \times \mathbb{Z}_{[0,\infty)} \to \mathbb{R}$. To clear any confusion, we note that the summation term in (2) could have been started at any initial time $t_0 \geq 0$. We will use the resolvent equation, (see [10]) combined with Lyapunov functionals and fixed point theory to obtain boundedness of solutions and their asymptotic behaviors of (2). One of the major difficulties when using a suitable Lyapunov functional on Volterra summation equations is relating the solution back to that Lyapunov functional. Using Lyapunov functionals does not come for free. One must first construct such a function that implies meaningful information regarding the behavior of solutions. Such construction is an art, rather than a science. Lyapunov functions/functionals method were first implemented for ordinary differential equations, and then later on they were extended to integro-differential equations and functional differential equations. Thanks to Elaydi, in the last thirty years, Lyapunov functions/functionals were extended to all type of difference equations and Volterra difference equations, see [4] and the references therein. Since then the present author has published many papers, using the method of Lyapunof functionas/functionals to deal with boundedness, stability and the existence of periodic solutions of various kind of difference equations. However, the extension of Lyapunov method to Volterra summation equations has not been fully developed and this author has every intention of filling the void, which was initiated in [10]. Thus, the Sect. 2 of this chapter is a continuation of the work that was initiated in [10]. In [8] Messina and Vecchio displayed interesting Lyapunov functionals and studied the stability of the zero solution of Volterra integral dynamic equations under bounded and unbounded perturbations. In their work they derive different but interesting formula for the $\triangle$-derivative of absolute valued functions. For comprehensive work on the use and the construction of Lyapunov functionals we refer the reader to the book [9]. Moreover, for more on Volterra summation equations we refer to [2, 7]. In Sect. 3, we consider perturbed and unperturbed Volterra summation equations and use Lyapunov functionals to obtain exponential stability and boundedness of solutions. In Sect. 4, we consider the relationship between l_p-stability and exponential stability. For more on l_p-stability we refer to [6].

Let X denotes the set of functions $\phi : [0,t] \to \mathbb{R}$ and $\|\phi\| = \sup\{|\phi(s)| : 0 \leq s \leq t\}$. Adivar et al. [1] were the first to establish the existence of the resolvent, of an equation that is similar to (2) on time scales. Hence based on [10], the resolvent equation of (2) is given by

$$R(t,s) = C(t,s) - \sum_{u=s+1}^{t-1} R(t,u)C(u,s), \tag{4}$$

and consequently, the solution of (2) is

$$R(t,s) = C(t,s) - \sum_{u=s+1}^{t-1} R(t,u)C(u,s), \tag{5}$$

where $R : \mathbb{Z}^+ \times \mathbb{Z}_{[-1,\infty)} \to \mathbb{R}$. In [10] the emphases was on the size of $C(t,s)$ instead on its $\triangle$-difference, which is the case in the next theorem. Throughout this paper we use the notation $\triangle_t C(t,s)$ for partial difference with respect to t and $\triangle_{ts} C(t,s) = \triangle_t\big(\triangle_s C(t,s)\big)$. Also, we define the shift operator E by $Ez(t) = z(t+1)$. In the next theorem we construct a Lyapunov functional that we may call a perfect match for (2) since its $\triangle$-difference along the solutions is accomplished without using any type of inequalities.

2 Summation Equation

In this section we consider the summation equation given by (2) and use a Lyapunov functional coupled with its corresponding resolvent equation to obtain results regarding boundedness of solution.

Theorem 1 *Assume for $t \geq 1$ and $0 \leq s \leq t-1$, we have*

$$C(t,s) \geq 0,\ \triangle_s C(t-1,s-1) \geq 0,\ \triangle_t C(t-1,s-1) \leq 0,\ \triangle_{st} C(t,s-1) \leq 0. \tag{6}$$

Define the function

$$V(t) = \sum_{s=0}^{t-1} \triangle_s C(t-1,s-1)\Big(\sum_{u=s}^{t-1} x(u)\Big)^2 + C(t-1,-1)\Big(\sum_{u=0}^{t-1} x(u)\Big)^2. \tag{7}$$

(i) Let $\alpha \in (0,1)$ be a constant such that $\triangle_s C(t,t-1) + C(t,t-1) \leq \alpha$. Then

$$\triangle V(t) \leq a^2(t) - \beta x^2(t),\ \text{where } \beta = 1-\alpha. \tag{8}$$

In addition if $a \in l^2_{[0,\infty)}$, then so is x and $\sum_{s=0}^{t-1} R(t,s)a(s)$. Moreover, $V(t)$ is bounded.

(ii) Assume the existence of two positive constants D and L such that

$$\max_{t\in\mathbb{Z}^+} \sum_{s=0}^{t-1} \triangle_s C(t-1,s-1) = D,\ \text{and}\ \max_{t\in\mathbb{Z}^+} C(t,-1) = L, \tag{9}$$

then along the solutions of Eq. (2) *we have*

$$\Big(\sum_{s=0}^{t-1} R(t,s)a(s)\Big)^2 = (a(t)-x(t))^2 \le 2(D+L)V(t). \tag{10}$$

Remark 1 Note that (10) does not ask for $a \in l^2$. However, if $a \in l^2$ for all t and bounded, then both x and V are bounded.

Proof Let $V(t)$ be given by (7). Then we have, by applying $\triangle_t$ that

$$\begin{aligned}\triangle_t V(t) &= \sum_{s=0}^{t} \triangle_s C(t,s-1)\Big(\sum_{u=s}^{t} x(u)\Big)^2 + C(t,-1)\Big(\sum_{u=0}^{t} x(u)\Big)^2 \\ &\quad - \sum_{s=0}^{t-1} \triangle_s C(t-1,s-1)\Big(\sum_{u=s}^{t-1} x(u)\Big)^2 - C(t-1,-1)\Big(\sum_{u=0}^{t-1} x(u)\Big)^2.\end{aligned} \tag{11}$$

Next, we do some algebra on the side in order to simplify (11).

$$\begin{aligned}&\sum_{s=0}^{t} \triangle_s C(t,s-1)\Big(\sum_{u=s}^{t} x(u)\Big)^2 - \sum_{s=0}^{t-1} \triangle_s C(t-1,s-1)\Big(\sum_{u=s}^{t-1} x(u)\Big)^2 \\ &= \sum_{s=0}^{t} \triangle_s C(t,s-1)\Big[x(t)+\sum_{u=s}^{t-1} x(u)\Big]^2 - \sum_{s=0}^{t-1} \triangle_s C(t-1,s-1)\Big(\sum_{u=s}^{t-1} x(u)\Big)^2 \\ &= \triangle_s C(t,t-1)x^2(t) + \sum_{s=0}^{t-1} \triangle_s C(t,s-1)\Big[x(t)+\sum_{u=s}^{t-1} x(u)\Big]^2 \\ &\quad - \sum_{s=0}^{t-1} \triangle_s C(t-1,s-1)\Big(\sum_{u=s}^{t-1} x(u)\Big)^2 \\ &= \triangle_s C(t,t-1)x^2(t) + x^2(t)\sum_{s=0}^{t-1} \triangle_s C(t,s-1) + 2x(t)\sum_{s=0}^{t-1} \triangle_s C(t,s-1)\sum_{u=s}^{t-1} x(u) \\ &\quad + \sum_{s=0}^{t-1} \triangle_s C(t,s-1)\Big(\sum_{u=s}^{t-1} x(u)\Big)^2 - \sum_{s=0}^{t-1} \triangle_s C(t-1,s-1)\Big(\sum_{u=s}^{t-1} x(u)\Big)^2 \\ &= \triangle_s C(t,t-1)x^2(t) + \sum_{s=0}^{t-1} \triangle_{st} C(t,s-1)\Big(\sum_{u=s}^{t-1} x(u)\Big)^2 \\ &\quad + x^2(t)\sum_{s=0}^{t-1} \triangle_s C(t,s-1) + 2x(t)\sum_{s=0}^{t-1} \triangle_s C(t,s-1)\sum_{u=s}^{t-1} x(u).\end{aligned} \tag{12}$$

Similarly,

$$
\begin{aligned}
&C(t,-1)\Big(\sum_{u=0}^{t}x(u)\Big)^2 - C(t-1,-1)\Big(\sum_{u=0}^{t-1}x(u)\Big)^2\\
&= C(t,-1)\Big[x(t)+\sum_{u=0}^{t-1}x(u)\Big]^2 - C(t-1,-1)\Big(\sum_{u=0}^{t-1}x(u)\Big)^2\\
&= x^2(t)C(t,-1) + 2x(t)C(t,-1)\sum_{u=0}^{t-1}x(u)\\
&+ C(t,-1\Big(\sum_{u=0}^{t-1}x(u)\Big)^2 - C(t-1,-1)\Big(\sum_{u=0}^{t-1}x(u)\Big)^2\\
&= x^2(t)C(t,-1) + 2x(t)C(t,-1)\sum_{u=0}^{t-1}x(u)\\
&+ \triangle_t C(t-1,-1)\Big(\sum_{u=0}^{t-1}x(u)\Big)^2. \qquad (13)
\end{aligned}
$$

Finally, we make use of the summation by part formula; for any two sequences $y(t)$ and $z(t)$

$$\sum_{s=0}^{t-1} y(s)\triangle z(s) = y(s)z(s)|_{s=0}^{t} - \sum_{s=0}^{t-1} Ez(s)\triangle y(s).$$

With this in mind, we let $y(s) = \sum_{u=s}^{t-1} x(u)$ and $\triangle z(s) = \triangle_s C(t, s-1)$. Then $z(s) = C(t, s-1)$ and $\triangle y(s) = -x(s)$. Hence,

$$
\begin{aligned}
&2x(t)\sum_{s=0}^{t-1}\triangle_s C(t,s-1)\sum_{u=s}^{t-1}x(u))\\
&= 2x(t)\Big[C(t,s-1)\sum_{u=s}^{t-1}x(u)\Big|_{s=0}^{t} + \sum_{s=0}^{t-1}C(t,s)x(s)\Big]\\
&= 2x(t)\Big[0 - C(t,-1)\sum_{u=0}^{t-1}x(u) + \sum_{s=0}^{t-1}C(t,s)x(s)\Big]. \qquad (14)
\end{aligned}
$$

Thus, substituting (12)–(14) into (11) leads to

$$\begin{aligned}\triangle_t V(t) &= \triangle_s C(t,t-1)x^2(t) + \sum_{s=0}^{t-1} \triangle_{st} C(t,s-1)\Big(\sum_{u=s}^{t-1} x(u)\Big)^2 \\ &+ x^2(t)\sum_{s=0}^{t-1} \triangle_s C(t,s-1) + x^2(t)C(t,-1) \\ &+ 2x(t)C(t,-1)\sum_{u=0}^{t-1} x(u) + \triangle_t C(t-1,-1)\Big(\sum_{u=0}^{t-1} x(u)\Big)^2 \\ &+ 2x(t)C(t,-1)\sum_{u=0}^{t-1} x(u) + 2x(t)\sum_{s=0}^{t-1} C(t,s)x(s).\end{aligned}$$

Making use of (6) gives

$$\begin{aligned}\triangle_t V(t) &\le x^2(t)\Big[\triangle_s C(t,t-1) + C(t,-1) + \sum_{s=0}^{t-1} \triangle_s C(t,s-1)\Big] \\ &+ 2x(t)\sum_{s=0}^{t-1} C(t,s)x(s) \\ &= x^2(t)\Big[\triangle_s C(t,t-1) + C(t,t-1)\Big] + 2x(t)\sum_{s=0}^{t-1} C(t,s)x(s) \\ &\le \alpha x^2(t) + 2x(t)\Big[a(t) - x(t)\Big] \\ &\le \alpha x^2(t) + 2x(t)a(t) - 2x^2(t) \\ &\le \alpha x^2(t) + x^2(t) + a^2(t) - 2x^2(t) \\ &= (\alpha - 1)x^2(t) + a^2(t) \\ &= -\beta x^2(t) + a^2(t). \end{aligned} \tag{15}$$

Summing (15) from 0 to $t-1$ yields

$$0 \le V(t) - V(0) \le \sum_{s=0}^{t-1} a^2(s) - \beta \sum_{s=0}^{t-1} x^2(s) \tag{16}$$

which implies that if $a \in l^2_{[0,\infty)}$, then so is x, since C is bounded. Consequently, inequality (16) implies the boundedness of $V(t)$. This completes the proof of part (i).

Next we turn our attention to proving (ii). Assume (9) hold and by applying the Schwartz inequality we get

$$
\begin{aligned}
&\Big(\sum_{s=0}^{t-1}\triangle_s C(t-1,s-1)\sum_{u=s}^{t-1}x(u)\Big)^2\\
&\le \sum_{s=0}^{t-1}\triangle_s C(t-1,s-1)\sum_{s=0}^{t-1}\triangle_s C(t-1,s-1)\Big(\sum_{u=s}^{t-1}x(u)\Big)^2\\
&\le D\sum_{s=0}^{t-1}\triangle_s C(t-1,s-1)\Big(\sum_{u=s}^{t-1}x(u)\Big)^2\\
&\le D\sum_{s=0}^{t-1}\triangle_s C(t-1,s-1)\Big(\sum_{u=s}^{t-1}x(u)\Big)^2 + DC(t-1,-1)\Big(\sum_{u=0}^{t-1}x(u)\Big)^2\\
&= DV(t). \qquad (17)
\end{aligned}
$$

On the other hand, using a similar summation by parts on (14) yields

$$
\begin{aligned}
&\Big[\sum_{s=0}^{t-1}\triangle_s C(t,s-1)\sum_{u=s}^{t-1}x(u))\Big]^2\\
&= \Big[C(t,s-1)\sum_{u=s}^{t-1}x(u)\big|_{s=0}^{t} + \sum_{s=0}^{t-1}C(t,s)x(s)\Big]^2\\
&= \Big[-C(t,-1)\sum_{u=0}^{t-1}x(u) + \sum_{s=0}^{t-1}C(t,s)x(s)\Big]^2\\
&= \Big[a(t)-x(t)-C(t,-1)\sum_{u=0}^{t-1}x(u)\Big]^2\\
&\ge (1/2)\big(a(t)-x(t)\big)^2 - \Big[C(t,-1)\sum_{u=0}^{t-1}x(u)\Big]^2.
\end{aligned}
$$

Thus the above inequality gives

$$
\begin{aligned}
(1/2)\big(a(t)-x(t)\big)^2 &\le \Big[C(t,-1)\sum_{u=0}^{t-1}x(u)\Big]^2 + \Big[\sum_{s=0}^{t-1}\triangle_s C(t,s-1)\sum_{u=s}^{t-1}x(u))\Big]^2\\
&\le C(t,-1)C(t,-1)\Big(\sum_{u=0}^{t-1}x(u)\Big)^2 + DV(t)\\
&\le L\Big[C(t,-1)\Big(\sum_{u=0}^{t-1}x(u)\Big)^2 + \sum_{s=0}^{t-1}\triangle_s C(t-1,s-1)\Big(\sum_{u=s}^{t-1}x(u)\Big)^2\Big]\\
&\quad + DV(t)\\
&= (D+L)V(t).
\end{aligned}
$$

This completes the proof of (ii) and hence the Theorem.

We have the following Lemma.

Lemma 1 *Assume the hypothesis of Theorem 1 and let $V(t)$ be given by (7). Then*

$$V(t) \leq \frac{1}{\beta}\sum_{s=0}^{t-1} \triangle_s C(t-1,s-1)(t-s)\sum_{u=s}^{t-1} a^2(u)$$
$$+ \frac{1}{\beta}C(t-1,-1)t\sum_{u=0}^{t-1} a^2(u). \tag{18}$$

Proof Using Schwartz inequality in (7) gives

$$V(t) \leq \sum_{s=0}^{t-1} \triangle_s C(t-1,s-1)(t-s)\sum_{u=s}^{t-1} x^2(u) + C(t-1,-1)t\sum_{u=0}^{t-1} x^2(u). \tag{19}$$

Sum (15) from 0 to $t-1$ and obtain

$$0 \leq V(t) - V(s) \leq \sum_{u=s}^{t-1} a^2(s) - \beta\sum_{u=s}^{t-1} x^2(s)$$

which implies that

$$\sum_{u=s}^{t-1} x^2(s) \leq \frac{1}{\beta}\sum_{u=s}^{t-1} a^2(s).$$

Substituting the above inequality and (16) in (19) gives (18). This completes the proof.

3 Volterra Difference Equations

In this section we consider the perturbed scalar Volterra difference equation

$$x(t+1)) = \mu(t)x(t) + \sum_{s=0}^{t-1} h(t,s)x(s) + f(t), \tag{20}$$

and its homogenous counter part

$$x(t+1)) = \mu(t)x(t) + \sum_{s=0}^{t-1} h(t,s)x(s), \tag{21}$$

and show, under suitable conditions that, all its solutions are uniformly bounded and its zero solution is uniformly exponentially stable when $f(t)$ is identically zero. We assume the existence of an initial sequence $\phi : \mathbb{Z}^+ \to [0, \infty)$, that is bounded and $||\phi|| = \max_{0 \le s \le t_0} |\phi(s)|$, $t_0 \ge 0$ for fixed t_0. We begin with the following definition.

Definition 1 The zero solution of (21) is said to be exponentially stable if any solution $x(t, t_0, \psi)$ of (21) satisfies

$$|x(t, t_0, \psi)| \le C\Big(||\psi||, t_0\Big)\zeta^{\gamma(t-t_0)}, \quad \text{for all } t \ge t_0,$$

where ζ is constant with $0 < \zeta < 1$, $C : \mathbb{R}^+ \times \mathbb{Z}^+ \to \mathbb{R}^+$, and γ is a positive constant. The zero solution of (21) is said to be uniformly exponentially stable if C is independent of t_0.

Theorem 2 *Suppose there is a scalar sequence $\alpha : \mathbb{Z}^+ \to [0, \infty)$. Assume there are positive constants $a > 1$ and b such*

$$\alpha(s)a^{-b(t-s-1)} - \sum_{u=s}^{t-1} a^{-b(t-s-1)}|h(u, s)| > 0, \tag{22}$$

$$|\mu(t)| + |\alpha(t)| - |h(t, t)| - 1 \le -(1 - a^{-b}), \tag{23}$$

and for some positive constant M

$$\sum_{s=0}^{t-1}(1 - a^{-b})^{(t-s-1)}|f(s)| \le M, \quad for \;\; 0 \le t < \infty.$$

(i) If

$$\max_{t \ge t_0} \sum_{s=0}^{t} \Big(\alpha(s)a^{-b(t-s-1)} - \sum_{u=s}^{t} a^{-b(t-s-1)}|h(u, s)|\Big) < \infty$$

then all solutions of (20) are uniformly bounded and the zero solution of (21) is uniformly exponentially stable.
(ii) If for every $t_0 \ge 0$, there is a constant $M(t_0)$ depending on t_0 such that

$$\sum_{s=0}^{t_0-1} \alpha(s)a^{-b(t_0-s-1)} - \sum_{u=s}^{t_0-1} a^{-b(t_0-s-1)}|h(u, s)| < M(t_0),$$

then all solutions of (20) are bounded and the zero solution of (21) is exponentially stable.

Proof Consider the Lyapunov functional

$$V(t,x) = |x(t)| + \sum_{s=0}^{t-1}\Big[\alpha(s)a^{-b(t-s-1)} - \sum_{u=s}^{t-1} a^{-b(t-u-1)}|h(u,s)|\Big]|x(s)|. \tag{24}$$

Then along the solutions of (20) we have

$$\begin{aligned}\triangle V(t,x) \le\ & |\mu(t)||x(t)| + \sum_{s=0}^{t-1}|h(t,s)||x(s)| + |f(t)| \\ & + \sum_{s=0}^{t}\Big[\alpha(s)a^{-b(t-s)} - \sum_{u=s}^{t} a^{-b(t-u)}|h(u,s)|\Big]|x(s)| \\ & - \sum_{s=0}^{t-1}\Big[\alpha(s)a^{-b(t-s-1)} - \sum_{u=s}^{t-1} a^{-b(t-u-1)}|h(u,s)|\Big]|x(s)|.\end{aligned}$$

Next we try to simplify $\triangle V(t,x)$.

$$\begin{aligned}& \sum_{s=0}^{t}\Big[\alpha(s)a^{-b(t-s)} - \sum_{u=s}^{t} a^{-b(t-u)}|h(u,s)|\Big]|x(s)| \\ &= \sum_{s=0}^{t}\Big[\alpha(s)a^{-b(t-s)} - \sum_{u=s}^{t-1} a^{-b(t-u)}|h(u,s)| - |h(t,s)|\Big]|x(s)| \\ &= \sum_{s=0}^{t-1}\Big[\alpha(s)a^{-b(t-s)} - \sum_{u=s}^{t-1} a^{-b(t-u)}|h(u,s)| - |h(t,s)|\Big]|x(s)| \\ &+ \alpha(n)|x(t)| - |h(t,t)||x(t)| \\ &= a^{-b}\sum_{s=0}^{t-1}\Big[\alpha(s)a^{-b(t-s-1)} - \sum_{u=s}^{t-1} a^{-b(t-u-1)}|h(u,s)|\Big]|x(s)| \\ &- \sum_{s=0}^{t-1}|h(t,s)||x(s)| + \alpha(t)|x(t)| - |h(t,t)||x(t)|.\end{aligned}$$

Substituting the above expression into (25) and making use of (23) yield

$$\begin{aligned}\triangle V(t,x) &\le \big[|\mu(t)| + |\alpha(t)| - |h(t,t)| - 1\big]|x(t)| \\ &\quad - (1-a^{-b})\sum_{s=0}^{t-1}\Big[\alpha(s)a^{-b(t-s-1)} - \sum_{u=s}^{t-1} a^{-b(t-u-1)}|h(u,s)|\Big]|x(s)| + |f(t)| \\ &\le -(1-a^{-b})\Big[|x(t)| \\ &\quad + \sum_{s=0}^{t-1}\Big[\alpha(s)a^{-b(t-s)} - \sum_{u=s}^{t-1} a^{-b(t-u)}|h(u,s)|\Big]|x(s)| + |f(t)| \\ &= -(1-a^{-b})V(t,x) + |f(t)|. \end{aligned} \tag{25}$$

Set $\beta = (1-a^{-b}) \in (0,1)$ and apply the variation of parameters formula to get

$$\begin{aligned} V(t,x(t)) &\le (1-\beta)^{t-t_0}V(t_0,\phi) + \sum_{s=t_0}^{t-1}(1-\alpha)^{(t-s-1)}|f(s)| \\ &\le (1-\beta)^{t-t_0}||\phi||\Big[1 + \\ &\quad + \sum_{s=0}^{t_0-1}\Big[\alpha(s)a^{-b(t_0-s-1)} - \sum_{u=s}^{t_0-1} a^{-b(t_0-u-1)}|h(u,s)|\Big] \\ &\quad + \sum_{s=t_0}^{t-1}(1-\alpha)^{(t-s-1)}|f(s)|. \end{aligned} \tag{26}$$

The results readily follow from (26) and the fact that $|x(t)| \le V(t,x)$. This completes the proof.

Remark 2 We state that Theorem 2 can be easily extended to nonlinear Voleterra difference equations of the form

$$x(t+1)) = \mu(t)x(t) + \sum_{s=0}^{t-1} h(t,s)g(x(s)) + f(t),$$

under the assumption that $g(x) \le d|x|$, for some positive constant d.

4 l_p-Stability

In this section we state the definition of l_p-stability and state theorems under which it occurs. We begin by considering the non-autonomous nonlinear discrete system

$$x(n+1) = G(n, x(s);\ 0 \le s \le n) \stackrel{def}{=} G(n, x(\cdot)) \tag{27}$$

where $G : \mathbb{Z}^+ \times \mathbb{R}^k \to \mathbb{R}^k$ is continuous in x and $G(n, 0) = 0$. Let $C(n)$ denote the set of functions $\phi : [0, n] \to \mathbb{R}$ and $\|\phi\| = \sup\{|\phi(s)| : 0 \le s \le n\}$.
We say that $x(n) = x(n, n_0, \phi)$ is a solution of (27) with a bounded initial function $\phi : [0, n_0] \to \mathbb{R}^k$ if it satisfies (27) for $n > n_0$ and $x(j) = \phi(j)$ for $j \le n_0$.

Definition 2 The zero solution of (27) is stable (S) if for each $\varepsilon > 0$, there is a $\delta = \delta(n_0, \varepsilon) > 0$ such that $[n_0 \ge 0, \phi \in C(n_0),\ \|\phi\| < \delta]$ imply $|x(n, n_0, \phi)| < \varepsilon$ for all $n \ge n_0$. It is uniformly stable (US) if it is stable and δ is independent of n_0. It is asymptotically stable (AS) if it is (S) and $|x(n, n_0, \phi)| \to 0,$ as $n \to \infty$.

Definition 3 The zero solution of system (27) is said to be exponentially stable if any solution $x(n, n_0, \phi)$ of (27) satisfies

$$||x(n, n_0, \phi)|| \le C\Big(||\phi||, n_0\Big)a^{\eta(n-n_0)}, \quad \text{for all } n \ge n_0,$$

where a is constant with $0 < a < 1$, $C : \mathbb{R}^+ \times \mathbb{Z}^+ \to \mathbb{R}^+$, and η is a positive constant. The zero solution of (27) is said to be uniformly exponentially stable if C is independent of n_0.

Definition 4 The zero solution of system (27) is said to be l_p-stable if it is stable and if $\sum_{n=n_0}^{\infty} ||x(n, n_0, \phi)||^p < \infty$ for positive p.

We have the following elementary theorem and for its proof we refer to [6].

Theorem 3 *If the the zero solution of* (27) *is exponentially stable, then it is also* l_p*-stable.*

We caution that the l_p-stability is not uniform with respect to p as the next example shows. Also, it shows that (AS) does not imply l_p-stability for all p. To see this we consider the difference equation

$$x(n+1) = \frac{n}{n+1}x(n), \quad x(n_0) = x_0 \neq 0, \quad n_0 \ge 1$$

and its solution is given by

$$x(n) := x(n, n_0, x_0) = \frac{x_0 n_0}{n}.$$

Clearly the zero solution is (US) and (AS). However, for $n_0 = n$, we have

$$x(2n, n, x_0) = \frac{x_0 n}{2n} \to \frac{x_0}{2} \neq 0$$

which implies that the zero solution is not (UAS). Moreover,

$$\sum_{n=n_0}^{\infty} ||x(n, n_0, x_0)||^p \leq \sum_{n=n_0}^{\infty} |(\frac{x_0 n_0}{n})|^p = |x_0|^p (n_0)^p \sum_{n=n_0}^{\infty} (\frac{1}{n})^p,$$

which diverges for $0 < p \leq 1$ and converges for $p > 1$.

The next example shows that asymptotic stability does not necessary imply l_p stability for any $p > 0$. Let $g : [0, \infty) \to (0, \infty)$ with $\lim_{n\to\infty} g(n) = \infty$. Consider the non-autonomous difference equation

$$x(n+1) = \big[g(n)/g(n+1)\big]x(n), \;\; x(n_0) = x_0, \tag{28}$$

which has the solution $x(n, n_0, x_0) = \frac{g(n_0)}{g(n)} x_0$. It is obvious that as $n \to \infty$ the solution tends to zero, for fixed initial n_0 and the zero solution is indeed asymptotically stable. On the other hand

$$\sum_{n=n_0}^{\infty} ||x(n, n_0, x_0)||^p = [g(n_0)x_0]^p \sum_{n=n_0}^{\infty} \Big(\frac{1}{g(n)}\Big)^p, \tag{29}$$

which may not converge for any $p > 0$. For example, if we take

$$g(n) = log(n+2),$$

then from (29) we have

$$\sum_{n=n_0}^{\infty} ||x(n, n_0, x_0)||^p = [log(n_0+2)]^p ||x_0||^p \sum_{n=n_0}^{\infty} \Big(\frac{1}{log(n+2)}\Big)^p,$$

which is known to diverge for all $p \geq 0$.

The next theorem relates l_p stability to Lyapunov functionals. Again for its proof we refer to [6].

Theorem 4 *Let D be an open set in $\mathbb{R}^k$ with $0 \in D$. If there exists a continuous function $V : D \to [0, \infty)$ such that $V(0) = 0$ with $V(x) > 0$ if $x \neq 0$ and along the solutions of (27), V satisfies $\triangle V \leq -c||x||^p$, for some positive constants c and p, then the zero solution of (27) is l_p- stable.*

In the next two examples we establish that the l_p- stability depends on the type of Lyapunov functional that is being used. Moreover, there will be a price to pay if you want to obtain l_p- stability for higher values of p.

Example 1 Consider the scalar Volterra difference equation

$$x(n+1) = a(n)x(n) + \sum_{s=0}^{n-1} b(n, s) f(s, x(s)) \tag{30}$$

with f being continuous and there exists a constant λ_1 such that $f(n,x)| \leq \lambda_1|x|$. Assume there exists a positive α such that

$$|a(n)| + \lambda \sum_{s=n+1}^{\infty} |b(s,n)| + \lambda_1|b(n,n)| - 1 \leq -\alpha, \tag{31}$$

and for some positive constant λ which is to be specified later, we have

$$\lambda_1 \leq \lambda. \tag{32}$$

Then the zero solution of (30) is l_1-stable.

Proof Define the Lyapunov functional V by

$$V(n,x) = |x(n)| + \lambda \sum_{j=0}^{n-1} \sum_{s=n}^{\infty} |b(s,j)||x(j)|.$$

We have along the solutions of (30) that

$$\begin{aligned} \triangle V(t) &\leq \Big(|a(n)| + \lambda \sum_{s=n+1}^{\infty} |b(s,n)| + \lambda_1|b(n,n)| - 1\Big)|x(n)| \\ &+ (\lambda_1 - \lambda) \sum_{s=0}^{n-1} |b(n,s)||x(s)| \\ &\leq -\alpha|x(n)|. \end{aligned}$$

This implies the zero solution is stable and l_1-stable by Theorem 4. This completes the proof.

Example 2 Consider (30) and assume f is continuous with $|f(n,x)| \leq \lambda_1 x^2$. Assume there exists a positive constant α such that

$$a^2(n) + \lambda \sum_{s=n+1}^{\infty} |b(s,n)| + \lambda_1|a(n)| \sum_{s=0}^{n} |b(n,s)| - 1 \leq -\alpha, \tag{33}$$

and for some positive constant λ which is to be specified later, we have

$$\lambda_1|a(n)| + \lambda_1^2 \sum_{s=0}^{n-1} |b(n,s)| - \lambda \leq 0. \tag{34}$$

Then the zero solution of (30) is l_2-stable.

Proof define the Lyapunov functional V by

$$V(n,x) = x^2(n) + \lambda \sum_{j=0}^{n-1} \sum_{s=n}^{\infty} |b(s,j)| x^2(j).$$

We have along the solutions of (30) that

$$\begin{aligned}
\triangle V(t) &= \Big(a(n)x(n) + \sum_{s=0}^{n-1} b(n,s) f(s,x(s))\Big)^2 - x^2(n) \\
&+ \lambda x^2(n) \sum_{s=n+1}^{\infty} |b(s,n)| - \lambda \sum_{s=0}^{n-1} |b(n,s)| x^2(s) - x^2(n) \\
&\le a^2(n)x^2(n) + 2\lambda_1 |a(n)||x(n)| \sum_{s=0}^{n-1} |b(n,s)||x(s)| + \big(\sum_{s=0}^{n-1} b(n,s) f(s,x(s))\big)^2 \\
&+ \lambda x^2(n) \sum_{s=n+1}^{\infty} |b(s,n)| - \lambda \sum_{s=0}^{n-1} |b(n,s)| x^2(s) - x^2(n).
\end{aligned}$$

As a consequence of $2zw \le z^2 + w^2$, for any real numbers z and w we have

$$2\lambda_1 |a(n)||x(n)| \sum_{s=0}^{n-1} |b(n,s)||x(s)| \le \lambda_1 |a(n)| \sum_{s=0}^{n-1} |b(n,s)| (x^2(n) + x^2(s)).$$

Also, using Schwartz inequality we obtain

$$\begin{aligned}
\big(\sum_{s=0}^{n-1} b(n,s) f(s,x(s))\big)^2 &= \sum_{s=0}^{n-1} |b(n,s)|^{1/2} |b(n,s)|^{1/2} |f(s,x(s))| \\
&\le \sum_{s=0}^{n-1} |b(n,s)| \sum_{s=0}^{n-1} |b(n,s)| f^2(s,x(s)) \\
&\le \lambda_1^2 \sum_{s=0}^{n-1} |b(n,s)| \sum_{s=0}^{n-1} |b(n,s)| x^2(s).
\end{aligned}$$

Putting all together, we get

$$\begin{aligned}
\triangle V(t) &\le \Big(a^2(n) + \lambda \sum_{s=n+1}^{\infty} |b(s,n)| + \lambda_1 |a(n)| \sum_{s=0}^{n} |b(n,s)| - 1\Big) x^2(n) \\
&+ \Big(\lambda_1 |a(n)| + \lambda_1^2 \sum_{s=0}^{n-1} |b(n,s)| - \lambda\Big) \sum_{s=0}^{n-1} |b(n,s)| x^2 \\
&\le -\alpha x^2(n).
\end{aligned}$$

This implies the zero solution is stable and l_2-stable by Theorem 4. This completes the proof.

A quick comparison of (31) with (33) and (32) with (34) reveals that the conditions for the l_2 stability are more stringent.

References

1. Adivar, M., Raffoul, Y.: Existence of Resolvent for Volterra integral equations on time scales. Bull. Aust. Math. Soc. **82**, 139–155 (2010)
2. Adivar, M., Raffoul, Y.: Qualitative analysis of nonlinear Volterra integral equations on time scales using and Lyapunov functionals. Appl. Math. Comput. **273**, 258–266 (2016)
3. Bohner, M., Peterson, A.: Dynamic Equations on Time Scales. An Introduction with Applications. Birkhäuser, Boston (2001)
4. Elaydi, S.: An Introduction to Difference Equations, 3rd edn. Springer, New York (2005)
5. Kelley, W., Peterson, A.: Difference Equations An Introduction With Applications, 2nd edn. Academic Press, New York (2001)
6. Lakshmikantham, L., Trigiante, D.: Theory of Difference Equations: Numerical Methods and Applications. Academic Press, New York (1991)
7. Medina, R.: Solvability of discrete Volterra equations in weighted spaces. Dyn. Syst. Appl. **5**, 407–422 (1996)
8. Messina, E., Vecchio, A.: Stability analysis of linear Volterras equations on time scales under bounded perturbations. Appl. Math. Lett. **59**, 6–11 (2016)
9. Raffoul, Y.N.: Qualitative Theory of Volterra Difference Equations. Springer, Ney York (2018)
10. Raffoul, Y.N.: Qualitative analysis of nonconvolution Volterra summation equations. Int. J. Differ. Equ. **14**(1), 75–89 (2019)
11. Zhang, S.: Stability of neutral delay difference systems. Comput. Math. Appl. **42**, 291–299 (2001)

New Method of Smooth Extension of Local Maps on Linear Topological Spaces. Applications and Examples

Genrich Belitskii and VICTORIA RAYSKIN

Abstract The question of extension of locally defined maps to the entire space arises in many problems of analysis (e.g., local linearization of functional equations). A known classical method of extension of smooth local maps on Banach spaces uses smooth bump functions. However, such functions are absent in the majority of infinite-dimensional spaces. We suggest a new approach to localization of Banach spaces with the help of locally identical maps, which we call blid maps. In addition to smooth spaces, blid maps also allow to extend local maps on non-smooth spaces (e.g., $C^q[0, 1]$, $q = 0, 1, 2, ...$). For the spaces possessing blid maps, we show how to reconstruct a map from its derivatives at a point (see the Borel Lemma). We also demonstrate how blid maps assist in finding global solutions of cohomological equations having linear transformation of the argument. We present application of blid maps to local differentiable linearization of maps on Banach spaces. We discuss differentiable localization for metric spaces (e.g., $C^\infty(\mathbb{R})$), prove an extension result for locally defined maps and present examples of such extensions for the specific metric spaces. In conclusion, we formulate open problems.

Keywords Bump functions · Local maps · Map extensions

Research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-19-1-0399. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

G. Belitskii
Ben Gurion University of the Negev, Be'er Sheva, Israel
e-mail: Genrich@math.bgu.ac.il

V. RAYSKIN (✉)
New York Institute of Technology, Long Island, NY 11568, USA
e-mail: vrayskin@NYIT.edu

S. Baigent et al. (eds.), *Progress on Difference Equations and Discrete Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 341,
https://doi.org/10.1007/978-3-030-60107-2_19

1 Introduction

The subject of localization of maps goes back to the works of Sobolev [23] on generalized functions and of K. O. Friedrichs and D. A. Flanders on molifiers. Nowadays, the most frequently used analogous notions are the bump functions. Recall that a bump function on a space X at $x \in X$ is a map $h : X \to \mathbb{R}$ such that $|h| \leq 1$ on the entire X, $h = 1$ in a neighborhood of x and has bounded support.

There are many examples, where bump functions are used for the study of local properties of dynamical systems in $\mathbb{R}^n$. For instance, see [15, 24]. J. Palis in his work [16] considers bump functions in Banach spaces. He proves the existence of Lipschitz-continuous extensions of local maps with the help of Lipschitz-continuous bump functions. However, Nitecki [15] points out that it is unknown whether the smoothness of these extensions may be higher than Lipschitz.

Even though continuous bump functions exist in all Banach spaces, the majority of infinite-dimensional spaces do not have smooth bump functions. This is an obstacle in the local analysis of dynamical systems in infinite-dimensional spaces. Following Meshkov [12] we adopt the following

Definition 1 (*C^q-smooth spaces*) A space is called C^q-smooth, if it possesses a C^q-bump function.

