Applied Numerical Mathematics 156 (2020) 174–191

Contents lists available at ScienceDirect

Applied Numerical Mathematics

www.elsevier.com/locate/apnum

Estimation of the interpolation error for semiregular prismatic elements

Ali Khademi *, Jon Eivind Vatne

Department of Computer science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, P.O. Box 7030, Bergen, Norway

ARTICLE INFO

Article history:
Received 10 December 2019
Received in revised form 20 February 2020
Accepted 26 April 2020
Available online 4 May 2020

Keywords:
Interpolation error
Semiregular prismatic element
Maximum angle condition

ABSTRACT

In this paper we introduce the semiregularity property for a family of decompositions of a polyhedron into a natural class of prisms. In such a family, prismatic elements are allowed to be very flat or very long compared to their triangular bases, and the edges of quadrilateral faces can be nonparallel. Moreover, the triangular faces of each element are either parallel or skew to each other. To estimate the error of the interpolation operator defined on the finite space whose basis functions are defined on the general prismatic elements, we consider quadratic polynomials as the basis functions for that space which are bilinear on the reference prism. We then prove that under this modification of the semiregularity criterion, the interpolation error is of order $O(h)$ in the H^1-norm.

1. Introduction

The finite element method is one of the most flexible and powerful methods to solve numerically a wide variety of partial differential equations [3,13,11,14]. A fundamental problem is to estimate the error between the exact solution and its computable finite element approximation. This error can be bounded by the best approximation of the exact solution in the finite element space consisting of piecewise polynomial functions (see Céa's lemma [4]). Hence, it is important to estimate the interpolation errors.

In the process of estimation of the interpolation error, some constant times a power of the discretization parameter h appears. It is crucial that this constant does not blow up when h tends to zero. For linear elliptic boundary value problems in 2-dimensional space, Zlámal [15] introduced the minimum angle condition that guarantees a bound on the constant in the final error which comes from the estimation error of the defined interpolation operator. See also Synge [12]. Babuška and Aziz [2] proposed that the minimum angle condition can be relaxed to the maximum angle condition. In 3-dimensional space, the natural extension of the maximum angle condition for tetrahedral elements was proposed by Křížek [9]. Recently, the generalization of the maximum angle condition in d-dimensional spaces ($d \geq 2$), by means of $\sin_d$ [5], for d-simplices is introduced and extended in [7,8] and also the equivalence of the maximum angle condition and its generalized version is proved.

The maximum angle condition enables us to keep an optimal error whereas we are allowed to consider degenerating families of elements in order to cover the narrow or flat parts of a given bounded domain. For instance, in geophysical simulations [10], where the domain consists of horizontal triangles as a base and regular vertical layer, all finite prismatic

* Corresponding author.
E-mail addresses: ali.khademi@hvl.no, akhademi.math@gmail.com (A. Khademi), jon.eivind.vatne@hvl.no (J.E. Vatne).

https://doi.org/10.1016/j.apnum.2020.04.018

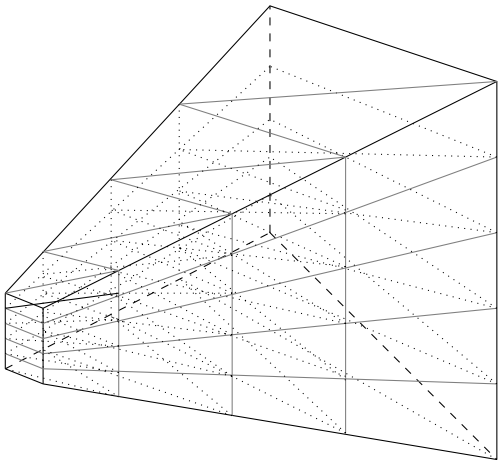


Fig. 1. Partition of frustum into prisms satisfying Definition 3.

elements are produced by the Cartesian product of triangles and the closed intervals called triangular prisms. For such simulation, high aspect-ratio for the elements must be allowed. Therefore, we [6] analyzed the behavior of the interpolation error under the maximum angle condition on the above prisms.

The aim of this paper is to estimate the interpolation error for a more general class of prismatic elements than previously considered in [6]. This class of elements naturally appear e.g. in some standard geometric models. In Fig. 1 an example of a frustum is given. We interpolate a given function by quadratic polynomials which are bilinear on the reference prism. To introduce general prismatic element, similar to [6] we consider the maximum angle condition for all dihedral angles. In addition, we assume that the ratios between the three edges that connect the triangular faces is bounded from below by some positive constant. Note that these ratios for triangular prisms are one. We relax the conditions from [6] to allow e.g. slanted or skew elements. In particular small deformations of the geometry from [6] are covered. We refer to [6] for further motivation and context.

We use the technique of reference element in several parts of the main proof in order to demonstrate that the interpolation error is of order $O(h)$ in the H^1-norm for sufficiently smooth functions and sufficiently small h. In that proof we use a positive lower bound for the Jacobian determinant. In our case, this determinant is a quadratic polynomial in three variables whose coefficients are expressed in terms of volumes of tetrahedra formed by the vertices of the prism.

The outline of the paper is as follows. First, in Section 2, we introduce notations and give some definitions. In addition, we propose an extension of the semiregularity property that allows us to consider some degenerate families of prismatic elements. In Section 3, we obtain a positive lower bound for the Jacobian determinant in Theorem 6, since we use the technique of the reference element to prove the main result. In Section 4 we prove Theorem 7 which states that the interpolation error is of the order $O(h)$ under the extended semiregularity condition, followed by some conclusions in Section 5.

2. Main definitions and geometric preliminaries

We will consider meshes whose elements are defined in this definition:

Definition 1. A straight-side, triangular based prism is a convex polyhedron with six vertices, two triangular faces and three quadrilateral (convex planar) faces. Furthermore, each quadrilateral is incident to the other four faces. The two triangles are not incident. See Fig. 2 (right). In this paper, we will refer to this as a general prism.

We define general prismatic meshes as follows:

Definition 2. A general prismatic mesh $\mathcal{P}_h$ of a bounded polyhedral domain is a face-to-face partition whose elements are general prisms, where h is the maximum diameter of all elements in the mesh.

The following lemma helps us to order the vertices of the prism $\mathcal{P}$. For more details, see also Remark 1.

Lemma 1. *The three edges which connect the two triangular faces of $\mathcal{P}$ are either parallel or if we extend these edges in one direction then they meet each other at some point.*

Proof. The planes containing the three quadrilateral faces intersect in one point or this intersection is empty. □

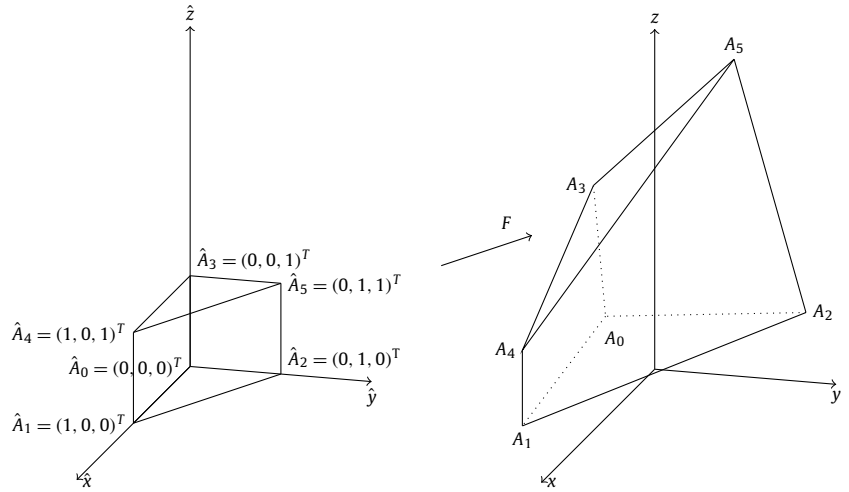


Fig. 2. The reference prism $\hat{\mathcal{P}}$ (left) and an arbitrary prismatic element $\mathcal{P}$ (right). Further, the mapping F is given by formula (9).

Remark 1. We order the vertices of $\mathcal{P}$ similar to [9, pp. 517–518], in such a way that the non-parallel edges $\overline{A_3A_0}$, $\overline{A_4A_1}$ and $\overline{A_5A_2}$ satisfy Lemma 1 and the triangular face $A_3A_4A_5$ is closer to the intersection point than the triangle $A_0A_1A_2$. See Fig. 2 (right). If the edges are parallel, then we do not need to order the vertices similar to the nonparallel case except that the vertices A_0 and A_3 are the end points of the same edge connecting the triangular faces. Furthermore, we assume that in any case the maximum angle for the triangular base is at vertex A_0.

We now define the modification of the semiregularity property from [6] to our setting that will be used throughout the paper.

Definition 3. A family of general prismatic meshes $\mathcal{F} = \{\mathcal{P}_h\}_{h\to 0}$ is semiregular if there exist constants $\overline{\gamma} < \pi$, $c_1 > 0$, and $c_2 > 0$ such that the following conditions hold:

a) **Maximum angle condition** : For any $\mathcal{P} \in \mathcal{P}_h$ and any $\mathcal{P}_h \in \mathcal{F}$ let $\gamma_{\mathcal{P}}$ be the maximum angle of any triangular faces and dihedral angle between any two faces of $\mathcal{P}$. Then

$$\gamma_{\mathcal{P}} \leq \overline{\gamma}. \tag{1}$$

b) **Edge ratio condition** : For any $\mathcal{P} \in \mathcal{P}_h$ and any $\mathcal{P}_h \in \mathcal{F}$ let L_{min} and L_{max} be the minimum and maximum lengths of the three edges connecting the triangular faces. Then

$$\frac{L_{min}}{L_{max}} \geq c_1.$$

c) **Tetrahedra ratio condition** : For any $\mathcal{P} \in \mathcal{P}_h$ and any $\mathcal{P}_h \in \mathcal{F}$ let the vertices of $\mathcal{P}$ be ordered as in Remark 1. Then

$$\frac{\mathrm{Vol}\mathcal{T}(A_0, A_3, A_4, A_5)}{\mathrm{Vol}\mathcal{T}(A_0, A_1, A_2, A_3)} \geq c_2.$$

Lemma 2. *The conditions a), b) and c) from Definition 3 are independent.*

Proof. In Figs. 3-5 we present examples showing the independence. In each figure, all vertices of triangles on the base and on the top of the considered prisms are denoted by $\bullet$ and $\bigcirc$, respectively.

Consider first a case in which *a)* fails, but *b)* and *c)* hold, see Fig. 3. Let $A_0 = (0,0,0)$, $A_1 = (-h, -h^2, 0)$, $A_2 = (h,0,0)$, $A_3 = (0,0,h)$, $A_4 = (-h, -h^2, h)$, and $A_5 = (h,0,h)$ be the vertices of the prism. In this case, $\measuredangle A_1A_0A_2 \to \pi$ as $h \to 0$, so condition *a)* fails. On the other hand, conditions *b)* and *c)* hold with $c_1 = 1$ and $c_2 = 1$.

Consider next a case in which *c)* fails but *a)* and *b)* hold, see Fig. 4. Let $A_0 = (0, \frac{\sqrt{3}}{3}h, 0)$, $A_1 = (-\frac{1}{2}h, -\frac{\sqrt{3}}{6}h, 0)$, $A_2 = (\frac{1}{2}h, -\frac{\sqrt{3}}{6}h, 0)$, $A_3 = (0, \frac{\sqrt{3}}{3}h^2, h)$, $A_4 = (-\frac{1}{2}h^2, -\frac{\sqrt{3}}{6}h^2, h)$, and $A_5 = (\frac{1}{2}h^2, -\frac{\sqrt{3}}{6}h^2, h)$. The triangles on the base and top of the prism are equilateral. Now, if h tends to zero, this family degenerates into a regular tetrahedron, so clearly condition *a)* holds. Further, condition *b)* with $c_1 = 1$ is fulfilled, meanwhile condition *c)* is violated, since the ratio of the volume of the two tetrahedra $\mathcal{T}(A_0, A_3, A_4, A_5)$ and $\mathcal{T}(A_0, A_1, A_2, A_3)$ is h^2.

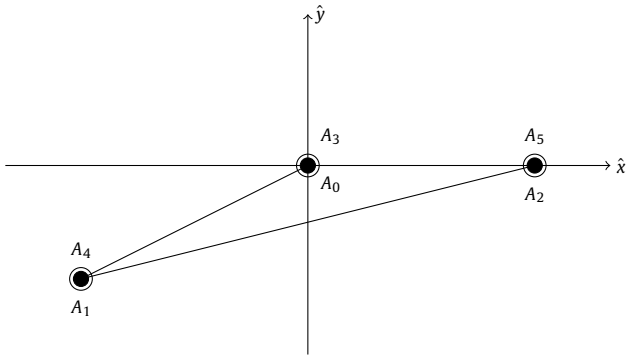


Fig. 3. Orthogonal projection of prism onto $\hat{x}\hat{y}$-plane with vertices $A_0 = (0, 0, 0)$, $A_1 = (-h, -h^2, 0)$, $A_2 = (h, 0, 0)$, $A_3 = (0, 0, h)$, $A_4 = (-h, -h^2, h)$, $A_5 = (h, 0, h)$.

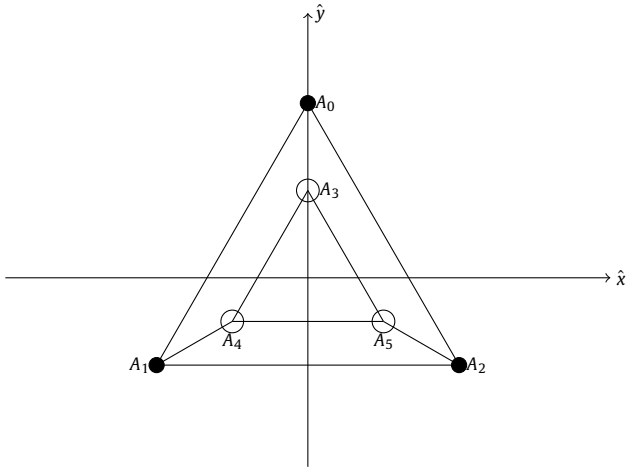


Fig. 4. Orthogonal projection of the prism onto $\hat{x}\hat{y}$-plane with vertices $A_0 = (0, \frac{\sqrt{3}}{3}h, 0)$, $A_1 = (-\frac{1}{2}h, -\frac{\sqrt{3}}{6}h, 0)$, $A_2 = (\frac{1}{2}h, -\frac{\sqrt{3}}{6}h, 0)$, $A_3 = (0, \frac{\sqrt{3}}{3}h^2, h)$, $A_4 = (-\frac{1}{2}h^2, -\frac{\sqrt{3}}{6}h^2, h)$, $A_5 = (\frac{1}{2}h^2, -\frac{\sqrt{3}}{6}h^2, h)$.

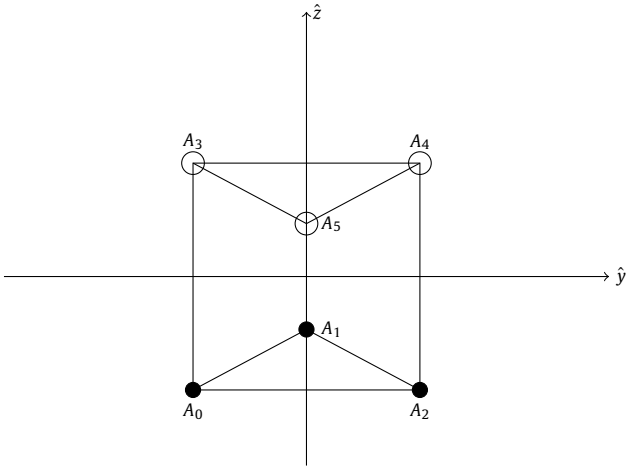


Fig. 5. Orthogonal projection of the prism onto $\hat{y}\hat{z}$-plane with vertices $A_0 = (0, -h, -h)$, $A_1 = (h, 0, -h^2)$, $A_2 = (0, h, -h)$, $A_3 = (0, -h, h)$, $A_4 = (0, h, h)$, $A_5 = (h, 0, h^2)$.

Finally, we consider a case in which b) fails, but the two other conditions hold. Let $A_0 = (0, -h, -h)$, $A_1 = (h, 0, -h^2)$, $A_2 = (0, h, -h)$, $A_3 = (0, -h, h)$, $A_4 = (0, h, h)$, and $A_5 = (h, 0, h^2)$, see Fig. 5. Now assume that $h \to 0$. Then the family degenerates into a pyramid, so clearly condition a) holds. Moreover, the family satisfies condition c) with $c_2 = 1$. But condition b) is violated, since $L_{min}/L_{max} = h$. □

The condition c) in Definition 3 implies bounds on ratios of the volumes of other tetrahedra as well. We will see this in Lemma 5.

To prove Lemma 5, we need the following lemmas from [9].

Lemma 3. *[9] Let $\zeta \le \eta \le \tau$ be angles of an arbitrary face of an arbitrary tetrahedron. Assume furthermore that $\tau \le \overline{\gamma}$. Then $\tau \ge \pi/3$ and*

$$\eta, \tau \in \Big[\frac{\pi - \overline{\gamma}}{2}, \overline{\gamma}\Big].$$

Lemma 4. *[9] Let A be an arbitrary vertex of an arbitrary tetrahedron $\mathcal{T}$ and let $\chi \le \psi \le \varphi$ be angles between faces passing through A. Assume furthermore that $\varphi \le \overline{\gamma}$. Then $\varphi > \pi/3$ and*

$$\psi, \varphi \in \Big(\frac{\pi - \overline{\gamma}}{2}, \overline{\gamma}\Big].$$

Lemma 5. *There exist positive constants $C_i(c_1, m)$, $i = 0, \dots, 3$, which depend only on c_1 and m such that*

i)

$$\mathrm{Vol}\big(\mathcal{T}(A_0, A_1, A_2, A_3)\big) \ge C_0(c_1, m) a b L_{max},$$

where

$$m := \min\Big(\sin(\frac{\pi - \overline{\gamma}}{2}), \sin(\overline{\gamma})\Big),$$

$a = \mid A_1A_0 \mid$, and $b = \mid A_2A_0 \mid$.

ii) *The ratio of the volumes of the tetrahedra $\mathcal{T}(A_0, A_1, A_2, A_4)$ and $\mathcal{T}(A_0, A_1, A_2, A_3)$ is bounded from below by $C_1(c_1, m)$.*

iii) *The ratio of the volumes of the tetrahedra $\mathcal{T}(A_0, A_1, A_2, A_5)$ and $\mathcal{T}(A_0, A_1, A_2, A_3)$ is bounded from below by $C_2(c_1, m)$.*

iv) *The ratio of the volumes of the tetrahedra $\mathcal{T}(A_0, A_1, A_2, A_3)$ and $\mathcal{T}(A_0, A_3, A_4, A_5)$ is bounded from below by $C_3(c_1, m)$.*

Proof. *i*) The rays $\overrightarrow{A_4A_1}$ and $\overrightarrow{A_3A_0}$ meet each other at some point or are parallel. One is depicted in Fig. 6, where the angle between the lines $\overrightarrow{A_4A_3}$ and $\overrightarrow{A_0A_3}$, denoted by θ, is not the smallest angle in the triangle $A_0A_3A_4$. Note that for other possibilities we have similar results. Lemma 3 implies that $\sin(\theta)$ is bounded from below by the positive constant m as in [9]. Then

$$\sin(\theta) = \sin(\pi - \theta) = \frac{\mid A_4M \mid}{\mid A_4A_3 \mid} \ge m,$$

and consequently

$$\frac{\mid A_1A_0 \mid}{\mid A_4A_3 \mid} \ge \frac{\mid A_4M \mid}{\mid A_4A_3 \mid} \ge m.$$

Hence,

$$\mid A_1A_0 \mid \ge m \mid A_4A_3 \mid, \tag{2}$$

and similarly

$$\mid A_2A_0 \mid \ge m \mid A_5A_3 \mid . \tag{3}$$

We denote the angles between the edges A_3A_0 and A_1A_0, and the edges A_3A_1 and A_1A_0, by α and β, respectively, see Fig. 6. Now, according to [9, pp. 517–518], Lemmas 3–4, and condition b), if α is greater than or equal to β, we get

$$\begin{aligned}\mathrm{Vol}\big(\mathcal{T}(A_0, A_1, A_2, A_3)\big) &\ge \frac{1}{6}m^3 \mid A_1A_0 \mid\mid A_2A_0 \mid\mid A_3A_0 \mid \\ &\ge \frac{1}{6}c_1m^3 \mid A_1A_0 \mid\mid A_2A_0 \mid L_{max}.\end{aligned}$$

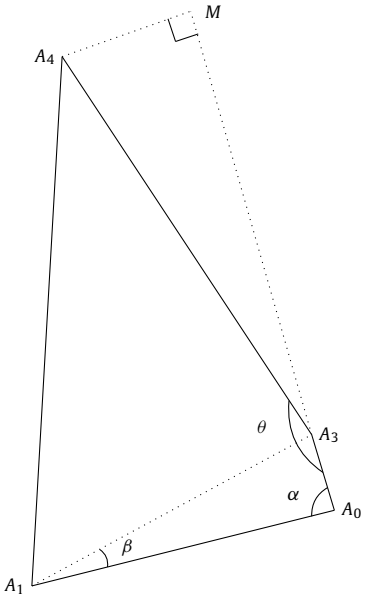


Fig. 6. Quadrilateral face of prism $\mathcal{P}$ made by vertices A_0, A_1, A_3, and A_4.

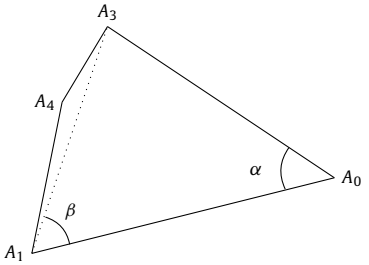


Fig. 7. Quadrilateral face of prism $\mathcal{P}$, where $\beta \geq \alpha$ and $| A_4A_1 | \leq | A_3A_1 |$.

Otherwise β is greater than α. In this case, either $| A_3A_1 | \geq | A_4A_1 |$ or $| A_4A_1 | > | A_3A_1 |$. First, we assume that $| A_3A_1 | \geq | A_4A_1 |$. Then

$$\begin{aligned}\mathrm{Vol}\big(\mathcal{T}(A_0, A_1, A_2, A_3)\big) &\geq \frac{1}{6} m^2 \sin(\alpha) \mid A_1A_0 \mid\mid A_2A_0 \mid\mid A_3A_0 \mid \\ &\geq \frac{1}{6} c_1 m^2 \sin(\alpha) \mid A_1A_0 \mid\mid A_2A_0 \mid L_{max}.\end{aligned}$$

To obtain a lower bound for $\sin(\alpha)$, it suffices to use the law of sines for the triangle $A_0A_1A_3$ (see Fig. 7), conditions a) and b), which implies

$$\sin(\alpha) = \sin(\beta) \frac{| A_3A_1 |}{| A_3A_0 |} \geq m \frac{| A_4A_1 |}{| A_3A_0 |} \geq mc_1, \tag{4}$$

and therefore

$$\mathrm{Vol}\big(\mathcal{T}(A_0, A_1, A_2, A_3)\big) \geq \frac{1}{6} c_1^2 m^3 \sin(\alpha) \mid A_1A_0 \mid\mid A_2A_0 \mid L_{max}.$$

Now, if $| A_4A_1 | > | A_3A_1 |$, we consider the triangle A_0A_1M, see Fig. 8 (which also defines $\delta = \pi - \alpha - \beta$). Then

$$\begin{aligned}\sin(\alpha) &= \sin(\beta + \beta_1) \frac{| MA_1 |}{| MA_0 |} \\ &\geq m \frac{| A_4A_1 |}{| A_3A_0 | + | A_4A_3 | \cos(\pi - \theta)} \\ &\geq m \frac{| A_4A_1 |}{| A_3A_0 | + m^{-1} | A_1A_0 |}.\end{aligned} \tag{5}$$

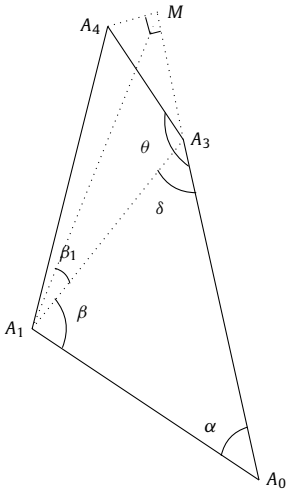


Fig. 8. Quadrilateral face of prism $\mathcal{P}$, where $\beta \geq \alpha$ and $\mid A_4A_1 \mid > \mid A_3A_1 \mid$.

Note that for the above inequalities we used Lemma 3, since $\alpha < \beta + \beta_1$, and (2), respectively. Writing the law of sines for the triangle $A_0A_1A_3$ once again, leads to

$$\mid A_1A_0 \mid = \frac{\sin(\delta)}{\sin(\beta)} \mid A_3A_0 \mid \leq m^{-1} \mid A_3A_0 \mid . \tag{6}$$

Substitute the right-hand side of (6) into (5), we have

$$\sin(\alpha) \geq \frac{m^3}{(1+m^2)} \frac{\mid A_4A_1 \mid}{\mid A_3A_0 \mid} \geq \frac{c_1 m^3}{(1+m^2)} \geq \frac{1}{2} c_1 m^3,$$

and consequently

$$\mathrm{Vol}\big(\mathcal{T}(A_0, A_1, A_2, A_3)\big) \geq \frac{1}{12} c_1^2 m^5 \mid A_1A_0 \mid\mid A_2A_0 \mid L_{max}.$$

ii) To estimate a lower bound for the ratio of the volumes of the tetrahedra $\mathcal{T}(A_0, A_1, A_2, A_4)$ and $\mathcal{T}(A_0, A_1, A_2, A_3)$, if $\angle A_4A_1A_0$ is greater than or equal to $\angle A_4A_0A_1$, condition *b*) implies

$$\frac{\mathrm{Vol}\big(\mathcal{T}(A_0, A_1, A_2, A_4)\big)}{\mathrm{Vol}\big(\mathcal{T}(A_0, A_1, A_2, A_3)\big)} \geq \frac{m^3 \mid A_1A_0 \mid\mid A_2A_0 \mid\mid A_4A_1 \mid}{\mid A_1A_0 \mid\mid A_2A_0 \mid L_{max}} \geq c_1 m^3.$$

Otherwise, exchanging the indices of the vertices A_0, A_3, A_4, A_1 in Fig. 8 into $1, 4, 3, 0$, respectively, and following the same proof as in part *i*), there exists a positive constant $C^*(c_1, m)$ such that

$$\mathrm{Vol}\big(\mathcal{T}(A_0, A_1, A_2, A_4)\big) \geq C^*(c_1, m) \mid A_1A_0 \mid\mid A_2A_0 \mid L_{max}.$$

iii) The proof is same as in parts *i*) and *ii*).

iv) From part *i*), (2) and (3), we have

$$\frac{\mathrm{Vol}\big(\mathcal{T}(A_0, A_1, A_2, A_3)\big)}{\mathrm{Vol}\big(\mathcal{T}(A_0, A_3, A_4, A_5)\big)} \geq \frac{C_0(c_1, m) \mid A_1A_0 \mid\mid A_2A_0 \mid L_{max}}{\mid A_4A_3 \mid\mid A_5A_3 \mid L_{max}} \geq m^2 C_0(c_1, m). \quad \square$$

In what follows, we use the standard denotation $W_p^k(\Omega)$, $k = 0, 1, \ldots, p \geq 1$, for Sobolev spaces with norms $\|.\|_{k,p} = \|.\|_{k,p,\Omega}$ and seminorms $|.|_{k.p} = |.|_{k,p,\Omega}$. The symbol $C(\overline{\Omega})$ stands for the space of continuous functions over $\overline{\Omega}$.

To prove the main result of the paper we will employ the technique based on a transfer of the prism $\mathcal{P} \in \mathcal{P}_h$ onto the reference prism $\hat{\mathcal{P}} = \hat{\mathcal{K}} \times \hat{\mathcal{I}}$, where $\hat{\mathcal{K}}$ is the triangular base and $\hat{\mathcal{I}}$ is the altitude of $\hat{\mathcal{P}}$.

The vertices $\hat{A}_0, \ldots, \hat{A}_5$ of the prism $\hat{\mathcal{P}}$ are given in Fig. 2 (left). The associated basis functions $\hat{\phi}_0, \ldots, \hat{\phi}_5$ for bilinear functions are

$$\begin{aligned}
\hat{\phi}_0(\hat{x},\hat{y},\hat{z}) &= (1-\hat{x}-\hat{y})(1-\hat{z}),\\
\hat{\phi}_1(\hat{x},\hat{y},\hat{z}) &= \hat{x}(1-\hat{z}),\\
\hat{\phi}_2(\hat{x},\hat{y},\hat{z}) &= \hat{y}(1-\hat{z}),\\
\hat{\phi}_3(\hat{x},\hat{y},\hat{z}) &= (1-\hat{x}-\hat{y})\hat{z},\\
\hat{\phi}_4(\hat{x},\hat{y},\hat{z}) &= \hat{x}\hat{z},\\
\hat{\phi}_5(\hat{x},\hat{y},\hat{z}) &= \hat{y}\hat{z}.
\end{aligned} \tag{7}$$

The prismatic interpolant $\hat{\pi}_{\hat{\mathcal{P}}}$ of the function $\hat{u}$ defined on $\hat{\mathcal{P}}$ is constructed as follows:

$$\hat{\pi}_{\hat{\mathcal{P}}}\hat{u} = \sum_{i=0}^{5} \hat{u}(\hat{A}_i)\hat{\phi}_i. \tag{8}$$

By definition, $\hat{\pi}_{\hat{\mathcal{P}}}\hat{u}(\hat{A}_i) = \hat{u}(\hat{A}_i)$, $i = 0,\ldots,5$, for any $\hat{u} \in C(\hat{\mathcal{P}})$.

Let

$$F(\hat{x},\hat{y},\hat{z}) = \sum_{i=0}^{5} A_i\hat{\phi}_i(\hat{x},\hat{y},\hat{z}). \tag{9}$$

Equation (9) defines a mapping $F : \hat{\mathcal{P}} \to \mathcal{P}$, which is a bijection from the prism $\hat{\mathcal{P}}$ onto the prism $\mathcal{P}$. Hence we can define ϕ_i on $\mathcal{P}$ such that

$$\phi_i(A) = \hat{\phi}_i(\hat{A}) = \hat{\phi}_i(F^{-1}(A)), \text{ for all points } A \text{ of } \mathcal{P} \in \mathcal{P}_h.$$

With any prismatic mesh $\mathcal{P}_h$ we associate the finite element space

$$V_h = \{u \in C(\overline{\Omega}) \mid u|_{\mathcal{P}} \in Q(\mathcal{P}) \quad \forall \mathcal{P} \in \mathcal{P}_h\},$$

where $Q(\mathcal{P}) = \{\varphi \mid \varphi = \sum_{i=0}^{5} c_i\phi_i\}$. For similar cases, see [1], Section 5.3. Then the interpolation operator $\pi_h : C(\overline{\Omega}) \to V_h$ is uniquely determined by the requirement

$$\pi_h u(A_i) = u(A_i) \text{ for } A_i,\ i = 0,\ldots,5 \text{ of } \mathcal{P} \in \mathcal{P}_h. \tag{10}$$

Consider $\mathcal{B}$ be a (3×5) matrix whose entries are denoted by B_{ij},

$$\mathcal{B} = \left[\ \mathcal{B}_{1:\mathcal{P}} \ \middle|\ \mathcal{B}_{2:\mathcal{P}}\ \right],$$

where

$$\begin{aligned}
\mathcal{B}_{1:\mathcal{P}} &= \begin{bmatrix} B_{11} & B_{12} & B_{13}\\ B_{21} & B_{22} & B_{23}\\ B_{31} & B_{32} & B_{33}\end{bmatrix}\\
&= \begin{bmatrix} A_{1,x}-A_{0,x} & A_{2,x}-A_{0,x} & A_{3,x}-A_{0,x}\\ A_{1,y}-A_{0,y} & A_{2,y}-A_{0,y} & A_{3,y}-A_{0,y}\\ A_{1,z}-A_{0,z} & A_{2,z}-A_{0,z} & A_{3,z}-A_{0,z}\end{bmatrix},
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{B}_{2:\mathcal{P}} &= \begin{bmatrix} B_{14} & B_{15}\\ B_{24} & B_{25}\\ B_{34} & B_{35}\end{bmatrix}\\
&= \begin{bmatrix} A_{4,x}-A_{0,x}-(B_{11}+B_{13}) & A_{5,x}-A_{0,x}-(B_{12}+B_{13})\\ A_{4,y}-A_{0,y}-(B_{21}+B_{23}) & A_{5,y}-A_{0,y}-(B_{22}+B_{23})\\ A_{4,z}-A_{0,z}-(B_{31}+B_{33}) & A_{5,z}-A_{0,z}-(B_{32}+B_{33})\end{bmatrix}.
\end{aligned}$$

Let $\hat{\mathrm{J}}$ denote the Jacobian of the mapping F. Then

$$\hat{\mathrm{J}} = \begin{bmatrix} \frac{\partial x}{\partial \hat{x}} & \frac{\partial x}{\partial \hat{y}} & \frac{\partial x}{\partial \hat{z}}\\ \frac{\partial y}{\partial \hat{x}} & \frac{\partial y}{\partial \hat{y}} & \frac{\partial y}{\partial \hat{z}}\\ \frac{\partial z}{\partial \hat{x}} & \frac{\partial z}{\partial \hat{y}} & \frac{\partial z}{\partial \hat{z}}\end{bmatrix}$$

$$= \begin{bmatrix} B_{11} + B_{14}\hat{z} & B_{12} + B_{15}\hat{z} & B_{13} + B_{14}\hat{x} + B_{15}\hat{y} \\ B_{21} + B_{24}\hat{z} & B_{22} + B_{25}\hat{z} & B_{23} + B_{24}\hat{x} + B_{25}\hat{y} \\ B_{31} + B_{34}\hat{z} & B_{32} + B_{35}\hat{z} & B_{33} + B_{34}\hat{x} + B_{35}\hat{y} \end{bmatrix}. \tag{11}$$

In order to obtain the rate of convergence of the interpolation operator, we will estimate an upper bound for $|\det(\hat{\mathrm{J}})|^{-1}$, which plays the key role in the proof of Theorem 7. We will show that the lower bound of $|\det(\hat{\mathrm{J}})|$ depends on the volumes of tetrahedra in the prism $\mathcal{P}$.

3. Jacobian determinant

For prisms [6] the determinant of the Jacobian is a constant. We see that for the general prisms, according to (11), $\det(\hat{\mathrm{J}})$ is a polynomial in terms of $\hat{x}$, $\hat{y}$, and $\hat{z}$. To show that $\det(\hat{\mathrm{J}}) \neq 0$, using the linearity property of the determinant, the Jacobian determinant has the explicit form

$$\det(\hat{\mathrm{J}}) = \mathbf{A} + \mathbf{B}\hat{x} + \mathbf{C}\hat{y} + \mathbf{D}\hat{z} - \mathbf{E}\hat{x}\hat{z} - \mathbf{F}\hat{y}\hat{z} + \mathbf{G}\hat{z}^2, \tag{12}$$

where

$$\begin{aligned}
\mathbf{A} &= \begin{vmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \\ B_{31} & B_{32} & B_{33} \end{vmatrix} \\
&= 6\mathrm{Vol}\big(\mathcal{T}(A_0, A_1, A_2, A_3)\big), \\
\mathbf{B} &= \begin{vmatrix} B_{11} & B_{12} & B_{14} \\ B_{21} & B_{22} & B_{24} \\ B_{31} & B_{32} & B_{34} \end{vmatrix} \\
&= 6\mathrm{Vol}\big(\mathcal{T}(A_0, A_1, A_2, A_4)\big) - \mathbf{A} \\
&= \mathbf{B}_1 - \mathbf{A}, \\
\mathbf{C} &= \begin{vmatrix} B_{11} & B_{12} & B_{15} \\ B_{21} & B_{22} & B_{25} \\ B_{31} & B_{32} & B_{35} \end{vmatrix} \\
&= 6\mathrm{Vol}\big(\mathcal{T}(A_0, A_1, A_2, A_5)\big) - \mathbf{A} \\
&= \mathbf{C}_1 - \mathbf{A}, \\
\mathbf{D} &= \begin{vmatrix} B_{12} & B_{13} & B_{14} \\ B_{22} & B_{23} & B_{24} \\ B_{32} & B_{33} & B_{34} \end{vmatrix} - \begin{vmatrix} B_{11} & B_{13} & B_{15} \\ B_{21} & B_{23} & B_{25} \\ B_{31} & B_{33} & B_{35} \end{vmatrix} \\
&= 6\big\{\mathrm{Vol}\big(\mathcal{T}(A_0, A_2, A_3, A_4)\big) + \mathrm{Vol}\big(\mathcal{T}(A_0, A_1, A_5, A_3)\big)\big\} - 2\mathbf{A} \\
&= \mathbf{D}_1 + \mathbf{D}_2 - 2\mathbf{A}, \\
\mathbf{E} &= \begin{vmatrix} B_{11} & B_{14} & B_{15} \\ B_{21} & B_{24} & B_{25} \\ B_{31} & B_{34} & B_{35} \end{vmatrix} \\
&= 6\big\{\mathrm{Vol}\big(\mathcal{T}(A_0, A_3, A_1, A_5)\big) - \mathrm{Vol}\big(\mathcal{T}(A_0, A_4, A_1, A_5)\big)\big\} + \mathbf{B} \\
&= \mathbf{D}_2 - \mathbf{E}_1 + \mathbf{B}, \\
\mathbf{F} &= \begin{vmatrix} B_{12} & B_{14} & B_{15} \\ B_{22} & B_{24} & B_{25} \\ B_{32} & B_{34} & B_{35} \end{vmatrix} \\
&= 6\big\{\mathrm{Vol}\big(\mathcal{T}(A_0, A_2, A_3, A_4)\big) - \mathrm{Vol}\big(\mathcal{T}(A_0, A_2, A_5, A_4)\big)\big\} + \mathbf{C} \\
&= \mathbf{D}_1 - \mathbf{F}_1 + \mathbf{C}, \\
\mathbf{G} &= \begin{vmatrix} B_{13} & B_{14} & B_{15} \\ B_{23} & B_{24} & B_{25} \\ B_{33} & B_{34} & B_{35} \end{vmatrix} \\
&= 6\big\{\mathrm{Vol}\big(\mathcal{T}(A_0, A_3, A_4, A_5)\big) - \mathrm{Vol}\big(\mathcal{T}(A_0, A_3, A_4, A_2)\big) - \mathrm{Vol}\big(\mathcal{T}(A_0, A_3, A_1, A_5)\big)\big\} + \mathbf{A} \\
&= \mathbf{G}_1 - \mathbf{D}_1 - \mathbf{D}_2 + \mathbf{A}.
\end{aligned}$$

Therefore,

$$\begin{aligned}\det(\hat{\mathrm{J}}) = {} & \mathbf{A}(1-\hat{x}-\hat{y})(1-\hat{z}) + \mathbf{B}_1\hat{x}(1-\hat{z}) + \mathbf{C}_1\hat{y}(1-\hat{z}) + \mathbf{D}_1\hat{z}(1-\hat{y}) \\ & + \mathbf{D}_2\hat{z}(1-\hat{x}) + \mathbf{E}_1\hat{x}\hat{z} + \mathbf{F}_1\hat{y}\hat{z} + \mathbf{G}_1\hat{z}^2 + \mathbf{A}\hat{z}^2 - \mathbf{A}\hat{z} - (\mathbf{D}_1+\mathbf{D}_2)\hat{z}^2. \end{aligned} \tag{13}$$

Theorem 6. *Let $\mathcal{F} = \{\mathcal{P}_h\}_{h\to 0}$ be a semiregular family of general prisms of a bounded polygonal domain. Then, there exists a positive constant $\bar{C}(c_1, c_2, m)$ which depends on c_1, c_2 and m, such that*

$$|\det(\hat{\mathrm{J}})|^{-1} \le \bar{C}(c_1, c_2, m)\big(abL_{max}\big)^{-1}. \tag{14}$$

Proof. For a fixed $\hat{z} = \hat{z}_0$, $\det(\hat{\mathrm{J}})$ is linear, and thus attains its maximum and minimum at vertices of the triangle $0 \le \hat{x}, \hat{y} \le 1, \hat{x}+\hat{y} \le 1, \hat{z} = \hat{z}_0$. Therefore it is enough to consider the restriction of $\det(\hat{\mathrm{J}})$ to the three vertical lines. Then the extremal values of $\det(\hat{\mathrm{J}})$ can be found at one of these points: the six vertices of the prism $\hat{A}_i, i = 0, \dots, 5$, as well as if points $(0, 0, -\mathbf{D}/2\mathbf{G})$, $(1, 0, (\mathbf{E}-\mathbf{D})/2\mathbf{G})$ and $(0, 1, (\mathbf{F}-\mathbf{D})/2\mathbf{G})$ are in the domain of definition.

Now, if the minimum value of $\det(\hat{\mathrm{J}})$ occurs at one of the vertices $\hat{A}_i, i = 0, \dots, 5$, then

$$\min_{\{\hat{A}_0,\dots,\hat{A}_5\}} \det(\hat{\mathrm{J}}) = \min\{\mathbf{A}, \mathbf{B}_1, \mathbf{C}_1, \mathbf{G}_1, \mathbf{G}_1+\mathbf{E}_1-\mathbf{D}_2, \mathbf{G}_1+\mathbf{F}_1-\mathbf{D}_1\}. \tag{15}$$

On the right-hand side of (15), all terms are six times the volume of a tetrahedron. Indeed,

$$\begin{aligned}\mathbf{G}_1+\mathbf{E}_1-\mathbf{D}_2 &= 6\Big\{\mathrm{Vol}\big(\mathcal{P}(A_0,\dots,A_5)\big) - \mathrm{Vol}\big(\mathcal{T}(A_0,A_1,A_2,A_5)\big)\Big\} - \mathbf{D}_2 \\ &= 6\Big\{\mathrm{Vol}\big(\mathcal{P}(A_0,\dots,A_5)\big) \\ &\qquad - \big\{\mathrm{Vol}\big(\mathcal{T}(A_0,A_1,A_2,A_5)\big) + \mathrm{Vol}\big(\mathcal{T}(A_0,A_1,A_5,A_3)\big)\big\}\Big\} \\ &= 6\mathrm{Vol}\big(\mathcal{T}(A_1,A_4,A_3,A_5)\big), \end{aligned} \tag{16}$$

and

$$\begin{aligned}\mathbf{G}_1+\mathbf{F}_1-\mathbf{D}_1 &= 6\Big\{\mathrm{Vol}\big(\mathcal{P}(A_0,\dots,A_5)\big) - \mathrm{Vol}\big(\mathcal{T}(A_0,A_1,A_2,A_4)\big)\Big\} - \mathbf{D}_1 \\ &= 6\Big\{\mathrm{Vol}\big(\mathcal{P}(A_0,\dots,A_5)\big) \\ &\qquad - \big\{\mathrm{Vol}\big(\mathcal{T}(A_0,A_1,A_2,A_4)\big) + \mathrm{Vol}\big(\mathcal{T}(A_0,A_2,A_3,A_4)\big)\big\}\Big\} \\ &= 6\mathrm{Vol}\big(\mathcal{T}(A_3,A_4,A_5,A_2)\big). \end{aligned} \tag{17}$$

Now, Lemma 5 provides the lower bounds for $\mathbf{A}$, $\mathbf{B_1}$, and $\mathbf{C_1}$. In addition, condition c) and part i) of Lemma 5 imply that

$$\mathbf{G}_1 \ge c_2 C_0(c_1, m)abL_{max}.$$

Using the same proof as in Lemma 5 for (16) and (17), we obtain the lower bounds which consist of constants in terms of c_1 and m, times abL_{max}.

Now if the critical point $P_{(c)} = (0, 0, -\mathbf{D}/2\mathbf{G})$ is a point, where $\det(\hat{\mathrm{J}})$ has a minimum value, we have

$$\det(\hat{\mathrm{J}})(P_{(c)}) = \mathbf{A} - \frac{\mathbf{D}^2}{4\mathbf{G}}. \tag{18}$$

Due to the valid interval of $\hat{z}$, there are two possibilities for $\mathbf{D}$ and $\mathbf{G}$, $\mathbf{D} > 0, \mathbf{G} < 0$ or $\mathbf{D} < 0, \mathbf{G} > 0$. When $\mathbf{G} < 0$, we obtain $\det(\hat{\mathrm{J}})(P_{(c)}) \ge \mathbf{A}$. For $\mathbf{D} < 0$,

$$\det(\hat{\mathrm{J}})(P_{(c)}) = \frac{1}{4\mathbf{G}}\big\{4\mathbf{A}\mathbf{G}_1 - (\mathbf{D}_1+\mathbf{D}_2)^2\big\}. \tag{19}$$

If $\mathbf{A} \le \mathbf{G}_1$, we get

$$\begin{aligned}\det(\hat{\mathrm{J}})(P_{(c)}) &\ge \frac{1}{4\mathbf{G}}\big\{2\mathbf{A} - (\mathbf{D}_1+\mathbf{D}_2)\big\}\big\{2\mathbf{A}+\mathbf{D}_1+\mathbf{D}_2\big\} \\ &\ge \lambda_1\big\{2\mathbf{A}+\mathbf{D}_1+\mathbf{D}_2\big\} \\ &> 2\lambda_1\mathbf{A}, \end{aligned} \tag{20}$$

where

$$0 < \lambda_1 = \frac{2\mathbf{A} - (\mathbf{D}_1 + \mathbf{D}_2)}{4\mathbf{G}} = \frac{1}{2} P_{(c)} \leq \frac{1}{4}.$$

When λ_1 tends to zero, consequently $P_{(c)}$ tends to $(0, 0, 0)$, and due to Lemma 5, the family of functions $\det(\hat{\mathbf{J}})(P_{(c)})$ for all $\mathcal{P} \in \mathcal{P}_h \in \mathcal{F}$ is equicontinuous and by (18) we obtain

$$\det(\hat{\mathbf{J}})(P_{(c)}) \to \mathbf{A}.$$

Otherwise, $\mathbf{G}_1 < \mathbf{A}$ and

$$\det(\hat{\mathbf{J}})(P_{(c)}) \geq \frac{1}{4\mathbf{G}} \big\{2\mathbf{G}_1 - (\mathbf{D}_1 + \mathbf{D}_2)\big\}\big\{2\mathbf{G}_1 + \mathbf{D}_1 + \mathbf{D}_2\big\}$$

Since in this case, condition $2\mathbf{G}_1 - (\mathbf{D}_1 + \mathbf{D}_2) < 0$ leads to $P_{(c)}$ be outside of the domain, then the valid condition is $2\mathbf{G}_1 - (\mathbf{D}_1 + \mathbf{D}_2) > 0$. Hence,

$$\begin{aligned} \det(\hat{\mathbf{J}})(P_{(c)}) &\geq \lambda_2 \big\{2\mathbf{G}_1 + \mathbf{D}_1 + \mathbf{D}_2\big\} \\ &> 2\lambda_2 \mathbf{G}_1, \end{aligned} \tag{21}$$

where

$$0 < \lambda_2 = \frac{2\mathbf{G}_1 - (\mathbf{D}_1 + \mathbf{D}_2)}{4\mathbf{G}} < \frac{1}{4}.$$

When λ_2 tends to zero, $\mathbf{G}_1$ and $(\mathbf{D}_1 + \mathbf{D}_2)/2$ tend together, and from definition of $\mathbf{G}$ we have

$$2\mathbf{G} \to 2\mathbf{A} - (\mathbf{D}_1 + \mathbf{D}_2) = -\mathbf{D}. \tag{22}$$

This means that $P_{(c)}$ tends to $(0, 0, 1)$ and by (18) and condition c), the family of functions $\det(\hat{\mathbf{J}})(P_{(c)})$ for all $\mathcal{P} \in \mathcal{P}_h$ and $\mathcal{P}_h \in \mathcal{F}$ is equicontinuous, and $\det(\hat{\mathbf{J}})(P_{(c)})$ tends to

$$A + \frac{1}{2}\mathbf{D} = \frac{1}{2}(\mathbf{D}_1 + \mathbf{D}_2) \to \mathbf{G}_1 \geq c_2 \mathbf{A}. \tag{23}$$

Now, let $P_{(c)} = (1, 0, (\mathbf{E} - \mathbf{D})/2\mathbf{G})$ be a critical point, where the Jacobian matrix has a minimum value. Then

$$\det(\hat{\mathbf{J}})(P_{(c)}) = \mathbf{B}_1 - \frac{(\mathbf{E} - \mathbf{D})^2}{4\mathbf{G}}. \tag{24}$$

Since $(\mathbf{E} - \mathbf{D})/2\mathbf{G} \in (0, 1)$, then we have either $\mathbf{D} > \mathbf{E}, \mathbf{G} < 0$ or $\mathbf{D} < \mathbf{E}, \mathbf{G} > 0$. For the first case, from (24), we get

$$\det(\hat{\mathbf{J}})(P_{(c)}) \geq \mathbf{B}_1. \tag{25}$$

For the case that $\mathbf{D} < \mathbf{E}, \mathbf{G} > 0$, if $\mathbf{B}_1 \leq \mathbf{G}$, we have

$$\begin{aligned} \det(\hat{\mathbf{J}})(P_{(c)}) &= \frac{1}{4\mathbf{G}}\{4\mathbf{B}_1\mathbf{G} - (\mathbf{E} - \mathbf{D})^2\} \\ &\geq \frac{1}{4\mathbf{G}}\{2\mathbf{B}_1 - (\mathbf{E} - \mathbf{D})\}\{2\mathbf{B}_1 + \mathbf{E} - \mathbf{D}\} \\ &\geq \frac{1}{4\mathbf{G}}\{2\mathbf{B}_1 - (\mathbf{E} - \mathbf{D})\}\{\mathbf{E} - \mathbf{D}\} \\ &\geq \lambda_3\{\mathbf{B}_1 + \mathbf{E}_1 + \mathbf{D}_1 - \mathbf{A}\}, \end{aligned}$$

where

$$0 < \lambda_3 = \frac{\mathbf{E} - \mathbf{D}}{4\mathbf{G}} < \frac{1}{2}.$$

Furthermore,

$$\begin{aligned} \mathbf{B}_1 + \mathbf{E}_1 + \mathbf{D}_1 - \mathbf{A} &= 6\Big\{\mathrm{Vol}\big(\mathcal{T}(A_0, A_1, A_2, A_4)\big) + \mathrm{Vol}\big(\mathcal{T}(A_0, A_4, A_1, A_5)\big) \\ &\quad + \mathrm{Vol}\big(\mathcal{T}(A_0, A_2, A_3, A_4)\big) - \mathrm{Vol}\big(\mathcal{T}(A_0, A_1, A_2, A_3)\big)\Big\} \\ &= 6\Big\{\mathrm{Vol}\big(\mathcal{P}(A_0, \ldots, A_5)\big) - \mathrm{Vol}\big(\mathcal{T}(A_2, A_5, A_3, A_4)\big) \\ &\quad + \mathrm{Vol}\big(\mathcal{T}(A_0, A_4, A_1, A_5)\big) - \mathrm{Vol}\big(\mathcal{T}(A_0, A_1, A_2, A_3)\big)\Big\}. \end{aligned}$$

Using

$$\begin{aligned}\mathrm{Vol}\big(\mathcal{T}(A_2, A_5, A_3, A_4)\big) + \mathrm{Vol}\big(\mathcal{T}(A_0, A_1, A_2, A_3)\big)\\ = \mathrm{Vol}\big(\mathcal{P}(A_0, \ldots, A_5)\big) - \mathrm{Vol}\big(\mathcal{T}(A_3, A_4, A_2, A_1)\big),\end{aligned}$$

implies

$$\mathbf{B}_1 + \mathbf{E}_1 + \mathbf{D}_1 - \mathbf{A} = 6\Big\{\mathrm{Vol}\big(\mathcal{T}(A_0, A_4, A_1, A_5)\big) + \mathrm{Vol}\big(\mathcal{T}(A_3, A_4, A_2, A_1)\big)\Big\}. \tag{26}$$

Hence

$$\det(\hat{\mathrm{J}})(P_{(c)}) \geq 6\lambda_3 \mathrm{Vol}\big(\mathcal{T}(A_0, A_4, A_1, A_5)\big), \tag{27}$$

and a same proof as in Lemma 5, parts *i*) and *ii*), to obtain a lower bound for $\mathrm{Vol}_{(3)}\big(\mathcal{T}(A_0, A_4, A_1, A_5)\big)$ implies the desirable result. Further, if $\lambda_3 \to 0$, then $\mathbf{E} - \mathbf{D} \to 0$ and $P_{(c)} \to (1, 0, 0)$, and according to Lemma 5, the family of Jacobian determinant at $P_{(c)}$ for all $\mathcal{P} \in \mathcal{P}_h \in \mathcal{F}$ is equicontinuous. Therefore (24) yields

$$\det(\hat{\mathrm{J}})(P_{(c)}) \to \mathbf{B}_1.$$

The other case is $\mathbf{D} < \mathbf{E}$, $\mathbf{G} > 0$ and $\mathbf{G} < \mathbf{B}_1$. Since the third coordinate of $P_{(c)}$ must be in $(0, 1)$, we have

$$(\mathbf{E} - \mathbf{D}) < 2\mathbf{G}. \tag{28}$$

Hence, we use $(\mathbf{E} - \mathbf{D})^2 < 4\mathbf{G}^2$ to obtain

$$\begin{aligned}\det(\hat{\mathrm{J}})(P_{(c)}) &= \frac{1}{4\mathbf{G}}\{4\mathbf{B}_1\mathbf{G} - (\mathbf{E} - \mathbf{D})^2\}\\ &\geq \mathbf{B}_1 - \mathbf{G}\\ &= (1 - \lambda_4)\mathbf{B}_1,\end{aligned}$$

where

$$0 < \lambda_4 = \frac{\mathbf{G}}{\mathbf{B}_1} < 1.$$

If λ_4 tends to zero, then $\mathbf{G}$ tends to zero. By (28), $\mathbf{D} \to \mathbf{E}$ (or conversely), and from (25) we conclude that $\det(\hat{\mathrm{J}})(P_{(c)}) \to \mathbf{B}_1$. When $\lambda_4 \to 1$, then $\mathbf{G} \to \mathbf{B}_1$ and (24) tends to

$$\begin{aligned}\mathbf{B}_1 - \frac{(\mathbf{E} - \mathbf{D})^2}{4\mathbf{B}_1} &= \frac{1}{4\mathbf{B}_1}\{2\mathbf{B}_1 - (\mathbf{E} - \mathbf{D})\}\{2\mathbf{B}_1 + (\mathbf{E} - \mathbf{D})\}\\ &\geq \frac{1}{2}\{2\mathbf{B}_1 - (\mathbf{E} - \mathbf{D})\}\\ &\geq 3\mathrm{Vol}\big(\mathcal{T}(A_0, A_4, A_1, A_5)\big).\end{aligned} \tag{29}$$

Note that for the last inequality we used (26). Now, the same argument for (27) implies (14).