Consider $X = l_p$. If $p = 2n$, then $h(||x||^p)$ is a C^∞-bump function at 0, where h is a bump function on $\mathbb{R}$. However, it is known that l_1 space does not have C^1 bump functions (e.g., [12]). Consequently, $C[0, 1]$ does not have smooth bump functions (this follows from Banach-Mazur Embedding Theorem, see [2]).

In order to allow smooth localization of Banach spaces, we define analogs of bump functions, which we call blid maps (Sect. 2). C^q-smooth blid maps exist not only on all C^q-smooth spaces, but also on some Banach spaces, which are not C^q-smooth. We present specific examples of blid maps for such spaces.

The general topological spaces, such as $C^\infty(\mathbb{R})$ and $C^\infty([0, 1])$, are frequently discussed in a context of partial differential equations. For this reason, we also discuss how to apply our ideas to linear topological spaces. In Sect. 3 we define the blid-differentiable property for topological spaces, present examples of spaces with such property and prove a theorem which asserts existence of global differentiable extension of locally defined maps.

We also discuss applications (Sect. 4) of the localization of the spaces to the questions of solvability of smooth cohomological equations and to the local differentiable linearization on Banach spaces. The proofs of these results are based on an extension of the well-known Borel Lemma to Banach spaces, which can be found in the same section.

We conclude our paper (Sect. 5) with a few more examples and open questions regarding the existence of smooth blid-maps for some non-smooth spaces, Whitney Extension Problems for non-smooth spaces and existence of Banach spaces without differentiable blid maps.

2 Banach Spaces

First, let X be a real Banach space, and Y be a real or complex Banach space. We will discuss smooth local maps $f : X \to Y$ and a possibility of smooth extension of the maps. Since Banach spaces are equipped with norms, we can consider Fréchet derivatives. In the topics related to Banach spaces, we will assume that differentiation is defined in Fréchet sense.

The map's extension is usually not unique and can be studied in the context of the equivalence class of f, i.e. a germ $[f]$. Recall (see [14]) that a germ $[f]$ at $x \in X$ is the equivalence class of local maps, such that any pair of the class members coincide on some neighborhood of x. Each element of the class is called a representative of a germ. Occasionally, we denote a germ $[f]$ as f. In the future, without loss of generality, we will assume that $x = 0$. We are interested in the question of existence of a global representative of the germ.

Consider a C^q germ. Does there exist a C^q global representative of the germ? Suppose there exists a representative with q bounded derivatives. Does there exist a global representative which also has q bounded derivatives? To answer these question, we introduce special maps, discussed below.

Definition 2 (*C^q-smooth blid maps*) A C^q map $H : X \to X$ is called a C^q blid-map at 0 for a Banach space X if there exists a neighborhood U, $0 \in U \subset X$, such that $H_{|U} = id$ and $\sup_x ||H(x)||_X < \infty$. In other words, the map H is a **B**ounded **I**ocally **ID**entical map on X.

The idea of extensions with the help of blids first appears in [1], later in [4]. The Definition 2 was introduced in [5] and was motivated with the following example.

Example 1 The C^∞ germ, defined in the neighborhood of $0 \in C[0, 1]$

$$f(x) = \int_0^1 \frac{dt}{1 - x(t)}$$

has a global C^∞ representative:

$$\int_0^1 \frac{dt}{1 - h(x(t))x(t)}.$$

Here the blid map $H(s) = h(s)s$ is defined with the help of a bump function h, such that $h(s) = 1$ on $|s| < 1/3$ and 0 on $|s| > 1/2$. It is easy to see that $H = id$ when $|s| < 1/3$ and $|H| < 1/3$ on $C[0, 1]$, i.e. satisfies the definition of a blid map.

In [5], we generalize the idea of smooth extension of a locally defined map via composition of the map with a smooth blid-map. This method allows us to prove the Borel Lemma for Banach spaces. Many questions related to local dynamics can be addressed with the help of this theorem.

Theorem 1 ([5]) *Let a space X possesses a C^q-blid map H. Then for every Banach space Y and any C^q-germ f at zero from X to Y there exists a global C^q-representative. Moreover, if all derivatives of H are bounded, and f contains a local representative bounded together with all its derivatives, then it has a global one with the same property.*

Obviously, if a space is C^q-smooth, it possesses a C^q-blid map. However, there are examples of Banach spaces that have blid-maps, but do not have bump functions of the same smoothness. We will illustrate this idea with the following examples (for details and proofs see [5]) of blid-maps in various Banach spaces:

Example 2 Suppose X has a C^q bump function $h : X \to \mathbb{R}$. Then, $H(x) = h(x)x$ is a $C^q(X)$ blid map.

In the following 3 examples h is a $C^\infty(\mathbb{R})$ bump function.

Example 3 Let $X = C[0, 1]$. Then, $H : X \to X$ defined by $H(x)(t) = h(x(t))x(t)$ is a $C^\infty(X)$ blid map.

Example 4 More generally, suppose $X = C(M)$ where M is a compact space. Then, $H(x)(t) = h(x(t))x(t)$ is a $C^\infty(X)$-blid map.

Example 5 Let $X = C^q[0, 1]$. Then a $C^\infty(X)$-blid map $H(x)(t)$ can be defined via

$$H(x)(t) = \sum_{j=0}^{q-1} \frac{t^j}{j!} h(x^{(j)}(0))x^{(j)}(0) + \int_0^t dt_1 \int_0^{t_1} dt_2 \ldots \int_0^{t_{q-1}} h\left(x^{(q)}(s)\right) x^{(q)}(s)\, ds.$$

There are also some examples of subspaces, where blid maps can be constructed:

Example 6 Suppose X possess a C^q-blid map H, and a subspace X_1 of X is H-invariant. Then the restriction $H_1 = H|X_1$ is a C^q-blid map on X_1.

Example 7 Assume $\pi : X \to X$ is a bounded projector and X possess C^q-blid map H. Then, the restriction $\pi(H)|Im(\pi)$ is a C^q-blid map on $Im(\pi)$, while the restriction $(H - \pi(H))|Ker(\pi)$ is a C^q-blid map on $Ker(\pi)$. Consequently, if $X_1 \subset X$ is a subspace, such that there exists another subspace of X, so that these two form a complementary pair, then X_1 possesses a blid map.

3 Linear Topological Spaces

As we noted in Sect. 1 localization on topological linear spaces (e.g., $C^\infty(D)$, where D is some smooth manifold) is important for the study of partial differential equations.

It is not always possible to define Fréchet differentiability on a linear topological space. For this reason, we will use weaker notions of differentiation. As we have seen

in Sect. 2, an extension of maps with the help of blids requires composition. Thus, we will discuss differentiability that satisfies the Chain Rule (in particular, we cannot use Gâteaux derivative). We will sometimes work with bounded-differentiability, sometimes the concept of stronger compact (Hadamard) differentiability, and finally (if it can be defined) the strongest form of differentiability, Fréchet differentiability. Let us recall these definitions (see [22]).

Let X and Y be linear topological spaces.

Definition 3 (*Bounded differentiability*) The map $f : X \to Y$ is bounded-differentiable at $x \in X$, if for every bounded subset $S \subset X$ and every $h \in S$ and $t \in \mathbb{R}$

$$f(x + th) - f(x) = tAh + r(th)$$

with

$$r(th)/t \to 0$$

uniformly in h as $t \to 0$ (and A is called the derivative).

Definition 4 (*Compact (Hadamard) differentiability*) The map $f : X \to Y$ is compact (Hadamard) differentiable at $x \in X$, if

$$f(x + t_n h_n) - f(x) = t_n Ah + o(t_n)$$

as $t_n \to 0$, and $h_n \to h$ (and A is called the derivative).

If both X and Y are Banach spaces with the norms $||.||_1$ and $||.||_2$ respectively, then Fréchet differentiation is well-defined.

Definition 5 (*Fréchet differentiability*) The map f is Fréchet differentiable at 0 if (in the notation of Definition 3)

$$\lim_{h \to 0} ||r(h)||_2 / ||h||_1 = 0.$$

These definitions define the same derivative A whenever it exists, and differ only by the definition of the remainder term. In what follows differentiable means one of the above three differentiability types.

Definition 6 A space X satisfies a blid-differentiable property if for every neighborhood $U \subset X$ of 0 there is a differentiable map H defined on X, locally coinciding with the identity map, such that $H(X) \subset U$.

Let us recall that a neighborhoods base of zero is a system $B = \{V_\alpha\}$ of neighborhoods of 0, such that for any neighborhood $U \subset X$ of 0 there exists some $V_\beta \in B$, $V_\beta \subset U$.

Therefore, if there is a neighborhoods base B such that for every V_α from B there exists local identity H_α, $H_\alpha(X) \subset V_\alpha$, then X satisfies the blid-property.

Proposition 1 (Extension of Local Maps) *If X satisfies blid-differentiable property, then every differentiable germ $f : X \to Y$ has a global differentiable representative.*

Proof Let f be a local representative of the germ defined on a neighborhood $U \subset X$ of zero. Let $H : X \to X$ be a differentiable local identity map such that $H(X) \subset U$. Then the map

$$F(x) = f(H(x)), \quad x \in X$$

is a global representative of the germ. □

Let X be a metric space with a metric d. Here we consider germs of maps from X into an arbitrary linear topological space Y. Instead of Fréchet differentiation (which is not defined for all metric spaces) we use bounded and compact (Hadamard) differentiation. The neighborhoods base B can be chosen as a collection $\{B_c\}_c = \{x \in X : d(x, 0) < c\}_c$. Then the space X satisfies the differentiable-blid property if for every c there exists a differentiable, local identity map $H_c : X \to X$ such that $d(H_c(x), 0) < c$ for all x, i.e., $H_c(X) \subset B_c$.

In particular, if topology on X is defined by countable collection of norms $||x||_k$, then the metric can be written as

$$d(x, y) := \sum_{k=0}^{\infty} \frac{1}{2^k} \cdot \frac{||x - y||_k}{||x - y||_k + 1}.$$

It can always be assumed that $||x||_k$ are monotonically increasing.

Proposition 2 *Suppose for every $k = 0, 1, \ldots$ there exists a global differentiable local identity map $\mathcal{H}_k$ such that*

$$\sup_x ||\mathcal{H}_k(x)||_k < \infty.$$

Then X satisfies the differentiable blid property.

Proof For a given $c > 0$ choose any

$$k > 1 - \ln c / \ln 2 \tag{1}$$

and let $\mathcal{H}_k$ be such that

$$||\mathcal{H}_k(x)||_k < N, \ x \in X.$$

Set

$$H_c(x) = \frac{c}{4N} \mathcal{H}_k \left(\frac{4N}{c} x \right).$$

Then inequality (1) and the fact that $||x||_j$ is monotonically increasing with j imply that

$$d(H_c(x), 0) < c,$$

i.e. $H_c(X) \in B_c$. □

In the following subsections, we present the examples of the spaces with differentiable blid property and state the existence of extension of locally defined maps on these spaces.

3.1 The Space of Smooth Functions on the Real Line

The space $X = C^q(\mathbb{R})$ $(0 \leq q < \infty)$ of all smooth functions on $\mathbb{R}$ can be endowed with the collection of norms

$$||x||_k = \max_{t\in[-k,k]} \max_{l\leq q} |x^{(l)}(t)|.$$

Lemma 1 *The space X possesses the bounded- (consequently compact-) differentiable blid property.*

Proof Let $h(u)$ be a C^∞-bump function on $\mathbb{R}$. Note, $a = \sup_{u\in\mathbb{R}} h(u)u < \infty$. Then

$$H(x)(t) = \begin{cases} h(x(t))x(t), \ q = 0 \\ \sum_{j=0}^{q-1} \frac{t^j}{j!} h(x^{(j)}(0))x^{(j)}(0) + \int_0^t dt_1 \int_0^{t_1} dt_2 ... \int_0^{t_{q-1}} h\left(x^{(q)}(s)\right) x^{(q)}(s)\, ds, \ q \geq 1 \end{cases}$$

is differentiable local identity map, and

$$||H(x)||_k < ae^k, \ k = 0, 1, ..., \ x \in X.$$

□

Corollary 1 *Every bounded- (consequently compact-) differentiable germ at* $0 \in C^q(\mathbb{R})$ *has a global differentiable (in the corresponding sense) representative.*

3.2 The Space of Infinitely Differentiable Functions on a Closed Interval

The space $X = C^\infty[0, 1]$ is endowed with the collection of norms

$$||x||_k = \max_{j\leq k} \max_{t\in[0,1]} |x^{(j)}(t)|.$$

Lemma 2 *The space X possesses the bounded- (consequently compact-) differentiable property.*

Proof Let $h(u)$ be the same bump function on $\mathbb{R}$ as above. Then

$$H_0(x)(t) = h(x(t))x(t)$$

is a differentiable local identity map, and

$$||H_0(x)||_0 < c.$$

for some positive constant a. Further, let $k > 0$. Then

$$H_k(x)(t) = \sum_{j=0}^{k-1} \frac{t^j}{j!} h(x^{(j)}(0))x^{(j)}(0) + \int_0^t dt_1 \int_0^{t_1} dt_2 ... \int_0^{t_{k-1}} h\left(x^{(k)}(s)\right) x^{(k)}(s)\, ds.$$

is differentiable local identity map, and

$$||H_k(x)||_k < ce^k, \quad k = 0, 1, ..., \quad x \in X. \qquad \square$$

Corollary 2 *Every bounded- (consequently compact-) differentiable germ at $0 \in C^\infty[0, 1]$ has a global representative.*

3.3 The Space of Infinitely Differentiable Functions on the Real Line

The space $X = C^\infty(\mathbb{R})$ is endowed with the collection of norms

$$||x||_k = \max_{j \le k} \max_{t \in [-k,k]} |x^{(j)}(t)|, \ k = 0, 1, 2, ...$$

Lemma 3 *The space X possesses the bounded- (consequently compact-) differentiable property.*

Proof Let $h(u)$ be the same bump function on $\mathbb{R}$ as above. Then

$$H_0(x)(t) = h(x(t))x(t)$$

is a differentiable local identity map, and

$$||H_0(x)||_0 < c.$$

Further, let $k > 0$. Then

$$H_k(x)(t) = \sum_{p=0}^{k-1} \frac{t^p}{j!} h(x^{(p)}(0))x^{(p)}(0) + \int_0^t dt_1 \int_0^{t_1} dt_2 ... \int_0^{t_{k-1}} h\left(x^{(k)}(s)\right) x^{(k)}(s)\, ds.$$

is a differentiable local identity map, and

$$||H_k(x)||_k < ce^k, \quad k = 0, 1, ..., \quad x \in X. \qquad \square$$

Corollary 3 *Every bounded- (consequently compact-) differentiable germ at* $0 \in C^\infty(\mathbb{R})$ *has a global representative.*

4 Applications

Frequently, in questions of local analysis and local dynamical systems bump functions are used. For example, they are used for normal forms conjugation [24], in the proofs of the Borel Lemma [15], and the Whitney Extension results [26].

Since blid maps substitute bump functions, they allow localization of a broader class of spaces. First, the blid maps were used in [1] for a smooth conjugation of two C^∞ diffeomorphisms on some Banach spaces. In the later works [4, 19] we discussed conditions for when two C^∞ diffeomorphisms on some Banach spaces are locally C^∞-conjugate. Below, we discuss applications of blid maps to differentiable linearization (without non-resonance assumption), and applications to cohomological equations. For the proofs of these results we need the Borel Lemma extended to Banach spaces. With the help of blid-maps we are able to prove the Borel Lemma for Banach spaces.

4.1 The Borel Lemma

In this section we state the version of the Borel Lemma proved in [5]. For finite dimensional X, the Borel Lemma [15] is a particular case of the celebrated Whitney theorem [26] on the extensions of functions beyond a closed set. The use of blid-maps in our proofs is analogous to the use of the bump functions in the proofs of finite-dimensional case. The infinite dimensional version of the proof also requires some estimates on the growth of the derivatives of the blid-maps.

Theorem 2 (The Borel lemma) *Let a Banach space X possess a C^∞-blid map with bounded derivatives of all orders. Then for any Banach space Y and any sequence $\{P_j\}_{j=0}^\infty$ of continuous homogeneous polynomial maps from X to Y there is a C^∞-map $f : X \to Y$ with bounded derivatives of all orders such that $P_j(x) = f^{(j)}(0)(x)^j$ is satisfied for all $j = 0, 1, ...$*

Here (in the notations of [6]) $f^{(j)}(0)$ is the j-linear map and $f^{(j)}(0)(x)^j$ is the value of this map at the point $\underbrace{x, x, ..., x}_{j}$.

4.2 Cohomological Equations

In this section we outline the main ideas of the application of the blid maps to the solutions of cohomological equations. For the detailed discussion please see [5]. Given a map $F : X \to X$, (X is a Banach space) we want to find a C^∞ $g : X \to \mathbb{C}$, that satisfies the following cohomological equation:

$$g(Fx) - g(x) = f(x) \tag{2}$$

For a broad overview of various versions of the equation see the works of Lyubich (e.g., [13]). Also, for a discussion of smooth cohomological equations we recommend the book [3].
In our example, we will assume that F is linear and denote it by A.
Define a homogenious polinomial map $P_n(x) = f^{(n)}(0)(x)^n$. We will search for homogeneous, degree n, polynomial solutions $Q_n(x)$ ($n = 1, 2...$) such that

$$(L_n - id)\,Q_n(x) = P_n(x), \quad n = 1, 2, 3...., \tag{n}$$

where $L_n Q_n(x) = (Q_n(Ax))^{(n)}$.

If for every n equation (n) is solvable, we call the cohomological equation (2) formally solvable. Then we can use Borel Lemma to reduce the Eq. (2) to the equation in flat functions (that are the functions with 0 Taylor coefficients at the origin). Then, applying some decomposition results (see [5]) for the space X, we can formulate conditions for the solvability of the original cohomological equation:

Theorem 3 ([5]) *Let $A : X \to X$ be a hyperbolic linear automorphism, where a Banch space X possesses a C^∞-blid map with bounded derivatives on X. If all derivatives of f are bounded on every bounded subset, and the cohomological equation (2) is formally solvable at zero (i.e. each n-th equation has continuous solution, $n = 1, 2, ...$), then there exists a global C^∞-solution $g(x)$.*

4.3 *Differentiable Linearization Without Non-resonance Assumption*

Local linearization and normal forms are convenient simplification of complex dynamics. In this section we discuss differentiable linearization on Banach spaces. Following the approach of Poincaré [18], for a diffeomorphism $F : X \to X$ (X a Banach space) with a fixed point 0, we would like to find a smooth transformation Φ defined in a neighborhood of 0 such that $\Phi \circ F \circ \Phi^{-1}$ has a simplified (polynomial) form [18] called the normal form. If $\Phi \circ F \circ \Phi^{-1} = DF = \Lambda$ is linear, the conjugation is called linearization. There are two major questions in this area of research: how to improve smoothness of the conjugation Φ, and how to lower the assumption on the smoothness of the given diffeomorphism F.

Hartman [10] and Grobman [8] independently showed that if Λ is hyperbolic, then for a diffeomorphism F there exists a local homeomorphism Φ such that $\Phi \circ F \circ \Phi^{-1} = \Lambda$. Different proofs were given by Pugh in [17]. A higher regularity of Φ has been an active area of research (see, for example, [9, 16, 25, 27]).

The first attempt to answer the question of differentiability of Φ at the fixed point 0 under hyperbolicity assumption was made in [25], but an error was found and discussed in [20]. Later, in [9], Guysinsky, Hasselblatt and Rayskin presented a correct proof. However, it was restricted to $F \in C^\infty$ (or more precisely, it was restricted to $F \in C^k$, where k is defined by complicated expression). It was conjectured in the paper [9] that the result is correct for $F \in C^2$ (as it was originally claimed in [25]).

Zhang, Lu and Zhang, in their Theorem 7.1 published in [27] showed that for a Banach space diffeomorphism F with a hyperbolic fixed point and α-Hölder DF, the local conjugating homeomorphism Φ is differentiable at the fixed point. Moreover,

$$\Phi(x) = x + O(||x||^{1+\beta}) \text{ and } \Phi^{-1}(x) = x + O(||x||^{1+\beta})$$

as $x \to 0$, for certain $\beta \in (0, \alpha]$.

There are two additional assumptions behind this theorem. The first one is the spectral band width inequality. The authors explain that this inequality is sharp if the spectrum has at most one connected component inside of the unit circle in X, and at most one connected component outside of the unit circle in X. The precise formulation of the spectral band width condition is somewhat bulky and we present it in the Appendix. It is important (and it is pointed out in [27]) that this is not a non-resonance condition. The latter is required for generic linearization of higher smoothness.

The second assumption is the assumption that the Banach space must possess smooth bump functions. It is conjectured in the paper that the second assumption is a necessary condition.

In this section we explain that this conjecture is not correct (see Theorem 4). The bump function condition can be replaced with the less restrictive blid map condition. Blid maps allow to reformulate Theorem 7.1 in the following way:

Theorem 4 *Let X be a Banach space possessing a differentiable blid map with bounded derivative. Suppose $F : X \to X$ is a diffeomorphism with a hyperbolic fixed point, DF is α-Hölder, and the spectral band width condition is satisfied. Then, there exists local linearizing homeomorphism Φ which is differentiable at the fixed point. Moreover,*

$$\Phi(x) = x + O(||x||^{1+\beta}) \text{ and } \Phi^{-1}(x) = x + O(||x||^{1+\beta})$$

as $x \to 0$, for certain $\beta \in (0, \alpha]$.

In particular, we have the following

Corollary 4 *Let $X = C^q[0, 1]$. Suppose $F : X \to X$ is a diffeomorphism with a hyperbolic fixed point, DF is α-Hölder, and the spectral band width condition is satisfied. Then, the local conjugating homeomorphism Φ is differentiable at the fixed point. Moreover,*

$$\Phi(x) = x + O(||x||^{1+\beta}) \text{ and } \Phi^{-1}(x) = x + O(||x||^{1+\beta})$$

as $x \to 0$, for certain $\beta \in (0, \alpha]$.

Below we sketch a proof of Theorem 4

Proof Zhang, Lu and Zhang showed that for the conclusion of their Theorem 7.1 it is enough to satisfy the inequalities 1 and 2 (see 4 below), which are called condition (7.6) in their paper.

In order to apply the blid maps instead of bump functions to the inequalities (4), it is sufficient to construct a bounded blid map, which has only first-order bounded derivative. That is, let blid map $H(x) : X \to X$ be as follows:

$$\begin{array}{l} 1.\ H(x) = x \text{ for } ||x|| < 1 \\ 2.\ H \in C^1 \text{ and } ||H^{(j)}(x)|| \le c_j,\ j = 0, 1. \end{array} \tag{3}$$

The condition (7.6) of [27] is:

$$\begin{array}{l} 1.\ \sup_{x \in X} ||DF(x) - \Lambda|| \le \delta_\eta \\ 2.\ \sup_{x \in V \setminus \{0\}} \{||DF(x) - \Lambda||/||x||^\alpha\} < \infty. \end{array} \tag{4}$$

Here δ_η is some small constant, and V is a neighborhood of 0.

Let $DF - \Lambda = f$. Define for $\delta > 0$

$$\tilde{f}(x) := f\,(\delta H(x/\delta))$$

We will show that if f satisfies (4), then so does $\tilde{f}$.

$$\sup_{x\in X}||D\tilde{f}(x)|| \le \sup_{x\in X}||Df(x)|| \cdot \sup_{x\in X}||DH(x)|| \le \delta_\eta \cdot c_1 = \tilde{\delta_\eta}.$$

Here $\tilde{\delta_\eta}$ can be made as small as necessary via appropriate choice of δ_η.

Thus, the first inequality of (7.6) holds for $\tilde{f}$. For the second inequality we have the following estimate:

$$\frac{||D\tilde{f}(x)||}{||x||^\alpha} \le \frac{||Df\,(\delta H(x/\delta))\,||}{||\delta H(x/\delta)||^\alpha} \cdot \left(\frac{||\delta H(x/\delta)||}{||x||}\right)^\alpha.$$

The second multiple is bounded, because for small x (say, $||x/\delta|| < \epsilon$ for some $\epsilon > 0$) we have

$$\frac{||\delta H(x/\delta)||}{||x||} < c_1 + o(1),$$

while for $||x/\delta|| \ge \epsilon$

$$\frac{||\delta H(x/\delta)||}{||x||} < c_0/\epsilon.$$

I.e., $\frac{||\delta H(x/\delta)||}{||x||}$ is less than some constant m. Then,

$$\sup_{x\in V\setminus O}\frac{||D\tilde{f}(x)||}{||x||^\alpha} \le \sup_{0<||x||<\delta c_0}\{||Df(x)||/||x||^\alpha\} \cdot \sup_{x\in X}||DH(x)|| \cdot m^\alpha$$

$$= \sup_{0<||x||<\delta c_0}\{||Df(x)||/||x||^\alpha\}\, c_1 \cdot m^\alpha.$$

This quantity is bounded by Mc_1m^α if δ is sufficiently small and M is defined as $\sup_{x\in V\setminus\{0\}}\{||DF(x) - \Lambda||/||x||^\alpha\}$. □

5 More Examples and Open Questions

One of the important questions of local analysis on Banach spaces is the following. Do Banach spaces without smooth blid maps exist? Recently, affirmative answer was presented in [7] (also see [11]). The authors of [7, 11] proved that there exist Banach spaces that do not allow C^2-extension (and hence the C^2-blid map).

Question 1 For which spaces do smooth blid maps exist? Do they exist on l_p, with non-even p?

Question 2 Are there Banach spaces without differentiable blid maps?

In the Theorem 1 we considered a C^q-germ at a point. For such germs the existence of a local representative with bounded derivatives implies the existence of the global one with the same properties.

How can we extend germs of maps defined at a closed subset $S \subset X$? For this construction we need to define smooth blid maps at S and germs at S. More precisely, generalizing the definition of germs at a point, we will say that maps f_1 and f_2 from neighborhoods U_1 and U_2 of S into Y are *equivalent*, if they coincide in a (smaller) neighborhood of S. Every equivalence class is called a *germ at* S. We pose the same question. Given a C^q-germ at S, does there exist a global representative? Assume there exist a C^q-map $H : X \to X$ whose image $H(X)$ is contained in a neighborhood U of S and which is equal to the identity map in a smaller neighborhood. Such maps we call *smooth blid maps at* S. Then every local map f defined in U can be extended on the whole X. It suffices to set $F(x) = f(H(x))$.

In the next example, we construct the map H for a segment (in particular, for a ball).

Example 8 Let $S(A)$ be a set of all functions $x \in C[0, 1]$ whose graphs $(t, x(t))$ are contained in a closed $A \subset \mathbb{R}^2$, where A is chosen in such a way that $S(A) \neq \emptyset$. Let $h(t, x)$ be a C^∞-function, which is equals to 1 in a neighborhood of A and vanishes outside of a bigger set. Then, for an arbitrary $y \in C[0, 1]$

$$H_y(x)(t) = y(t) + h(t, x(t))(x(t) - y(t))$$

is a C^∞-blid map for $S(A)$.

If $A = \{\{t, x\} : \min(\psi(t), \phi(t)) \leq x \leq \max(\psi(t), \phi(t))\}$ for some $\phi, \psi \in C[0, 1]$, then $S(A)$ can be thought of as a segment $[\phi, \psi] \subset C[0, 1]$.

In particular, given $z \in C[0, 1]$ and a constant $r > 0$, setting $\phi = z - r$ and $\psi = z + r$, we obtain the ball $B_r(z) = \{x : ||x - z|| \leq r\} \subset C[1, 0]$.

Every C^q-germ at $[\phi, \psi] \subset C[0, 1]$ contains a global representative.

Note, this example has an obvious generalization to segments and balls in $C^k[0, 1]$.

Question 1 and Example 8 bring us to the next question.

Question 3 For which pairs (S, X) do similar constructions exist? In particular, can a smooth blid map be constructed for any bounded subset S of a space X possessing a smooth blid map? For example, we do not know whether a smooth blid map can be constructed for a sphere $S = \{x \in C[0, 1] : ||x|| = r\}$.

Example 9 The Borel lemma for finite-dimensional spaces is a particular case of the well-known Whitney extension theorem from a closed set $S \subset \mathbb{R}^n$. There are other variations of extension questions among the Whitney Extension Problems. They can be applied to fitting smooth functions and manifolds to data. Fitting manifolds to data is related to the Whitney extension problem for the infinite-dimensional case. Some cases of these Extension Problems are solved in [21] with the help of the blid map ideas.

In Sect. 3 we presented several examples of linear topological spaces with the differentiable blid property.

Question 4 Which linear topological spaces have the differentiable blid property?

Linearization is a convenient simplification in the study of local dynamics. In some cases partial differential equations can be studied in terms of operators on linear topological spaces. Thus, there arises the question of differentiable linearization.

Question 5 Is it possible to generalize the Theorem 4 for the case of linear topological spaces (e.g., space of C^∞ functions), which posses differentiable blid property?

6 Appendix

Here we formulate the spectral band width condition.

Assume that $x = 0$ is a hyperbolic fixed point of F, $\Lambda = DF(0)$ and the spectrum

$$\sigma(\Lambda) = \sigma_- \cap \sigma_+,$$

where $\sigma_- = \{\lambda \in \sigma(\Lambda) : |\lambda| < 1\}$ and $\sigma_+ = \{\lambda \in \sigma(\Lambda) : |\lambda| > 1\}$.

The sets $\sigma_\pm$ can be written as the union of disjoint sets:

$$\sigma_- = \sigma_1 \cap ... \cap \sigma_p \text{and} \sigma_+ = \sigma_{p+1} \cap ... \cap \sigma_d, \tag{5}$$

where $d \in \mathbb{N}$, $p \in \{1, ..., d\}$ and the numbers $\lambda_i^- := \inf\{|\lambda| : \lambda \in \sigma_i\}$, $\lambda_i^+ := \sup\{|\lambda| : \lambda \in \sigma_i\}$ $(i = 1, ..., d)$ satisfy

$$0 < \lambda_1^- \le \lambda_1^+ < ... < \lambda_p^- \le \lambda_p^+ < 1 < \lambda_{p+1}^- \le \lambda_{p+1}^+ < ... < \lambda_d^- \le \lambda_d^+. \tag{6}$$

Then the spectral band inequality can be written as

$$\begin{array}{l} \lambda_i^+/\lambda_i^- < (\lambda_p^+)^{-\alpha}, \ i = 1, ..., p \\ \lambda_j^+/\lambda_j^- < (\lambda_{p+1}^-)^{\alpha}, \ j = p+1, ..., d. \end{array} \tag{7}$$

References

1. Belitskii, G.: The Sternberg theorem for a Banach space. Funct. Anal. Appl. **18**, 238–239 (1984). http://link.springer.com/article/10.1007%2FBF01086163. (MR 0757253 (86b:58097))
2. Banach, Stefan: Théorie des opérations linéaires. Monografie Matematyczne, Warszawa (1932)
3. Belitskii, G., Tkachenko, V.: One-dimensional functional equations. In: Operator Theory: Advances Applications. vol. 144. Birkhäauser (2003)
4. Belitskii, G., Rayskin, V.: Equivalence of families of diffeomorphisms on Banach spaces. Math. preprint archive, UT Austin, 07–71. https://www.ma.utexas.edu/mp_arc-bin/mpa?yn=07-71
5. Belitskii, G., Rayskin, V.: A New Method of Extension of Local Maps of Banach Spaces, Applications and Examples. Contemporary Mathematics, AMS series, p. 733 (2019)

6. Cartan, H.: Calcul Différentiel. Hermann, Paris, pp. 178 (1967). (MR 0223194 (36 #6243))
7. D'Alessandro, S., Hajek, P.: Polynomial algebras and smooth functions in Banach spaces. J. Funct. Anal, **266**, 1627–1646 (2014)
8. Grobman, D.M.: Homeomorphisms of systems of differential equations. Doklady Akademii Nauk SSSR **128**, 880–881 (1959)
9. Guysinsky, M., Hasselblatt, B., Rayskin, V.: Differentiability of the Hartman-Grobman linearization. Discret. Contin. Dyn. Syst. **9**(4), 979–984 (2003)
10. Hartman, P.: A lemma in the theory of structural stability of differential equations. Proc. A.M.S. **11**(4), 610–620 (1960)
11. Hajek, P., Johanis, M.: Smooth Analysis in Banach Spaces. Walter de Gruyter, GmbH, Berlin, p. 497 (2014)
12. Meshkov, V.Z.: Smoothness properties in Banach spaces. Studia Mathe. **63**, 111–123 (1978). http://matwbn.icm.edu.pl/ksiazki/sm/sm63/sm6319.pdf. (MR 0511298 (80b:46027))
13. Lyubich, Yu.: The cohomological equations in nonsmooth categories. arXiv:1211.0229v1 [math. FA] 1 Nov.2012, https://arxiv.org/pdf/1211.0229.pdf
14. Narasimhan, R.: Analysis on real and complex manifolds. Masson & Cie/North-Holland (1973)
15. Nitecki, Z.: Differentiable Dynamics. An Introduction to the Orbit Structure of Diffeomorphisms. The MIT Press, Camridge, Mass.-London, pp. xv+282 (1971). (MR 0649788 (58 #31210))
16. Palis, J.: Local Structure of Hyperbolic Fixed Points in Banach Space. Anais da Academia Brasileira de Ciencias **40**, 263–266 (1968)
17. Pugh, C.C.: On a theorem of P. Hartman. Am. J. Math. **91**, 363–367 (1969)
18. Poincaré, H., Nouvelles, Les Méthodes, de la Mécanique Celeste: English translation, New Methods of Celestial Mechanics, History of Modern Physics and Astronomy 13, p. 1993. Am. Inst. Phys. (1892)
19. Rayskin, V.: Theorem of Sternberg-Chen modulo central manifold for Banach spaces. Ergod. Theory Dyn. Syst. **29**(6), 1965–1978 (2009). https://doi.org/10.1017/S0143385708000989. (MR 2563100 (2011a:37038))
20. Rayskin, V.: α-Hölder linearization. J. Differ. Equ. **2**(147), 271–284 (1998)
21. Rayskin, V.: Whitney's and Seeley's type of extensions for maps defined on some Banach spaces (2019). https://arxiv.org/abs/1910.14248. (preprint)
22. Schechter, M.: Differentiation in abstract spaces. J. Differ. Equ. **55**, 330–345 (1984)
23. Sobolev, S.: Sur un théorème d'analyse fonctionnelle. Rec. Math. [Mat. Sbornik] N.S. **4**(46), 471–497 (1938). (3)
24. Sternberg, S.: On the structure if local homeomorphisms of Euclidean n-space II. Am. J. Math. **80**, 623–631 (1958)
25. van Strien, S.: Smooth linearization of hyperbolic fixed points without resonance conditions. J. Differ. Equ. **85**(1), 66–90 (1990)
26. Whitney, H.: Analytic extensions of functions defined in closed sets. Trans. Am. Math. Soc. **36**(1), 63–89 (1934)
27. Zhang, W., Lu, K., Zhang, W.: Dfifferentiability of the Conjugacy in the Harman-Grobman Theorem. Trans. Am. Math. Soc. **369**(7), 4995–5030 (2017)

QRT-Families of Degree Four Biquadratic Curves Each of Them Has Genus Zero, Associated Dynamical Systems

Guy Bastien and MARC ROGALSKI

Abstract In the Congress ICDEA2019 in London, we give two examples of QRT-families of biquadratic curves $Q_1(x, y) - \lambda Q_2(x, y) = 0$, with Q_1 of degree 4 and Q_2 of degree 2, each of them has genus zero; these examples contrast with many examples published of QRT-families, where almost all curves have genus one. After a brief summary of these examples (the details will be published in Sarajevo Journal of Mathematics), we give an example with Q_1 of degree 4 and Q_2 of degree 3. We prove that, for the QRT-map T associated to this family, the orbit of every point not in the union of three lines and an hyperbola converges to a fixed point. Finally we present an example with Q_1 and Q_2 of degrees 4, where there are some bifurcations in the behaviour of the QRT-map.

Keywords QRT maps · Genus of curves · Dynamical systems

1 Introduction, the Results

In the Congress ICDEA2019 in London, we introduced two QRT-families of biquadratic curves (of degree 2 in x and in y) $Q_1(x, y) - \lambda Q_2(x, y) = 0$ each of them has genus 0. In these examples, Q_1 was of degree 4, but Q_2 was of degree 2 only. The associated QRT-map (see the classical definition in Sect. 4) in $\mathbb{R}^2 \setminus \{(x, y)|x = y\}$ has two different behaviours in two regions: convergent orbits in a region, periodicity or density in a curve in the other region. In the present paper, we present first a summary of these examples (the details of proofs will be published in [7]).

Then we present a case of a QRT-family with Q_1 of degree 4 and Q_2 of degree 3, such that every curve of it has genus zero. And we prove that for the QRT-map

G. Bastien · M. ROGALSKI (✉)
Institut Mathématique de Jussieu-Paris Rive Gauche and CNRS, 4 place Jussieu, 75005 Paris, France
e-mail: marc.rogalski@imj-prg.fr

G. Bastien
e-mail: guy.bastien@imj-prg.fr

S. Baigent et al. (eds.), *Progress on Difference Equations and Discrete Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 341,
https://doi.org/10.1007/978-3-030-60107-2_20

associated to this family, which is defined outside of the union of the three lines $x + y = 0$, $x = 1$ and $y = 1$, and the hyperbola $xy = 1$, the orbit of every point not in this set converges to the point $(1, 1)$.

This situation contrasts with most of the classical QRT-families studied in some papers: see [1, 4–6, 8, 9, 11, 13],…, where almost all curves have genus 1, *id est* are elliptic, and where the dynamical system has an infinity of periods for a dense set of initial points. So it was necessary in these cases to use tools which become from algebraic geometry of elliptic curves, such as Weierstrass' function and the chord-tangent law on a regular cubic curve. This will not be necessary for our examples (Fig. 1).

In the last section, we present without proof (it is analogous to this one in [7] or presented in Sects. 3 and 4) an example with Q_1 and Q_2 of degrees 4.

2 Summary of the Two First Examples Presented in ICDEA2019

The two examples are the following:

$$x^2y^2 - 5xy(x + y) + 16(x^2 + y^2) - 20(x + y) + 16 - \lambda(x - y)^2 = 0, \tag{1}$$

$$x^2y^2 - 5xy(x + y) - 24(x^2 + y^2) - 20(x + y) + 16 - \lambda(x - y)^2 = 0. \tag{2}$$

For the origin of these examples, see [4, 7]. The results are the following:

Theorem 1 *Every curve of each QRT families (1) and (2) is not reducible and is of genus* 0*, except for two values of* λ *for which the curve is reducible. The exceptional values of* λ *are* 10 *and* 11 *for the family (1), and* -19 *and* -10 *for the family (2).*

Define the two functions

$$\begin{aligned} G_1(x, y) &:= \frac{x^2y^2 - 5xy(x + y) + 16(x^2 + y^2) - 20(x + y) + 16}{(x - y)^2}, \\ G_2(x, y) &:= \frac{x^2y^2 - 5xy(x + y) - 24(x^2 + y^2) - 20(x + y) + 16}{(x - y)^2}. \end{aligned} \tag{3}$$

Theorem 2 *For the family (1), suppose* $G_1(M_0) \notin \{10, 11\}$*. If* $G_1(M_0) < 11$*, then the sequence of points* $T^n(M_0)$ *converges to the point* $D = (2, 2)$*; if* $G_1(M_0) > 11$*, then the sequence of points* $T^n(M_0)$ *is periodic or is dense in the curve* $\mathcal{C}_\lambda$ *which passes through* M_0*.*

Theorem 3 *For the family (2), suppose* $G_2(M_0) \notin \{-19, -10\}$*. If* $G_2(M_0) < -19$*, then the sequence of points* $T^n(M_0)$ *converges to* $D' = (-2, -2)$*; if* $G_2(M_0) > -19$*, then the sequence of points* $T^n(M_0)$ *is periodic or is dense in the curve* $\mathcal{C}_\lambda$ *which passes through* M_0*.*

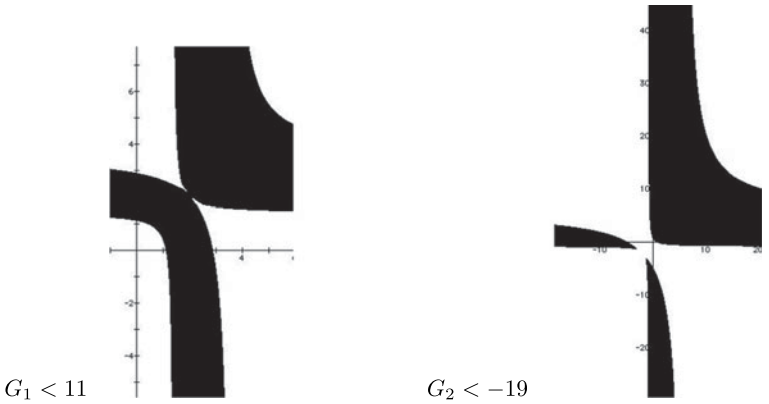

Fig. 1 The regions defined by the values of the G_i

3 The New Example, Proof that the Curves Have Genus Zero

The example is the following: we consider the family of real curves $\mathcal{C}_\lambda$ with equations

$$B_\lambda := \left[x^2y^2 - xy(x+y) + (x^2+y^2) - (x+y) + 1\right] - \lambda(x+y)(x-1)(y-1) = 0 \tag{4}$$

where $\lambda \in \mathbb{R}$.

First we remark that for $\lambda = -1$ the curve $\mathcal{C}_{-1}$ split in a double hyperbola with equation

$$(xy-1)^2 = 0. \tag{5}$$

Moreover we remark that the point $D = (1, 1)$ is on every curve $\mathcal{C}_\lambda$. So it is simpler to make the change of variables

$$x = u + 1, \qquad y = v + 1, \tag{6}$$

so that the new curve $\widetilde{\mathcal{C}_\lambda}$ has equation

$$\widetilde{B}_\lambda := u^2v^2 + uv(u+v) + u^2 + v^2 - \lambda uv(u+v+2) = 0. \tag{7}$$

The point D becomes the point $O = (0, 0)$. First we search the singular points at infinity, which are given by the easy result:

Lemma 1 *The points at infinity are H and V in directions horizontal and vertical, and they are double points.*
If $-1 < \lambda < 3$ the tangents (asymptotes) at these points are complex and distinct.
If $\lambda < -1$ and if $\lambda > 3$, the asymptotes are real and distinct.

If $\lambda = -1$, there are double asymptotes $v = -1$ and $u = -1$.
If $\lambda = 3$, there are double asymptotes $v = 1$ and $u = 1$.
So if $\lambda \notin \{-1, 3\}$, H and V are ordinary singular points of multiplicity 2.

One has also easily, with formula (7).

Lemma 2 *The point O is an ordinary singular points of $\widetilde{\mathcal{C}}_\lambda$ of multiplicity 2 if $|\lambda| \neq 1$, which is isolated (complex distinct tangents) if $|\lambda| < 1$, and with real distinct tangents if $|\lambda| > 1$. If $\lambda = -1$ there is a double tangent $(u+v)^2 = 0$, and if $\lambda = 1$ the double tangent is $(u-v)^2 = 0$.*

Now we search another singular points. We have

$$\begin{aligned} \frac{d\widetilde{B}_\lambda}{du} - \frac{d\widetilde{B}_\lambda}{dv} &= (u-v)\Big((\lambda-1)(u+v) + 2(\lambda+1) - 2uv\Big), \\ u\frac{d\widetilde{B}_\lambda}{du} - v\frac{d\widetilde{B}_\lambda}{dv} &= (u-v)\Big(2(u+v) - (\lambda-1)uv\Big). \end{aligned} \tag{8}$$

So we search if a point (t, t) may be a singular point. We must have $\widetilde{B}_\lambda(t, t) = 0$ and $\frac{d\widetilde{B}_\lambda}{du}(t, t) = 0$, so we have as solution $t = 0$, which gives again the point O. If $t \neq 0$ when the equations

$$t^2 - 2(\lambda-1)t + 2(1-\lambda) = 0 \quad \text{and} \quad 2t^2 - 3(\lambda-1) = 0$$

must have a common root. But it is easy to see that this is possible only for $\lambda = -1$ or $\lambda = 1$. In the first case, it is evident because all the points of $\widetilde{\mathcal{C}}_{-1}$ are singular. For $\lambda = 1$ we have again the point O.

If $u \neq v$, we must find a common root of the two second members of (8), and see if it is a solution of $\widetilde{B}_\lambda(u, v) = 0$. We put $u + v = s$ and $uv = p$, and the three equations become

$$\begin{aligned} &p^2 - (\lambda-1)ps + s^2 - 2(\lambda+1)p = 0, \\ &(\lambda-1)s - 2p + 2(\lambda+1) = 0, \\ &2s - (\lambda-1)p = 0. \end{aligned} \tag{9}$$

The two last equations give s and p if $\lambda \neq 3$, we put their values in the first equation and obtain

$$\frac{4(\lambda+1)}{\lambda-3} = 0. \tag{10}$$

This is possible only for $\lambda = -1$, the case of the double curve $(xy-1)^2 = 0$ or, in the new coordinates, $(uv+u+v)^2 = 0$.

In fine if $\lambda = 3$, then the two last equations in (9) are not compatible. In conclusion, if $\lambda \neq -1$, there is no other singular point except O, H and V, and they are ordinary singular points with multiplicity 2 if $\lambda \neq 1$.

If $|\lambda| \neq 1$, the formula for the genus g of an algebraic curve of degree d

$$g = \frac{(d-1)(d-2)}{2} - \sum_{p \in P} \frac{\mu_p(\mu_p - 1)}{2}, \tag{11}$$

which is true if the curve is not reducible, and where P is the set of all singular points p supposed ordinary and of multiplicity μ_p (see [10]), gives for genus the number 0. So it remains to see if the curves are not reducible.

Necessary, by the symmetry with respect to the diagonal and the biquadratic character of the curve, the only possibilities are that one has $u^2v^2 - (\lambda - 1)uv(u + v) + u^2 + v^2 - 2\lambda uv$ identical to

$$(u^2 + \alpha u + \gamma)(v^2 + \alpha v + \gamma), \quad \text{or to}$$

$$(uv + \alpha(u + v) + \gamma)(uv + \beta(u + v) + \delta), \quad \text{or to}$$

$$(uv + \alpha u + \beta v + \gamma)(uv + \beta u + \gamma v + \gamma).$$

It is easy to see that each of these cases is impossible or gives again the curve $(uv + u + v)^2 = 0$ with $\lambda = -1$.

For $\lambda = 1$, the curve $\widetilde{\mathcal{C}_1}$ has for equation $u^2v^2 + (u - v)^2 = 0$, that it is reducible as two complex conic curves.

So we have proved the essential result.

Theorem 4 *If $|\lambda| \neq 1$, the curve $\mathcal{C}_\lambda$ is not reducible and of genus zero. If $\lambda = -1$ it is reducible as $(xy - 1)^2 = 0$. If $\lambda = 1$, it is reducible as $(x - 1)^2(y - 1)^2 + (x - y)^2 = 0$ (two complex conic curves except the real point $D = (1, 1)$).*

At last we prove the final result of this section.

Proposition 1 (a) *If $-1 < \lambda \leq 1$, the curve $\mathcal{C}_\lambda$ has no real point, except the point D.*
(b) *The intersection of the set $\{Q_2(x, y) = 0\} = \{x = 1\} \cup \{y = 1\} \cup \{x + y = 0\}$ with each curve $\mathcal{C}_\lambda$ in the real domain is exactly the point $D = (1, 1)$.*
(c) *The intersection of the set $\{xy = 1\}$ with a curve $\mathcal{C}_\lambda$ with $\lambda \neq -1$ reduces to the point $D = (1, 1)$.*

Proof The second and third assertions are easy. So we prove assertion (a).

For a point (u, v) of a curve $\widetilde{\mathcal{C}_\lambda}$ with equation $R_1(u, v) - \lambda R_2(u, v) = 0$ we have $\lambda = \dfrac{R_1(u, v)}{R_2(u, v)}$. We shall prove that we have $\lambda = \dfrac{R_1(u, v)}{R_2(u, v)} > 1$ or $\lambda = \dfrac{R_1(u, v)}{R_2(u, v)} \leq -1$, with $R_1(u, v) = u^2v^2 + uv(u + v) + u^2 + v^2$ and $R_2(u, v) = uv(u + v) + 2uv$.

If $R_2(u, v) > 0$, then $\lambda > 1$: in fact, $u^2v^2 + uv(u + v) + u^2 + v^2 > uv(u + v + 2)$, because $u^2v^2 + (u - v)^2 > 0$ (except at O). If $R_2(u, v) < 0$, then one has $\lambda < -1$. Because $u^2v^2 + uv(u + v) + u^2 + v^2 > -uv(u + v + 2)$, or $(uv + u + v)^2 > 0$, which is true, except for the curve $\widetilde{\mathcal{C}_{-1}}$. □

In the following Figure, we give some examples of the forms of the curves $\mathcal{C}_\lambda$.

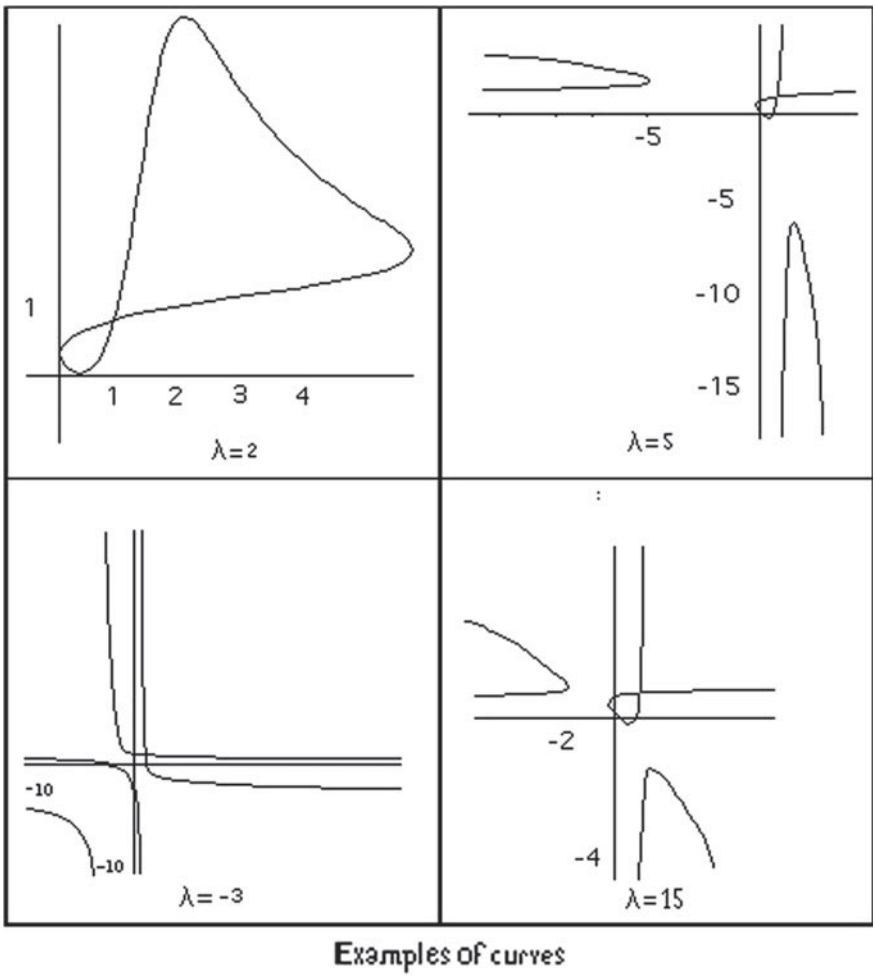

Fig. 2 Different forms of curves

4 The Associated Dynamical System

Recall what is the QRT-map T associated to the QRT-family (4). Let be M a point in the region where $Q_2(x, y) \neq 0$, that is the set $Z := \mathbb{R}^2 \setminus \Big[\{x + y = 0\} \cup \{x = 1\} \cup \{y = 1\}\Big]$. The horizontal line which passes to M cuts again the curve $\mathcal{C}_\lambda$ of the family which passes to M at a point M_1, and the vertical line through M_1 cuts $\mathcal{C}_\lambda$ at the image $T(M)$. It is possible that one of these points are at infinity, *id est* in H or in V. Moreover the image by T of a point in $\{xy = 1\}$ is always V. So we consider

the set $\overline{Z} := (Z \cup \{H, V\}) \setminus \{xy = 1\}$. It results of Proposition 1(b)(c) that the map T sends the set $\overline{Z}$ into itself. So it is possible to study the dynamical system $(\overline{Z}, T)$.

First we make the change of variables

$$x = \frac{1-X}{Y}, \quad y = \frac{1-Y}{X}, \qquad \text{or} \qquad X = \frac{x-1}{xy-1}, \quad Y = \frac{y-1}{xy-1}, \tag{12}$$

obtained by splitting the point D. Putting the new variables in the equation of $\mathcal{C}_\lambda$, we obtain a double line with equation $(X + Y - 1)^2 = 0$ and a family of conics, equations of them are (for $\lambda \neq -1$)

$$X^2 + Y^2 - X - Y + \frac{1}{\lambda + 1} = 0. \tag{13}$$

With the new change of variables

$$X = U + 1/2, \quad Y = V + 1/2, \tag{14}$$

and putting

$$a^2 = \frac{\lambda - 1}{2(\lambda + 1)}, \quad a > 0 \ (|\lambda| > 1), \tag{15}$$

we obtain a family of circles Γ_a of centre $(0, 0)$ and radius a:

$$U^2 + V^2 = a^2. \tag{16}$$

Now we search the conjugated $\widetilde{T}$ of the map T by this transformation (12). We remark that the horizontal lines become the pencil of lines passing through the point $\widetilde{H} = (-1/2, 1/2)$, and that the vertical lines become the pencil of lines passing through the point $\widetilde{V} = (1/2, -1/2)$. So the map $\widetilde{T}$ is defined geometrically by the following procedure: if $M \in \Gamma_a$, the line $(M\widetilde{H})$ cuts Γ_a at M_1, and the line $(M_1\widetilde{V})$ cuts again Γ_a at the point $\widetilde{T}(M)$ (remark that $\widetilde{H}$ and $\widetilde{V}$ are not on Γ_a).

We use the parametrisation of the circles

$$U(t) = a\frac{1-t^2}{1+t^2}, \quad V(t) = a\frac{2t}{1+t^2}, \tag{17}$$

and some easy computations with *Maple* gives the parameter s of the image $\widetilde{T}(M)$:

$$s = h(t) := \frac{(2a^2 - 2a + 1)t - 2a}{2a^2 + 2a + 1 - 2at}. \tag{18}$$

This map h is an homographic map (also called Moebius map, because $2a^2 - 1 \neq 0$), which has two fixed points: $1 + \sqrt{2}$ and $1 - \sqrt{2}$. So it exists a number k independant of t such that we have

$$\frac{h(t)-(1+\sqrt{2})}{h(t)-(1-\sqrt{2})}=k\frac{t-(1+\sqrt{2})}{t-(1-\sqrt{2})}. \tag{19}$$

The n-iteration of the map h is associated to the number k^n, so we have to determine k by the formulas (18) and (19). The computation gives

$$k=\left(\frac{1+\sqrt{2}a}{1-\sqrt{2}a}\right)^2>1. \tag{20}$$

So $k^n\to\infty$ when $n\to\infty$, and then $h^n(t)\to 1-\sqrt{2}$ when $n\to\infty$.

We return to the old coordinates : $U\to a\frac{\sqrt{2}-1}{2-\sqrt{2}}$, $V\to a\frac{1-\sqrt{2}}{2-\sqrt{2}}$; then the limit of (X,Y) is $\left(1/2+a\frac{\sqrt{2}-1}{2-\sqrt{2}},1/2-a\frac{\sqrt{2}-1}{2-\sqrt{2}}\right)$, and finally $(x,y)\to(1,1)$.

In fine we have proved the final result.

Theorem 5 *For every point M of the set $\mathbb{R}^2\setminus\Big[\{x+y=0\}\cup\{x=1\}\cup\{y=1\}\cup\{xy=1\}\Big]$, the point $T^n(M)$ converges to the point $D=(1,1)$.*

So we see that this result contrasts with this one of Sect. 2 (or of [7]) where in a region of the plane the dynamical system has an infinity of periods.

5 An Example with Q_1 and Q_2 of Degrees Four

Without proof, we give another example $Q_1-\lambda Q_2=0$ with Q_1 and Q_2 of degrees four. This is the following

$$G_\lambda(x,y):=x^2y^2+x^2+y^2-\lambda xy(xy-1)=0. \tag{21}$$

For $\lambda\notin\{-2,1,2\}$, each of the associated curves has the points O, H and V which are ordinary singular points with multiplicity two. Moreover if $\lambda\notin\{-2,1,2\}$, then the curves are not reducible. So their genus are exactly zero.

For $\lambda\in\{-2,1\}$, the curves are reducible to two complex conic curves or to two complex lines. For $\lambda=2$, the curve reduces to two real hyperbolas. And for $\lambda\in[-2,1]$ the curves have only complex points, except the point O.

For the associated QRT-map T defined in the complement of the set $\{xy=0\}\cup\{xy=1\}$, there are two regions: this one where one has $\left|\dfrac{x^2y^2+x^2+y^2}{xy(xy-1)}\right|>2$, and then if M is in this region $T^n(M)\to O=(0,0)$; and the region where one has $1<\dfrac{x^2y^2+x^2+y^2}{xy(xy-1)}<2$, and then every integer larger than some integer N is the period of some points in this region, and the orbits of some points are dense in the curves $\mathcal{C}_\lambda$ which contain them, the two kinds of points being dense in this second region.

References

1. Bastien G., Mañosa V., Rogalski M.: On periodic solutions of 2-periodic Lyness' equations. Int. J. Bifurc. Chaos **23**(4) (2013)
2. Bastien G., Rogalski M.: A biquadratic system of two order one difference equations: periods, chaotic behavior of the associated dynamical system. Int. J. Bifurc. Chaos **22**(11) (2012)
3. Bastien, G., Rogalski, M.: On some algebraic difference equations $u_{n+2}u_n = \psi(u_{n+1})$ in $\mathbb{R}_*^+$, related to families of conics or cubics: generalization of the Lyness' sequences. J. Math. Anal. Appl. **300**, 303–333 (2004)
4. Bastien G., Rogalski M.: Behavior of orbits and periods of a dynamical system in $\mathbb{R}_*^2$ associated to a special QRT-map. DCDIS-A **27**, 81–130 (2020)
5. Bastien G., Rogalski M.: A QRT-system of two order one homographic difference equations: conjugation to rotations, periods of periodic solutions, sensitiveness to initial conditions, In: Alsedà i Soler L., Cushing J., Elaydi S., Pinto A. (eds.) Difference Equations, Discrete Dynamical Systems and Applications. ICDEA 2012. Springer Proceedings in Mathematics and Statistics, vol. 180. Springer, Berlin (2017)
6. Bastien, G., Rogalski, M.: On the algebraic difference equations $u_{n+2} + u_n = \psi(u_{n+1})$ in $\mathbb{R}$, related to a family of elliptic quartics in the plane. J. Math. Anal. Appl. **326**, 822–844 (2007)
7. Bastien G., Rogalski M.: Dynamical systems associated with QRT families of degree four biquadratic curves each of them of genus zero. To appear in Sarajevo Journal of Mathematics
8. Duistermaat J.J.: Discrete Integrable Systems. QRT Maps and Elliptic Surfaces. Springer Monographs in Mathematics (2010)
9. Jogi, D., Roberts, J.A.G., Vivaldi, F.: An algebraic geometric approach to integrable maps of the plane. J. Phys. A Math. Gen. **39**, 1133–1149 (2006)
10. Perrin, D.: Géométrie algébrique. CNRS Editions, EDP Sciences (2001)
11. Pettigrew, J., Roberts, J.A.G.: Characterizing singular curves in parametrized families of biquadratics. J. Phys. A Math. Theor. **41**, 115203 (2008)
12. Quispel, G.R.W., Roberts, J.A.G., Thompson, C.J.: Integrable mappings and soliton equations. Phys. Lett. A **126**, 419–421 (1988)
13. Zeemann, E.C.: Geometric unfolding of a difference equation (1996). (unpublished paper)

Stability of Discrete-Time Coupled Oscillators via Quotient Dynamics

Brian Ryals

Abstract We examine a discrete-time version of the Kuramoto Model for coupled oscillators. Phase-locked states of N coupled oscillators correspond to invariant circles on $\mathbb{T}^N$, and can be viewed as fixed points of a quotient dynamical system. We will discuss how to define and classify the stability of these phase-locked states via the corresponding equilibria in the quotient system, and give some examples where the quotient mapping helps identify the phase-locked families and their stability.

Keywords Kuramoto model · Coupled oscillators · Asymptotic stability

1 Introduction

Perhaps the most historically significant coupled oscillator model is the Kuramoto model, which features N oscillators rotating around a circle in continuous time. Inspired by observations made by Winfree [20], Kuramoto's original model [14] had the form

$$\dot{\theta}_i = \omega_i - \sum_{j=1}^{N} \Gamma_{ij}(\theta_i - \theta_j),\ i = 1, \ldots, N\ .$$

Kuramoto then made the following simplification on the coupling functions: He set $\Gamma_{ij}(\theta_i - \theta_j) = \frac{K}{N}\sin(\theta_i - \theta_j)$ for all i, j. The coupling constant K is nonnegative and is independent of i and j. With this, Kuramoto's model becomes

$$\dot{\theta}_i = \omega_i - \frac{K}{N}\sum_{j=1}^{N} \sin(\theta_i - \theta_j),\ i = 1, \ldots, N\ . \tag{1}$$

B. Ryals (✉)
California State University, Bakersfield, Bakersfield, CA 93311, USA
e-mail: bryals@csub.edu

S. Baigent et al. (eds.), *Progress on Difference Equations and Discrete Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 341,
https://doi.org/10.1007/978-3-030-60107-2_21

The version with the sine coupling is usually what is meant by the "The Kuramoto Model." Classical studies of this model in the infinite N limit may be found in [1, 6, 14, 16, 19]. Rigorous results for finite N can be found in [4, 5, 12, 15]. More recent works include [7–9, 17].

In this chapter, we will consider a *discrete-time* adaption of the Kuramoto model. Our main goal is to put the ideas of synchronization, phase-locking, and the stability of such solutions in a rigorous setting. We will discuss different viewpoints of the model and show how the dynamics may be reduced to a quotient dynamical system. We will conclude with some examples to illustrate the usefulness of our quotient system.

Throughout this chapter, we adopt the following notation. We view

$$\mathbb{S} \cong \mathbb{R}/\left(\mathbb{Z} + \frac{1}{2}\right)$$

so that $\mathbb{S}$ is a circle with circumference 1 and let

$$p : \mathbb{R} \to \mathbb{S}$$

be the corresponding projection map. We view $(-1/2, 1/2]$ as a fundamental domain and for each $x \in \mathbb{S}$, we let $\hat{x} \in \mathbb{R}$ be given by

$$\hat{x} = p^{-1}(x) \cap (-1/2, 1/2] \,. \tag{2}$$

We view $\mathbb{S}$ as being positively oriented and adopt additive notation on the circle, so that if $x, y \in \mathbb{S}$, then $x + y = p(\hat{x} + \hat{y})$.

The N-dimensional torus is defined similarly. We view

$$\mathbb{T}^N = \mathbb{S} \times \mathbb{S} \times \cdots \times \mathbb{S} \cong \mathbb{R}^N/(\mathbb{Z} + 1/2)^N \,.$$

As each component of $\mathbb{T}^N$ is an element of $\mathbb{S}$, addition is understood to be defined by addition of their respective components.

The difference equation model to be considered in the rest of this chapter is

$$\theta_i(t+1) = \theta_i(t) + \omega_i - p\left(\frac{K}{N}\sum_{j=1}^{n} g\big(\theta_i(t) - \theta_j(t)\big)\right), \; i = 1, \ldots, N \,.$$

Here we take discrete steps $t \in \mathbb{Z}^+$, the coupling constant K is nonnegative, N is the number of oscillators, denoted as $\theta_1, \ldots, \theta_N$ with each $\theta_i \in \mathbb{S}$, and $\omega_i \in \mathbb{S}$ corresponds to a rigid rotation in the uncoupled ($K = 0$) system. Further, p is the projection map described above, and $g : \mathbb{S} \to \mathbb{R}$ is the coupling function which we will assume to be an odd function.

2 Viewpoints of the Model and a Quotient Dynamical System

There are a couple of ways to visualize the dynamics. The first is the *marked particle* viewpoint where we view, as Kuramoto did in the continuous model, all N oscillators as N particles rotating along a lone circle. The ideas of synchronization and phase-locking are perhaps more intuitive in this marked particle viewpoint, see Fig. 1. With synchronization, every oscillator collapses to the same point and rotates around the circle together thereafter (row 1 of Fig. 1). Phase-locking is similar, with oscillators having their phase differences approaching a constant, rotating in unison with these phase differences remaining locked (row 2 of Fig. 1). A lack of either is sometimes called incoherence (row 3 of Fig. 1). In some coupled oscillator systems, chimera states (not pictured), where some oscillators phase-lock while others act incoherently, have been observed [17].

The other *torus* viewpoint is to view the whole phase space on $\mathbb{T}^N$ and view the collection of oscillators as a single point, with the dynamics being an orbit of points on the torus. That is, we view the equation as $\theta(t+1) = F(\theta(t))$ where $F : \mathbb{T}^N \to \mathbb{T}^N$, $\theta(t) = \big(\theta_1(t), \ldots, \theta_N(t)\big)$. Here we denote the i-th component by $F_i(\theta)$, which is given by

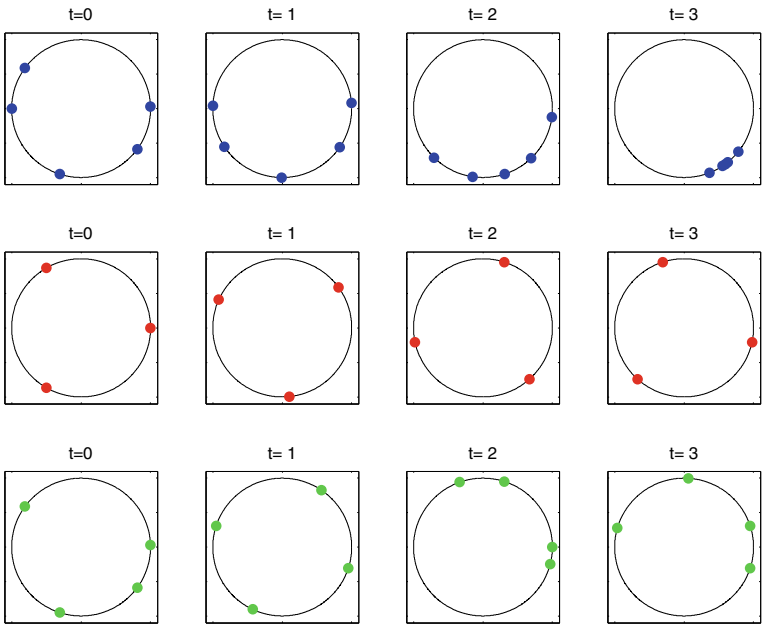

Fig. 1 Examples of different long term phenomena of coupled oscillators. In the top row, the oscillators synchronize. In the middle row, the oscillators are phase-locked. In the bottom row, the oscillators are incoherent

$$F_i(\theta) = \theta_i + \omega_i - p\left(\frac{K}{N}\sum_{j=1}^{n} g(\theta_i - \theta_j)\right) .$$

A *synchronized solution* is any orbit $\theta(t)$ where for all $t \in \mathbb{Z}^+$ we have $\theta_i(t) = \theta_j(t)$ for all $i, j \in 1, \ldots, N$. Similarly, a *phase-locked solution* is any orbit $\theta(t)$ where for each $t \in \mathbb{Z}^+$, we have $\theta_i(t) - \theta_j(t) = \psi_{ij}$ for all $i, j \in 1, \ldots, N$, with ψ_{ij} independent of the time parameter t. Of course, if all $\psi_{ij} = 0$, then the phase-locked solution is a synchronized solution.

There are two important observations to be made for these phase-locked solutions, both of which will motivate our study of a quotient system to be described shortly. The first observation is that these phase-locked solutions cannot be asymptotically stable in the usual sense. To see this, observe that if $\theta(t)$ is a phase-locked solution, then any rotation of θ is also a distinct phase-locked solution. More generally, we have the following Lemma.

Lemma 1 *Let $\theta \in \mathbb{T}^N$. Let $\overline{\Omega} = (\Omega, \Omega, \ldots, \Omega)$ where $\Omega \in \mathbb{S}$. Let m be a positive integer. Then*

$$F^m\left(\theta + \overline{\Omega}\right) = F^m(\theta) + \overline{\Omega} .$$

Proof This is just a computation. For $m = 1$ we have that the i-th entry of $F\left(\theta + \overline{\Omega}\right)$ is given by

$$\begin{aligned} F_i(\theta + \overline{\Omega}) &= \theta_i + \Omega + \omega_i + p\left(\frac{K}{N}\sum_{j=1}^{n} g(\theta_i + \Omega - \theta_j - \Omega)\right) \\ &= \theta_i + \Omega + \omega_i + p\left(\frac{K}{N}\sum_{j=1}^{n} g(\theta_i - \theta_j)\right) \\ &= F_i(\theta) + \Omega . \end{aligned}$$

The statement for $m > 1$ follows by induction. □

This implies we cannot have the usual notion of asymptotic stability for our phase-locked solutions. For if we have any phase-locked solution, we can translate all of its components by an arbitrary small constant, and this new solution will never return to the original solution. However, from a practical point of view, if a perturbation of a phase-locked solution is asymptotic to a different phase-locked solution with an identical structure, then in some sense that solution should be stable. Our quotient system, to be defined shortly, will address this paradox by treating all rotations of a configuration in the marked particle viewpoint as identical.

The other important observation is to notice that neither synchronized solutions nor phase-locked solutions are necessarily fixed points of F, as they still rotate with their locked-phases. However, by equating rotations, we can turn the problem of finding phase-locked solutions into one of finding fixed points of a quotient map.

Let us now be more precise. For $x, y \in \mathbb{T}^N$, where $x = (x_1, \ldots, x_N)$ and $y = (y_1, \ldots, y_N)$, we define $x \sim y$ if for some $\Omega \in \mathbb{S}$, we have that $x_i + \Omega = y_i$ for $i = 1, \ldots, N$. Notice that Lemma 1 implies that if $x \sim y$, then $F(x) \sim F(y)$, so that the equivalence relation respects the dynamics. This gives a quotient space $\mathbb{T}^N / \sim \equiv \tilde{\mathbb{T}}^N$ with a projection map $q : \mathbb{T}^N \to \tilde{\mathbb{T}}^N$ that sends a point $x \in \mathbb{T}^N$ to its equivalence class $\tilde{x} \in \tilde{\mathbb{T}}^N$. We will study the induced quotient mapping $\tilde{F} : \tilde{\mathbb{T}}^N \to \tilde{\mathbb{T}}^N$ where $\tilde{F} = q \circ F \circ q^{-1}$.

Note that the preimage of a single point $\tilde{x} \in \tilde{\mathbb{T}}^N$ given by $q^{-1}(\tilde{x})$ is an entire circle in $\mathbb{T}^N$. Consequently, if $\tilde{x}$ is a fixed point of $\tilde{F}$, so that $\tilde{F}(\tilde{x}) = \tilde{x}$, then the corresponding circle $q^{-1}(\tilde{x}) \in \mathbb{T}^N$ is invariant under F, i.e. $F\left(q^{-1}(\tilde{x})\right) = q^{-1}(\tilde{x})$. We will soon show that these invariant circles are the phase-locked solutions of the original mapping.

We remark that it is straightforward to show that $\tilde{\mathbb{T}}^N$ is homeomorphic to $\mathbb{T}^{N-1}$ using the following well known topological lemma.

Lemma 2 *Let X and Y be compact Hausdorff spaces and let h be a continuous function from X onto Y. Let $\sim$ be an equivalence relation on X such that $x \sim y$ if and only if $h(x) = h(y)$. Then $X/\sim$ is homeomorphic to Y.*

For a proof of the lemma, see e.g. [10].

Corollary 1 *The quotient space $\tilde{\mathbb{T}}^N$ is homeomorphic to $\mathbb{T}^{N-1}$.*

Proof An onto function $h : \mathbb{T}^N \to \mathbb{T}^{N-1}$ will be constructed such that $x \sim y$ if and only if $h(x) = h(y)$. Let $x \in \mathbb{T}^N$ be such that $x = (x_1, \ldots, x_N)$ with each $x_i \in \mathbb{S}$. Let

$$h(x) = h(x_1, \ldots, x_N) = (x_1 - x_N, x_2 - x_N, \ldots, x_{N-1} - x_N) \in \mathbb{T}^{N-1} .$$

The function h is onto since

$$h(a_1, \ldots, a_{N-1}, 0) = (a_1, \ldots, a_{N-1})$$

for any $(a_1, \ldots, a_{N-1}) \in \mathbb{T}^{N-1}$. Further, if $h(x) = h(y)$, then

$$(x_1 - x_N, x_2 - x_N, \ldots, x_{N-1} - x_N) = (y_1 - y_N, y_2 - y_N, \ldots, y_{N-1} - y_N) ,$$

so that $x_i = y_i - y_N + x_N = y_i + \Omega$ for all i, with $\Omega = x_N - y_N \in \mathbb{S}$, so that $x \sim y$.

If $x \sim y$, then $(x_1, \ldots, x_N) = (y_1 + \Omega, \ldots, y_N + \Omega)$, and

$$h(x) = (x_1 - x_N, x_2 - x_N, \ldots, x_{N-1} - x_N)$$

while

$$h(y) = (y_1 + \Omega - y_N - \Omega, y_2 + \Omega - y_N - \Omega, \ldots, y_{N-1} + \Omega - y_N - \Omega) = h(x) .$$

Thus, the result follows from the preceding lemma. □

The next lemma says that phase-locked solutions are exactly the trajectories that correspond to an invariant circle on $\mathbb{T}^N$.

Lemma 3 *Let $\theta(t)$ be an orbit of $\theta(t+1) = F(\theta(t))$. Then $\theta(t)$ is a phase-locked solution if and only if there exists an $\tilde{x}$ such that $\theta(t) \in q^{-1}(\tilde{x})$ for all t.*

Proof Recall that a phase-locked solution is any orbit $\theta(t)$ where we have for any $i, j \in \{1, \ldots, N\}$ that

$$\theta_i(t) - \theta_j(t) = \psi_{ij} \tag{3}$$

with ψ_{ij} independent of the time parameter t.