Finally, for $P_{(c)} = (0, 1, (\mathbf{F} - \mathbf{D})/2\mathbf{G})$,

$$\det(\hat{\mathrm{J}})(P_{(c)}) = \mathbf{C}_1 - \frac{(\mathbf{F} - \mathbf{D})^2}{4\mathbf{G}}. \tag{30}$$

Moreover, $\mathbf{D} > \mathbf{F}$, $\mathbf{G} < 0$, or $\mathbf{D} < \mathbf{F}$, $\mathbf{G} > 0$, since $(\mathbf{F} - \mathbf{D})/2\mathbf{G} \in (0, 1)$.

Let $\mathbf{D} > \mathbf{F}$ and $\mathbf{G} < 0$. By (30), we then have

$$\det(\hat{\mathrm{J}})(P_{(c)}) \geq \mathbf{C}_1.$$

Now, let $\mathbf{D} < \mathbf{F}$, $\mathbf{G} > 0$. If $\mathbf{C}_1 < \mathbf{G}$ we have

$$\begin{aligned}\det(\hat{\mathrm{J}})(P_{(c)}) &= \frac{1}{4\mathbf{G}}\{4\mathbf{C}_1\mathbf{G} - (\mathbf{F} - \mathbf{D})^2\}\\ &\geq \frac{1}{4\mathbf{G}}\{2\mathbf{C}_1 - (\mathbf{F} - \mathbf{D})\}\{2\mathbf{C}_1 + \mathbf{F} - \mathbf{D}\}\\ &\geq \frac{1}{4\mathbf{G}}\{2\mathbf{C}_1 - (\mathbf{F} - \mathbf{D})\}\{\mathbf{F} - \mathbf{D}\}\\ &\geq \lambda_5\{\mathbf{C}_1 + \mathbf{F}_1 + \mathbf{D}_2 - \mathbf{A}\},\end{aligned} \tag{31}$$

where

$$0 < \lambda_5 = \frac{\mathbf{F}-\mathbf{D}}{4\mathbf{G}} < \frac{1}{2}.$$

Using

$$\begin{aligned}\mathrm{Vol}\big(\mathcal{T}(A_1, A_4, A_5, A_3)\big) + \mathrm{Vol}\big(\mathcal{T}(A_0, A_1, A_2, A_3)\big)\\ = \mathrm{Vol}\big(\mathcal{P}(A_0, \ldots, A_5)\big) - \mathrm{Vol}\big(\mathcal{T}(A_1, A_2, A_5, A_3)\big),\end{aligned}$$

implies

$$\begin{aligned}\mathbf{C}_1 + \mathbf{F}_1 + \mathbf{D}_2 - \mathbf{A} &= 6\Big\{\mathrm{Vol}\big(\mathcal{T}(A_0, A_1, A_2, A_5)\big) + \mathrm{Vol}\big(\mathcal{T}(A_0, A_2, A_5, A_4)\big)\\ &\quad + \mathrm{Vol}\big(\mathcal{T}(A_0, A_1, A_5, A_3)\big) - \mathrm{Vol}\big(\mathcal{T}(A_0, A_1, A_2, A_3)\big)\Big\}\\ &= 6\Big\{\mathrm{Vol}\big(\mathcal{P}(A_0, \ldots, A_5)\big) - \mathrm{Vol}\big(\mathcal{T}(A_1, A_4, A_5, A_3)\big)\\ &\quad + \mathrm{Vol}\big(\mathcal{T}(A_0, A_2, A_5, A_4)\big) - \mathrm{Vol}\big(\mathcal{T}(A_0, A_1, A_2, A_3)\big)\Big\}\\ &= 6\Big\{\mathrm{Vol}\big(\mathcal{T}(A_0, A_2, A_5, A_4)\big) + \mathrm{Vol}\big(\mathcal{T}(A_1, A_2, A_5, A_3)\big)\Big\}. \qquad (32)\end{aligned}$$

Then we can obtain the lower bound for (31) as follows.

$$\det(\hat{\mathrm{J}})(P_{(c)}) \geq 6\lambda_5 \mathrm{Vol}\big(\mathcal{T}(A_0, A_2, A_5, A_4)\big).$$

Extending the proof of Lemma 5, one comes to (14).

Now, if $\lambda_5 \to 0$, then $\mathbf{F}-\mathbf{D} \to 0$, and as a result $P_{(c)} \to (0,1,0)$. Similar to previous cases, due to Lemma 5, here the family of Jacobian determinant at $P_{(c)}$ for all prisms belonging to $\mathcal{F}$ is also equicontinuous and we get

$$\det(\hat{\mathrm{J}})(P_{(c)}) \to \mathbf{C}_1. \qquad (33)$$

For the case that $\mathbf{G} < \mathbf{C}_1$,

$$\begin{aligned}\det(\hat{\mathrm{J}})(P_{(c)}) &= \frac{1}{4\mathbf{G}}\{4\mathbf{C}_1\mathbf{G} - (\mathbf{F}-\mathbf{D})^2\}\\ &\geq \mathbf{C}_1 - \mathbf{G}\\ &= (1-\lambda_6)\mathbf{C}_1,\end{aligned}$$

where

$$0 < \lambda_6 = \frac{\mathbf{G}}{\mathbf{C}_1} < 1.$$

If λ_6 approaches to zero, then $\mathbf{G} \to 0$, which implies (33).

To end the proof, let λ_6 tend to 1. Then $\mathbf{G} \to \mathbf{C}_1$, and consequently by (32), (30) tends to

$$\begin{aligned}\mathbf{C}_1 - \frac{(\mathbf{F}-\mathbf{D})^2}{4\mathbf{C}_1} &= \frac{1}{4\mathbf{C}_1}\{2\mathbf{C}_1 - (\mathbf{F}-\mathbf{D})\}\{2\mathbf{C}_1 + (\mathbf{F}-\mathbf{D})\}\\ &\geq \frac{1}{2}\{2\mathbf{C}_1 - (\mathbf{F}-\mathbf{D})\}\\ &\geq 3\mathrm{Vol}_{(3)}\big(\mathcal{T}(A_0, A_2, A_5, A_4)\big). \quad \square \qquad (34)\end{aligned}$$

4. Interpolation error

Theorem 7. *Let $u \in W_2^3(\Omega)$ and $\mathcal{F} = \{\mathcal{P}_h\}_{h\to 0}$ be a family of semiregular prismatic partitions of $\bar{\Omega}$. Then, there exists a positive constant C^*, independent of the diameter $h(\mathcal{P})$, such that*

$$|u - \pi_h u|_{1,2,\Omega} \leq C^*\big\{h(\mathcal{P})\,|u|_{2,2,\Omega} + h^2(\mathcal{P})\,|u|_{3,2,\Omega}\big\} \qquad (35)$$

Proof. From the definition of semi-norm we have

$$|u-\pi_{\mathcal{P}}u|^2_{1,2,\mathcal{P}} = \int_{\mathcal{P}} \left(\left| \frac{\partial}{\partial x}(u-\pi_{\mathcal{P}}u) \right|^2 + \left| \frac{\partial}{\partial y}(u-\pi_{\mathcal{P}}u) \right|^2 + \left| \frac{\partial}{\partial z}(u-\pi_{\mathcal{P}}u) \right|^2 \right) dX. \tag{36}$$

To estimate (36), first we will estimate it on $\hat{\mathcal{P}}$. Then, from equation (40) in [6], we have

$$\int_{\hat{\mathcal{P}}} \left| \frac{\partial}{\partial \hat{x}}(\hat{u}-\hat{\pi}_{\hat{\mathcal{P}}}\hat{u}) \right|^2 d\hat{X} \le 12 \int_{\hat{\mathcal{P}}} \left(\left| \frac{\partial^2 \hat{u}}{\partial \hat{x}^2} \right|^2 + \left| \frac{\partial^2 \hat{u}}{\partial \hat{x}\partial \hat{y}} \right|^2 + \left| \frac{\partial^2 \hat{u}}{\partial \hat{x}\partial \hat{z}} \right|^2 + \left| \frac{\partial^3 \hat{u}}{\partial \hat{x}\partial \hat{y}\partial \hat{z}} \right|^2 \right) d\hat{X}, \tag{37}$$

where

$$\begin{aligned}
\left| \frac{\partial^2 \hat{u}}{\partial \hat{x}^2} \right|^2 &= \left| \hat{J}^2_{(11)} \frac{\partial^2 u}{\partial x^2} + \hat{J}^2_{(21)} \frac{\partial^2 u}{\partial y^2} + \hat{J}^2_{(31)} \frac{\partial^2 u}{\partial z^2} \right. \\
&\quad \left. + 2\left\{ \hat{J}_{(11)}\hat{J}_{(21)} \frac{\partial^2 u}{\partial x \partial y} + \hat{J}_{(11)}\hat{J}_{(31)} \frac{\partial^2 u}{\partial x \partial z} + \hat{J}_{(21)}\hat{J}_{(31)} \frac{\partial^2 u}{\partial y \partial z} \right\} \right|^2 \\
&\le 24 \left\{ \hat{J}^4_{(11)} \left| \frac{\partial^2 u}{\partial x^2} \right|^2 + \hat{J}^4_{(21)} \left| \frac{\partial^2 u}{\partial y^2} \right|^2 + \hat{J}^4_{(31)} \left| \frac{\partial^2 u}{\partial z^2} \right|^2 + \hat{J}^2_{(11)}\hat{J}^2_{(21)} \left| \frac{\partial^2 u}{\partial x \partial y} \right|^2 \right. \\
&\quad \left. + \hat{J}^2_{(11)}\hat{J}^2_{(31)} \left| \frac{\partial^2 u}{\partial x \partial z} \right|^2 + \hat{J}^2_{(21)}\hat{J}^2_{(31)} \left| \frac{\partial^2 u}{\partial y \partial z} \right|^2 \right\}.
\end{aligned}$$

For the last inequalities we used the so-called sum of squares inequality

$$\left(\sum_{j=1}^{s} a_j \right)^2 \le s \sum_{j=1}^{s} a_j^2.$$

In the remaining computations, we will use C as an unspecified positive constant. It is not necessarily the same in two lines of a computation, for instance in equation (38). We get

$$\begin{aligned}
\left| \frac{\partial^2 \hat{u}}{\partial \hat{x}\partial \hat{y}} \right|^2 &\le C \left\{ \hat{J}^2_{(11)}\hat{J}^2_{(12)} \left| \frac{\partial^2 u}{\partial x^2} \right|^2 + \hat{J}^2_{(21)}\hat{J}^2_{(22)} \left| \frac{\partial^2 u}{\partial y^2} \right|^2 + \hat{J}^2_{(31)}\hat{J}^2_{(32)} \left| \frac{\partial^2 u}{\partial z^2} \right|^2 \right. \\
&\quad + \left(\hat{J}^2_{(12)}\hat{J}^2_{(21)} + \hat{J}^2_{(11)}\hat{J}^2_{(22)} \right) \left| \frac{\partial^2 u}{\partial x \partial y} \right|^2 + \left(\hat{J}^2_{(12)}\hat{J}^2_{(31)} + \hat{J}^2_{(32)}\hat{J}^2_{(11)} \right) \left| \frac{\partial^2 u}{\partial x \partial z} \right|^2 \\
&\quad \left. + \left(\hat{J}^2_{(22)}\hat{J}^2_{(31)} + \hat{J}^2_{(21)}\hat{J}^2_{(32)} \right) \left| \frac{\partial^2 u}{\partial y \partial z} \right|^2 \right\}, \\
\left| \frac{\partial^2 \hat{u}}{\partial \hat{x}\partial \hat{z}} \right|^2 &\le C \left\{ \hat{J}^2_{(11)}\hat{J}^2_{(13)} \left| \frac{\partial^2 u}{\partial x^2} \right|^2 + \hat{J}^2_{(21)}\hat{J}^2_{(23)} \left| \frac{\partial^2 u}{\partial y^2} \right|^2 + \hat{J}^2_{(31)}\hat{J}^2_{(33)} \left| \frac{\partial^2 u}{\partial z^2} \right|^2 \right. \\
&\quad + \left(\hat{J}^2_{(13)}\hat{J}^2_{(21)} + \hat{J}^2_{(11)}\hat{J}^2_{(23)} \right) \left| \frac{\partial^2 u}{\partial x \partial y} \right|^2 + \left(\hat{J}^2_{(13)}\hat{J}^2_{(31)} + \hat{J}^2_{(33)}\hat{J}^2_{(11)} \right) \left| \frac{\partial^2 u}{\partial x \partial z} \right|^2 \\
&\quad \left. + \left(\hat{J}^2_{(23)}\hat{J}^2_{(31)} + \hat{J}^2_{(21)}\hat{J}^2_{(33)} \right) \left| \frac{\partial^2 u}{\partial y \partial z} \right|^2 \right\}, \\
\left| \frac{\partial^3 \hat{u}}{\partial \hat{x}\partial \hat{y}\partial \hat{z}} \right|^2 &\le C \left\{ \hat{J}^2_{(11)}\hat{J}^2_{(12)}\hat{J}^2_{(13)} \left| \frac{\partial^3 u}{\partial x^3} \right|^2 + \hat{J}^2_{(21)}\hat{J}^2_{(22)}\hat{J}^2_{(23)} \left| \frac{\partial^3 u}{\partial y^3} \right|^2 \right. \\
&\quad + \hat{J}^2_{(31)}\hat{J}^2_{(32)}\hat{J}^2_{(33)} \left| \frac{\partial^3 u}{\partial z^3} \right|^2 + \left(\hat{J}^2_{(12)}\hat{J}^2_{(11)}\hat{J}^2_{(23)} + (\hat{J}^2_{(12)}\hat{J}^2_{(21)} + \hat{J}^2_{(11)}\hat{J}^2_{(22)})\hat{J}^2_{(13)} \right) \left| \frac{\partial^3 u}{\partial x^2 \partial y} \right|^2 \\
&\quad + \left(\hat{J}^2_{(11)}\hat{J}^2_{(12)}\hat{J}^2_{(33)} + (\hat{J}^2_{(12)}\hat{J}^2_{(31)} + \hat{J}^2_{(11)}\hat{J}^2_{(32)})\hat{J}^2_{(13)} \right) \left| \frac{\partial^3 u}{\partial x^2 \partial z} \right|^2 \\
&\quad + \left((\hat{J}^2_{(12)}\hat{J}^2_{(21)} + \hat{J}^2_{(11)}\hat{J}^2_{(22)})\hat{J}^2_{(23)} + \hat{J}^2_{(22)}\hat{J}^2_{(21)}\hat{J}^2_{(13)} \right) \left| \frac{\partial^3 u}{\partial x \partial y^2} \right|^2 \\
&\quad + \left((\hat{J}^2_{(12)}\hat{J}^2_{(21)} + \hat{J}^2_{(11)}\hat{J}^2_{(22)})\hat{J}^2_{(33)} + (\hat{J}^2_{(12)}\hat{J}^2_{(31)} + \hat{J}^2_{(32)}\hat{J}^2_{(11)})\hat{J}^2_{(23)} + (\hat{J}^2_{(22)}\hat{J}^2_{(31)} + \hat{J}^2_{(32)}\hat{J}^2_{(21)})\hat{J}^2_{(13)} \right) \left| \frac{\partial^3 u}{\partial x \partial y \partial z} \right|^2 \\
&\quad + \left(\hat{J}^2_{(22)}\hat{J}^2_{(21)}\hat{J}^2_{(33)} + (\hat{J}^2_{(22)}\hat{J}^2_{(31)} + \hat{J}^2_{(32)}\hat{J}^2_{(21)})\hat{J}^2_{(23)} \right) \left| \frac{\partial^3 u}{\partial y^2 \partial z} \right|^2
\end{aligned}$$

$$+\left(\hat{J}^2_{(32)}\hat{J}^2_{(31)}\hat{J}^2_{(13)}+(\hat{J}^2_{(12)}\hat{J}^2_{(31)}+\hat{J}^2_{(32)}\hat{J}^2_{(21)})\hat{J}^2_{(33)}\right)\left|\frac{\partial^3 u}{\partial x\partial z^2}\right|^2$$
$$+\left(\hat{J}^2_{(32)}\hat{J}^2_{(31)}\hat{J}^2_{(23)}+(\hat{J}^2_{(22)}\hat{J}^2_{(31)}+\hat{J}^2_{(32)}\hat{J}^2_{(21)})\hat{J}^2_{(33)}\right)\left|\frac{\partial^3 u}{\partial y\partial z^2}\right|^2\Big\}.$$

To estimate the upper bounds for $\left|\frac{\partial^2\hat{u}}{\partial\hat{x}\partial\hat{y}}\right|^2$, $\left|\frac{\partial^2\hat{u}}{\partial\hat{x}\partial\hat{z}}\right|^2$, and $\left|\frac{\partial^3\hat{u}}{\partial\hat{x}\partial\hat{y}\partial\hat{z}}\right|^2$, we denote the length of the segments $\overrightarrow{A_4A_3}$ and $\overrightarrow{A_5A_3}$ by c and d respectively. Now, we have

$$(A_{1,x}-A_{0,x})^2+(A_{1,y}-A_{0,y})^2\le a^2,\qquad (A_{4,x}-A_{3,x})^2+(A_{4,y}-A_{3,y})^2\le c^2,$$
$$(A_{2,x}-A_{0,x})^2+(A_{2,y}-A_{0,y})^2\le b^2,\qquad (A_{5,x}-A_{3,x})^2+(A_{5,y}-A_{3,y})^2\le d^2,$$

which imply

$$\begin{aligned}
\left|\hat{J}_{(11)}\right| &= \left|B_{11}+(A_{4,x}-A_{0,x}-(B_{11}+B_{13}))\tilde{z}\right|\\
&\le \left|A_{1,x}-A_{0,x}\right|(1-\tilde{z})+\left|A_{4,x}-A_{3,x}\right|\tilde{z}\\
&\le a+c,\\
\left|\hat{J}_{(21)}\right| &= \left|B_{21}+(A_{4,y}-A_{0,y}-(B_{21}+B_{23}))\tilde{z}\right|\\
&\le \left|A_{1,y}-A_{0,y}\right|(1-\tilde{z})+\left|A_{4,y}-A_{3,y}\right|\tilde{z}\\
&\le a+c,\\
\left|\hat{J}_{(31)}\right| &= \left|B_{31}+(A_{4,z}-A_{0,z}-(B_{31}+B_{33}))\tilde{z}\right|\\
&\le \left|(A_{1,z}-A_{0,z}\right|(1-\tilde{z})+\left|A_{4,z}-A_{3,z}\right|\tilde{z}\\
&\le a+c,
\end{aligned}$$

and similarly

$$\left|\hat{J}_{(12)}\right|\le b+d,\qquad \left|\hat{J}_{(22)}\right|\le b+d,\qquad \left|\hat{J}_{(32)}\right|\le b+d.$$

Now the upper bounds of $|\,J_{(i,j)}\,|$, $i=1,2,3$, $j=1,2$ can be expressed in terms of a, b, and m as follows.

$$\left\{\left|\hat{J}_{(11)}\right|,\left|\hat{J}_{(21)}\right|,\left|\hat{J}_{(31)}\right|\right\}\le(1+m^{-1})a,$$
$$\left\{\left|\hat{J}_{(12)}\right|,\left|\hat{J}_{(22)}\right|,\left|\hat{J}_{(32)}\right|\right\}\le(1+m^{-1})b.$$

Moreover

$$\begin{aligned}
\left|\hat{J}_{(13)}\right| &\le \left|A_{3,x}-A_{0,x}\right|+\left|A_{4,x}-A_{1,x}\right|+\left|A_{5,x}-A_{2,x}\right|\le 3L_{max},\\
\left|\hat{J}_{(23)}\right| &\le \left|A_{3,y}-A_{0,y}\right|+\left|A_{4,y}-A_{1,y}\right|+\left|A_{5,y}-A_{2,y}\right|\le 3L_{max},\\
\left|\hat{J}_{(33)}\right| &\le \left|A_{3,z}-A_{0,z}\right|+\left|A_{4,z}-A_{1,z}\right|+\left|A_{5,z}-A_{2,z}\right|\le 3L_{max}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\left|\frac{\partial^2\hat{u}}{\partial\hat{x}^2}\right|^2 &\le 24(1+m^{-1})^4a^4\left\{\left|\frac{\partial^2 u}{\partial x^2}\right|^2+\left|\frac{\partial^2 u}{\partial y^2}\right|^2+\left|\frac{\partial^2 u}{\partial z^2}\right|^2+\left|\frac{\partial^2 u}{\partial x\partial y}\right|^2+\left|\frac{\partial^2 u}{\partial x\partial z}\right|^2+\left|\frac{\partial^2 u}{\partial y\partial z}\right|^2\right\}\\
&= 24(1+m^{-1})^4a^4\sum_{|\beta|=2}|\,D^\beta u\,|^2,
\end{aligned}$$

and

$$\left|\frac{\partial^2\hat{u}}{\partial\hat{x}\partial\hat{y}}\right|^2\le 24(1+m^{-1})^4a^2b^2\sum_{|\beta|=2}|\,D^\beta u\,|^2,$$
$$\left|\frac{\partial^2\hat{u}}{\partial\hat{x}\partial\hat{z}}\right|^2\le 6^3(1+m^{-1})^2a^2L^2_{max}\sum_{|\beta|=2}|\,D^\beta u\,|^2,$$
$$\left|\frac{\partial^3\hat{u}}{\partial\hat{x}\partial\hat{y}\partial\hat{z}}\right|^2\le 10\times 18^2(1+m^{-1})^4a^2b^2L^2_{max}\sum_{|\beta|=3}|\,D^\beta u\,|^2.$$

Using Theorem 6, (37) can be expressed as follows.

$$\begin{aligned}
&\int_{\hat{\mathcal{P}}} \left| \frac{\partial}{\partial \hat{x}} (\hat{u} - \hat{\pi}_{\hat{\mathcal{P}}} \hat{u}) \right|^2 d\hat{X} \\
&\quad \leq C \int_{\mathcal{P}} |\det(\hat{\mathrm{J}})|^{-1} (1+m^{-1})^2 a^2 \Big\{ \big((1+m^{-1})^2 (a^2+b^2) + L_{max}^2\big) \sum_{|\beta|=2} \left|D^\beta u\right|^2 \\
&\qquad + \big((1+m^{-1})^2 b^2 L_{max}^2\big) \sum_{|\beta|=3} \left|D^\beta u\right|^2 \Big\} dX \\
&\quad \leq C \int_{\mathcal{P}} |\det(\hat{\mathrm{J}})|^{-1} a^2 \Big\{ h^2(\mathcal{P}) \sum_{|\beta|=2} \left|D^\beta u\right|^2 + h^4(\mathcal{P}) \sum_{|\beta|=3} \left|D^\beta u\right|^2 \Big\} dX. \qquad (38)
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&\int_{\hat{\mathcal{P}}} \left| \frac{\partial}{\partial \hat{y}} (\hat{u} - \hat{\pi}_{\hat{\mathcal{P}}} \hat{u}) \right|^2 d\hat{X} \\
&\quad \leq C \int_{\mathcal{P}} |\det(\hat{\mathrm{J}})|^{-1} (1+m^{-1})^2 b^2 \Big\{ \big((1+m^{-1})^2 (a^2+b^2) + L_{max}^2\big) \sum_{|\beta|=2} \left|D^\beta u\right|^2 \\
&\qquad + \big((1+m^{-1})^2 a^2 L_{max}^2\big) \sum_{|\beta|=3} \left|D^\beta u\right|^2 \Big\} dX \\
&\quad \leq C \int_{\mathcal{P}} |\det(\hat{\mathrm{J}})|^{-1} b^2 \Big\{ h^2(\mathcal{P}) \sum_{|\beta|=2} \left|D^\beta u\right|^2 + h^4(\mathcal{P}) \sum_{|\beta|=3} \left|D^\beta u\right|^2 \Big\} dX. \qquad (39)
\end{aligned}$$

From equation (45) in [6], we get

$$\begin{aligned}
&\int_{\hat{\mathcal{P}}} \left| \frac{\partial}{\partial \hat{z}} (\hat{u} - \hat{\pi}_{\hat{\mathcal{P}}} \hat{u}) \right|^2 d\hat{X} \\
&\quad \leq 2 \int_{\hat{\mathcal{P}}} \left| \frac{\partial^2 \hat{u}}{\partial \hat{z}^2} \right|^2 d\hat{X} + C \int_{\hat{\mathcal{P}}} \left(\left| \frac{\partial^3 \hat{u}}{\partial \hat{x}^2 \partial \hat{z}} \right|^2 + \left| \frac{\partial^3 \hat{u}}{\partial \hat{x} \partial \hat{y} \partial \hat{z}} \right|^2 + \left| \frac{\partial^3 \hat{u}}{\partial \hat{y}^2 \partial \hat{z}} \right|^2 \right) d\hat{X}.
\end{aligned}$$

Since

$$\begin{aligned}
&\left| \frac{\partial^2 \hat{u}}{\partial \hat{z}^2} \right|^2 \leq 6 \times 18^2 L_{max}^4 \sum_{|\beta|=2} \left|D^\beta u\right|^2, \\
&\left| \frac{\partial^3 \hat{u}}{\partial \hat{x}^2 \partial \hat{z}} \right|^2 \leq 10 \times 18^2 (1+m^{-1})^4 a^4 L_{max}^2 \sum_{|\beta|=3} \left|D^\beta u\right|^2, \\
&\left| \frac{\partial^3 \hat{u}}{\partial \hat{y}^2 \partial \hat{z}} \right|^2 \leq 10 \times 18^2 (1+m^{-1})^4 b^4 L_{max}^2 \sum_{|\beta|=3} \left|D^\beta u\right|^2,
\end{aligned}$$

we get

$$\begin{aligned}
&\int_{\hat{\mathcal{P}}} \left| \frac{\partial}{\partial \hat{z}} (\hat{u} - \hat{\pi}_{\hat{\mathcal{P}}} \hat{u}) \right|^2 d\hat{X} \\
&\quad \leq C \int_{\mathcal{P}} |\det(\hat{\mathrm{J}})|^{-1} L_{max}^2 \Big\{ L_{max}^2 \sum_{|\beta|=2} \left|D^\beta u\right|^2 + (1+m^{-1})^4 \big(a^4 + b^4 + a^2 b^2\big) \sum_{|\beta|=3} \left|D^\beta u\right|^2 \Big\} dX \\
&\quad \leq C \int_{\mathcal{P}} |\det(\hat{\mathrm{J}})|^{-1} L_{max}^2 \Big\{ h^2(\mathcal{P}) \sum_{|\beta|=2} \left|D^\beta u\right|^2 + h^4(\mathcal{P}) \sum_{|\beta|=3} \left|D^\beta u\right|^2 \Big\} dX. \qquad (40)
\end{aligned}$$

Now, we estimate (36) as follows.

$$
\begin{aligned}
&|u-\pi_{\mathcal{P}}u|^2_{1,2,\mathcal{P}}\\
&=\int_{\hat{\mathcal{P}}}|\det(\hat{\mathrm{J}})|\Big(\Big|\mathrm{J}^{-1}_{(11)}\frac{\partial}{\partial\hat{x}}(\hat{u}-\hat{\pi}_{\hat{\mathcal{P}}}\hat{u})+\mathrm{J}^{-1}_{(21)}\frac{\partial}{\partial\hat{y}}(\hat{u}-\hat{\pi}_{\hat{\mathcal{P}}}\hat{u})+\mathrm{J}^{-1}_{(31)}\frac{\partial}{\partial\hat{z}}(\hat{u}-\hat{\pi}_{\hat{\mathcal{P}}}\hat{u})\Big|^2\\
&\quad+\Big|\mathrm{J}^{-1}_{(12)}\frac{\partial}{\partial\hat{x}}(\hat{u}-\hat{\pi}_{\hat{\mathcal{P}}}\hat{u})+\mathrm{J}^{-1}_{(22)}\frac{\partial}{\partial\hat{y}}(\hat{u}-\hat{\pi}_{\hat{\mathcal{P}}}\hat{u})+\mathrm{J}^{-1}_{(32)}\frac{\partial}{\partial\hat{z}}(\hat{u}-\hat{\pi}_{\hat{\mathcal{P}}}\hat{u})\Big|^2\\
&\quad+\Big|\mathrm{J}^{-1}_{(13)}\frac{\partial}{\partial\hat{x}}(\hat{u}-\hat{\pi}_{\hat{\mathcal{P}}}\hat{u})+\mathrm{J}^{-1}_{(23)}\frac{\partial}{\partial\hat{y}}(\hat{u}-\hat{\pi}_{\hat{\mathcal{P}}}\hat{u})+\mathrm{J}^{-1}_{(33)}\frac{\partial}{\partial\hat{z}}(\hat{u}-\hat{\pi}_{\hat{\mathcal{P}}}\hat{u})\Big|^2\Big)d\hat{X}\\
&\le 3\int_{\hat{\mathcal{P}}}|\det(\hat{\mathrm{J}})|\Big(\Big(|\mathrm{J}^{-1}_{(11)}|^2+|\mathrm{J}^{-1}_{(12)}|^2+|\mathrm{J}^{-1}_{(13)}|^2\Big)\Big|\frac{\partial}{\partial\hat{x}}(\hat{u}-\hat{\pi}_{\hat{\mathcal{P}}}\hat{u})\Big|^2\\
&\quad+\Big(|\mathrm{J}^{-1}_{(21)}|^2+|\mathrm{J}^{-1}_{(22)}|^2+|\mathrm{J}^{-1}_{(23)}|^2\Big)\Big|\frac{\partial}{\partial\hat{y}}(\hat{u}-\hat{\pi}_{\hat{\mathcal{P}}}\hat{u})\Big|^2\\
&\quad+\Big(|\mathrm{J}^{-1}_{(31)}|^2+|\mathrm{J}^{-1}_{(32)}|^2+|\mathrm{J}^{-1}_{(33)}|^2\Big)\Big|\frac{\partial}{\partial\hat{z}}(\hat{u}-\hat{\pi}_{\hat{\mathcal{P}}}\hat{u})\Big|^2\Big)d\hat{X}.
\end{aligned}
\tag{41}
$$

Theorem 6 and computations of cofactors lead to

$$
\begin{aligned}
&\big\{|\mathrm{J}^{-1}_{(11)}|,|\mathrm{J}^{-1}_{(12)}|,|\mathrm{J}^{-1}_{(13)}|\big\}\le\frac{6}{|\det(\hat{\mathrm{J}})|}(1+m^{-1})bL_{max}\le C\big(\frac{1}{a}\big),\\
&\big\{|\mathrm{J}^{-1}_{(21)}|,|\mathrm{J}^{-1}_{(22)}|,|\mathrm{J}^{-1}_{(23)}|\big\}\le\frac{6}{|\det(\hat{\mathrm{J}})|}(1+m^{-1})aL_{max}\le C\big(\frac{1}{b}\big),\\
&\big\{|\mathrm{J}^{-1}_{(31)}|,|\mathrm{J}^{-1}_{(32)}|,|\mathrm{J}^{-1}_{(33)}|\big\}\le\frac{2}{|\det(\hat{\mathrm{J}})|}(1+m^{-1})^2ab\le C\big(\frac{1}{L_{max}}\big).
\end{aligned}
\tag{42}
$$

Using (42) for (41) yields

$$
|u-\pi_{\mathcal{P}}u|^2_{1,2,\mathcal{P}}\le C\int_{\hat{\mathcal{P}}}\Big(\frac{1}{a^2}\Big|\frac{\partial}{\partial\hat{x}}(\hat{u}-\hat{\pi}_{\hat{\mathcal{P}}}\hat{u})\Big|^2+\frac{1}{b^2}\Big|\frac{\partial}{\partial\hat{y}}(\hat{u}-\hat{\pi}_{\hat{\mathcal{P}}}\hat{u})\Big|^2+\frac{1}{L^2_{max}}\Big|\frac{\partial}{\partial\hat{z}}(\hat{u}-\hat{\pi}_{\hat{\mathcal{P}}}\hat{u})\Big|^2\Big)d\hat{X},
$$

and by (38), (39), and (40) we deduce

$$
|u-\pi_{\mathcal{P}}u|^2_{1,2,\mathcal{P}}\le C\big\{\big(h(\mathcal{P})\big)^2|u|^2_{2,2,\mathcal{P}}+\big(h(\mathcal{P})\big)^4|u|^2_{3,2,\mathcal{P}}\big\},
$$

which implies (35). □

5. Conclusion

In this paper, we proposed the combination of the edge and tetrahedra ratio conditions with the maximum angle condition in three dimensional space, as the natural version of semiregularity for possibly degenerating families of prismatic elements. We have shown that the new semiregularity condition property guarantees that an optimal order of interpolation error is preserved.

In future work, we plan to estimate interpolation errors for pyramidal elements under similar conditions.

References

[1] O. Axelsson, V.A. Barker, Finite Element Solution of Boundary Value Problems: Theory and Computation, Academic Press, New York, 1984.
[2] I. Babuška, A.K. Aziz, On the angle condition in the finite element method, SIAM J. Numer. Anal. 13 (1976) 214–226.
[3] I. Babuška, R. Tempone, G. Zouraris, Galerkin finite element approximations of stochastic elliptic partial differential equations, SIAM J. Numer. Anal. 42 (2004) 800–825.
[4] P.G. Ciarlet, The Finite Element Method for Elliptic Problems, North-Holland, Amsterdam, 1978.
[5] F. Eriksson, The law of sines for tetrahedra and *n*-simplices, Geom. Dedic. 7 (1979) 71–80.
[6] A. Khademi, S. Korotov, J.E. Vatne, On interpolation error on degenerating prismatic elements, Appl. Math. 63 (2018) 237–258.
[7] A. Khademi, S. Korotov, J.E. Vatne, On equivalence of maximum angle conditions for tetrahedral finite element meshes, in: V. Garanzha, L. Kamenski, H. Si (Eds.), Proceedings of the 9th International Conference on Numerical Geometry, Grid Generation and Scientific Computing, NUMGRID 2018, Moscow, Russia, in: Lecture Notes in Computional Science and Engineering, vol. 131, Springer, 2019, pp. 101–108.
[8] A. Khademi, S. Korotov, J.E. Vatne, On the generalization of the Synge-Křížek maximum angle condition for *d*-simplices, J. Comput. Appl. Math. 358 (2019) 29–33.

[9] M. Křížek, On the maximum angle condition for linear tetrahedral elements, SIAM J. Numer. Anal. 29 (1992) 513–520.
[10] A.T.T. Mcrae, G.-T. Bercea, L. Mitchell, D.A. Ham, C.J. Cotter, Automated generation and symbolic manipulation of tensor product finite elements, SIAM J. Sci. Comput. 38 (2016).
[11] P. Šolín, K. Segeth, I. Doležel, Higher-Order Finite Element Methods, Chapman & Hall/CRC, New York, 2004.
[12] J.L. Synge, The Hypercircle in Mathematical Physics: A Method for the Approximate Solution of Boundary Value Problems, Cambridge University Press, New York, 1957.
[13] V. Thomée, Galerkin Finite Element Methods for Parabolic Problems, Springer, New York, 1997.
[14] Y. Yan, Galerkin finite element methods for stochastic parabolic partial differential equations, SIAM J. Numer. Anal. 43 (4) (2005) 1363–1384.
[15] M. Zlámal, On the finite element method, Numer. Math. 12 (1968) 394–409.

mathematics

MDPI

Article

Visual Cryptography Scheme with Essential Participants

Peng Li *, Liping Yin and Jianfeng Ma

Department of Mathematics and Physics, North China Electric Power University, Baoding 071003, China; yinlipingytlx@163.com (L.Y.); jianfma@163.com (J.M.)

* Correspondence: peng.li@ncepu.edu.cn

Received: 23 April 2020; Accepted: 18 May 2020; Published: 22 May 2020

Abstract: Visual cryptography scheme (VCS) shares a binary secret image into multiple shadows printed on transparencies. Stacking shadows can visually decode the secret image without computational resources. Specifically, a (k, n) threshold VCS ((k, n)-VCS) shares a secret image into n shadows, stacking any k shadows can reveal the secret image by human visual system, while any less than k shadows cannot decode any information regarding the secret image. In practice, some participants (essentials) play more important roles than others (non-essentials). In this paper, we propose a (t, s, k, n) VCS with essential participants (so called (t, s, k, n)-EVCS). The secret image is shared into n shadows with s essentials and n-s non-essentials. Any k shadows, including at least t essentials, can reveal the secret image. The proposed scheme is constructed from a monotonic (K, N)-VCS. The condition and optimal choice of (K, N)-VCS to construct (t, s, k, n)-EVCS are given by solving integer programming model. The experimental results are conducted to verify the feasibility of our scheme.

Keywords: visual secret sharing; secret image sharing; visual cryptography; integer programming; essential shadows

1. Introduction

Visual cryptography scheme (VCS) is a technique for sharing a secret image among the participants. The revealing process of the secret image can be implemented by stacking operation without computation. The first VCS was proposed by Naor and Shamir [1] in 1994. A (k, n) threshold VCS ((k, n)-VCS) shares a binary secret image into n shadows printed on transparencies, which are assigned to n participants, respectively. Stacking any k shadows can reveal the secret image by human visual system without computation. The advantage of VCS is its easy revealing. Stacking shadows without computational resources can reveal the secret image. However, the disadvantages are the large shadow size expansion and the degraded visual quality of the revealed image. Many researchers were dedicated on improving performance of VCS [2–4], and proposed VCS with different properties, like VCS for color images [5–7], VCS for multiple secret images [8,9], VCS with meaningful shadows [10,11], and random grid-based VCS (RGVCS) [12–15], et.al.

A (k, n)-VCS is called the monotonic VCS if it can reveal the secret image by stacking more than k shadows. Otherwise, it is called the non-monotonic VCS. Jin et al. proposed progressive VCS [16]. Stacking more shadows can decode secret image with better visual quality. Most of existing VCSs do not distinguish the roles of each shadow. However, in practice, some shadows are more important than others according to the status of the participants. Arumugam et al. [17] proposed (k, n)-VCS with one essential participant and n-1 non-essential participants. In the revealing process, any k participants, including the essential one, can reveal the secret image. Without the essential one, the secret image cannot be revealed, even with all other non-essentials. Guo et al. [18] extended the scheme

of Arumugam et al. [17] and proposed a (t, k, n)-VCS with t essential participants, namely (t, k, n)-EVCS that is constructed from a known (k-t, n-t)-VCS and a known optimal (t, t)-VCS. A qualified set of shadows should contain k shadows, including t essential ones.

Another category of secret image sharing scheme is based on polynomial. Thien and Lin [19] proposed a (k, n) threshold secret image sharing (SIS) scheme. The secret pixels are embedded into the coefficients of a (k−1)-degree polynomial to generate shadow pixels. With any k shadows, the secret image can be decoded by Lagrange interpolation. When compared with VCS, polynomial based SIS can reveal the original grayscale secret image by computation. Many researchers proposed many polynomial based SIS [20–23]. Li et al. [24] first presented the concept of secret image sharing scheme with essential participants (ESIS), and proposed (t, s, k, n)-ESIS. For a (t, s, k, n)-ESIS, the secret image is shared into n shadows with s essentials and n-s non-essentials. A qualified set of shadows should contain k shadows, including at least t essentials. Many researchers proposed different (t, s, k, n)-ESIS schemes to achieve smaller shadow size and equal shadow size [25–28]. Liu et al. [29] combined the scalable secret image sharing scheme with ESIS, so that k or more shadows, including at least t essential shadows, can gradually restore the secret image, while restoring the whole secret image requires the participation of all s essential shadows.

In this paper, we propose general (t, s, k, n) visual cryptography scheme with essential participants (EVCS). For a (t, s, k, n)-EVCS, the secret image is shared into n shadows with s essentials and n-s non-essentials. Stacking any k shadows, including at least t essentials, can reveal the secret image. The proposed (t, s, k, n)-EVCS is constructed from a monotonic (K, N)-VCS based on integer programming. When compared with ESIS, the revealing process of the proposed EVCS does not need complicated mathematical operation. EVCS has potential application when some participants are accorded special privileges due to their status or importance, e.g., heads of government, managers of company, high-level corporate officers, major employers, etc. For example, in a nuclear-powered submarine under the ocean, the missile launch code is shared by (2, 2, 4, 6)-EVCS into two essential shadows for the commander and the executive commander and four non-essentials for four other decision members. The missile launch code can be decoded if and only if at least four participants, including the commander and the executive commander, have the agreement on the launch of the missile, and stacking their shadows. EVCS can also be applied in key exchange or key distribution when exchanging message in a public secure network [30,31].

The layout of this paper is as follows. In next section, we present some preliminaries of VCS. In Section 3, we propose our (t, s, k, n)-EVCS based on integer programming. Section 4 provides experimental results and comparisons and Section 5 concludes the paper.

2. Related Works

In this section, we briefly introduce relevant concepts of (k, n)-VCS and Yan et al.'s random grid based VCS (RGVCS) [32].

2.1. Access Structure of (k, n)-VCS

Let $P = \{1, 2, \cdots, n\}$ be the set of all participants and 2^P is the power set of P. Let qualified sets Γ_{Qual} be the collection of the set of participants that can recover the secret, forbidden sets Γ_{Forb} be the collection of the set of participants that cannot recover the secret. $(\Gamma_{Qual}, \Gamma_{Forb})$ constitutes an access structure, where $\Gamma_{Qual} \subseteq 2^P$, $\Gamma_{Forb} \subseteq 2^P$, and $\Gamma_{Qual} \cap \Gamma_{Forb} = \varnothing$.

Definition 1 ([33]). *A (k, n)-VCS with access structure $(\Gamma_{Qual}, \Gamma_{Forb})$ is monotonic if the following conditions are satisfied.*

(1) *Γ_{Qual} is monotonic increasing, i.e. if a subset of Q can reveal the secret, then the participants in Q can reveal the secret as well.*

(2) *Γ_{Forb} is monotonic decreasing, i.e. if $F \in \Gamma_{Forb}$ cannot reveal the secret, then any subset of F cannot reveal the secret as well.*

(3) $\Gamma_{Qual} \cup \Gamma_{Forb} = 2^p$

For a (k, n)-VCS, a qualified set should contain at least k participants. Stacking any k shadows can reveal the secret image. If stacking more than k shadows can still reveal the secret, the (k, n)-VCS is also called monotonic (k, n)-VCS.

Usually, a (k, n)-VCS is constructed by a pair of matrices, called basis matrices M_0 and M_1. Let $S \subseteq P$, $M|S$ is a submatrix generated by restricting matrix M on the rows of S. Let $OR(M)$ denote the vector generated by performing OR operation on the rows of matrix M. Let $w(a)$ denote the Hamming weight of vector a. Formally, we have the definition of monotonic (k, n)-VCS, as follows.

Definition 2. *Two binary matrices M_0 and M_1 with the size $n \times m$ can be used as basis matrices of a monotonic (k, n)-VCS if and only if the following conditions satisfied.*

(Contrast condition). For any $S \subseteq P$ and $|S| \geq k$, we have $w(OR(M_0|S)) < w(OR(M_1|S))$.

(Security condition). For any $S \subseteq P$ and $1 \leq |S| < k$, we have $w(OR(M_0|S)) = w(OR(M_1|S))$.

For a (k, n)-VCS with basis matrices M_0 and M_1, if the secret pixel is white (resp. black), permute the columns of M_0 (resp. M_1), and then assign its n rows to n shadows, respectively. Since each shadow receives m pixels for sharing each secret pixel, the shadows size is m times of the secret image. m is also called size expansion. The visual quality of the revealed image is usually degraded, and it is evaluated by the contrast defined, as follows.

$$\alpha = (w(OR(M_1\big|S)) - w(OR(M_0\big|S)))/m$$

where $S \subseteq P$ and $|S| \geq k$. By contrast condition of the definition of VCS, we know that α is larger than 0 and no more than 1. When the contrast is 1, the revealed image has the perfect visual quality. The larger value of the contrast, the better visual quality of the revealed image.

Example 1. *The example of (3, 4)-VCS.*

Here we show a (3, 4)-VCS while using the following basis matrices presented in [33].

$$M_0 = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix} \text{ and } M_1 = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The size expansion m is 6, which means the generated shadows have the size six times of the secret image, as we can see from the basis matrices. The contrast value α when stacking three shadows is 1/6. When stacking four shadows, the contrast value α is increased to 1/3. Therefore, the (3, 4)-VCS with above basis matrices is monotonic. Figure 1 shows the experimental results of (3, 4)-VCS. Stacking any three or four shadows can reveal the secret image.

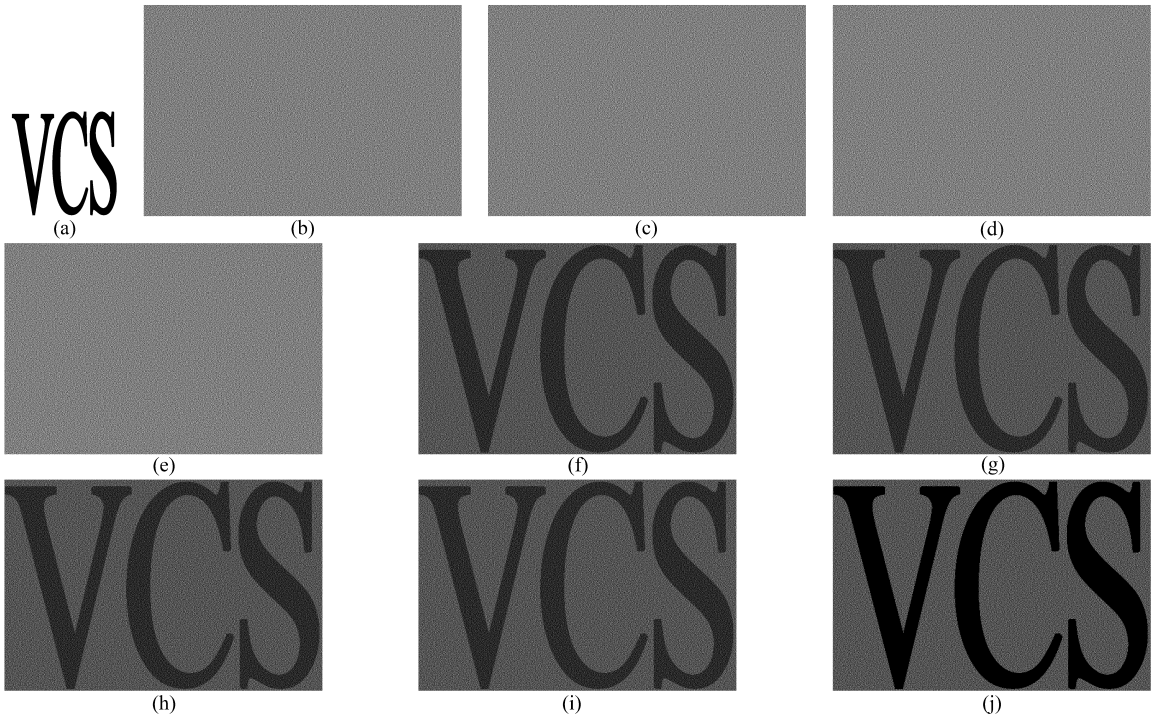


Figure 1. The experimental results of (3, 4)-VCS. (**a**) the secret image; (**b**–**e**) four generated shadows; (**f**) revealed image by shadow 1, 2 and 3; (**g**) revealed image by shadow 1, 2, and 4; (**h**) revealed image by shadow 1, 3, and 4; (**i**) revealed image by shadow 2, 3, and 4; and, (**j**) revealed image by shadow 1, 2, 3, and 4.

2.2. Yan et al.'s RGVCS

Kafri and Keren first presented RG-based VCS [34]. Each shadow is noise-like and it has the same size as the secret image. The revealing operation is also stacking shadows. First, we briefly introduce the generation of (2, 2)-RGVCS, as described below.

Step 1: Randomly generate the first shadow SC_1 with the same size as secret image S.

Step 2: Calculate the corresponding $SC_2(i, j)$ according to $S(i, j)$(the value of pixels in the secret image), as described in Equation (1).

$$SC_2(i,j) = \begin{cases} SC_1(i,j) & if\ S(i,j) = 0 \\ \overline{SC_1(i,j)} & if\ S(i,j) = 1 \end{cases} \tag{1}$$

Step 3: Repeat Step2 until all pixels in S are processed.

Finally, the revealed image obtained by stacking shadows ($S' = SC_1 \otimes SC_2$ as in Equation (2), where $\otimes$ denotes the Boolean OR operation) SC_1 and SC_2 can be directly recognized by the human visual system.

$$S'(i,j) = SC_1(i,j) \otimes SC_2(i,j) = \begin{cases} SC_1(i,j) \otimes SC_1(i,j) & if\ S(i,j) = 0 \\ SC_1(i,j) \otimes \overline{SC_1(i,j)} = 1 & if\ S(i,j) = 1 \end{cases} \tag{2}$$

Many (k, n)-RGVCS schemes have been proposed based on (2, 2)-RGVCS. Their similarity is to repeat the above process for the first k bits, but the difference is the disposal of the last n-k bits. Yan et al. [32] proposed a novel (k, n)-RGVCS, which makes full use of n-k random bits to improve the visual quality of the recovered image. Their (k, n)-RGVCS is also a progressive VCS. Better visual quality of the revealed secret image will be gained by stacking more shadows. The algorithm of Yan et al.'s (k, n)-RGVCS is given, as follows (Algorithm 1).

Algorithm 1. Yan et al.'s RGVCS.

Input: secret image S, the threshold parameters (k, n)
Output: n shadows $SC_1, SC_2, \ldots, SC_n$

A1-1: For each pixel $S(i, j)$ in the secret image S, repeat Steps 2–4.
A1-2: Apply the above conventional (2, 2)-RGVCS to encrypt the pixel $S(i, j)$, then b_1 and b'_2 are obtained. b'_2 is encrypted in the same way. Repeat the above operation until $b_1, b_2, \ldots, b'_k (= b_k)$ are obtained.
A1-3: For $b_l (k + 1 \le l \le n)$, if $l \bmod k = x, (0 \le x \le k - 1)$, then $b_l = b_k$.
A1-4: Redistribute $b_1, b_2, \ldots, b_n$ to $SC_1(i, j), SC_2(i, j), \ldots, SC_n(i, j)$ randomly.
A1-5: Output n shadows $SC_1, SC_2, \ldots, SC_n$.

Example 2. *The experiment of Yan et al.'s (3, 6)-RGVCS*

An experiment of (3, 6) threshold of scheme with secret image "VCS" is conducted in order to demonstrate the Yan et al.'s algorithm. Figure 2a shows the secret image. Figure 2b–g show six shadows. Figure 2h shows the revealed image by 2 shadows. Figure 2i–l show the revealed image by stacking 3, 4, 5, and 6 shadows, respectively. Apparently, better visual quality of the recovered secret will be gained by stacking more shadow images. The results show that Yan et al.'s scheme satisfies monotonicity, as described in Definition 1.

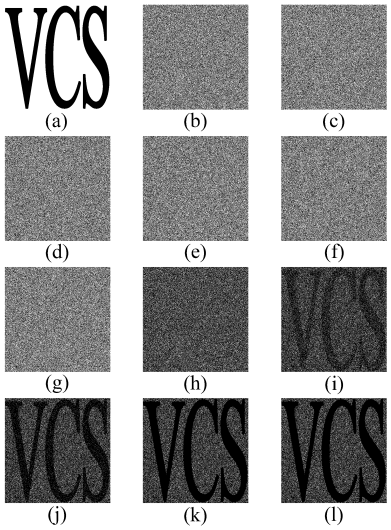


Figure 2. An experiment of Yan et al.'s (3, 6)-VCS. (**a**) secret image; (**b**–**g**) six shadows; (**h**) revealed image by two shadows; (**i**) revealed image by three shadows; (**j**) revealed image by four shadows; (**k**) revealed image by five shadows; and, (**l**) revealed image by six shadows.