First, assume that $\theta(t)$ is a phase-locked solution. Then for all $t \geq 0$, we have, by taking $j = 1$ in Eq. (3) above for all i, that

$$\begin{aligned}\big(\psi_{11}, \psi_{21}, \ldots, \psi_{N1}\big) &= \big(\theta_1(t) - \theta_1(t), \theta_2(t) - \theta_1(t), \ldots, \theta_N(t) - \theta_1(t)\big) \\ &= \big(\theta_1(t), \theta_2(t), \ldots, \theta_N(t)\big) - \big(\theta_1(t), \theta_1(t), \ldots, \theta_1(t)\big)\,.\end{aligned}$$

Then for any times τ_1, τ_2, we have that

$$\begin{aligned}&(\theta_1(\tau_1), \ldots, \theta_N(\tau_1)) - (\theta_1(\tau_1), \theta_1(\tau_1), \ldots, \theta_1(\tau_1)) \\ &= (\theta_1(\tau_2), \ldots, \theta_N(\tau_2)) - (\theta_1(\tau_2), \theta_1(\tau_2), \ldots, \theta_1(\tau_2))\,.\end{aligned}$$

But then $\theta_i(\tau_1) = \theta_i(\tau_2) + c_{\tau_1,\tau_2}$ where $c_{\tau_1,\tau_2} = \theta_1(\tau_2) - \theta_1(\tau_1)$ for all i. Consequently $\theta(\tau_1) \sim \theta(\tau_2)$ and since τ_1, τ_2 are arbitrary, it follows that $\theta_t \in q^{-1}\big(q(\theta(0))\big)$ for all t.

Now assume there exists a $\tilde{x}$ such that $\theta(t) \in q^{-1}(\tilde{x})$ for all t. Then $\theta(t) \sim \theta(0)$ for all t, so that $\theta_i(t) = \theta_i(0) + c(t)$ for some $c(t) \in \mathbb{S}$ for all i. But then $\theta_i(t) - \theta_j(t) = \theta_i(0) - \theta_j(0)$ for every t, with $\theta_i(0) - \theta_j(0)$ independent of t. □

We remark that in the proof of Lemma 3 we use the difference

$$\big(\theta_1(t), \theta_2(t), \ldots, \theta_N(t)\big) - \big(\theta_1(t), \theta_1(t), \ldots, \theta_1(t)\big)\,.$$

An intuitive interpretation of this computation in the marked particle viewpoint is that we mark one of the particles and rotate our point of view after teach time step so that this particle is always in the same location. As a consequence of Lemma 3, we have the following.

Corollary 2 *Let $\theta(t)$ be an orbit of $\theta(t+1) = F(\theta(t))$. Then $\theta(t)$ is a phase-locked solution if and only if $q(\theta(t))$ is a fixed point of $\tilde{F}$ for all t.*

Proof If $\theta(t)$ is a phase-locked solution, then by Lemma 3 there exists a $\tilde{x}$ such that $\theta(t) \in q^{-1}(\tilde{x})$ for all t. Then

$$q^{-1}(\tilde{x}) = \{x \in \mathbb{T}^N : x = \theta(0) + \big(\Omega, \Omega, \ldots, \Omega\big),\ \Omega \in \mathbb{S}\}\,.$$

Then for any $x \in q^{-1}(\tilde{x})$, we have

$$F(x) = F\left(\theta(0) + \left(\Omega, \Omega, \ldots, \Omega\right)\right) = F(\theta(0)) + \left(\Omega, \Omega, \ldots, \Omega\right).$$

Then

$$q(F(x)) = q\left(F(\theta(0)) + \left(\Omega, \Omega, \ldots, \Omega\right)\right) = q\left(F(\theta(0))\right) = \tilde{x}\,.$$

It follows that $\tilde{F}(\tilde{x}) = q(F(q^{-1}(\tilde{x}))) = \tilde{x}$.

If instead $q(\theta(t))$ is a fixed point of $\tilde{F}$ for all t, then $\theta(t) \in q^{-1}(\tilde{x})$ for all t and the result follows by Lemma 3. □

Lemma 3 and Corollary 2 imply that phase-locked solutions come in entire circles in the phase space, and that each phase-locked solution is part of a one parameter family of solutions. Moreover, these circles correspond to a fixed point of the induced quotient dynamical system $\tilde{F}$, so to find all of the phase-locked solutions (and thus also the synchronized solutions), we only need to find all the fixed points of $\tilde{F}$. In the next section, we will explore how to construct this map in more detail, as well as discuss the stability of these phase-locked solutions.

3 Stability of Phase-Locked Solutions

As noted in the previous section, since phase-locked solutions come in whole circles, it is impossible to obtain the usual notion of asymptotic stability. We will consider instead the stability of the entire family of phase-locked solutions. Informally, we can think of perturbing a phase-locked solution, and asking if the future of this perturbation is asymptotic to the future of some phase-locked solution in the same family. As this entire family is a single point in $\tilde{\mathbb{T}}^N$, we are really asking for asymptotic stability of a fixed point of $\tilde{F}$.

In symbols, by a *family of phase-locked solutions* we mean the one-parameter set $\theta(t) + \overline{\Omega}$ with $\theta(t) \in \mathbb{T}^N$ and our parameter $\overline{\Omega} = (\Omega, \ldots, \Omega)$ with $\Omega \in \mathbb{S}$. Here $\theta(t)$ satisfies $\theta(t+1) = F(\theta(t))$ and corresponds to a fixed point for $\tilde{F}$, so that $q(\theta(t)) = \tilde{x}$ for all t. We define this family of phase-locked solutions to be asymptotically stable if the fixed point $\tilde{x}$ of $\tilde{F}$ is asymptotically stable.

For differentiable maps in $\mathbb{R}^N$, we often use eigenvalues to identify equilibria that are asymptotically stable. For smooth g, let us now see how to both explicitly write out the quotient mapping using coordinates, as well as how to extract a matrix from our quotient map so that we can perform the usual type of analysis.

Let $\tilde{x} \in \tilde{\mathbb{T}}^N$. Then, by Corollary 1, $\tilde{x}$ has a representation $(x_1, \ldots, x_{N-1}) \in \mathbb{T}^{N-1}$. Since one of the points $x \in q^{-1}(\tilde{x})$ has its last component equal to 0, we can choose the coordinates in such a way that the first $N-1$ coordinates match with $(x_1, \ldots, x_{N-1}) \in \mathbb{T}^{N-1}$ for each $\tilde{x} \in \tilde{\mathbb{T}}^N$ (see also the proof of Corollary 1). That is, we have that

$$q^{-1}(\tilde{x}) = \{x \in \mathbb{T}^n \ : \ x = (x_1, \ldots, x_{N-1}, 0) + \left(\Omega, \Omega, \ldots, \Omega\right), \ \Omega \in \mathbb{S}\} \, .$$

Denoting $\overline{\Omega} = \left(\Omega, \Omega, \ldots, \Omega\right)$ and $\overline{x} = (x_1, \ldots, x_{N-1}, 0)$, we have by Lemma 1 that

$$q\left(F\left(q^{-1}(\tilde{x})\right)\right) = q\left(F(\overline{x} + \overline{\Omega})\right) = q\left(F(\overline{x}) + \overline{\Omega}\right) = q\left(F\left(\overline{x}\right)\right) \, .$$

Notice the last component

$$F_N(\overline{x}) = \omega_N + p\left(\frac{K}{N}\sum_{j=1}^{N-1} g(x_j)\right) \, .$$

Then consider

$$x^* = \overline{x} - \left(\omega_N + p\left(\frac{K}{N}\sum_{j=1}^{N-1} g(x_j)\right), \ldots, \omega_N + p\left(\frac{K}{N}\sum_{j=1}^{N-1} g(x_j)\right)\right)$$

so that x^* is the same as $\overline{x}$ but rotated so that the last component of $F(x^*)$ is 0. Since $\overline{x} \sim x^*$, we have $q(F(\overline{x})) = q(F(x^*))$, so it suffices to write out the mapping $F(x^*)$. We have for $i = 1, \ldots, N-1$ that

$$F_i(x^*) = x_i + \omega_i - \omega_N - p\left(\frac{K}{N}g(x_i) + \frac{K}{N}\sum_{j=1}^{N-1} g(x_i - x_j)\right) - p\left(\frac{K}{N}\sum_{j=1}^{N-1} g(x_i)\right)$$

while by construction

$$F_N(x^*) = 0 \, .$$

It follows that $\tilde{F}(\tilde{x}) = q \circ F \circ q^{-1}(\tilde{x})$ has as its $N-1$ components

$$\tilde{F}_i(\tilde{x}) = x_i + \omega_i - \omega_N - p\left(\frac{K}{N}g(x_i) + \frac{K}{N}\sum_{j=1}^{N-1} g(x_i - x_j)\right) - p\left(\frac{K}{N}\sum_{j=1}^{N-1} g(x_i)\right)$$

where $(x_1, \ldots, x_{N-1}) \in \mathbb{T}^{N-1}$ is the coordinate representation of $\tilde{x}$.

Further, we may identify the tangent space of any $\tilde{x} \in \mathbb{T}^{N-1}$ with $\mathbb{R}^{N-1}$. There, the matrix representation of $D\tilde{F}(x)$ is an $(N-1) \times (N-1)$ matrix with the i-th diagonal entry given by

$$1 - \frac{2K}{N}g'(x_i) - \frac{K}{N}\sum_{j=1, j\neq i}^{N-1} g'(x_i - x_j)$$

and the ij-th entry for $i \neq j$ given by

$$\frac{K}{N} g'(x_i - x_j) - \frac{K}{N} g'(x_j) \, .$$

Eigenvalues may be then computed at a fixed point $\tilde{x}$, and the eigenvalues λ satisfying $|\lambda| < 1$ is, as usual, equivalent to asymptotic stability for $\tilde{x}$ under the mapping $\tilde{F}$.

4 Examples

In this section we will give two examples where the coordinates for the quotient map and the resulting matrix are computed explicitly.

4.1 Two Oscillators with a Sinusoidal Coupling

Consider the system with two oscillators and the coupling function

$$g(x) = \sin(2\pi \hat{x})$$

where $\hat{x}$ is defined in Eq. (2). Then the system takes the form

$$\theta_1(t+1) = \theta_1(t) + \omega_1 - p\left(\frac{K}{2} \sin\left(2\pi(\widehat{\theta_1 - \theta_2})\right)\right)$$
$$\theta_2(t+1) = \theta_2(t) + \omega_2 - p\left(\frac{K}{2} \sin\left(2\pi(\widehat{\theta_2 - \theta_1})\right)\right)$$

The corresponding quotient map $\tilde{F}$ gives the rather simple equation

$$x(t+1) = x(t) + \omega_1 - \omega_2 - p\left(\frac{K}{2} \sin(2\pi \hat{x}(t))\right) - p\left(\frac{K}{2} \sin(2\pi \hat{x}(t))\right) \, .$$

Phase locked solutions correspond to fixed points of this quotient map, given by

$$\omega_1 - \omega_2 - p\left(\frac{K}{2} \sin(2\pi \hat{x})\right) - p\left(\frac{K}{2} \sin(2\pi \hat{x})\right) = 0 \, .$$

Fixed points are easily found by solving the 1D equation in $\mathbb{R}$ given by

$$K \sin(2\pi \hat{x}) = \widehat{\omega_1 - \omega_2} \, .$$

The stability of such a solutions is determined by whether

$$1 - 2K\pi \cos(2\pi \hat{x}) \in (-1, 1) \, .$$

Phase locked solutions are then found by computing $q^{-1}(x)$, which yields the family of phase-locked solutions determined by $(x + a, a)$ where $a \in \mathbb{S}$ with phase difference x.

We see that we only have phase-locked solutions if K is sufficiently large; in particular, we need $K \geq |\widehat{\omega_1 - \omega_2}|$. In the case of $K > |\widehat{\omega_1 - \omega_2}|$ there are two solutions $\hat{x}_1$ and $\hat{x}_2$ where $|\hat{x}_1| < \frac{1}{4}$ and $\frac{1}{4} < |\hat{x}_2| \leq \frac{1}{2}$. We see that the phase-locked solution corresponding to $\hat{x}_2$ is always unstable, as $\cos(2\pi\hat{x}_2) \leq 0$.

For $\hat{x}_1$, we see that the expression

$$1 - 2K\pi\cos(2\pi\hat{x}_1) = 1 - 2\pi K\sqrt{1 - \left(\frac{\widehat{\omega_1 - \omega_2}}{K}\right)^2}.$$

The expression inside the squareroot is positive by assumption, so the above expression is real and always smaller than 1. A routine computation shows that this expression is only greater than -1 if $K < \sqrt{\frac{1}{\pi^2} + \left(\widehat{\omega_1 - \omega_2}\right)^2}$.

We summarize our findings in the following Theorem.

Theorem 1 *The system*

$$\theta_1(t+1) = \theta_1(t) + \omega_1 - p\left(\frac{K}{2}\sin\left(2\pi(\widehat{\theta_1 - \theta_2})\right)\right)$$
$$\theta_2(t+1) = \theta_2(t) + \omega_2 - p\left(\frac{K}{2}\sin\left(2\pi(\widehat{\theta_2 - \theta_1})\right)\right)$$

has both a unique family of unstable phase-locked solutions and a unique family of asymptotically stable phase-locked solutions if

$$|\widehat{w_1 - w_2}| < K < \sqrt{\frac{1}{\pi^2} + \left(\widehat{\omega_1 - \omega_2}\right)^2}.$$

If instead

$$K < |\widehat{w_1 - w_2}|$$

then there are no phase-locked solutions, and if

$$K > \sqrt{\frac{1}{\pi^2} + \left(\widehat{\omega_1 - \omega_2}\right)^2}$$

then there are two phase-locked families of unstable phase-locked solutions.

Let us look at an example. Let us consider $\hat{\omega}_1 = \frac{1}{4}, \hat{\omega}_2 = \frac{1}{8}$, so that $|\widehat{w_1 - w_2}| = 0.125$ and $\sqrt{\frac{1}{\pi^2} + \left(\widehat{\omega_1 - \omega_2}\right)^2} \approx 0.34197$. A plot of $\theta_1(t)$ and $\theta_2(t)$ for

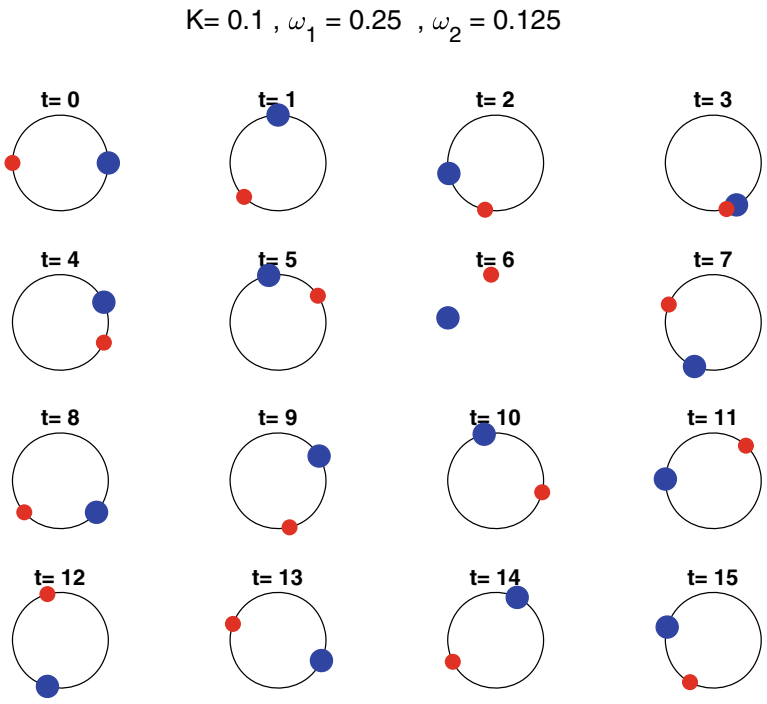

Fig. 2 A plot of the first 16 iterates when $K = 0.1$, $\hat{\omega}_1 = \frac{1}{4}$, and $\hat{\omega}_2 = \frac{1}{8}$ is shown. No phase-locking is observed, as expected for these values as there are no fixed points in the quotient map. The two oscillators act incoherently for all t

$t = 0, 1, \ldots, 15$ for coupling values $K = 0.1$, $K = 0.25$, and $K = 0.4$ are shown in Figs. 2, 3, and 4 respectively. In the figures, θ_1 is shown as the larger dot. The initial condition used was to start θ_1 and θ_2 on opposite sides of the circle. With $K < |\widehat{w_1 - w_2}|$ in Fig. 2, no phase-locking is observed, and the oscillators act as if uncoupled, with θ_1 rotating at a faster frequency and lapping θ_2. When $K = 0.25$ we are in the stability range $|\widehat{w_1 - w_2}| < K < \sqrt{\frac{1}{\pi^2} + \left(\widehat{\omega_1 - \omega_2}\right)^2}$. The orbits are depicted in Fig. 3, and we see that by $t = 5$ the two oscillators begin to lock their phases and rotate in unison thereafter. When $K = 0.4$, shown in Fig. 4, we have $K > \sqrt{\frac{1}{\pi^2} + \left(\widehat{\omega_1 - \omega_2}\right)^2}$. There the phase-locked families are unstable, and we instead observe a period 2 phase-locked orbit, where even t are locked at a different phase than odd t.

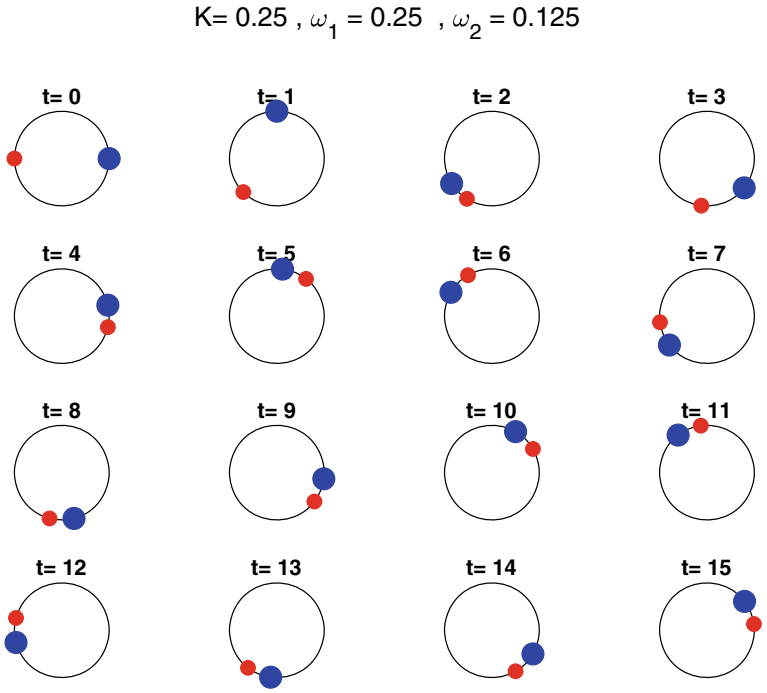

Fig. 3 A plot of the first 16 iterates when $K = 0.25$ $\hat{\omega}_1 = \frac{1}{4}$, and $\hat{\omega}_2 = \frac{1}{8}$ is shown. For these parameter values the quotient map has a stable fixed point. The oscillators quickly become phase-locked with the phase-difference rapidly converging to $\frac{1}{12}$ as t grows. They will continue to rotate in unison with this fixed separation thereafter

4.2 *Five Oscillators with Equal Rotation Frequencies and a Piecewise Linear Coupling*

Let us consider the system where $\omega_i = \omega,\ i = 1, \ldots, 5$ and the piecewise linear coupling $g_0 : \mathbb{S} - \{p(1/2)\} \to \mathbb{R}$ with $g_0(x) = \hat{x}$, with the $\hat{}$ operation defined in Eq. (2). A plot of $g_0 \circ p : \mathbb{R} \to \mathbb{R}$ is shown in Fig. 5. It has discontinuities whenever $p(x) = 1/2$. The coupling function used in this system is also found in [13]. There the intuition behind this map is given, namely that $g_0(x_i - x_j)$ measures what they call the "oriented distance" between x_i and x_j, and this distance is undefined when two points are directly opposite, explaining why we exclude one of the points in the domain of g_0. See Fig. 6 for some examples.

Nonetheless, we can still look for phase-locked solutions that stay away from the discontinuity set. Our equations take the form

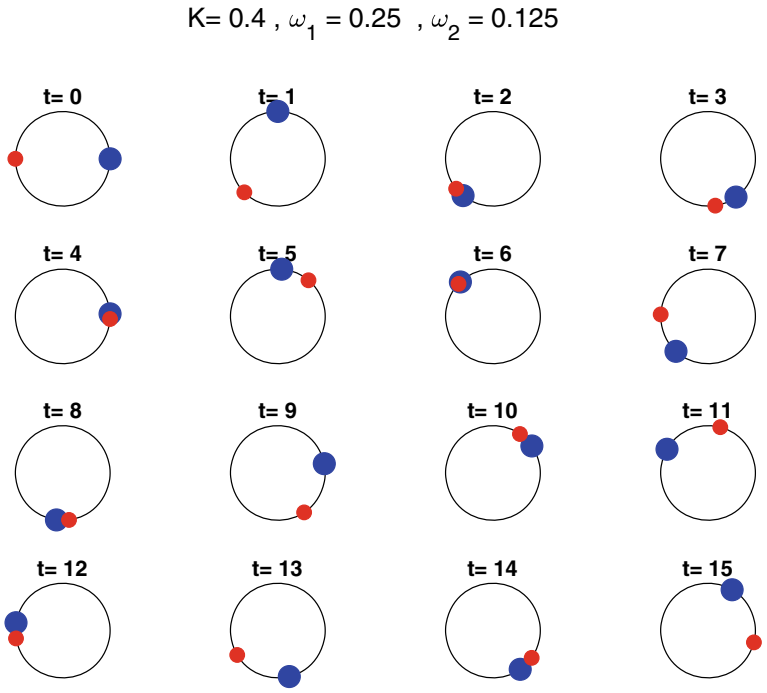

Fig. 4 A plot of the first 16 iterates when $K = 0.4$, $\hat{\omega}_1 = \frac{1}{4}$, and $\hat{\omega}_2 = \frac{1}{8}$ is shown. For these values both fixed points of the quotient map are unstable. The oscillators appear to settle on a stable period 2 phase-locked orbit where odd t have a phase-difference of $\widehat{\theta_1 - \theta_2} \approx 0.205$ while even t have $\widehat{\theta_1 - \theta_2} \approx -0.054$. This suggests the initially stable fixed point undergoes a period doubling bifurcation in the quotient map as K crosses the stability threshold of 0.34197. This period 2 orbit in the quotient map corresponds to the continued rotation with alternating phase-differences seen here

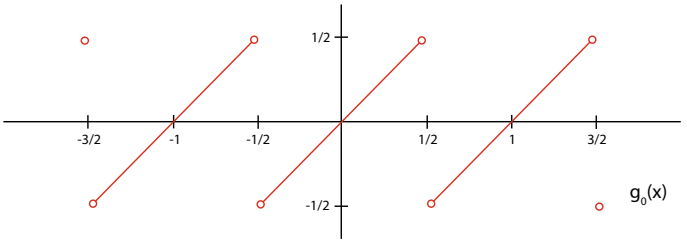

Fig. 5 A plot of $g_0 \circ p$ is shown. The function is discontinuous whenever $p(x) = 1/2$. As required, the function depicted is odd with period 1

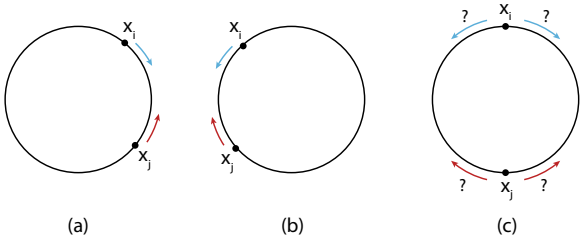

Fig. 6 The coupling function g_0 between two points on the circle gives the oriented distance between the two oscillators. The orientation of the attraction of the two oscillators is shown. In (*c*) this orientation is undefined, as is the case for any two antipodal points

$$\theta_i(t+1) = \theta_i(t) + \omega - p\left(\frac{K}{5}\sum_{j=1}^{5} g_0(\theta_i(t) - \theta_j(t))\right) .$$

To look for phase-locked solutions, we look at the four-dimensional quotient mapping with equations

$$x_i(t+1) = x_i(t) - p\left(\frac{K}{5}g_0(x_i(t)) + \frac{K}{5}\sum_{j=1}^{4} g_0(x_i(t) - x_j(t))\right) - p\left(\frac{K}{5}\sum_{j=1}^{4} g_0(x_i(t))\right) .$$

Fixed points are obtained whenever, for all $i = 1, 2, 3, 4$, the sums

$$\frac{K}{5}g_0(x_i) + \frac{K}{5}\sum_{j=1}^{4} g_0(x_i - x_j) + \frac{K}{5}\sum_{j=1}^{4} g_0(x_i)$$

yield an integer. Since g_0 maps to $(-1/2, 1/2]$, for any $K < 2$, the only possible integer is 0. Consequently, the expression can only be made an integer if

$$g_0(x_i) + \sum_{j=1}^{4} g_0(x_i - x_j) + \sum_{j=1}^{4} g_0(x_i) = 0 .$$

This sum can be simplified by noting that the only three possibilities for $g_0(x_i - x_j) = \widehat{x_i - x_j}$ are given by

$$g_0(x_i - x_j) = \begin{cases} \hat{x}_i - \hat{x}_j + 1, & \hat{x}_i - \hat{x}_j \le -\frac{1}{2} \\ \hat{x}_i - \hat{x}_j, & -\frac{1}{2} < \hat{x}_i - \hat{x}_j \le \frac{1}{2} \\ \hat{x}_i - \hat{x}_j - 1, & \hat{x}_i - \hat{x}_j > \frac{1}{2} \end{cases}$$

and that $g_0(x_i) = \hat{x}_i$. After some omitted calculations, one obtains fixed points of the form

$$x = (0, 0, 0, 0),\ x = (-2/5, -2/5, 2/5, 2/5),\ x = (-2/5, -1/5, 1/5, 2/5)$$

as well as the remaining 28 permutations of these solutions. Of note is that the first of these solutions corresponds to a synchronized trajectory. Each of these is easily translated to a phase-locked family in the original system.

Let us now determine their stability. Since $g'(x) = 1$, the Jacobian matrix is constant, with diagonal entries given by

$$1 - \frac{K}{5} - \frac{K}{5} \sum_{j=1, j\neq i}^{4} (1) - \frac{K}{5} = 1 - K.$$

Since the off diagonal entries are just 0, the Jacobian is a diagonal 4×4 matrix with diagonal entries $1 - K$. We conclude that all of these phase-locked solutions, including the synchronized state, are stable whenever $K < 2$. If $K > 2$, these solutions still satisfy the equations, but are unstable. There may also be other fixed points of the quotient map when $K > 2$. This leaves us with the following Theorem.

Theorem 2 *The model described by the equations*

$$\theta_i(t+1) = \theta_i(t) + \omega - p\left(\frac{K}{5}\sum_{j=1}^{5} g_0(\theta_i(t) - \theta_j(t))\right),\ i = 1, 2, 3, 4, 5$$

has 31 stable phase-locked families if $K < 2$. They are given explicitly by

$$(a, a, a, a, a)\ ,$$

the six distinct permutations of the first four coordinates of

$$\left(-\frac{2}{5} + a, -\frac{2}{5} + a, \frac{2}{5} + a, \frac{2}{5} + a, a\right)\ ,$$

and the 24 permutations of the first four coordinates of

$$\left(-\frac{2}{5} + a, -\frac{1}{5} + a, \frac{1}{5} + a, \frac{2}{5} + a, a\right)$$

where $a \in \mathbb{S}$. All 31 of these phase-locked families are unstable if $K > 2$.

We emphasize again that when we say a phase-locked family S is asymptotically stable, such as one of the families in Theorem 2, we do not mean that $x \in S$ is asymptotically stable in the classical sense. It is not true that if $x \in S$ and y is a sufficiently small perturbation of x, that $|F^n(x) - F^n(y)| \to 0$ as $n \to \infty$. What is true is that there exists a $z \in S$ such that $|F^n(z) - F^n(y)| \to 0$, but in general $z \neq x$.

5 Conclusion and Discussion

We have analyzed a discrete time model on $\mathbb{T}^N$ for a system of coupled oscillators, with an emphasis on phase-locking and synchronization. By identifying all rotations of a configuration of oscillators, we defined a quotient space homeomorphic to $\mathbb{T}^{N-1}$ with a quotient map whose fixed points correspond to families of phase-locked and synchronized solutions in the original system. Moreover, the stability of phase-locked families can be determined by examining the stability of the fixed point in the quotient space.

It is appropriate to compare the results here to that of the continuous time Kuramoto model given by Eq. 1. Though this is an oversimplification, the results may be broken into two cases, that of finite N and that of the continuum $N \to \infty$ limit. We compare only to the former and comment that there are many stability results in the continuum limit, see [7] for a thorough review of this case.

For finite N, stability results for Eq. 1 are analogous to those in this article in the sense that they use a rotating frame to translate the problem to one of fixed points, as we have done with the quotient mapping. In some cases, such as in [4], results are also given on the size of the basins of attraction by measuring what proportion of initial conditions will phase-lock. In contrast, the stability results in this article are *local* and we have not provided any estimates on the basin of attraction of the stable phase-locked families.

In the example with two oscillators in Sect. 4.1, for the parameter range where there is a single stable fixed point in the quotient map, numerical simulations suggest that the lone stable family attracts almost all initial conditions. It is likely that some of the global convergence results in difference equations on $[0, \infty)^n$ (see for instance [2, 3, 11, 18]) can be extended to systems of coupled oscillators on the torus. Future work on the global properties of phase-locking is planned.

References

1. Acebrón, J., Bonilla, L., Vicente, P., Conrad, J., Ritort, F., Spigler, R.: The Kuramoto model: a simple paradigm for synchronization phenomena. Rev. Mod. Phys. **77** (2005)
2. Baigent, S., Hou, Z.: Global stability of discrete-time competitive population models. J. Differ. Equ. Appl. **8**, 1–19 (2017)
3. Cabral Balreira, E., Elaydi, S., Luis, R.: Global stability of higher dimensional monotone maps. J. Diff. Equ. Appl. **23**(12), 2037–2071 (2017)
4. Bronski, J., DeVille, L., Park, M.: Fully synchronous solutions and the synchronization phase transition for the finite-N Kuramoto model. Chaos **22** (2012)
5. Canale, E., Monzon, P.: Almost global synchronization of symmetric Kuramoto coupled oscillators. In: Systems Structure and Control, vol. 8, pp. 167–190. InTech Education and Publishing (2008)
6. Crawford, J.: Amplitude expansions for instabilities in populations of globally-coupled oscillators. J. Stat. Phys. **74**, 1047–1084 (1994)
7. Dietert, H., Fernandez, B.: The mathematics of asymptotic stability in the Kuramoto model. Proc. Ro. Soc. A **474** (2018)

8. Dietert, H., Fernandez, B., Gérard-Varet, D.: Landau damping to partially locked states in the Kuramoto model. Commun. Pure Appl. Math. **71**, 953–993 (2018)
9. Fernandez, B., Gérard-Varet, D., Giacomin, G.: Landau damping in the Kuramoto Model. Ann. Henri Poincaré **17**, 1793—1823 (2016)
10. Gamelin, T., Greene, R.: Introduction to Topology, 2nd edn. Dover Publications (1999)
11. Hirsch, M.W.: On existence and uniqueness of the carrying simplex for competitive dynamical systems. J. Biol. Dyn. **2**(2), 169–179 (2008)
12. Jadbabaie, A., Motee, N., Barahona, M.: On the stability of the Kuramoto model of coupled nonlinear oscillators. In: American Control Conference, vol. 5, pp. 4296–4301 (2004)
13. Koiller, J., Young, L.-S.: Coupled map networks. Nonlinearity **23**(5), 1121–1141 (2010)
14. Kuramoto, Y.: Chemical Oscillations, Waves, and Turbulence. Springer, Berlin (1984)
15. Lin, Z., Francis, B., Maggiore, M.: State agreement for continuous-time coupled nonlinear systems. SIAM J. Control Optim. **46**(1), pp. 288–307 (2007)
16. Mirollo, R., Strogatz, S.: Stability of incoherence in a population of coupled oscillators. J. Stat. Phys. **63**, 613–635 (1991)
17. Panaggio, M.J., Abrams, D.M.: Chimera states: coexistence of coherence and incoherence in networks of coupled oscillators. Nonlinearity **28**, R67–R87 (2015)
18. Smith, H.: Planar competitive and cooperative difference equations. J. Differ. Equ. Appl. **3**(5–6), 335–357 (1998)
19. Strogatz, S.: From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators. Phys. D **143**, 1–20 (2000)
20. Winfree, A.: Biological rhythms and the behavior of populations of coupled oscillators. J. Theor. Biol. **16**, 15–42 (1967)

Reaching a Consensus via Krause Mean Processes in Multi-agent Systems: Quadratic Stochastic Operators

Tuncay Candan, MANSUR SABUROV, and Ünal Ufuktepe

Abstract A multi-agent system is a system composed of multiple interacting so-called *intelligent agents* who possibly have different information and/or diverging interests. The agents could be robots, humans or human teams. Opinions are at the basis of human behavior, and can be seen as the internal state of individuals that drives a certain action. Opinion dynamics is a process of individual opinions, in which a group of interacting agents continuously fuse their opinions on the same issue based on established rules to reach a *consensus* in the final stage. To some extent, *the Krause mean process* is a general model of opinion sharing dynamics in which the opinions are represented by vectors. In this paper, we present an opinion sharing dynamics by using positive quadratic stochastic operators and establish the consensus in the system.

Keywords Krause mean process · Quadratic stochastic operator · Cubic stochastic matrix · Consensus

1 Introduction

A multi-agent system is a system composed of multiple interacting so-called *intelligent agents* who possibly have different information and/or diverging interests. The agents could be robots, humans or human teams. The humans are complex individuals whose behaviors are governed by many aspects, related to social context, culture, law and other factors. In spite of these many factors, human societies are

T. Candan · M. SABUROV (✉) · Ü. Ufuktepe
College of Engineering and Technology, American University of the Middle East, Kuwait, Kuwait
e-mail: Mansur.Saburov@aum.edu.kw

T. Candan
e-mail: Tuncay.Candan@aum.edu.kw

Ü. Ufuktepe
e-mail: Unal.Ufuktepe@aum.edu.kw

S. Baigent et al. (eds.), *Progress on Difference Equations and Discrete Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 341,
https://doi.org/10.1007/978-3-030-60107-2_22

characterized by stunning global regularities in which we can see transitions from disorder to order. These macroscopic phenomena naturally call for a mathematical model to understand social behavior, i.e., a model to understand regularities at large scale as collective effects of the interaction among single individuals. Opinions are at the basis of human behavior, and can be seen as the internal state of individuals that drives a certain action. *Opinion dynamics* is a fusion process of individual opinions, in which a group of interacting agents continuously fuse their opinions on the same issue based on established fusion rules to reach a *consensus*, *polarization*, or *fragmentation* in the final stage.

In sociology, different mathematical models have been constructed to study the evolution of the opinions of a group of interacting individuals. The majority of the concerned models are linear. Typically researchers are more focused on the consensus problem and try to find out how to reach it. Historically, an idea of reaching consensus for a structured time-invariant and synchronous environment was introduced by DeGroot [6]. Later, Chatterjee and Seneta [5] generalized DeGroot's model for a structured time-varying and synchronous environment. In these models, an opinion sharing dynamics of a structured time-varying synchronous multi-agent system is presented by *the backward product* of square stochastic matrices. Meanwhile, a non-homogeneous Markov chain is presented by *the forward product* of square stochastic matrices. Therefore, the consensus in the multi-agent system and the ergodicity of the Markov chain are *dual problems* to each other. Since that time, the consensus which is the most ubiquitous phenomenon of multi-agent systems becomes popular in various scientific communities, such as biology, physics, control engineering and, social science (see [4, 14, 25, 26, 28, 43–45]). Recently, some nonlinear models have been constructed to characterize the opinion dynamics in social communities (see [12, 13, 17–20]). A more general model of the opinion sharing dynamics is the *Krause mean process* in which the opinions are represented by vectors. The reader may refer to the monograph [21] for a complete exposition of the Krause mean process. In the series of papers [32–37], the correlation between the Krause mean processes and *quadratic stochastic processes* was established.

A quadratic stochastic process (see [7, 41]) is the simplest *nonlinear Markov chain*. The analytic theory of the quadratic stochastic process generated by cubic stochastic matrices was established in [7, 41]. Historically, a quadratic stochastic operator (in short QSO) was first introduced by Bernstein [3]. The quadratic stochastic operator was considered an important source of analysis for the study of dynamical properties and modeling in various fields such as biology [15, 22], physics [46], and control system [32–37]. The fixed point sets and omega limiting sets of quadratic stochastic operators defined on the finite-dimensional simplex were studied in the references [38–40]. Ergodicity and chaotic dynamics of quadratic stochastic operators on the finite dimensional simplex were studied in the papers [29–31]. A long self-contained exposition of recent achievements and open problems in the theory of quadratic stochastic operators and processes was presented in the survey paper [9].

In this paper, we are aiming to establish a consensus in the multi-agent system in which an opinion sharing dynamics is presented by positive quadratic stochastic

operators associated with *positive cubic doubly stochastic matrices*. We also show that the proposed nonlinear protocol generates the Krause mean process.

It is also worth mentioning that there are also many recent research papers on this topic done in time scale calculus, fractional calculus (see [1, 10, 11, 23, 24]).

2 The Krause Mean Processes

We first review a general model of opinion sharing dynamics of the multi-agent system presented in [12] which encompasses all classical models of opinion sharing dynamics [2, 5, 6]. Consider a group of m individuals $\mathbf{I}_m := \{1, \ldots, m\}$ acting together as a team or committee, each of whom can specify his/her own subjective distribution for some given task. It is assumed that if the individual i is informed of the distributions of each of the other members of the group then he/she might wish to revise his/her subjective distribution to accommodate the information.

Let $\mathbf{x}(t) = (x_1(t), \ldots, x_m(t))^T$ be the subjective distributions of the multi-agent system at the time t where $x_i(t) \geq 0$ for all $1 \leq i \leq m$. Let $p_{ij}(t, \mathbf{x}(t))$ denote the weight that the individual i assigns to $x_j(t)$ when he/she makes the revision at the time $t+1$. It is assumed that $p_{ij}(t, \mathbf{x}(t)) \geq 0$ and $\sum_{j=1}^{m} p_{ij}(t, \mathbf{x}(t)) = 1$. After being informed of the subjective distributions of the other members of the group, the individual i revises his/her own subjective distribution from $x_i(t)$ to $x_i(t+1) = \sum_{j=1}^{m} p_{ij}(t, \mathbf{x}(t)) x_j(t)$.

Let $\mathbb{P}(t, \mathbf{x}(t))$ denote an $m \times m$ row-stochastic matrix whose (ij) element is $p_{ij}(t, \mathbf{x}(t))$. *A general model of the structured time-varying synchronous system* is defined as follows

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

We may then obtain all classical models [2, 5, 6, 12, 13] by choosing suitable row-stochastic matrices $\mathbb{P}(t, \mathbf{x}(t))$.

We say that *a consensus* is reached in the structured time-varying synchronous multi-agent system (1) if $\mathbf{x}(t)$ converges to $\mathbf{c} = (c, \ldots, c)^T$ as $t \to \infty$. It is worth mentioning that the consensus $\mathbf{c} = \mathbf{c}(\mathbf{x}(0))$ might depend on an initial opinion $\mathbf{x}(0)$.

A more general model of the opinion sharing dynamics is the *Krause mean process* in which the opinions are represented by vectors. The reader may refer to an excellent monograph by Krause [21] for a detailed exposition of mean processes.

Let S be a non-empty convex subset of $\mathbb{R}^d$ and S^m be the $m-$fold Cartesian product of S. A sequence $\{\mathbf{x}(t)\}_{t=0}^{\infty} \subset S^m$, $\mathbf{x}(t) = (x_1(t), \ldots, x_m(t))^T$ is called a *Krause mean process* on S^m if $x_i(t+1) \in \mathbf{conv}\{x_1(t), \ldots, x_m(t)\}$ for all $1 \leq i \leq m$ and for all $t = 0, 1, \ldots$. In other words, a sequence $\{\mathbf{x}(t)\}_{t=0}^{\infty} \subset S^m$ is the *Krause*

mean process if $\mathbf{conv}\{x_1(t+1), \ldots, x_m(t+1)\} \subset \mathbf{conv}\{x_1(t), \ldots, x_m(t)\}$ for all $t = 0, 1, \ldots$ where $\mathbf{conv}\{A\}$ is a convex hull of a set A. A mapping $T : S^m \to S^m$ is called a *Krause mean operator* if its trajectory $\{\mathbf{x}(t)\}_{t=0}^{\infty}$, $\mathbf{x}(t) = T^t(\mathbf{x}(0))$ starting from any initial point $\mathbf{x}(0) \in S^m$ generates a Krause mean process on S^m.

It is worth mentioning that the nonlinear model of opinion sharing dynamics given by (1) is a Krause mean process due to the fact that the action of a stochastic matrix $\mathbb{P} = (p_{ij})_{i,j=1}^m$ on a vector $\mathbf{x} = (x_1, \ldots, x_m)^T$ can be viewed as formation of arithmetic means $(\mathbb{P}\mathbf{x})_i = \sum_{j=1}^{m} p_{ij}x_j$ with weights p_{ij}. The various kinds of nonlinear models of mean processes have been studied in the series of papers [12, 13, 17–20].

3 The Quadratic Stochastic Processes

Let $\mathbf{I}_m := \{1, \ldots, m\}$ be a finite set and $\{\mathbf{e}_k\}_{k=1}^m$ be the standard basis of the space $\mathbb{R}^m$. Suppose that $\mathbb{R}^m$ is equipped with the l_1-norm $\|\mathbf{x}\|_1 := \sum_{k=1}^m |x_k|$ where $\mathbf{x} = (x_1, \ldots, x_m)^T \in \mathbb{R}^m$. We say that $\mathbf{x} \geq 0$ (respectively, $\mathbf{x} > 0$) if $x_k \geq 0$ (respectively, $x_k > 0$) for all $k \in \mathbf{I}_m$. Let

$$\mathbb{S}^{m-1} = \left\{\mathbf{x} \in \mathbb{R}^m : \mathbf{x} \geq 0,\ \|\mathbf{x}\|_1 = 1\right\}$$

be the $(m-1)-$dimensional standard simplex. An element of the simplex $\mathbb{S}^{m-1}$ is called *a stochastic vector*. Let $\mathbf{c} = (\frac{1}{m}, \ldots, \frac{1}{m})^T$ be the center of the simplex $\mathbb{S}^{m-1}$. Let $\mathrm{int}\mathbb{S}^{m-1} = \{\mathbf{x} \in \mathbb{S}^{m-1} : \mathbf{x} > 0\}$ and $\partial\mathbb{S}^{m-1} = \mathbb{S}^{m-1} \setminus \mathrm{int}\mathbb{S}^{m-1}$ be, respectively, an interior and boundary of the simplex $\mathbb{S}^{m-1}$.

Let us now provide some necessary definitions of non-homogeneous Markov chains and quadratic stochastic processes by following the papers [7, 8, 27, 41, 42].

Let $\mathbb{P} = (p_{ij})_{i,j=1}^m$ be a matrix, $\mathbf{p}_{i\bullet} := (p_{i1}, \ldots, p_{im})$, and $\mathbf{p}_{\bullet j} := (p_{1j}, \ldots, p_{mj})^T$ for any $i, j \in \mathbf{I}_m$. A square matrix $\mathbb{P} = (p_{ij})_{i,j=1}^m$ is called *row-stochastic* (respectively, *column-stochastic*) if $\mathbf{p}_{i\bullet}$ (respectively, $\mathbf{p}_{\bullet j}$) is a stochastic vector for all $i \in \mathbf{I}_m$ (respectively, for all $j \in \mathbf{I}_m$). We say that $\mathbb{P} \geq 0$ (respectively, $\mathbb{P} > 0$) if $\mathbf{p}_{i\bullet} \geq 0$ (respectively, $\mathbf{p}_{i\bullet} > 0$) for all $i \in \mathbf{I}_m$.

During the last few decades, the huge efforts have been made to construct various necessary and/or sufficient conditions for the ergodicity of non-homogeneous Markov chains (see [27, 42] and references therein). One of the major areas of study in non-homogeneous Markov chains is that of finding conditions under which a chain is weakly/strongly ergodic. A basic technique for doing this is to establish that all finite products are regular and then require some condition on the size of the positive entries in the transition matrices [42]. In looking for sets of square stochastic matrices which can be used in forming weakly/strongly ergodic non-homogeneous Markov chains, one needs to find subsets of regular square stochastic matrices which form semi-groups. A set of *scrambling* square stochastic matrices is one of these sets (for details see [42]). A stochastic matrix $\mathbb{P} = (p_{ij})_{i,j=1}^m$ is called *scrambling* if for any

i, j there is k such that $p_{ik}p_{jk} > 0$, i.e., any two rows of the square stochastic matrix are not orthogonal. One of the classical results in the theory of linear Markov chains states that a stochastic matrix is strongly ergodic if and only if its some power is a scrambling matrix.

A family of square row-stochastic matrices

$$\left\{\mathbb{P}^{[r,t]} = \left(p_{ik}^{[r,t]}\right)_{i,k=1}^{m} : r, t \in \mathbb{N}, t - r \geq 1\right\}$$

is called *a discrete time non-homogeneous Markov chain* if for any natural numbers r, s, t with $r < s < t$ the following condition, known as *the Chapman–Kolmogorov equation*, is satisfied

$$p_{ik}^{[r,t]} = \sum_{j=1}^{m} p_{ij}^{[r,s]} p_{jk}^{[s,t]}, \qquad 1 \leq i, k \leq m. \tag{2}$$

A linear operator $\mathscr{L}^{[r,t]} : \mathbb{S}^{m-1} \to \mathbb{S}^{m-1}$ associated with the square row-stochastic matrix $\mathbb{P}^{[r,t]} = \left(p_{ik}^{[r,t]}\right)_{i,k=1}^{m}$

$$\left(\mathscr{L}^{[r,t]}(\mathbf{x})\right)_k = \sum_{i=1}^{m} x_i p_{ik}^{[r,t]}, \qquad 1 \leq k \leq m, \tag{3}$$

is called *a linear stochastic operator* (*a Markov operator*) (see [27, 42]).

Notice that the Chapman–Kolmogorov equation can be written in the following form

$$\mathscr{L}^{[r,t]} = \mathscr{L}^{[s,t]} \circ \mathscr{L}^{[r,s]}, \qquad r < s < t. \tag{4}$$

Let $\mathscr{P} = (p_{ijk})_{i,j,k=1}^{m}$ be a cubic matrix (see [7, 8, 41]) and $\mathbf{p}_{ij\bullet} := (p_{ij1}, \ldots, p_{ijm})$ be a vector for all $1 \leq i, j \leq m$. A cubic matrix $\mathscr{P} = (p_{ijk})_{i,j,k=1}^{m}$ is called *stochastic* if $\mathbf{p}_{ij\bullet}$ is a stochastic vector for all $1 \leq i, j \leq m$.

A family of cubic stochastic matrices

$$\left\{\mathscr{P}^{[r,t]} = \left(p_{ijk}^{[r,t]}\right)_{i,j,k=1}^{m} : p_{ijk} = p_{jik},\ r, t \in \mathbb{N},\ t - r \geq 1\right\}$$

with an initial distribution $\mathbf{x}^{(0)} \in \mathbb{S}^{m-1}$ is called *a discrete time quadratic stochastic process* if for any natural numbers r, s, t with $r < s < t$ one of the following conditions, the so-called *nonlinear Chapman–Kolmogorov equations*, is satisfied

(A) $\quad p_{ijk}^{[r,t]} = \sum_{\alpha,\beta=1}^{m} p_{ij\alpha}^{[r,s]} x_\beta^{(s)} p_{\alpha\beta k}^{[s,t]}, \qquad 1 \leq i, j, k \leq m;$

(B) $$p_{ijk}^{[r,t]} = \sum_{\alpha,\beta,\gamma,\delta=1}^{m} x_\alpha^{(r)} p_{i\alpha\beta}^{[r,s]} x_\gamma^{(r)} p_{j\gamma\delta}^{[r,s]} p_{\beta\delta k}^{[s,t]}, \qquad 1 \le i, j, k \le m;$$

where $x_k^{(\nu)} = \sum_{i,j=1}^{m} x_i^{(0)} x_j^{(0)} p_{ijk}^{[0,\nu]}$.

We remark that the conditions (**A**) and (**B**) are not equivalent to each other. The reader may refer to [7, 41] for the exposition of quadratic stochastic processes. The reasons why the condition (**A**) is homogeneous degree one in $\mathbf{x}$ and the condition (**B**) is homogeneous degree two in $\mathbf{x}$ were explained in the papers [7, 41].

A nonlinear operator $\mathscr{Q}^{[r,t]} : \mathbb{S}^{m-1} \to \mathbb{S}^{m-1}$ associated with the cubic stochastic matrix $\mathscr{P}^{[r,t]} = \left(p_{ijk}^{[r,t]}\right)_{i,j,k=1}^{m}$

$$\left(\mathscr{Q}^{[r,t]}(\mathbf{x})\right)_k = \sum_{i,j=1}^{m} x_i x_j p_{ijk}^{[r,t]}, \qquad 1 \le k \le m. \tag{5}$$

is called *a quadratic stochastic operator* (*a nonlinear Markov operator*), Obviously, we have that $\mathbf{x}^{(\nu)} = \mathscr{Q}^{[0,\nu]}(\mathbf{x}^{(0)})$.

Notice that the nonlinear Chapman–Kolmogorov equation can be written in the following form

$$\mathscr{Q}^{[r,t]}(\mathbf{x}^{(r)}) = \mathscr{Q}^{[s,t]}\left(\mathscr{Q}^{[r,s]}(\mathbf{x}^{(r)})\right), \quad r < s < t. \tag{6}$$

We define the following stochastic vectors and square row-stochastic matrices associated with the cubic stochastic matrix $\mathscr{P} = (p_{ijk})_{i,j,k=1}^{m}$

$$\begin{aligned} \mathbf{p}_{ij\bullet} &:= (p_{ij1}, p_{ij2}, \ldots, p_{ijm}), && 1 \le i, j \le m, \\ \mathbb{P}_{i\bullet\bullet} &:= (p_{ijk})_{j,k=1}^{m}, && 1 \le i \le m, \\ \mathbf{P}_{\mathbf{x}} &:= \sum_{i=1}^{m} x_i \mathbb{P}_{i\bullet\bullet}, && \mathbf{x} \in \mathbb{S}^{m-1}. \end{aligned}$$

It is easy to check that the quadratic stochastic operator has the following vector and matrix forms

$$\mathscr{Q}(\mathbf{x}) = \sum_{i,j=1}^{m} x_i x_j \mathbf{p}_{ij\bullet} \qquad \text{(Vector form)} \tag{7}$$

$$\mathscr{Q}(\mathbf{x}) = \mathbf{x}^T \mathbf{P}_{\mathbf{x}} = \sum_{i=1}^{m} x_i \left(\mathbf{x}^T \mathbb{P}_{i\bullet\bullet}\right) \qquad \text{(Matrix form)} \tag{8}$$

See Sect. 5 for some examples.

Remark 1 Recall (see [16]) that a continuous mapping $\mathscr{M} : \mathbb{S}^{m-1} \to \mathbb{S}^{m-1}$ is called *a nonlinear Markov operator* if one has that $\mathscr{M}(\mathbf{x}) = \mathbf{x}^T \mathbb{M}_{\mathbf{x}}$ for any $\mathbf{x} \in \mathbb{S}^{m-1}$ where $\mathbb{M}_{\mathbf{x}} = \left(p_{ij}(\mathbf{x})\right)_{i,j=1}^{m}$ is a row-stochastic matrix depends on $\mathbf{x} \in \mathbb{S}^{m-1}$ (it introduces nonlinearity). The quadratic stochastic operator $\mathscr{Q} : \mathbb{S}^{m-1} \to \mathbb{S}^{m-1}$ given by (7) is indeed a nonlinear Markov operator since it can be written in the matrix form $\mathscr{Q}(\mathbf{x}) = \mathbf{x}^T \mathbf{P}_{\mathbf{x}}$ for any $\mathbf{x} \in \mathbb{S}^{m-1}$ defined by (8). It is worth mentioning that there are some nonlinear Markov operators which are not polynomial (see [16]). Therefore, the set of all quadratic (polynomial) stochastic operators cannot cover the set of all nonlinear Markov operators.

4 Krause Mean Processes via Quadratic Stochastic Operators

In this section, we establish some correlation with the Krause mean processes and quadratic stochastic operators. We first introduce some notions and notations.

Definition 1 A cubic matrix $\mathscr{P} = (p_{ijk})_{i,j,k=1}^{m}$ is called *stochastic* if one has that

$$\sum_{k=1}^{m} p_{ijk} = 1, \quad p_{ijk} \geq 0, \quad \forall\, 1 \leq i, j, k \leq m.$$

Definition 2 A cubic matrix $\mathscr{P} = (p_{ijk})_{i,j,k=1}^{m}$ is called *doubly stochastic* if one has that

$$\sum_{j=1}^{m} p_{ijk} = \sum_{k=1}^{m} p_{ijk} = 1, \quad p_{ijk} \geq 0, \quad \forall\, 1 \leq i, j, k \leq m.$$

Remark 2 In this paper, we *do not* require the condition $p_{ijk} = p_{jik}$ for all $i, j, k \in \mathbf{I}_m$.

Let $\mathscr{P} = (p_{ijk})_{i,j,k=1}^{m}$ be a cubic doubly stochastic matrix and $\mathbb{P}_{\bullet\bullet k} = (p_{ijk})_{i,j=1}^{m}$ be a square matrix for fixed $k \in \mathbf{I}_m$. It is clear that $\mathbb{P}_{\bullet\bullet k} = (p_{ijk})_{i,j=1}^{m}$ is also a square stochastic matrix. In the sequel, we write $\mathscr{P} = (\mathbb{P}_{\bullet\bullet 1} | \mathbb{P}_{\bullet\bullet 2} | \cdots | \mathbb{P}_{\bullet\bullet m})$ for the cubic doubly stochastic matrix.

We define a quadratic stochastic operator $\mathscr{Q} : \mathbb{S}^{m-1} \to \mathbb{S}^{m-1}$ associated with the cubic doubly stochastic matrix $\mathscr{P} = (\mathbb{P}_{\bullet\bullet 1} | \mathbb{P}_{\bullet\bullet 2} | \cdots | \mathbb{P}_{\bullet\bullet m})$ as follows

$$(\mathscr{Q}(\mathbf{x}))_k = \sum_{i,j=1}^{m} p_{ijk} x_i x_j, \qquad 1 \leq k \leq m. \tag{9}$$

We also define a linear stochastic operator $\mathscr{L}_k : \mathbb{S}^{m-1} \to \mathbb{S}^{m-1}$ associated with the square stochastic matrix $\mathbb{P}_{\bullet\bullet k} = (p_{ijk})_{i,j=1}^{m}$ as

$$\left(\mathscr{L}_k(\mathbf{x})\right)_j = \left(\mathbf{x}^T \mathbb{P}_{\bullet\bullet k}\right)_j = \sum_{i=1}^{m} p_{ijk}\, x_i, \qquad 1 \le j \le m. \tag{10}$$

It follows from (9) and (10) that

$$\left(\mathscr{Q}(\mathbf{x})\right)_k = \sum_{j=1}^{m} \left(\sum_{i=1}^{m} p_{ijk} x_i \right) x_j = \sum_{j=1}^{m} \left(\mathscr{L}_k(\mathbf{x})\right)_j x_j = \left(\mathscr{L}_k(\mathbf{x}), \mathbf{x}\right), \quad 1 \le k \le m$$

where $(\cdot, \cdot)$ stands for the standard inner product of two vectors.

Therefore, the quadratic stochastic operator $\mathscr{Q} : \mathbb{S}^{m-1} \to \mathbb{S}^{m-1}$ given by (9) can be written as follows

$$\mathscr{Q}(\mathbf{x}) = \left(\left(\mathscr{L}_1(\mathbf{x}), \mathbf{x}\right), \ldots, \left(\mathscr{L}_m(\mathbf{x}), \mathbf{x}\right) \right)^T \tag{11}$$

where $\mathscr{L}_k : \mathbb{S}^{m-1} \to \mathbb{S}^{m-1}$ is defined by (10) for all $k \in \mathbf{I}_m$.

We now define an $m \times m$ matrix as follows

$$\mathbb{P}(\mathbf{x}) = \begin{pmatrix} \left(\mathscr{L}_1(\mathbf{x})\right)_1 & \left(\mathscr{L}_1(\mathbf{x})\right)_2 & \cdots & \left(\mathscr{L}_1(\mathbf{x})\right)_m \\ \left(\mathscr{L}_2(\mathbf{x})\right)_1 & \left(\mathscr{L}_2(\mathbf{x})\right)_2 & \cdots & \left(\mathscr{L}_2(\mathbf{x})\right)_m \\ \vdots & \vdots & \ddots & \vdots \\ \left(\mathscr{L}_m(\mathbf{x})\right)_1 & \left(\mathscr{L}_m(\mathbf{x})\right)_2 & \cdots & \left(\mathscr{L}_m(\mathbf{x})\right)_m \end{pmatrix}. \tag{12}$$

We show that $\mathbb{P}(\mathbf{x})$ is doubly stochastic matrix for every $\mathbf{x} \in \mathbb{S}^{m-1}$. In fact we know that $\mathbb{P}(\mathbf{x}) = \left(p_{kj}(\mathbf{x})\right)_{k,j=1}^{m}$ where

$$p_{kj}(\mathbf{x}) = \left(\mathscr{L}_k(\mathbf{x})\right)_j = \sum_{i=1}^{m} p_{ijk} x_i. \tag{13}$$

Therefore, we have that

$$\sum_{k=1}^{m} p_{kj}(\mathbf{x}) = \sum_{k=1}^{m} \left(\sum_{i=1}^{m} p_{ijk} x_i \right) = \sum_{i=1}^{m} \left(\sum_{k=1}^{m} p_{ijk} \right) x_i = \sum_{i=1}^{m} x_i = 1,$$

$$\sum_{j=1}^{m} p_{kj}(\mathbf{x}) = \sum_{j=1}^{m} \left(\sum_{i=1}^{m} p_{ijk} x_i \right) = \sum_{i=1}^{m} \left(\sum_{j=1}^{m} p_{ijk} \right) x_i = \sum_{i=1}^{m} x_i = 1.$$

Hence, it follows from (11) and (12) that

$$\mathscr{Q}(\mathbf{x}) = \mathbb{P}(\mathbf{x})\mathbf{x} \tag{14}$$

and we call it *a matrix form* of the quadratic stochastic operator (9) associated with the cubic doubly stochastic matrix.

Remark 3 There is a relation between the matrix forms (8) and (14) of the quadratic stochastic operators. In fact, it is easy to check that for any $i \in \mathbf{I}_m$ and $\mathbf{x} \in \mathbb{S}^{m-1}$ one has

$$\mathbb{P}(\mathbf{e}_i) = (\mathbb{P}_{i\bullet\bullet})^T, \qquad \mathbb{P}(\mathbf{x}) = \mathbf{P}_{\mathbf{x}}^T, \qquad \mathscr{Q}(\mathbf{x}) = \mathbf{x}^T\mathbf{P}_{\mathbf{x}} = \left(\mathbf{P}_{\mathbf{x}}^T\mathbf{x}\right)^T = (\mathbb{P}(\mathbf{x})\mathbf{x})^T \quad (15)$$

We now present the nonlinear opinion sharing dynamics of the multi-agent system.

PROTOCOL A: Let $\mathscr{P} = (\mathbb{P}_{\bullet\bullet 1}|\mathbb{P}_{\bullet\bullet 2}|\cdots|\mathbb{P}_{\bullet\bullet m})$ be a cubic doubly stochastic matrix and let $\mathscr{Q} : \mathbb{S}^{m-1} \to \mathbb{S}^{m-1}$ be a quadratic stochastic operators associated with the cubic doubly stochastic matrix $\mathscr{P} = (\mathbb{P}_{\bullet\bullet 1}|\mathbb{P}_{\bullet\bullet 2}|\cdots|\mathbb{P}_{\bullet\bullet m})$. Suppose that an opinion sharing dynamics of the multi-agent system is generated by the quadratic stochastic operators as follows

$$\mathbf{x}^{(n+1)} = \mathscr{Q}\left(\mathbf{x}^{(n)}\right), \quad \mathbf{x}^{(0)} \in \mathbb{S}^{m-1} \quad (16)$$

where $\mathbf{x}^{(n)} = \left(x_1^{(n)}, \ldots, x_m^{(n)}\right)^T$ is the subjective distribution after n revisions.

Definition 3 We say that the multi-agent system presented by PROTOCOL A eventually reaches a consensus if $\{\mathbf{x}^{(n)}\}_{n=0}^{\infty}$ converges to the center $\mathbf{c} = (\frac{1}{m}, \ldots, \frac{1}{m})^T$ of the simplex $\mathbb{S}^{m-1}$ for any $\mathbf{x}^{(0)} \in \mathbb{S}^{m-1}$.

It follows from (14) that the opinion sharing dynamics of the multi-agent system given by PROTOCOL A can be written as

$$\mathbf{x}^{(n+1)} = \mathbb{P}\left(\mathbf{x}^{(n)}\right)\mathbf{x}^{(n)}, \quad \mathbf{x}^{(0)} \in \mathbb{S}^{m-1} \quad (17)$$

where $\mathbf{x}^{(n)} = \left(x_1^{(n)}, \ldots, x_m^{(n)}\right)^T$ is the subjective distribution after n revisions. This means that, due to the matrix form (1), the opinion sharing dynamics of the multi-agent system given by PROTOCOL A generates a Krause mean process.

Consequently, we have shown the following result.

Proposition 1 *Let $\mathscr{P} = (\mathbb{P}_{\bullet\bullet 1}|\mathbb{P}_{\bullet\bullet 2}|\cdots|\mathbb{P}_{\bullet\bullet m})$ be a cubic doubly stochastic matrix and $\mathscr{Q} : \mathbb{S}^{m-1} \to \mathbb{S}^{m-1}$ be the associated quadratic stochastic operator. Then the opinion sharing dynamics of the multi-agent system given by* PROTOCOL A *generates the Krause mean process.*

We are now ready to state the main result of this paper.

Theorem 1 *Let $\mathscr{P} = (\mathbb{P}_{\bullet\bullet 1}|\mathbb{P}_{\bullet\bullet 2}|\cdots|\mathbb{P}_{\bullet\bullet m})$ be a cubic doubly stochastic matrix and let $\mathscr{Q} : \mathbb{S}^{m-1} \to \mathbb{S}^{m-1}$ be the associated quadratic stochastic operator. If $\mathscr{P} > 0$, i.e., $p_{ijk} > 0$ for any $i, j, k \in \mathbf{I}_m$ then the opinion sharing dynamics of the multi-agent system given by* PROTOCOL A *eventually reaches a consensus.*

Proof Let $\mathscr{P} = (\mathbb{P}_{\bullet\bullet 1}|\mathbb{P}_{\bullet\bullet 2}|\cdots|\mathbb{P}_{\bullet\bullet m}) > 0$ be the positive cubic doubly stochastic matrix. Let $\{\mathbf{x}^{(n)}\}_{n=0}^{\infty}$, $\mathbf{x}^{(n+1)} = \mathscr{Q}(\mathbf{x}^{(n)})$ be a trajectory of the associated quadratic stochastic operator $\mathscr{Q} : \mathbb{S}^{m-1} \to \mathbb{S}^{m-1}$ starting from an initial point $\mathbf{x}^{(0)} \in \mathbb{S}^{m-1}$. According to the definition, the multi-agent system eventually reaches a consensus if $\{\mathbf{x}^{(n)}\}_{n=0}^{\infty}$ converges to the center $\mathbf{c} = (\frac{1}{m}, \ldots, \frac{1}{m})^T$ of the simplex $\mathbb{S}^{m-1}$.

Let $\delta(\mathbb{P}) = \frac{1}{2} \max\limits_{i_1, i_2} \sum\limits_{j=1}^{m} |p_{i_1 j} - p_{i_2 j}|$ be Dobrushin's ergodicity coefficient of a square stochastic matrix $\mathbb{P} = (p_{ij})_{i,j=1}^{m}$ (see [42]). Then we have that

$$\mathbf{x}^{(n+1)} = \mathbb{P}\left(\mathbf{x}^{(n)}\right)\mathbf{x}^{(n)} = \mathbb{P}\left(\mathbf{x}^{(n)}\right)\cdots\mathbb{P}\left(\mathbf{x}^{(1)}\right)\mathbb{P}\left(\mathbf{x}^{(0)}\right)\mathbf{x}^{(0)} \tag{18}$$

where $\mathbb{P}(\mathbf{x})$ is the square doubly stochastic matrix defined by (12). We setup for any two integer numbers $s > r$

$$\mathbb{P}^{[\mathbf{x}^{(s)}, \mathbf{x}^{(r)}]} := \mathbb{P}\left(\mathbf{x}^{(s)}\right)\mathbb{P}\left(\mathbf{x}^{(s-1)}\right)\cdots\mathbb{P}\left(\mathbf{x}^{(r+1)}\right)\mathbb{P}\left(\mathbf{x}^{(r)}\right).$$

We then obtain for any $n \geq r \geq 0$ that

$$\mathbf{x}^{(n+1)} = \mathbb{P}^{[\mathbf{x}^{(n)}, \mathbf{x}^{(0)}]}\mathbf{x}^{(0)} = \mathbb{P}^{[\mathbf{x}^{(n)}, \mathbf{x}^{(r)}]}\mathbf{x}^{(r)}.$$

Since $\mathscr{P} = (p_{ijk})_{i,j,k=1}^{m} > 0$ is the positive cubic doubly stochastic matrix, the square doubly stochastic matrices $\mathbb{P}_{1\bullet\bullet}^{T}, \ldots, \mathbb{P}_{m\bullet\bullet}^{T}$ are positive. It means that $\mathbb{P}_{1\bullet\bullet}^{T}, \ldots, \mathbb{P}_{m\bullet\bullet}^{T}$ are the scrambling matrices (see Sect. 3), i.e.,

$$\delta\left(\mathbb{P}_{i\bullet\bullet}^{T}\right) < 1, \quad \forall\, 1 \leq i \leq m.$$

We let

$$\lambda := \max_{1 \leq i \leq m} \{\delta\left(\mathbb{P}_{i\bullet\bullet}^{T}\right)\} < 1.$$

We then obtain that

$$\delta\left(\mathbb{P}(\mathbf{x})\right) = \delta\left(\mathbf{P}_{\mathbf{x}}^{T}\right) \leq \lambda < 1, \quad \forall\, \mathbf{x} \in \mathbb{S}^{m-1}. \tag{19}$$

This means that $\mathbb{P}(\mathbf{x})$ is also a scrambling (positive) matrix. Hence, we have that

$$\delta\left(\mathbb{P}^{[\mathbf{x}^{(n)}, \mathbf{x}^{(0)}]}\right) \leq \lambda^n, \quad \lim_{n\to\infty} \delta\left(\mathbb{P}^{[\mathbf{x}^{(n)}, \mathbf{x}^{(0)}]}\right) = 0$$

Therefore, the backward product of doubly stochastic matrices $\{\mathbb{P}_{\mathbf{x}^{(n)}}\}_{n=0}^{\infty}$ is strongly ergodic (see [42]), i.e., $\lim\limits_{n\to\infty} \mathbb{P}^{[\mathbf{x}^{(n)}, \mathbf{x}^{(0)}]} = m\mathbf{c}^T\mathbf{c}$ and

$$\lim_{n\to\infty} \mathbf{x}^{(n+1)} = \lim_{n\to\infty} \mathbb{P}^{[\mathbf{x}^{(n)}, \mathbf{x}^{(0)}]}\mathbf{x}^{(0)} = \mathbf{c}, \quad \mathbf{x}^{(0)} \in \mathbb{S}^{m-1},$$

where $\mathbf{c} = (\frac{1}{m}, \ldots, \frac{1}{m})^T$. This completes the proof.

Remark 4 Let us now compare the contribution of this paper with some previous results. In the series of the papers [32–37], we always assume *triple* stochasticity of cubic (hyper) matrices. However, in this paper we only assume *double* stochasticity of cubic matrices. Since we *did not* require the condition $p_{ijk} = p_{jik}$ for all $i, j, k \in \mathbf{I}_m$, in general, the double stochasticty does not imply the triple stochasticity of cubic matrices. In this sense, the result of this paper generalizes and extends these previous results.

5 An Example

We consider the following cubic doubly stochastic matrix $\mathscr{P} = (\mathbb{P}_{1\bullet\bullet}|\mathbb{P}_{2\bullet\bullet}|\mathbb{P}_{3\bullet\bullet})$ where $\mathbb{P}_{1\bullet\bullet}, \mathbb{P}_{2\bullet\bullet}, \mathbb{P}_{3\bullet\bullet}$ are square doubly stochastic matrices given as

$$\mathbb{P}_{1\bullet\bullet} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \quad \mathbb{P}_{2\bullet\bullet} = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix} \quad \mathbb{P}_{3\bullet\bullet} = \begin{pmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{pmatrix}$$

The following quadratic stochastic operator $\mathscr{Q} : \mathbb{S}^2 \to \mathbb{S}^2$ presents PROTOCOL A

$$\mathscr{Q}(\mathbf{x}) = x_1 \left(\mathbf{x}^T \mathbb{P}_{1\bullet\bullet}\right) + x_2 \left(\mathbf{x}^T \mathbb{P}_{2\bullet\bullet}\right) + x_3 \left(\mathbf{x}^T \mathbb{P}_{3\bullet\bullet}\right) = \mathbf{x}^T \mathbf{P}_{\mathbf{x}} = \left(\mathbb{P}\left(\mathbf{x}\right)\mathbf{x}\right)^T \quad (20)$$

where $\mathbf{P}_{\mathbf{x}} = x_1\mathbb{P}_{1\bullet\bullet} + x_2\mathbb{P}_{2\bullet\bullet} + x_3\mathbb{P}_{3\bullet\bullet}$ and $\mathbb{P}\left(\mathbf{x}\right) = \mathbf{P}_{\mathbf{x}}^T$ are the square doubly stochastic matrices.

It was shown in [32–37] that if the square doubly stochastic matrices $\mathbb{P}_{1\bullet\bullet}$, $\mathbb{P}_{2\bullet\bullet}, \mathbb{P}_{3\bullet\bullet}$ are positive and

$$\mathbb{P}_{1\bullet\bullet} + \mathbb{P}_{2\bullet\bullet} + \mathbb{P}_{3\bullet\bullet} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \quad (21)$$

then the consensus is established in the system described by PROTOCOL A. However, Theorem 1 improves this result. Namely, without the constraint (21), the consensus is still established in the system described by PROTOCOL A if the square doubly stochastic matrices $\mathbb{P}_{1\bullet\bullet}, \mathbb{P}_{2\bullet\bullet}, \mathbb{P}_{3\bullet\bullet}$ are (only) positive. In this sense, the result of this paper generalizes and extends all results of the papers [32–37]. It is worth mentioning that if the matrices $\mathbb{P}_{1\bullet\bullet}, \mathbb{P}_{2\bullet\bullet}, \mathbb{P}_{3\bullet\bullet}$ are merely scrambling then we may not have a consensus in the system (see [38] for some supporting examples).

Acknowledgements This work was supported by American University of the Middle East, Kuwait. The authors are greatly indebted to the reviewer for several useful suggestions and comments which improved the presentation of this paper.