3. The Proposed (*t*, *s*, *k*, *n*)-EVCS Based on Integer Programming

3.1. The Definition of (t, s, k, n)-EVCS

In traditional (k, n)-VCS, a qualified subset of participants should have any k or more participants. The roles of each participant are the same. However, there are many examples in practical situations where some participants are given privileges because of their status or importance, such as heads of government, company managers, etc. Therefore, it is reasonable for us to consider giving special powers to some participants in VCS. The proposed (t, s, k, n)-EVCS shares the secret image into n shadows with s essentials and n-s non-essentials. Stacking any k shadows, including at least t essential one, can reveal the secret image. A qualified subset of participants should have at least k shadows, including t essentials. Let $EP = \{1, 2, \ldots, s\}$ and $NEP = \{s + 1, s + 2, \ldots, n\}$ denote the set of essential

participants and non-essential participants, respectively. Subsequently, we can derive all qualified sets Γ_{Qual} of (t, s, k, n)-EVCS, as follows.

$$\Gamma_{Qual} = \{Q | Q \subseteq P, |Q| \geq k \text{ and } |Q \backslash NEP| \geq t\} \tag{3}$$

If a subset of participant does not belong to the qualified sets Γ_{Qual}, it belongs to forbidden sets. Hence, we have forbidden sets Γ_{Forb} of (t, s, k, n)-EVCS.

$$\Gamma_{Forb} = \left\{ S \middle| S \subseteq P \text{ and } S \notin \Gamma_{Qual} \right\} \tag{4}$$

Subsequently, (t, s, k, n)-EVCS can be defined if and only if the access structure satisfies Equations (3) and (4).

We only consider non-trivial EVCS, which cannot be reduced to a threshold VCS. For the relationships among t, s, k, and n of (t, s, k, n)-EVCS, we have the following facts.

(1) t, s, k and n are all integers no less than 1, and $t \leq s \leq n$, $t \leq k \leq n$.
(2) $k > t$. Otherwise, (t, s, k, n)-EVCS is reduced to (t, s)-VCS.
(3) $k < n$. Otherwise, (t, s, k, n)-EVCS is reduced to (n, n)-VCS.
(4) If $s = n$, (t, s, k, n)-EVCS is reduced to (t, n)-VCS. Hence, $s < n$. If $s = n - 1$, then there is only one non-essential participant. A qualified set of participants contains k members, including at least t essentials and k - t non-essentials. We have $k - t \leq 1$. Since $k > t$, then $k = t + 1$. If $s = t$, then $k = s + 1 = n$, and (t, s, k, n)-EVCS is reduced to (n, n)-VCS. Otherwise, $s \geq t + 1$, which means that there are more than t essentials. Since there is only one non-essential participant, any $t + 1$ participants must contain at least t essentials and they can reveal the secret image. Afterwards, (t, s, k, n)-EVCS is reduced to $(t + 1, n)$-VCS. Overall, we have $s \leq n$-2.
(5) $k - t < n - s$. The number of non-essentials is $n - s$, and the largest number of non-essentials in a qualified set is $k - t$. Obviously, $k - t \leq n - s$. If $k - t = n - s$, then any k participants will contain at least $k - (n - s) = k - (k - t) = t$ essentials. Subsequently, (t, s, k, n)-EVCS is reduced to (k, n)-VCS. Therefore, we have $k - t < n - s$.

Finally, we have the relationships among t, s, k and n of (t, s, k, n)-EVCS are shown, as follows:

$$\begin{cases} t \leq s \leq n - 2 \\ t < k < n \\ k - t < n - s \\ t, s, k \text{ and } n \text{ are integers no less than } 1 \end{cases} \tag{5}$$

3.2. Constructing (t, s, k, n)-EVCS Based on Integer Programming

The idea for constructing (t, s, k, n)-EVCS is that we generate the shadows by a monotonic (K, N)-VCS. The secret image is first shared into N shadows by (K, N)-VCS. Subsequently, each essential (non-essential) shadow of EVCS is obtained by the superposition of ω_1 (ω_2) shadows of VCS. Obviously, we have

$$N = s\omega_1 + (n - s)\omega_2 \tag{6}$$

Since essential shadow is more important than the non-essential shadow of EVCS, ω_1 must be larger than ω_2.

Figure 3 shows the diagram of generating shadows of EVCS by the shadows of a monotonic VCS.

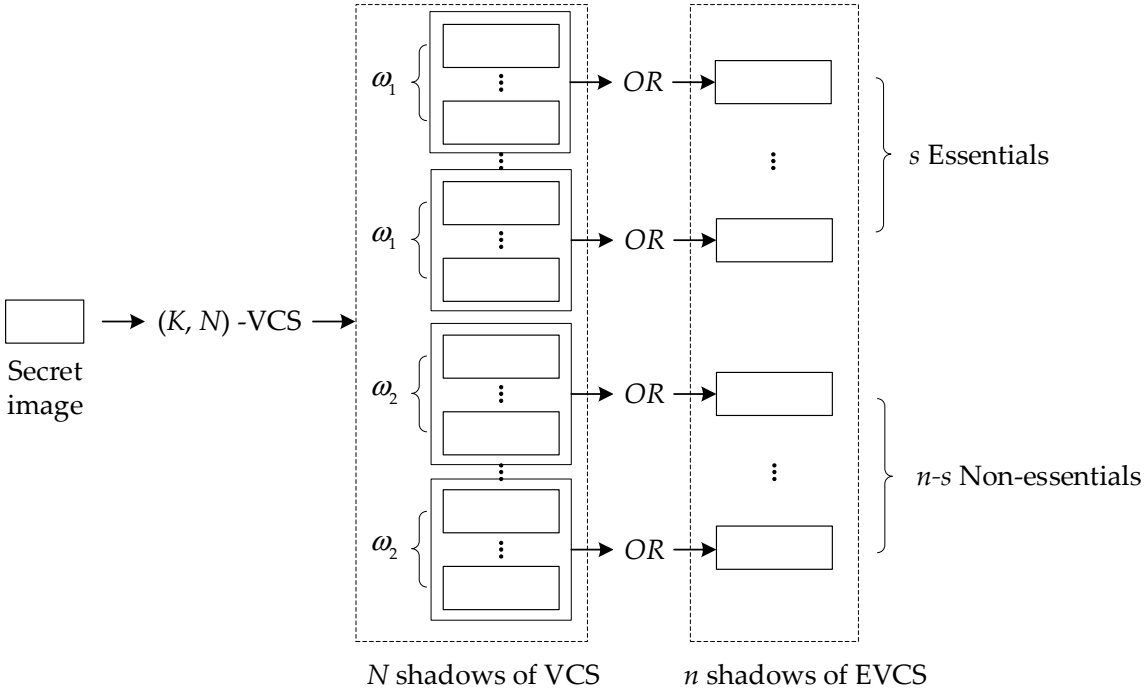


Figure 3. The diagrammatical representation of the proposed (*t*, s, *k*, *n*)-EVCS.

For (*K*, *N*)-VCS, a qualified subset of shadows should have at least *K* shadows. Each essential (non-essential) shadow of EVCS represents the stacking result of ω_1 (ω_2) shadows of (*K*, *N*)-VCS. A qualified subset of shadows of EVCS should contribute no less than *K* shadows of (*K*, *N*)-VCS to satisfy the contrast condition of (*K*, *N*)-VCS. Any forbidden subset of shadows of EVCS should contribute less than *K* shadows of (*K*, *N*)-VCS to satisfy the security condition of (*K*, *N*)-VCS. Therefore, our task is determining the proper values of ω_1, ω_2, *K*, and *N* to be used for constructing (*t*, *s*, *k*, *n*)-EVCS. In this paper, we get the values of ω_1, ω_2, *K*, and *N* by solving an integer programming model.

For (*t*, *s*, *k*, *n*)-EVCS, we first need to build the relationship among the values of ω_1, ω_2, *K*, and *N*. These parameters should satisfy the following conditions.

(1) For all parameters ω_1, ω_2, *K*, and *N* to make sense, we need to restrict that $\omega_1 \geq 1$, $\omega_2 \geq 1$, $K \geq 1$, and $N \geq K$.

(2) Essential shadows are more important than non-essential shadows. In another word, an essential shadow can contribute more shadows of VCS than a non-essential shadow. Hence, ω_1 must be larger than ω_2. That is

$$\omega_1 - \omega_2 \geq 1 \tag{7}$$

(3) By Equation (3), we have that a qualified set should contain any *k* shadows, including at least *t* essentials. In another word, *k* shadows of EVCS, including *t* essential ones, can contribute at least *K* shadows of VCS. Subsequently, we have

$$t\omega_1 + (k-t)\omega_2 \geq K \tag{8}$$

Obviously, Equation (8) guarantees that any *k* shadows of EVCS, including more than *t* essential ones, can also contribute at least *K* shadows of VCS.

(4) By Equations (3) and (4), the secret image cannot be recovered with less than *t* essential shadows. In another word, the threshold value *K* of (*K*, *N*)-VCS is still not satisfied, even if $t-1$ essential shadows and all *n-s* non-essential shadows are gathered. Subsequently, we have the following inequality.

$$(t-1)\omega_1 + (n-s)\omega_2 \leq K-1 \tag{9}$$

(5) By Equations (3) and (4), the secret image cannot be recovered with less than k shadows.

If $s \geq k$, then any $k-1$ essential shadows cannot contribute enough shadows of VCS. In order to satisfy the security condition of VCS, we have

$$(k-1)\omega_1 \leq K-1 \tag{10}$$

If $s < k$, then any s essential shadows and k-$s-1$ non-essential shadows cannot contribute enough shadows of VCS. To satisfy the security condition of VCS, we have

$$s\omega_1 + (k-1-s)\omega_2 \leq K-1 \tag{11}$$

For (K, N)-VCS, the larger values of K and N may reduce the visual quality of the revealed image, and complicate the sharing process. Therefore, we want to obtain as small values of K and N as possible. In general, the objective function is:

$$\min K+N \tag{12}$$

We generate the following integer programming models (IPM) by combining the constraint conditions and objective function.

(IPM I): When $s < k$, the corresponding integer programming model is:

$$\begin{cases} \min K + s\omega_1 + (n-s)\omega_2 \\ t\omega_1 + (k-t)\omega_2 - K \geq 0 \\ (t-1)\omega_1 + (n-s)\omega_2 - K \leq -1 \\ s\omega_1 + (k-1-s)\omega_2 - K \leq -1 \\ \omega_1 - \omega_2 \geq 1 \\ \omega_1 \geq 1 \\ \omega_2 \geq 1 \\ K \geq 1 \end{cases} \tag{13}$$

(IPM II): When $s \geq k$, the corresponding integer programming model is:

$$\begin{cases} \min K + s\omega_1 + (n-s)\omega_2 \\ t\omega_1 + (k-t)\omega_2 - K \geq 0 \\ (t-1)\omega_1 + (n-s)\omega_2 - K \leq -1 \\ (k-1)\omega_1 - K \leq -1 \\ \omega_1 - \omega_2 \geq 1 \\ \omega_1 \geq 1 \\ \omega_2 \geq 1 \\ K \geq 1 \end{cases} \tag{14}$$

3.3. Determine the Parameters by Solving IPMs

We need to solve IPM I or IPM II to determine the values of ω_1, ω_2, K and N. Before we solve IPM, we divide the relationship among t, s and k into six cases: (Case 1) $t = s$, $s < k$; (Case 2) $t \neq s$, $s < k$; (Case 3) $s - t = 1$, $s = k$; (Case 4) $s - t \neq 1$, $s = k$; (Case 5) $k - t = 1$, $s > k$; (Case 6) $k - t = 1$, $s > k$. Figure 4 shows the diagram of the division for different cases. For Case 1 and Case 2, we need to solve IPM I. For the other cases, we need to solve IPM II.

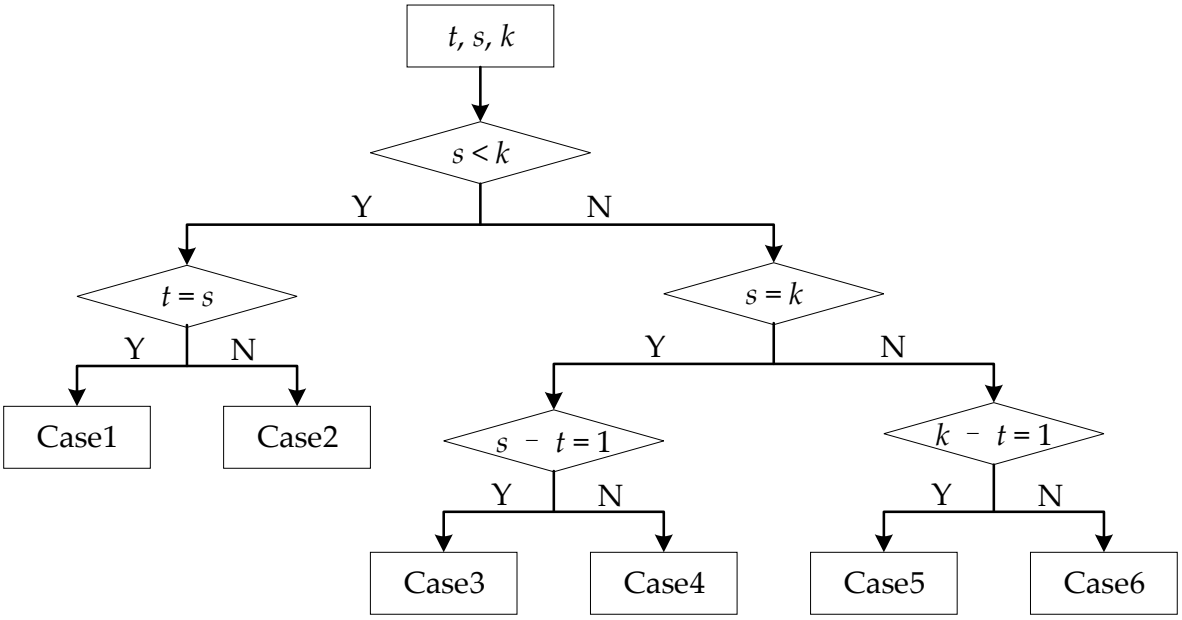


Figure 4. The diagram of the division for different cases.

Now we solve IPM according to the six cases, respectively.

(Case 1) t = s, s < k.

For this case, we need to solve IPM I. First, we convert IPM I into standard form by generalizing the sighs of ω_1, ω_2 and K to x_1, x_2, and x_3. Subsequently, we have new IPM as follows.

$$\begin{cases} \max -tx_1 + (t-n)x_2 - x_3 \\ tx_1 + (k-t)x_2 - x_3 - x_4 = 0 \\ (1-t)x_1 + (t-n)x_2 + x_3 - x_5 = 1 \\ -tx_1 + (t+1-k)x_2 + x_3 - x_6 = 1 \\ x_1 - x_2 - x_7 = 1 \\ x_1 - x_8 = 1 \\ x_2 - x_9 = 1 \\ x_3 - x_{10} = 1 \\ x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10} \geq 0 \end{cases} \tag{15}$$

where x_4, x_5, x_6, x_7, x_8, x_9, and x_{10} are non-negative residual variables (slack variables). We use the dual simplex method to solve above IPM. Equation (15) is converted to the following form to obtain the initial feasible basis of the dual problem.

$$\begin{cases} \max -tx_1 + (t-n)x_2 - x_3 \\ -tx_1 + (t-k)x_2 + x_3 + x_4 = 0 \\ (t-1)x_1 + (n-t)x_2 - x_3 + x_5 = -1 \\ tx_1 + (k-t-1)x_2 - x_3 + x_6 = -1 \\ -x_1 + x_2 + x_7 = -1 \\ -x_1 + x_8 = -1 \\ -x_2 + x_9 = -1 \\ -x_3 + x_{10} = -1 \\ x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9 x_{10} \geq 0 \end{cases} \tag{16}$$

Establish the initial simplex table for IPM, as shown in Table 1.

Table 1. The initial simplex table of integer programming models (IPM) I for Case 1.

	$c_j \rightarrow$		-t	$t-n$	-1	0	0	0	0	0	0	0
C_B	X_B	b	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
0	x_4	0	$-t$	$t-k$	1	1	0	0	0	0	0	0
0	x_5	-1	$t-1$	$n-t$	[[-1]	0	1	0	0	0	0	0
0	x_6	-1	t	$k-t-1$	-1	0	0	1	0	0	0	0
0	x_7	-1	-1	1	0	0	0	0	1	0	0	0
0	x_8	-1	-1	0	0	0	0	0	0	1	0	0
0	x_9	-1	0	-1	0	0	0	0	0	0	1	0
0	x_{10}	-1	0	0	-1	0	0	0	0	0	0	1
	$c_j - z_j$		$-t$	$t-n$	-1	0	0	0	0	0	0	0

From Table 1 it can be seen that the solution of the dual problem corresponding to the row of checking number is feasible. Since some numbers in column b is negative, iterative operation is required. Since the values of b_i are equal, the non-basic variable with the smallest subscript in $\boldsymbol{X_B}$ is selected as the leaving variable, i.e. x_5. Check the coefficients $a_{lj}(j = 1, 2, \ldots, 10)$ of the row of a_l in the simplex table, if all $a_{lj} \geq 0$, there is no feasible solution, and the calculation is terminated. If $a_{lj} < 0$, calculate $\theta = \min_j\left(\frac{c_j - z_j}{a_{lj}} \middle| a_{lj} < 0\right) = \frac{c_k - z_k}{a_{lk}}$, and the non-basic variable x_k of the column corresponding to rule of θ is the entering variable. Calculating according to the above steps, we obtain $\theta = \min\left\{-, -, \frac{-1}{-1}\right\} = 1$, so x_3 is the entering variable. "-1" is the pivot element at the intersection of the column and row where the variables are entering and leaving. The iteration is performed according to the calculation steps of dual simplex method, and Table 2 shows the results.

Table 2. Simplex table of IPM I for Case 1 after one iteration.

	$c_j \rightarrow$		$-t$	$t-n$	-1	0	0	0	0	0	0	0
$\boldsymbol{C_B}$	$\boldsymbol{X_B}$	$\boldsymbol{b}$	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
0	x_4	-1	[-1]	$n-k$	0	1	1	0	0	0	0	0
-1	x_3	1	$1-t$	$t-n$	1	0	-1	0	0	0	0	0
0	x_6	0	1	$k-n-1$	0	0	-1	1	0	0	0	0
0	x_7	-1	-1	1	0	0	0	0	1	0	0	0
0	x_8	-1	-1	0	0	0	0	0	0	1	0	0
0	x_9	-1	0	-1	0	0	0	0	0	0	1	0
0	x_{10}	0	$1-t$	$t-n$	0	0	-1	0	0	0	0	1
	$c_j - z_j$		$1-2t$	$2(t-n)$	0	0	-1	0	0	0	0	0

From Table 2 it can be seen that the dual problem is still a feasible solution, and there are still negative components in column $\boldsymbol{b}$. Repeat the above iterative steps until the numbers in column $\boldsymbol{b}$ are all non-negative and the test numbers are all non-positive, as shown in Table 3.

Table 3. The final simplex table of IPM I for Case 1.

	$c_j \rightarrow$		$-t$	$T-n$	-1	0	0	0	0	0	0	0
$\boldsymbol{C_B}$	$\boldsymbol{X_B}$	$\boldsymbol{b}$	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
$-t$	x_1	$n-k+1$	1	0	0	$k-n-1$	-1	$k-n$	0	0	0	0
-1	x_3	$t(n-k)+k$	0	0	1	$t(k-n)-k+1$	$-t$	$t(k-n+1)-k$	0	0	0	0
$t-n$	x_2	1	0	1	0	-1	0	-1	0	0	0	0
0	x_7	$n-k-1$	0	0	0	$k-n$	-1	$k-n+1$	1	0	0	0
0	x_8	$n-k$	0	0	0	$k-n-1$	-1	$k-n$	0	1	0	0
0	x_9	0	0	0	0	-1	0	-1	0	0	1	0
0	x_{10}	$k+t(n-k)-1$	0	0	0	$t(k-n)-k+1$	$-t$	$t(k-n+1)-k$	0	0	0	1
	$c_j - z_j$		0	0	0	$2t(k-n)+1-k-n$	$-2t$	$2t(k-n+1)-n-k$	0	0	0	0

The numbers in column $\boldsymbol{b}$ are all non-negative and the test numbers are all non-positive, as shown in Table 3. Therefore, the optimal solution of the problem is $X^* = (n-k+1, 1, t(n-k)+k, 0, 0, 0, n-k-1,$

$n-k$,0,$k+t(n-k)-1$). Additionally, since $tx_1+(n-t)x_2=N$, then $t(n-k+1)+n-t=N$. From what has been discussed above, any (t, t, k, n)-EVCS can be constructed by a monotonic (K, N)-VCS = ($t(n-k)+k$, $t(n-k+1)+n-t$)-VCS with $\omega_1=n-k+1$ and $\omega_2=1$.

(Case 2) t ≠ s, s < k.

For this case, we need to solve IPM I with the same method. Table 4 shows the simplex table obtained after two iterations.

Table 4. Simplex table after two iterations for Case 2.

	$c_j\rightarrow$		−s	$s-n$	−1	0	0	0	0	0	0	0
C_B	X_B	b	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
−s	x_1	1	1	$k+s-t-n$	0	−1	−1	0	0	0	0	0
−1	x_3	t	0	$t-k-t(t+n-k-s)$	1	$1-t$	$-t$	0	0	0	0	0
0	x_6	$t-s-1$	0	$[(t+n-k-s)(s-t)+t-s-1]$	0	$S-t+1$	$s-t$	1	0	0	0	0
0	x_7	0	0	$1-t-n+k+s$	0	−1	−1	0	1	0	0	0
0	x_8	0	0	$k+s-t-n$	0	−1	−1	0	0	1	0	0
0	x_9	−1	0	−1	0	0	0	0	0	0	1	0
0	x_{10}	$t-1$	0	$t-k-t(t+n-k-s)$	0	$1-t$	$-t$	0	0	0	0	1
	c_j-z_j		0	$2(s-n)+(s-t-1)(k+s-t-n)$	0	$1-s-t$	$-s-t$	0	0	0	0	0

Since $t\neq s$, with the relationship between s and t, we have $t-s-1<-1$. Subsequently, the linear programming has a solution if and only if $(t+n-k-s)(s-t)+t-s-1<0$. Continue to iterate. It is calculated that the elements in column $\boldsymbol{b}$ are $1-\frac{(t-s-1)(k+s-t-n)}{(t+n-k-s)(s-t)+t-s-1}$, $t-\frac{(t-s-1)[t-k-t(t+n-k-s)]}{(t+n-k-s)(s-t)+t-s-1}$, $\frac{t-s-1}{(t+n-k-s)(s-t)+t-s-1}$, $-\frac{(t-s-1)(1-t-n+k+s)}{(t+n-k-s)(s-t)+t-s-1}$, $-\frac{(t-s-1)(k+s-t-n)}{(t+n-k-s)(s-t)+t-s-1}$, $\frac{t-s-1}{(t+n-k-s)(s-t)+t-s-1}-1$, $t-1-\frac{(t-s-1)[t-k-t(t+n-k-s)]}{(t+n-k-s)(s-t)+t-s-1}$, respectively. They are obviously all non-negative, except $-\frac{(t-s-1)(1-t-n+k+s)}{(t+n-k-s)(s-t)+t-s-1}$, which needs further discussion. If $1-t-n+k+s\leq 0$, i.e. $-\frac{(t-s-1)(1-t-n+k+s)}{(t+n-k-s)(s-t)+t-s-1}\geq 0$, then the calculation is terminated. It can be known from conditions $(t+n-k-s)(s-t)+t-s-1<0$ and $1-t-n+k+s\leq 0$ that $1\leq t+n-k-s<2$, namely, $t+n-k-s=1$. Thus, the above simplex table can be simplified into the following table.

The elements in column $\boldsymbol{b}$ are all non-negative and the checking numbers are all non-positive, as shown in Table 5. Therefore, the optimal solution of the problem is $X^*=(s-t+2,\ s-t+1,\ t-k(t-s-1)$,0,0,0,0, $s-t+1$, and $t-k(t-s-1)-1$, $s-t)$. From what has been discussed above, (t, s, k, n)-EVCS of Case 2 can be constructed by a monotonic (K, N)-VCS = ($t-k(t-s-1)$, $s(s-t+2)+(n-s)(s-t+1)$)-VCS with ω_1=$s-t+2$ and ω_2=$s-t+1$. If $1-t-n+k+s>0$, we have known that $k+s<n+t$, then $0<t+n-k-s<1$, the absence of t, s, k, and n makes this condition satisfy.

Table 5. The final simplex table of IPM I for Case 2.

		$c_j\rightarrow$	$-t$	$t-n$	−1	0	0	0	0	0	0	0
C_B	X_B	b	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
−s	x_1	$S-t+2$	1	0	0	$t-s-2$	$t-s-1$	−1	0	0	0	0
−1	x_3	$T-k(t-s-1)$	0	0	1	$1-t-k(s-t+1)$	$-t-k(s-t)$	$-k$	0	0	0	0
$s-n$	x_2	$1+s-t$	0	1	0	$t-s-1$	$t-s$	−1	0	0	0	0
0	x_7	0	0	0	0	−1	−1	0	1	0	0	0
0	x_8	$1+s-t$	0	0	0	$t-s-2$	$t-s-1$	−1	0	1	0	0
0	x_9	$t-k(t-s-1)-1$	0	0	0	$t-s-1$	$t-s$	−1	0	0	1	0
0	x_{10}	$s-t$	0	0	0	$1-t-k(s-t+1)$	$-t-k(s-t)$	$-k$	0	0	0	1
		c_j-z_j	0	0	0	$(t-s-1)(k+n)-s-t+1$	$(t-s)(k+n)-s-t$	$-k-n$	0	0	0	0

(Case 3) s − t = 1, s = k.

For this case, we need to solve IPM II with the same method. Table 6 shows the simplex table obtained after three iterations.

Table 6. Simplex table of IPM II for Case 3 after three iterations.

C_B	$c_j \rightarrow$ X_B	b	$-s$ x_1	$s-n$ x_2	-1 x_3	0 x_4	0 x_5	0 x_6	0 x_7	0 x_8	0 x_9	0 x_{10}
$-s$	x_1	$n-s$	1	0	0	$s-n$	-1	$s-n+1$	0	0	0	0
-1	x_3	$t(n-s)+1$	0	0	1	$t(s-n)$	$-t$	$-1-t(n-s-1)$	0	0	0	0
$s-n$	x_2	1	0	1	0	-1	0	-1	0	0	0	0
0	x_7	$n-s-2$	0	0	0	$s-n+1$	[−1]	$s-n+2$	1	0	0	0
0	x_8	$n-s-1$	0	0	0	$s-n$	-1	$s-n+1$	0	1	0	0
0	x_9	0	0	0	0	-1	0	-1	0	0	1	0
0	x_{10}	$t(n-s)$	0	0	0	$t(s-n)$	$-t$	$-1-t(n-s-1)$	0	0	0	1
	$c_j - z_j$		0	0	0	$(s+t+1)(s-n)$	$-s-t$	$s(s-n+2)-t(n-s-1)-1-n$	0	0	0	0

If $n-s \geq 2$, i.e. $n-s-2 \geq 0$, then the calculation is terminated and we have (K, N)-VCS = $(t(n-s)+1, s(n-s)+n-s)$-VCS with $\omega_1 = n-s$ and $\omega_2 = 1$. Otherwise, $n-s-2 < 0$ does not satisfy the conditions that are given in Equation (5).

Similarly, for Case 4, Case 5, and Case 6, the same analysis method is used to solve the corresponding IPM. Finally, we have the solutions of the corresponding IPM with different conditions. as shown in Table 7.

Table 7. The solutions of IPM for all cases.

Conditions			Solution (ω_1, ω_2,K)	Case
$s < k$	$t = s$	$V' \geq 0$ *	$(n-k+1, 1, t(n-k)+k)$	Case1
	$t \neq s$	$V' < 0$	$-$ $(s-t+2, s-t+1, t-k(t-s-1))$	Case2
$s = k$	$s-t=1$	$V'' \geq 0$ *	$(n-s, 1, t(n-s)+1)$	Case3
	$s-t \neq 1$	$V'' < 0$	$-$ $(s-t+1, s-t, t-s(t-s))$	Case4
$s > k$	$k-t=1$	$V''' \geq 0$ *	$(n-s, 1, t(n-s)+1)$	Case5
	$k-t \neq 1$	$V''' < 0$	$-$ $(k-t+1, k-t, t-k(t-k))$	Case6

* Where $V' = (t+n-k-s)(s-t)+t-s-1$, $V'' = (t+n-k-s)(s-t)+s-n$, $V''' = (t+n-k-s)(k-t)+s-n$.

From Table 7, for the most cases, we can find the solutions (ω_1, ω_2, K) of the corresponding IPM. Since N can be calculated by Equation (6), we can construct (t, s, k, n)-EVCS by the corresponding monotonic (K, N)-VCS. For some common cases of (t, s, k, n)-EVCS, we list the solutions of the corresponding IPM, the values of ω_1, ω_2, K, and N in Table 8.

Table 8. The solutions of IPM for some specific essential participants (EVCSs).

(t, s, k, n)-EVCS	(K, N)-VCS	ω_1	ω_2	Case
(1, 1, 2, 3)-EVCS	(3, 4)-VCS	2	1	1
(1, 1, 2, 4)-EVCS	(4, 6)-VCS	3	1	1
(1, 1, 2, 5)-EVCS	(5, 8)-VCS	4	1	1
(1, 1, 3, 5)-EVCS	(5, 7)-VCS	3	1	1
(1, 1, 3, 6)-EVCS	(6, 9)-VCS	4	1	1
(1, 2, 2, 4)-EVCS	(3, 6)-VCS	2	1	3
(1, 2, 2, 5)-EVCS	(4, 9)-VCS	3	1	3
(1, 2, 2, 6)-EVCS	(5, 12)-VCS	4	1	3
(1, 2, 3, 5)-EVCS	(7, 12)-VCS	3	2	2
(1, 2, 3, 6)-EVCS	–	–	–	2
(1, 3, 2, 5)-EVCS	(3, 8)-VCS	2	1	5
(1, 3, 2, 6)-EVCS	(4, 12)-VCS	3	1	5
(1, 3, 3, 6)-EVCS	(7, 15)-VCS	3	2	4
(1, 3, 3, 7)-EVCS	–	–	–	4
(1, 4, 3, 7)-EVCS	(7, 18)-VCS	3	2	6
(1, 4, 3, 8)-EVCS	–	–	–	6
(2, 2, 3, 4)-EVCS	(5, 6)-VCS	2	1	1
(2, 2, 3, 5)-EVCS	(7, 9)-VCS	3	1	1
(2, 2, 3, 6)-EVCS	(9, 12)-VCS	4	1	1
(2, 2, 4, 5)-EVCS	(6, 7)-VCS	2	1	1
(2, 2, 4, 6)-EVCS	(8, 10)-VCS	3	1	1
(2, 3, 3, 5)-EVCS	(5, 8)-VCS	2	1	3
(2, 3, 3, 6)-EVCS	(7, 12)-VCS	3	1	3
(2, 3, 4, 6)-EVCS	(10, 15)-VCS	3	2	2
(2, 4, 3, 6)-EVCS	(5, 10)-VCS	2	1	5

4. Experimental results and Comparison

4.1. Experimental Results

In this subsection, we conduct two experiments to verify the feasibility of the proposed scheme.

Example 3. *The experiment of the proposed (1, 1, 2, 3)-EVCS.*

By Table 8, we can generate our (1, 1, 2, 3)-EVCS by a monotonic (3, 4)-VCS with $\omega_1 = 2$, and $\omega_2 = 1$. First, we share the secret image into four shadows by a monotonic (3, 4)-VCS. Subsequently, the first two shadows are used to generate the essential shadow of (1, 1, 2, 3)-EVCS by OR operation. Additionally, the left two shadows are treated as two non-essential shadows of (1, 1, 2, 3)-EVCS, respectively. Finally, we have three shadows of (1, 1, 2, 3)-EVCS with one essential and two non-essentials.

The chosen (3, 4)-VCS used to construct (1, 1, 2, 3)-EVCS can be any monotonic (3, 4)-VCS proposed by researchers. In this example, we choose two monotonic (3, 4)-VCS separately to implement (1, 1, 2, 3)-EVCS. First, we use monotone (3, 4)-VCS in Example 1 to implement (1, 1, 2, 3)-EVCS. We already show the experiment of (3, 4)-VCS in Example 1. Figure 1 shows the four generated shadows of (3, 4)-VCS. Now, we can generate the essential shadow of (1, 1, 2, 3)-EVCS, as shown in Figure 1a, by performing OR operation on the Figure 1b,c. The rest two shadows Figure 1d,e are treated as two non-essential shadows of (1, 1, 2, 3)-EVCS as shown in Figure 5b,c. For (1, 1, 2, 3)-EVCS, the qualified sets are {1,2}, {1,3}, and {1,2,3}. We show the revealed image by different qualified sets of shadows in Figure 5d,e and g. As we can see, the secret image can only be revealed with at least two shadows, including the essential one. Without the essential shadow, stacking two non-essentials cannot reveal the secret image, as shown in Figure 5f. Since the (3, 4)-VCS has the size expansion 6 and contrast loss of the revealed image. Each shadow of (1, 1, 2, 3)-EVCS has the size six times of the secret image. The visual quality of the revealed image is also degraded.

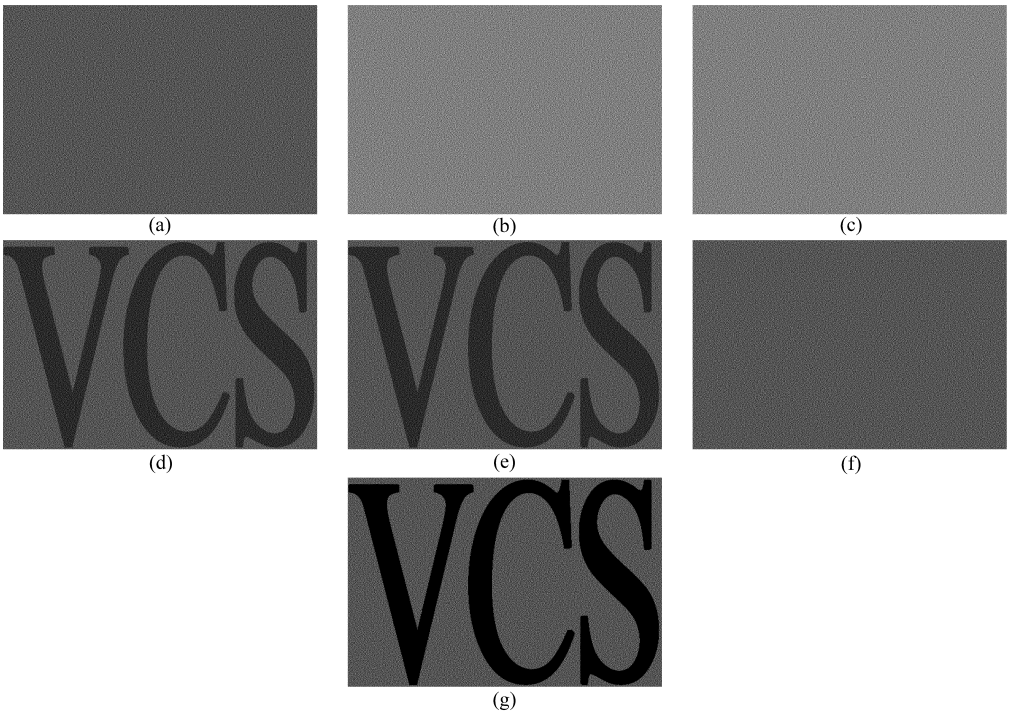


Figure 5. The experimental results of proposed scheme for (1, 1, 2, 3)-EVCS. (**a**) essential shadow; (**b**,**c**) two non-essential shadows; (**d**) revealed image by shadow 1 and 2; (**e**) revealed image by shadow 1 and 3; (**f**) revealed image by shadow 2 and 3; and, (**g**) revealed image by all shadows.

Second, we choose Yan et al.'s (3, 4)-RGVCS to implement (1, 1, 2, 3)-EVCS. First, a secret image (Figure 1a) is shared into four shadows by Yan et al.'s (3, 4)-RGVCS. With the same method, performing OR operation on the first two shadows can obtain the essential shadow of (1, 1, 2, 3)-EVCS. The other two shadows of (3, 4)-RGVCS are treated as two non-essentials, respectively. Since (3, 4)-RGVCS generates shadows without size expansion, all of the shadows of (1, 1, 2, 3)-EVCS have the same size as the secret image. Figure 6a–c show the three generated shadows of (1, 1, 2, 3)-EVCS. In the revealing process, stacking qualified set of shadows can reveal the secret image. Figure 6d–g show the revealed images with different shadows. As we can see, without the essential shadow, we cannot reveal the secret image by the shadows. EVCS based on RGVCS can achieve better performance over that based on traditional VCS since RGVCS has the advantage of no size expansion.

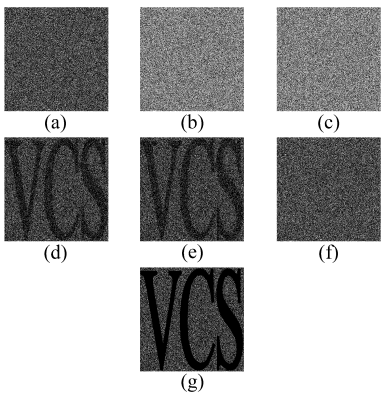


Figure 6. The experiment of the proposed (1, 2, 3)-EVCS. (**a**) essential shadow; (**b**,**c**) two non-essential shadows; (**d**) revealed image by shadow 1 and 2; (**e**) revealed image by shadow 1 and 3; (**f**) revealed image by shadow 2 and 3; (**g**) revealed image by all three shadows.

Example 4. *The experiment of the proposed (1,2,2,4)-EVCS.*

(1, 2, 2, 4)-EVCS can be implemented by (3, 6)-VCS according to the solution of the aforementioned IPM. We still adopt Yan et al. 's (3, 6) -RGVCS to implement (1, 2, 2, 4)-EVCS in order to achieve better performance. Since $s = 2$ and $\omega_1 = 2$, we can get two essential shadows by performing the OR operation twice on any two shadows. It should be noted that the operands of twice OR operations are non-overlapping from each other, i.e. the same shadow of VCS can only participate in OR operation for one time in generating shadow of EVCS and cannot participate in the formation of two shadows of EVCS at the same time. Figure 2 shows the experimental results of Yan et al.'s (3, 6)-RGVCS. Here, we generate the first essential shadow by performing OR operation on Figure 2b,c, and the second essential shadow is generated by performing OR operation on Figure 2d,e. The remaining two shadows of RGVCS are considered as two non-essential shadows of (1, 2, 2, 4)-EVCS.

Figure 7 shows the experimental results. Figure 7a,b are two essential shadows and Figure 7c,d show two non-essential shadows, which have the same size of the secret image. Figure 7e–j illustrate the recovered image by stacking any two shadows. Since none of the essential shadows are included in {3, 4}, Figure 7j is as cluttered as random noise and does not show any information about the secret image. Figure 7k–n show the revealed image recovered by any three shadows and the last one is revealed by all shadows. Stacking two or more shadows that include any one or two essential shadows can reveal the secret image. Reconstruction without the essential shadow cannot get obtain information regarding the secret.

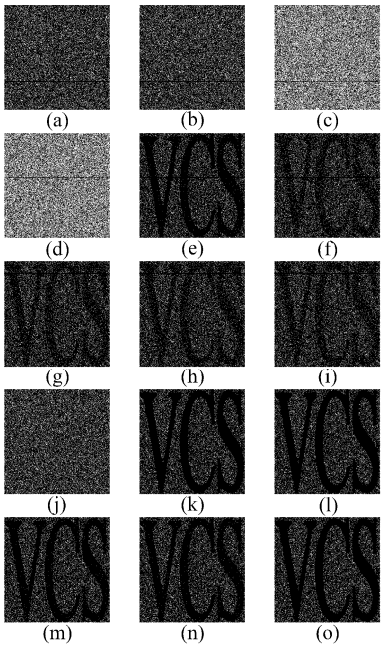


Figure 7. An experiment of the proposed (1,2,2,4)-EVCS. (**a**,**b**) two essential shadows; (**c**,*d*) two non-essential shadows; (**e**) revealed image by shadow 1 and 2; (**f**) revealed image by shadow 1 and 3; (**g**) revealed image by shadow 1 and 4; (**h**) revealed image by shadow 2 and 3; (**i**) revealed image by shadow 2 and 4; (**j**) revealed image by shadow 3 and 4; (**k**) revealed image by shadow 1, 2, and 3; (l) revealed image by shadow 1, 2 and 4; (**m**) revealed image by shadow 1, 3, and 4; (**n**) revealed image by shadow 2, 3, and 4; and, (**o**) revealed image by four shadows.

4.2. Comparison and Discussion

This subsection compares the proposed scheme with some literature schemes in terms of functionalities, as shown in Table 9. Both [24] and [25] are polynomial-based ESIS schemes that can reveal secret image perfectly, while they suffers from the disadvantage of heavy computation that secret information cannot be obtained by superimposing shadow images. In addition, these two schemes have the problem of unequal sizes of essential shadow and non-essential shadow, and the concatenation of sub-shadows. However, scheme [17,18,28] and the proposed scheme do not have

these two problems. Among them, scheme [17,18] and [28] are only applicable to (t, k, n), while the proposed scheme is applicable to a wider range. When compared with the polynomial-based ESIS scheme, the proposed scheme does not require complicated mathematical calculations in the secret reconstruction process. Most importantly, our scheme can select the appropriate (k, n)-VCS according to the actual needs. With the improvement of the (k, n)-VCS scheme, our scheme will also achieve better visual effects. The threshold condition refers to that the number of shadows in a qualified set should be no less than a threshold number. The essentiality condition refers to that a qualified set of shadows should contain at least a certain number of essentials. All of the mentioned schemes in Table 9 satisfy the threshold condition. When compared with the general VCS [32,35], our scheme not only satisfies the threshold condition, but it also satisfies the essentiality condition.

Table 9. The comparison of functionality among the literature schemes and proposed EVCS.

Schemes	Construction Method	Size Expansion	Concatenation of Sub-shadows	Essentiality	Decoding Operation	Stacking-to-see
Scheme [17]	VCS	Large	No	Yes	OR	Yes
Scheme [18]	VCS	Large	No	Yes	OR	Yes
Scheme [24]	PSIS	Small	Yes	Yes	Lagrange's interpolation	No
Scheme [25]	PSIS	Small	Yes	Yes	Lagrange's interpolation	No
Scheme [26]	PSIS	Small	No	Yes	Birkhoff Interpolation	No
Scheme [28]	PSIS	Small	No	Yes	Lagrange's interpolation	No
Scheme [32]	RGVCS	Small	No	No	OR	Yes
Scheme [35]	XVCS	Medium	Yes	No	XOR	Yes
Proposed scheme	VCS	Alternative	No	Yes	OR	Yes

In general, the reconstructed image of VCS is not completely consistent with the secret image. The size expansion and visual quality are commonly used to measure the performance of VCS. The proposed scheme has high flexibility that a (t, s, k, n)-EVCS can be constructed utilizing any monotone (k, n)-VCS. The performance of EVCS is determined by the performance of chosen VCS. For example, if we construct (t, s, k, n)-EVCS by Yan et al.'s (k, n)-RGVCS, we can achieve (t, s, k, n)-EVCS without size expansion. In Example 3, we adopt a traditional (k, n)-VCS and a non-size-expansion (k, n)-RGVCS in order to validate our scheme, respectively. From Example 3, we know that EVCS based on RGVCS has better size expansion than that based on traditional VCS. Arumugam et al. [17] proposed (1, 1, k, n)-EVCS and Guo et al. [18] proposed (t, t, k, n)-EVCS. Both schemes in [17,18] are special cases of (t, s, k, n)-EVCS, and construct the basis matrices of EVCS from the basis matrices of VCS. The main disadvantage of EVCS in [17,18] is the large size expansion, which is not convenient for the storage and transmission. We compare the experiment results among Arumugam et al.'s (1, 1, 3, 4)-VCS, Guo et al.'s (1, 1, 3, 4)-VCS, and the proposed (1, 1, 3, 4)-EVCS. Figure 8 shows the experimental results. The sizes of the revealed secret images of the three schemes are six, six, and one times of the secret image, respectively. Figure 8b,d show the revealed secret images of the three schemes, respectively. As we can see, the revealed images of Arumugam et al.'s (1, 1, 3, 4)-VCS, Guo et al.'s (1, 1, 3, 4)-VCS have large size expansion, while the proposed (1, 1, 3, 4)-EVCS has no size expansion, namely, the size of revealed image is the same as the original secret image. In addition, our proposed scheme can realize general (t, s, k, n)-EVCS.

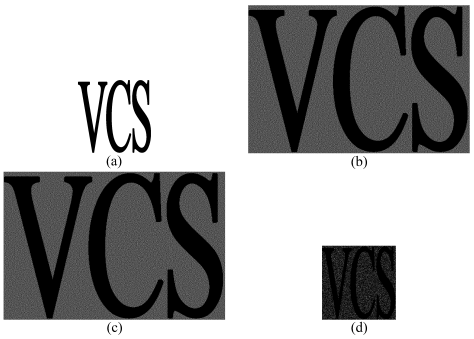


Figure 8. The experimental results of Arumugam et al.'s (1, 1, 3, 4)-VCS, Guo et al.'s (1, 1, 3, 4)-VCS and the proposed (1, 1, 3, 4)-EVCS. (**a**) the secret image; and, (**b**–**d**) the revealed image of the three schemes, respectively.

In this paper, we propose EVCS based on existing monotonic VCS. From the definition of (K, N)-VCS, the contrast condition guarantees that no information about the secret can be revealed with less than K shadows. When we construct (t, s, k, n)-EVCS from a monotonic (K, N)-VCS, the security condition of EVCS is derived from that of VCS. In Section 3.2, we show the construction method of EVCS from a monotonic (K, N)-VCS, where w_1 (resp. w_2) shadows of VCS are stacked together as an essential (resp. non-essential) shadow of EVCS. Combining with the access structures of EVCS and VCS, some constraints should be satisfied when determining the values of w_1 and w_2, as shown in Equations (9)–(11). Therefore, a forbidden set of EVCS cannot contribute at least K shadows of VCS, and then they cannot reveal any information regarding the secret image. In other words, the security level of the proposed EVCS is the same as the VCS.

For the security condition of (K, N)-VCS, no information regarding the secret image can be revealed with less than K shadows. With less than K shadows, the secret pixel is revealed as a black pixel or a white pixel with the same probability. From the view point of information theory, the entropy has the largest value. For the contrast condition of (K, N)-VCS, any K shadows can reveal the secret image by stacking operation. Each shadow can be viewed as the key to decode the secret image. Since all of the shadows have the same size, VCS can be also viewed as a one-time pad system. The Shannon theories have already proven that one-time pad system is a perfect secret system. Therefore, the secret image cannot be revealed with less than K shadows, even with computational resources. To the best of our knowledge, there is no cryptoanalysis scheme for VCS while using machine learning or deep learning algorithms.

4.3. Applications of EVCS

VCS is technique for sharing a binary secret image among the participants. The main advantage of VCS is that the revealing process does not need the computer resources. VCS has potential application when the collective decision making is required and the computer resources is not available. For example, in the battlefield, VCS shares the military instruction from the commander is shared into multiple shadows. Each shadow is delivered to a soldier. Since the environment of the battlefield is not predictable, the soldiers can decode the military instruction by stacking their shadows without any computational resources. VCS with essential participants (EVCS) divide the participants into two groups: essentials with higher status and non-essentials with lower status. In the revealing process, (t, e, k, n)-EVCS requires k participants, including t essentials to stacking their shadows. EVCS has the potential application when some participants are accorded special privileges due to their status or importance, e.g., heads of government, managers of company, high-level corporate officers, major employers, etc.

EVCS also has potential application in key distribution when exchanging message in a public secure network [30,31]. The trapped users may have different social attributes. Hence, they can be divided as essentials or non-essentials according to their attributes. Before message exchange, users

need to obtain the correct password in order to ensure the confidentiality of communication. The password usually consists of letters or numbers. It is suitable to share the password by EVCS. In addition, EVCS has the advantage of easy-decoding without computation. The environment of the various emergency events, e.g., natural disaster, terrorist attacks, etc., is terrible. The computational resources may be also limited. EVCS is a perfect way to decode secret by stacking shadows without computer. The decoding process is simple, and do not need any cryptography knowledge.

In realistic implementation, the access structure of EVCS should be open to public. The values of t, s, k, and n are known to the participants. In addition, the essentials and non-essentials are credible and known to everyone. With the above known information, the participants can confirm whether they can constitute a qualified set. When the participants of a qualified set are gathered and their shadows are collected, the secret information can be easily decoded by stacking their shadows.

5. Conclusions

In this paper, we proposed a construction method for (t, s, k, n)-EVCS with essential participants. The proposed EVCS is constructed from a monotonic VCS that is based on integer programming. When cmpared with literature EVCS, the proposed EVCS might achieve no size expansion if we adopt RGVCS to generate shadows. By solving the corresponding integer programming model, we give the condition and optimal choice of (K, N)-VCS to construct (t, s, k, n)-EVCS. The proposed EVCS also has the advantage of easy decoding since VCS can reveal the secret image by stacking shadows. The experimental results show the feasibility of our scheme. The construction method of general (t, s, k, n)-EVCS scheme with better performance needs further study.

Author Contributions: Conceptualization, P.L. and L.Y.; methodology, L.Y.; formal analysis, J.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Natural Science Foundation of Hebei Province (Grant number: F2019502173), National Natural Science Foundation of China (Grant number: 61602173) and the Fundamental Research Funds for Central Universities (Grant number: 2019MS116).