References

1. Almeida, R., Girejko, E., Machado, L., Malinowska, A., Martins, N.: Application of predictive control to the Hegselmann-Krause model. Math. Meth. Appl. Sci. **41**, 9191–9202 (2018)
2. Berger, R.L.: A necessary and sufficient condition for reaching a consensus using DeGroot's method. J. Amer. Stat. Assoc. **76**, 415–418 (1981)
3. Bernstein, S.: Solution of a mathematical problem connected with the theory of heredity. Ann. Math. Statist. **13**, 53–61 (1942)
4. Cao, M., Morse, A.S., Anderson, B.D.O.: Reaching a consensus in a dynamically changing environment: a graphical approach. SIAM J. Control Optim. **47**(2), 575–600 (2008)
5. Chatterjee, S., Seneta, E.: Towards consensus: some convergence theorems on repeated averaging. J. Appl. Prob. **14**, 89–97 (1977)
6. De Groot, M.H.: Reaching a consensus. J. Amer. Stat. Assoc. **69**, 118–121 (1974)
7. Ganihodzhaev, N.: On stochastic processes generated by quadratic operators. J. Theoretical Prob. **4**, 639–653 (1991)
8. Ganikhodjaev, N., Akin, H., Mukhamedov, F.: On the ergodic principle for Markov and quadratic stochastic processes and its relations. Linear Algebra App. **416**, 730–741 (2006)
9. Ganikhodzhaev, R., Mukhamedov, F., Rozikov, U.: Quadratic stochastic operators and processes: results and open problems. Inf. Dim. Anal. Quan. Prob. Rel. Top. **14**(2), 279–335 (2011)
10. Girejko, E., Machado, L., Malinowska, A., Martins, N.: Krause's model of opinion dynamics on isolated time scales. Math. Meth. Appl. Sci. **39**, 5302–5314 (2016)
11. Girejko, E., Machado, L., Malinowska, A., Martins, N.: On consensus in the Cucker-Smale type model on isolated time scales. Discrete Contin. Dyn. Syst. S **11**(1), 77–89 (2018)
12. Hegselmann, R., Krause, U.: Opinion dynamics and bounded confidence: models, analysis and simulation. J. Art. Soc. Social Sim. **5**(3), 1–33 (2002)
13. Hegselmann, R., Krause, U.: Opinion dynamics driven by various ways of averaging. Comp. Econ. **25**, 381–405 (2005)
14. Jadbabaie, A., Lin, J., Morse, A.S.: Coordination of groups of mobile autonomous agents using nearest neighbor rules. IEEE Trans. Autom. Control **48**(6), 985–1001 (2003)
15. Kesten, H.: Quadratic transformations: A model for population growth I. Adv. App. Prob. **2**, 1–82 (1970)
16. Kolokoltsov, V.: Nonlinear Markov Processes and Kinetic Equations. Cambridge University Press (2010)
17. Krause, U.: A discrete nonlinear and non-autonomous model of consensus formation. In: Elaydi, S., et al. (eds.) Communications in Difference Equations, pp. 227–236. Gordon and Breach, Amsterdam (2000)
18. Krause, U.: Compromise, consensus, and the iteration of means. Elem. Math. **64**, 1–8 (2009)
19. Krause, U.: Markov chains, Gauss soups, and compromise dynamics. J. Cont. Math. Anal. **44**(2), 111–116 (2009)
20. Krause, U.: Opinion dynamics—local and global. In: Liz, E., Manosa, V. (eds.) Proceedings of the Workshop Future Directions in Difference Equations, pp. 113–119. Universidade de Vigo, Vigo (2011)
21. Krause, U.: Positive Dynamical Systems in Discrete Time: Theory, Models, and Applications. Walter de Gruyter (2015)
22. Lyubich, Y.I.: Mathematical Structures in Population Genetics. Springer (1992)
23. Lu, J., Yu, X., Chen, G., Yu, W.: Complex Systems and Networks: Dynamics. Springer, Controls and Applications (2016)
24. Malinowska, A., Odzijewicz, T.: Optimal control of discrete-time fractional multi-agent systems. J. Comput. Appl. Math. **339**, 258–274 (2018)
25. Moreau, L.: Stability of multiagent systems with time-dependent communication links. IEEE Trans. Autom. Control **50**(2), 169–182 (2005)
26. Olfati-Saber, R., Murray, R.M.: Consensus problems in networks of agents with switching topology and time-delays. IEEE Trans. Autom. Control **49**(9), 1520–1533 (2004)

27. Pulka, M.: On the mixing property and the ergodic principle for non-homogeneous Markov chains. Linear Algebra App **434**, 1475–1488 (2011)
28. Ren, W., Beard, R.W.: Consensus seeking in multiagent systems under dynamically changing interaction topologies. IEEE Trans. Autom. Control **50**(5), 655–661 (2005)
29. Saburov, M.: Ergodicity of nonlinear Markov operators on the finite dimensional space. Non. Anal. Theo. Met. Appl. **143**, 105–119 (2016)
30. Saburov, M.: Ergodicity of **p**−majorizing quadratic stochastic operators. Markov Processes Relat. Fields **24**(1), 131–150 (2018)
31. Saburov, M.: Ergodicity of **p**−majorizing nonlinear Markov operators on the finite dimensional space. Linear Algebra Appl. **578**, 53–74 (2019)
32. Saburov, M., Saburov, Kh: Reaching a consensus in multi-agent systems: a time invariant nonlinear rule. J. Educ. Vocat. Res. **4**(5), 130–133 (2013)
33. Saburov, M., Saburov, Kh: Mathematical models of nonlinear uniform consensus. ScienceAsia **40**(4), 306–312 (2014)
34. Saburov, M., Saburov, Kh: Reaching a nonlinear consensus: polynomial stochastic operators. Inter. J. Cont. Auto. Sys. **12**(6), 1276–1282 (2014)
35. Saburov, M., Saburov, Kh: Reaching a nonlinear consensus: a discrete nonlinear time-varying case. Inter. J. Sys. Sci. **47**(10), 2449–2457 (2016)
36. Saburov, M., Saburov, Kh: Reaching consensus via polynomial stochastic operators: a general study. Springer Proc. Math. Statist. **212**, 219–230 (2017)
37. Saburov, M., Saburov, Kh: Mathematical models of nonlinear uniformly consensus II. J. Appl. Nonlinear Dyn. **7**(1), 95–104 (2018)
38. Saburov, M., Yusof, N.A.: Counterexamples to the conjecture on stationary probability vectors of the second-order Markov chains. Linear Algebra Appl. **507**, 153–157 (2016)
39. Saburov, M., Yusof, N.: The structure of the fixed point set of quadratic operators on the simplex. Fixed Point Theory **19**(1), 383–396 (2018)
40. Saburov, M., Yusof, N.: On uniqueness of fixed points of quadratic stochastic operators on a 2D simplex. Methods Funct. Anal. Topol. **24**(3), 255–264 (2018)
41. Sarymsakov, T., Ganikhodjaev, N.: Analytic methods in the theory of quadratic stochastic processes. J. Theoretical Prob. **3**, 51–70 (1990)
42. Seneta, E.: Nonnegative Matrices and Markov Chains. Springer (1981)
43. Touri, B., Nedić, A.: Product of random stochastic matrices. IEEE Trans. Autom. Control **59**(2), 437–448 (2014)
44. Tsitsiklis, J.N.: Problems in Decentralized Decision Making and Computation. Ph.D. thesis, Department of Electrical Engineering and Computer Science, MIT (1984)
45. Tsitsiklis, J., Bertsekas, D., Athans, M.: Distributed asynchronous deterministic and stochastic gradient optimization algorithms. IEEE Trans. Autom. Control **31**(9), 803–812 (1986)
46. Ulam, S.: A Collection of Mathematical Problems. New-York, London (1960)

Global Attractivity For a Volterra Difference Equation

Kaori Saito

Abstract Sufficient conditions for the global asymptotic stability of nonlinear Volterra difference equations of convolution type, Logistic type and Volterra systems, which appear as models in science and engineering, are obtained by applying the technique of comparison method, semi-cycle theory and Liapunov functions, without the method of Z-transform.

Keywords Global asymptotic stability · Nonlinear Volterra difference equations · Comparison method · Semi-cycle theory · Liapunov functions

1 Introduction

The stability theory of nonlinear Volterra difference equations has many interesting applications in science and engineering, especially control theory, population dynamics and others [2–5, 7–10]. Moreover, interesting results and many references on stability and boundedness of solutions of Volterra difference equations may be found in [10]. Recently, Elaydi [2] surveyed some of the fundamental results on the stability and asymptotic stability of linear Volterra difference equations. The method Z-transform is heavily utilized in equations of convolution type. However, for nonlinear Volterra difference equations, it is well known that this method does not work well. Therefore, in this paper, we study the stability and asymptotic stability of nonlinear Volterra difference equations of convolution type, which is based on our manuscript [11].

Let $\mathbf{Z}$ denotes the set of all integers. For any $p, q \in \mathbf{Z}$ such that $p < q$, we define $\mathbf{Z} \ni [p, \infty) = \{p, p+1, p+2, ...\}$, $\mathbf{Z} \ni [p, q] = \{p, p+1, ..., q\}$ and $x_n = x(n)$ for $n \in [0, \infty)$.

K. Saito (✉)
Department of Industrial Information, Iwate Prefectural University, Miyako College, 1-5-1 Kanan, Miyako, Iwate 027-0039, Japan
e-mail: saitok@iwate-pu.ac.jp

S. Baigent et al. (eds.), *Progress on Difference Equations and Discrete Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 341,
https://doi.org/10.1007/978-3-030-60107-2_23

First, we consider the global asymptotic stability of the Volterra difference equation of the convolution type

$$x_{n+1} = ax_n + \sum_{j=0}^{n} b_{n-j} g(x_j), \quad n = 0, 1, 2, \ldots, \tag{1}$$

where a is a constant such that $1 > a \geq 0$, and each constants $b_j > 0$ such that

$$\sum_{j=0}^{\infty} b_j = b < \infty. \tag{2}$$

Moreover, $g(x)$ is a positive continuous monotone function such that $g(0) \geq 0$ and $0 < g^{'}(x) \leq 1$ for all $x \in \mathbf{R}^+ = (0, \infty)$.

Let $\{x_n\}$ be the solution of equation (1) with initial condition $x_0 \geq 0$. Then, from (1), $x_n \geq 0$, $n = 0, 1, 2, \ldots$. In what follows, we need the following definitions of stability (cf. [10]).

Definition 1 The bounded solution $y(n)$ of Eq. (1) with respect to initial condition y_0 is said to be;

(i) Stable (in short, S) if for any $\epsilon > 0$ there exists a $\delta(\epsilon) > 0$ such that if $|x_0 - y_0| < \delta(\epsilon)$, then $|x_n - y_n| < \epsilon$ for all $n \geq 0$, where $x(n)$ is a solution of (1) through $(0, x_0)$ such that $x_0 \geq 0$.

(ii) Asymptotically stable (in short, AS) if it is S and there exists a $\delta_0 > 0$ such that if $|x_0 - y_0| < \delta_0$, then $|x_n - y_n| \to 0$ as $n \to \infty$, where $x(n)$ is a solution of (1) through $(0, x_0)$ such that $x_0 \geq 0$.

(iii) Global attractor (in short, GA) if any initial condition x_0 (≥ 0) of Eq. (1), then $|x_n - y_n| \to 0$ as $n \to \infty$, where $x(n)$ is a solution of (1) through $(0, x_0)$ such that $x_0 \geq 0$.

(iv) Globally asymptotically stable (in short, GAS) if it is S, and GA, that is $|x_n - y_n| \to 0$ as $n \to \infty$, where $x(n)$ is a solution of (1) through $(0, x_0)$ such that $x_0 \geq 0$.

For (i) and (ii) in the above Definition 1, actually, the S is weaker than the AS and also, from (ii) and (iv), the AS is weaker than the GAS as [3] shows.

We can see the linearized equation of (1)

$$y_{n+1} = ay_n + \sum_{j=0}^{n} b_{n-j} y_j, \quad n = 0, 1, 2, \ldots, \tag{3}$$

with $y_0 \geq 0$. We have the following Lemmas that may proved using mathematical induction.

Lemma 1 *Let $\{x_n\}$ and $\{y_n\}$ be solutions of the Eqs. (1) and (3), respectively, such that*

$$0 \le x_0 \le y_0.$$

Then

$$0 \le x_n \le y_n \quad for \quad n = 0, 1, 2, \ldots.$$

Lemma 2 *let $\{\bar{x}_n\}$ be the solution of equation (1) with initial condition $\bar{x}_0 = 0$, and let $\{x_n\}$ be any solution of equation (1) with initial condition $x_0 \ge 0$. Then $\bar{x}_n \le x_n$ for $n = 1, 2, \ldots$.*

If we set $w_n = x_n - \bar{x}_n$,then from Lemma 2, we have $w_n \ge 0$

2 Global Asymptotic Stability

Theorem 1 *Under the above assumptions of section 1, we assume that condition (2) holds with $a + b < 1$, and let $\{\bar{x}_n\}$ be the solution of equation (1) with initial condition $\bar{x}_0 = 0$. Then $\{\bar{x}_n\}$ is a globally asymptotically stable solution of equation (1).*

Proof The substitution $w_n = x_n - \bar{x}_n$ transforms Eq. (1) into

$$w_{n+1} = aw_n + \sum_{j=0}^{n} b_{n-j} G^{-1} g(w_j), \quad n = 0, 1, 2, \ldots, \tag{4}$$

where function $G^{-1} := G^{-1}(x_j, \bar{x}_j) = \frac{g(x_j) - g(\bar{x}_j)}{g(w_j)}$, $g(w_j) \ne 0$ for $j = 0, 1, 2, \ldots$ and $0 < G^{-1} < 1$.

To do the proof of Theorem 1, it suffices to show that the zero solution of Eq. (4) is globally asymptotically stable. We first consider a function V defined by

$$V(w_n) = (1 - b)^{-1} \Big(w_n + \sum_{r=0}^{n-1} \sum_{s=n}^{\infty} b_{s-r} G^{-1} g(w_r) \Big). \tag{5}$$

Since for each $r = 0, 1, \ldots, n-1$ the series $\sum_{s=n}^{\infty} b_{s-r}$ converges and $\{w_n\}$ is nonnegative sequence, it follows that for every integer $n \ge 0$ the function V is well-defined and nonnegative. The function V plays the role of a "Liapunov function". Since $(1 - b)^{-1} > 1$, it is easy to see that for all integers $n \ge 0$,

$$V(w_n) \ge w_n. \tag{6}$$

Next, we prove that for every nonnegative solution $\{w_n\}$ of Eq. (4),

$$\Delta V(w_n) = V(w_{n+1}) - V(w_n) \le -dG^{-1}w_n, \quad n = 0, 1, \ldots, \tag{7}$$

where d is a constant such that $0 < d \le 1$. Using the facts that $\{w_n\}$ is a nonnegative solution of Eq. (4), $\bar{x}_n \ge 0$, and $g(w) \le w$ for $w \ge 0$, we find

$$\begin{aligned}
\Delta V(w_n) &= V(w_{n+1}) - V(w_n) \\
&= (1-b)^{-1}\Big(w_{n+1} + \sum_{r=0}^{n}\sum_{s=n+1}^{\infty} b_{s-r}G^{-1}g(w_r)\Big) \\
&\quad - (1-b)^{-1}\Big(w_n + \sum_{r=0}^{n-1}\sum_{s=n}^{\infty} b_{s-r}G^{-1}g(w_r)\Big) \\
&= (1-b)^{-1}\Big\{aw_n + \sum_{j=0}^{n} b_{n-j}G^{-1}g(w_j) \\
&\quad + \sum_{r=0}^{n}\sum_{s=n+1}^{\infty} b_{s-r}G^{-1}g(w_r)\Big\} \\
&\quad - (1-b)^{-1}\Big(w_n + \sum_{r=0}^{n-1}\sum_{s=n}^{\infty} b_{s-r}G^{-1}g(w_r)\Big) \\
&\le (1-b)^{-1}\Big\{aw_n + \sum_{j=0}^{n} b_{n-j}G^{-1}g(w_j) \\
&\quad + \sum_{r=0}^{n-1}\sum_{s=n+1}^{\infty} b_{s-r}G^{-1}g(w_r) + \sum_{s=n+1}^{\infty} b_{s-n}G^{-1}g(w_n) \\
&\quad - w_n - \sum_{r=0}^{n-1}\sum_{s=n+1}^{\infty} b_{s-r}G^{-1}g(w_r) - \sum_{r=0}^{n-1} b_{n-r}G^{-1}g(w_r)\Big\} \\
&= (1-b)^{-1}\Big(aw_n + b_0G^{-1}g(w_n) + \sum_{s=n+1}^{\infty} b_{s-n}G^{-1}g(w_n) - w_n\Big) \\
&\le (1-b)^{-1}\Big(\sum_{s=n}^{\infty} b_{s-n} - (1-a)\Big)G^{-1}g(w_n) \ \text{(by } G^{-1} < 1 \text{ and } g(w_n) \le w_n\text{)}.
\end{aligned}$$

Since $\sum_{s=n}^{\infty} b_{s-n} = b$, it follows from the above inequality that

$$\Delta V(w_n) \le (1-b)^{-1}(b - (1-a))G^{-1}g(w_n) = -dG^{-1}g(w_n), \tag{8}$$

where $d = (1-(a+b))/(1-b)$, $0 < d \le 1$, and (8) proved. From (8), it follows that the sequence $\{V(w_n)\}$ is non-increasing for all nonnegative solutions $\{w_n\}$ of Eq. (4) and so $\{V(w_n)\}$ is convergent. Thus, there exists an $\alpha \ge 0$ such that $V(w_n) \to \alpha < \infty$ for $n \to \infty$. Letting $n \to \infty$ into (8) we obtain

$$\lim_{n\to\infty} \Delta V(w_n) = \lim_{n\to\infty} (V(w_{n+1}) - V(w_n)) = 0 \le -d \lim_{n\to\infty} G^{-1} g(w_n),$$

which implies, $\lim_{n\to\infty} w_n = 0$.

To complete the proof of Theorem 1 it remains to establish the local stability of the zero solution. From (5) and (7) it follows from $G_0^{-1} < 1$ and $g(w_0) < w_0$ that

$$\begin{aligned} w_n &\le V(w_n) \le V(w_0) \\ &= (1-b)^{-1}(w_0 + \sum_{s=0}^{\infty} b_s G_0^{-1} g(w_0)) \quad (\text{where } G_0^{-1} = \frac{g(x_0) - g(\bar{x}_0)}{g(w_0)}) \\ &\le \frac{1+b}{1-b} w_0 \end{aligned}$$

from which the stability of the zero solution of Eq. (4) follows and then, the zero solution of Eq. (1) is globally asymptotically stable. Thus, the proof of theorem is complete.

3 A Volterra Difference Equation of Logistic Type

Next, we consider global attractivity of a positive equilibrium point of a nonlinear Volterra difference equation of logistic type, modeling a population of a single species, such that the present population size is affected by the sizes of earlier times due to resource availability (cf. for differential equations, see in [1, 6]).

Recently, Elaydi et al [5] have employed Liapunov-Razumikhin function techniques to investigate the stability of nonlinear functional difference equations. Then they applied their results to investigate the stability of generalized discrete logistic equations and solved some open problems [3, 5] that were raised by Kocic and Ladas [9]. In section 4, we shall give a new proof (cf. [8]) of the extended result in [5] by employing the idea of semi-cycle theory of Kocic and Ladas in [9].

We consider the following difference equation

$$x_{n+1} = x_n\{r - ax_n - \sum_{l=1}^{m} b_l g(x_{n-k_l})\}, \qquad n = 0, 1, \ldots \tag{9}$$

where $r, a, b, b_l \in \mathbf{R}$, $g(x)$ is a positive continuous monotone increasing function on $\mathbf{R}^+$, $g(0) = 0$ and,

$$\begin{aligned} 3 > r > 1, \;\; a > 0, \;\; b_l \ge 0, \quad \infty > b =: \sum_{l=1}^{m} b_l, \;\; a > b, \\ k_1, \ldots, k_m \in \mathbf{Z}^+, \;\; \text{and } k = \max(k_1, \ldots, k_m). \end{aligned} \tag{10}$$

Then, it is easy to see that Eq. (9) has a unique positive equilibrium point $x^* > 0$, which satisfies

$$ax^* + bg(x^*) = r - 1. \tag{11}$$

In the case where $g(x)$ is a linear function; $g(x) = px + q, \ \ p, q > 0$, we have

$$x^* = \frac{r - 1 - bq}{a + bp}, \quad \text{whenever} \quad r > 1 + bq.$$

Here x_n denotes the density of the population at time n and we assume the existence of positive solution $\{x_n\}$ for Eq. (9) whenever the initial conditions are such that $0 \leq x_{-j} \leq x^*$ for $j = 0, 1, 2, \ldots, k$.

4 Global Attractor

In this section we study global attractivity of the positive equilibrium point x^* of Eq. (9). First, we need the concept of the semi-cycle of a sequence (cf. [9]).

Definition 2 A positive semi-cycle of a solution $\{x_n\}$ of Eq. (9) consists of a string of terms $\{x_l, x_{l+1}, \ldots, x_m\}$, all greater than x^*, with $l \geq -k$ and $m \leq \infty$ and such that

$$either \quad l = -k \quad or \quad l > -k \quad and \quad x_{l-1} \leq x^*$$

and

$$either \quad m = \infty \quad or \quad m < \infty \quad and \quad x_{m+1} \leq x^*.$$

A negative semi-cycle of a solution $\{x_n\}$ of Eq. (9) consists of a string of terms $\{x_l, x_{l+1}, \ldots, x_m\}$, all less than x^*, with $l \geq -k$ and $m \leq \infty$ and such that

$$either \quad l = -k \quad or \quad l > -k \quad and \quad x_{l-1} \geq x^*$$

and

$$either \quad m = \infty \quad or \quad m < \infty \quad and \quad x_{m+1} \geq x^*.$$

Let

$$\{x_{p_i+1}, x_{p_i+2}, \ldots, x_{p_{i+1}}\}$$

be the ith positive semi-cycle of solution $\{x_n\}$ (i.e. $x_{p_i} \leq x^*$), and let

$$\{x_{q_i+1}, x_{q_i+2}, \ldots, x_{q_{i+1}}\}$$

be the ith negative semi-cycle of solution $\{x_n\}$ (i.e. $x_{q_i} \geq x^*$). Let x_{M_i} and x_{m_i} be the extreme values in these two semi-cycles, respectively, with the smallest possible indices M_i and m_i.

To prove Theorem 2, we need the following Lemmas.

Lemma 3 *$F \in C([0, \infty), (0, \infty))$ and F is non increasing in $[0, \infty)$.*

Lemma 4 *$F^2(x) > x$ for $0 < x < x^*$.*

Lemma 5 *We have*

$$M_i - p_i \leq k + 1 \quad \text{and} \quad m_i - q_i \leq k + 1. \tag{12}$$

Lemma 6 *Let $\lambda = \liminf_{n\to\infty} x_n = \liminf_{i\to\infty} x_{m_i}$. For ϵ $(0 < \epsilon < \lambda)$, we have*

$$x_{M_i} \leq G(\lambda - \epsilon, x_{p_i}) \tag{13}$$

where G is given by

$$\begin{aligned} G(x, y) = {} & y(r - ay - \sum_{l=1}^{m} b_l g(x))(r - ax^* - \sum_{l=1}^{m-1} b_l g(x^*) - b_m g(y)) \\ & \times (r - ax^* - \sum_{l=1}^{m} b_l g(x))^{k-1}, \end{aligned} \tag{14}$$

where x^* is only one fixed point of $F^2(x^*)$.

Lemma 7 *Equation (9) is permanent, that is, there exist numbers α and β with $0 < \alpha \leq \beta < \infty$ such that for any initial conditions $x_{-k}, \ldots, x_0 \in (0, \infty)$ there is a positive integer n_1 which depends on the initial conditions such that*

$$\alpha \leq x_n \leq \beta \text{ for } n \geq n_1. \tag{15}$$

The main result in this section is the following:

Theorem 2 *In addition to (10), suppose that x^* is the only fixed point of F^2. Then, the positive solution $\{x_n\}$ of Eq. (9) satisfies*

$$\lim_{n\to\infty} x_n = x^*.$$

Here x^* is the one in (11), and we set

$$F(x) = \begin{cases} \max_{x\leq y\leq x^*} G(x, y) & (0 \leq x \leq x^*), \\ \min_{x^*\leq y\leq x} G(x, y) & (x > x^*), \end{cases}$$

where $G(x, y)$ is satisfies (14).

Remark 1 The hypothesis of, only one fixed point x^* with $F^2(x) = x$, Theorem 2 is natural condition since $F(x^*) = x^*$ and F is non increasing function by Lemma 3. Moreover, we are able to drop this assumption of our theorem. However, We do not have the proof for it. On the other hand, condition $3 > r$ in (10) is sharp condition, which was given in [5] and the numerical test.

Now, we will start to prove the theorem. The idea of this proof is based on [8, 9].

4.1 Proof of Theorem 2

Assume for the sake of contradiction that Eq. (9) has a positive solution $\{x_n\}$ which is eventually nonnegative or is eventually non positive about x^*. We will assume that $\{x_n\}$ is eventually nonnegative. The case where $\{x_n\}$ is eventually non positive is similar and will be omitted. We claim that

$$\lim_{n\to\infty} x_n = x^*. \tag{16}$$

Let n_0 be an integer such that

$$x_{n-k_l} \geq x^* \qquad for \quad n \geq n_0 + k, \ \ l = 1, \ldots, m.$$

Then, by (8) and (11),

$$\begin{aligned} x_{n+1} &= x_n(r - ax_n - \sum_{l=1}^{m} b_l g(x_{n-k_l})) \\ &\leq x_n(r - ax^* - \sum_{l=1}^{m} b_l g(x^*)) \\ &= x_n. \end{aligned}$$

Thus the sequence $\{x_n\}$ is monotone decreasing for $n \geq n_0 + k$. Then there is an $\alpha \geq 0$ such that $\lim_{n\to\infty} x_n = \alpha$. If $\alpha > x^*$, by taking limits in (9),

$$1 = r - a\alpha - \sum_{l=1}^{m} b_l g(\alpha).$$

This is a contradiction by the uniqueness of x^*, and hence (16) holds. Therefore it remains to establish (16) when solution $\{x_n\}$ is the case of except for the above statement. To this end, we obtain (12) in Lemma 5, that is

$$M_i - p_i \leq k + 1 \qquad and \qquad m_i - q_i \leq k + 1.$$

We will prove (16) for positive semi-cycles. The proof for negative semi-cycles is similar and will be omitted. Suppose $M_i - p_i \leq k+1$ is not true. Then $M_i - p_i > k+1$. We set

$$\begin{aligned} \lambda &= \liminf_{n\to\infty} x_n = \liminf_{i\to\infty} x_{m_i}, \\ \mu &= \limsup_{n\to\infty} x_n = \limsup_{i\to\infty} x_{M_i}. \end{aligned} \tag{17}$$

which in view of (15) exist and are such that

$$0 < \alpha_1 \leq \lambda \leq x^* \leq \mu < \beta_1.$$

To complete the proof it suffices to show that

$$\lambda = \mu = x^*.$$

From (17) it follows that if $\mu \in (0, \infty)$ and $\epsilon \in (0, \lambda)$ are given, then there exists $n_2 \in N$ such that

$$\lambda - \epsilon \leq x_{n-k_l} \leq \mu \text{ for } n \geq n_2 + k, \ \ l = 1, \ldots, m.$$

We now have that, by Lemma 6,

$$x_{M_i} \leq G(\lambda - \epsilon, x_{p_i}). \tag{18}$$

Since $\lambda - \epsilon < x_{p_i} \leq x^*$, it follows from (18) that

$$x_{M_i} \leq G(\lambda - \epsilon, x_{p_i}) \leq \max_{\lambda-\epsilon \leq y \leq x^*} G(\lambda - \epsilon, y) = F(\lambda - \epsilon). \tag{19}$$

Therefore, as $\epsilon > 0$ is arbitrary, $x_{M_i} \leq F(\lambda)$ and so from (17)

$$\mu \leq F(\lambda).$$

In a similar way we can show that

$$\lambda \geq F(\mu).$$

By applying Lemma 3,

$$F(\mu) \leq \lambda \leq x^* \leq \mu \leq F(\lambda). \tag{20}$$

Then, we can show that $\lambda = x^*$. If we have that $\lambda < x^* \leq \mu$, from (19), Lemma 3 and 4,

$$\lambda \geq F(\mu) \geq F(F(\lambda)) = F^2(\lambda) > \lambda.$$

This is a contradiction. By the same argument, we can show that $\mu = x^*$. Thus, it finally follows that (16) is true and the proof is complete.

5 Volterra Difference Systems

Finally, we consider the following Volterra difference system.

$$\begin{cases} x_{n+1} = a_1 x_n + \sum_{j=0}^{n} b_{(1)n-j} g(y_j), \\ y_{n+1} = a_2 y_n + \sum_{j=0}^{n} b_{(2)n-j} g(x_j), \quad n = 0, 1, 2, \ldots, \end{cases} \qquad (E_0)$$

where a_1, a_2 are positive constants such that $0 < a_i < 1$, for $i = 1, 2$ and each constants $b_{(i)j} > 0$ for $i = 1, 2$ such that

$$a^* = \max\{a_1, a_2\} < 1 \text{ and,}$$
$$\sum_{j=0}^{\infty} b_{(i)j} = b_i^* < \infty \quad \text{for } i = 1, 2, \text{ and } b^* = \max\{b_1^*, b_2^*\} < 1. \qquad (21)$$

We can rewrite equation (E_0) to the following equation (E).

$$\begin{cases} w_{n+1} = a_1 w_n + \sum_{j=0}^{n} b_{(1)n-j} G_z^{-1} g(z_j), \\ z_{n+1} = a_2 z_n + \sum_{j=0}^{n} b_{(2)n-j} C_w^{-1} g(w_j), \quad n = 0, 1, 2, \ldots, \end{cases} \qquad (E)$$

where functions $G_w^{-1} := G_w^{-1}(x_j, \bar{x}_j) = \frac{g(x_j)-g(\bar{x}_j)}{g(w_j)}$, $g(w_j) \neq 0$ for $j = 0, 1, 2, \ldots$ and $0 < G_w^{-1} < 1$, and $G_z^{-1} := G_z^{-1}(y_j, \bar{y}_j) = \frac{g(y_j)-g(\bar{y}_j)}{g(z_j)}$, $g(z_j) \neq 0$ for $j = 0, 1, 2, \ldots$ and $0 < G_z^{-1} < 1$.

Theorem 3 *Assume that condition (21) holds with $a^* + b^* < 1$, and let $(\{\bar{x}_n\}, \{\bar{y}_n\})$ be the solution of equation (E) with initial condition $\bar{x}_0 = 0$ and $\bar{y}_0 = 0$. Then $(\{\bar{x}_n\}, \{\bar{y}_n\})$ is a globally asymptotically stable solution of equation (E).*

Remark 2 (cf. [2, 7]). It is natural to extend equation (E_0) to the general system of

$$x_{n+1} = Ax_n + \sum_{l=0}^{n} B_{n-l} g(x_l), \quad n = 0, 1, 2, \ldots \quad 0 \leq l \leq n, \qquad (22)$$

where $x \in \mathbf{R}^k$ and $A = (a_{ij})$ is a $k \times k$ real matrix such that $1 > |A| := a^{**} \geq 0$, and $B_n = (b_{(ij)n})$ is a $k \times k$ real matrix defined on $\mathbf{Z}^+$ such that

$$|b_{(ij)n}| > 0, \quad 1 \le i, j \le k, \text{ and } \sum_{n=0}^{\infty} |b_{(ij)n}| = b^{**} < \infty. \tag{23}$$

Moreover, $g(x)$ is a positive continuous monotone function such that $|g(0)| \ge 0$ and $0 < |g'(x)| \le 1$ for all nonnegative $x \in \mathbf{R}^{k(+)}$.

Then, we can obtain the similar stability result of Theorem 3 for Eq. (22) by using the extended Liapunov function:

$$V(w_n) = (1 - b^{**})^{-1}\{\sum_{i=1}^{k}(w_{(i)n} + \sum_{j=1}^{k}\sum_{r=0}^{n-1}\sum_{s=n}^{\infty} b_{(ij)s-r} G^{-1} g(w_{(j)r})\}$$

for all $j = 1, 2, \ldots, k$.

Theorem 4 *Under the above assumptions, we assume that condition (23) holds with* $a^{**} + b^{**} < 1$, *and let* $\{\hat{x}_n\}$ *be the solution of equation (22) with initial condition* $\hat{x}_0 = 0$. *Then* $\{\hat{x}_n\}$ *is a globally asymptotically stable solution of equation (22).*

References

1. Burton, T.A.: Stability and Periodic Solutions of Ordinary and Functional Differential Equations. Academic Press, INC. (1982)
2. Elaydi, S.: Stability and asymptoticity of Volterra difference equations; a Progress report. J. Comput. Appl. Math. **228**(2):504–513
3. Elaydi, S.: An Introduction to Difference Equations, 3rd edn. Springer, New York (2005)
4. Elaydi, S., Murakami, S.: Asymptotic stability versus exponential stability in linear Volterra difference equations of convolution type. J. Differ. Equ. Appl. **2**, 401–410 (1996)
5. Elaydi, S., Kocic, V.L., Li, J.: Global stability of nonlinear delay difference equations. J. Differ. Equ. Appl. **2**, 87–96 (1996)
6. Gopalsamy, K.: Time lags in Richardson's arms race model. J. Social. Biol. Struc. **4**, 303–317 (1981)
7. Gopalsamy, K.: Stability and Oscillations in Delay Differential Equations of Population Dynamics. Kluwer Academic Publishers (1992)
8. Hamaya, Y., Shinohara, Y.: On the remark to Elaydi's paper. Far East J. Dyn. Syst. **7**(2), 161–173 (2005)
9. Kocic, V.L., Ladas, G.: Global Behavior of Nonlinear Difference Equations of Higher Order with Applications. Kluwer Academic Publishers (1993)
10. Raffoul, Y.N.: Qualitative Theory of Volterra Difference Equations. Springer (2018)
11. Saito, K., Hamaya, Y.: Global attractivity for Volterra difference equations of convolution type, submitted

Bifurcation Scenarios Under Symbolic Template Iterations of Flat Top Tent Maps

Luís Silva

Abstract The behavior of orbits for iterated flat top maps has been widely studied since the dawn of discrete dynamics as a research field. However, little is known about orbit behavior if the map changes along with the iterations. In this work we consider a family of flat top tent maps and investigate in which ways the iteration pattern (symbolic template) can affect the structure of the bifurcation scenarios.

Keywords Nonautonomous dynamical systems · Bifurcations · Piecewise smooth maps · Stunted tent maps

1 Introduction

To our knowledge, the first paper dedicated to the study of flat top tent maps was [7]. If a dynamic process is generated by a one-dimensional map, then insertion of a flat segment on the map will often lead to a superstable periodic orbit. This mechanism has been widely used in the control of chaos on one-dimensional systems in areas as diverse as cardiac dynamics (see [6]), telecommunications or electronic circuits (see [1, 5, 12] and references therein). Families of flat top tent maps have also been used as models to study related families of differentiable maps, since they are closely related with symbolic dynamics and are rich enough to encompass in a canonical way all possible kneading data and all possible itineraries, see [8, 9].

Parameters in real world situations very often are not constant with time. In that cases, the evolutionary equations have to depend explicitly on time, through time-dependent parameters or external inputs. Then the classical theory of autonomous dynamical systems is no longer applicable and we get into the field of nonautonomous dynamical systems. The time dependence may be periodic or not. Nonautonomous

L. Silva (✉)
Departmental Area of Mathematics, ISEL-Lisboa, Rua Conselheiro Emídio Navarro 1, 1959-007 Lisboa, Portugal
e-mail: lfs@adm.isel.pt

S. Baigent et al. (eds.), *Progress on Difference Equations and Discrete Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 341,
https://doi.org/10.1007/978-3-030-60107-2_24

periodic dynamical systems can be used, for example, to model populations with periodic forcing, see [4].

In [11] we studied the local bifurcation structure of a family of 2-periodic nonautonomous dynamical systems, generated by the alternate iteration of two flat top tent maps.

In [10], it was introduced the idea of iteration pattern. It was considered the iteration scheme $x_{n+1} = f_{c_n}(x_n)$, where $(c_n)_{n\in\mathbb{N}_0} \in \{c_0, c_1\}^{\mathbb{N}_0}$ and $(c_0, c_1) \in \mathbb{C}^2$, for the complex logistic family $f_c(z) = z^2 + c,\ \ z \in \mathbb{C}$, and studied how the iteration pattern (symbolic template) can affect the topology of the Julia and Mandelbrot sets.

In this work we will consider an analogous iteration scheme for a family of flat top tent maps and investigate in which ways the iteration pattern (symbolic template) can affect the structure of the bifurcation scenarios.

2 Template Iterations of Flat Top Tent Maps

We will consider the family of flat top tent maps $f_u : [-1, 1] \to [-1, 1], u \in [-1, 1]$, such that

$$f_u(x) = \begin{cases} 2x + 1, & if \ \ -1 \le x \le (u-1)/2 \\ u, & if \ \ (u-1)/2 < x < (1-u)/2 \\ -2x + 1, & if \ \ (1-u)/2 \le x \le 1 \end{cases}.$$

We study iterations of two different functions, f_{u_0} and f_{u_1}, according to a general binary sequence $s \in \{0, 1\}^{\mathbb{N}_0}$ (template), in which

- the "zero" positions correspond to iterating the function f_{u_0}, and
- the "one" positions correspond to iterating the function f_{u_1}.

For fixed parameters $u_0,\ u_1 \in [-1, 1]$, and a fixed binary sequence $s = (s_n)_n \in \{0, 1\}^{\mathbb{N}_0}$, one can define the s-template orbit for any $x_0 \in [-1, 1]$ as the sequence

$$o^s_{u_0,u_1}(x_0) = (x_n)_{n\ge 0} : x_{n+1} = f_{u_{s_n}}(x_n).$$

Through this work we fix $u_1 = 1$, so f_{u_1} is the usual tent map and the parameters space is

$$[-1, 1] \times \{0, 1\}^{\mathbb{N}_0},$$

Definition 1 For fixed $s \in \{0, 1\}^{\mathbb{N}_0}$, $u \in [-1, 1]$, a point $x \in [-1, 1]$ is said to be periodic for s, u if its orbit $o^s_u(x)$ is a periodic sequence. Moreover, we say that a periodic point x for s, u is stable if there is a neighborhood J of x such that $\sigma(o^s_u(y)) = \sigma(o^s_u(x))$ for all $y \in J$, where σ is the shift map.

Remark 1 It follows immediately (see Lemma 1 in [5]) from the definition of stability that a periodic point x is stable if and only if $\sigma(o^s_u(x)) = \sigma(o^s_u(0))$. So, to study the stable periodic behavior we just have to study $o^s_u(0)$.

Remark 2 Stable periodic orbits are not the only kind of attractors. Indeed, the orbit of 0 always attracts a set of positive Lebesgue measure and if s is periodic then the set of initial values converging to this orbit is dense in $[-1, 1]$. If 0 is mapped after some iteration steps in to a point of an unstable periodic orbit, then this orbit will attract a set of positive Lebesgue measure. These orbits are commonly known as Milnor attractors and will be studied in a future work.

Throw this work we will the notation $(X_1 \dots X_n)^p$, $0 < p \leq \infty$ for the concatenation p times of the finite sequence $X_1 \dots X_n$, if $p = \infty$ then we are considering a periodic infinite sequence.

Considering the periodic template sequence $s = (011)^\infty$, and printing the orbits of 0, varying $u \in [-1, 1]$ we obtain the following bifurcation scenario, see Fig. 1

We will use Symbolic dynamics to describe the bifurcation scenarios.

Definition 2 Define the symbolic address of a point $x \in [-1, 1]$, as

$$ad(x) = \begin{cases} L, & if \ \ x < 0 \\ 0, & if \ \ x = 0 \\ R, & if \ \ x > 0 \end{cases}$$

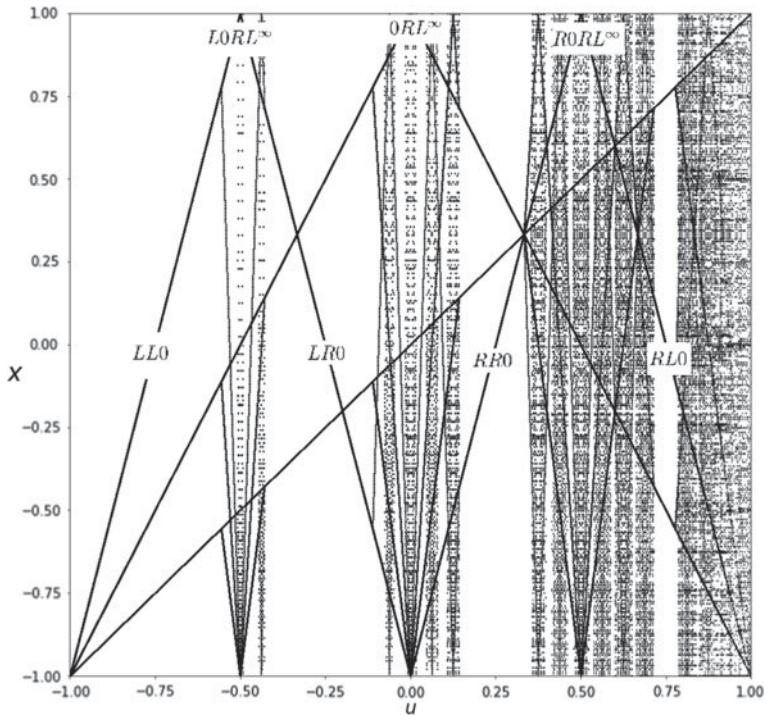

Fig. 1 Bifurcation scenario with template $s = (011)^\infty$, $u \in [-1, 1]$. For each parameter u we calculated 600 iterates with initial value $x_0 = 0$, ignored the first 100 and ploted the last 500

Definition 3 For fixed $s \in \{0,1\}^{\mathbb{N}_0}$, $u \in [-1,1]$ and a point $x \in [-1,1]$ with orbit

$$o_u^s(x) = (x_n)_{n\in\mathbb{N}},$$

define the itinerary of x as

$$I_u^s(x) = ad(x_1)ad(x_2)\ldots.$$

Let Σ be the set of sequences $X_1 \ldots X_n$ such that $X_j \in \{L, 0, R\}$ for all $j \leq n \leq +\infty$ and $\Sigma' = X_1 \ldots X_n \in \Sigma$ such that $X_n = 0$ and $X_j \neq 0$ for all $j < n$. We say that $X \in \Sigma'$ has length $|X| = n \leq +\infty$.

To each symbol $X_j \neq 0$ we associate a sign,

$$\epsilon(L) = + \ and \ \epsilon(R) = -.$$

Define

$$-L = R \ and \ \ -R = L.$$

Let $n_R(X_1 \ldots X_k) = \#\{X_j : 1 \leq j \leq k \text{ and } X_j = R\}$.

For $X \in \Sigma'$ and $0 < j < |X|$, define

$$\epsilon_j(X) = + \text{ (resp. } -\text{) if } n_R(X_1 \ldots X_j) \text{ is even (resp. odd).}$$

and

$$\epsilon(X) = \epsilon_{|X|-1}(X).$$

Considering the natural order relation $L < 0 < R$, we will introduce an order structure in Σ':

$X < Y$ if and only if there exists $r < \min\{|X|, |Y|\}$, such that $X_j = Y_j$ for all $j < r$ and $\epsilon_{r-1}(X)X_r < \epsilon_{r-1}(Y)Y_r$.

Definition 4 For $s \in \{0,1\}^{\mathbb{N}_0}$ and $u \in [-1,1]$, define the kneading data

$$K_u^s = I_u^s(0).$$

Let $\sigma(X_1X_2\ldots) = X_2\ldots$ be the usual shift map on Σ.

From now on we will restrict ourselves to the family of periodic sequences $s = (01^{p-1})^\infty$, $p \in \mathbb{N}$.

Definition 5 A sequence $X \in \Sigma$ is admissible for $(01^{p-1})^\infty$ if the following conditions are verified:

1. $\sigma^{pn}(X) \leq X$ for all $n \in \mathbb{N}$.
2. If $X_{pk} = 0$ for some k, then $X = (X_1 \ldots X_{pk-1}0)^\infty$.
3. If $X_j = 0$ for some j such that $p \nmid j$ then $\sigma^j(X) = RL^\infty$.

Denote the set of admissible sequences for $(01^{p-1})^\infty$ by $\Sigma^{01^{p-1}}$.

From now on we identify the periodic sequences $(X_1 \dots X_{pk-1}0)^\infty$ with the corresponding finite sequence $X_1 \dots X_{pk-1}0 \in \Sigma'$.

Proposition 1 *Let $X \in \Sigma$ and $s = (01^{p-1})^\infty$, then $X = K_u^s$ for some $u \in [-1, 1]$ if and only if $X \in \Sigma^{01^{p-1}}$.*

Proof Condition 2 follows from the fact that, if $X = K_u^s$ and $X_{pk} = 0$ then 0 is periodic with period pk, while Condition 3 follows from the fact that $(f_u \circ f_1)(0) = -1$ is a fixed point for all f_u, $u \in [-1, 1]$. The rest of the proof follows analogously to the proof of Theorem 2.2 in [3]. □

3 Characterization of the Bifurcation Scenarios

We will now use the kneading data to describe the bifurcation scenarios. Basically there are two main symbolic structures involved:

- $*$-product, described in Theorem 1 and Proposition 3;
- period adding, described in Theorem 2.

The idea of symbolic $*$ product was introduced in [2] for the kneading data of the quadratic family and since then it has been widely used to interpret renormalization at a symbolic level and describe period-doubling and, more generally, box-within-a-box structures in bifurcation scenarios.

Definition 6 Let $X = X_1 \dots X_{pn-1}0 \in \Sigma'$. Define

- $X^L = X_1 \dots X_{pn-1}\epsilon(X)L$.
- $X^R = X_1 \dots X_{pn-1}\epsilon(X)R$.
- $X^0 = X$.

Remark 3 It is immediate to see that, for all $X \in \Sigma'$, $\epsilon(X^L) = +$ and $\epsilon(X^R) = -$.

Definition 7 (*-product) Let $X \in \Sigma'$ and $Y = Y_1 \dots Y_n \in \Sigma$ then

$$X * Y = X^{Y_1} \dots X^{Y_n}.$$

Definition 8 A sequence $X \in \Sigma'$ is unimodal if $\sigma^n(X) \leq X$ for all $n < |X|$.

Denote the set of unimodal sequences by Σ^U.

Remark 4 Σ^U is the set of kneading sequences realized by the unimodal family, see [7].

The following two propositions are analogous, respectively, to Theorem 3.5 and Proposition 3.3 in [3] and the proofs follow analogously with the necessary adaptations.

Proposition 2 *Let $X \in \Sigma'$ and $Y \in \Sigma$, then $X * Y \in \Sigma^{01^{p-1}}$ if and only if $X \in \Sigma^{01^{p-1}}$ and $Y \in \Sigma^U$.*

Proposition 3 *Let $X \in \Sigma' \cap \Sigma^{01^{p-1}}$ and $Y_1, Y_2 \in \Sigma^U$ such that $Y_1 < Y_2$, then $X * Y_1 < X * Y_2$.*

Theorem 1 *Let $X \in \Sigma' \cap \Sigma^{01^{p-1}}$ and $X < Z \leq X * (RL^\infty)$, then $Z \in \Sigma^{01^{p-1}}$ if and only if $Z = X * Y$ for some $Y \in \Sigma^U$.*

Proof $X = (X_1 \dots X_{pk-1}0)^\infty$ and $X * RL^\infty = X^R(X^L)^\infty$, so the conditions imply that $Z = X^R \dots$.

Let us suppose that there are $j \geq 0$, minimal, and $1 \leq i < pk$, such that

$$\sigma^{pkj}(Z) = X_1 \dots X_{i-1} Y_i \dots$$

with $Y_i \neq X_i$. From the admissibility of Z, $X_1 \dots X_{i-1} Y_i < X_1 \dots X_{i-1} X_i$.

If $Z = X^R(X^L)^n X_1 \dots X_{i-1} Y_i \dots$, then, since $\epsilon(X^R(X^L)^n) = -$, we would have $Z > X * RL^\infty$ and this contradicts the hypothesis.

If, for some $n > 0$ and $m \geq 0$, $Z = X^R(X^L)^n X^R \dots X^R(X^L)_m X_1 \dots X_{i-1} Y_i \dots$, then admissibility of Z implies that $m \leq n$ but in that case $X^R(X^L)^m X_1 \dots X_{i-1} Y_i \dots > X^R(X^L)_n X_1 \dots X_{i-1} X_i \dots$ and this violates the admissibility of Z. □

From the previous theorem, for each $X \in \Sigma' \cap \Sigma^{01^{p-1}}$, the symbolic interval $[X * (L^\infty), X * (RL^\infty)] \cap \Sigma^{01^{p-1}}$ is a copy (box-within-a-box) of the space of unimodal kneading data Σ^U (in particular it contains the period-doubling sequences, see Fig. 2). These are the reducible kneading data and it can be proved that the corresponding maps are renormalizable to flat top tent maps.

Now, the period-adding structure will generate the irreducible kneading data. This kind of structure was studied in [5] for discontinuous flat top tent maps.

In our context it can be represented by the following infinite directed acyclic graph.

Consider the 2^{p-1} finite words in Σ', $B^1 < \dots < B^{2^{p-1}}$ such that $|B^j| = p$ for all j. These are the source vertices of the graph.

For $i = 1, \dots, 2^{p-1} - 1$, let

$$r_i = \min\{j : B^i_j \neq B^{i+1}_j\},$$

and

$$l(B^i) = B^i_1 \dots B^i_{r_i - 1} 0 R L^\infty.$$

From the point of view of graph theory the sequences $l(B^i)$ correspond to sink vertices(i.e., vertices with no outgoing edges). The sinks $l(B^i)$ correspond to kneading data $K_u^{01^{p-1}}$ such that $(f_1)^{r_i - 1} \circ f_u(0) = 0$.

Obviously

$$B^1 < l(B^1) < B^2 < \dots < B^{2^{p-1}-1} < l(B^{2^{p-1}-1}) < B^{2^{p-1}}.$$

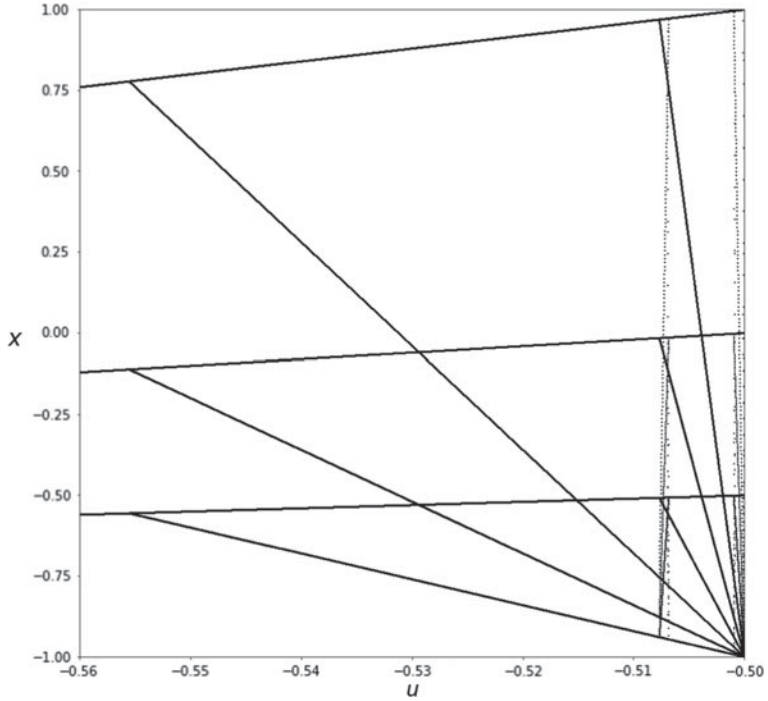

Fig. 2 Bifurcation scenario with template $s = (011)^{\infty}$, $u \in [-0.56, -0.5]$. We can observe the orbits corresponding to the period doubling sequence, $LL0$, $LL0 * R0$, $LL0 * RLR0, \ldots$

For $p = 3$ we have

$$B^1 = LL0,\ B^2 = LR0,\ B^3 = RR0,\ B^4 = RL0,$$

and

$$l(B^1) = L0RL^{\infty},\ l(B^2) = 0RL^{\infty},\ l(B^3) = R0RL^{\infty},$$

see Fig. 1.

We say that $X \in \Sigma'$ has a repeated prefix if there exists $m \in \mathbb{N}$ such that

$$X_{pm+1} \ldots X_{|X|-1} = X_1 \ldots X_{|X|-pm-1}.$$

For any set $\mathcal{W} \subset \Sigma'$, set

$$\mathcal{W} = \mathcal{W}^{\mathcal{P}} \cup \mathcal{W}^{\mathcal{NP}},$$

where $\mathcal{W}^{\mathcal{P}}$ is the subset of sequences $X \in \mathcal{W}$ such that X has a repeated prefix and $\mathcal{W}^{\mathcal{NP}} = \mathcal{W} \backslash \mathcal{W}^{\mathcal{P}}$.

We will now build recursively the levels of the period-adding graph. Denote by $\mathcal{L}_k = \mathcal{B}_k \cup l(\mathcal{B}_k)$ the k-th level in the graph, where $\mathcal{B}_k$ are the vertices with outgoing edges and $l(\mathcal{B}_k)$ the sinks in level k.

So the first level is

$$\mathcal{L}_1 = \mathcal{B}_1 \cup l(\mathcal{B}_1)$$

where

$$\mathcal{B}_1 = \{B^i,\ i = 1, \ldots, 2^{p-1}\} \text{ and } l(\mathcal{B}_1) = \{l(B^i),\ i = 1, \ldots, 2^{p-1} - 1\}.$$

Next, for general k,

$$\begin{aligned}\mathcal{B}_{k+1} = &\{(X^L)^n Y,\ X^R Y : X \in \mathcal{B}_k^{\mathcal{NP}} \text{ and } Y < X\} \bigcup \\ &\{(X^L)^n Y : X \in \mathcal{B}_k^{\mathcal{P}}, \epsilon(X) = \epsilon_{|X|-pm-1}(X) \text{ and } Y < X\} \bigcup \\ &\{X^R Y : X \in \mathcal{B}_k^{\mathcal{P}}, \epsilon(X) = \epsilon_{|X|-pm-1}(X) \text{ and } \sigma^{|X|-pm}(X) < Y < X\} \bigcup \\ &\{(X^L)^n Y : X \in \mathcal{B}_k^{\mathcal{P}}, \epsilon(X) \neq \epsilon_{|X|-pm-1}(X) \text{ and } \sigma^{|X|-pm}(X) < Y < X\} \bigcup \\ &\{X^R Y : X \in \mathcal{B}_k^{\mathcal{P}}, \epsilon(X) \neq \epsilon_{|X|-pm-1}(X) \text{ and } Y < X\},\end{aligned}$$

where Y, k', n, m are such that $k' \leq k$, $Y \in \mathcal{B}_{k'}$, $n \in \mathbb{N}$, and $X_{pm+1} \ldots X_{|X|-1} = X_1 \ldots X_{|X|-pm-1}$ if $X \in \mathcal{B}_k^{\mathcal{P}}$,

$$\begin{aligned}l(\mathcal{B}_{k+1}) = &\{(X^L)^n l(Y) : X \in \mathcal{B}_k, Y \in \mathcal{B}_{k'}, k' \leq k \text{ and } (X^L)^n Y \in \mathcal{B}_{k+1}\} \bigcup \\ &\{X^R l(Y) : X \in \mathcal{B}_k, Y \in \mathcal{B}_{k'}, k' \leq k \text{ and } X^R Y \in \mathcal{B}_{k+1}\}.\end{aligned}$$

and $\mathcal{L}_{k+1} = \mathcal{B}_{k+1} \cup l(\mathcal{B}_{k+1})$.

There exists one edge (X, Y) from X to Y iff $X \in \mathcal{L}_k$, $Y \in \mathcal{L}_{k+1}$ and $Y = (X^L)^n Z$ or $Y = X^R Z$ with $Z \in \mathcal{L}_{k'}$, $k' \leq k$.

The descendants (respectively, ancestors) of a vertex X are all vertices $Y \neq X$ such that there is a directed path from X to Y (respectively, from Y to X).

The reachable set from X, $\mathcal{T}(X)$, is the vet of vertices containing X and all its descendants.

Let $\mathcal{PA} = \underset{k}{\cup}\mathcal{L}_k$ be the set of period-adding sequences, the following Theorem follows directly from the construction.

Theorem 2 $\mathcal{PA} \subset \Sigma^{01^{p-1}}$

Moreover, let $X, Y, Z, Z' \in \mathcal{PA}$, then:

- *If $X < Y$, $\mathcal{T}(X) \cap \mathcal{T}(Y) = \emptyset$, $Z \in \mathcal{T}(X)$ and $Z' \in \mathcal{T}(Y)$ then $Z < Z'$.*
- *If $X \in \mathcal{B}_k$, $Z, Z' \in \mathcal{B}_{k'}$ with $k' \leq k$ are such that $X^L Z$ and $X^L Z'$ belong to $\mathcal{B}_{k+1}$ and $Z < Z'$ then, for all $n \in \mathbb{N}$*

$$(X^L)^n Z < (X^L)^n l(Z) < (X^L)^{n+1} Z < X^L Z' < X.$$

- *If $X \in \mathcal{B}_k$, $Z, Z' \in \mathcal{B}_{k'}$ with $k' \leq k$ are such that $X^R Z$ and $X^R Z'$ belong to $\mathcal{B}_{k+1}$ and $Z < Z'$ then, for all $n \in \mathbb{N}$*

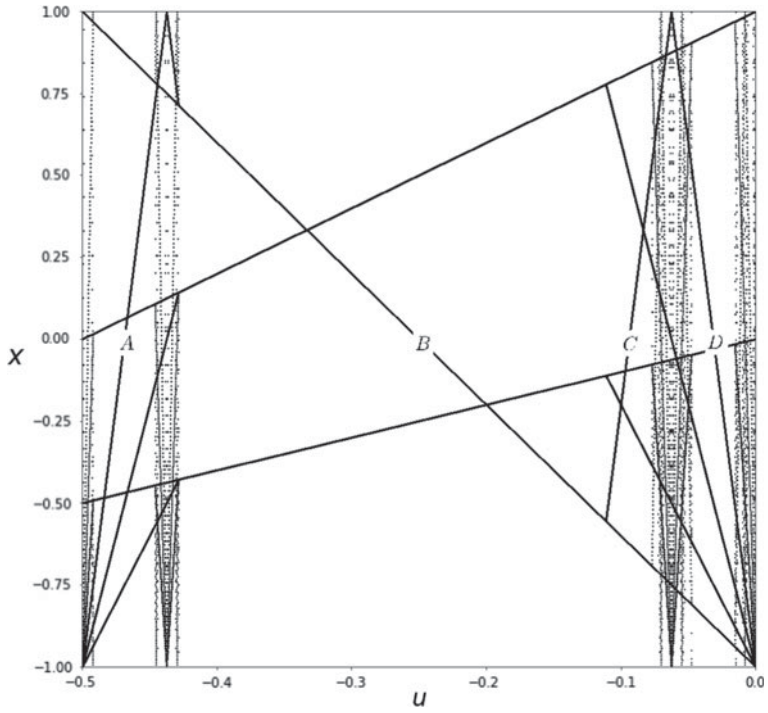

Fig. 3 Template $s = (011)^{\infty}$, $u \in [-0.5, 0]$. Here we can observe the interlacement of $*$-product and period adding, we pointed the parameters corresponding to the sequences $A = LR0^L LL0 < B = LR0 < C = LR0 * R0 < D = LR0^R LL0$

$$X < X^R Z' < X^R l(Z) < X^R Z.$$

Now we will study the interaction between the two stuctures, see Fig. 3.

Proposition 4 *If $Z \in X * \Sigma^U$ for some $X \in \Sigma^{01^{p-1}}$ then $Z^L W \notin \Sigma^{01^{p-1}}$ for any $W : W < X$.*

Proof Let $Z = X * Y$. We will make the proof considering $\epsilon(Z) = +$, being completely analogous when $\epsilon(Z) = -$.

1. $\epsilon(X) = +$.
   $\epsilon(Z) = \epsilon(X) = +$ implies that $n_R(Y) > 1$, so

   $$Y = RL^k \dots RL^{k'} 0$$

   with $k' < k$ ($k' \neq k$ because $RL^k 0 > RL^k R \dots$), then $\epsilon(XR(XL)^{k'}) = -$, $Z_{|X|(k'+2)} = L$ and $XR(XL)^{k'} XLW > XR(XL)^{k'} XLX \dots = Z$.

2. $\epsilon(X) = -$.
   This case follows immediately because $X_1 \dots X_{|X|-1} LW > X_1 \dots X_{|X|-1} LX_1 \dots$.

□

Proposition 5 *If $Z \in X * \Sigma^U$ for some $X \in \Sigma^{01^{p-1}}$ and $Z^R W \in \Sigma^{01^{p-1}}$ for any $W < X$ then $Z = X * RL^k 0$ and $Z^R W > X * RL^\infty$.*

Proof Let $Z = X * Y$. We will first prove that, if $Y = RL^k 0$, then $Z^R W \in \Sigma^{01^{p-1}}$.

$Z = X^R (X^L)^k X 0$, so if $\epsilon(X) = +$ then $\epsilon(Z) = -$ and $Z^R W = X^R (X^L)^k X^L W \in \Sigma^{01^{p-1}}$ because $X^L < X^R$.

If $\epsilon(X) = -$ then $\epsilon(Z) = +$ and $Z^R W = X^L (X^R)^k X^R W \in \Sigma^{01^{p-1}}$ because $X^R < X^L$.

Let us now suppose that $Y = RL^k \dots RL^{k'} 0$ with $k' < k$.

If $\epsilon(X) = \epsilon(Z)$ then $\sigma^{|X|(|Y|-1)}(Z^R W) = X^R W > X^R X \dots = Z$.

Analogously, if $\epsilon(X) \neq \epsilon(Z)$ then $\sigma^{|X|(|Y|-k'-2)}(Z^R W) = X^R (X^L)^{k'+1} W > X^R (X^L)^{k'+1} X \dots = Z$. □

Finally, the following result follows from Theorems 1 and 2.

Proposition 6 *If $X \in \mathcal{B}_k$, $Y \in \mathcal{B}_{k'}$, $k' \leq k$ are such that $(X^L)^n Y \in \mathcal{B}_{k+1}$ then $(X^L)^n (Y * Z) \in \Sigma^{01^{p-1}}$, for any $Z \in \Sigma^U$.*

Analogously, if $X \in \mathcal{B}_k$, $Y \in \mathcal{B}_{k'}$, $k' \leq k$ are such that $X^R Y \in \mathcal{B}_{k+1}$ then $X^R (Y * Z) \in \Sigma^{01^{p-1}}$, for any $Z \in \Sigma^U$.

References

1. Avrutin, V., Futter, B., Gardini, L., Schanz, M.: Unstable orbits and Milnor attractors in the discontinuous flat top tent map, ESAIM: PROCEEDINGS Danièle Fournier-Prunaret. Laura Gardini and Ludwig Reich, Editors **36**, 126–158 (2012)
2. Derrida, B., Gervois, A., Pomeau, Y.: Iteration of endomorphisms on the real axis and representation of numbers, pp. 305–356. XXIX, Ann. Inst. Henri Poincaré A (1978)
3. Franco, N., Silva, L., Simões, P.: Symbolic dynamics and renormalization of nonautonomous k periodic dynamical systems. J. Differ. Equ. Appl. **19**, 27–38 (2013)
4. Franke, J., Yakubu, A.: Population models with periodic recruitment functions and survival rates. J. Differ. Equ. Appl. **11**, 1169–1184 (2005)
5. Futter, B., Avrutin, V., Schanz, M.: The discontinuous flat top tent map and the nested period incrementing bifurcation structure. Chaos Solitons Fractals **45**, 465–482 (2012)
6. Glass, L., Zeng, W.: Bifurcations in flat-topped maps and the control of cardiac chaos. Int. J. Bifurcation Chaos **4**, 1061–1067 (1994)
7. Metropolis, N., Stein, M.L., Stein, P.R.: On finite limit sets for transformations on the unit interval. J. Comb. Theory A **15**(1), 25–44 (1973)
8. Milnor, J., Tresser, C.: On entropy and monotonicity for real cubic maps. Comm. Math. Phys. **209**, 123–178 (2000)
9. Rădulescu, A.: The connected isentropes conjecture in a space of quartic polynomials. Discrete Contin. Dyn. Syst. **19**, 139–175 (2007)

10. Rădulescu, A, Pignatelli, A.: Symbolic template iterations of complex quadratic maps. Nonlinear Dyn. **84** (4), 2025–2042 (2016)
11. Silva, L., Rocha, J.L., Silva, M.T.: Bifurcations of 2-periodic nonautonomous stunted tent systems. Int. J. Bifurcation Chaos **27**, 1730020 (2017) [17 pages]
12. Wagner, C., Stoop, R.: Renormalization approach to optimal limiter control in 1-D Chaotic systems. J. Statist. Phys. **106**, 97–106 (2002)

Linear Operators Associated with Differential and Difference Systems: What Is Different?

Petr Zemánek

Abstract The existence of a densely defined operator associated with (time-reversed) discrete symplectic systems is discussed and the necessity of the development of the spectral theory for these systems by using linear relations instead of operators is shown. An explanation of this phenomenon is provided by using the time scale calculus. In addition, the density of the domain of the maximal linear relation associated with the system is also investigated.

Keywords Discrete symplectic system · Linear hamiltonian differential system · Linear relations · Multi-valuedness · Densely defined operator · Time scale

1 Introduction

The study of the spectral theory of linear operators acting on a (finite or infinite dimensional) Hilbert space is a classical topic in functional analysis. The development of this theory for operators associated with differential equations or systems can be seen (from the mathematical point of view) as one of the cornerstones in the mathematical physics. Roughly speaking, quantum mechanics is Hilbert space theory (or vice versa), see e.g. [27, 28]. However, from [3, 4, 6, 10, 14] we may observe that even difference equations or systems should not be ignored in this direction. Hence, it is not very surprising that the spectral theory of linear operators associated with difference equations or systems attracts more and more attention in the last two decades. Nevertheless, it remains significantly underdeveloped except for some special cases as for the Jacobi and CMV operators in [11, 15, 26]. In the present note we aim to point out a phenomenon concerning the foundations of the theory of linear operators given by certain differential and difference expressions, which

P. Zemánek (✉)
Faculty of Science, Department of Mathematics and Statistics, Masaryk University, Kotlářská 2, 61137 Brno, Czech Republic
e-mail: zemanekp@math.muni.cz
URL: http://www.math.muni.cz/∼zemanekp/

S. Baigent et al. (eds.), *Progress on Difference Equations and Discrete Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 341,
https://doi.org/10.1007/978-3-030-60107-2_25

is from time to time overlooked by some authors (including the unification of these theories based on the time scale calculus).

In our recent works [12, 31] we established the foundations of this theory in connection with the time-reversed discrete symplectic systems depending linearly on the spectral parameter, i.e.,

$$z_k(\lambda) = \mathbb{S}_k(\lambda) z_{k+1}(\lambda), \quad k \in \mathcal{I}_{\mathbb{z}}, \tag{S_λ}$$

where $\lambda \in \mathbb{C}$ is the spectral parameter and $\mathbb{S}_k(\lambda) := \mathcal{S}_k + \lambda \mathcal{V}_k$ with the $2n \times 2n$ complex-valued matrices $\mathcal{S}_k$ and $\mathcal{V}_k$ such that

$$\begin{gathered} \mathcal{S}_k^* \mathcal{J} \mathcal{S}_k = \mathcal{J}, \quad \mathcal{V}_k^* \mathcal{J} \mathcal{S}_k \text{ is Hermitian,} \\ \mathcal{V}_k^* \mathcal{J} \mathcal{V}_k = 0, \quad \text{and} \quad \Psi_k := \mathcal{J} \mathcal{S}_k \mathcal{J} \mathcal{V}_k^* \mathcal{J} \geq 0 \end{gathered} \tag{1}$$

for the skew-symmetric $2n \times 2n$ matrix $\mathcal{J} := \left(\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}\right)$ and the superscript $*$ meaning the conjugate transpose. Furthermore, $\mathcal{I}_{\mathbb{z}}$ denotes a discrete interval, which is finite or unbounded from above. It should be also pointed out that the assumptions concerning the matrix $\mathcal{V}_k$ in (1) or Ψ_k in (2) below are naturally forced by the *Lagrange identity*, which is the main tool in the study of square summable solutions of system (S_λ), see [12, Theorem 2.5]. The term "symplectic" refers to the fact that the first three conditions in (1) are equivalent to the symplectic-type equality $\mathbb{S}_k^*(\bar{\lambda}) \mathcal{J} \mathbb{S}_k(\lambda) = \mathcal{J}$ for all $(k, \lambda) \in \mathcal{I}_{\mathbb{z}} \times \mathbb{C}$. Discrete symplectic systems were established in [7] as the natural generalization of the second order Sturm–Liouville difference equations and as the proper discrete analogue of linear Hamiltonian differential systems (4) below, see also [1]. They play an important role in the discrete Hamiltonian mechanics, numerical analysis of Hamiltonian differential systems, discrete variational theory, numerical optimal control, and in the theory of continued fractions.

Although system (S_λ) is determined by the pair of matrices $\mathcal{S}_k, \mathcal{V}_k$ satisfying (1) it can be alternatively given by the pair $\mathcal{S}_k, \Psi_k$ satisfying

$$\mathcal{S}_k^* \mathcal{J} \mathcal{S}_k = \mathcal{J}, \quad \Psi_k^* = \Psi_k \geq 0, \quad \text{and} \quad \Psi_k \mathcal{J} \Psi_k = 0, \tag{2}$$

because it follows from (1) that $\mathcal{V}_k = -\mathcal{J} \Psi_k \mathcal{S}_k$ for all $k \in \mathcal{I}_{\mathbb{z}}$. In any case, the matrix Ψ_k is absolutely crucial in study of the spectral theory for system (S_λ), because it appears as the weight matrix in the associated semi-inner product and it enables us to write system (S_λ) by using the linear map $\mathscr{L}(z)_k := \mathcal{J}(z_k - \mathcal{S}_k z_{k+1})$ as

$$\mathscr{L}(z(\lambda))_k = \lambda \Psi_k z_k(\lambda).$$

However we illustrated in [12, Example 5.1 and Remark 5.3] that this natural map *may not* give rise to a linear operator. Hence, the results of [12, 31] were phrased by using the concept of linear relations instead of operators. At the same time, without going into further details and only for completeness, we derived a sufficient condition for the existence of a densely defined (minimal) operator associated with

system (S_λ), see [12, Theorem 5.4]. In the present paper we return to this result and give "a very explicit characterization" of system (S_λ) satisfying the latter condition. More specifically, we show that (except the trivial case $\Psi_k \equiv 0$) the natural maximal operator associated with the map $\mathscr{L}(\cdot)$ is *never well-defined* (i.e., it must be *multivalued*) and the density of the domain of the corresponding minimal operator is *always violated*, see Corollary 1. This fact means that the approach based on linear relations is the most proper way for the development of the spectral theory for the discrete symplectic mapping $\mathscr{L}(\cdot)$. In addition, we attempt to explain this phenomenon by using the time scale calculus, from which we shall see the "singularity" of the truly possible single-valuedness and density in the purely continuous time case, see Theorems 3 and 4. Finally, we discuss also the density of the domain of the maximal linear relation.

The paper is organized as follows. In the next section, for a better insight into the problem at hand, we summarize the situation in the cases of the second order Sturm–Liouville differential and difference equations and the linear Hamiltonian differential systems. The main result is established in Sect. 3 and the time scale explanation is provided in Sect. 4.

2 Motivation

The traditional approach to the spectral theory requires the existence of a densely defined operator, because only in that case the classical adjoint operator is well-defined. In the simplest case, we can take the operator associated with the second order Sturm–Liouville differential expression

$$(\tau_{\mathrm{csl}}\, y)(t) := \frac{1}{w(t)}\Big\{-\big[p(t)\,y'(t)\big]' + q(t)\,y(t)\Big\} \tag{3}$$

acting on the interval $[a, b)$, where $-\infty < a < b \le \infty$ and the coefficients $p, q, w : [a, b) \to \mathbb{R}$ are (locally) integrable on $[a, b)$ with $p(t) \neq 0$ and $w(t) > 0$ for almost all $t \in [a, b)$. If we denote by L^2_W the Hilbert space of (equivalence classes of) measurable functions $y : [a, b) \to \mathbb{C}$ such that the function $w\,|y|^2$ is integrable over $[a, b)$, then the corresponding maximal operator $T^{\mathrm{csl}}_{\max}\, y := \tau_{\mathrm{csl}}\, y$ is generated by τ_{csl} on the domain

$$\operatorname{dom} T^{\mathrm{csl}}_{\max} := \Big\{y \in L^2_W \mid y \text{ and } p\,y' \text{ are (locally) absolutely continuous in} [a, b) \text{ and } \tau_{\mathrm{csl}}\, y \in L^2_W\Big\},$$

while the minimal operator is defined as $T^{\mathrm{csl}}_{\min} := \overline{T^{\mathrm{csl}}_0}$, i.e., as the closure of the pre-minimal operator, which is given by the restriction of the maximal operator to

$$\operatorname{dom} T_0^{\mathrm{csl}} := \left\{ y \in \operatorname{dom} T_{\max}^{\mathrm{csl}} \mid y \text{has compact support in} [a,b) \right\}.$$

Then it can be shown that $\operatorname{dom} T_0^{\mathrm{csl}}$ is dense in L_W^2 and $(T_0^{\mathrm{csl}})^* = (T_{\min}^{\mathrm{csl}})^* = T_{\max}^{\mathrm{csl}}$ with the superscript $*$ denoting now the adjoint operator, cf. [29, Chap. 3].

The natural generalization of τ_{csl} leads us to operators associated with the linear Hamiltonian differential system

$$-\mathcal{J} z'(t,\lambda) = [H(t) + \lambda W(t)] z(t,\lambda), \quad t \in [a,b), \tag{4}$$

where $H, W : [a,b) \to \mathbb{C}^{2n\times 2n}$ are Hermitian matrix-valued (locally) integrable functions, and $W(t) \geq 0$ for almost all $t \in [a,b)$. At this moment we should mention one big difference between system (4) and its discrete counterpart represented by system (S_λ), for which explanation we refer to [25]. There is no restriction on the invertibility of the matrix $W(t)$, while the third condition in (1) implies that the matrices $\mathcal{V}_k$ and Ψ_k must be singular over $\mathcal{I}_{\mathbb{Z}}$.

If we proceed in the same way as before, we obtain the maximal operator $T_{\max}^{\mathrm{H}}$ with

$$\operatorname{dom} T_{\max}^{\mathrm{H}} := \left\{ z \in L_W^2 \cap \mathrm{AC} \mid \text{it holds } (\tau_{\mathrm{H}} z)(t) = W(t) f(t) \text{ for some } f \in L_W^2 \right\},$$

where $(\tau_{\mathrm{H}} y)(t) := -\mathcal{J} z'(t) - H(t) z(t)$, the symbol AC denotes the set of all $2n$-vector-valued (locally) absolutely continuous functions on $[a,b)$, and L_W^2 means the Hilbert space of (equivalence classes of) $2n$-vector-valued square integrable functions, i.e., it consists of all measurable functions $z : (a,b) \to \mathbb{C}^{2n}$ such that

$$\int_a^b z^*(t)\, W(t)\, z(t)\, \mathrm{d}t < \infty.$$

Then we put

$$T_{\max}^{\mathrm{H}} z := f$$

and for the domain of the pre-minimal operator we consider only $z \in \operatorname{dom} T_{\max}^{\mathrm{H}}$ with compact support in (a,b). However, in contrast to the previous case, we need to employ an additional assumption, otherwise it is possible to have system (4) such that the corresponding maximal operator is not well-defined (multivalued) and the density of the domain of the pre-minimal operator is violated. It typically reads as

$$\left.\begin{array}{c} \textit{whenever } (\tau_{\mathrm{H}} z)(t) = W(t) f(t) \textit{ for some pair } z \in L_W^2 \cap \mathrm{AC} \textit{ and} \\ f \in L_W^2 \textit{ such that } W(t) z(t) \equiv 0, \textit{ then } z(t) \equiv 0 \textit{ on } (a,b). \end{array}\right\} \tag{C}$$

Condition (C) generalizes the classical *Atkinson* (or *definiteness*) *condition*, see [5, Inequality (9.1.6)], and it is satisfied, e.g., when $W(t) > 0$ on (a,b) or if the weight matrix $W(t)$ has a very special structure such as

$$W(t) = \begin{pmatrix} W_1(t) & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad H(t) = \begin{pmatrix} A(t) & B(t) \\ B^*(t) & C(t) \end{pmatrix}$$

with the $n \times n$ blocks being such that $A(t) = A^*(t)$, $C(t) = C^*(t)$, $\det C(t) \neq 0$, and $W_1(t) > 0$ on (a, b). We note that the latter case includes also the situation when the second order equation $(\tau_{\mathrm{csl}}\, y)(t, \lambda) = \lambda\, y(t, \lambda)$ is written in the form of system (4). To the best of the author's knowledge, condition (C) appeared for the first time in [20, Theorem 7.6], where it was used to guarantee the density of yet another set (domain) associated with τ_{H}, see [20, Definition 7.1] and compare with [16, Hypothesis 2.2].