Acknowledgments: The authors would like to thank the editor and the anonymous reviewers for their valuable comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Naor, M.; Shamir, A. Visual Cryptography. *Lect. Notes Comput. Sci.* **1994**, *950*, 1–12.
2. Liu, F.; Guo, T.; Wu, C.K.; Qian, L. Improving the visual quality of size invariant visual cryptography scheme. *J. Vis. Commun. Image Represent.* **2012**, *23*, 331–342. [CrossRef]
3. Fu, Z.; Yu, B. Optimal pixel expansion of deterministic visual cryptography scheme. *Multimed. Tools Appl.* **2014**, *73*, 1177–1193. [CrossRef]
4. Li, P.; Ma, J.; Yin, L.; Ma, Q. A Construction Method of (2, 3) Visual Cryptography Scheme. *IEEE Access* **2020**, *8*, 32840–32849. [CrossRef]
5. Cimato, S.; Prisco, R.D.; Santis, A.D. Optimal Colored Threshold Visual Cryptography Schemes. *Des. Codes Cryptogr.* **2005**, *35*, 311–335. [CrossRef]
6. Liu, F.; Wu, C.K.; Lin, X.J. Colour visual cryptography schemes. *IET Inf. Secur.* **2008**, *2*, 151–165. [CrossRef]
7. Dutta, S.; Adhikari, A.; Ruj, S. Maximal contrast color visual secret sharing schemes. *Des. Codes Cryptogr.* **2019**, *87*, 1699–1711. [CrossRef]
8. Shyu, S.J.; Huang, S.Y.; Lee, Y.K.; Wang, R.Z.; Chen, K. Sharing multiple secrets in visual cryptography. *Pattern Recognit.* **2007**, *40*, 3633–3651. [CrossRef]
9. Chen, C.C.; Wu, W.J. A secure Boolean-based multi-secret image sharing scheme. *J. Syst. Softw.* **2014**, *92*, 107–114. [CrossRef]
10. Tsai, D.S.; Chen, T.; Horng, G. On generating meaningful shares in visual secret sharing scheme. *Imaging Sci. J.* **2008**, *56*, 49–55. [CrossRef]

11. Shyu, S.J. Threshold Visual Cryptographic Scheme with Meaningful Shares. *IEEE Signal Process. Lett.* **2014**, *21*, 1521–1525. [CrossRef]
12. Shyu, S.J. Image encryption by random grids. *Pattern Recognit.* **2007**, *40*, 1014–1031. [CrossRef]
13. Chen, T.H.; Tsao, K.H. User-Friendly Random-Grid-Based Visual Secret Sharing. *IEEE Trans. Circuits Syst. Video Technol.* **2011**, *21*, 1693–1703. [CrossRef]
14. Wu, X.; Sun, W. Random grid-based visual secret sharing with abilities of OR and XOR decryptions. *J. Vis. Commun. Image Represent.* **2013**, *24*, 48–62. [CrossRef]
15. Hu, H.; Shen, G.; Liu, Y.; Fu, Z.; Yu, B. Improved schemes for visual secret sharing based on random grids. *Multimed. Tools Appl.* **2019**, *78*, 12055–12082. [CrossRef]
16. Jin, D.; Yan, W.-Q.; Kankanhalli, M.S. Progressive color visual cryptography. *J. Electron. Imaging* **2005**, *14*, 033019. [CrossRef]
17. Arumugam, S.; Lakshmanan, R.; Nagar, A.K. On (k,n)*-visual cryptography scheme. *Des. Codes Cryptogr.* **2014**, *71*, 153–162. [CrossRef]
18. Guo, T.; Liu, F.; Wu, C.K.; Ren, Y.W.; Wang, W. On (k, n) Visual Cryptography Scheme with t Essential Parties. *Lect. Notes Comput. Sci.* **2013**, *8317*, 56–68.
19. Thien, C.C.; Lin, J.-C. Secret image sharing. *Comput. Graph.* **2002**, *26*, 765–770. [CrossRef]
20. Wu, Z.; Liu, Y.-N.; Wang, D.; Yang, C.-N. An Efficient Essential Secret Image Sharing Scheme Using Derivative Polynomial. *Symmetry* **2019**, *11*, 69. [CrossRef]
21. Liu, Y.; Yang, C.; Wang, Y.; Lei, Z.; Ji, W. Cheating Identifiable Secret Sharing Scheme Using Symmetric Bivariate Polynomial. *Inf. Sci.* **2018**, *453*, 21–29. [CrossRef]
22. Zhou, X.; Lu, Y.; Yan, X.; Wang, Y.; Liu, L. Lossless and Efficient Polynomial-Based Secret Image Sharing with Reduced Shadow Size. *Symmetry* **2018**, *10*, 249. [CrossRef]
23. Liu, Y.-X.; Yang, C.-N.; Wu, C.-M.; Sun, Q.-D.; Bi, W. Threshold changeable secret image sharing scheme based on interpolation polynomial. *Multimed. Tools Appl.* **2019**, *78*, 18653–18667. [CrossRef]
24. Li, P.; Yang, C.N.; Wu, C.C.; Kong, Q.; Ma, Y. Essential secret image sharing scheme with different importance of shadows. *J. Vis. Commun. Image Represent.* **2013**, *24*, 1106–1114. [CrossRef]
25. Yang, C.N.; Li, P.; Wu, C.C.; Cai, S.R. Reducing shadow size in essential secret image sharing by conjunctive hierarchical approach. *Signal Process. Image Commun.* **2015**, *31*, 1–9. [CrossRef]
26. Li, P.; Yang, C.N.; Zhou, Z. Essential secret image sharing scheme with the same size of shadows. *Digit. Signal Process.* **2016**, *50*, 51–60. [CrossRef]
27. Li, P.; Liu, Z. An Improved Essential Secret Image Sharing Scheme with Smaller Shadow Size. *Int. J. Digit. Crime Forensics* **2018**, *10*, 78–94. [CrossRef]
28. Peng, L.; Liu, Z.; Yang, C.N. A construction method of (t,k,n)-essential secret image sharing scheme. *Signal Process. Image Commun.* **2018**, *65*, 210–220.
29. Liu, Y.; Yang, C. Scalable secret image sharing scheme with essential shadows. *Signal Process. Image Commun.* **2017**, *58*, 49–55. [CrossRef]
30. Tsiropoulou, E.; Koukas, K.; Papavassiliou, S. A socio-physical and mobility-aware coalition formation mechanism in public safety networks. *EAI Endorsed Trans. Future Internet* **2018**, *4*, 154176. [CrossRef]
31. Thai, M.; Wu, W.; Xiong, H. *Big Data in Complex and Social Networks*; CRC Press Taylor & Francis Group: Boca Raton, FL, USA, 2016; pp. 125–182.
32. Yan, X.; Liu, X.; Yang, C.N. An enhanced threshold visual secret sharing based on random grids. *J. Real Time Image Process.* **2018**, *14*, 61–73. [CrossRef]
33. Ateniese, G.; Blundo, C.; Santis, A.D.; Stinson, D.R. Visual Cryptography for General Access Structures. *Inf. Comput.* **1996**, *129*, 86–106. [CrossRef]
34. Kafri, O.; Keren, E. Encryption of pictures and shapes by random grids. *Opt. Lett.* **1987**, *12*, 377–379. [CrossRef]
35. Shen, G.; Liu, F.; Fu, Z.; Yu, B. Perfect contrast XOR-based visual cryptography schemes via linear algebra. *Des. Codes Cryptogr.* **2017**, *85*, 15–37. [CrossRef]

mathematics

MDPI

Article

Market Volatility of the Three Most Powerful Military Countries during Their Intervention in the Syrian War

Viviane Naimy [1], José-María Montero [2], Rim El Khoury [1,*] and Nisrine Maalouf [3]

[1] Faculty of Business Administration and Economics, Notre Dame University—Louaize, Zouk Mikayel, Zouk Mosbeh 72, Lebanon; vnaimy@ndu.edu.lb
[2] Department of Political Economy and Public Finance, Economic and Business Statistics, and Economic Policy, Faculty of Law and Social Sciences, University of Castilla-La Mancha, 45071 Toledo, Spain; Jose.mlorenzo@uclm.es
[3] Financial Risk Management—Faculty of Business Administration and Economics, Notre Dame University—Louaize, Zouk Mikayel, Zouk Mosbeh 72, Lebanon; nisrinemaalouf1@gmail.com
* Correspondence: rkhoury@ndu.edu.lb

Received: 8 April 2020; Accepted: 17 May 2020; Published: 21 May 2020

Abstract: This paper analyzes the volatility dynamics in the financial markets of the (three) most powerful countries from a military perspective, namely, the U.S., Russia, and China, during the period 2015–2018 that corresponds to their intervention in the Syrian war. As far as we know, there is no literature studying this topic during such an important distress period, which has had very serious economic, social, and humanitarian consequences. The Generalized Autoregressive Conditional Heteroscedasticity (GARCH (1, 1)) model yielded the best volatility results for the in-sample period. The weighted historical simulation produced an accurate value at risk (VaR) for a period of one month at the three considered confidence levels. For the out-of-sample period, the Monte Carlo simulation method, based on student t-copula and peaks-over-threshold (POT) extreme value theory (EVT) under the Gaussian kernel and the generalized Pareto (GP) distribution, overstated the risk for the three countries. The comparison of the POT-EVT VaR of the three countries to a portfolio of stock indices pertaining to non-military countries, namely Finland, Sweden, and Ecuador, for the same out-of-sample period, revealed that the intervention in the Syrian war may be one of the pertinent reasons that significantly affected the volatility of the stock markets of the three most powerful military countries. This paper is of great interest for policy makers, central bank leaders, participants involved in these markets, and all practitioners given the economic and financial consequences derived from such dynamics.

Keywords: GARCH; EGARCH; VaR; historical simulation approach; peaks-over-threshold; EVT; student t-copula; generalized Pareto distribution

1. Introduction

Political uncertainty occurs due to many factors like elections and changes in the government or parliament, changes in policies, strikes, minority disdain, foreign intervention in national affairs, and others. In many cases, these uncertainties lead to further complications affecting the economy and the financial market of the concerned country. Accordingly, the currency could devaluate, prices of assets, commodities, and stocks could fluctuate, and the growth of the economy could be hindered. From this perspective, countries strive to keep political risks controlled to be able to endure the cost or consequence of any sudden political unrest. This is one of the main reasons behind the intervention of powerful countries in the political and military affairs of less powerful countries, which is usually done at a high cost. This paper studies the impact of the intervention of the three most powerful military

countries in the world, namely, the United States, Russia, and China (Figure 1), in the Syrian war on their market volatility.

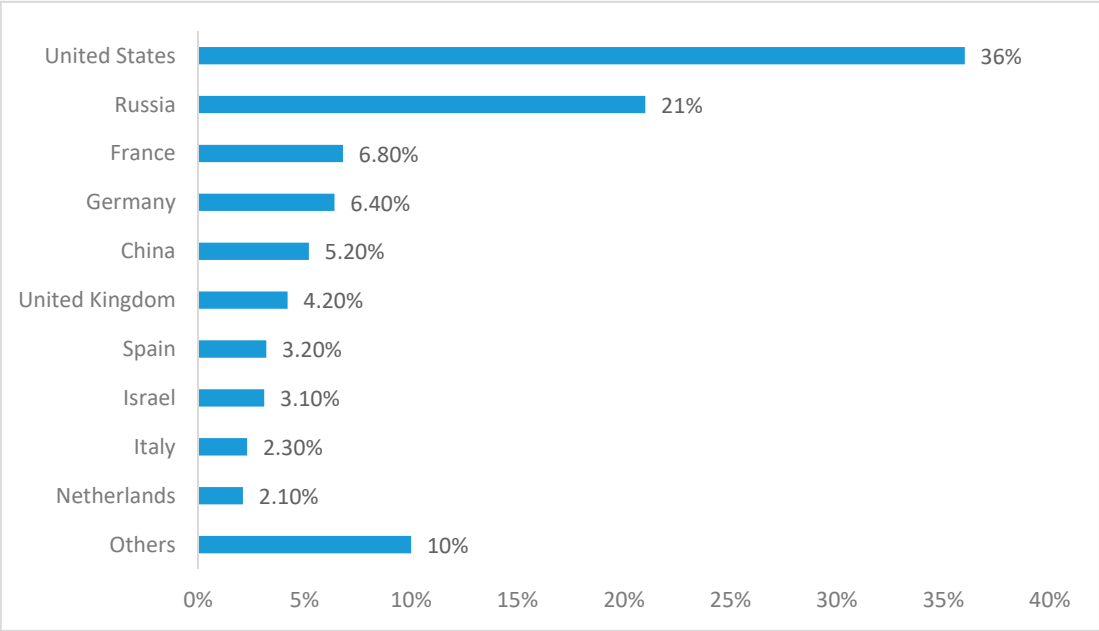


Figure 1. Global share of major arms exports by the 10 largest exporters, 2014–2018.

In March 2011, large peaceful protests broke in Syria to call for economic and political reforms with few armed protesters, leading to man arrests. Events evolved into violent acts using artillery and aircrafts, antigovernment rebels, terrorist and extremist attacks, suicide attacks, explosive operations, the intervention of foreign countries, chemical weapons, and others leading to a humanitarian crisis. In 2015, Russia started supporting the Syrian president through financial aid and military support [1]. In the meantime, the United States was providing support for the local Syrians. Later on, the United States and Russia increased their intervention in the war mainly through arms and aircrafts, each supporting their own political interests and allies. By the same token, China's involvement was shifting from humanitarian assistance and weapon exports [2] to armed forces and increased weapon exports to support its allies' objectives during this war [3].

Table 1 shows countries with the highest military spending in the world for 2016, 2017, and 2018. The U.S. spends the highest budget in the world on defense forces. This expenditure rounded up to USD 649 billion during 2018 based on information from the Stockholm International Peace Research Institute [4]. In fact, the defense spending of the United States alone is higher than the sum of that of the next eight countries in the ranking. These countries include China, Russia, Saudi Arabia, India, France, UK, Japan, and Germany. The country with the second highest defense expenditure is China with USD 250 billion in 2018 compared to USD 228 billion in 2017. As for Russia, its expenditure reached USD 61.4 billion in 2018 compared to USD 66.5 billion in 2017. Figure 1 represents the 10 largest arms exporters in the world between 2013 and 2017 [5]. Besides having the highest budgets for defense, the U.S. and Russia are also the top exporters of weapons, and China is among the top five worldwide countries. Based on these facts, the importance of the U.S., China, and Russia among military countries is highly reinforced. For this reason, we opted to study the dynamics of their financial markets to comprehend the risks and opportunities they might face, which would affect their worldwide exposure.

Table 1. Countries with the highest military spending worldwide in 2016–2018 (In Billion USD).

In USD Billion	2016	2017	2018
USA	600.1	605.8	648.8
China	216.0	227.8	250.0
Russia	69.2	66.5	61.4
Saudi Arabia	63.7	70.4	67.6
India	56.6	64.6	66.5
France	57.4	60.4	63.8

Source: Stockholm International Peace Research Institute (SIPRI), 2019.

To this end, measuring the effect of their intervention in the Syrian war on their financial market volatility is of great importance for policy makers, central bank leaders, analysts, and practitioners because there is a complete absence in the literature of studies that involve the volatility of the financial markets of the U.S., China, and Russia together. Many studies, however, explored the volatility of these countries during different periods and using different volatility models.

In his paper, Wei [6] forecasted the Chinese stock market volatility using non-linear Generalized Autoregressive Conditional Heteroscedasticity (GARCH) models such as the quadratic GARCH (QGARCH) and the Glosten, Jagannathan, and Runkle GARCH (GJR GRACH) models. The author studied seven-year data for the Shanghai Stock Exchange Composite (HSEC) and the Shenzhen Stock Exchange Component (ZSEC). The QGARCH outperformed the linear GARCH model. Furthermore, Lin and Fei [7] concluded that the nonlinear asymmetric power GARCH (APGARCH) model outperformed other GARCH models on different time scales in estimation of the "long memory property of the Shanghai and Shenzhen stock markets". Recently, Lin [8] studied the volatility of the SSE Composite Index using GARCH models during the period 2013–2017. The asymmetric exponential GARCH (EGARCH (1, 1)) model outperformed the symmetric ones in the forecasting results.

Value at risk (VaR), extreme value theory (EVT), and expected shortfall (ES) models were also used by Wang et al. [9], who implemented an EVT based VaR and ES to estimate the exchange rate risk of the Chinese currency (CNY). They found that the EVT-based VaR estimation produces accurate results for the currency exchange rate risks of EUR/CNY and JPY/CNY. However, EVT underestimated this risk for both exchange rates. Chen et al. [10] estimated VaR and ES by applying EVT on 13 worldwide stock indices. They concluded that China ranks first for VaR and ES with negative returns and ranks third for positive returns with high levels of risk.

A new strategy to estimate daily VaR based on the autoregressive fractionally integrated moving average model (ARFIMA), the multifractal volatility (MFV) model, and EVT was implemented by Wei et al. [11] for the Chinese stock market using high-frequency intraday quotes of the Shanghai Stock Exchange Component (SSEC). This hybrid ARFIMA-MFV-EVT strategy was compared to a number of popular linear and nonlinear GARCH-type-EVT models, i.e., the RiskMetrics, GARCH, IGARCH, and EGARCH models. Although GARCH-type models showed a good performance, VaR results obtained from the ARFIMA-MFV-EVT method outperformed several of them widely used in the literature. Furthermore, Hussain and Li [12] focused on the effect of extreme returns in stock markets on risk management by studying the SSEC index and by using the block maxima (Minima) method (BMM), instead of the popular peaks-over-threshold (POT) method, with various time intervals of extreme daily returns. Three well-known distributions in extreme value theory, i.e., generalized extreme value (GEV), generalized logistic (GL), and generalized Pareto distributions (GP), were employed to model the SSEC index returns. Results showed that GEV and GL distributions are found to be appropriate for the modeling of the extreme upward and downward market movements for China.

Another comparative study, conducted by Hou and Li [13], investigated the transmission of information between the U.S. and China's index futures markets using an asymmetric dynamic conditional correlation GARCH (DCC GARCH) approach. They found that the correlation between U.S. and Chinese index futures markets increases with the rise of negative shocks in these markets, and that the U.S. index futures market is more efficient in terms of price adjustment, since it is older

and more mature. On the other hand, Awartani and Corradi [14] focused on the role of asymmetries in the prediction of the volatility of the S&P 500 Composite Price Index. They examined the relative out-of-sample predictive ability of different GARCH-type models. First, they performed pairwise comparisons of various models against GARCH (1, 1). Then, they carried out a joint comparison of all models. They found that for the case of the one-step ahead pairwise comparison, GARCH (1, 1) is beaten by the asymmetric GARCH models. A similar finding applies to different longer forecast horizons. In the multiple comparison case, GARCH (1, 1) is only beaten when compared against the class of asymmetric GARCH. Another interesting finding is that the RiskMetrics exponential smoothing seems to be the worst model in terms of predictive ability. Furio and Climent [15] studied extreme movements in the return of S&P 500, FTSE 100, and NIKKEI 225 using GARCH-type models and EVT estimates. Results pointed out that more accurate estimates are derived from EVT calculations in both the in-sample and out-of-sample, when compared to less accurate estimates using the GARCH models.

As can be deduced from the review of the above literature, the question of how devastating wars, with indirect consequences all around the world, affect the volatility of financial markets of countries supporting them (and others, of course) might be of core importance for those directly or indirectly involved in such markets. Therefore, the main research question is the following: are the volatility dynamics of those countries affected by an event of the importance of the Syrian war? This paper fills this gap through evaluating the results of a number of traditional volatility models of the GARCH-type family and using EVT and historical simulation (HS) to estimate the VaR of these markets during the Syrian war period.

S&P 500 (Standard & Poor's), SSEC (Shanghai Stock Exchange Composite) and MICEX (Moscow Interbank Currency Exchange) are used to assess the financial markets' volatility of the U.S., China, and Russia, respectively. The period of study extends from 2015 to 2018. The in-sample period extends from 5 January 2015 until 30 December 2016 as it refers to the beginning of the direct and indirect intervention of the chosen countries in the war in Syria [1]; the out-of-sample interval is 3 January 2017–31 May 2018.

The paper is structured as follows: Section 2 reviews the methodology and the specificities of the applied econometric models, and Section 3 shows the estimated GARCH-type models considered and the selection process. This section also depicts the results related to the calculation of VaR using HS volatility and the "peaks-over-threshold" (POT) EVT model under the GP distribution. Section 4 concludes and discusses the empirical findings.

2. Econometric Models

As previously outlined, we use the GARCH (1, 1) and EGARCH (1, 1) as competing models to measure the volatility of the financial markets of the U.S., Russia, and China. GARCH models are commonly used by financial institutions to obtain volatility and correlation forecasts of asset and risk factor return. We use the symmetric normal GARCH given its strength to provide short- and medium-term volatility forecasts. We also use EGARCH, the asymmetric GARCH model, which is widely recognized in providing a better in-sample fit than other types of GARCH processes and avoids the need for any parameter constraints (see [16,17] for details on other GARCH-type models). The exponentially weighted moving average (EWMA) model is not used because it does not account for mean reversion and overvalue volatility after severe price fluctuation [18]. As said in the introductory section, for VaR estimation with high confidence intervals, we apply EVT [19], and more specifically GEV, GL, and GP distributions. We decided to use EVT because of its ability to provide good estimates and serve of help in situations where high confidence levels are needed, since EVT has proven to be a robust way of smoothing and extrapolating the tails of an empirical distribution [20]. The EVT implementation in this paper is based on a multivariate analysis to accurately measure the VaR of the portfolio composed of the U.S., Russia, and China stock markets. We also estimate the VaR of the portfolio using HS for comparison.

2.1. GARCH Model

The pioneering work of Engle [21], where the Autoregressive Conditional Heteroscedasticity (ARCH) model (that relates the current level of volatility to p past squared error terms) was introduced, constitutes the main pillar of modern financial econometrics. However, the ARCH strategy has some limitations, including the typically required 5–8 lagged error terms to adequately model conditional variance. That was the reason for this model to be generalized by Bollerslev [22], giving rise to the generalized ARCH (GARCH) model, by adding lagged conditional variance, which acts as a smoothing term. In practical terms, the GARCH (p, q) model builds on the ARCH (p) by including q lags of the conditional variance. Therefore, a GARCH specification uses the weighted average of long-run variance, the predicted variance for the current period, and any new information in this period, as captured by the squared residuals, to forecast a future variance. More specifically, the general GARCH (p, q) model is as shown in Equation (1):

$$\sigma_t^2 = \gamma V_L + \sum_{i=1}^{p} \alpha_i u_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 \tag{1}$$

where σ_t^2 is the time $t-1$ conditional variance, V_L is the long run average variance, σ_{t-j}^2 are the lags of the conditional variance, and u_{t-i}^2 are the lagged squared error terms. $u_t = \sigma_t e_t$ with e_t *i.i.d.* $N(0, 1)$. Coefficients γ, α_i and β_j are the weights for V_L and the lags of the conditional variance and the squared error terms, respectively, and their estimates are obtained by Maximum Likelihood.

GARCH (1, 1) is the most used model of all GARCH models. It can be written as follows:

$$\sigma_t^2 = \gamma V_L + \alpha_1 u_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \tag{2}$$

or, alternatively,

$$\sigma_t^2 = \omega + \alpha_1 u_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \tag{3}$$

where $\omega = \gamma V_L$. Coefficients in the GARCH specification sum up to the unity and have to be restricted for the conditional variances to be uniformly positive. In the case of the GARCH (1, 1) such restrictions are: $\omega > 0$, $\alpha_1 \geq 0$ and $\beta_1 \geq 0$. In addition, the requirement for stationarity is $1 - \alpha_1 - \beta_1 > 0$. The unconditional variance can be shown to be $E\left(\sigma_t^2\right) = \omega/(1 - \alpha_1 - \beta_1)$.

2.2. EGARCH Model

The EGARCH model was proposed by Nelson [23] to capture the leverage effects observed in financial series and represents a major shift from the ARCH and GARCH models. The EGARCH specification does not model the variance directly, but its natural logarithm. This way, there is no need to impose sign restrictions on the model parameters to guarantee that the conditional variance is positive. In addition, EGARCH is an asymmetric model in the sense that the conditional variance depends not only on the magnitude of the lagged innovations but also on their sign. This is how the model accounts for the different response of volatility to the upwards and downwards movement of the series of the same magnitude. More specifically, EGARCH implements a function $g(e_t)$ of the innovations e_t, which are *i.i.d.* variables with zero mean, so that the innovation values are captured by the expression $|e_t| - E|e_t|$.

An EGARCH (p, q) is defined as:

$$\log \sigma_t^2 = \omega + \sum_{j=1}^{q} \beta_j \log \sigma_{t-j}^2 + \sum_{j=1}^{p} \theta_i g\left(e_{t-j}\right) \tag{4}$$

where $g(e_t) = \delta e_t + \alpha(|e_t| - E|e_t|)$ are variables *i.i.d.* with zero mean and constant variance. It is through this function that depends on both the sign and magnitude of e_t, that the EGARCH model captures

the asymmetric response of the volatility to innovations of different sign, thus allowing the modeling of a stylized fact of the financial series: negative returns provoke a greater increase in volatility than positive returns do.

The innovation (standardized error divided by the conditional standard deviation) is normally used in this formulation. In such a case, $E|e| = \sqrt{2/\pi}$ and the sequence $g(e_t)$ is time independent with zero mean and constant variance, if finite. In the case of Gaussianity, the equation for the variance in the model EGARCH (1, 1) is:

$$\log \sigma_t^2 = \omega + \beta \log \sigma_{t-1}^2 + \delta e_{t-1} + \alpha\left(|e_{t-1}| - \sqrt{\frac{2}{\pi}}\right). \tag{5}$$

Stationarity requires $\left|\beta\right| < 1$, the persistence in volatility is indicated by β, and δ indicates the magnitude of the leverage effect. δ is expected to be negative, which implies that negative innovations have a greater effect on volatility than positive innovations of the same magnitude. As in the case of the standard GARCH specification, maximum likelihood is used for the estimation of the model.

2.3. EVT

EVT deals with the stochastic behavior of extreme events found in the tails of probability distributions, and, in practice, it has two approaches. The first one relies on deriving block maxima (minima) series as a preliminary step and is linked to the GEV distribution. The second, referred to as the peaks over threshold (POT) approach, relies on extracting, from a continuous record, the peak values reached for any period during which values exceed a certain threshold and is linked to the GP distribution [24]. The latter is the approach used in this paper.

The generalized Pareto distribution was developed as a distribution that can model tails of a wide variety of distributions. It is based on the POT method which consists in the modelling of the extreme values that exceed a particular threshold. Obviously, in such a framework there are some important decisions to take: (i) the threshold, μ; (ii) the cumulative function that best fits the exceedances over the threshold; and (iii) the survival function, that is, the complementary of the cumulative function.

The choice of the threshold implies a trade-off bias-variance. A low threshold means more observations, which probably diminishes the fitting variance but probably increases the fitting bias, because observations that do not belong to the tail could be included. On the other hand, a high threshold means a fewer number of observations and, maybe, an increment in the fitting variance and a decrement in the fitting bias.

As for the distribution function that best fits the exceedances over the threshold, let us suppose that $F(x)$ is the distribution function for a random variable X, and that threshold μ is a value of X in the right tail of the distribution; let y denote the value of the exceedance over the threshold μ. Therefore, the probability that X lies between μ and $\mu + y$ $(y > 0)$ is $F(\mu + y) - F(\mu)$ and the probability for X greater than μ is $1 - F(\mu)$. Writing the exceedances (over a threshold μ) distribution function $F^{\mu}(y)$ as the probability that X lies between μ and $\mu + y$ conditional on $X > \mu$, and taking into account the identity linking the extreme and the exceedance: $X = Y + \mu$, it follows that:

$$F^{\mu}(y) = P(Y \leq y \left| X > \mu\right.) = P(\mu < X \leq \mu + y \left| X > \mu\right.) = \frac{F(x) - F(\mu)}{1 - F(\mu)} \tag{6}$$

and that

$$1 - F^{\mu}(y) = 1 - \frac{F(x) - F(\mu)}{1 - F(\mu)} = \frac{1 - F(x)}{1 - F(\mu)} \tag{7}$$

In the case that the parent distribution F is known, the distribution of threshold exceedances also would be known. However, this is not the practical situation, and approximations that are broadly applicable for high values of the threshold are sough. Here is where Pickands–Balkema–de Haan

theorem ([25,26]) comes into play. Once the threshold has been estimated, the conditional distribution $F^{\mu}(y)$ converges to the GP distribution. It is known that $F^{\mu}(y) \to G_{\xi,\sigma}(y)$ as $\mu \to \infty$, with

$$G_{\xi,\sigma}(y) = \begin{cases} 1 - \left(1 + \xi \frac{y}{\sigma}\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0 \\ 1 - e^{-\frac{y}{\sigma}} & \text{if } \xi = 0 \end{cases} \tag{8}$$

where $\sigma > 0$ and $y \geq 0$ if $\xi \geq 0$ and $0 \leq y \leq -\sigma/\xi$ if $\xi < 0$. ξ is a shape parameter that determines the heaviness of the tail of the distribution, and σ is a scale parameter. When $\xi = 0$, $G_{\xi,\sigma}(y)$ reduces to the exponential distribution with expectation $\exp(\sigma)$; in the case that $\xi < 0$, it becomes a Uniform $(0, \sigma)$; finally, $\xi > 0$ leads to the Pareto distribution of the second kind [27]. In general, ξ has a positive value between 0.1 and 0.4. The GP distribution parameters are estimated via maximum likelihood.

Once the maximum likelihood estimates are available, a specific GP distribution function is selected, and an analytical expression for *VaR* with a confidence level q can be defined as a function of the GP distribution parameters:

$$VaR_{\hat{q}} = \mu + \frac{\hat{\sigma}(\mu)}{\hat{\xi}} \left(\frac{N}{N_{\mu}} (1-q)^{-\hat{\xi}} - 1 \right) \tag{9}$$

where N is the number of observations in the left tail and N_{μ} is the number of excesses beyond the threshold μ.

$$VaR_{\hat{q}} = \mu + \frac{\hat{\sigma}(\mu)}{\hat{\xi}} \left(\frac{N}{N_{\mu}} (1-q)^{-\hat{\xi}} - 1 \right) \tag{10}$$

3. Results

3.1. Descriptive Statistics

Data for 3 years were extracted from the Bloomberg platform for the three selected stock market indices and were manipulated to derive the return from the closing prices corresponding to each index. For the in-sample period, 458 daily observations were studied compared to 315 for the out-of-sample forecast period. It is important to note that in November 2017, the name of the MICEX index (composed of Russian stocks of the top 50 largest issues in the Moscow Exchange) was officially changed to the MOEX Russia Index, representing the "Russian stock market benchmark" [28]. Table 2 lists the descriptive statistics of S&P 500, SSEC, and MICEX for the in- and out-of-sample periods. Surprisingly, S&P 500 and the MICEX behaved similarly in terms of return during the in-sample and out-of-sample periods.

Table 2. Descriptive Statistics of S&P 500, SSEC, and MICEX: 5 January 2015–30 December 2016 and 3 January 2017–31 May 2018.

Stock Markets	S&P 500 (In-Sample)	SSEC (In-Sample)	MICEX (In-Sample)	S&P 500 (Out-of-Sample)	SSEC (Out-of-Sample)	MICEX (Out-of-Sample)
Mean	0.022%	−0.017%	0.017%	0.060%	−0.001%	0.054%
Standard Deviation	1.0%	2.10%	1.2%	0.711%	0.774%	1.025%
Skewness	1.0%	2.10%	1.2%	0.711%	0.774%	1.025%
Kurtosis	3.33	4.62	0.97	7.93	4.83	12.45
Median	0.01%	0.10%	0.04%	0.06%	0.08%	−0.05%
Minimum	−4.0%	−10.8%	−4.4%	−4.18%	−4.14%	−8.03%
Maximum	4.7%	6.0%	4.4%	2.76%	2.15%	3.89%
1st Quartile	−0.41%	−0.66%	−0.65%	−0.18%	−0.35%	−0.51%
3rd Quartile	0.49%	0.87%	0.85%	0.34%	0.40%	0.54%

In reference to the two chosen time periods, the skewness of the returns of the three indices is close to 0 and the returns display excess in kurtosis. This implies that the distributions of returns are not normal as confirmed by Jarque-Bera normality tests (Table 3). The distributions of returns are stationary according to the augmented Dicky–Fuller (ADF) test applied to the three indices (Table 4).

Table 3. Jarque-Bera Normality Test of S&P 500, SSEC, and MICEX.

Stock Markets	In-Sample		Out-of-Sample	
	Score	*p*-Value	Score	*p*-Value
S&P 500	198.66	0.000	865.78	0.000
SSEC	483.75	0.000	352.63	0.000
MICEX	18.12	0.000	2027.03	0.000

Note: *p*-value refers to Jarque–Bera normality test, Ho: the index return is normally distributed.

Table 4. ADF Stationarity Test.

	Critical Values at 5%	STAT	*p*-Value	STAT	*p*-Value	STAT	*p*-Value
		S&P 500		SSEC		MICEX	
No Constant	−1.9	−28.4	0.001	−12.8	0.001	−26.9	0.001
Constant Only	−2.9	−28.4	0.001	−12.8	0.001	−11.6	0.001
Constant and Trend	−1.6	−28.4	0.000	−12.7	0.000	−11.6	0.000
Constant, Trend, and Trend^2	−1.6	−28.4	0.000	−12.7	0.000	−11.6	0.000

Note: ADF *p*-value refers to the augmented Dickey–Fuller unit root test, Ho: the index return has a unit root.

3.2. GARCH (1, 1) and EGARCH (1, 1) Results

GARCH (1, 1) and EGARCH (1, 1) parameters were estimated using the daily returns of each index. Results from the normal distribution, the student's t-distribution and the Generalized Error Distribution (GED) were derived. The goodness of fit test and residual analysis were then performed to ensure that the assumptions of the applied models were all met. The model parameters were estimated by maximum likelihood. Table 5 presents a summary of such estimates for GARCH (1, 1) and EGARCH (1,1).

Table 5. Estimates of the Parameters of GARCH (1,1) and EGARCH (1,1) (with GED).

	S&P 500	SSEC	MICEX
		GARCH (1, 1)	
Long-run mean (μ)	0.000249	−0.00009	0.00080
Omega (ω)	8.5218×10^{-6}	1.1035×10^{-6}	2.9655×10^{-6}
ARCH component (α)	0.1972	0.0569	0.0653
GARCH component (β)	0.7105	0.9331	0.9065
		EGARCH (1, 1)	
Long-run mean (μ)	0.00030	−0.00009	0.00142
Omega (ω)	−0.72879	−0.08598	−0.87318
ARCH component (α)	0.04766	0.20530	0.30477
Leverage coefficient (δ)	−0.29149	−0.01362	0.05345
GARCH component (β)	0.92471	0.98689	0.90237

Note: The GED was selected after checking the average, standard deviation, skewness, kurtosis, the noise, and ARCH tests corresponding to the three distributions.

3.3. Best Volatility Model Selection for the In- and Out-of-Sample Periods

The root mean square error (RMSE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE) are implemented to choose the best volatility model. For S&P 500, we compared the estimated volatility to the implied volatility. However, this was not possible for SSEC and MICEX due the absence of data. Results depicted in Table 6 reveal the superiority of GARCH (1, 1) in estimating the volatility of the three countries in the in-sample period, which coincides with the peak period of the Syrian war. As for the out-of-sample period, while GARCH (1, 1) ranks first for S&P 500, EGARCH (1, 1) ranks first for the SSEC and MICEX with a difference in RMSE of around 0.002 and 0.01 units respectively as compared to GARCH (1, 1). Volatilities estimated with the superior volatility models in comparison to the realized volatilities for the out-of-sample period are depicted in Figure 2.

Table 6. In-Sample Period Error Statistics.

S&P 500 during In-Sample Period (from 5 January 2015 till 30 December 2016)						
	RMSE	**Rating**	**MAE**	**Rating**	**MAPE**	**Rating**
Implied Vol.	0.047694	2	0.039028	2	0.00377	3
GARCH (1, 1)	0.040995	1	0.028954	1	0.002376	1
EGARCH (1, 1)	0.049947	3	0.039066	3	0.003261	2
SSEC during In-Sample Period (from 5 January 2015 till 30 December 2016)						
GARCH (1, 1)	0.09567	1	0.080424	1	0.002879	1
EGARCH (1, 1)	0.11588	2	0.090075	2	0.003134	2
MICEX during In-Sample Period (from 5 January 2015 till 30 December 2016)						
GARCH (1, 1)	0.184631	1	0.173261	1	0.004775	1
EGARCH (1, 1)	0.188566	2	0.179723	2	0.004997	2
S&P 500 during Out-of-Sample (from 3 January 2017 till 31 May 2018)						
Implied Vol.	0.04809	3	0.04325	3	0.006092	3
GARCH (1, 1)	0.040024	1	0.035135	1	0.004896	1
EGARCH (1, 1)	0.047078	2	0.041034	2	0.005689	2
SSEC during Out-of-Sample (from 3 January 2017 till 31 May 2018)						
GARCH (1, 1)	0.08478	2	0.080437	2	0.004033	2
EGARCH (1, 1)	0.07113	1	0.063535	1	0.00322	1
MICEX during Out-of-Sample (from 3 January 2017 till 31 May 2018)						
GARCH (1, 1)	0.071894	2	0.064616	2	0.002909	2
EGARCH (1, 1)	0.069381	1	0.061135	1	0.00271	1

Note: $RME = \sqrt{\sum_{t=1}^{n}(f-Y)^2/n}$; $MAE = \sum_{t=1}^{n}|f-Y|/n$; $MAPE = 100\sum_{t=1}^{n}\left|\frac{f-Y}{Y}\right|/n$, where n is the number of periods, Y is the true value and f is the prediction value. The best model is the one that has a minimum value of *RMSE*, *MAE* and *MAPE*.

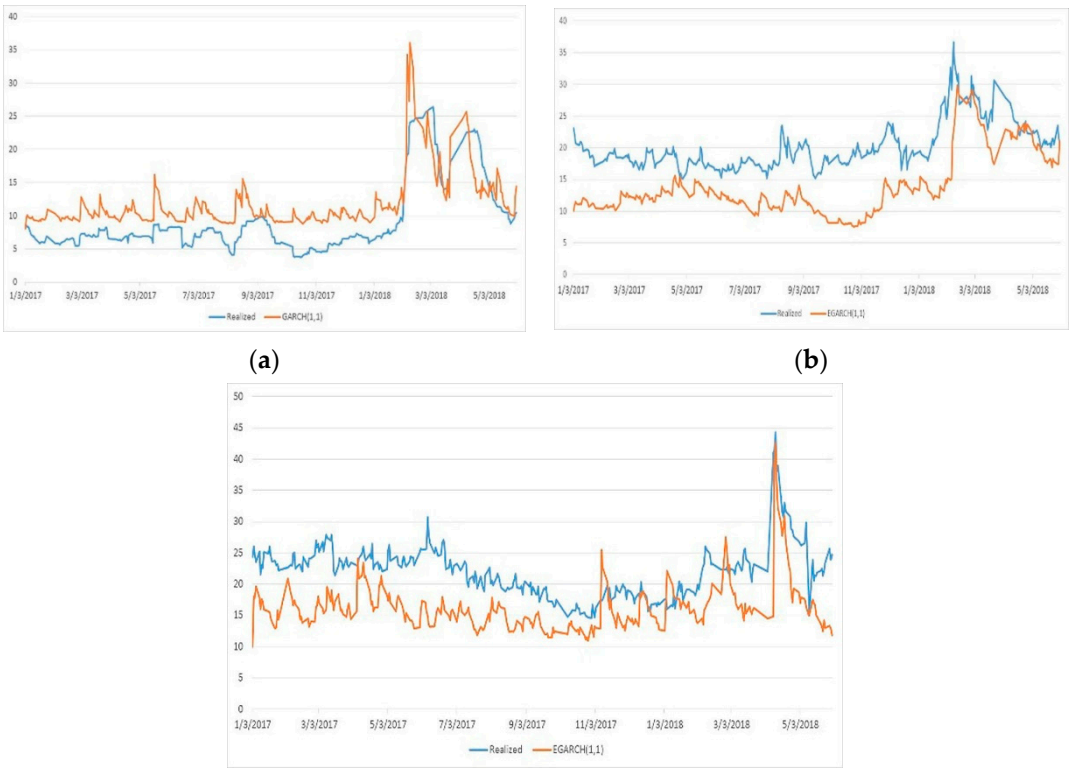


Figure 2. Realized Volatility vs. Volatilities Estimated with the Superior Volatility Models–Out-of-Sample Period: (**a**) for S&P 500. (**b**) for SSEC. (**c**) for MICEX.

3.4. VaR Output

Based on the superior model corresponding to each index, a portfolio of volatility updates was established for each sample period. First, the historical simulation approach was implemented. It involved incorporating volatility in updating the historical return. Because the volatility of a market variable may vary over time, we modified the historical data to reflect the variation in volatility. This approach uses the variation in volatility in a spontaneous way to estimate VaR by including more recent information. Second, a Monte Carlo simulation method including student t-copula and EVT was applied to the created portfolio composed of the three markets to estimate VaR with different confidence levels. The filtered residuals of each return series were extracted using EGARCH. The Gaussian kernel estimate was used for the interior marginal cumulative distribution function (CDF) and the generalized Pareto distribution (GP) was applied to estimate the upper and lower tails. The student t-copula was also applied to the portfolio's data in order to reveal the correlation among the residuals of each index. This process led to the estimation of the portfolio's VaR over a horizon of one month and confidence levels of 90%, 95%, and 99%. Table 7 summarizes all the VaR estimates calculated for the in-sample and out-of-sample periods using HS and EVT compared to the Real VaR. The visual illustrations of the relevant outcomes related to the logarithmic returns of the selected stock indices, the auto-correlation function (ACF) of returns and of the squared returns, the filtered residuals and the filtered conditional standard deviation, the ACF of standardized residuals, and the upper tail of standardized residuals for both periods, are presented in Appendix A (Figures A1 and A2).

Table 7. VaR Summary Results.

Outcome	HS (Volatility Weighted)	EVT	Real VaR
In-Sample			
90% VaR	0.68%	2.93%	1.31%
95% VaR	1.03%	4.83%	1.68%
99% VaR	2.31%	8.39%	2.38%
Out-of-Sample			
90% VaR	0.48%	3.76%	0.73%
95% VaR	0.71%	5.47%	0.94%
99% VaR	1.65%	9.60%	1.33%

It is apparent that VaR with a confidence level of 99%, using HS and EVT, overrates the risk for the three countries during both periods. Furthermore, the HS VaR results are closer to the Real VaR results compared to those of the EVT VaR. This is not altogether surprising since the EVT method is concerned with studying the behavior of extremes within these markets rather than simply fitting the curve. Therefore, the above output represents a benchmark that can be extrapolated beyond the data during stress periods.

4. Discussion

This paper revealed original common points among the most powerful military countries in the world regarding the behavior of their financial markets during the period 2015–2018, which corresponds to their intervention in the Syrian war. First, the returns of S&P 500 and MICEX were quite similar during the in-sample and out-of-sample periods. Second, the GARCH (1, 1) was found to be the best volatility model for the in-sample period for S&P 500, MICEX, and SSEC, outperforming EGARCH (1, 1). The incorporation of the GARCH (1, 1) specification to the HS produced an accurate VaR for a period of one month, at the three confidence levels, compared to the real VaR.

EVT VaR results are consistent with those found by Furio and Climent [15] and Wang et al. [9], who highlighted the accuracy of studying the tails of loss severity distribution of several stock markets. Furthermore, part of our results corroborates the work of Peng et al. [29], who showed that EVT GP distribution is superior to certain GARCH models implemented on the Shanghai Stock Exchange Index.

The GP distribution highlighted the behavior of "extremes" for the U.S., Russian, and Chinese financial markets, which is of great importance since it emphasizes the risks and opportunities inherent to the dynamics of their markets and also underlines the uncertainty corresponding to their worldwide exposure.

Expected EVT VaR values of 3.76%, 5.47%, and 9.60%, at 90%, 95%, and 99% confidence levels, respectively (for the out-of-sample period), might appear overstated. However, uncertainty layers are all the way inherent and our results are, naturally, subject to standard error. For comparison purposes, and in order to make a relevant interpretation of the tail distribution of returns corresponding to these markets, we opted to derive the EVT VaR of a portfolio of stock indices pertaining to non-military countries, namely Finland, Sweden, and Ecuador, for the same out-of-sample period of study (January 2017 to May 2018). These countries were chosen randomly based on the similarities in income groups when compared to the selected military countries. While Finland and Sweden are both classified as high income like the U.S., Ecuador is classified as upper middle income like Russia and China [30]. Also, the selection of these non-military countries follows the same structure of the capital market development found at the military countries which, when combined, find their average ratio of stock market capitalization to GDP is 77.23%, compared to 76.46% for Finland, Sweden, and Ecuador [30,31]. Finally, when comparing the ratio of private credit ratio to the stock market capitalization ratio corresponding to each country, we notice that the former is higher than the latter for Ecuador, Sweden, Finland, Russia, and China [32]. Only the U.S. depends mostly on its stock market to finance its

economy. The portfolio is composed of the OMX Helsinki index, the ECU Ecuador General index, and the OMX 30 Sweden index. The GP distribution was used to estimate the upper and lower tails. Remarkably, EVT VaR results were 2.23%, 3.49%, and 6.45% at the 90%, 95%, and 99% confidence levels, respectively, well below the estimates found for the U.S., Russian, and Chinese stock markets. Consequently, it can be concluded that the intervention in the Syrian war may have been one of the latent and relevant factors that affected the volatility of the stock markets of the selected military countries. This conclusion is reinforced by the fact that the EVT VaR was higher by 40%, 26%, and 32%, at the 90%, 95%, and 99% confidence levels, respectively, compared to the VaR of the portfolio constituted of the selected non-military countries. However, it can neither be deduced nor confirmed that the intervention in the Syrian war is the sole source, and more specifically, the trigger of the significant increase in the volatility of the American, Russian, and Chinese stock markets. Answering this question requires further lines of future research that involves incorporating a number of control covariates and using a different modeling methodology.

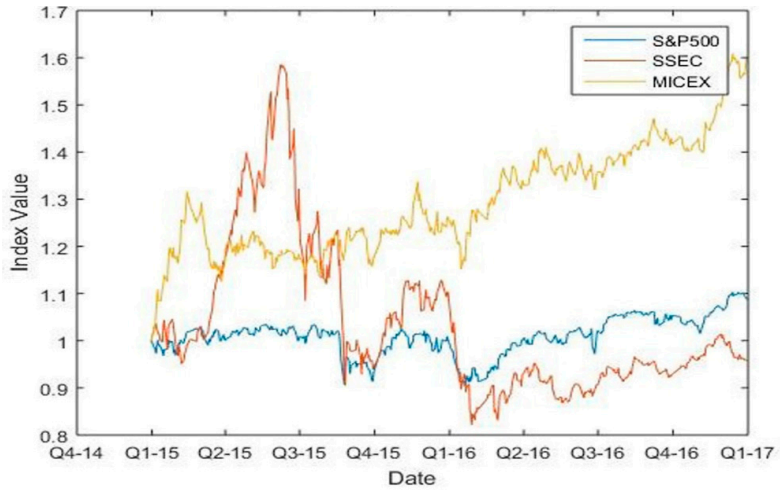
Although we covered a significant number of observations, our results are subject to errors; it is never possible to have enough data when implementing the extreme value analysis, since the tail distribution inference remains less certain. Introducing hypothetical losses to our historical data to generate stress scenarios is of no interest to this study and falls outside its main objective. It would be interesting to repeat the same study with the same selected three military countries during the period 2018–2020, which is expected to be the last phase of the Syrian war given that the Syrian army reached back to the border frontier with Turkey and that the Syrian Constitutional Committee Delegates launched meetings in Geneva to hold talks on the amendment of Syria's constitution.

Author Contributions: Conceptualization, V.N.; methodology, V.N., N.M., and J.-M.M.; formal analysis, J.-M.M.; resources, N.M.; writing—original draft preparation, V.N., N.M., and J.-M.M.; writing—review and editing, V.N., J.-M.M., and R.E.K.; visualization, R.E.K.; supervision, V.N.; project administration, J.-M.M. and V.N.; funding acquisition, J.-M.M. All authors have read and agreed to the published version of the manuscript.

Funding: José-María Montero benefited from the co-funding by the University of Castilla-La Mancha (UCLM) and the European Fund for Regional Development (EFRD) to the Research Group "Applied Economics and Quantitative Methods" (ECOAPP&QM): Grant 2019-GRIN-26913 2019.

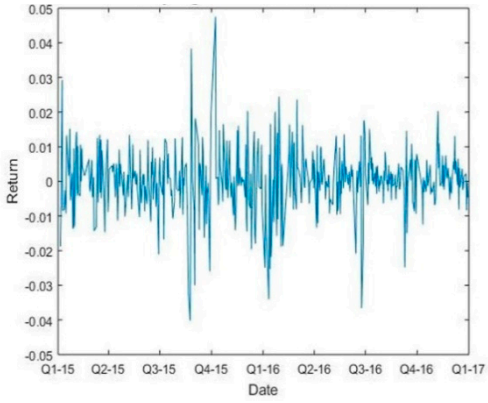
Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

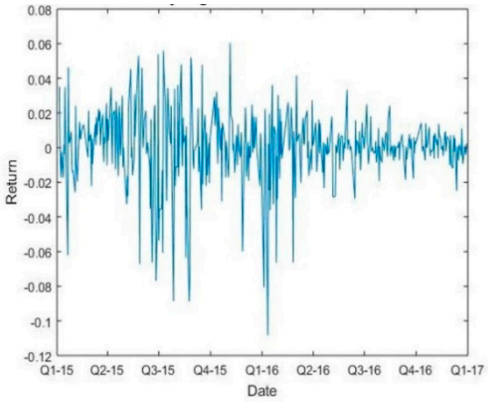


Relative Daily Index Closings of the In-Sample Portfolio

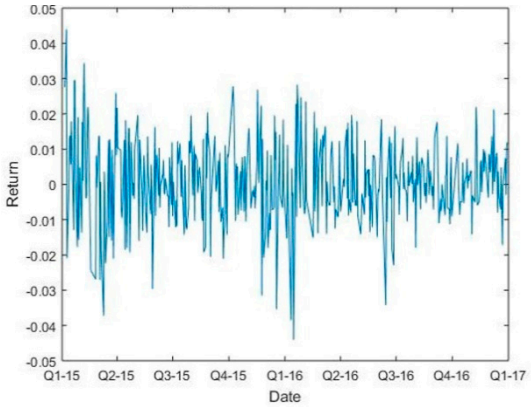
Figure A1. *Cont.*



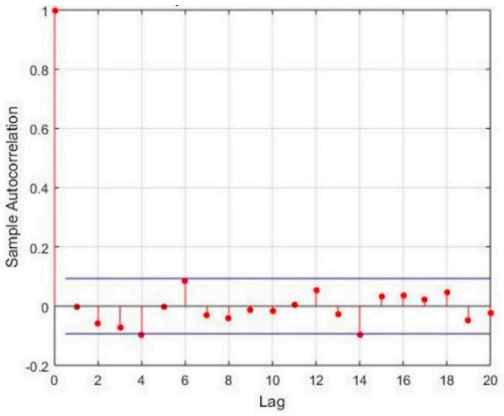
Daily Logarithmic Returns of S&P 500



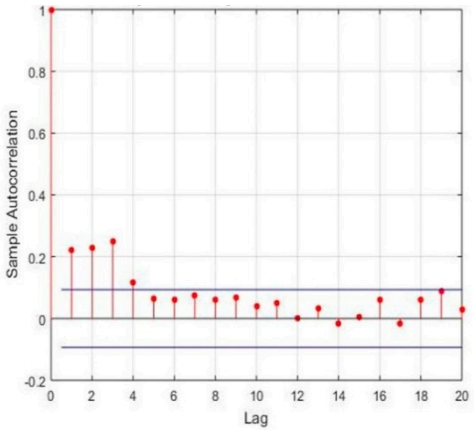
Daily Logarithmic Returns of SSEC



Daily Logarithmic Returns of MICEX

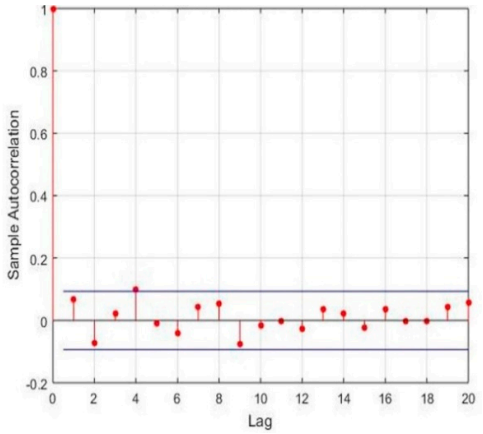


ACF of Returns of S&P 500

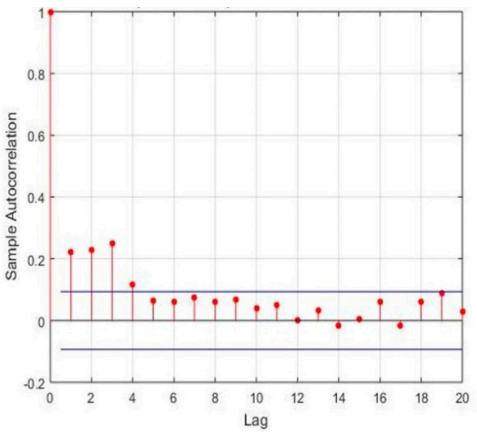


ACF of Squared Returns of S&P 500

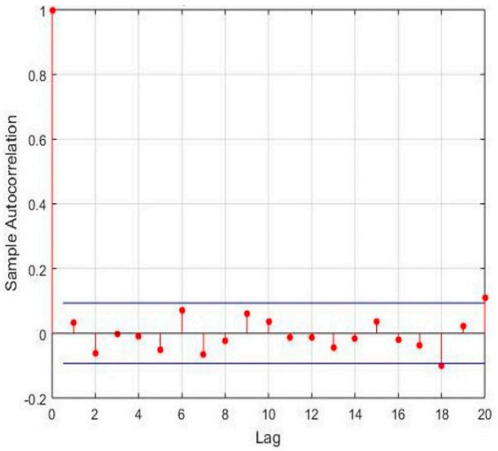
Figure A1. *Cont.*



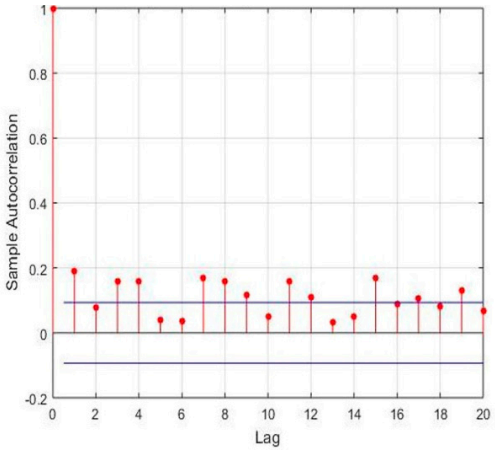
ACF of Returns of SSEC



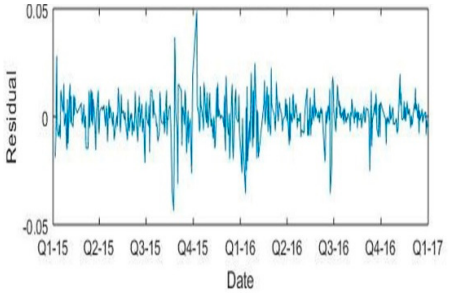
ACF of Squared Returns of SSEC



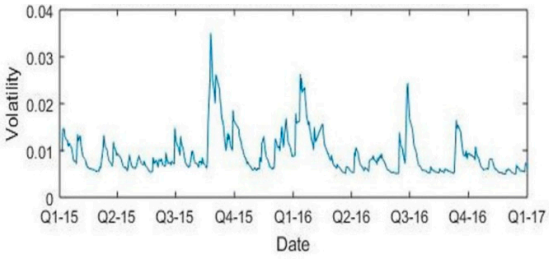
ACF of Returns of MICEX



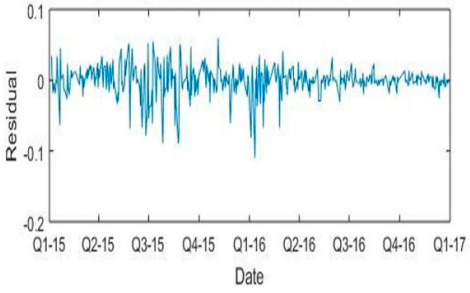
ACF of Squared Returns of MICEX



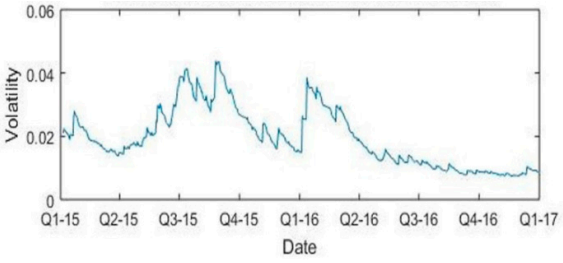
Filtered Residuals of S&P 500



Filtered Conditional Standard Deviation of S&P 500

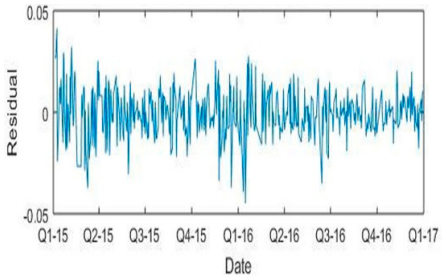


Filtered Residuals of SSEC

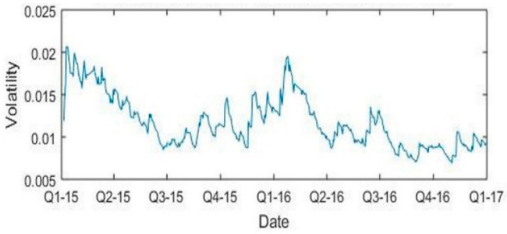


Filtered Conditional Standard Deviation of SSEC

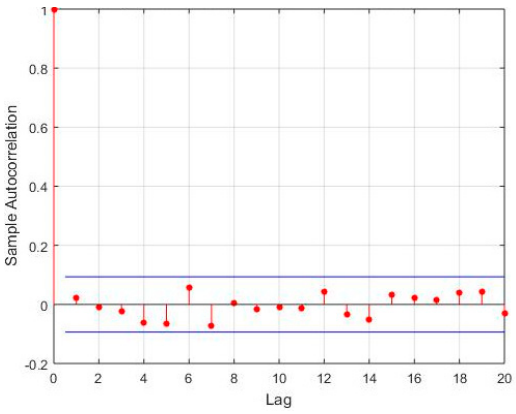
Figure A1. *Cont.*



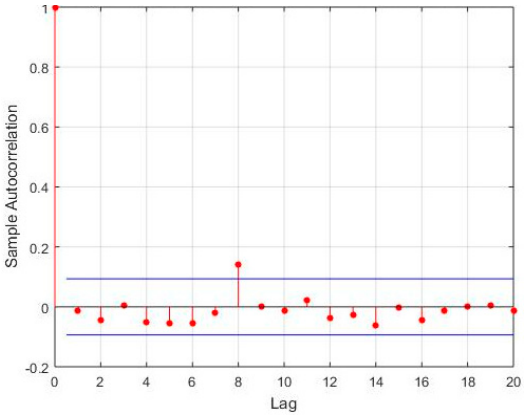
Filtered Residuals of MICEX



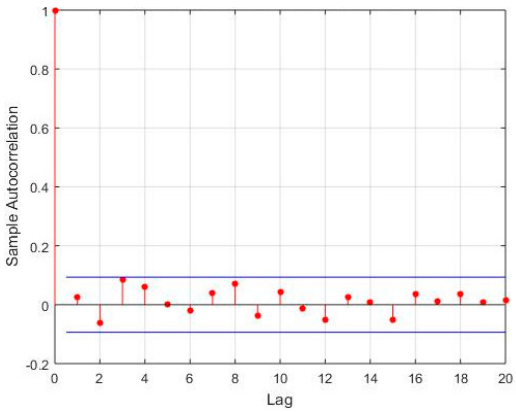
Filtered Conditional Standard Deviation of MICEX