The discrete counterpart of τ_{csl} is provided by the second order Sturm–Liouville difference expression or by the three term recurrence relation, i.e.,

$$(\tau_{\mathrm{dsl}}\, y)_k := \frac{1}{w_k}\Big[- \nabla\big(p_k\, \Delta y_k\big) + q_k\, y_k \Big] = \frac{1}{w_k}\big(- p_{k-1}\, y_{k-1} + r_k\, y_k - p_k\, y_{k+1}\big),$$

where $k \in \mathcal{I}_{\mathbb{Z}} = [0, N+1)_{\mathbb{Z}} := [0, N+1) \cap \mathbb{Z}$ for a given $N \in \mathbb{N} \cup \{0, \infty\}$, the symbols Δ and ∇ mean the forward and backward difference operators, respectively, and $\{p_k\}_{k=-1}^{N}$, $\{q_k\}_{k=0}^{N}$, $\{w_k\}_{k=0}^{N}$, $\{r_k\}_{k=0}^{N}$ are real-valued sequences such that $w_k > 0$ and $r_k = p_{k-1} + p_k + q_k$ for all $k \in \mathcal{I}_{\mathbb{Z}}$. In addition, for $N \neq \infty$ we put $\mathcal{I}_{\mathbb{Z}}^{+} := [0, N+1]_{\mathbb{Z}}$, otherwise $\mathcal{I}_{\mathbb{Z}}^{+} := \mathcal{I}_{\mathbb{Z}}$, and in both cases we let $\mathcal{I}_{\mathbb{Z}}^{\pm} := \mathcal{I}_{\mathbb{Z}}^{+} \cup \{-1\}$. The maximal operator associated with τ_{dsl} is acting on the domain

$$\operatorname{dom} T_{\max}^{\mathrm{dsl}} := \{ y \in \ell_w^2 \mid \tau_{\mathrm{dsl}}\, y \in \ell_w^2 \} \quad \text{with} \quad T_{\max}^{\mathrm{dsl}}\, y := \tau_{\mathrm{dsl}}\, y,$$

where ℓ_w^2 denotes the Hilbert space of equivalence classes of complex-valued sequences $\{y_k\}_{k \in \mathcal{I}_{\mathbb{Z}}^{\pm}}$ such that $\sum_{k \in \mathcal{I}_{\mathbb{Z}}} w_k\, |y_k|^2 < \infty$. However, in contrast to the continuous time case, it was shown in [23, p. 904] that the maximal operator is always multivalued under the classical assumption $p_k \neq 0$ for all $k \in \mathcal{I}_{\mathbb{Z}} \cup \{-1\}$, which guarantees the equivalence between equation $(\tau_{\mathrm{dsl}}\, y)_k = \lambda\, y_k$ and system (S_λ) with $z_k = \left(\begin{smallmatrix} y_k \\ -p_{k-1}\, y_{k-1} \end{smallmatrix} \right)$ and the coefficient matrices

$$\mathcal{S}_k = \begin{pmatrix} 0 & -1/p_k \\ p_k & 1 + (p_{k-1} + q_k)/p_k \end{pmatrix} \quad \text{and} \quad \mathcal{V}_k = \begin{pmatrix} 0 & 0 \\ 0 & -w_k/p_k \end{pmatrix}.$$

This multi-valuedness can be suppressed if we put $p_{-1} = 0$ (and $p_N = 0$ if N is finite), in which case the recurrence relation τ_{dsl} can be expressed as the multiplication by a tridiagonal (Jacobi) matrix, see [9] and also [21, Remark 10]. In the next section we will see that the situation in the setting of system (S_λ) is even worse.

3 Main Result

Before we establish the main result for system (S_λ) with the coefficients specified in (1) or (2), we need to recall some fundamental results from the theory of linear relations, which was established as a suitable tool for the study of multivalued or non-densely defined linear operators in a Hilbert space, cf. [2]. A *(closed) linear relation* $\mathcal{T}$ in a Hilbert space $\mathcal{H}$ over the field of complex numbers $\mathbb{C}$ with the inner product $\langle\cdot,\cdot\rangle$ is a (closed) linear subspace of the product space $\mathcal{H}^2 := \mathcal{H}\times\mathcal{H}$, i.e., the Hilbert space of all ordered pairs $\{z, f\}$ such that $z, f \in \mathcal{H}$. By $\overline{\mathcal{T}}$ we mean the closure of $\mathcal{T}$. The *domain* and the *multivalued part* of $\mathcal{T}$ are, respectively, defined as

$$\operatorname{dom}\mathcal{T} := \{z\in\mathcal{H} \mid \{z,f\}\in\mathcal{T}\} \quad \text{and} \quad \operatorname{mul}\mathcal{T} := \big\{f\in\mathcal{H} \mid \{0,f\}\in\mathcal{T}\big\}.$$

A linear relation $\mathcal{T}$ is the graph of a linear operator in $\mathcal{H}$ if and only if the subspace $\operatorname{mul}\mathcal{T}$ is trivial.

The *adjoint* $\mathcal{T}^*$ of the linear relation $\mathcal{T}$ is the closed linear relation given by

$$\mathcal{T}^* := \big\{\{y,g\}\in\mathcal{H}^2 \mid \langle z,g\rangle = \langle f,y\rangle \text{ for all } \{z,f\}\in\mathcal{T}\big\}.$$

The definition of $\mathcal{T}^*$ reduces to the standard definition for the graph of the adjoint operator when $\mathcal{T}$ is a densely defined operator. The following proposition is crucial for our present treatment, see [2, Proposition 3.32] and also [13, Theorem 1].

Proposition 1 *Let T be a linear relation in $\mathcal{H}^2$. Then* $\operatorname{dom} T$ *is dense in $\mathcal{H}$ if and only if the adjoint T^* is single-valued, i.e.,* $\operatorname{mul} T^* = \{0\}$.

Let us denote by ℓ^2_Ψ the linear space of all complex $2n$-vector-valued sequences defined on $\mathcal{I}_\mathbb{z}^+$, which are square summable with respect to the weight Ψ_k, i.e.,

$$\ell^2_\Psi := \big\{\{z\}_{k\in\mathcal{I}_\mathbb{z}^+} \mid\ z_k\in\mathbb{C}^{2n} \text{ and } \|z\|_\Psi < \infty\big\},$$

where $\|z\|_\Psi := \sqrt{\langle z,\ z\rangle_\Psi}$ is the natural semi-norm determined by the semi-inner product with the weight Ψ_k, i.e., $\langle z,\ v\rangle_\Psi := \sum_{k\in\mathcal{I}_\mathbb{z}} z_k^*\Psi_k v_k$. As the consequence of the singularity of Ψ_k, it follows that ℓ^2_Ψ is not a Hilbert space. However, the quotient space $\tilde{\ell}^2_\Psi$ obtained after factoring out the kernel of the semi-norm, i.e.,

$$\tilde{\ell}^2_\Psi := \ell^2_\Psi\big/\{z\in\ell^2_\Psi \mid \|z\|_\Psi = 0\}, \tag{5}$$

is a Hilbert space. Henceforth, the equivalence class corresponding to a sequence $z\in\ell^2_\Psi$ will be written by using the brackets $[\cdot]$, i.e., $z\in[z]\in\tilde{\ell}^2_\Psi$. One can easily observe that we have $z^{[1]}, z^{[2]}\in[z]$ if and only if $\Psi_k z_k^{[1]} = \Psi_k z_k^{[2]}$ for all $k\in\mathcal{I}_\mathbb{z}$. We also need to define the subspace

$$\ell^2_{\Psi,0} := \begin{cases} \big\{z\in\ell^2_\Psi \mid z_0 = 0 = z_{N+1}\big\} & \text{if } N\in\mathbb{N}\cup\{0\},\\ \big\{z\in\ell^2_\Psi \mid z\text{has a compact support in}\,\mathcal{I}_\mathbb{z} \text{ and } z_0 = 0\big\} & \text{if } N=\infty, \end{cases}$$

With this notation, we introduce the *maximal linear relation* $T_{\max}$ as a subspace of the product space $\tilde{\ell}^{2\times 2}_{\Psi} := \tilde{\ell}^2_{\Psi} \times \tilde{\ell}^2_{\Psi}$ given by

$$T_{\max} := \big\{\{[z],[f]\} \in \tilde{\ell}^{2\times 2}_{\Psi} \mid \text{there exists } \varkappa \in [z] \text{ such that } \mathscr{L}(\varkappa)_k = \Psi_k f_k \text{ for all } k \in \mathcal{I}_{\mathbb{z}}\big\}$$

and we emphasize that it does not depend on the choice of the representative $f \in [f]$. Similarly, we define the *pre-minimal linear relation*

$$T_0 := \big\{\{[z],[f]\} \in \tilde{\ell}^{2\times 2}_{\Psi} \mid \text{there exists } \varkappa \in [z] \cap \ell^2_{\Psi,0} \text{ such that } \mathscr{L}(\varkappa)_k = \Psi_k f_k \text{ for all } k \in \mathcal{I}_{\mathbb{z}}\big\},$$

which evidently satisfies $T_0 \subseteq T_{\max}$. The closure of T_0 is said to be the *minimal linear relation*, i.e.,

$$T_{\min} := \overline{T_0},$$

and the following equalities for these three linear relations were established in [12, Theorem 5.10].

Theorem 1 *The linear relations $T_{\max}$, T_0, and $T_{\min}$ as defined above satisfy*

$$T_0^* = T_{\min}^* = T_{\max}.$$

The latter statement together with Proposition 1 implies that dom T_0 and dom $T_{\min}$ are dense subsets of $\tilde{\ell}^2_{\Psi}$ if and only if mul $T_{\max} = \{[0]\}$, i.e., $T_{\max}$ is a (graph of a) linear operator. Equivalently, it means that there is no $z \in [0]$ such that equality $\mathscr{L}(z)_k = \Psi_k f_k$ is satisfied for all $k \in \mathcal{I}_{\mathbb{z}}$ and some $f \notin [0]$. In particular, this is true when $z \equiv 0$ on $\mathcal{I}^+_{\mathbb{z}}$ is the only representative of the class $[0]$ such that the equality $\mathscr{L}(z)_k = \Psi_k f_k$ is satisfied for all $k \in \mathcal{I}_{\mathbb{z}}$ and some $f \in \ell^2_{\Psi}$. As we mentioned in the introductory section, this condition was proposed in [12, Theorem 5.4], compare also with condition (C) for system (4). But is it even possible? The following theorem shows that the answer is *negative for every nontrivial choice* of the weight matrices Ψ_k.

Theorem 2 *The condition* mul $T_{\max} = \{[0]\}$ *holds if and only if $\Psi_k \equiv 0$ on the discrete interval $\mathcal{I}_{\mathbb{z}}$.*

Proof If $\Psi_k \equiv 0$ on $\mathcal{I}_{\mathbb{z}}$, then $\tilde{\ell}^2_{\Psi} = \{[0]\}$ and the statement is trivial. On the other hand, let $\Psi_k \not\equiv 0$ on $\mathcal{I}_{\mathbb{z}}$ and denote by $m \in \mathcal{I}_{\mathbb{z}}$ the first index such that $\Psi_m \neq 0$, i.e., $\Psi_k = 0$ for all $k \in \mathcal{I}_{\mathbb{z}} \cap (-\infty, m)_{\mathbb{z}}$. Moreover, let $\xi \in \mathbb{C}^{2n} \setminus \operatorname{Ker} \Psi_m$ be arbitrary, i.e., $\Psi_m \xi \neq 0$. If $m = 0$, then the pair

$$z_k = \begin{cases} -\mathcal{J}\Psi_0\xi, & k = 0, \\ 0, & k \in \mathcal{I}^+_{\mathbb{z}}/\{0\}, \end{cases} \qquad f_k = \begin{cases} \xi, & k = 0, \\ 0, & k \in \mathcal{I}^+_{\mathbb{z}}/\{0\}, \end{cases}$$

satisfies $\mathscr{L}(z)_k = \Psi_k f_k$ for all $k \in \mathcal{I}_{\mathbb{z}}$ and simultaneously

$$\|z\|^2_{\Psi} = -\xi^*\Psi_0\mathcal{J}\Psi_0\mathcal{J}\Psi_0\xi \overset{(2)}{=} 0, \qquad \|f\|^2_{\Psi} = \xi^*\Psi_0\xi \neq 0.$$

Therefore, we have $\{[z],[f]\} \in T_{\max}$ for the corresponding equivalence classes with $[z] = [0]$ and $[f] \neq [0]$, which shows that $\operatorname{mul} T_{\max} \neq \{[0]\}$. Similarly, one can verify directly that for $m > 0$ the pair

$$z_k = \begin{cases} -(\Pi_{j=k}^{m-1} \mathcal{S}_j)\, \mathcal{J} \Psi_m \xi, & k \in [0, m-1]_{\mathbb{Z}} \\ -\mathcal{J} \Psi_m \xi, & k = m, \\ 0, & k \in \mathcal{I}_{\mathbb{Z}}^+ \cap [m+1, \infty)_{\mathbb{Z}}, \end{cases} \qquad f_k = \begin{cases} \xi, & k = m, \\ 0, & k \in \mathcal{I}_{\mathbb{Z}}^+ / \{m\}, \end{cases}$$

satisfies $\mathscr{L}(z)_k = \Psi_k f_k$ for all $k \in \mathcal{I}_{\mathbb{Z}}$ and

$$\|z\|_{\Psi}^2 = \sum_{\substack{k \in \mathcal{I}_{\mathbb{Z}} \\ k \geq m}} z_k^* \Psi_k z_k = -\xi^* \Psi_m \mathcal{J} \Psi_m \mathcal{J} \Psi_m \xi \overset{(2)}{=} 0, \qquad \|f\|_{\Psi}^2 = \xi^* \Psi_m \xi \neq 0.$$

Hence again $\{[z],[f]\} \in T_{\max}$ for the corresponding equivalence classes with $[z] = [0]$ and $[f] \neq [0]$, i.e., $\operatorname{mul} T_{\max} \neq \{[0]\}$.

As the direct consequence of Theorem 2 we get the main result concerning the fundamental characterization of $T_{\max}$ and the domains of T_0 and $T_{\min}$. It shows that the development of the spectral theory for discrete symplectic systems in [12, 31] by using the linear relations instead of operators is not only fruitful because of its generality but it is, in fact, *necessary*. In contrast to the continuous time case and condition (C), it is not possible to "fix" it by any additional condition.

Corollary 1 *The maximal linear relation $T_{\max}$ is always multivalued and the sets* $\operatorname{dom} T_0$ *and* $\operatorname{dom} T_{\min}$ *are never dense in $\tilde{\ell}_{\Psi}^2$ but for $\Psi_k \equiv 0$ on $\mathcal{I}_{\mathbb{Z}}$.*

From the proof of Theorem 2 one can observe that the permanent multi-valuedness of $T_{\max}$ is caused by the singularity of the weight matrices Ψ_k. For nonsingular weight matrices in the setting of discrete symplectic systems it is necessary to have at least a quadratic dependence on the spectral parameter, which was studied in [24]. Alternatively and less generally, we could consider the linear Hamiltonian difference systems instead of (S_λ). But, although such systems allow nonsingular weight matrices, they lead to the same conclusion because of the presence of a partial shift, cf. [22].

Finally, Proposition 1 applied to the linear relation $T = T_{\max}$ yields the dependence of the density of $\operatorname{dom} T_{\max}$ in $\tilde{\ell}_{\Psi}^2$ on the single-valuedness of $T_{\min}$, because it holds $T_{\max}^* = T_{\min}^{**} = T_{\min}$. Is there a nontrivial case where this is not possible? For example, let us take $N \geq 1$, $\mathcal{S}_k \equiv I$ on $\mathcal{I}_{\mathbb{Z}}$, $\Psi_0 = \Psi_1 \neq 0$, and $\Psi_k = 0$ for all $k \in [2, N+1)_{\mathbb{Z}}$. Then $\Psi_0 \xi \neq 0$ for some $\xi \in \mathbb{C}^{2n}$ and the pair

$$z_k = \begin{cases} \mathcal{J} \Psi_0 \xi, & k = 1, \\ 0, & k \in \mathcal{I}_{\mathbb{Z}}^+ / \{1\}, \end{cases} \qquad f_k = \begin{cases} \xi, & k = 0, \\ -\xi, & k = 1, \\ 0, & k \in \mathcal{I}_{\mathbb{Z}} / \{0, 1\} \end{cases}$$

satisfies $\mathscr{L}(z)_k = \Psi_k f_k$ on $\mathcal{I}_{\mathbb{Z}}$. Simultaneously, we have $z \in \ell_{\Psi,0}^2$, $z \in [0]$, and $f \notin [0]$, which yields that $\{[z],[f]\} \in T_0 \subseteq T_{\min}$. Hence $\operatorname{mul} T_{\min} \neq \{[0]\}$ and so the set $\operatorname{dom} T_{\max}$ is not dense in $\tilde{\ell}_{\Psi}^2$.

On the other hand, if system (S_λ) is *definite* on the discrete interval $\mathcal{I}_\mathbb{z}$, i.e., there exists $\lambda \in \mathbb{C}$ and a finite discrete subinterval $[a, b]_\mathbb{z} \subseteq \mathcal{I}_\mathbb{z}$ such that every nontrivial solution of (S_λ) satisfies $\sum_{k=a}^{b} z_k^*(\lambda)\Psi_k z_k(\lambda) > 0$, then for every $\{[z], [f]\} \in T_{\max}$ there exists a unique $\hat{z} \in [z]$ satisfying $\mathscr{L}(\hat{z})_k = \Psi_k f_k$ on $\mathcal{I}_\mathbb{z}$ and the minimal linear relation admits the representation

$$T_{\min} = \begin{cases} \big\{\{[z], [f]\} \in T_{\max} \mid \hat{z}_0 = 0 = \lim_{k\to\infty} \hat{z}_k^* \mathcal{J} \hat{w}_k \\ \qquad\qquad\qquad\qquad \text{for all } [w] \in \operatorname{dom} T_{\max}\big\}, \\ \big\{\{[z], [f]\} \in T_{\max} \mid \hat{z}_0 = 0 = \hat{z}_{N+1}\big\}, \end{cases} \tag{6}$$

see [12, Theorem 5.2] and [31, Theorem 3.2]. In that case the equality $\operatorname{mul} T_{\min} = \{[0]\}$ is true, e.g., for the choice

$$\mathcal{S}_k = \begin{pmatrix} \mathcal{A}_k & \mathcal{B}_k \\ \mathcal{C}_k & \mathcal{D}_k \end{pmatrix} \quad \text{and} \quad \Psi_k = \begin{pmatrix} \mathcal{W}_k & 0 \\ 0 & 0 \end{pmatrix}$$

with the $n \times n$ blocks such that $\mathcal{W}_k > 0$ and $\det \mathcal{B}_k \neq 0$ on $\mathcal{I}_\mathbb{z}$, compare with condition (C). Indeed, if $\{[z], [f]\} \in T_{\min}$ with $\|\hat{z}\|_\Psi = 0$ and $\hat{z} = (\hat{x}^* \ \ \hat{u}^*)^*$, then the given form of Ψ_k and the expression of $T_{\min}$ given above imply that $\hat{x}_k \equiv 0$ on $\mathcal{I}_\mathbb{z}^+$. Consequently, the block structure of $\mathcal{S}_k$, the invertibility of $\mathcal{B}_k$, and (6) yield also $\hat{u} \equiv 0$ on $\mathcal{I}_\mathbb{z}^+$, i.e., $\hat{z} \equiv 0$ on $\mathcal{I}_\mathbb{z}^+$. Therefore $[f] = [0]$, which shows the single-valuedness of $T_{\min}$ and simultaneously the density of $\operatorname{dom} T_{\max}$ in $\widetilde{\ell}_\Psi^2$.

4 Time Scale Explanation

In this final section we attempt to provide an explanation of the phenomenon concerning the density of $T_{\min}$ by using the time scale calculus, which was developed for the simultaneous study of differential and difference equations and many cases "in between". Henceforth we suppose that the reader is familiar with the foundations of the time scale calculus as it can be found in the original works [17–19] by Hilger or in the monograph [8] by Bohner and Peterson. In particular, by a *time scale* $\mathbb{T}$ we mean an arbitrary nonempty closed subset of real numbers. Every time scale is equipped with the *forward jump operator* and *graininess function* defined respectively as

$$\sigma(t) := \inf\{s \in \mathbb{T},\ s > t\} \quad \text{and} \quad \mu(t) := \sigma(t) - t.$$

If the time scale $\mathbb{T}$ has a right-scattered maximum M, then $\mathbb{T}^\kappa := \mathbb{T}/\{M\}$. Otherwise we let $\mathbb{T}^\kappa := \mathbb{T}$. The *time scale Δ-derivative* $f^\Delta(t)$ is defined for all $t \in \mathbb{T}^\kappa$ in such a way that, in the case $\mathbb{T} = \mathbb{R}$, we get the classical derivative, while for $\mathbb{T} = \mathbb{Z}$ it reduces to the forward difference operator. In addition, we recall two useful formulas

$$f^\sigma(t) = f(t) + \mu(t) f^\Delta(t) \quad \text{and} \quad \int_t^{\sigma(t)} f(\tau)\,\Delta\tau = \mu(t) f(t) \tag{7}$$

for any $t \in \mathbb{T}^\kappa$, where we apply the abbreviation $f^\sigma(t) := f(\sigma(t))$.

Systems (4) and (S_λ) are the simplest examples of the time-reversed symplectic dynamic system

$$z^\Delta(t,\lambda) = [\mathcal{S}(t) + \lambda \mathcal{V}(t)] z^\sigma(t,\lambda), \quad t \in \mathbb{T}^\kappa, \tag{$S_\lambda^{\mathbb{T}}$}$$

where the coefficients $\mathcal{S}, \mathcal{V} : \mathbb{T}^\kappa \to \mathbb{C}^{2n\times 2n}$ are piecewise rd-continuous functions (i.e., from C_{prd}) on $\mathbb{T}^\kappa$ and such that the function $\mathbb{S}(t,\lambda) := \mathcal{S}(t) + \lambda\mathcal{V}(t)$ satisfies the symplectic-type identity

$$\mathbb{S}^*(t,\bar\lambda)\mathcal{J} + \mathcal{J}\mathbb{S}(t,\lambda) - \mu(t)\mathbb{S}^*(t,\bar\lambda)\mathcal{J}\mathbb{S}(t,\lambda) = 0. \tag{8}$$

for all $(t,\lambda) \in \mathbb{T}^\kappa \times \mathbb{C}$, see [30, Section 4] for more details. More specifically, if $\mathbb{T} = [a,b)$ then $\sigma(t) \equiv t$ and system (8) corresponds to (4) with $H(t) := -\mathcal{J}\mathcal{S}(t)$ and $W(t) := -\mathcal{J}\mathcal{V}(t)$. In the case $\mathbb{T} = \mathcal{I}_\mathbb{Z}$ we have $\sigma(t) \equiv t+1$, and so system (8) corresponds to (S_λ) with the coefficients $\mathcal{S}_k := I - \mathcal{S}(k)$ and $\mathcal{V}_k := -\mathcal{V}(k)$. The existence and uniqueness of a solution being piecewise rd-continuously differentiable on $\mathbb{T}$ (i.e., from C^1_{prd}) of any initial value problem associated with ($S_\lambda^{\mathbb{T}}$) is guaranteed by the relation

$$[I - \mu(t)\mathbb{S}(t,\lambda)]^{-1} = -\mathcal{J}[I - \mu(t)\mathbb{S}^*(t,\bar\lambda)]\mathcal{J}, \tag{9}$$

which follows directly from (8). Furthermore, by (7)–(9) we obtain that system ($S_\lambda^{\mathbb{T}}$) can be equivalently written as

$$z^\Delta(t,\lambda) = \mathcal{S}(t)z^\sigma(t,\lambda) + \lambda\mathcal{J}\Psi(t)z(t), \text{ where } \Psi(t) := \mathcal{J}\mathcal{V}(t)\mathcal{J}[I - \mu(t)\mathcal{S}^*(t)]\mathcal{J}$$

is such that

$$\Psi^*(t) = \Psi(t) \quad \text{and} \quad \mu(t)\Psi(t)\mathcal{J}\Psi(t) = 0. \tag{10}$$

With these assumptions, we could define the space of square integrable functions with respect to the weight $\Psi(t) \geq 0$, the corresponding Hilbert space, the maximal and minimal linear relations, and establish a connection between them. However, it is not needed for the present investigation and it will be done in our subsequent work. At this moment we focus only on the existence of a function $z \in C^1_{prd}(\mathbb{T})$ such that $\|z\|_\Psi := \int_\mathbb{T} z^*(t)\Psi(t)z(t)\,\Delta t = 0$ and

$$-\mathcal{J}[z^\Delta - \mathcal{S}(t)z^\sigma(t)] = \Psi(t)f(t), \quad t \in \mathbb{T}^\kappa, \tag{11}$$

for some $f \in C_{prd}(\mathbb{T}^\kappa)$ with $\|f\|_\Psi := \int_{\mathbb{T}^\kappa} f^*(t)\Psi(t)f(t)\,\Delta t \neq 0$. In this context we remark that every solution of (11) can be expressed as

$$z(t) = \Phi(t)\Big[\eta + \mathcal{J}\int_a^t \Phi^*(\tau)\Psi(\tau)f(\tau)\,\Delta\tau\Big],$$

where $\eta = \Phi^{-1}(a)z(a)$ and $\Phi(\cdot)$ denotes an arbitrary fundamental matrix of the homogeneous system $(S_0^{\mathbb{T}})$ satisfying the condition $\Phi(t)\mathcal{J}\Phi^*(t) = \mathcal{J}$ for some (and hence for any) $t \in \mathbb{T}$. Before we analyze the problem in question, we summarize the basic assumptions concerning system $(S_\lambda^{\mathbb{T}})$. Note that we require $\mathbb{T} \neq [a, b)$ as it was already discussed in detail.

Hypothesis 1 A time scale $\mathbb{T}$ is given with $a := \inf \mathbb{T} > -\infty$ and at least one point $t \in \mathbb{T}^\kappa$ is right-scattered. We have $\mathbb{S}(t, \lambda) := \mathcal{S}(t) + \lambda\mathcal{V}(t)$ on $\mathbb{T}^\kappa \times \mathbb{C}$, where the $2n \times 2n$ matrix-valued functions $\mathcal{S}, \mathcal{V} : \mathbb{T}^\kappa \to \mathbb{C}^{2n\times 2n}$ are piecewise rd-continuous on $\mathbb{T}^\kappa$ such that condition (8) holds and $\Psi(t) := \mathcal{J}\mathcal{V}(t)\mathcal{J}[I - \mu(t)\mathcal{S}^*(t)]\mathcal{J} \geq 0$ for all $t \in \mathbb{T}^\kappa$.

If a is a right-scattered point and $\Psi(a) \neq 0$, then one can readily verify that the pair

$$z(t) := \begin{cases} -\mu(a)\mathcal{J}\Psi(a)\xi, & t = a, \\ 0, & t \in \mathbb{T}/\{a\}, \end{cases} \quad \text{and} \quad f(t) := \begin{cases} \xi, & t = a, \\ 0, & t \in \mathbb{T}^\kappa/\{a\}, \end{cases}$$

is such that (11) holds and by (7), we see that

$$\|z\|_\Psi = \int_a^{\sigma(a)} z^*(t)\Psi(t)z(t)\,\Delta t = -\mu^3(a)\xi^*\Psi(a)\mathcal{J}\Psi(a)\mathcal{J}\Psi(a)\xi \overset{(10)}{=} 0,$$

$$\|f\|_\Psi = \int_a^{\sigma(a)} f^*(t)\Psi(t)f(t)\,\Delta t = \mu(a)\xi^*\Psi(a)\xi \neq 0$$

for any $\xi \in \mathbb{C}^{2n} \setminus \operatorname{Ker}\Psi(a)$. This simple observation justifies the following statement.

Theorem 3 *Let Hypothesis 1 be satisfied and, in addition, the left-end point a be right-scattered with $\Psi(a) \neq 0$. Then there exist $z \in \mathrm{C}^1_{\mathrm{prd}}(\mathbb{T})$ and $f \in \mathrm{C}_{\mathrm{prd}}(\mathbb{T}^\kappa)$ such that equation (11) is satisfied, $\|z\|_\Psi = 0$, and $\|f\|_\Psi \neq 0$.*

On the pother hand, let a be right-dense and $t_0 \in \mathbb{T}^\kappa$ be an arbitrary right-scattered point such that $\Psi(t_0) \neq 0$. If we put

$$z(t) := \Phi(t)\Big[-\mu(t_0)\mathcal{J}\Psi(t_0)\xi + \mathcal{J}\int_a^t \Phi^*(\tau)\Psi(\tau)f(\tau)\,\Delta\tau\Big]$$

and

$$f(t) := \begin{cases} \xi, & t = t_0, \\ 0, & t \in \mathbb{T}^\kappa/\{t_0\}, \end{cases}$$

where the fundamental matrix $\Phi(\cdot)$ is determined by the initial condition $\Phi(t_0) = I$ and the vector $\xi \in \mathbb{C}^{2n} \setminus \in \operatorname{Ker}\Psi(t_0)$ is arbitrary, then we have

$$\|f\|_\Psi = \int_{t_0}^{\sigma(t_0)} f^*(t)\Psi(t)f(t)\,\Delta t \overset{(7)}{=} \mu(t_0)\xi^*\Psi(t_0)\xi \neq 0$$

and simultaneously

$$z(t) = \begin{cases} -\mu(t_0)\,\Phi(t)\mathcal{J}\Psi(t_0)\xi, & t \in [a,t_0]\cap\mathbb{T}, \\ 0, & t \in [\sigma(t_0),\infty)\cap\mathbb{T}, \end{cases}$$

which yields that

$$\begin{aligned}\|z\|_\Psi &= \left(\int_a^{t_0} + \int_{t_0}^{\sigma(t_0)}\right) z^*(t)\Psi(t)z(t)\,\Delta t \\ &= -\mu^2(t_0)\xi^*\Psi(t_0)\mathcal{J}\left(\int_a^{t_0}\Phi^*(t)\Psi(t)\Phi(t)\,\Delta t\right)\mathcal{J}\Psi(t_0)\xi,\end{aligned}$$

because $\int_{t_0}^{\sigma(t_0)} z^*(t)\Psi(t)z(t)\,\Delta t = -\mu^3(t_0)\xi^*\Psi(t_0)\mathcal{J}\Psi(t_0)\mathcal{J}\Psi(t_0)\xi = 0$ by (10). Consequently we get the following statement, which shows the sporadic nature of densely defined operators associated with system $(\mathrm{S}_\lambda^{\mathbb{T}})$ as in the case $\mathbb{T} = [a,b)$, i.e., for the linear Hamiltonian differential system (4). Actually, this fact is (again) closely connected with the necessary singularity of the weight matrix $\Psi(\cdot)$ at every right-scattered point, which follows from (8), see the second condition in (10).

Theorem 4 *Let Hypothesis 1 be satisfied and, in addition, the left-end point a be right-dense. If there exist a right-scattered point $t_0 \in \mathbb{T}^\kappa$ with $\Psi(t_0) \neq 0$ and a vector $\xi \in \mathbb{C}^{2n} \setminus \operatorname{Ker}\Psi(t_0)$ such that*

$$\xi^*\Psi(t_0)\mathcal{J}\left(\int_a^{t_0}\Phi^*(t)\Psi(t)\Phi(t)\,\Delta t\right)\mathcal{J}\Psi(t_0)\xi = 0,$$

then system (11) possesses a solution $z \in \mathrm{C}^1_{\mathrm{prd}}(\mathbb{T})$ with $\|z\|_\Psi = 0$ for some function $f \in \mathrm{C}_{\mathrm{prd}}(\mathbb{T}^\kappa)$ with $\|f\|_\Psi \neq 0$.

Acknowledgements The research was supported by the Czech Science Foundation under Grant GA16-00611S. The author is grateful to the anonymous referee for a detailed reading of the manuscript and her/his comments.

References

1. Ahlbrandt, C.D., Peterson, A.C.: Discrete Hamiltonian Systems: Difference Equations, Continued Fractions, and Riccati Equations, Kluwer Texts in the Mathematical Sciences, vol. 16. Kluwer Academic Publishers Group, Dordrecht (1996). ISBN 0-7923-4277-1
2. Arens, R.: Operational calculus of linear relations. Pacific J. Math. **11**, 9–23 (1961)
3. Asahi, T.: Spectral theory of the difference equations. Progr. Theoret. Phys. **36**, 55–96 (1966)

4. Asahi, T., Kashiwamura, S.: Spectral theory of the difference equations in isotopically disordered harmonic chains. Progr. Theoret. Phys. **48**, 361–371 (1972)
5. Atkinson, F.V.: Discrete and Continuous Boundary Problems, Mathematics in Science and Engineering, vol. 8. Academic Press, New York (1964)
6. Bender, C.M., Mead, L.R., Milton, K.A.: Discrete time quantum mechanics. Comput. Math. Appl. **28**(10–12), 279–317 (1994)
7. Bohner, M., Došlý, O.: Disconjugacy and transformations for symplectic systems. Rocky Mountain J. Math. **27**(3), 707–743 (1997)
8. Bohner, M., Peterson, A.C.: Dynamic Equations on Time Scales: An Introduction with Applications. Birkhäuser Verlag, Boston (2001). ISBN 0-8176-4225-0
9. Brown, B.M., Christiansen, J.S.: On the Krein and Friedrichs extensions of a positive Jacobi operator. Expo. Math. **23**(2), 179–186 (2005)
10. Bruce, S.: Discrete time in quantum mechanics. Phys. Rev. A (3) **64**(1), 014103, 4 pp. (electronic) (2001)
11. Clark, S.L., Gesztesy, F., Zinchenko, M.: Weyl-Titchmarsh theory and Borg-Marchenko-type uniqueness results for CMV operators with matrix-valued Verblunsky coefficients. Oper. Matrices **1**(4), 535–592 (2007)
12. Clark, S.L., Zemánek, P.: On discrete symplectic systems: associated maximal and minimal linear relations and nonhomogeneous problems. J. Math. Anal. Appl. **421**(1), 779–805 (2015)
13. Coddington, E.A.: Extension Theory of Formally Normal and Symmetric Subspaces, Memoirs of the American Mathematical Society, vol. 134. American Mathematical Society, Providence (1973)
14. Dobrev, V.K., Doebner, H.-D., Twarock, R.: Quantum mechanics with difference operators. Rep. Math. Phys. **50**(3), 409–431 (2002)
15. Gesztesy, F., Zinchenko, M.: Weyl-Titchmarsh theory for CMV operators associated with orthogonal polynomials on the unit circle. J. Approx. Theory **139**(1–2), 172–213 (2006)
16. Gesztesy, F., Zinchenko, M.: Renormalized oscillation theory for Hamiltonian systems. Adv. Math. **311**, 569–597 (2017)
17. Hilger, S.: Ein Maßkettenkalkül mit Anwendung auf Zentrumsmannigfaltigkeiten (in German, [Calculus on Measure Chains with Application to Central Manifolds]), Ph.D. dissertation, University of Würzburg, Würzburg (1988)
18. Hilger, S.: Analysis on measure chains—a unified approach to continuous and discrete calculus. Results Math. **18**(1–2), 18–56 (1990)
19. Hilger, S.: Differential and difference calculus–unified!, in "Proceedings of the Second World Congress of Nonlinear Analysts, Part 5 (Athens, 1996)". Nonlinear Anal. **30**(5), 2683–2694 (1997)
20. Krall, A.M.: $M(\lambda)$ theory for singular Hamiltonian systems with one singular point. SIAM J. Math. Anal. **20**(3), 664–700 (1989)
21. Malamud, M.M.: On a formula for the generalized resolvents of a non-densely defined Hermitian operator. Ukrainian Math. J. **44** (1993), no. 12, 1522–1547. Translated from: Ukraïn. Mat. Zh. 44 (1992), no. 12, 1658–1688 (Russian)
22. Ren, G.: On the density of the minimal subspaces generated by discrete linear Hamiltonian systems. Appl. Math. Lett. **27**, 1–5 (2014)
23. Shi, Y., Sun, H.: Self-adjoint extensions for second-order symmetric linear difference equations. Linear Algebra Appl. **434**(4), 903–930 (2011)
24. Šimon Hilscher, R., Zemánek, P.: Generalized Lagrange identity for discrete symplectic systems and applications in Weyl–Titchmarsh theory. In: "Theory and Applications of Difference Equations and Discrete Dynamical Systems", Proceedings of the 19th International Conference on Difference Equations and Applications (Muscat, 2013), Z. AlSharawi, J. Cushing, and S. Elaydi (editors), Springer Proceedings in Mathematics & Statistics, Vol. 102, pp. 187–202, Springer, Berlin (2014)
25. Šimon Hilscher, R., Zemánek, P.: Limit point and limit circle classification for symplectic systems on time scales. Appl. Math. Comput. **233**, 623–646 (2014)

26. Teschl, G.: Jacobi Operators and Completely Integrable Nonlinear Lattices, Mathematical Surveys and Monographs, vol. 72. American Mathematical Society, Providence (2000). ISBN 0-8218-1940-2
27. Teschl, G.: Mathematical Methods in Quantum Mechanics: With applications to Schrödinger operators, Graduate Studies in Mathematics, vol. 99. American Mathematical Society, Providence (2009). ISBN 978-0-8218-4660-5
28. von Neumann, J.: Mathematical Foundations of Quantum Mechanics, translated by Robert T. Princeton University Press, Princeton, Beyer (1955)
29. Weidmann, J.: Spectral Theory of Ordinary Differential Operators. Lecture Notes in Mathematics, vol. 1258. Springer, Berlin (1987)
30. Zemánek, P.: Non-limit-circle and limit-point criteria for symplectic dynamic systems on time scales, submitted
31. Zemánek, P., Clark, S.L.: Characterization of self-adjoint extensions for discrete symplectic systems. J. Math. Anal. Appl. **440**(1), 323–350 (2016)