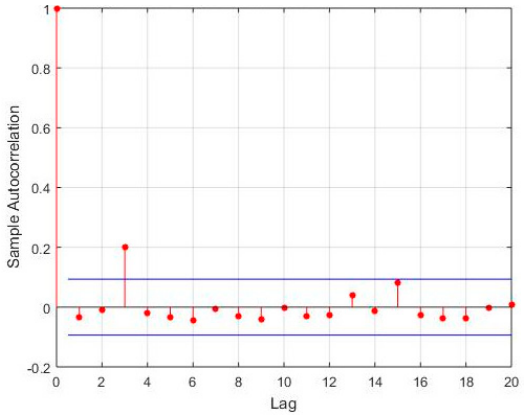
ACF of Standardized Residuals of S&P 500



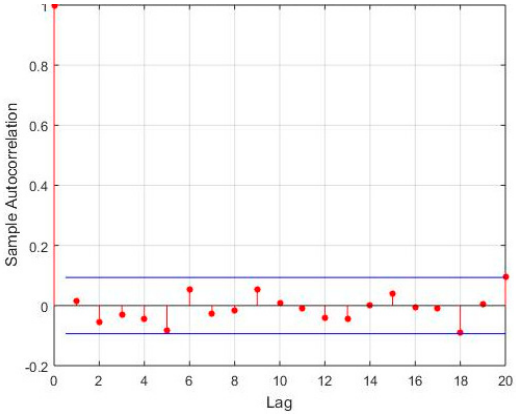
ACF of Squared Standardized Residuals of S&P 500



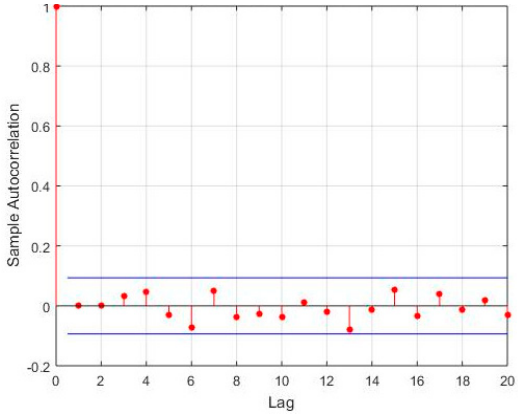
ACF of Standardized Residuals of SSEC



ACF of Squared Standardized Residuals of SSEC

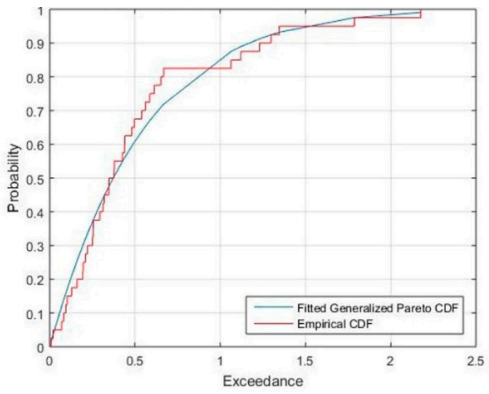


ACF of Standardized Residuals of MICEX

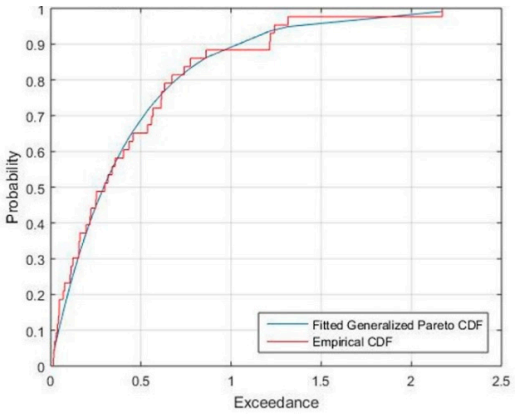


ACF of Squared Standardized Residuals of MICEX

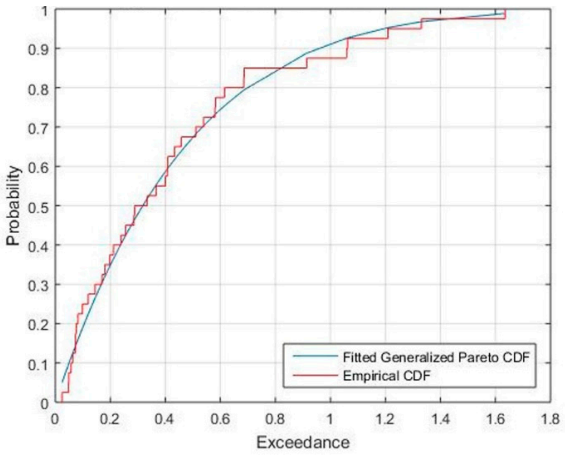
Figure A1. *Cont.*



S&P 500 Upper Tail of Standardized Residuals



SSEC Upper Tail of Standardized Residuals

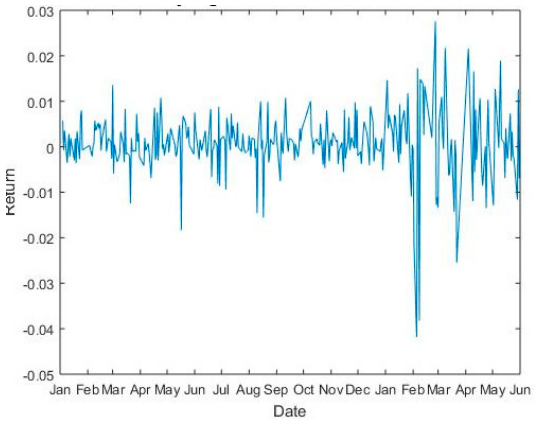


MICEX Upper Tail of Standardized Residuals

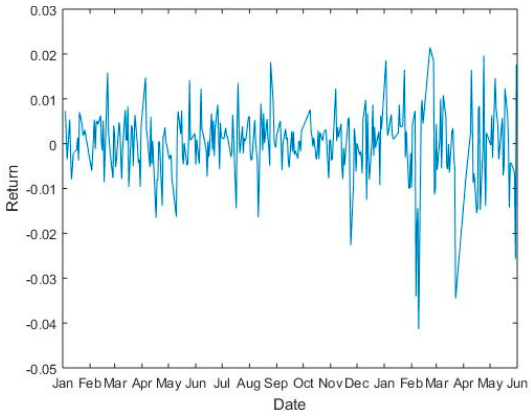
Figure A1. In-Sample VaR Figures.

Relative Daily Index Closings of the Out-of-Sample Portfolio

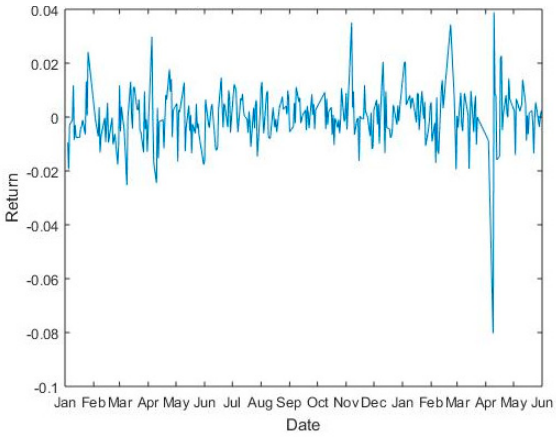
Figure A2. *Cont.*



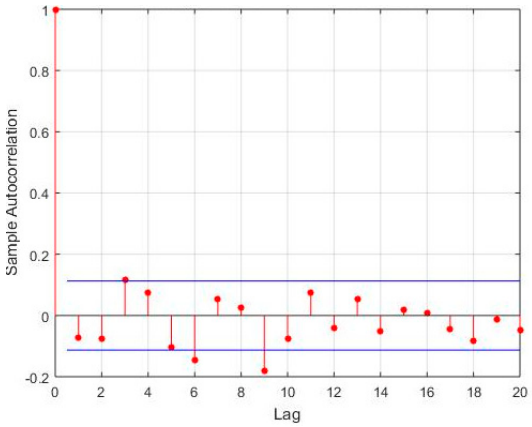
Daily Logarithmic Returns of S&P 500



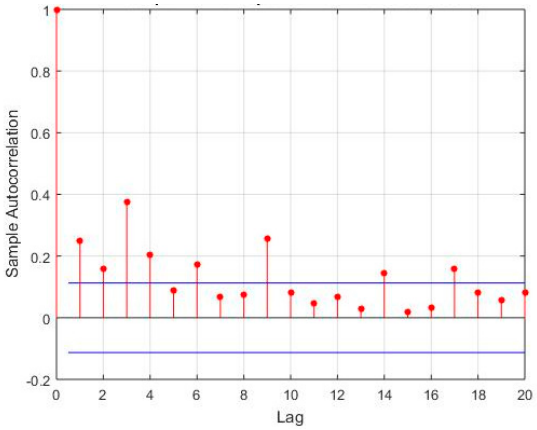
Daily Logarithmic Returns of SSEC



Daily Logarithmic Returns of MICEX

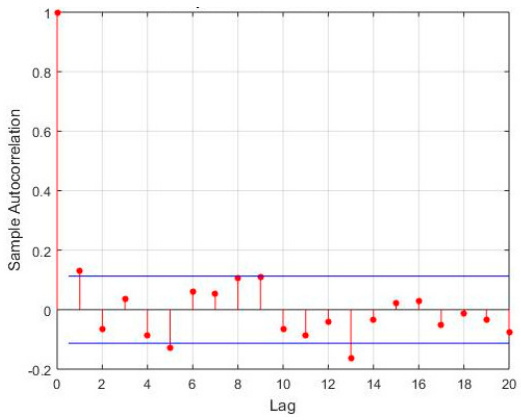


ACF of Returns of S&P 500

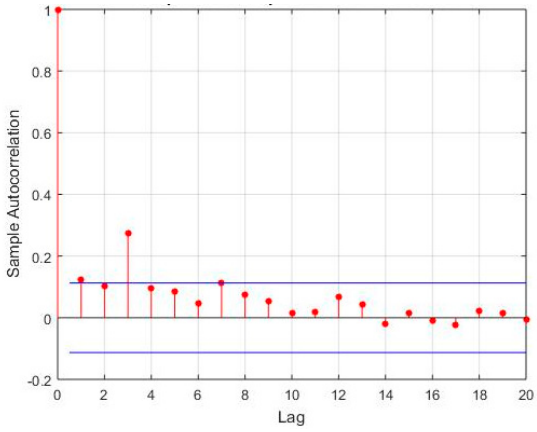


ACF of Squared Returns of S&P 500

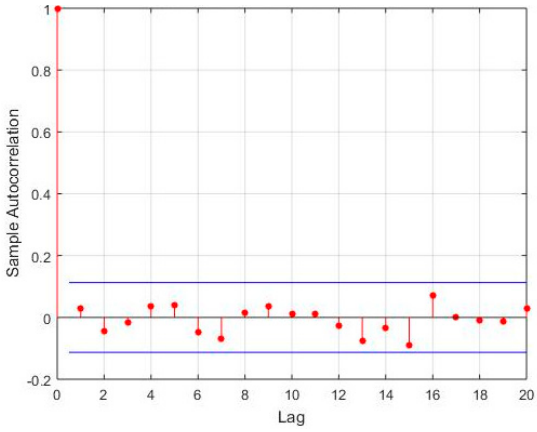
Figure A2. *Cont.*



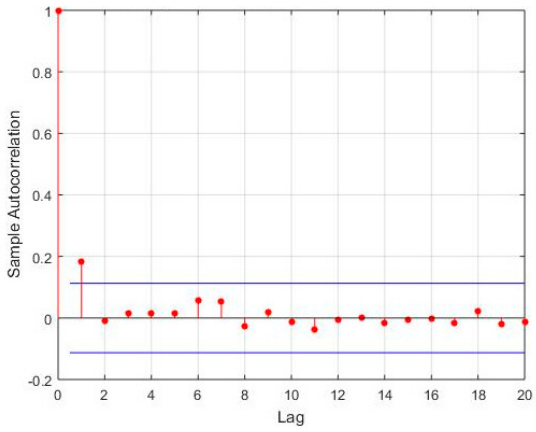
ACF of Returns of SSEC



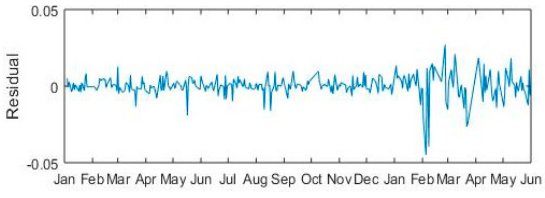
ACF of Squared Returns of SSE



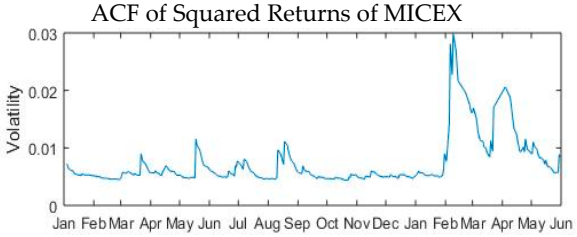
ACF of Returns of MICEX



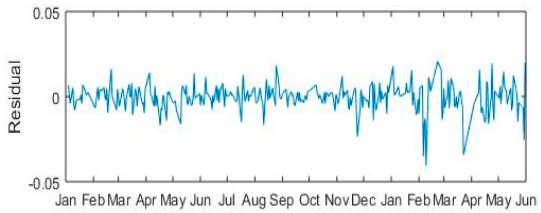
ACF of Squared Returns of MICEX



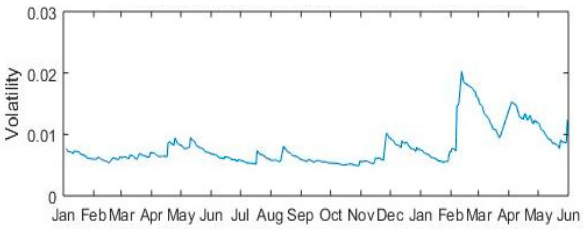
Filtered Residuals of S&P 500



Filtered Conditional Standard Deviation of S&P 500

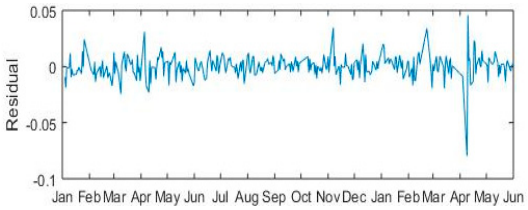


Filtered Residuals of SSEC

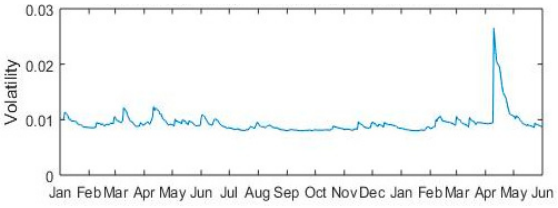


Filtered Conditional Standard Deviation of SSEC

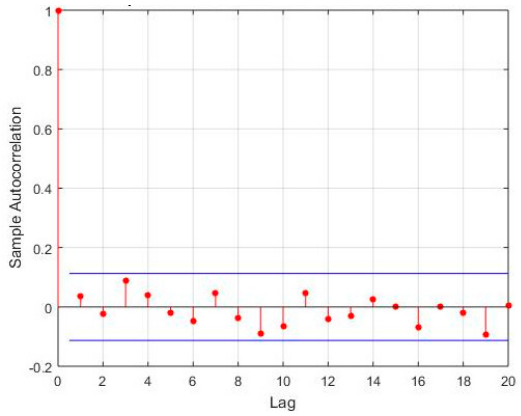
Figure A2. *Cont.*



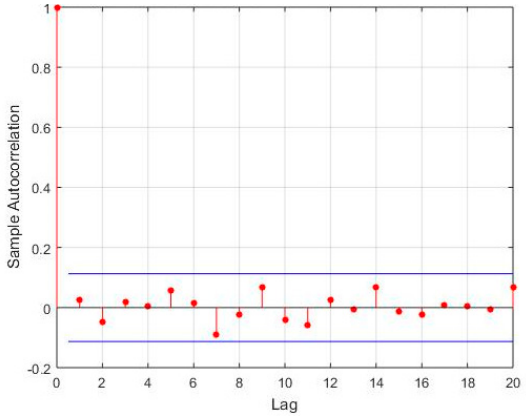
Filtered Residuals of MICEX



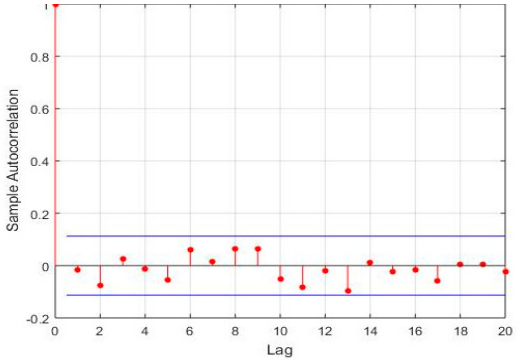
Filtered Conditional Standard Deviation of MICEX



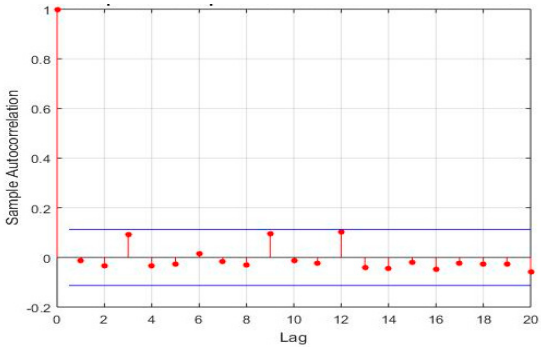
ACF of Standardized Residuals of S&P 500



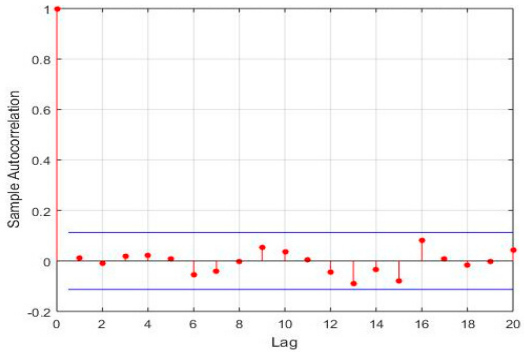
ACF of Squared Standardized Residuals of S&P 500



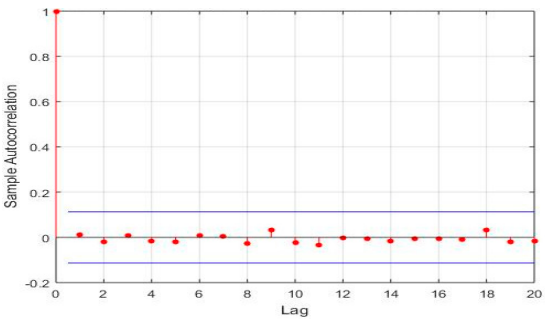
ACF of Standardized Residuals of SSEC



ACF of Squared Standardized Residuals of SSEC

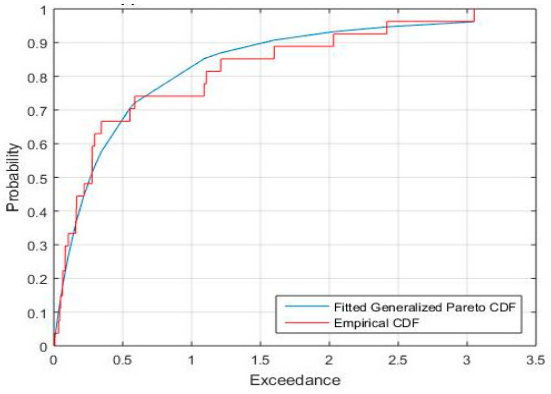


ACF of Standardized Residuals of MICEX

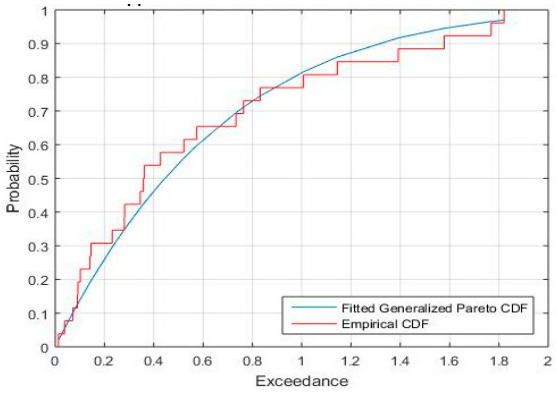


ACF of Squared Standardized Residuals of MICEX

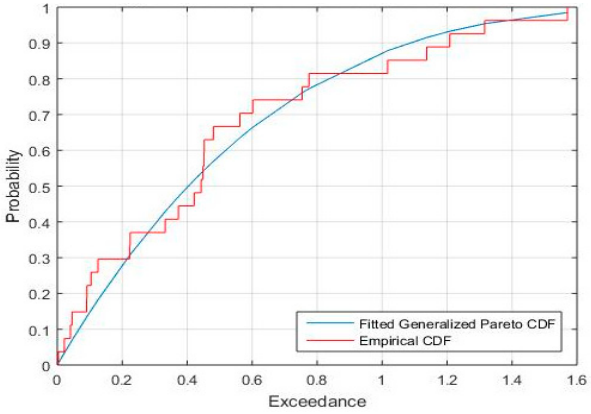
Figure A2. *Cont.*



S&P 500 Upper Tail of Standardized Residuals



SSEC Upper Tail of Standardized Residuals



MICEX Upper Tail of Standardized Residuals

Figure A2. Out-of-Sample VaR Figures.

References

1. Humud, C.E.; Blanchar, C.M.; Nikitin, M.B.D. *Armed Conflict in Syria: Overview and U.S. Response*; CRS: Washington, DC, USA, 2017.
2. Swaine, M. *Chinese Views of the Syrian Conflict*; Carnegie Endowment for International Peace: Washington, DC, USA, 2012.
3. O'Conor, T. China May Be the Biggest Winner of All If Assad Takes over Syria. *Newsweek*, 19 January 2018.
4. SIPRI. *Trends in Military Expenditures, 2018*; SIPRI: Solna, Sweden; Stockholm, Sweden, 2019.
5. SIPRI. *Trends in International Arms Transfers, 2018*; SIPRI: Solna, Sweden; Stockholm, Sweden, 2019.
6. Wei, W. Forecasting Stock Market Volatility with Non-Linear GARCH Models: A Case for China. *Appl. Econ. Lett.* **2002**, *9*, 163–166. [CrossRef]
7. Lin, X.; Fei, F. Long Memory Revisit in Chinese Stock Markets: Based on GARCH-Class Models and Multiscale Analysis. *Econ. Model.* **2013**, *31*, 265–275. [CrossRef]
8. Lin, Z. Modelling and Forecasting the Stock Market Volatility of SSE Composite Index Using GARCH Models. *Future Gener. Comput. Syst.* **2018**, *79*, 960–972. [CrossRef]
9. Wang, Z.; Wu, W.; Chen, C.; Zhou, Y. The Exchange Rate Risk of Chinese Yuan: Using VaR and ES Based on Extreme Value Theory. *J. Appl. Stat.* **2010**, *37*, 265–282. [CrossRef]
10. Chen, Q.; Giles, D.E.; Feng, H. The Extreme-Value Dependence between the Chinese and Other International Stock Markets. *Appl. Financ. Econ.* **2012**, *22*, 1147–1160. [CrossRef]
11. Wei, Y.; Chen, W.; Lin, Y. Measuring Daily Value-at-Risk of SSEC Index: A New Approach Based on Multifractal Analysis and Extreme Value Theory. *Phys. A Stat. Mech. Appl.* **2013**, *392*, 2163–2174. [CrossRef]

12. Hussain, S.I.; Li, S. Modeling the Distribution of Extreme Returns in the Chinese Stock Market. *J. Int. Financ. Mark. Inst. Money* **2015**, *34*, 263–276. [CrossRef]
13. Hou, Y.; Li, S. Information Transmission between U.S. and China Index Futures Markets: An Asymmetric DCC GARCH Approach. *Econ. Model.* **2016**, *52*, 884–897. [CrossRef]
14. Awartani, B.M.A.; Corradi, V. Predicting the Volatility of the S&P-500 Stock Index via GARCH Models: The Role of Asymmetries. *Int. J. Forecast.* **2005**, *21*, 167–183. [CrossRef]
15. Furió, D.; Climent, F.J. Extreme Value Theory versus Traditional GARCH Approaches Applied to Financial Data: A Comparative Evaluation. *Quant. Financ.* **2013**, *13*, 45–63. [CrossRef]
16. Trinidad Segovia, J.E.; Fernández-Martínez, M.; Sánchez-Granero, M.A. A Novel Approach to Detect Volatility Clusters in Financial Time Series. *Phys. A Stat. Mech. Appl.* **2019**, *535*, 122452. [CrossRef]
17. Ramos-requena, J.P.; Trinidad-segovia, J.E.; Sánchez-granero, M.Á. An Alternative Approach to Measure Co-Movement between Two Time Series. *Mathematics* **2020**, *8*, 261. [CrossRef]
18. Naimy, V.Y.; Hayek, M.R. Modelling and Predicting the Bitcoin Volatility Using GARCH Models. *Int. J. Math. Model. Numer. Optim.* **2018**, *8*, 197–215. [CrossRef]
19. Embrechts, P.; Resnick, S.I.; Samorodnitsky, G. Extreme Value Theory as a Risk Management Tool. *N. Am. Actuar. J.* **1999**, *3*, 30–41. [CrossRef]
20. Jorion, P. *Value at Risk: The New Benchmark for Managing Financial Risk*, 3rd ed.; McGraw-Hill: New York, NY, USA, 2007. [CrossRef]
21. Engle, R.F. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica* **1982**, *50*, 987–1007. [CrossRef]
22. Bollerslev, T. Generalized Autoregressive Conditional Heteroskedasticity. *J. Econom.* **1986**, *31*, 307–327. [CrossRef]
23. Nelson, D. Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrica* **1991**, *59*, 347–370. [CrossRef]
24. McNeil, A. Extreme Value Theory for Risk Managers. In *Internal Modelling and CAD II*; Risk Waters Books: London, UK, 1999; pp. 93–113.
25. Pickands, J. Statistical Inference Using Extreme Order Statistics. *Ann. Stat.* **1975**, *3*, 119–131.
26. Balkema, A.A.; de Haan, L. Residual Life Time at Great Age. *Ann. Probab.* **1974**, *2*, 792–804. [CrossRef]
27. Lee, W. *Applying Generalized Pareto Distribution to the Risk Management of Commerce Fire Insurance*; Working Paper; Tamkang University: New Taipei, Taiwan, 2009.
28. Russian Benchmark Officially Renamed the MOEX Russia Index. Available online: https://www.moex.com/n17810 (accessed on 3 April 2020).
29. Peng, Z.X.; Li, S.; Pang, H. *Comparison of Extreme Value Theory and GARCH Models on Estimating and Predicting of Value-at-Risk*; Working Paper; Wang Yanan Institute for Studies in Economics, Xiamen University: Xiamen, China, 2006.
30. Beck, T.; Demirguc-Kunt, A.; Levine, R.E.; Cihak, M.; Feyen, E. *Financial Development and Structure Dataset*. (updated September 2014). Available online: https://www.worldbank.org/en/publication/gfdr/data/financial-structure-database (accessed on 2 April 2020).
31. Market Capitalziation: % of GDP. Available online: https://www.ceicdata.com/en/indicator/market-capitalization--nominal-gdp (accessed on 2 April 2020).
32. Beck, T.; Demirguc-Kunt, A.; Levine, R.E.; Cihak, M.; Feyen, E. *Financial Development and Structure Dataset*. (updated September 2019). Available online: https://www.worldbank.org/en/publication/gfdr/data/financial-structure-database (accessed on 2 April 2020).

mathematics

MDPI

Article

Analysis of a Model for Coronavirus Spread

Youcef Belgaid [1], Mohamed Helal [1] and Ezio Venturino [2,*,†]

[1] Laboratory of Biomathematics, Univ. Sidi Bel Abbes, P.B. 89, Sidi Bel Abbes 22000, Algeria; youcef.belgaid@univ-sba.dz (Y.B.); mohamed.helal@univ-sba.dz (M.H.)

[2] Dipartimento di Matematica "Giuseppe Peano", Università di Torino, via Carlo Alberto 10, 10123 Torino, Italy

* Correspondence: ezio.venturino@unito.it; Tel.: +39-011-670-2833

† Member of the INdAM research group GNCS.

Received: 3 April 2020; Accepted: 15 May 2020; Published: 19 May 2020

Abstract: The spread of epidemics has always threatened humanity. In the present circumstance of the Coronavirus pandemic, a mathematical model is considered. It is formulated via a compartmental dynamical system. Its equilibria are investigated for local stability. Global stability is established for the disease-free point. The allowed steady states are an unlikely symptomatic-infected-free point, which must still be considered endemic due to the presence of asymptomatic individuals; and the disease-free and the full endemic equilibria. A transcritical bifurcation is shown to exist among them, preventing bistability. The disease basic reproduction number is calculated. Simulations show that contact restrictive measures are able to delay the epidemic's outbreak, if taken at a very early stage. However, if lifted too early, they could become ineffective. In particular, an intermittent lock-down policy could be implemented, with the advantage of spreading the epidemics over a longer timespan, thereby reducing the sudden burden on hospitals.

Keywords: dynamical systems; compartment model; epidemics; basic reproduction number; stability

MSC: 92D30; 92D25

1. Introduction

The coronavirus infection has been spreading for a few months. Authorities in several countries have relied on scientific tools for fighting the epidemics. With the lack of a vaccine, distancing methods have been forced on populations to avoid the transmission by direct contact. In laboratories, possible vaccines are being developed, but at the moment they are still at the experimental stage [1]. Meanwhile several models, mathematical, statistical and computer-science-based, are being developed to study the disease and contribute to fighting it.

Models for the spread of epidemics are classic, and an excellent presentation is [2]. In general, the total population is partitioned into at least two classes, susceptibles and infectives, with migrations from the former to the latter by disease propagation through direct or indirect contact, if the disease is transmissible. Additionally, if it can be overcome but causes relapses, the infected can become susceptible again, after maybe going through an intermediate class of being recovered. More sophisticated versions include quarantined and exposed individuals. Some of these classes will be considered also in the present study and illustrated in detail before the model formulation process.

In [3] the disease evolution forecast in several of the most affected countries is attempted, using for that purpose, parameter estimation techniques to calibrate the model. The involved compartments are susceptibles, asymptomatic individuals and symptomatic ones, which in turn are partitioned into reported and unreported cases. In [4] a simple SIRI model is considered, in which the recovered could still contribute to the disease spreading. The model is then extended to account for a possible vaccine,

which, unfortunately, at present is not yet available, although several laboratories worldwide are trying to develop and test it, as mentioned above. In [5] a dynamic model for the diffusion of Covid-19 has been proposed. The transmission network is made by the bats–hosts–reservoir–people compartments; compare also [1]. As it amounts to about 14 differential equations and 25 parameters, it is rather complex. From it, the authors have obtained a simplified version, consisting of six compartments and 13 parameters. Then, the disease basic reproduction number has been calculated.

Our aim here is the mathematical analysis of a slightly modified version of the simpler model in [5]. The most important change accounts for the fact that asymptomatic people may indeed turn into fully symptomatic and infectious individuals. This feature also distinguishes the system introduced here from the one studied in [6], which, however, contains more compartments. The main aim of that study is the forecast of the epidemic spread in various cities in China, considering, additionally, weather data, which finally indicate that higher humidity favors the containment of a coronavirus epidemic. Our focus in the first part of this investigation is the theoretical analysis of the proposed system, and then we perform some preliminary simulations with realistic parameter values. More extended simulations will be devoted to a subsequent study.

The analysis of dynamical systems usually considers the possible equilibria that can be attained, assessing their feasibility and stability, and possible connections between them. For more details on these issues we refer the reader to classical texts, such as [7–9].

The paper is organized as follows. The main findings are outlined in the next section, which also discusses the results of numerical simulations. Section 3 contains an evaluation of their implications under various distancing policies. We formulate the model in Section 4, where we also analyze it mathematically, showing boundedness of the trajectories, establishing an expression for the disease basic reproduction number, finding its equilibria and assessing their local stability, and global stability is established just for the disease-free equilibrium. The section ends with the details on the numerical simulations.

2. Results

2.1. Theoretical Findings

The main analytical findings of this investigation are summarized in Tables 1 and 2. The expressions of B_T, C_T,H_T, D_T and R_0 are given by Equations (3) and (6).

The model (1) allows only three possible equilibria; the disease-free state C_0, where only susceptibles thrive; an equilibrium without symptomatic infected, which occurs only for a very particular case, when the exposed individuals become all asymptomatic infected; and finally, the endemic equilibrium C^*. All these equilibria are locally asymptotically stable, if suitable, rather complicated conditions, hold. Among the endemic and the disease-free equilibrium bistability is impossible, since they are related via a transcritical bifurcation.

Table 1. Equilibria of the model (1) and their feasibility conditions.

Equilibrium	Populations	Feasibility
$C_0 = (S_0, 0, 0, 0, 0, 0)$	$S_0 = \dfrac{\Lambda}{d_p}$	–
$C_I = (S_I, E_I, 0, A_I, R_I)$	$S_I = \dfrac{\Lambda - B_T E_I}{d_p}$	
	$E_I = \dfrac{1}{B_T}\left(\Lambda - \dfrac{d_p B_T C_T H_T}{\beta_I \omega_p' D_T}\right)$	$\Lambda > \dfrac{d_p B_T C_T H_T}{\beta_I \omega_p' D_T}$
	$A_I = \left(\dfrac{\omega_p'}{H_T}\right) E_I$	(resp. $R_0 > 1$)
	$R_I = \dfrac{\gamma_p'}{d_p}\left(\dfrac{\omega_p'}{H_T}\right) E_I$	$\alpha = 1$ and $\xi = 0$

Table 1. *Cont.*

Equilibrium	Populations	Feasibility
$C^* = (S^*, E^*, I^*, A^*, R^*)$	$S^* = \frac{\Lambda - B_T E^*}{d_p}$	$\Lambda > \frac{d_p B_T C_T H_T}{\beta_I \left[(1-\alpha)\omega_p H_T + \alpha\omega'_p D_T\right]}$
	$E^* = \frac{1}{B_T}\left(\Lambda - \frac{d_p B_T C_T H_T}{\beta_I \left[(1-\alpha)\omega_p H_T + \alpha\omega'_p D_T\right]}\right)$	(resp. $R_0 > 1$)
	$I^* = \left(\frac{(1-\alpha)\omega_p H_T + \alpha\omega'_p \xi}{C_T H_T}\right) E^*$	$\alpha \neq 1$ or $\xi \neq 0$
	$A^* = \left(\frac{\alpha\omega'_p}{H_T}\right) E^*$	
	$R^* = \left[\frac{\gamma_p}{d_p}\left(\frac{(1-\alpha)\omega_p H_T + \alpha\omega'_p \xi}{C_T H_T}\right) + \frac{\gamma'_p}{d_p}\left(\frac{\alpha\omega'_p}{H_T}\right)\right] E^*$	

Table 2. Stability conditions of the equilibria of the model (1).

Point	Coefficients	Stability
C_0	$a_2 = B_T + C_T + H_T$	
	$a_1 = H_T[B_T + C_T] + B_T C_T - \beta_I S_0((1-\alpha)\omega_p + k\alpha\omega'_p)$	$\Lambda < \frac{d_p B_T C_T H_T}{\beta_I \left[(1-\alpha)\omega_p H_T + \alpha\omega'_p D_T\right]}$
	$a_0 = B_T C_T H_T - \beta_I S_0 \left((1-\alpha)\omega_p H_T + \alpha\omega'_p D_T\right)$	(resp. $R_0 < 1$)
C_I	$b_3 = \beta_I k A_I + d_p + H_T + B_T + C_T,$	
	$b_2 = [\beta_I k A_I + d_p](H_T + B_T + C_T)$	$\Lambda > \frac{d_p B_T C_T H_T}{\beta_I \omega'_p D_T}$
	$+B_T C_T + H_T(B_T + C_T) - \omega'_p k \beta_I S_I$	
	$b_1 = [\beta_I k A_I + d_p][B_T C_T + H_T(B_T + C_T)]$	(resp. $R_0 > 1$)
	$+H_T B_T C_T - \omega'_p(k d_p + D_T)\beta_I S_I$	
	$b_0 = [\beta_I k A_I + d_p] H_T B_T C_T - d_p \omega'_p D_T \beta_I S^*$	$\alpha = 1$ and $\xi = 0$
C^*	$c_3 = \beta_I(I^* + kA^*) + d_p + H_T + B_T + C_T,$	
	$c_2 = [\beta_I(I^* + kA^*) + d_p](H_T + B_T + C_T)$	$\Lambda > \frac{d_p B_T C_T H_T}{\beta_I \left[(1-\alpha)\omega_p H_T + \alpha\omega'_p D_T\right]}$
	$+B_T C_T + H_T(B_T + C_T) - [\alpha\omega'_p k + (1-\alpha)\omega_p]\beta_I S^*$	(resp. $R_0 > 1$)
	$c_1 = [\beta_I(I^* + kA^*) + d_p][B_T C_T + H_T(B_T + C_T)] + H_T B_T C_T$	$\alpha \neq 1$ or $\xi \neq 0$
	$-[\alpha\omega'_p(k d_p + D_T) + (1-\alpha)\omega_p(d_p + H_T)]\beta_I S^*$	
	$c_0 = [\beta_I(I^* + kA^*) + d_p] H_T B_T C_T$	
	$-d_p[\alpha\omega'_p D_T + (1-\alpha)\omega_p H_T]\beta_I S^*$	

2.2. Simulations Results

We have performed some simulations with the parameter values listed in Table 3, to simulate various implementations of the distancing policy, which actually is in current use in several countries. The simulations may not be fully realistic, but our point is to investigate their qualitative behavior, not to give quantitative forecasts.

Table 3. Parameters values.

Parameter	Value	Parameter	Value
Λ	500	d_p	8.2×10^{-6}
γ_p	1.764	γ'_p	0.6024
ξ	0.1	k	$0.1 \in [0{:}005; 0{:}2]$
μ	0.001	α	$0.15 \in [0.01, 0.3]$
ω_p	0.1	ω'_p	0.1
ν	0		

We look at the influence that the time of starting the restrictive measures has on the disease spread, while keeping fixed the time of their lifting. We next investigate the effect of the time at which the restricting measures are lifted.

Now comparing the results for the start of implementation at $t_1 = 1$ and $t_1 = 10$, and ending them at the same time, it is seen that the earlier the measures are taken, the better it is, because the epidemic's outbreak is kept in check. In Figure 1 the epidemic outbreak starts around time 30, immediately after lifting the restrictions, while in Figure 2 the initial surge before the measures are implemented is damped by their implementation, and after their lifting the outbreak occurs. Both figures use $t_2 = 30$. The same result is seen using $t_2 = 90$ as the time for removing the restrictions; compare Figures 3 and 4. In Figure 3 nothing apparently happens until time 100 because of the restrictions. When they are lifted, the epidemic spreads. In Figure 4 there is a small peak at the onset of the contagion, immediately curbed by the containment measures, lasting as long as they are in use. In spite of their longer implementation, the outbreak occurs nevertheless with the peak at the same time as in Figure 3.

The investigation of different timings for introducing and relaxing the distancing measures shows that a late implementation has no effect, as the peak of the epidemic occurs and then these measures are ineffective, independently of how long they are implemented. An earlier implementation followed by their subsequent lifting leads to a secondary peak at some time later, the occurrence of which seems to be related to the time for which the measures are implemented; the longer the latter, the longer the delay in the secondary outbreak. However, the number of affected people remains the same.

Unfortunately, the result of the simulations indicates that essentially the whole population gets affected by the disease. Only the timings differ, if distancing measures are taken.

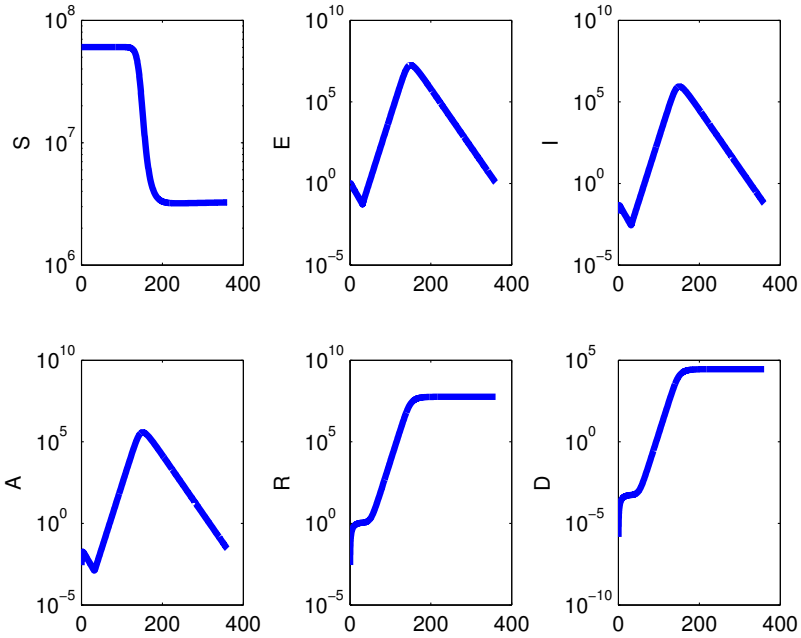


Figure 1. Using a semilogarithmic scale for the vertical axis, we show the results of starting the restrictions at time $t_1 = 1$, using $\beta_I = 10^{-10}$ and lifting them at time $t_2 = 30$, returning to $\beta_I = 10^{-7}$ one month later, over a one year timespan for the model with demographics (1). Left to right and top to bottom, the subpopulations are: S, E, I, A, R and D, the disease-related deceased class.

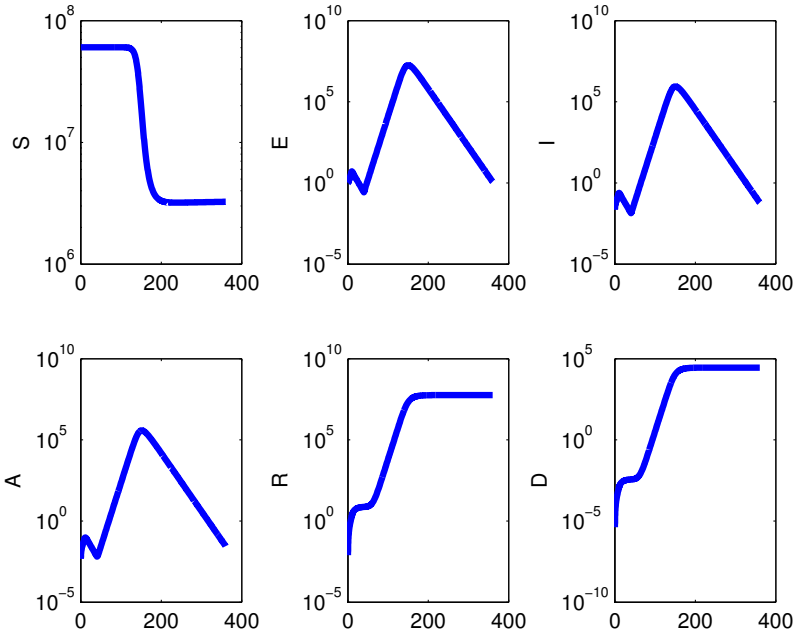


Figure 2. Using a semilogarithmic scale for the vertical axis, we show the results of starting the restrictions at time $t_1 = 10$, using $\beta_I = 10^{-10}$ and lifting them at time $t_2 = 30$, returning to $\beta_I = 10^{-7}$ one month later, over a one year timespan for the model with demographics (1). Left to right and top to bottom, the subpopulations are: S, E, I, A, R and D, the disease-related deceased class.

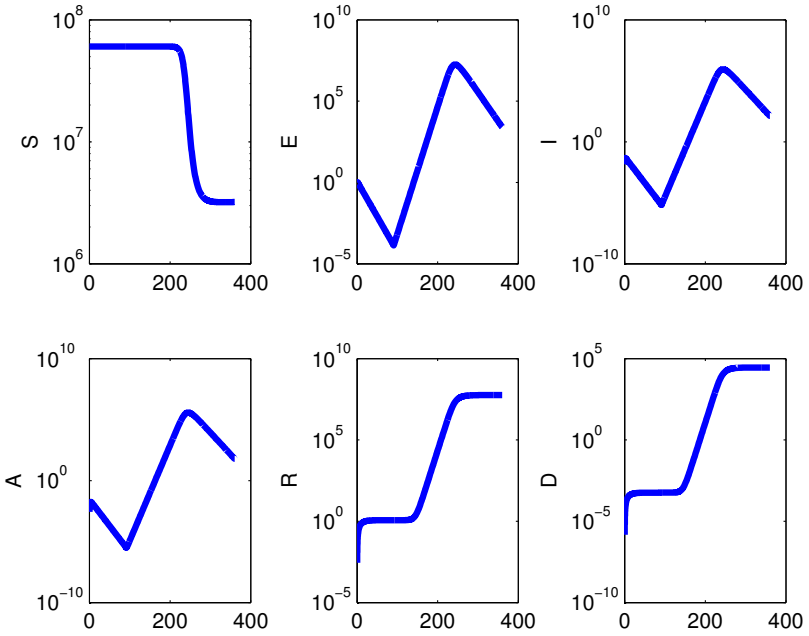


Figure 3. Using a semilogarithmic scale for the vertical axis, we show the results of starting the restrictions at time $t_1 = 1$, using $\beta_I = 10^{-10}$ and lifting them at time $t_2 = 90$, returning to $\beta_I = 10^{-7}$ three months later, over a one year timespan for the model with demographics (1). Left to right and top to bottom, the subpopulations are: S, E, I, A, R and D, the disease-related deceased class.

Thus, if the measures are implemented too late, independently of the time at which they are removed, the outbreak occurs and their subsequent application becomes, therefore, irrelevant, as it cannot be kept in check any longer; compare Figures 2 and 4. On the other hand, by implementing them at the early stages of the contagion process, the outbreak can be delayed, as long as these measures are

implemented, as can be seen from Figures 1 and 3. If they are lifted, the final results of the epidemic's outbreak are essentially the same as if they were not at all implemented, in terms of the number of people being affected by the disease and with possible ultimate fatal consequences; compare the peaks of all the infected classes in Figures 1–5 also with the results in Figure 6 where no measures are taken to prevent the epidemic from spreading.

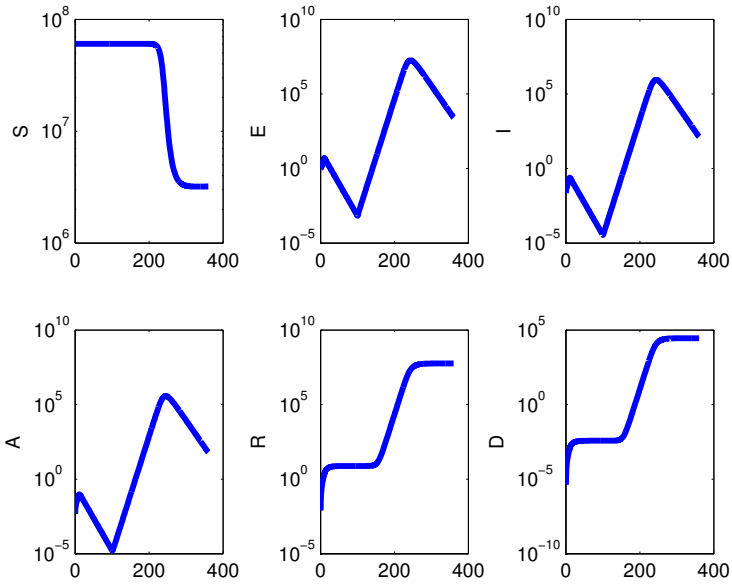


Figure 4. Using a semilogarithmic scale for the vertical axis, we show the results of starting the restrictions at time $t_1 = 10$, using $\beta_I = 10^{-10}$ and lifting them at time $t_2 = 90$, returning to $\beta_I = 10^{-7}$ three months later, over a one year timespan for the model with demographics (1). Left to right and top to bottom, the subpopulations are: S, E, I, A, R and D, the disease-related deceased class.

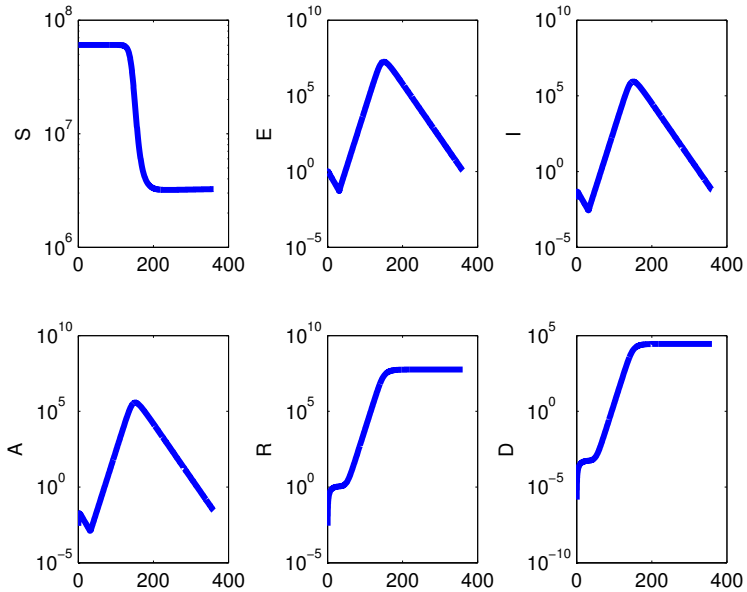


Figure 5. Using a semilogarithmic scale for the vertical axis, we show the results of absolute isolation, starting at time $t_1 = 1$ setting $\beta_I = 0$ and lifting it at time $t_2 = 30$, returning to $\beta_I = 10^{-7}$ one month later, over a one year timespan for the model with demographics (1). Left to right and top to bottom, the subpopulations are: S, E, I, A, R and D, the disease-related deceased class.

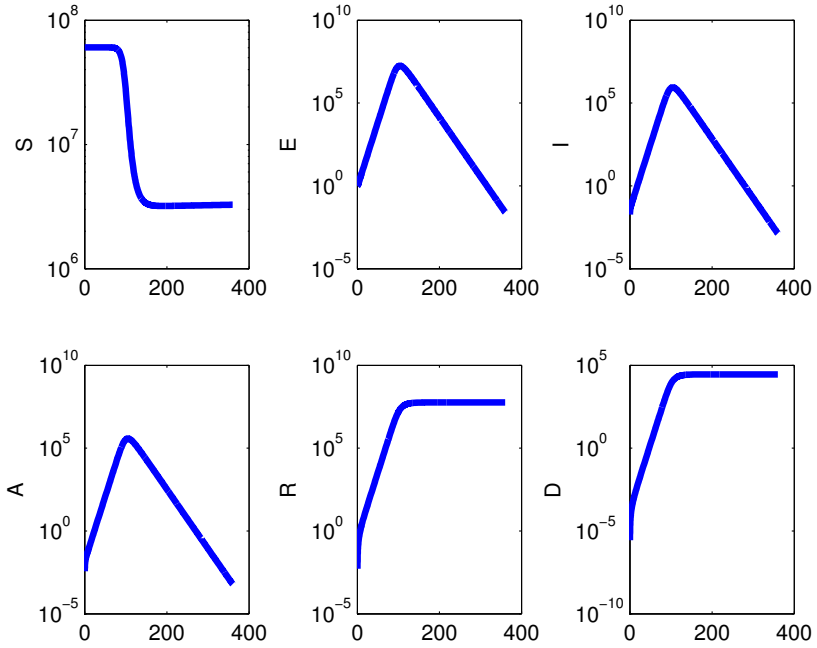


Figure 6. The epidemic's effect on the population in the absence of measures for $\beta_I = 10^{-7}$, on a semilogarithmic scale, over a period of one year. Left to right and top to bottom: S, E, I, A, R and D, the disease-related deceased class.

An intermittent lock-down policy, simulated as an alternative way of coping with the outbreak, might be important to render the burden on hospitalizations smaller, as it tends to spread the epidemic over a longer timespan.

For the particular situation in Italy, note that patient number 1 was diagnosed on 21 February, and the distancing measures in the area were in place starting the following days up to about two weeks later, and then extended to the whole country. Incidentally, patient number 0, the initial carrier of the disease, has never been identified, although there are some speculations. However, in the current news, it is reported that the virus was already circulating yet not known of in Northern Italy in January, which means that additional time had elapsed before the restrictions were applied.

Thus, apparently, these results are negative as for the possibility of containing the spread in the long run, in line with what is hinted in [10], with the exception of the intermittent distancing measures policy, which may spread the epidemic's effects over longer timespans. However, there are some assumptions in the model that make it too crude, so that we plan a deeper subsequent study. In particular, here the results depend on homogeneous mixing, which for a large country is hardly the case. Secondly, this is a continuous model, for which the compartments are depleted only asymptotically. Thus it is not possible to prevent the class of infectives from vanishing in finite time, so that even a small residual in it would start the epidemic's outbreak again. Therefore the somewhat negative results obtained might hopefully be better off in practice. Suitable modifications of the model along these lines will be the subject of a further investigation.

3. Discussion

We have investigated a simple model for the coronavirus pandemic. The steady states, apart from a symptomatic-infected-free point, which is unlikely to exist, are the disease-free equilibrium and the endemic state. The model differs from other current models that are being studied for a few features. From the simplified model that appears in [5], because our formulation contains less equations, it does not consider the viruses compartment, and above all, we allow disease-related mortality, which apparently is missing in the cited paper. Furthermore, we allow the progression of asymptomatic individuals to the class of fully symptomatic. This feature certainly distinguishes it also from [6],

where asymptomatics recover or become diagnosed with the disease, but do not spread it any longer. In the present situation in Italy our assumption is very realistic.

There is no possibility of bistability in our situation, as the two fully meaningful equilibria are related to each other via a transcritical bifurcation. The disease-free equilibrium is also globally asymptotically stable, if it is locally asymptotically stable. An expression for the basic reproduction number is established, with a possibly realistic numerical value [11,12].

The simulations show that containment measures could be effective in delaying the epidemic's outbreaks if taken at a very early stage, but when lifted the outbreaks would occur anyway and affect almost the whole population. However, this last statement should be mitigated by the drawbacks inherent in the model's assumptions, as mentioned in the previous section, thereby leaving hope that in practice it will not occur, if the measures are properly implemented.

We next discuss in detail the various different restriction policies that we have simulated.

3.1. Epidemic with a Lock-Down

In this case, in particular, assuming for the disease transmission coefficient the reference value $\beta_I = 10^{-7}$, we reduce it to $\beta_I = 10^{-10}$ during the interval $[t_1, t_1 + t_2]$ and reinstate the standard value afterwards; we monitor the epidemic's evolution over six months. Figures 1–5 show the results of different choices for t_1 and t_2. Containment measures are effective as long as they are implemented, if they are taken early enough, before the epidemic attains its peak.

Since reducing the transmission by one order of magnitude means that to infect a susceptible with rate β_I, it is necessary for only one infected; with $\beta_I/10$, 10 infected would be necessary. Thus since the lock-down is not perfect, as for instance, some essential activities like food production are still going on, a hypothetical reasonable estimate for the contact reduction is three orders of magnitude. A comparison with a different, milder reduction, $\beta_I = 10^{-8}$ is made, showing essentially no difference in the results, see Figure 7.

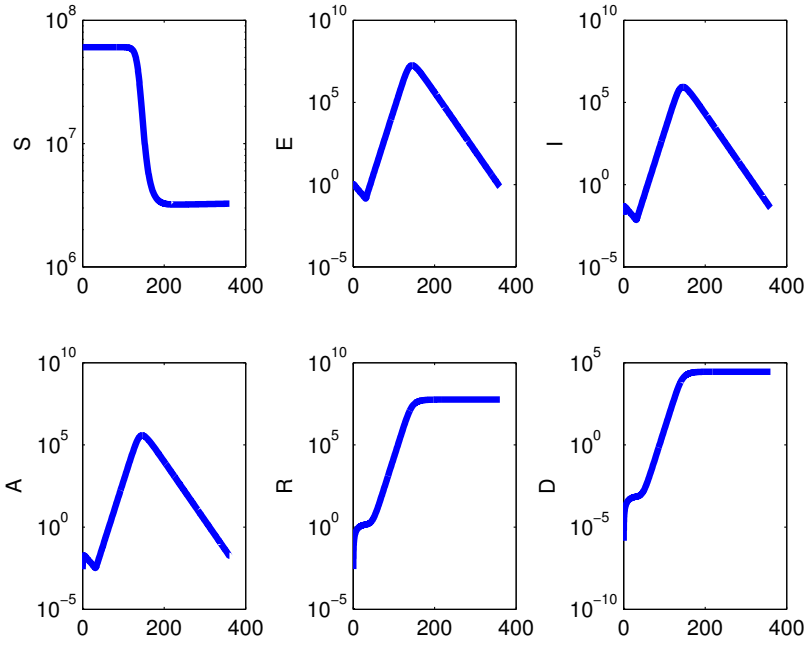


Figure 7. On a semilogarithmic scale, the total populations with the lock-down policy, implemented from time 1 up to time 30, with the milder reduced contact rate $\beta_I = 10^{-8}$, after which $\beta_I = 10^{-7}$ resumes. The simulation runs over a one year timespan for the simplified model (1). Left to right and top to bottom, the subpopulations are: S, E, I, A, R and D, the disease-related deceased class.

3.2. Epidemic with Total Isolation

We changed also the policy to an improbable absolute confinement of every individual in the population, reducing the transmission to exactly zero. The results show no change with respect to those of the lock-down policy. We report only Figure 5, which is identical to Figure 1. The same occurs in the cases contemplated by Figures 2–4.

3.3. The Simplified No-Demographics Model

We repeated the simulations for the model (1) in which we set $\Lambda = d_p = 0$. In the simulations we observed some small changes in the susceptibles behavior, with respect to the full model with vital dynamics. Figures 8 and 9 are the counterparts of the Figures 1 and 2. The ultimate impact of the epidemic is essentially the same; compare in particular, the curves of recovered and deceased. For the total isolation case, Figure 10 shows the same features; compare it with Figure 5. Similar considerations hold for the various remaining cases, and therefore, the pictures are not reported.

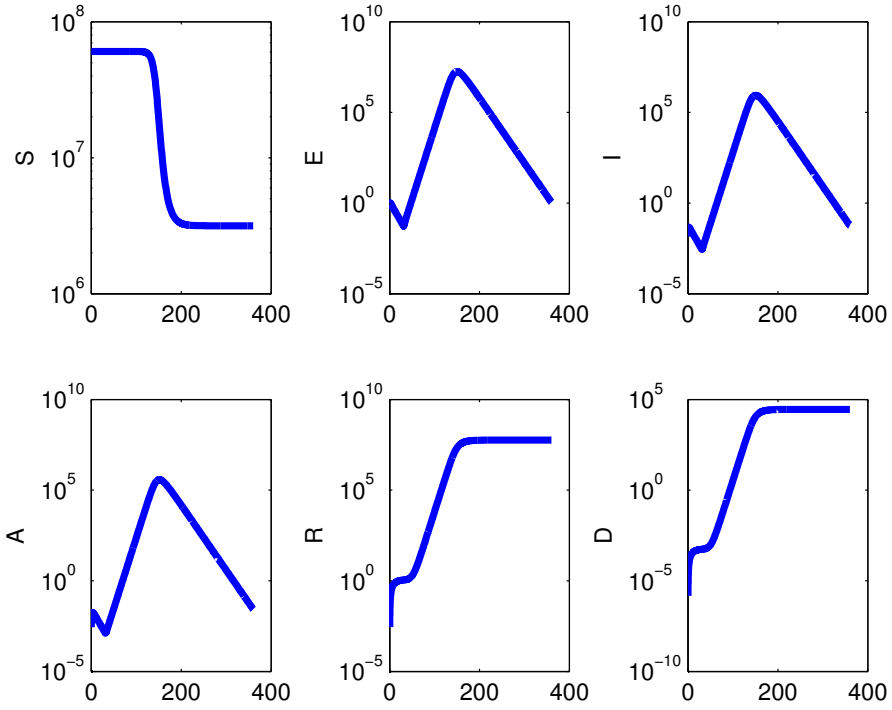


Figure 8. Using a semilogarithmic scale for the vertical axis, we show the results of starting the restrictions at time $t_1 = 1$, setting $\beta_I = 10^{-10}$ and lifting them at time $t_2 = 30$, returning to $\beta_I = 10^{-7}$ one month later, over a one year timespan for the simplified model with no demographics (1) where we take $\Lambda = 0$, $d_p = 0$. Left to right and top to bottom, the subpopulations are: S, E, I, A, R and D, the disease-related deceased class.

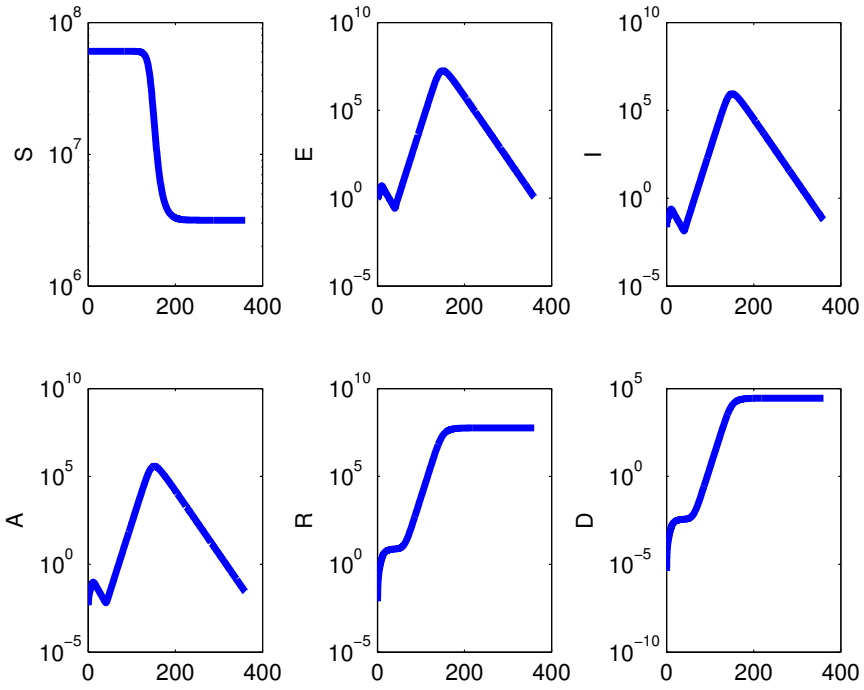


Figure 9. Using a semilogarithmic scale for the vertical axis, we show the results of starting the restrictions at time $t_1 = 10$, setting $\beta_I = 10^{-10}$ and lifting them at time $t_2 = 30$, returning to $\beta_I = 10^{-7}$ one months later, over a one year timespan for the simplified model with no demographics (1) where we take $\Lambda = 0$, $d_p = 0$. Left to right and top to bottom, the subpopulations are: S, E, I, A, R and D, the disease-related deceased class.

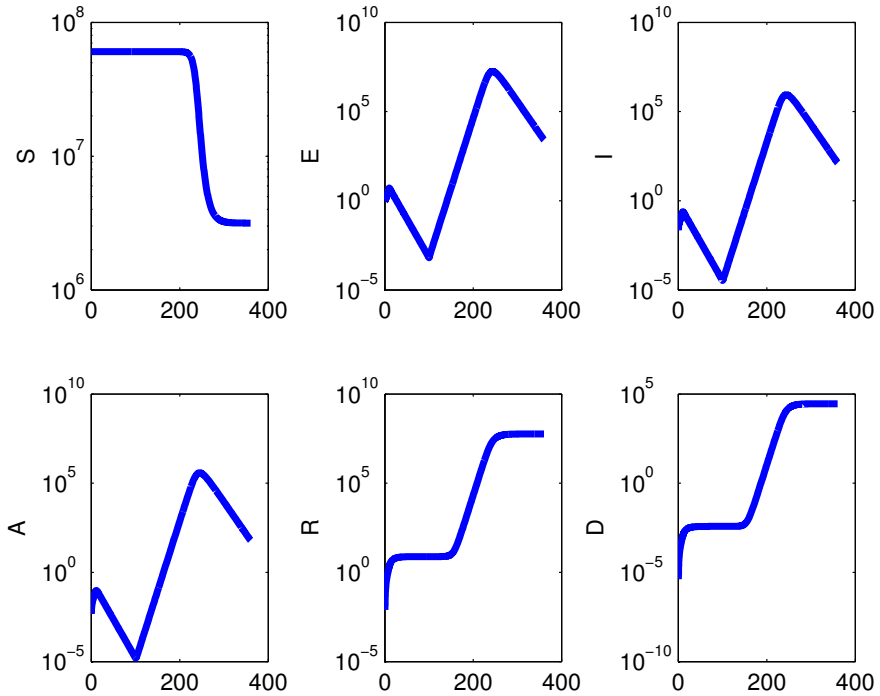


Figure 10. Using a semilogarithmic scale for the vertical axis, we show the results of absolute isolation, $\beta_I = 0$ starting at time $t_1 = 10$, setting $\beta_I = 0$ and lifting it at time $t_2 = 90$, returning to $\beta_I = 10^{-7}$ three months later, over a one year timespan for the model with no demographics (1) where $\Lambda = d_p = 0$. Left to right and top to bottom, the subpopulations are: S, E, I, A, R and D, the disease-related deceased class.

3.4. Investigation of Different Timings for Restrictions' Introduction and Lifting

A further study has been carried out to assess the impact of the time until taking action on the containment measures. All the possible different combinations of simple restriction or total isolation as well as the presence or the absence of demographic effects give essentially the same results. Therefore we present only the results for some selected alternatives, giving the plots in semilogarithmic or total population values, but stressing that for the options not considered, the figures would be the same.

In case the first restriction measure is taken too late, specifically at time $t = 120$, and followed by lifting it either one month or three months later, the epidemic occurs and the measures have no effect whatsoever; see Figure 11, where measures are kept for three months.

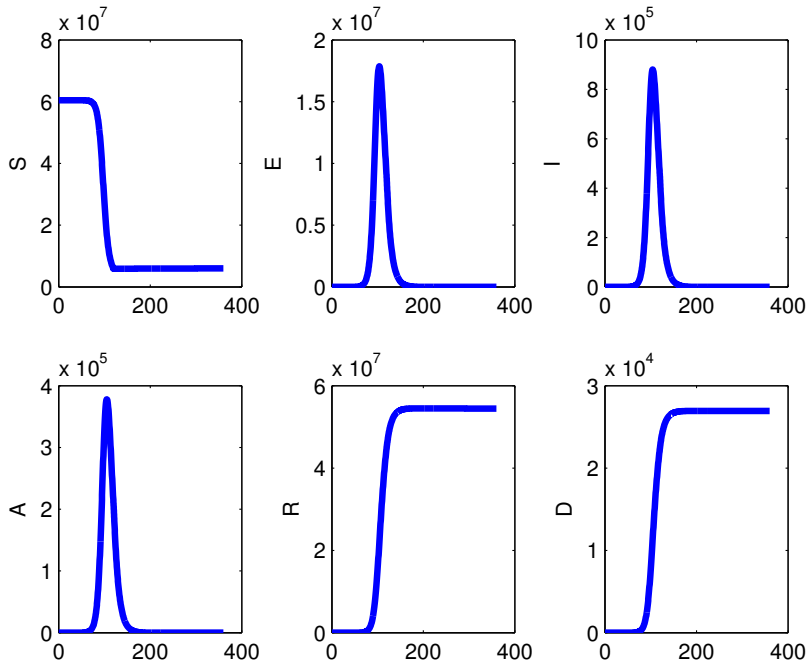


Figure 11. The total populations with the lock-down policy, implemented from time 120 up to time 210, with the reduced contact rate $\beta_I = 10^{-10}$, after which $\beta_I = 10^{-7}$ resumes. The simulation runs over a one year timespan for the model (1). Left to right and top to bottom, the subpopulations are: S, E, I, A, R and D, the disease-related deceased class.

Beginning the restrictions after three months from the start of the epidemic and removing them one month afterwards, causes a second peak about two months later; i.e., six months after the onset of the disease spreading (Figure 12), with a higher number of affected individuals. If instead the lock-down is implemented for three months, the second peak is delayed further, occurring about three months later, Figure 13. Although the pictures are shown on different population scales, absolute values and semilogarithmic, a comparison of the heights of the peaks for the various types of infected subpopulations indicates no difference. Hence, these policies cannot significantly influence the number of people ultimately affected by the disease.

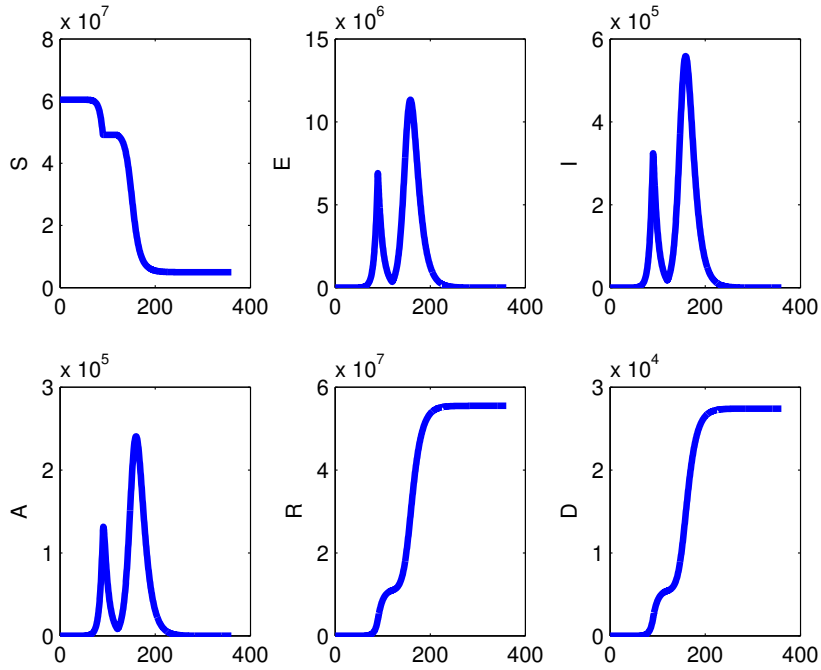


Figure 12. The total populations with the lock-down policy, implemented from time 90 up to time 120, with the total isolation policy $\beta_I = 0$, after which $\beta_I = 10^{-7}$ resumes. The simulation runs over a one year timespan for the simplified model (1) with no demographics, $\Lambda = d_p = 0$. Left to right and top to bottom, the subpopulations are: S, E, I, A, R and D, the disease-related deceased class.

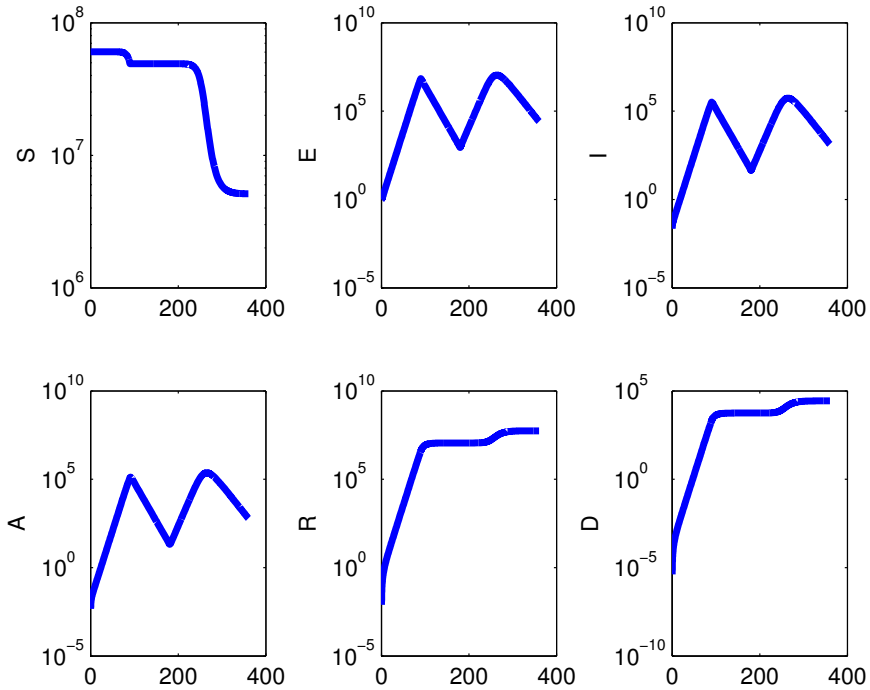


Figure 13. The semilogarithmic plot of the epidemic's spread with the lock-down policy, implemented from time 90 up to time 180, with the reduced contact rate $\beta_I = 10^{-10}$, after which $\beta_I = 10^{-7}$ resumes. The simulation runs over a one year timespan for the model (1) with demographics. Left to right and top to bottom, the subpopulations are: S, E, I, A, R and D, the disease-related deceased class.

3.5. The Intermittent Lock-Down Policy

We finally simulated a policy that attempts to assess the number of infectives at regular times, with a of period one week. If they exceed a threshold, taken to be 10, the lock-down is implemented for a week, and then lifted. Figures 14 and 15 show the results for the case with vital dynamics and in the case of $\Lambda = d_p = 0$. Note that susceptibles in both cases are at a constant value, the vertical scale being extremely small. The infected are kept below the threshold, and the periodic recurrences of the epidemic somewhat change its final impact, as the curves of recovered are reduced by about two orders of magnitude, and above all, the ones of the deceased decrease by about four orders, with respect to the ones found with the one-time lock-down policy. The other relevant change is that here the phenomenon is observed over a longer timespan. Thus the cumulative effects are spread out over a much longer time. This will have some importance to lessening the burden on hospitals. Figure 16 shows the results if the check policy starts immediately at time 1 rather than after a week.

Comparing the population values with the intermittent policy with the one time lock-down, done early enough and implemented for one month, the final outcomes are milder than the latter. Thus the intermittency allows the control of the outbreaks. Susceptibles are almost depleted in the one-time policy; with the intermittent one, however, they are essentially spared from getting the disease; compare Figures 17 and 18.

The intermittency has also been checked with different time intervals. Comparing Figures 19–22, it is seen that the more frequent the checks are implemented, the lower are the peaks in the exposed class, which in turn leads to a smaller cumulative number of recovered and fatalities, at least comparing the policies for the one- and two-weeks alternatives, Figures 19 and 20. For the longer intervals between the checks, again the peaks are higher, the longer the timespan, but it is observed that as time elapses, their heights tend to decrease; see Figures 21 and 22.

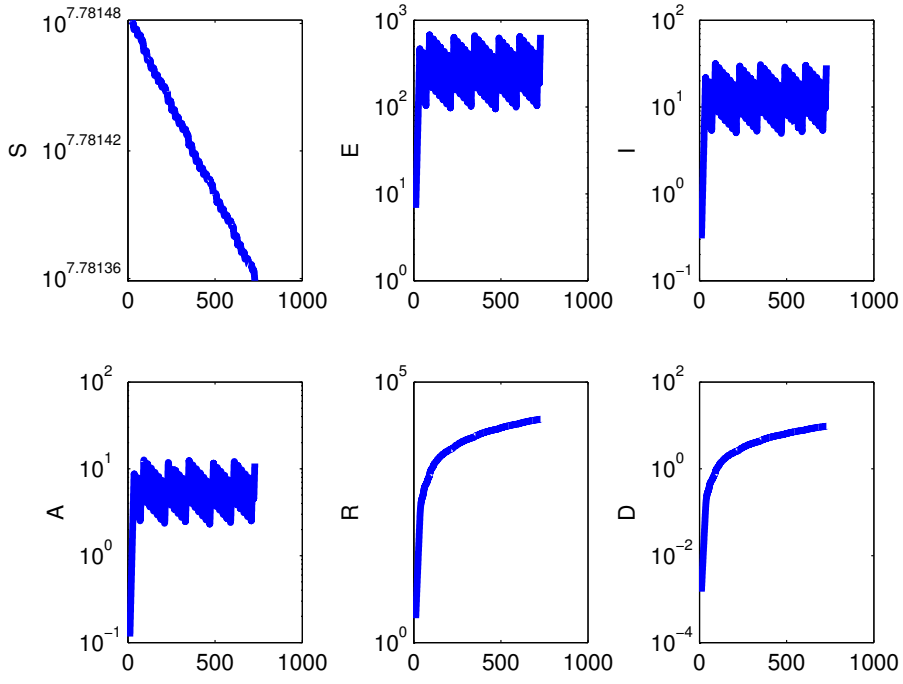


Figure 14. Using a semilogarithmic scale for the vertical axis, we show the results of the intermittent lock-down policy. Here the population is checked every week, starting after a week. If the number of infected is above a small threshold (here taken to be 10) the reduced contact rate $\beta_I = 10^{-10}$ is resumed for a week. The simulation runs over two years timespan for the model with demographics (1). Left to right and top to bottom, the subpopulations are: S, E, I, A, R and D, the disease-related deceased class.

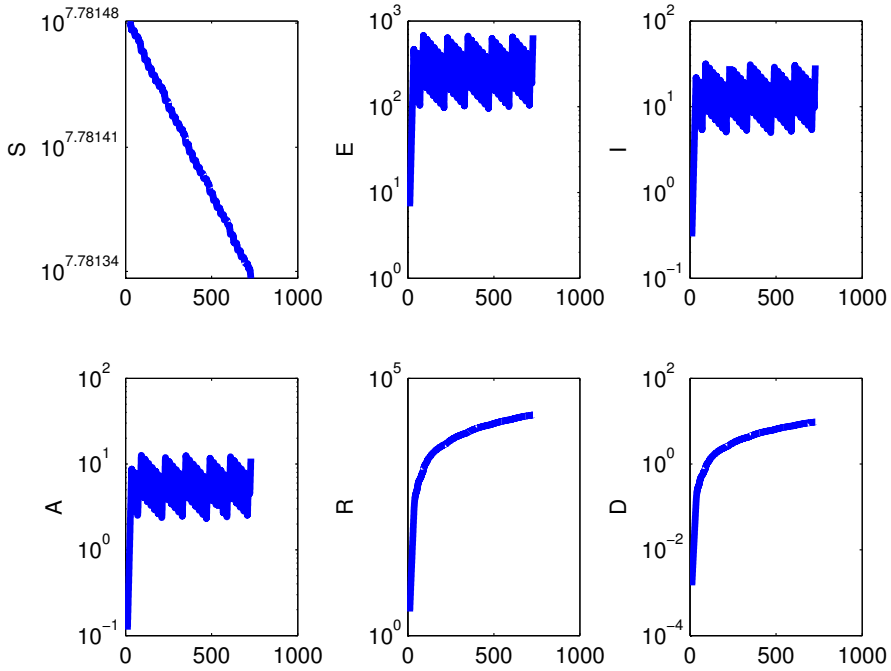


Figure 15. Using a semilogarithmic scale for the vertical axis, we show the results of the intermittent lock-down policy. Here the population is checked every week, starting after a week. If the number of infected is above a small threshold (here taken to be 10) the reduced contact rate $\beta_I = 10^{-10}$ is resumed for a week. The simulation runs over two years timespan for the simplified model with no demographics (1) where we take $\Lambda = 0$, $d_p = 0$. Left to right and top to bottom, the subpopulations are: S, E, I, A, R and D, the disease-related deceased class.

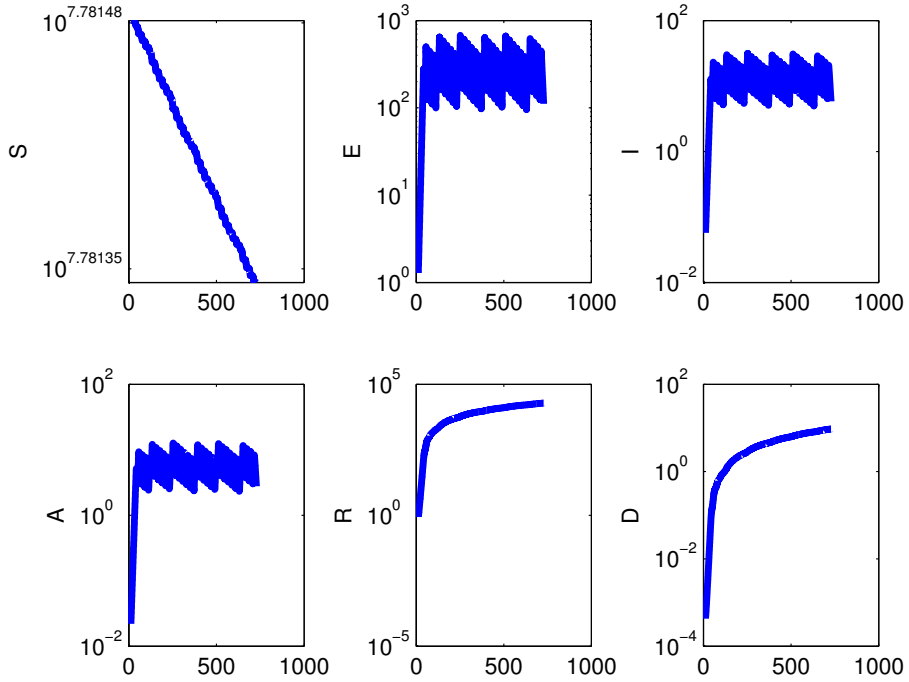


Figure 16. Using a semilogarithmic scale for the vertical axis, we show the results of the intermittent lock-down policy, implemented from time 1. Here the population is checked every week. If the number of infected is above a small threshold (here taken to be 10) the reduced contact rate $\beta_I = 10^{-10}$ is resumed for a week. The simulation runs over two years timespan for the simplified model with no demographics (1) where we take $\Lambda = 0$, $d_p = 0$. Left to right and top to bottom, the subpopulations are: S, E, I, A, R and D, the disease-related deceased class.

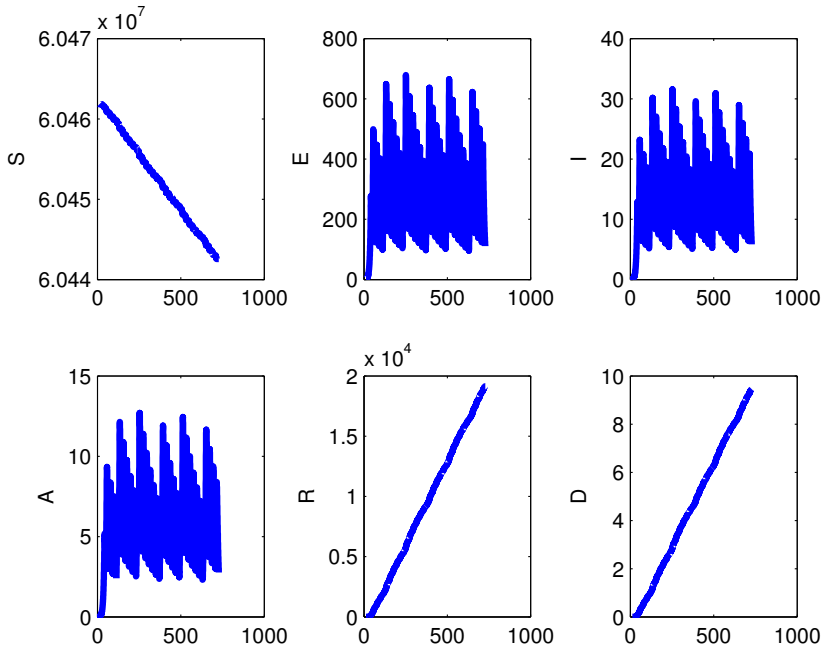


Figure 17. The total populations with the intermittent lock-down policy, implemented from time 1. Here the population is checked every week. If the number of infected is above a small threshold (here taken to be 10) the reduced contact rate $\beta_I = 10^{-10}$ is resumed for a week. The simulation runs over two years timespan for the simplified model with no demographics (1) where we take $\Lambda = 0$, $d_p = 0$. Left to right and top to bottom, the subpopulations are: S, E, I, A, R and D, the disease-related deceased class.

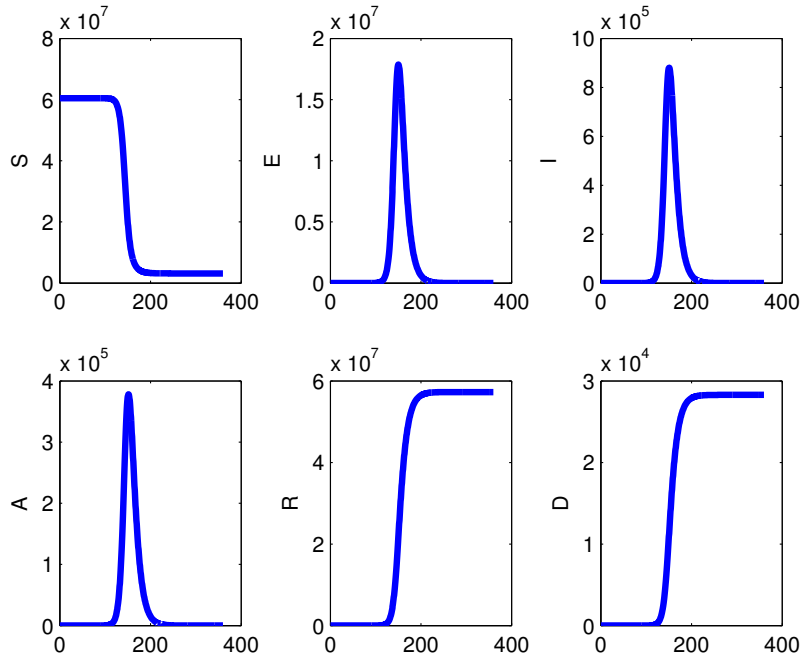


Figure 18. The total populations with the lock-down policy, implemented from time 1 up to time 30, with the reduced contact rate $\beta_I = 10^{-10}$, after which $\beta_I = 10^{-7}$ resumes. The simulation runs over a one year timespan for the simplified model with no demographics (1) where we take $\Lambda = 0$, $d_p = 0$. Left to right and top to bottom, the subpopulations are: S, E, I, A, R and D, the disease-related deceased class.

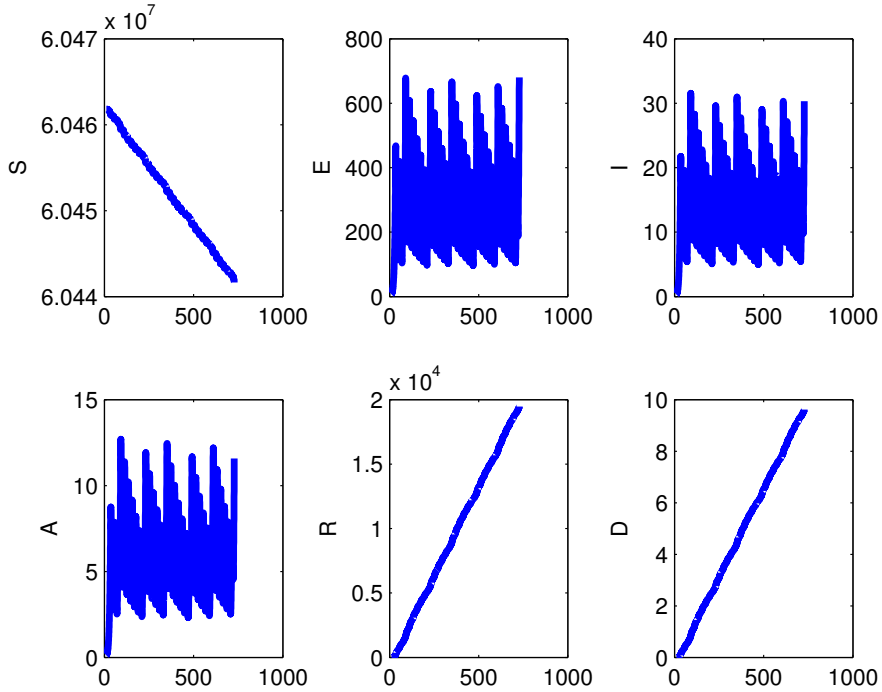


Figure 19. The population values with the lock-down policy, implemented after the first week with the reduced contact rate $\beta_I = 10^{-10}$, after which $\beta_I = 10^{-7}$ resumes. The check for possible repeated implementation is implemented every week afterwards. The simulation runs over a two year timespan for the model (1) with no demographics, i.e., $\Lambda = d_p = 0$. Left to right and top to bottom, the subpopulations are: S, E, I, A, R and D, the disease-related deceased class.

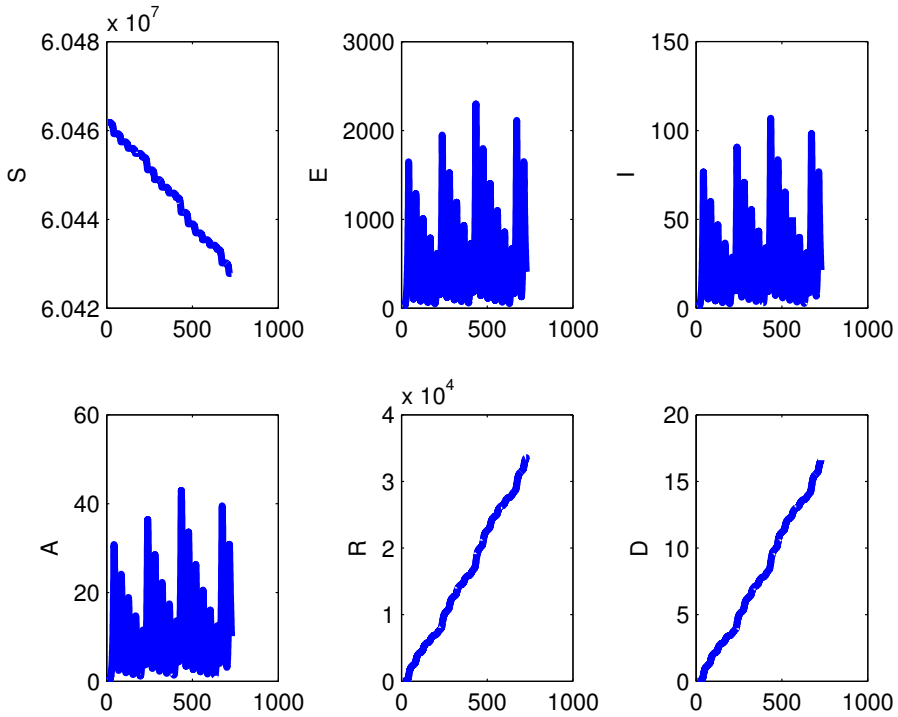


Figure 20. The population values with the lock-down policy, implemented after the first two weeks with the reduced contact rate $\beta_I = 10^{-10}$, after which $\beta_I = 10^{-7}$ resumes. The check for possible repeated implementation is implemented every two weeks afterwards. The simulation runs over a two year timespan for the model (1) with no demographics, i.e., $\Lambda = d_p = 0$. Left to right and top to bottom, the subpopulations are: S, E, I, A, R and D, the disease-related deceased class.

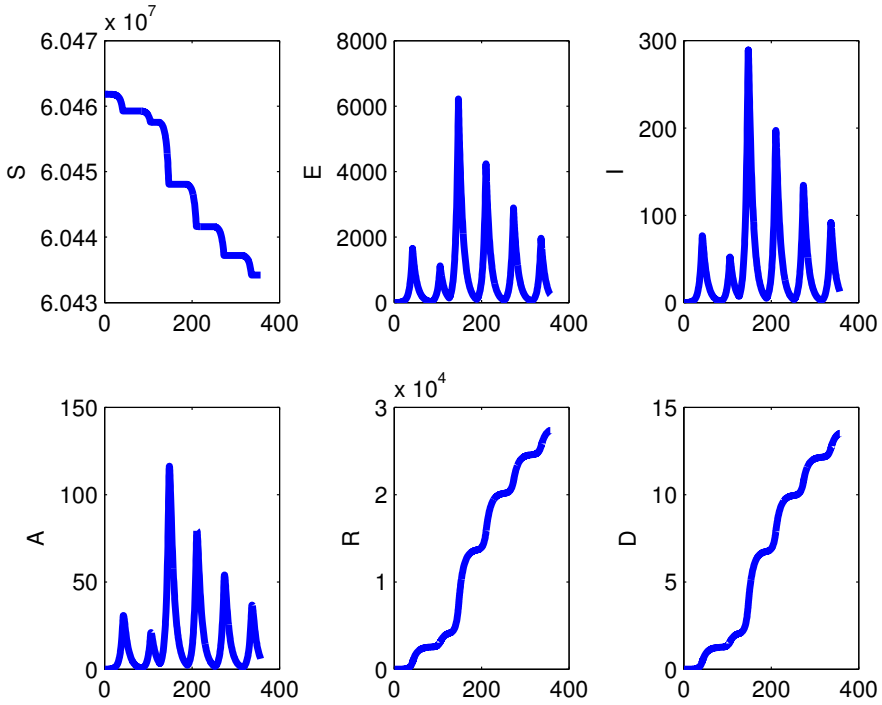


Figure 21. The population values with the lock-down policy, implemented after the first thee weeks with the reduced contact rate $\beta_I = 10^{-10}$, after which $\beta_I = 10^{-7}$ resumes. The check for possible repeated implementation is implemented every three weeks afterwards. The simulation runs over a two year timespan for the model (1) with no demographics, i.e., $\Lambda = d_p = 0$. Left to right and top to bottom, the subpopulations are: S, E, I, A, R and D, the disease-related deceased class.

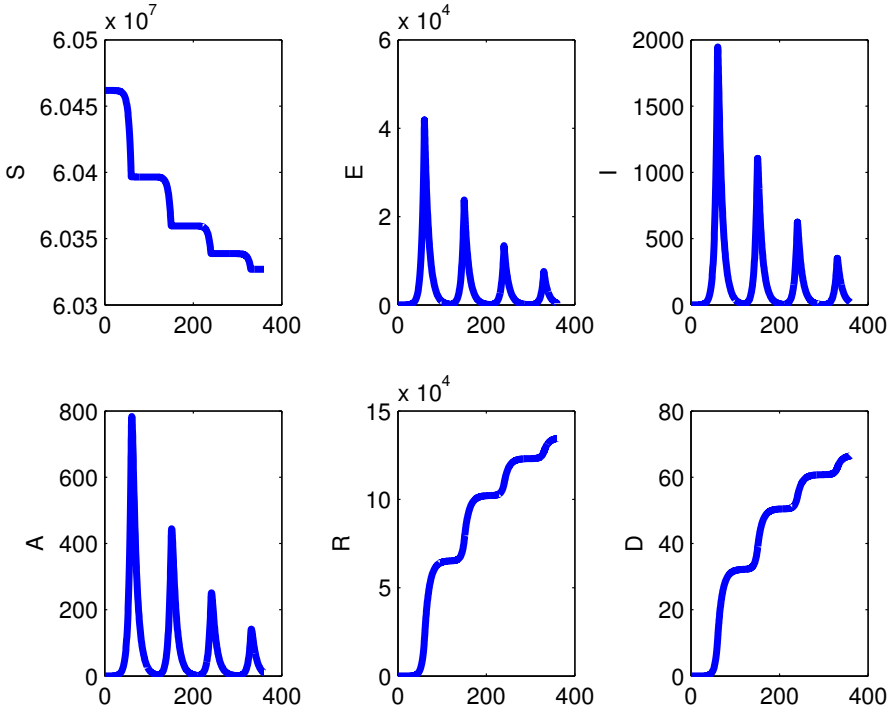


Figure 22. The population values with the lock-down policy, implemented after the first month with the reduced contact rate $\beta_I = 10^{-10}$, after which $\beta_I = 10^{-7}$ resumes. The check for possible repeated implementation is implemented every month afterwards. The simulation runs over a two years timespan for the model (1) with no demographics, i.e., $\Lambda = d_p = 0$. Left to right and top to bottom, the subpopulations are: S, E, I, A, R and D, the disease-related deceased class.

4. Materials and Methods

Here we develop a mathematical model of coronavirus, which is a zoonotic disease. Its biological characteristics indicate that the virus transmission occurred first from infected animals to humans [5], and then spread among populations worldwide by contact with infected individuals, to make it a pandemic.

Let $N(t)$ denote the total population. It is partitioned into the following five disjoint classes of individuals:

$S(t)$: The susceptible class, the individuals who have not yet been exposed to the virus.

$E(t)$: The exposed class, describing people who have become in contact with the virus, but are in incubation period and not yet able to spread the disease; possible presymptomatic individuals that can transmit the infection [13–15] are assumed to have already moved to the asymptomatic class defined below.

$I(t)$: The symptomatic infectious class, individuals that manifest symptoms and can spread the disease.

$A(t)$: The asymptomatic infectious class; those persons that can spread the disease even without having explicit symptoms.

$R(t)$: The removed class, that includes the people that recovered from the disease.

Thus, $N(t) = S(t) + E(t) + I(t) + A(t) + R(t)$.

The basic mechanisms underlying the model are shown in Figure 23. The model is formulated taking into account all the possible interactions among the compartments that were described above.

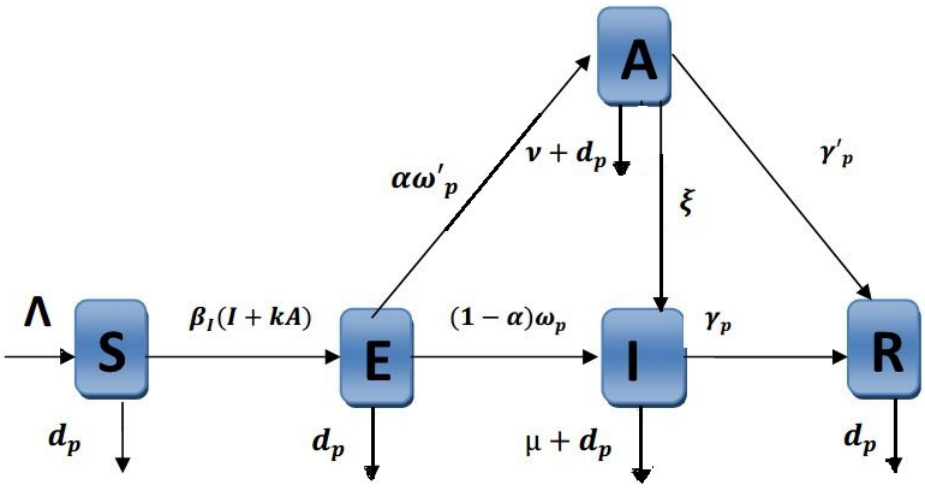


Figure 23. The basic interactions among the compartments.

Under the quasi-steady-state assumption of the total human population, we impose that susceptible individuals are recruited at the constant rate Λ, become infected by direct contact with an infectious individual at rate β_I, which is scaled by a factor k to account for the possibility that the latter is asymptomatic. Finally, all human individuals are subject to natural mortality d_p. These considerations are incorporated in the first equation of the system (1).

Individuals that contract the disease are accounted for in the second equation of (1). They become exposed, i.e., they cannot yet spread the virus, which needs an incubation period within the body of its hosts. In this class enter the susceptibles that were contaminated in the two ways described earlier. People leave it by becoming infectious, and either showing symptoms, thereby migrating into class I, or not, therefore, finding themselves in class A. The progression rates into these two classes are ω_p and ω'_p. Furthermore, we assume that a fraction α becomes asymptomatic and $1 - \alpha$ instead will manifest symptoms.

The third equation models the symptomatic infectious, recruited from the exposed class at rate $(1 - \alpha)\omega$ as described above. Furthermore, there could be asymptomatic individuals that become

symptomatic at rate ξ. They leave this class by either progressing to the recovered class at rate γ_p, or dying, naturally or by causes related to the disease at rate μ.

The asymptomatic individuals modeled in the fourth equation appear from the exposed ones, and leave the class by overcoming the disease at rate γ'_p, dying naturally or by disease-related causes at rate ν, or eventually showing the symptoms, for which they migrate into class I.

Recovered individuals are those that have healed from the disease. They are subject only to natural mortality. We assume that they have also become immune so that they are unaffected if become in contact with the infectious.

Note that in the simulations also the cumulative class of disease-related deceased people is shown, although the dead are not explicitly accounted for in the model. They indeed represent a sink, and thus do not contribute to the disease propagation. Incidentally, instead, in cultures where the deceased are kept for a while before burial and become in contact with the relatives, it may be necessary to introduce this class in the model, as another potential source of infection.

Taking into account the above considerations, the model dynamics is regulated by the following system of nonlinear ordinary differential equations:

$$\begin{aligned}
\frac{dS}{dt} &= \Lambda - \beta_I S(I + kA) - d_p S, \\
\frac{dE}{dt} &= \beta_I S(I + kA) - (1-\alpha)\omega_p E - \alpha\omega'_p E - d_p E, \\
\frac{dI}{dt} &= (1-\alpha)\omega_p E - (\gamma_p + d_p + \mu) I + \xi A, \\
\frac{dA}{dt} &= \alpha\omega'_p E - (\gamma'_p + d_p + \nu) A - \xi A, \\
\frac{dR}{dt} &= \gamma_p I + \gamma'_p A - d_p R,
\end{aligned} \tag{1}$$

or alternatively, excluding completely the demographic features, by setting $\Lambda = 0$ and $d_p = 0$ in (1). All the parameters are nonnegative and their meaning is summarized in Table 4. Note that in view of the definitions,

$$\frac{1}{\omega_p}, \quad \frac{1}{\omega'_p} \quad \frac{1}{\gamma_p}, \quad \frac{1}{\gamma'_p},$$

represent respectively the incubation period before manifesting symptoms, the latent period before becoming asymptomatic infectious, the infectious period for symptomatic infection and the infectious period for asymptomatic infection.

Table 4. Model parameters and their meaning.

Λ	susceptibles recruitment rate
d_p	natural mortality
β_I	disease transmission rate
k	transmissibility ratio between asymptomatics and symptomatics
μ	disease-related mortality for infected
ν	disease-related mortality for asymptomatics
ω_p	progression rate from exposed to symptomatic
ω'_p	progression rate from exposed to asymptomatic
α	fraction of exposed that turn asymptomatic
ξ	progression rate from asymptomatic to symptomatic
γ_p	recovery rate from symptomatic infection
γ'_p	recovery rate from asymptomatic infection

Theorem 1. *The system trajectories are bounded. Letting*

$$M = \max\left\{N(0), \frac{\Lambda}{d_p}\right\}$$

the set

$$\Gamma = \{(S, E, I, A, R) : S + E + I + A + R \leq M, \quad S > 0,\ E \geq 0,\ I \geq 0,\ A \geq 0,\ R \geq 0\}. \tag{2}$$

represents their ultimate attractor. In particular, if $N(0) < \Lambda d_p^{-1}$, $M = \Lambda d_p^{-1}$.

Proof. From the system (1) it follows that the total population evolves as follows:

$$\frac{dN}{dt} + d_p N = \Lambda - \nu A - \mu I \leq \Lambda.$$

Solving the differential inequality easily gives

$$N(t) \leq N(0)\exp(-d_p t) + \frac{\Lambda}{d_p}[1 - \exp(-d_p t)] \leq M,$$

so that all subpopulations, being nonnegative, are bounded as well. □

Note that Γ is positively invariant since all solutions of system (1) originating in Γ remain there for all $t > 0$, in view of the existence and uniqueness of its solutions.

4.1. System's Equilibria Assessment

The equilibrium points of the model are obtained by equating the right hand side of system (1) to zero. The solution of the so-obtained algebraic system gives three equilibrium points: the coronavirus-free equilibrium $C_0 = (S_0, 0, 0, 0, 0,)$, the coronavirus-symptomatic-infected-free equilibrium $C_I = (S_I, E_I, 0, A_I, R_I)$ with conditions $\alpha = 1$ and $\xi = 0$, and the fully coronavirus endemic equilibrium $C^* = (S^*, E^*, I^*, A^*, R^*)$ when either $\alpha \neq 1$ or $\xi \neq 0$. Specifically, for the former two we have:

$$S_0 = \frac{\Lambda}{d_p}, \quad E_I = \frac{1}{B_T}\left(\Lambda - \frac{d_p B_T C_T H_T}{\beta_I \omega_p' D_T}\right), \quad S_I = \frac{\Lambda - B_T E^*}{d_p}, \quad A_I = \left(\frac{\omega_p'}{H_T}\right) E_I, \quad R_I = \left(\frac{\gamma_p'}{d_p}\frac{\omega_p'}{H_T}\right) E_I,$$

where

$$B_T = (1-\alpha)\omega_p + \alpha\omega_p' + d_p, \quad C_T = \gamma_p + \mu + d_p, \quad D_T = \xi + k(\gamma_p + \mu + d_p), \quad H_T = \gamma_p' + \nu + \xi + d_p. \tag{3}$$

The feasibility conditions for C_I are

$$\Lambda > \frac{d_p B_T C_T H_T}{\beta_I \omega_p' D_T}, \quad \alpha = 1 \quad \text{and} \quad \xi = 0. \tag{4}$$

For the fully endemic equilibrium we find

$$E^* = \frac{1}{B_T}\left(\Lambda - \frac{d_p B_T C_T H_T}{\beta_I\left[(1-\alpha)\omega_p H_T + \alpha\omega_p' D_T\right]}\right), \qquad S^* = \frac{\Lambda - B_T E^*}{d_p},$$

$$I^* = \left(\frac{(1-\alpha)\omega_p H_T + \alpha\omega_p'\xi}{C_T H_T}\right)E^*, \qquad A^* = \left(\frac{\alpha\omega_p'}{H_T}\right)E^*$$

$$\text{and } R^* = \left(\frac{\gamma_p}{d_p}\left(\frac{(1-\alpha)\omega_p H_T + \alpha\omega_p'\xi}{C_T H_T}\right) + \frac{\gamma_p'}{d_p}\left(\frac{\alpha\omega_p'}{H_T}\right)\right)E^*,$$

with feasibility condition

$$\Lambda > \frac{d_p B_T C_T H_T}{\beta_I\left[(1-\alpha)\omega_p H_T + \alpha\omega_p' D_T\right]}, \qquad \text{and either} \quad \alpha \neq 1 \quad \text{or} \quad \xi \neq 0. \tag{5}$$

4.2. The Basic Reproduction Number

The basic reproduction number R_0 for system (1) is found using the next generation matrix method [16]. The reduced system of (1) may be written in compact form as: $X' = F(X) - V(X)$ where $X = (E, I, A)$

$$F(E,I,A) = \begin{pmatrix} \beta_I S(I + kA) \\ 0 \\ 0 \end{pmatrix}, \quad V(E,I,A) = \begin{pmatrix} -(1-\alpha)\omega_p E - \alpha\omega_p' E - d_p E \\ (1-\alpha)\omega_p E - (\gamma_p + d_p + \mu)I + \xi A \\ \alpha\omega_p' E - (\gamma_p' + d_p + \nu)A - \xi A \end{pmatrix}.$$

The Jacobian matrices of $F(X)$ and $V(X)$ at the disease-free equilibrium point C_0 are

$$J_F(C_0) = \begin{pmatrix} 0 & \beta_I S_0 & \beta_I S_0 k \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

and

$$J_V(C_0) = \begin{pmatrix} -B_T & 0 & 0 \\ (1-\alpha)\omega_p & -C_T & \xi \\ \alpha\omega_p' & 0 & -H_T \end{pmatrix}.$$

We find that

$$J_V^{-1}(C_0) = \begin{pmatrix} \frac{-1}{B_T} & 0 & 0 \\ \frac{-[(1-\alpha)\omega_p H_T + \alpha\omega_p'\xi]}{C_T B_T H_T} & \frac{-1}{C_T} & \frac{-\xi}{C_T H_T} \\ \frac{-\alpha\omega_p'}{B_T H_T} & 0 & \frac{-1}{H_T} \end{pmatrix}.$$

The next generation matrix is

$$-J_F(C_0)J_V^{-1}(C_0) = \begin{pmatrix} \beta_I S_0 \frac{(1-\alpha)\omega_p H_T + \alpha\omega_p' D_T}{C_T B_T H_T} & \frac{\beta_I S_0}{C_T} & \frac{\beta_I S_0 D_T}{C_T H_T} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Thus

$$R_0 = \rho(-J_F(C_0)J_V^{-1}(C_0)) = \beta_I S_0 \frac{(1-\alpha)\omega_p H_T + \alpha\omega_p' D_T}{C_T B_T H_T}. \tag{6}$$

The conditions (4) (resp. (5)) are equivalent to $R_0 > 1$ for $\alpha = 1$ and $\xi = 0$ (resp.$R_0 > 1$ for either $\alpha \neq 1$ or $\xi \neq 0$).

We have the following theorem

Theorem 2. *System (1) has the following equilibria:*

1. *The coronavirus-free equilibrium* $C_0 = (S_0, 0, 0, 0, 0) = \left(\frac{\Lambda}{d_p}, 0, 0, 0, 0\right)$ *which exists always.*
2. *In addition, if* $R_0 > 0$ *then system (1) admits another nontrivial equilibrium, in fact:*
   When $\alpha = 1$ *and* $\xi = 0$*, it is the coronavirus-symptomatic-infected-free equilibrium* $C_I = (S_I, E_I, I_I, A_I, R_I)$.
   When either $\alpha \neq 1$ *or* $\xi \neq 0$*, it is the fully coronavirus endemic equilibrium* $C^* = (S^*, E^*, I^*, A^*, R^*)$.

4.3. System's Equilibria Stability

4.3.1. Local Stability

In this subsection we investigate the local stability of the system's equilibria.

Theorem 3. *Letting*

$$\begin{aligned} a_2 &= B_T + C_T + H_T, \\ a_1 &= H_T[B_T + C_T] + B_T C_T - \beta_I S_0((1-\alpha)\omega_p + k\alpha\omega_p') \\ a_0 &= B_T C_T H_T - \beta_I S_0 \left((1-\alpha)\omega_p H_T + \alpha\omega_p' D_T\right). \end{aligned} \tag{7}$$

1. *The coronavirus-free equilibrium* $C_0 = (S_0, 0, 0, 0, 0)$ *of the system (1) is locally asymptotically stable if*

$$\Lambda < \frac{d_p}{\beta_I} \frac{B_T C_T H_T}{(1-\alpha)\omega_p H_T + \alpha\omega_p' D_T}, \text{ (resp. } R_0 < 1). \tag{8}$$

2. *If* $\Lambda > \frac{d_p}{\beta_I} \frac{B_T C_T H_T}{(1-\alpha)\omega_p H_T + \alpha\omega_p' D_T}$, *(resp.* $R_0 > 1$*), then the coronavirus-free equilibrium* C_0 *of the system (1) is unstable.*

Proof. The Jacobian matrix of system (1) at the coronavirus-free equilibrium C_0 is:

$$J(C_0) = \begin{pmatrix} -d_p & 0 & -\frac{\beta_I \Lambda}{d_p} & -\frac{k\beta_I \Lambda}{d_p} & 0 \\ 0 & -B_T & \frac{\beta_I \Lambda}{d_p} & \frac{k\beta_I \Lambda}{d_p} & 0 \\ 0 & (1-\alpha)\omega_p & -C_T & \xi & 0 \\ 0 & \alpha\omega_p' & 0 & -H_T & 0 \\ 0 & 0 & \gamma_p & \gamma_p' & -d_p \end{pmatrix}.$$

At point C_0, the eigenvalues of J are $-d_p$ of multiplicity order two and the roots of the following characteristic polynomial of a three by three submatrix of J whose coefficients a_i, $i = 0, \ldots, 2$ are given in (7):

$$\lambda^3 + a_2\lambda^2 + a_1\lambda + a_0 = 0. \tag{9}$$

It is evident that $a_2 > 0$. From condition (8) the following inequalities are also satisfied

$$\begin{aligned} a_0 &= B_T C_T H_T - \beta_I S_0 \left[(1-\alpha)\omega_p H_T + \alpha\omega_p' D_T\right] \\ &= \frac{\beta_I}{d_p} \left[(1-\alpha)\omega_p H_T + \alpha\omega_p' D_T\right] \left(\frac{d_p}{\beta_I} \frac{B_T C_T H_T}{\left[(1-\alpha)\omega_p H_T + \alpha\omega_p' D_T\right]} - \Lambda\right) > 0, \end{aligned}$$

$$\begin{aligned}\left[(1-\alpha)\omega_p H_T+\alpha\omega'_p D_T\right]a_1 &= [H_T(B_T+C_T)+B_TC_T]\left[(1-\alpha)\omega_p H_T+\alpha\omega'_p D_T\right]\\ &\quad -\beta_I S_0\left[(1-\alpha)\omega_p H_T+\alpha\omega'_p D_T\right][(1-\alpha)\omega_p+k\alpha\omega'_p]\\ &= [H_T(B_T+C_T)+B_TC_T]\left[(1-\alpha)\omega_p H_T+\alpha\omega'_p D_T\right]\\ &\quad +[a_0-B_TC_TH_T][(1-\alpha)\omega_p+k\alpha\omega'_p]\\ &= H_T(B_T+C_T)(1-\alpha)\omega_p H_T+(H_TC_T+B_TC_T)\alpha\omega'_p\xi\\ &\quad +a_0[(1-\alpha)\omega_p+k\alpha\omega'_p]>0\end{aligned}$$

and

$$\begin{aligned}\left[(1-\alpha)\omega_p H_T+\alpha\omega'_p D_T\right]a_1a_2 &= H_T(B_T+C_T)(1-\alpha)\omega_p H_Ta_2+(H_TC_T+B_TC_T)\alpha\omega'_p\xi\, a_2\\ &\quad +a_0[(1-\alpha)\omega_p+k\alpha\omega'_p]a_2\\ &= H_T(B_T+C_T)(1-\alpha)\omega_p H_Ta_2+(H_TC_T+B_TC_T)\alpha\omega'_p\xi\, a_2\\ &\quad +a_0(1-\alpha)\omega_p(B_T+C_T+H_T)+a_0k\alpha\omega'_p(B_T+C_T+H_T)\\ &> B_TC_T\alpha\omega'_p\xi a_2+a_0(1-\alpha)\omega_p H_T+a_0\alpha\omega'_p(kC_T+\xi)-a_0\alpha\omega'_p\xi\\ &> B_TC_TH_T\alpha\omega'_p\xi+a_0[(1-\alpha)\omega_p H_T+\alpha\omega'_p D_T]-B_TC_TH_T\alpha\omega'_p\xi\\ &= a_0[(1-\alpha)\omega_p H_T+\alpha\omega'_p D_T].\end{aligned}$$

Thus, $a_i>0$, $i=0,\ldots,2$ and $a_2a_1>a_0$.

Then, according to the Routh–Hurwitz criterion, all the roots of the characteristic Equation (9) have negative real parts. Therefore, the coronavirus-free equilibrium point C_0 is locally asymptotically stable under condition (8). □

Since we can deduce the stability of the coronavirus symptomatic infected-free equilibrium C_I from the stability of the coronavirus endemic equilibrium C^* simply by taking $\alpha=1$ and $\xi=0$ in the latter, we now just analyze the coronavirus endemic equilibrium C^*.

Theorem 4. *Let*

$$\begin{cases} c_3 &= \beta_I(I^*+kA^*)+d_p+H_T+B_T+C_T>0,\\ c_2 &= [\beta_I(I^*+kA^*)+d_p](H_T+B_T+C_T)+B_TC_T+H_T(B_T+C_T)\\ &\quad -[\alpha\omega'_pk+(1-\alpha)\omega_p]\beta_IS^*,\\ c_1 &= [\beta_I(I^*+kA^*)+d_p][B_TC_T+H_T(B_T+C_T)]+H_TB_TC_T\\ &\quad -[\alpha\omega'_p(kd_p+D_T)+(1-\alpha)\omega_p(d_p+H_T)]\beta_IS^*,\\ c_0 &= [\beta_I(I^*+kA^*)+d_p]H_TB_TC_T-d_p[\alpha\omega'_pD_T+(1-\alpha)\omega_pH_T]\beta_IS^*.\end{cases} \tag{10}$$

The coronavirus endemic equilibrium C^* is locally asymptotically stable if

$$\Lambda>\frac{d_p}{\beta_I}\frac{B_TC_TH_T}{(1-\alpha)\omega_pH_T+\alpha\omega'_pD_T},\ (\textit{ resp. } R_0>1). \tag{11}$$

Proof. The Jacobian matrix of system (1) at the coronavirus endemic equilibrium C^* is:

$$J(C^*)=\begin{pmatrix} -\beta_I(I^*+kA^*)-d_p & 0 & -\beta_IS^* & -\beta_IS^*k & 0\\ \beta_I(I^*+kA^*) & -B_T & \beta_IS^* & \beta_IS^*k & 0\\ 0 & (1-\alpha)\omega_p & -C_T & \xi & 0\\ 0 & \alpha\omega'_p & 0 & -H_T & 0\\ 0 & 0 & \gamma_p & \gamma'_p & -d_p \end{pmatrix}.$$

At point C^*, the eigenvalues of J are $-d_p$ and the roots of the characteristic polynomial of a three by three submatrix of J. The characteristic equation, in which the coefficients c_i, $i = 0, \ldots, 3$ are given in (10), is:

$$\lambda^4 + c_3\lambda^3 + c_2\lambda^2 + c_1\lambda + c_0 = 0. \tag{12}$$

It is evident that $c_3 > 0$. From condition (11) the following inequalities are also satisfied.

$$\begin{aligned}
c_0 &= [\beta_I(I^* + kA^*) + d_p]H_T B_T C_T - \beta_I[(1-\alpha)\omega_p H_T + \alpha\omega_p' D_T]d_p S^* \\
&= \beta_I\left[(1-\alpha)\omega_p H_T + \alpha\omega_p'(\xi + kC_T)\right]B_T E^* + d_p H_T C_T B_T \\
&\quad -\beta_I[(1-\alpha)\omega_p H_T + \alpha\omega_p' D_T](\Lambda - B_T E^*) \\
&= \beta_I\left[(1-\alpha)\omega_p H_T + \alpha\omega_p' D_T\right]\left(2B_T E^* + \frac{d_p H_T C_T B_T}{\beta_I[(1-\alpha)\omega_p H_T + \alpha\omega_p' D_T]} - \Lambda\right) \\
&= \beta_I\left[(1-\alpha)\omega_p H_T + \alpha\omega_p' D_T\right]B_T E^* > 0,
\end{aligned}$$

$$\begin{aligned}
c_1 &= [\beta_I(I^* + kA^*) + d_p][B_T C_T + H_T(B_T + C_T)] + H_T B_T C_T \\
&\quad -[(1-\alpha)\omega_p(d_p + H_T) + \alpha\omega_p'(kd_p + D_T)]\beta_I S^* \\
&= \left[\beta_I\left(\frac{(1-\alpha)\omega_p H_T + \alpha\omega_p' D_T}{C_T H_T}\right)E^* + d_p\right][B_T C_T + H_T(B_T + C_T)] \\
&\quad -\beta_I[(1-\alpha)\omega_p + \alpha\omega_p' k](\Lambda - B_T E^*) \\
&= \beta_I\left(\frac{(1-\alpha)\omega_p H_T(B_T + C_T)}{C_T} + \frac{\alpha\omega_p'\xi B_T}{C_T} + \frac{\alpha\omega_p' D_T(B_T + H_T)}{H_T}\right)E^* \\
&\quad +d_p[B_T C_T + H_T(B_T + C_T)] - \beta_I[(1-\alpha)\omega_p + \alpha\omega_p' k](\Lambda - 2B_T E^*) \\
&= \beta_I\left(\frac{(1-\alpha)\omega_p H_T(B_T + C_T)}{C_T} + \frac{\alpha\omega_p'\xi B_T}{C_T} + \frac{\alpha\omega_p' D_T(B_T + H_T)}{H_T}\right)E^* \\
&\quad +d_p[B_T C_T + H_T(B_T + C_T)] - \beta_I[(1-\alpha)\omega_p + \alpha\omega_p' k]\left(\frac{2d_p B_T C_T H_T}{\beta_I\left[(1-\alpha)\omega_p H_T + \alpha\omega_p' D_T\right]} - \Lambda\right) \\
&= \beta_I[(1-\alpha)\omega_p H_T + \alpha\omega_p' D_T]\left(\frac{C_T H_T + B_T(C_T + H_T)}{C_T H_T}\right)E^* \\
&\quad +d_p\left(\frac{(B_T + C_T)H_T^2(1-\alpha)\omega_p + H_T B_T\alpha\omega_p'\xi + C_T(B_T + H_T)\alpha\omega_p' D_T}{\left[(1-\alpha)\omega_p H_T + \alpha\omega_p' D_T\right]}\right) > 0,
\end{aligned}$$

$$
\begin{aligned}
c_2 &= [\beta_I(I^*+kA^*)+d_p](H_T+B_T+C_T)+B_TC_T+H_T(B_T+C_T)\\
&\quad -[\alpha\omega_p'k+(1-\alpha)\omega_p]\beta_I S^*\\
&= \left[\beta_I\left(\frac{(1-\alpha)\omega_pH_T+\alpha\omega_p'D_T}{C_TH_T}\right)E^*+d_p\right](H_T+B_T+C_T)+B_TC_T+H_T(B_T+C_T)\\
&\quad -[\alpha\omega_p'k+(1-\alpha)\omega_p]\beta_I\frac{\Lambda-B_TE^*}{d_p}\\
&= \beta_I\left(\frac{(1-\alpha)\omega_p(H_T+B_T)}{C_T}+\frac{\alpha\omega_p'\xi(H_T+B_T+C_T)}{C_TH_T}+\frac{\alpha\omega_p'k(B_T+C_T)}{H_T}\right)E^*\\
&\quad +d_p(H_T+B_T+C_T)+B_TC_T+H_T(B_T+C_T)\\
&\quad +\beta_I[(1-\alpha)\omega_p+\alpha\omega_p'k]\left(\frac{(d_p+B_T)E^*-\Lambda}{d_p}\right)\\
&= \beta_I\left(\frac{(1-\alpha)\omega_p(H_T+B_T)}{C_T}+\frac{\alpha\omega_p'\xi(H_T+B_T+C_T)}{C_TH_T}+\frac{\alpha\omega_p'k(B_T+C_T)}{H_T}\right)E^*\\
&\quad +d_p(H_T+B_T+C_T)+B_TC_T+H_T(B_T+C_T)\\
&\quad +\frac{\beta_I[(1-\alpha)\omega_p+\alpha\omega_p'k]}{B_T}\left(\Lambda-\frac{(d_p+B_T)B_TC_TH_T}{\beta_I\left[(1-\alpha)\omega_pH_T+\alpha\omega_p'D_T\right]}\right)\\
&= \beta_I(B_T+C_T+H_T)\left(\frac{(1-\alpha)\omega_pH_T+\alpha\omega_p'D_T}{C_TH_T}\right)E^*\\
&\quad +d_p(H_T+B_T+C_T)+C_TH_T+B_T\left(\frac{(1-\alpha)\omega_pH_T^2+H_T\alpha\omega_p'\xi+\alpha\omega_p'C_TD_T}{\left[(1-\alpha)\omega_pH_T+\alpha\omega_p'D_T\right]}\right)>0
\end{aligned}
$$

and

$$
\begin{aligned}
c_1(c_3c_2-c_1) &= \beta_I\left(\frac{(1-\alpha)\omega_pH_T+\alpha\omega_p'D_T}{C_TH_T}\right)E^*c_1c_2\\
&\quad +\beta_I\left[(H_T+C_T+d_p)H_T+C_T(C_T+d_p)\right]\left(\frac{\left[(1-\alpha)\omega_pH_T+\alpha\omega_p'D_T\right]}{C_TH_T}\right)E^*c_1\\
&\quad +\beta_I(H_T+B_T+C_T+d_p)B_T\left(\frac{(1-\alpha)\omega_pH_T+\alpha\omega_p'D_T}{C_TH_T}\right)E^*c_1\\
&\quad +(H_T+B_T+C_T)(H_T+B_T+C_T+d_p)d_pc_1+C_TH_T(C_T+H_T+B_T)c_1\\
&\quad +B_T(B_T+C_T+H_T)\left(\frac{[(1-\alpha)\omega_pH_T^2+\alpha\omega_p'C_TD_T]+\alpha\omega_p'\xi H_T}{\left[(1-\alpha)\omega_pH_T+\alpha\omega_p'D_T\right]}\right)c_1\\
&> \beta_I^2B_TE^{*2}\left\{\beta_I\left[(1-\alpha)\omega_pH_T+\alpha\omega_p'D_T\right]\left(\frac{(1-\alpha)\omega_pH_T+\alpha\omega_p'D_T}{C_TH_T}\right)^2E^*\right.\\
&\quad \left.+2(d_p+H_T+B_T+C_T)\left[(1-\alpha)\omega_pH_T+\alpha\omega_p'D_T\right]\left(\frac{(1-\alpha)\omega_pH_T+\alpha\omega_p'D_T}{C_TH_T}\right)\right\}\\
&\quad +\beta_IB_T(d_p+H_T+B_T+C_T)^2\left[(1-\alpha)\omega_pH_T+\alpha\omega_p'D_T\right]E^*\\
&= c_0c_3^2.
\end{aligned}
$$

Thus, $c_i > 0, i = 0, \ldots, 3$ and $c_1(c_3c_2 - c_1) > c_0c_3^2$. Then, according to the Routh–Hurwitz criterion, all the roots of the characteristic Equation (12) have negative real parts. Therefore, the coronavirus endemic equilibrium point C^* is locally asymptotically stable under condition (11). □

From Theorem 4 the following result is reached.

Theorem 5. *Let*

$$\begin{cases} b_3 &= \beta_I k A_I + d_p + H_T + B_T + C_T > 0, \\ b_2 &= [\beta_I k A_I + d_p](H_T + B_T + C_T) + B_T C_T + H_T(B_T + C_T) \\ & \quad -\omega_p' k \beta_I S_I, \\ b_1 &= [\beta_I k A_I + d_p][B_T C_T + H_T(B_T + C_T)] + H_T B_T C_T \\ & \quad -\omega_p'(k d_p + D_T)\beta_I S_I, \\ b_0 &= [\beta_I k A_I + d_p] H_T B_T C_T - d_p \omega_p' D_T \beta_I S^*. \end{cases} \tag{13}$$

The coronavirus symptomatic-infected-free equilibrium C_I of the system (1) is locally asymptotically stable if

$$\Lambda > \frac{d_p}{\beta_I} \frac{B_T C_T H_T}{\omega_p' D_T}, \ (\textit{ resp. } R_0 > 1). \tag{14}$$

Proof. The result can easily obtained from Theorem 4 by taking $\alpha = 1$ and $\xi = 0$. □

Additionally, from the previous discussion, we can claim the following result:

Theorem 6. *There is a transcritical bifurcation between C_0 and C^*.*

4.3.2. Global Stability

Next, we address the issue of global stability of the coronavirus–free equilibrium, employing as a tool a suitably constructed Lyapunov function and La Salle's Invariance Principle.

Theorem 7. *The coronavirus-free equilibrium C_0 of model (1) is globally asymptotically stable if*

$$\Lambda < \frac{d_p B_T C_T H_T}{\beta_I[(1-\alpha)\omega_p H_T + D_T \alpha \omega_p']}, \ (\textit{ resp. } R_0 < 1). \tag{15}$$

Proof. First, the four equations of (1) are independent of R, therefore, the last equation of (1) can be omitted without loss of generality. Hence, let us consider the following function:

$$P = \frac{1}{2S_0}(S - S_0)^2 + E + \frac{B_T}{[(1-\alpha)\omega_p H_T + D_T \alpha \omega_p']}(H_T I + D_T A) \tag{16}$$

It is easily seen that the above function is nonnegative and also $P = 0$ if and only if $S = S_0$, $E = 0$, $I = 0$ and $A = 0$. Further, calculating the time derivative of P along the positive solutions of (1), we find:

$$\begin{aligned}
\frac{dP}{dt} &= \frac{1}{S_0}(S-S_0)(-\beta_I S(I+kA) - d_p(S-S_0)) + \beta_I S(I+kA) - B_T E \\
&\quad + \frac{B_T H_T((1-\alpha)\omega_p E - C_T I + \xi A)}{[(1-\alpha)\omega_p H_T + D_T \alpha \omega_p']} + \frac{B_T D_T(\alpha\omega_p' E - H_T A)}{[(1-\alpha)\omega_p H_T + D_T \alpha \omega_p']} \\
&= -\frac{d_p}{S_0}(S-S_0)^2 + \beta_I[2S - \frac{S^2}{S_0} - S_0](I+kA) + \beta_I S_0(I+kA) \\
&\quad - \frac{B_T C_T H_T I}{[(1-\alpha)\omega_p H_T + D_T \alpha \omega_p']} - \frac{B_T(-\xi + D_T)H_T A}{[(1-\alpha)\omega_p H_T + D_T \alpha \omega_p']} \\
&= -\frac{d_p}{S_0}(S-S_0)^2 + \beta_I[2S - \frac{S^2}{S_0} - S_0](I+kA) \\
&\quad + \left(\beta_I S_0 - \frac{B_T C_T H_T}{[(1-\alpha)\omega_p H_T + D_T \alpha \omega_p']}\right)(I+kA) \\
&= -\frac{d_p}{S_0}(S-S_0)^2 + \beta_I[2S - \frac{S^2}{S_0} - S_0](I+kA) \\
&\quad + \frac{\beta_I}{d_p}\left(\Lambda - \frac{d_p B_T C_T H_T}{\beta_I[(1-\alpha)\omega_p H_T + D_T \alpha \omega_p']}\right)(I+kA).
\end{aligned}$$

From condition (15) we can show that the coefficients of the term $I + kA$ in the last equality are negative. Further, we have $2S - \frac{S^2}{S_0} - S_0 = -\frac{S^2 - 2SS_0 + S_0^2}{S_0} = -\frac{(S-S_0)^2}{S_0} \leq 0$ for all $S \geq 0$. Thus, we have $\frac{dP}{dt} \leq 0$ for all $(S, E, I, A) \in \mathbb{R}^4_+$ and $\frac{dP}{dt} = 0$ if and only if $(S, E, I, A) = (S_0, 0, 0, 0)$. Thus, the only invariant set contained in $\mathbb{R}^4_+$ is $\{(S_0, 0, 0, 0)\}$. Hence, La Salle's theorem implies convergence of the solutions (S, E, I, A) to $(S_0, 0, 0, 0)$. From the last equation if (1) we can show obviously that R converge also to 0. Therefore C_0 is globally asymptotically stable if $R_0 < 1$. □

4.4. Numerical Simulations

The calculation of the value of R_0 according to (6) with the parameter values used in the numerical simulations gives $R_0 = 3.1402$, in line with the current estimates [11,12].

4.4.1. Simulations Methodology

We use a simple own-developed driver code calling the Matlab intrinsic routine ode45, implementing the classical Runge–Kutta 45 integration method for ordinary differential equations.

At first, we consider only the demographic simulation and show that the population is essentially at the same level during a year. This fact is substantiated also by the simulation results, for which there is scant difference between those of the model (1) and the ones obtained by using its no-demographic counterpart, where Λ and d_p are both set to zero.

We then perform three sets of simulations describing different possible scenarios. The first one considers lock-down, i.e., decreasing the contact rate significantly, but not to zero, as some essential activities are still open. Then the total isolation policy, for which the contact rate is set to zero. Finally an intermittent closure policy, for which when infectives reappear in a significant way, temporary lock-down measures resume again.

4.4.2. Data Acquisition

We use data published on official websites about the epidemic's spread in Italy collected between 29 January and 28 March 2020, a period that spans 61 days, incremented by more recent information [17].

Using the day as the base time unit, we assume that the average incubation period lies in the interval between two and 14 days, with a mean of 8 days. Based on the percentage of the reported symptomatic infected patients, the proportion of symptomatic in the infected class α is estimated to be in the interval $[0.01, 0.3]$. The correction k for asymptomatics to diffuse the disease is set in the range $k \in [0:005; 0:2]$. There have been 27,359 deaths between 15 February and 29 April [17], with changes in the number of fatalities every day. Dividing the fatal cases by the timespan, one gets 370 daily fatalities, which gives a rate 0.0027. Using this value in the simulation, puts the total losses to about 10^5. But we observed that apparently children hardly get the disease, the younger and adult people have it generally in a mild form and fatalities occur mainly for the elderly people, compare with Figure 3 of [18]. In view of the fact that there is no age structure in this model, we corrected this value by taking a third of the above result to set the disease mortality rate at the final value $\mu = 0.001$, which gives a reasonable estimate for the losses in the timespan, in rough agreement with the actual tallies. We neglect altogether mortality for the asymptomatics, setting $\nu = 0$. Based on the officially published data we estimate $\gamma_p = 0.1764$, $\gamma_p' = 0.6024$. For the initial values, the total population is obtained from the report published by the official cite of worldometers [19], $S = 60461826$. To avoid demographic effects, we set the susceptible recruitment rate Λ in order that on the timespan of the simulation the total population N does not change much.

4.4.3. The Pure Demographic Case

We simulate first the population model without disease. In so doing, we varied the parameter Λ until a satisfactory behavior of N, the total population was found. With $\Lambda = 500$ there is little variation of N during a whole year, the population remains roughly stable around the level $60,400,000$, see Figure 24. In this way the demographic effects are sort of removed, and we can concentrate mainly on the epidemics. Actually, the number of newborns per day in Italy would be about four times higher, but as mentioned, we just would like here to hide the demographics from the simulations and not have a picture more adherent to reality.

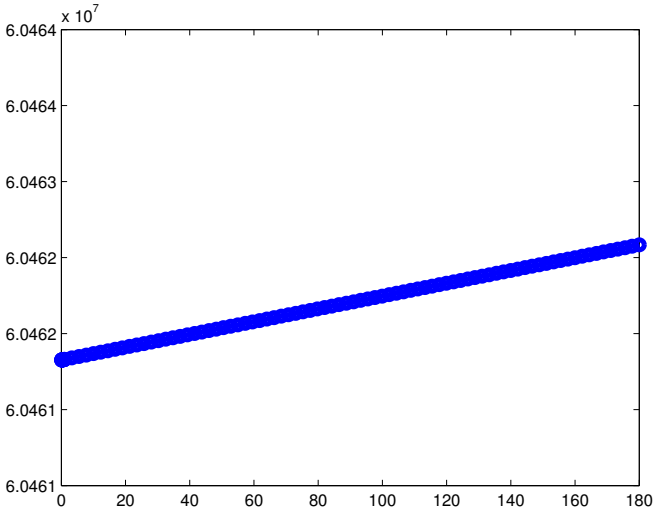


Figure 24. The susceptible population behavior over a year, without disease. It does not vary much as the vertical scale is rather small, the range of variation being around 3000, over a population of 60×10^6.

4.4.4. Epidemics Spread in the Absence of Measures

Here we introduced the disease, with incidence $\beta_I = 10^{-7}$. The result is shown in Figure 25 for absolute numbers, and in Figure 6 in semilogarithmic scale. In this case no measures are assumed to be taken to counteract the epidemics. These results are reported in order be able to compare the simulations with restrictions to what would happen if the containment measures were not taken.

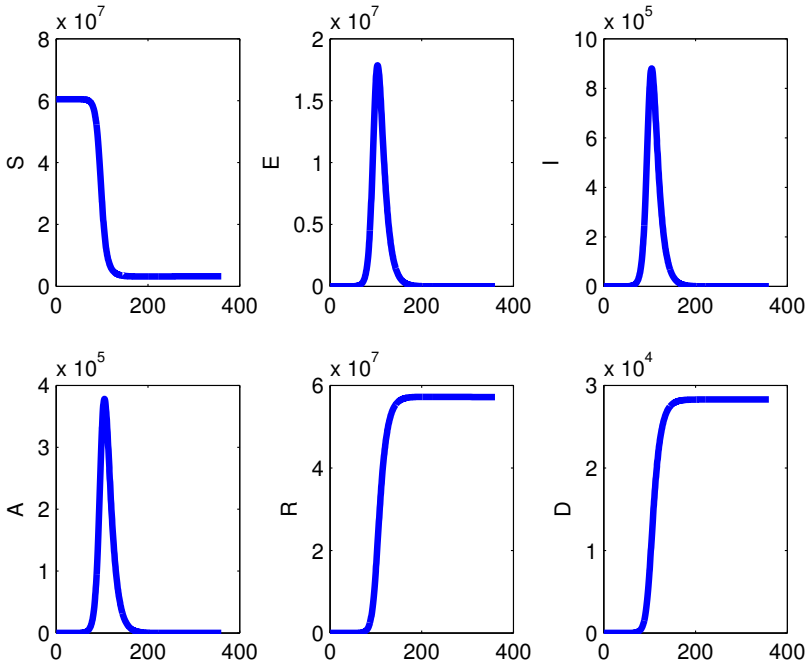


Figure 25. The epidemic's effect on the population in the absence of measures for $\beta_I = 10^{-7}$, over a period of one year. Left to right and top to bottom, the total population sizes: S, E, I, A, R and D, the disease-related deceased class.

4.4.5. Containment Measures for the Epidemics

Finally, we considered the introduction of the distancing policy. It is assumed to start at time t_1 and end at time $t_1 + t_2$. Two forms of containment measures are considered, substantially reducing the contact rate, or even setting it equal to zero, meaning the extreme measure of total individuals isolation.

In particular, we present the experience of using the reference value of the contact rate $\beta_I = 10^{-7}$, then reducing it to $\beta_I = 10^{-10}$ during the interval $[t_1, t_1 + t_2]$. We then reset it to its previous reference value after time $t_1 + t_2$. We monitored the epidemics evolution over six months.

The alternative, milder choice $\beta_I = 10^{-8}$ is also used, for comparison.

The simulations are then repeated with total isolation, setting $\beta_I = 0$ during the implementation of the restrictions.

A comparison of the results with the model obtained by disregarding the demographic parameters, i.e., setting $\Lambda = d_p = 0$ is also performed in the same way as done for the model (1).

Different timings for taking both the first restriction measure and for lifting it are then investigated, using all the above alternatives.

Finally an intermittent restrictive policy is examined, for which when the infected are observed to trespass a threshold, distancing measures are taken. Here again lock-down or total isolation produce essentially the same results. The use of different timings for the introduction of the restrictions is also scrutinized.

Author Contributions: Model formulation, Y.B. and E.V.; methodology, Y.B., M.H. and E.V.; formal analysis, Y.B., M.H. and E.V.; writing—original draft preparation, Y.B.; simulations, writing—review and editing, E.V. All authors have read and agreed to the published version of the manuscript.

Funding: E.V. was partially funded by the local research project "Questioni attuali di approssimazione numerica e loro applicazioni" of the Dipartimento di Matematica, Universitá di Torino. This research has been undertaken within the framework of the COST Action: CA 16227—Investigation and Mathematical Analysis of Avant-garde Disease Control via Mosquito Nano-Tech-Repellents. This work was partially supported by the Ministry of Higher Education and Scientific Research of Algeria (MESRS) and General Direction of Scientific Research and Technological Development (DGRSDT) through Research Project-University Formation (PRFU: C00L03UN220120190001 and PRFU: C00L03UN220120180004).

Acknowledgments: The authors would like to thank the reviewers for their valuable comments and suggestions that greatly improved the presentation of this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cecconi, M.; Forni, G.; Mantovani, A. *COVID-19: An Executive Report, "Commissione Salute, March 25th"*; Accademia Nazionale dei Lincei: Roma, Italy, 2020.
2. Herbert, W. Hethcote, The Mathematics of Infectious Diseases. *SIAM Rev.* **2000**, *42*, 599–653. [CrossRef]
3. Magal, P.; Webb, G. Predicting the number of reported and unreported cases for the COVID-19 epidemic in South Korea, Italy, France and Germany. *medRxiv* **2020**. [CrossRef]
4. Buonomo, B. Effects of information-dependent vaccination behavior on coronavirus outbreak: Insights from a SIRI model. *Ricerche di Matematica* **2020**, doi:10.1007/s11587-020-00506-8 [CrossRef]
5. Chen, T.M.; Rui, J.; Wang, Q.P.; Zhao, Z.Y.; Cui, J.A.; Yin, L. A mathematical model for simulating the transmission of Wuhan novel Coronavirus. *bioRxiv* **2020**. [CrossRef]
6. Jia, J.; Ding, J.; Liu, S.; Liao, G.; Li, J.; Duan, B.; Wang, G.; Zhang, R. Modeling the control of Covid-19: Impact of policy interventions and meteorological factors. *Electron. J. Differ. Equ.* **2020**, *2020*, 1–24. Available online: https://www.researchgate.net/publication/339786734 (accessed on 27 February 2020).
7. Murray, J.D. *Mathematical Biology*; Springer: Berlin/Heidelberg, Germany, 2002.
8. Perko, L. *Differential Equations and Dynamical Systems*; Springer: Berlin/Heidelberg, Germany, 2001.
9. Strogatz, S. *Nonlinear Dynamics and Chaos*; Perseus Books: Reading, MA, USA, 1994.
10. Adamik, B.; Bawiec, M.; Bezborodov, V.; Bock, W.; Bodych, M.; Burgard, J.; Götz, T.; Krueger, T.; Migalska, A.; Pabjan, B.; et al. Mitigation and herd immunity strategy for COVID-19 is likely to fail. *medRxiv* **2020**, doi:10.1101/2020.03.25.20043109 [CrossRef]
11. Liu, G.; Wilder-Smith, A.; Rocklöv, J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *J. Travel Med.* **2020**, *27*, taaa021. doi:10.1093/jtm/taaa021. [CrossRef] [PubMed]
12. COVID-19 Pandemic in Italy. Available online: https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Italy (accessed on 29 April 2020).
13. Banerjee, M.; Tokarev, A.; Volpert, V. Immuno-epidemiological model of two-stage epidemic growth. *Math. Model. Nat. Phenom.* **2020**, *15*, 27. [CrossRef]
14. Kochańczyk, M.; Grabowski, F.; Lipniacki, T. Dynamics of COVID-19 pandemic at constant and time-dependent contact rates. *Math. Model. Nat. Phenom.* **2020**, *15*, 28 [CrossRef]
15. Volpert, V.; Banerjee, M.; d'Onofrio, A.; Lipniacki, T.; Petrovskii, S.; Tran, V.C. Coronavirus—Scientific insights and societal aspects. *Math. Model. Nat. Phenom.* **2020**, *15*, E2. [CrossRef]
16. Van den Driessche, P.; Watmough, J. A simple SIS epidemic model with a backward bifurcation. *J. Math. Biol.* **2000**, *40*, 525–540. [CrossRef] [PubMed]
17. Italy Coronavirus: Cases and Deaths–Worldometer. Available online: https://www.worldometers.info/coronavirus/country/italy/ (accessed on 27 February 2020).
18. Verity, R.; Okell, L.C.; Dorigatti, I.; Winskill, P.; Whittaker, C.; Imai, N.; Cuomo-Dannenburg, G.; Thompson, H.; Walker, P.G.; Fu, H.; et al. Estimates of the severity of coronavirus disease 2019: A model-based analysis. *Lancet* **2020**, doi:10.1016/S1473-3099(20)30243-71. [CrossRef]
19. Reported Cases and Deaths by Country, Territory, or Conveyance. Available online: https://www.worldometers.info/coronavirus/#countries (accessed on 27 February 2020).

mathematics

MDPI

Article

Lanchester Models for Irregular Warfare

Moshe Kress

Operations Research Department, Naval Postgraduate School, Monterey, CA 93943, USA; mkress@nps.edu

Received: 25 March 2020; Accepted: 4 May 2020; Published: 7 May 2020

Abstract: Military operations research and combat modeling apply mathematical models to analyze a variety of military conflicts and obtain insights about these phenomena. One of the earliest and most important set of models used for combat modeling is the Lanchester equations. Legacy Lanchester equations model the mutual attritional dynamics of two opposing military forces and provide some insights regarding the fate of such engagements. In this paper, we review recent developments in Lanchester modeling, focusing on contemporary conflicts in the world. Specifically, we present models that capture irregular warfare, such as insurgencies, highlight the role of target information in such conflicts, and capture multilateral situations where several players are involved in the conflict (such as the current war in Syria).

Keywords: Lanchester models; irregular warfare; target information; multilateral conflicts

1. Introduction

Military operations research and combat modeling apply mathematical models to analyze a variety of military conflicts and combat situations, and obtain insights about these phenomena [1,2]. One of the earliest and most important set of models used for combat modeling is the Lanchester equations [3]. While there are two major manifestations of Lanchester equations—deterministic and stochastic—we focus in this paper only on the deterministic equations.

Lanchester equations are systems of ordinary differential equations describing the mutual attrition that occurs continuously in time between two opposing forces engaged in violent confrontation. The equations involve state variables that represent the number of live combatants (or weapons) at any given time during the battle. Each equation expresses the rate of change in one state variable as a function of other state variables. There are several Lanchester models that differ in their underlying assumptions regarding the operational posture and/or the tactical situation. We describe the two most common models in Section 2.

Lanchester equations, and variations thereof, have been implemented in large-scale combat models [4,5] and used by military analysts and combat planners for studying force structure and combat courses of action.

The recent literature on Lanchester equations is quite wide in scope and ranges from fitting Lanchester models to data from historical battles [6–8], to using partial differential equations to capture spatial and interaction effects [9], and to developing approximate solutions to stochastic versions of Lanchester equations [10]. Lanchester models have also been used in biology [11], evolution [12], and even in the advertisement world [13] and video-gaming [14].

A particular body of research on Lanchester models, notable in the past ten years or so, focuses on irregular warfare, which is manifested in asymmetric conflicts such as insurgencies and multilateral engagements. Irregular warfare is also characterized by asymmetry in the mechanism that provides information, intelligence and situational awareness to the two sides. These types of irregular warfare represent most recent conflicts, such as those in Afghanistan and Syria.

In this paper, we review recent applications of Lanchester theory to irregular warfare. One type of application includes Lanchester models in which the situational awareness capabilities on both sides are asymmetric; one side can target better than the other and therefore gains an advantage. A second type of model relates to cases where the two sides of the conflict are profoundly different in terms of their force structures and their associated attritional dynamics. Such models describe many-on-one situations or scenarios where civilians have a significant effect on the way the conflict evolves. A third type of irregular conflict is when there are more than two sides competing for dominance. We call such conflicts multilateral conflicts. Consider a situation where several sides, Blue, Red, Brown, Green, etc., seek dominance in a region by fighting (or cooperating with) others. Such multilateral conflicts lend themselves to game-theoretic situation, which we discuss in this paper. A striking result characterizes the fate of such conflicts (see Section 6).

The paper is organized as follows: in Section 2, we present a brief introduction to Lanchester models, describing the legacy aimed-fire and un-aimed (area) fire models [3], as well as the ancient battle. Section 3 presents models that capture asymmetric battles such as insurgencies and one-on-many combat situations. In particular, we present the classical guerrilla model of Deitchman [15]. Section 4 focuses on the role of information and combat intelligence in Lanchester modeling. Section 5 describes Lanchester models that incorporate the presence of civilians during insurgencies and show their impact on tactics and battlefield outcomes. Section 6 presents two models of a multilateral conflict such as the one that has been going on for the past nine years in Syria. Section 7 summarizes the paper and points at some future research directions in Lanchester theory.

2. Legacy Lanchester Models

Frederick William Lanchester proposed in 1916 to model mutual attrition of two fighting forces by a set of ordinary differential equations (ODE) [3]. The state variables of the ODE represent fighting entities in the battlefield, and each of the ODEs capture the rate of decrease in a certain state variable as a function of the other state variables. These models were inspired by air-combat scenarios in World War I and are named after Lanchester, even though it appears that a Russian mathematician, named Osipov, developed similar models at about the same time [16]. While, in general, each side may have several types of fighting combatant, each with different attrition rates, we assume here, for simplifying the initial exposition, that each of the two sides comprises a homogeneous force. We relax this assumption later on.

Let $B = B(t)$ and $R = R(t)$ be state variables denoting the sizes of the surviving combatants at time t of the Blue force and the Red force, respectively. The aimed-fire model represents a combat situation where each combatant on the Blue (Red) side effectively reduces the force on the Red (Blue) side at a certain fixed attrition rate $\beta(\rho)$. Formally, the aimed-fire model is:

$$\begin{aligned} \dot{B} &= -\rho R \\ \dot{R} &= -\beta B \end{aligned} \tag{1}$$

By the separation of variables, one can obtain the state-equation

$$\beta(B_0^2 - B^2) = \rho(R_0^2 - R^2) \tag{2}$$

where B_0 and R_0 are the force sizes of Blue and Red, respectively, at the beginning of the battle. In particular, we obtain that the parity condition—the values of B_0, R_0, β and ρ such that the battle ends with mutual annihilation—is $\beta B_0^2 = \rho R_0^2$. The side with the larger product of attrition rate and the square of the initial force size wins the battle. We observe that while the attrition rate has a linear effect, the effect of size is quadratic; doubling the attrition rate doubles the effective attrition, but doubling the number of combatants has a quadratic effect. This is the reason the aimed-fire model is also called the Square Law. The Square Law underscores the importance of concentration of forces—a well-known principle in military tactics.

While there is a single Square Law Lanchester model, there are two Linear Law models. The first Linear Law model, also called the Ancient Battle, assumes that the battle comprises a collection of one-on-one duels, typical to battles in early history. In such a battle, there is no meaning for concentration of forces and the attrition is fixed. Thus, the pair of equations describing the ancient battle is simply

$$\begin{aligned} \dot{B} &= -\rho \\ \dot{R} &= -\beta \end{aligned} \tag{3}$$

and the state-equation is $\beta(B_0 - B) = \rho(R_0 - R)$. In this case the effect of force size is linear, hence the name Linear Law.

The second Linear Law describes un-aimed fire, where the effect of one's fire does not only depend on the size of its own surviving force but also on the density of the targets at the opposing force. As the fire is not aimed, the probability of acquiring a target on the other side depends on the number of such targets in a given area. This is the reason that this model is also called the Area Fire model. The pair of differential equations in this case is:

$$\begin{aligned} \dot{B} &= -\rho BR \\ \dot{R} &= -\beta BR \end{aligned} \tag{4}$$

with the same state equations $\beta(B_0 - B) = \rho(R_0 - R)$ as in the ancient battle.

The classical literature on Lanchester models has other variations of the basic two laws: Square and Linear [17]. Also, there are stochastic versions of the deterministic Lanchester models described above, which are essentially continuous-time Markov processes. While they capture the inherent stochasticity embedded in a battlefield, they are, in general, less common because of their computational complexity and their relatively limited capability to represent non-homogeneous combat situations such as irregular warfare.

3. Asymmetric Engagements

In all three battles described in Section 2—aimed fire, area fire and ancient battle—both sides apply the same type of tactics and firing techniques; the battles are symmetric. Asymmetric engagements occur when the two sides apply different tactics. One such asymmetric combat situation occurs when regular forces of a state fight guerrillas or insurgents who apply irregular warfare tactics. The first to capture this situation in a Lanchester setting was Deitchman, who developed a mixture of the direct-fire and area-fire models called the Guerrilla Warfare model [15]. On the one hand, the guerrillas, who are well hidden in an ambush, or mixed in the civilian population, use aimed fire at the regular forces, who are fully exposed to the guerrillas. On the other hand, the regular forces are "shooting in the brown" and thus can only apply area fire on the guerrillas; the effectiveness of the regular force depends on the density of the guerrillas' live combatants. As the number of guerrillas decreases with attrition, it is harder to acquire a live target and therefore the probability of hitting a live target decreases with this number. If B is the regular force and R represents the guerrillas, then the attrition equations are:

$$\begin{aligned} \dot{B} &= -\rho R \\ \dot{R} &= -\beta B \frac{R}{R_0} \end{aligned} \tag{5}$$

and the state equation is

$$\frac{\beta}{2}(B_0^2 - B^2) = \rho(R_0^2 - R_0 R) \tag{6}$$

—a mixture of the two Lanchester laws: the Square Law and the Linear Law.

The guerrillas, who are essentially hidden, have an advantage over the regular force, which is fully exposed. This advantage is manifested in the parity condition derived from (6):

$$\frac{\rho R_0^2}{\beta B_0^2} = \frac{1}{2} \tag{7}$$

That is, ceteris paribus, Blue (the regular force) will need to double its per-capita effectiveness (kill rate) or increase its initial force size by $\sqrt{2}$ to achieve parity with Red (the guerrillas). Deitchman model was extended by Schaffer [18] who used the model for analyzing new combat hardware. We further expand on this model in Section 4.

Another asymmetric combat situation is manifested in attacks typically conducted in narrow passages such as in mountainous regions or bridges [19]. The defending Red force is effectively deployed in an area dominating the mouth of the passage so that it can concentrate its fire on the approaching attacking Blue force, which moves in a single column because of the topographical constraints. Thus, Red can apply direct fire from all its units, while Blue can only fire from its front moving weapon. The Lanchester equations in this scenario are:

$$\begin{aligned} \dot{B} &= -\rho R \\ \dot{R} &= -\beta. \end{aligned} \tag{8}$$

The state equation is $\beta(B_0 - B) = \frac{\rho}{2}(R_0^2 - R^2)$ and the parity condition is:

$$\frac{\rho R_0^2}{\beta B_0} = 2 \tag{9}$$

To achieve parity, the attacking and disadvantageous Blue force will need an initial force size in the order of the square of the Red initial force to achieve parity.

A different manifestation of asymmetry in Lanchester models is when the two sides employ profoundly different tactics. Consider an aimed-fire situation where a homogeneous Blue force is engaged in battle with a heterogeneous Red force comprising n units $R_1, \ldots, R_n$. The n Red units are different in terms of fire-effectiveness and vulnerability. Let $\beta_i(\rho_i)$ denote the kill rate of Blue (R_i) against R_i (Blue), $i = 1, \ldots, n$. While Red employs all its n units against Blue, the latter has a dilemma: how to dynamically allocate its fire among its n rivals? In other words, at any given time t in the battle, what fraction $\alpha_i(t)$ of its force to allocate for engaging $R_i(t)$? The Lanchester equations in this case are:

$$\begin{aligned} \dot{B}(t) &= -\sum_{i=1}^{n} \rho_i R_i(t) \\ \dot{R}_i(t) &= -\alpha_i(t)\beta_i B(t), \quad i = 1, \ldots, n, \end{aligned} \tag{10}$$

where for all t $\sum_{i=1}^{n} \alpha_i(t) = 1$.

Unlike the models presented before, which are purely descriptive, the model in (10) is prescriptive; Blue has a decision problem of how to dynamically allocate its attacking effort. In other words, the question is, for any point in time t, what are the optimal values of $\alpha_i(t)$. Lin and MacKay [20] showed that the optimal tactics for Blue is such that for any point in time t during the engagement $\alpha_{i_t}(t) = 1$ for a certain Red unit i_t. That is, Blue should not spread out its effort but rather concentrate all its fire on one Red adversary at a time. Moreover, Blue should engage the Red units in the descending order of the products $\beta_i \rho_i$. At any given time, Blue should concentrate its fire on the adversary for which the "product" of its vulnerability and threat is the highest.

4. Target Information

As mentioned earlier, Lanchester's aimed fire model assumes perfect visibility of targets on both sides, while the area fire model assumes none—both sides shoot "in the brown", such that their effectiveness depends on the density of live targets in the area. But what happens if the situation is somewhere in between? What happens if some portion of the force is visible to the other side while the rest of the force remains concealed? How does the level of situational awareness regarding the opponent's targets affect the outcome of the battle? Kress and MacKay [21] addressed this question by introducing a parameter representing the level of situational awareness present at each of the two sides. These parameters are traded off with the firepower of each side. Formally, the target information available to Blue regarding the Red force is parameterized by μ, $0 \leq \mu \leq 1$, where $\mu = 0$ implies no target information, and $\mu = 1$ implies full visibility, which means perfect information about the location and state of Red's targets. Similarly, we define the target information available to Red about Blue's targets by υ, $0 \leq \upsilon \leq 1$. The pair of ODEs in this case is:

$$\begin{aligned} \dot{B} &= -\rho R \frac{\upsilon B_0 + (1-\upsilon)B}{B_0} \\ \dot{R} &= -\beta B \frac{\mu R_0 + (1-\mu)R}{R_0}. \end{aligned} \tag{11}$$

When $\mu = \upsilon = 0$ (no situational awareness on both side), the model in (11) becomes the Lanchester area fire model in Equation (4), and when $\mu = \upsilon = 1$ (perfect visibility), the model in (11) becomes the Lanchester aimed fire model in Equation (1). When $\mu = 0$ (1), $\upsilon = 1$ (0) we obtain Deitchman's guerrilla warfare model in Equation (5). A special case of (11) is when $\upsilon = 1$, and $0 \leq \mu \leq 1$, which is called the generalized Deitchman model [21]. This model represents a contemporary counter-insurgency operations where the state forces (Blue), which are controlling the area, are fully exposed to the insurgents (Red), while the visibility of the insurgents, who adopt a "strike-and-hide" tactics, depends on the effort μ the state invests in surveillance, reconnaissance and human informants. The insurgents can partially be detected by the advanced sensors and surveillance systems of the state, and by the aid of local collaborators. In this case

$$\frac{dB}{dR} = \frac{\rho R}{\beta B(\mu + (1-\mu)R/R_0)} \tag{12}$$

and the parity condition becomes

$$\frac{\rho R_0^2}{\beta B_0^2} = \frac{1-\mu}{2}\left(1 + \frac{\mu \log \mu}{1-\mu}\right)^{-1} \tag{13}$$

Indeed, when $\mu = 0$, Equation (13) becomes Equation (7), and when $\mu = 1$, by Taylor expansion, Equation (13) is the parity condition of the Square Law $\frac{\rho R_0^2}{\beta B_0^2} = 1$. For example, if Blue's investment in information gathering capabilities is such that $\mu = 0.5$ (50% of the insurgents' targets are exposed) then $\frac{\rho R_0^2}{\beta B_0^2}$ is slightly higher than 4/5 at parity; ceteris paribus, Blue needs to increase its kill-rate by less than 25% to achieve parity. If target information is poor, say $\mu = 0.1$, then Blue needs to enhance its kill-rate by more than 65% to achieve parity. However, the required kill-rate enhancement for Blue is less than 8% if $\mu = 0.8$. In general, information is less "valuable" than kill-rate; a small increase in kill-rate is more valuable than an equivalent proportional increase in target information. The dilemma between investing in information "bits" or lethal "shots" boils down to the relative costs of these capabilities.

If the per-capita cost of bits and shots are c_1 and c_2, respectively, and Blue has a budget C for these capabilities, then its optimization problem is

$$\begin{array}{c} Max\ \beta B_0^2 \frac{1-\mu}{2}\left(1+\frac{\mu \log \mu}{1-\mu}\right)^{-1} \\ st \\ c_1\mu + c_2\beta \le C \\ 0 \le \mu \le 1,\ \beta \ge 0. \end{array} \tag{14}$$

Kaplan et al. [22] generalized the Deitchman original model even further and replaced the linear "intelligence function" $\mu + (1-\mu)R/R_0$ with a monotone non-decreasing function $p(R)$, which represents the per-shot probability of successfully acquiring and engaging a Red target. The parity condition in this case becomes:

$$\beta B_0^2 = 2\rho \int_0^{R_0} \frac{x}{p(x)} dx. \tag{15}$$

In other words, the state force (Blue) wins over the insurgents (Red) if and only if the initial state force B_0 satisfies

$$B_0 > \sqrt{\frac{2\rho}{\beta} \int_0^{R_0} \frac{x}{p(x)} dx} \equiv \hat{B}. \tag{16}$$

It is shown in [22] that in the case Equation (15) holds, the size of the surviving Blue force soldiers, after Red is annihilated, is $\sqrt{B_0^2 - \hat{B}^2}$.

5. Civilian Population During Conflict

The asymmetric models described in Section 4 apply to irregular warfare where well-organized, military forces of the state confront low-signature guerrilla fighters. These models focus on the asymmetry in information and its impact on battlefield outcome. Another crucial component in irregular-warfare scenarios is the civilian population who, on the one hand, are subject to violent actions by the guerrillas, and on the other hand, may be a source of support and provider of hiding places for the guerrillas' fighters.

Consider guerrillas (Red) who persistently attack civilians on the Blue side. The objective of the state forces (Blue), in attacking the guerrillas, is to prevent this killing from happening [22]. In other words, if Blue does not win over Red, that is, the inequality in Equation (16) is reversed, or if Blue decides against engaging Red in the first place, then Red causes k civilian casualties to Blue. Now, the decision of Blue to attack Red not only depends on whether Blue can win the battle (i.e., (16) holds) but it also depends on the total number of casualties. Obviously, if $k = 0$ then Blue has no incentive to attack. With equivalent valuation of civilian and Blue combatants, the total number (civilians and combatants) of Blue casualties $d(B_0)$, given an initial Blue force size of B_0, is

$$d(B_0) = \begin{cases} k + B_0, & B_0 \le \hat{B} \\ B_0 - \sqrt{B_0^2 - \hat{B}^2}, & B_0 > \hat{B}, \end{cases} \tag{17}$$

where $\sqrt{B_0^2 - \hat{B}^2}$ is the number of Blue combatants that survive the battle. The Blue force will attack the Red insurgents if and only if the total number of casualties following an attack is smaller than that

number absent an attack, that is, $d(B_0) < k$. In order for this to happen, the Blue initial force needs to be sufficiently large. Specifically,

$$B_0 > \begin{cases} \frac{k}{2} + \frac{\hat{B}^2}{2k}, & k \le \hat{B} \\ \hat{B}, & k > \hat{B}. \end{cases} \tag{18}$$

From Equation (18), we see that the minimum size of Blue that justifies an attack is a decreasing function of k for $k \le \hat{B}$ and constant thereafter. This means that if the benefits of the successful attack are small (k is small), it may not be worthwhile for Blue to engage Red, even if Blue has sufficient troops to successfully do it ($B_0 > \hat{B}$). Without a significantly larger force, the number of Blue soldiers lost may exceed the number of civilian casualties averted by defeating Red.

In the spirit of the model in Equation (10), suppose that Red is spread out in n strongholds in different and mutually distant geographical regions, and Blue has to decide how to allocate its forces among the strongholds [22]. However, unlike the case in Equation (10) where the Blue force is dynamically allocated, here, because of tactical and operational considerations, Blue has to decide, ab initio, how to allocate its force; the surviving Blue force from a defeat of one Red stronghold cannot reinforce attacks on other strongholds. Let $\hat{B}_i$ denote the threshold in Equation (16) for stronghold i, and suppose $k_i > \hat{B}_i, \quad i = 1, \ldots, n$. If $B_0 > \sum_{i=1}^{n} \hat{B}_i$, then Blue should engage and win all n battles and its only concern is to maximize the total number of Blue survivors. This leads to the following optimization problem:

$$\begin{aligned} Max \ & \sum_{i=1}^{n} \sqrt{x_i^2 - \hat{B}_i^2} \\ & st \\ & \sum_{i=1}^{n} x_i = B_0. \end{aligned} \tag{19}$$

Using a standard optimization technique, we obtain the optimal solution for Equation (19) [22]:

$$x_i^* = \frac{\hat{B}_i}{\sum_{j=1}^{n} \hat{B}_j} B_0, \quad i = 1, \ldots, n. \tag{20}$$

If $B_0 \le \sum_{i=1}^{n} \hat{B}_i$, the Blue force cannot engage (and win) all n strongholds and the problem for Blue boils down to selecting the strongholds to engage. This problem can be framed as a non-linear knapsack-type model where the objective is to maximize the number of casualties averted [22]:

$$\begin{aligned} Max & \sum_{i=1}^{n} k_i y_i + \sqrt{B_0^2 - \left(\sum_{i=1}^{n} \hat{B}_i y_i \right)^2} \\ & st \\ & \sum_{i=1}^{n} \hat{B}_i y_i \le B_0, \qquad y_i \in \{0, 1\}. \end{aligned} \tag{21}$$

Here, y_i is 1 if stronghold i is selected to be attacked and 0 otherwise.

In the special case where all the n regions are homogeneous, that is, $k_i = k, \ \hat{B}_i = \hat{B}, \quad i = 1, \ldots, n$, then it can be shown that the optimal number of strongholds (regions) $n* < n$ to be attacked is

$$n* = \sqrt{\frac{k^2}{\hat{B}^2 + k^2}} \frac{B_0}{\hat{B}} \tag{22}$$

which is clearly smaller than the number of battles $\left\lfloor B_0/\hat{B} \right\rfloor$ that can be fought [22]. The optimal number of Blue soldiers allocated to each battle is

$$x* = \frac{B_0}{n*} = \sqrt{\frac{\hat{B}^2 + k^2}{k^2}}\hat{B}. \tag{23}$$

A different situation where civilians may be involved in counterinsurgency scenarios was described in [23]. In this paper, the authors studied how incomplete target information (see Section 4) not only affects the ability of the state forces (Blue) to acquire and target guerrilla insurgents (Red), but also causes collateral casualties among civilians, which, in turn, increase resentment towards the government forces among civilians and thus may generate recruits that reinforce the guerrillas. As before, let B and R denote the sizes of the state forces and guerrillas, respectively. Let P denote the size of the civilian population, which is very large compared to B and R and thus is assumed to remain constant throughout. Absent any target information, the signature of the guerrillas, who are part of the civilian population, as targets is measured by R/P, which is interpreted as the probability that a randomly selected target in the population is indeed a guerrilla. If $\mu \in [0,1]$ is the level of target information available to Blue (see Section 4), then $\dot{R} = -\beta B(\mu + (1-\mu)(R/P))$ (see Equation (11)). Let $\theta(C)$ denote the recruitment rate to the guerrillas from the civilian population, where C is the rate at which collateral casualties in the population are generated. That is,

$$C = \beta B(1-\mu)(1-R/P) \tag{24}$$

where $(1-\mu)(1-R/P)$ is the fraction of Blue's fire that is mistakenly directed against innocent civilians. $\theta(C)$ is monotone increasing in C, and we assume, without loss of generality, that $\theta(0) = 0$. With a reinforcement rate α to the state forces, the Lanchester model capturing this scenario is:

$$\begin{aligned} \dot{B} &= -\rho R + \alpha \\ \dot{R} &= -\beta B(\mu + (1-\mu)(R/P) + \theta(\beta B(1-\mu)(1-(R/P))). \end{aligned} \tag{25}$$

If target information μ is constant throughout and the recruiting rate is proportional to the collateral casualties, that is, $\theta(C) = \theta C$, then it can be shown [23] that if $\mu \leq \theta/(1+\theta)$ then the insurgency cannot be eradicated, regardless of the initial force sizes and the attrition rates. If $\alpha/\rho < P(1-((1-\mu)(1+\theta))^{-1})$, then the state loses to the insurgents, and if the opposite is true, then the state forces can only contain the insurgency at a constant level $P(1-((1-\mu)(1+\theta))^{-1})$.

If $\mu > \theta/(1+\theta)$, then the insurgency wins if R_0 exceeds a threshold which depends on all the other parameters in Equation (25), and the state forces (Blue) win otherwise.

Although a constant value of target information μ is somewhat reasonable, in reality it is more likely that it is dynamically changing as a function $\mu(B,R)$ of the two forces, which is monotone non-decreasing in R and B. Assuming that both μ and θ are continuously differentiable, considering the second equation in Equation (25) and using the implicit function theorem, there is a continuously differentiable function $r(B)$ satisfying

$$\begin{aligned} &\beta B(\mu(B,r(B)) + (1-\mu(B,r(B)))(r(B)/P)) \\ &-\theta(\beta B(1-\mu(B,r(B)))(1-r(B)/P)) = 0. \end{aligned} \tag{26}$$

Figure 1 plots a possible shape of the function $r(B)$. This function separates between the bottom region where the insurgency grows ($\dot{R} > 0$) and the upper region where it gets smaller. Also, the line $R = \frac{\alpha}{\rho}$ separates between the upper region where the state forces decrease in strength and the lower region where their forces increase. If for some range of C the recruitment to the insurgency accelerates with the number of per-unit-time collateral casualties, and the growth more than makes up the increase in attrition of the insurgents, then $r(B)$ may increase, as shown in Figure 1. However, as the number of collateral casualties increases, the number of recruits ebbs down due to the bounded size of the

population. It is shown [23] that $r(B) > 0$ for all B. That is, the insurgency can never be physically eradicated. Any trajectory in the $B \times R$ space that crosses the $r(B)$ curve from above is destined to bounce back. The best the state forces can do is contain the insurgency at a certain low level. The operational explanation for this is that when the insurgency R is small, the target information available to the state forces is poor, which leads to inadvertent collateral casualties among the civilian population when the state forces attack the insurgents. These innocent casualties cause anger among civilians, which generates new recruits to the insurgents. The increased attrition generated by more state forces is offset by the new recruits.

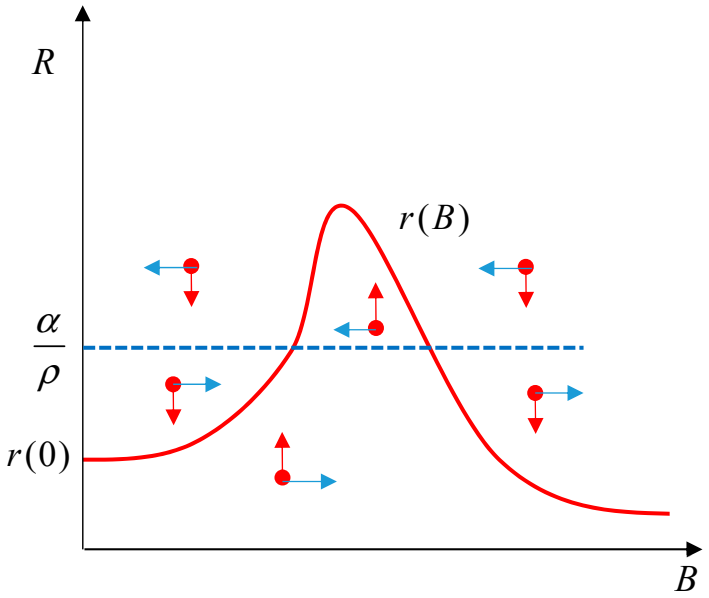


Figure 1. The function $r(B)$ and the regions in which state forces and insurgents increase or decrease.

One possible extension of the model in Equation (25) is to assume that the state forces (B) determine their rate of reinforcement by observing and responding to the size of the guerrillas (R). Thus α in Equation (25) is replaced by $\alpha(R)$. Suppose this function is linear; that is, $\alpha(R) = aR + b$, $(b \geq 0)$. In that case, if $\rho > a$, then we are back to the situation in (25) where ρ is simply replaced by $\rho - a$. If $\rho \leq a$, then the state forces keep growing and the guerrillas are led to their demise at a huge collateral cost to the civilian population in which the guerrillas are embedded. We conjecture that if $\alpha(R)$ is non-linear, then multiple equilibria may exist, depending on the shape of that function.

6. Multilateral Conflicts

As described thus far, legacy Lanchester equations essentially model the attrition between two opposing forces. They capture a duel, force-on-force, situation. However, recent, as well as some historical, conflicts involve more than two opposing forces. The Bosnian Civil War (Croatia, Bosnia Herzegovina, Serbia, NATO), the Iraq Civil War (Coalition Forces, Sunni Militia, Shia Militia), and most recently, the war in Syria (Assad Regime Forces, Free Syrian Army, Hezbollah, Kurds, Russia, Turkey) are just a few examples of such multilateral violent conflicts. Two recent papers extend the classical Lanchester theory to the case where the attritional conflict comprises more than two players. It is important to note a profound difference between two- and multiple-player Lanchester models. In a two-player (force-on-force) conflict, the legacy Lanchester models (i.e., Equations (1), (3) and (4) above) are purely descriptive; they simply capture the attrition on both sides as a function of the initial strengths (B_0, R_0) and the attrition rates (β, ρ) of the two players: Blue and Red. No decision is required, by either player, during the engagement. However, in a multiple-player conflict, each player has to decide how to allocate its strength among the other adversaries so as to maximize its own chances to be the victor. This decision, common to all other players, leads to a prescriptive model

where each one of n players ($n > 2$) has to dynamically allocate its existing strength among its $n-1$ adversaries. While the results are general, the rest of this section mostly focuses on the case $n = 3$.

The three-player (Blue, Red and Green) Lanchester direct-fire model is:

$$\begin{aligned} \dot{B}(t) &= -\alpha_{RB}(t)\rho_B R(t) - \alpha_{GB}(t)\gamma_B G(t) \\ \dot{R}(t) &= -\alpha_{BR}(t)\beta_R B(t) - \alpha_{GR}(t)\gamma_R G(t) \\ \dot{G}(t) &= -\alpha_{BG}(t)\beta_G B(t) - \alpha_{RG}(t)\rho_G R(t) \end{aligned} \tag{27}$$

where $\alpha_{BR}(t) + \alpha_{BG}(t) = \alpha_{RB}(t) + \alpha_{RG}(t) = \alpha_{GR}(t) + \alpha_{GB}(t) = 1, \quad \alpha_{ij}(t) \geq 0, \;\; i, j = B, R, G$ for all t.

Here ρ_B (ρ_G) is the effectiveness (attrition rate) of Red against Blue (Green). Similar notation applies to the other attrition rates of Blue (β) and Green (γ). The coefficients $\alpha_{ij}(t)$ are control variables indicating the fraction of the standing force of player i that should be directed against player j at time t, $i, j = B, R, G$.

Kress et. al. [24] considered the case where, due to tactical and geographical constraints, the three players have to determine their respective force allocations at the beginning of the battle, and they cannot change their allocations after the battle begins. That is, $\alpha_{ij}(t) = \alpha_{ij}$ for all t. The battle comprises two stages. The first stage is when all three players are alive and each engages the forces of the other two players according to its pre-determined force-allocation. There are three possible outcomes to this stage: (a) mutual annihilation (b) two of the three players are annihilated and a clear victor emerges, or (c) one player is annihilated and the battle transitions to the second stage where the two remaining players engage in a force-on-force battle as in Equation (1).

Without loss of generality, we assume that the initial forces of the three players are normalized; that is, $B_0 + R_0 + G_0 = 1$. Thus, the initial force sizes lie in the unit 2-simplex denoted by S.

We note that a simple characterization like the Square Law of the aimed-fire model (see the discussion following Equation (2)) does not exist for the multilateral case. However, for a given set of parameters α, β, ρ and γ, we can plot the regions of the initial forces that lead to victory, as shown in Figure 2. The starred point at the center is the point of mutual annihilation; the initial forces of Blue, Red and Green that are represented by this point lead to the demise of all. This is case (a) described above as a possible outcome of the first stage of the battle. This point is the three-player equivalent of the parity condition described in Section 2 regarding the force-on-force aimed-fire model. In that case, the equivalent annihilation point is the solution B_0 of $\beta B_0^2 = \rho(1 - B_0)^2$ where $R_0 = 1 - B_0$. The bold lines separate among the regions in which a player is the ultimate victor of the battle. The lower right region is where Blue is the victor, the upper region is where Red wins and the area close to the origin is where Green wins. Within a "win region", the dotted line separates between the two defeated players of the first stage. For example, the B region in which Blue is the final winner, the bottom sub-region, denoted R 1st, is where Red is defeated at the first stage. Initial force sizes located on a bold line lead to mutual annihilation, at the second stage, by the two adjacent players, after the third player was defeated at the first stage. Similarly, a point on a dotted line corresponds to mutual annihilation of the two adjacent players at the first stage. For example, a point on the dotted line in region R corresponds to the scenario in which Red defeated both Blue and Green at the first stage.

Suppose that Blue and Green are fierce opponents and both of them allocate an equal and large share of their force one against the other. Only a small fraction of their force is deployed against Red. Red is flexible to select how to balance its allocation of force against its two opponents. Figure 3 presents the winning regions of the three forces. In region i, $i = B, R, G$, player i wins regardless of the force allocation by the other players. In region R^*, Red can win if it optimizes its force allocation (α_{RB}, α_{RG}). In region X, Red loses but it is the "victor maker"; it can determine the winner between Blue and Green.

Perhaps a more interesting question is what happens when each player can dynamically and continuously decide its force allocation between its two opponents. Kress et al. [25] studied this question as a differential game where each player wishes to maximize its own surviving force minus

that of its enemies. The outcome of the analysis is surprising: either a player is strong enough to win over the other players combined in a coalition against itself, or all players are locked in a stalemate that leads to their mutual demise. In the case of three players, this conclusion stands in contrast to sequential-engagement scenarios in which the weakest player can achieve an advantage [26].

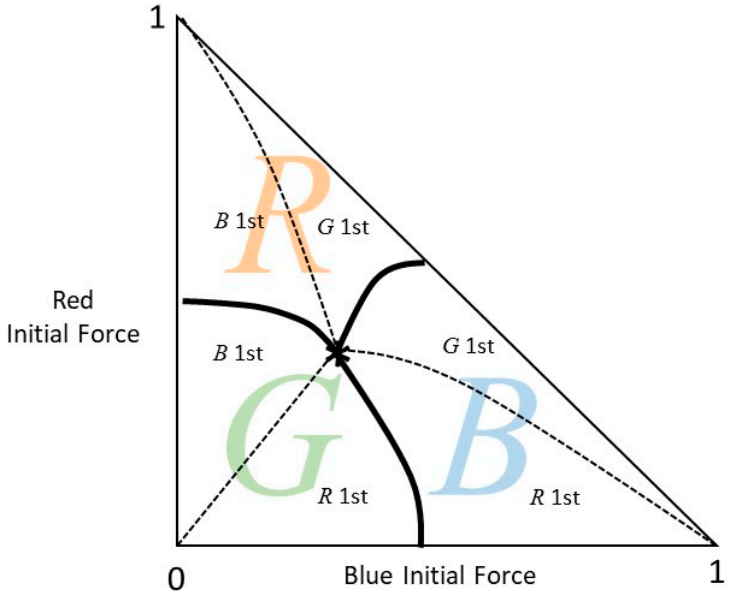


Figure 2. Winning regions. The solid curves represent points of mutual annihilation at the second stage by the two adjacent players, while the dotted lines indicate points of mutual annihilation at the first stage.

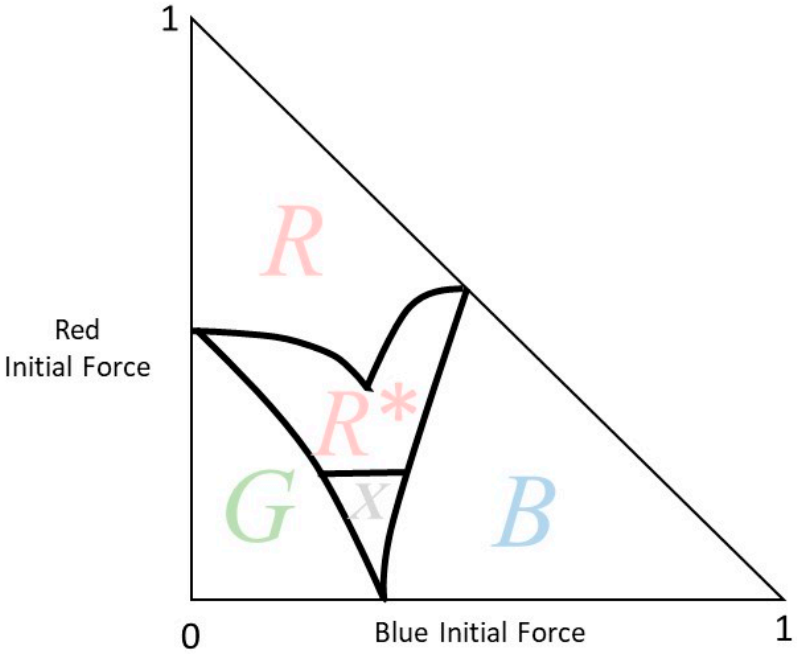


Figure 3. Unconditional winning regions, optimized winning region, and "victor maker" region for the case where the main battle is between Blue and Green.

We specialize now the notion of "win region", described above for the static fire-allocation case, and say that a player is dominant if it can defeat the alliance of all other players, regardless of the

fire-allocation of the members of the alliance. According to Lemma 1 in [25], in the special case of $n = 3$, if, without loss of generality, $\beta_R \rho_B \leq \beta_G \gamma_B$ then Blue is dominant if and only if

$$B_0^2 > \frac{\rho_B}{\beta_R} R_0^2 + \frac{\gamma_B}{\beta_G} G_0^2 + 2\frac{\rho_B}{\beta_G} R_0 G_0 \tag{28}$$

Blue is pseudo-dominant if Equation (28) is changed to equality.

The condition in Equation (28), when applied to each player, divides the non-negative quadrant into four disjoint regions D_B, D_R, D_G, N such that player i is dominant in region D_i, i = Blue, Red, Green, and the complement region N is the non-dominant region in which no player is dominant (see Figure 4). For example, OAQR marks the region D_G where Green is dominant. The surface OQR separates Green's dominant region from N. Similarly, OQP and ORP separate D_B and D_R from N, respectively. The initial states on the line OQ are where $R_0 = 0$ and B_0, G_0 are such that the duel between Blue and Green heads for mutual annihilation.

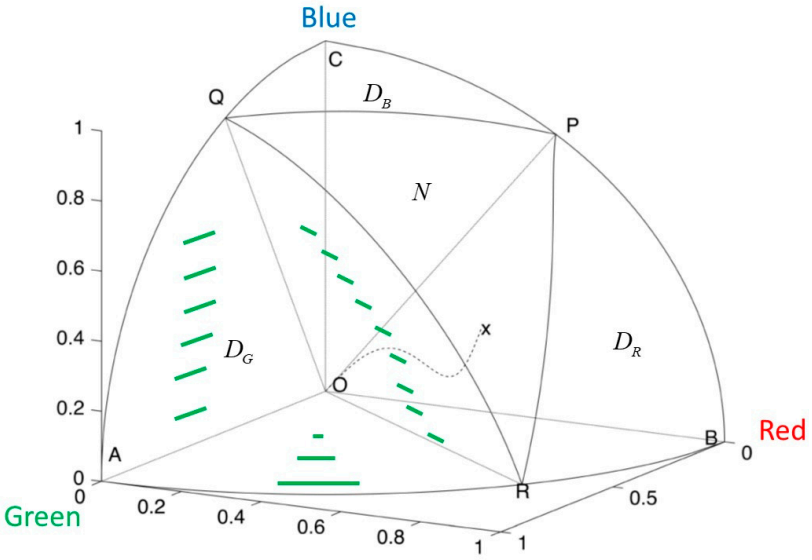


Figure 4. The case of $n = 3$ players where x marks initial force relative sizes and the dashed trajectory indicates the path to mutual annihilation.

If a state belongs to a dominant region, then the corresponding dominant player will use the optimal strategy, described at the end of Section 3 for the battle formalized in Equation (10), to guarantee a win. As for an initial state in region N, it is shown in [25] that for each such state there exists a fire-allocation α that leads to mutual annihilation. It is always possible, and fairly easy, to find fire strategies $\alpha(B(t), R(t), G(t))$ that do not shift the current balance of power, by keeping the ratios $B(t)/R(t), B(t)/G(t))$ and $R(t)/G(t)$ fixed throughout. If all players adopt such a strategy, then the resulting force trajectory is a straight line from x in Figure 4 toward (0,0,0). If some player tries to outwit the other players and divert from the status quo, then the relative force sizes may change over time, but such attempts are futile. Defining an appropriate n-person nonzero-sum game and using Nash equilibria, it is shown that such a curve must end up at the origin (see the dotted curve in Figure 4).

In conclusion, the insight from multilateral Lanchester conflicts is clear: either a single player is strong enough to beat all other opponents combined, or all players are destined to a prolonged attritional stalemate that culminates in mutual annihilation. The prolonged war in Syria is an example of such dynamics. Only the appearance of a dominant player (e.g., Russia) can end it with a victory.

7. Summary

Lanchester models of warfare have been around for over a century. They have played a major role in modeling and analyzing regular force-on-force engagements, in particular battles during

WWII. Recent conflict situations are not regular in the sense that they are profoundly asymmetric, they increasingly rely on data and information, they are affected by the behavior of civilians and they may involve more than two adversaries.

In this paper, we reviewed recent advances in modeling the aforementioned features of modern combat situations in the context of Lanchester theory. We described some important insights regarding the fate of these situations, as obtained from implementing such models. The main takeaways are: (a) guerrilla forces have an inherent advantage over state forces because the latter suffer from reduced level of situational awareness and are reluctant to massively hurt civilians among which guerrillas hide, (b) as a consequence of (a), the state forces cannot eradicate an insurgency unless it completely ignores civilian casualties, (c) in a situation of multiple adversaries, the state forces have an optimal target allocation plan that depends on the effectiveness of the state forces and the vulnerability of the adversaries, (d) in multilateral conflicts, either there exists a clear victorious player who wins the conflict even if all other sides collaborate in a coalition, or all players are dragged into a prolonged conflict with no victor.

Future conflicts and combat situations may include several characteristics that lend themselves to new Lanchester-type modeling. First, "soft kills", such as electronic and information warfare, will become more prevalent in future conflicts. This type of attrition is profoundly different than the legacy "hard kills" in which attrition is irreversible. Lanchester models accounting for a mix of soft and hard kills can help design policies and analyze the tradeoff between the two ways of projecting and enduring military force. Second, future combat will rely on the increased use of unmanned systems, which may change the battlefield landscape in the absence of human fear of death. "Attrition" will have a somewhat different meaning, and modeling it within a Lanchester framework would be challenging. Finally, the possible use of biological weapons of mass destruction may trigger an epidemic with significant effect on conflict outcomes. Combinations of two dynamic models—Lanchester and epidemic spread models such as SIR—will be needed to study such complex attritional situations.

Funding: This paper received no external funding.

Conflicts of Interest: The author declare no conflict of interest.

References

1. Washburn, A.; Kress, M. *Combat Modeling*; Springer: New York, NY, USA, 2009.
2. Kress, M. Modeling armed conflicts. *Science* **2012**, *336*, 865–869. [CrossRef] [PubMed]
3. Lanchester, F.W. *Aircraft in Warfare: The Dawn of the Fourth Arm*; Constable and Co.: London, UK, 1916.
4. Brown, G.; Washburn, A. The Fast Theater Model (FATHM). *Mil. Oper. Res.* **2007**, *12*, 33–45. [CrossRef]
5. Jones, H.W. *COSAGE User's Manual, Volume1-MainReport. Revision 4*; CAA-D-93-1-Rev-4; Center for Army Analyses: Fort Belvoir, VA, USA, 1995.
6. Sahni, M.; Das, S.K. Performance of maximum likelihood estimator for fitting Lanchester equations on Kursk Battle data. *J. Battlef. Technol.* **2015**, *18*, 23–30.
7. Lucas, T.W.; Turkes, T. Fitting Lanchester equations to the battles of Kursk and Ardennes. *Nav. Res. Logist.* **2004**, *51*, 95–116. [CrossRef]
8. MacKay, N.J.; Price, C.; Wood, A.J. Weight of Shell Must Tell: A Lanchestrian reappraisal of the Battle of Jutland. *History* **2016**, *101*, 536–563. [CrossRef]
9. Keane, T. Combat modelling with partial differential equations. *Appl. Math. Model.* **2011**, *35*, 2723–2735. [CrossRef]
10. Kim, D.; Moon, H.; Park, D.; Shin, H. An efficient approximate solution for stochastic. *J. Oper. Res. Soc.* **2017**, *68*, 1470–1481. [CrossRef]
11. Eldrige, S.A.; Mesterton-Gibbons, M. Lanchester's attrition models and fights among social animals. *Behav. Ecol.* **2003**, *14*, 719–723.
12. Johnson, D.D.P.; MacKay, N.J. Fight the Power: Lanchester's laws of combat in human evolution. *Evol. Hum. Behav.* **2013**, *36*, 152–163. [CrossRef]

13. Jorgensen, S.; Sigue, S. A Lanchester-Type Dynamic Game of Advertising and Pricing. In *Games in Management Science*; Sigue, P.-O., Taboubi, S., Pineau, S., Eds.; Springer: New York, NY, USA, 2020; pp. 1–14.
14. Stanescu, M.; Barriga, N.; Buro, M. Using Lanchester Attrition Laws for Combat Prediction in StarCraft. In Proceedings of the Eleventh AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE-15), Santa Cruz, CA, USA, 14–18 November 2015; pp. 86–92.
15. Deitchman, S. A Lanchester model of guerrilla war. *Oper. Res.* **1962**, *10*, 818–827. [CrossRef]
16. Helmbold, R.; Rehm, A. Translation of "The Influence of the Numerical Strength of Engaged Forces in Their Casualties". In *Naval Research Logistics*; Osipov, M., Ed.; Wiley: Hoboken, NJ, USA, 1995; Volume 42, pp. 435–490.
17. Taylor, J. *Lanchester Models of Warfare*; INFORMS: Rockville, MD, USA, 1983.
18. Schaffer, M.B. Lanchester models of Guerrillla engagements. *Oper. Res.* **1968**, *16*, 457–488. [CrossRef]
19. Gabriel, R.A. Lessons of war: The IDF in Lebanon. *Mil. Rev.* **1984**, *64*, 47–65.
20. Lin, K.Y.; MacKay, N.J. The optimal policy for the one-against-many heterogeneous lanchester model. *Oper. Res. Lett.* **2014**, *42*, 473–477. [CrossRef]
21. Kress, M.; MacKay, N. Bits or Shots in Combat? The Generalized Deitchman Model of Guerrilla Warfare. *Oper. Res. Lett.* **2014**, *42*, 102–108. [CrossRef]
22. Kaplan, E.; Kress, M.; Szechtman, R. Confronting Entrenched Insurgents. *Oper. Res.* **2010**, *58*, 329–341. [CrossRef]
23. Kress, M.; Szechtman, R. Why Defeating Insurgencies is Hard: The Effect of Intelligence in Counterinsurgency Operations—A Best Case Scenario. *Oper. Res.* **2009**, *57*, 578–585. [CrossRef]
24. Kress, M.; Caulkins, J.P.; Feichtinger, G.; Grass, D.; Seidl, A. Lanchester model for three-way combat. *Eur. J. Oper. Res.* **2018**, *264*, 46–54. [CrossRef]
25. Kress, M.; Lin, K.; MacKay, N. The Attrition Dynamics of Multilateral War. *Oper. Res.* **2018**, *66*, 950–956. [CrossRef]
26. Kilgour, D.M.; Brams, S.J. The truel. *Math. Mag.* **1997**, *70*, 315. [CrossRef]

J. Math. Anal. Appl. ••• (••••) ••••••

Contents lists available at ScienceDirect

Journal of Mathematical Analysis and Applications

www.elsevier.com/locate/jmaa

Modelling COVID-19 transmission in the United States through interstate and foreign travels and evaluating impact of governmental public health interventions

Nita H. Shah [a], Nisha Sheoran [a], Ekta Jayswal [a], Dhairya Shukla [b], Nehal Shukla [c], Jagdish Shukla [d,*], Yash Shah [e]

[a] *Department of Mathematics, Gujarat University, Ahmedabad, 380009, Gujarat, India*
[b] *Medical College of Georgia, 1120, 15th St, Augusta, GA 30912, USA*
[c] *Department of Mathematics, Columbus State University, 4225, University Avenue, Columbus, GA 31907, USA*
[d] *Department of Medical Education, Family Medicine Residency Program, 1900, 10th Avenue, Columbus, GA 31901, USA*
[e] *GCS Medical College, Ahmedabad, 380054, Gujarat, India*

ARTICLE INFO

Article history:
Received 26 May 2020
Available online xxxx
Submitted by S.G. Krantz

Keywords:
Covid-19
Travel
Transmission
Public health

ABSTRACT

Background: The first case of COVID-19 was reported in Wuhan, China in December 2019. The disease has spread to 210 countries and has been labelled as a pandemic by the World Health Organization (WHO). Modelling, evaluating, and predicting the rate of disease transmission is crucial in understanding optimal methods for prevention and control. Our aim is to assess the impact of interstate and foreign travel and public health interventions implemented by the United States government in response to the COVID-19 pandemic. **Methods**: A disjoint mutually exclusive compartmental model was developed to study transmission dynamics of the novel coronavirus. A system of nonlinear differential equations was formulated and the basic reproduction number R_0 was computed. Stability of the model was evaluated at the equilibrium points. Optimal controls were applied in the form of travel restrictions and quarantine. Numerical simulations were conducted. **Results**: Analysis shows that the model is locally asymptomatically stable, at endemic and foreigners free equilibrium points. Without any mitigation measures, infectivity and subsequent hospitalization of the population increased. When interstate and foreign travel was restricted and the population placed under quarantine, the probability of exposure and subsequent infection decreased significantly; furthermore, the recovery rate increased substantially. **Conclusion**: Interstate and foreign travel restrictions, in addition to quarantine, are necessary in effectively controlling the pandemic. The United States has controlled COVID-19 spread by implementing quarantine and restricting foreign travel. The government can further strengthen restrictions and reduce spread within the nation more effectively by implementing restrictions on interstate travel.

* Corresponding author.
E-mail addresses: nitahshah@gmail.com (N.H. Shah), sheorannisha@gmail.com (N. Sheoran), jagdish.shukla@piedmont.org (J. Shukla).

https://doi.org/10.1016/j.jmaa.2020.124896
0022-247X/

Table 1
Timeline of public health intervention implemented by the United States government.

Date	Action
17th *Jan*	Public health entry screening at 3 U.S. Airport.
31st *Jan*	Coronavirus declared Public health emergency, Chinese travel restrictions, restricted entry into U.S.A for foreign nationals who pose a risk of transmission, Funnel all flights from China to just 7 U.S. domestic airports.
29th *Feb*	Barred all travel to Iran, level 4 travel advisory to areas of Italy and South Korea.
11th *March*	Travel restriction for foreigners who visited Europe in the last 14 days.
14th *March*	Europe travel ban extended to UK and Ireland. Implement Social Distancing and Closure of teaching institutes in many states.
18th *March*	Temporary closure of U.S. Canada border for non-essential traffic.
19th *March*	Americans to avoid all international travel.
20th *March*	The U.S. and Mexico agreed to restrict non-essential cross border traffic. Closure of non-essential businesses and shelter in place order in NY.
24th *March*	Self-quarantine for 14 days for individuals who recently visited New York.
28th *March*	For residents in NY, NJ, CT, Avoid non-essential domestic travel for 2 weeks.
29th *March*	Social distancing extended through 30th April.
3rd *April*	All American wear non-medical, fabric or cloth masks to prevent asymptomatic spread of coronavirus.

1. Introduction

In December 2019, an unidentified pneumonia was found in Wuhan, Hubei province, China. The responsible virus was later identified as the novel coronavirus 2019, and the disease as coronavirus disease 2019 (COVID-19) [4]. COVID-19 is transmitted via direct contact with an infected person through respiratory droplets when a person coughs or sneezes, or by indirect contact with contaminated surfaces with respiratory droplets from infected person and then touching their eyes, nose or mouth [6]. Furthermore, the virus remains on surfaces from a few hours to several days and has an incubation period between 1-14 days. Consequently, the disease spread rapidly from Wuhan to all parts of the country and overseas.

Introduction and spread of COVID-19 within the United States is a direct result of transmission through foreign and interstate travel. The first known case of COVID-19 in the U.S.A. was confirmed on 20th January 2020 in a 35-year-old individual who had travelled from Wuhan to Washington state [6]. Soon, cases started appearing and rising in many other states of the U.S.A. including New York, New Jersey, Illinois, Florida, Georgia, Texas, Pennsylvania due to interstate travel. The CDC alarmed that hospitals may get overwhelmed by a large number of people seeking care at the same time due to widespread transmission of disease which may lead to otherwise preventable deaths (2020) [2]. In response to the oncoming epidemic the US government implemented the following regulations (Table 1).

While the United States has implemented numerous public health interventions, it has not implemented a ban on interstate travel. According to the World Health Organization (WHO) [12]. New cases of COVID-19 have emerged in 210 countries with 1,733, 945 confirmed cases and 106, 518 confirmed deaths globally as of 10th April 2020. The United States has implemented quarantine measures, close contact tracing, early testing for individuals with symptoms, hospitalization if needed, and closing of teaching institutes and non-essential businesses. Studies have shown that in other countries, the complete lockdown of travel has decreased the spread of the disease in the surrounding states ([1]; [10]). In order to prevent the transmission of COVID-19

Please cite this article in press as: N.H. Shah et al., Modelling COVID-19 transmission in the United States through interstate and foreign travels and evaluating impact of governmental public health interventions, J. Math. Anal. Appl. (2021), https://doi.org/10.1016/j.jmaa.2020.124896

Table 2
Model parameters and their interpretation.

Notation	Parameter description	Parametric values
B	Birth rate	0.0181
β_1	Rate at which U.S. population gets exposed to COVID-19 via interstate travel	0.0011
β_2	Rate at which U.S. population gets exposed to COVID-19 through contact with foreigners	0.0003
β_3	Rate of COVID-19 exposure through foreigners engaged in interstate travel	0.0012
β_4	Rate at which Interstate population goes for quarantine	0.0018
β_5	Rate at interstate population gets infected	0.0035
β_6	Rate at which foreigner quarantine themselves	0.000015
β_7	Rate at which foreigner gets infected	0.000001
β_8	Rate at which quarantine humans gets infected by COVID-19	0.0025
β_9	Rate at which infected humans gets hospitalized	0.0037
β_{10}	Rate at which exposed humans gets hospitalized	0.000002
β_{11}	Rate at which hospitalized humans gets recovered	0.0000003
μ	Death rate	0.000119
μ_c	Death rate due to COVID-19	0.0027

within the US, the mode of transmission must first be modelled and understood. Mathematical modelling is ideal for evaluating and predicting the rate of disease transmission. Data-driven mathematical modelling plays an important role in epidemic mitigation, in preparedness for future epidemic and in the evaluation of control effectiveness [13]. In this study, we adopted a disjoint mutually exclusive compartmental model to shed light on the transmission dynamics from foreign and interstate travel of the novel coronavirus and our aim is to assess the impact of public health interventions on infection by measuring basic reproduction number, contact rate, newly confirmed cases, total confirmed cases, total death. Our estimated parameters are largely in line with World Health Organization estimates and previous studies (2019).

2. Mathematical modelling

Mathematical modelling plays a vital role in determining dynamics of diseases. In this paper we consider a disjoint mutually exclusive compartmental model with compartments as follows: Exposed to COVID-19 E i.e. this compartment consists of individuals (Both foreign population and interstate population) who are exposed to COVID-19, next compartment is I_S i.e. transmission of COVID-19 through interstate travel, COVID-19 transmission through foreigners F - this compartment includes U.S. population which is exposed to COVID-19, Quarantined class Q, COVID-19 Infected I - this class includes infected population as well infectious population, hospitalized H - this class includes hospitalization of both COVID-19 infectives and also those who are exposed to COVID-19 and last compartment includes recovered population from hospitalized population denoted by R. Notations and parametric values used in the formulation of dynamical system model are given in the following Table 2.

This model considers new recruitment in the exposed class at the rate B and all the compartments have mortality rate μ. Here β_1 is the US population exposed to COVID-19 via interstate travel, and β_2 is the rate at which the US population gets exposed to COVID-19 through contact with foreigners. Next, β_3 is the rate of COVID-19 exposure through foreigners engaged in interstate travel. The US population engaging in interstate travel and foreigners quarantine themselves at the rate β_4 and β_6 respectively. Similarly, after getting exposed to COVID-19, US population engaging in interstate travel and foreigner population gets the infection joining infectious class I with the rate β_5 and β_7 respectively. Quarantined humans also get the infection at the rate β_8. Next, we assume infected population gets hospitalized at the rate β_9 joining H. We also assume population gets admitted to the hospital at the initial exposure of the disease at the rate

Please cite this article in press as: N.H. Shah et al., Modelling COVID-19 transmission in the United States through interstate and foreign travels and evaluating impact of governmental public health interventions, J. Math. Anal. Appl. (2021), https://doi.org/10.1016/j.jmaa.2020.124896

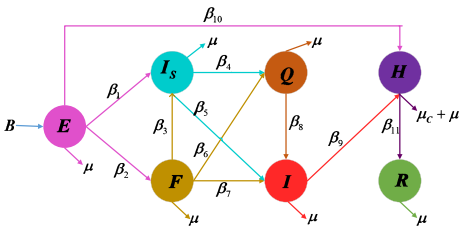


Fig. 1. Compartmental diagram showing movement of individuals from one compartment to another compartment.

β_{10}. Hospitalized patients after undergoing treatment gets recover joining R at the rate β_{11}. We take into consideration death due to COVID-19 μ_c, when the individual is hospitalized.

The Fig. 1 gives rise to the following set of non-linear ordinary differential equations

$$\begin{aligned}
\frac{dE}{dt} &= B - \beta_1 E I_S - \beta_2 E F - \beta_{10} E - \mu E \\
\frac{dI_S}{dt} &= \beta_1 E I_S + \beta_3 F - (\beta_4 + \beta_5 + \mu) I_S \\
\frac{dF}{dt} &= \beta_2 E F - (\beta_3 + \beta_6 + \beta_7 + \mu) F \\
\frac{dQ}{dt} &= \beta_4 I_S + \beta_6 F - (\beta_8 + \mu) Q \\
\frac{dI}{dt} &= \beta_5 I_S + \beta_7 F + \beta_8 Q - (\beta_9 + \mu) I_M \\
\frac{dH}{dt} &= \beta_9 I + \beta_{10} E - (\beta_{11} + \mu_C + \mu) I_M \\
\frac{dR}{dt} &= \beta_{11} H - \mu R
\end{aligned} \tag{1}$$

where, $N(t) = E(t) + I_S(t) + F(t) + Q(t) + I(t) + H(t) + R(t)$.

Adding all the differential equations of model, we get,

$$\frac{dN}{dt} \leq B - \mu(E + I_S + F + Q + I + H + R) \geq 0$$

Hence, $\frac{dN}{dt} \leq B - \mu N$.

So that $\lim_{t\to\infty} \text{supN} \leq \frac{B}{\mu}$.

Then, Feasible Region for the system is defined as

$$\Lambda = \left\{ (E, I_S, F, Q, I, H, R); E + I_S + F + Q + I + H + R \leq \frac{B}{\mu}, \right\} \tag{2}$$

with $E > 0, I_S > 0, F > 0, Q > 0, I > 0, H > 0, R > 0$.

This system has following equilibrium points

i. Foreigner free equilibrium point

ii. Endemic equilibrium point

3. Reproduction number

Basic reproduction number is defined as the total number of secondary infections in a total susceptible population. Here, we calculate the reproduction number using Diekmann et al., when the disease in its

Please cite this article in press as: N.H. Shah et al., Modelling COVID-19 transmission in the United States through interstate and foreign travels and evaluating impact of governmental public health interventions, J. Math. Anal. Appl. (2021), https://doi.org/10.1016/j.jmaa.2020.124896

endemic stage i.e. for this model it is defined as percentage of population infected by a single infection in a totally exposure situation [5]. We also compute the value of reproduction number R_F when there are no foreigners present in the total population.

$$F_1 = \begin{bmatrix} \beta_1 E & 0 & 0 & 0 & 0 & 0 & \beta_1 I_S \\ 0 & \beta_2 E & 0 & 0 & 0 & 0 & \beta_2 F \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$V_1 = \begin{bmatrix} \beta_4+\beta_5+\mu & -\beta_3 & 0 & 0 & 0 & 0 & 0 \\ 0 & l1 & 0 & 0 & 0 & 0 & 0 \\ -\beta_4 & -\beta_6 & \beta_8+\mu & 0 & 0 & 0 & 0 \\ -\beta_5 & -\beta_7 & -\beta_8 & \beta_9+\mu & 0 & 0 & 0 \\ 0 & 0 & 0 & -\beta_9 & \beta_{11}+\mu+\mu_C & 0 & -\beta_{10} \\ 0 & 0 & 0 & 0 & -\beta_{11} & \mu & 0 \\ \beta_1 E & \beta_2 E & 0 & 0 & 0 & 0 & l2 \end{bmatrix}$$

where, $l1 = \beta_3 + \beta_6 + \beta_7 + \mu$, $l2 = \beta_1 I_S + \beta_2 F + \beta_{10} + \mu$.

The reproduction number R_{E^*}, R_F is the spectral radius of $F_1 {V_1}^{-1}(E^*)$ and $F_1 {V_1}^{-1}(E^F)$ respectively. The value of $R_{E^*} = 81\%$ and $R_F = 1.10$.

4. Stability analysis

In this section we study the stability analysis of the model. Here we study Local stability of all the equilibrium points using Routh-Hurwitz criterion by Routh 1877 [9].

Theorem 1. *The foreigner free equilibrium point is locally asymptotically stable if* $(\beta_4 + \beta_5 + \mu) < \max\left\{\frac{B\beta_1}{\beta_{10}+\mu}, \frac{\beta_1(\beta_3+\beta_6+\beta_7+\mu)}{\beta_2}\right\}$.

Proof. The Jacobian of system (1) at Foreigner free equilibrium is as follows

$$J^F = \begin{bmatrix} -t_1-\beta_{10}-\mu & -(\beta_4+\beta_5+\mu) & \frac{-\beta_2(\beta_4+\beta_5+\mu)}{\beta_1} & 0 & 0 & 0 & 0 \\ t_1 & 0 & \beta_3 & 0 & 0 & 0 & 0 \\ 0 & 0 & t_2 & 0 & 0 & 0 & 0 \\ 0 & \beta_4 & \beta_6 & -t_3 & 0 & 0 & 0 \\ 0 & \beta_5 & \beta_7 & \beta_8 & -t_4 & 0 & 0 \\ \beta_{10} & 0 & 0 & 0 & \beta_9 & -t_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & \beta_{11} & -\mu \end{bmatrix}$$

$$t_1 = \frac{B\beta_1 - (\beta_{10}+\mu)(\beta_4+\beta_5+\mu)}{(\beta_4+\beta_5+\mu)}, \qquad t_2 = \frac{\beta_2(\beta_4+\beta_5+\mu)}{\beta_1} - (\beta_3+\beta_6+\beta_7+\mu),$$

$$t_3 = (\beta_8+\mu), \qquad t_5 = (\beta_{11}+\mu+\mu_C).$$

The eigen values of the Jacobian J^F are

$$\lambda_1 = t_2,\ \lambda_2 = -(\beta_8+\mu),\ \lambda_3 = -(\beta_9+\mu),\ \lambda_4 = -\mu,\ \lambda_5 = -(\beta_{11}+\mu+\mu_C),$$

$$\lambda_{6,7} = -\frac{1}{2}\left(t_1+\beta_{10}+\mu \pm \sqrt{\xi}\right),\ \xi = (\beta_{10}+\mu)^2 + 2\beta_{10}t_1 - 4t_1(\beta_4+\beta_5) - 2\mu t_1 + t_1^2$$

If it has imaginary roots i.e. $\xi < 0$. Then we have negative real part. Hence the theorem. But if $\xi \geq 0$, then eigen values are negative if $(\beta_4 + \beta_5 + \mu) < \max\left\{\frac{B\beta_1}{\beta_{10}+\mu}, \frac{\beta_1(\beta_3+\beta_6+\beta_7+\mu)}{\beta_2}\right\}$. Hence the Foreigner free equilibrium point is locally asymptotically stable.

Please cite this article in press as: N.H. Shah et al., Modelling COVID-19 transmission in the United States through interstate and foreign travels and evaluating impact of governmental public health interventions, J. Math. Anal. Appl. (2021), https://doi.org/10.1016/j.jmaa.2020.124896

Theorem 2. *The endemic equilibrium point is locally asymptotically stable if* $E^* < \max\left\{\frac{\beta_4+\beta_5+\mu}{\beta_1}, \frac{\beta_3+\beta_6+\beta_7+\mu}{\beta_2}\right\}$.

Proof. The Jacobian matrix of system (1) for endemic equilibrium is given by

$$J^* = \begin{bmatrix} -a_{11} & -\beta_1 E^* & -\beta_2 E^* & 0 & 0 & 0 & 0 \\ \beta_1 I_S^* & -a_{22} & \beta_3 & 0 & 0 & 0 & 0 \\ \beta_2 F^* & 0 & -a_{33} & 0 & 0 & 0 & 0 \\ 0 & \beta_4 & \beta_6 & -a_{44} & 0 & 0 & 0 \\ 0 & \beta_5 & \beta_7 & \beta_8 & -a_{55} & 0 & 0 \\ \beta_{10} & 0 & 0 & 0 & \beta_9 & -a_{66} & 0 \\ 0 & 0 & 0 & 0 & 0 & \beta_{11} & -\mu \end{bmatrix}$$

where, $a_{11} = \beta_1 I_S^* + \beta_2 F^* + \beta_{10} + \mu$, $a_{22} = -\beta_1 E^* + \beta_4 + \beta_5 + \mu$, $a_{33} = -\beta_2 E^* + \beta_3 + \beta_6 + \beta_7 + \mu$, $a_{44} = \beta_8 + \mu$, $a_{55} = \beta_9 + \mu$, $a_{66} = \beta_{11} + \mu_C + \mu$.

The characteristic polynomial for Jacobian J^* is

$$\lambda^7 + b_6\lambda^6 + b_5\lambda^5 + b_4\lambda^4 + b_3\lambda^3 + b_2\lambda^2 + b_1\lambda + b_0$$

where,

$$\begin{aligned}
b_0 &= a_{44}a_{55}a_{66}\mu(E^*(F^*\beta_2(a_{22}\beta_2 + \beta_1\beta_3)+) + a_{11}a_{22}a_{33}) \\
b_1 &= E^*F^*\beta_2^2 a_{66}\mu(a_{22}a_{44} + (a_{22} + a_{44})a_{55}) \\
&\quad + E^*F^*\beta_1\beta_2\beta_3\mu(a_{44}a_{55}a_{66} + a_{44}a_{55} + a_{44}a_{66} + a_{55}a_{66}) \\
&\quad + E^*I_S^*\beta_1^2\mu(a_{33}(a_{44} + a_{55}) + a_{44}a_{55}) + a_{11}a_{22}((a_{33}a_{44}a_{55}(a_{66} + \mu)) \\
&\quad + a_{66}\mu(a_{33}(a_{44} + a_{55}) + a_{44}a_{55})) + a_{33}a_{44}a_{55}a_{66}\mu(a_{11} + a_{22}) \\
&\quad + (a_{66} + \mu)E^*a_{44}a_{55}(F^*\beta_2^2 a_{22} + I_S\beta_1^2 a_{33}) \\
b_2 &= E^*F^*\beta_2^2((a_{66} + \mu)(a_{22}(a_{44} + a_{55}) + a_{44}a_{55}) + a_{66}\mu(a_{22} + a_{44} + a_{55}) + a_{22}a_{44}a_{55}) \\
&\quad + E^*F^*\beta_1\beta_2\beta_3((a_{66} + \mu)(a_{44} + a_{55}) + a_{66}\mu + a_{44}a_{55}) \\
&\quad + E^*I_S^*\beta_1^2((a_{66} + \mu)(a_{33}(a_{44} + a_{55}) + a_{44}a_{55}) + a_{66}\mu(a_{33} + a_{44} + a_{55}) + a_{33}a_{44}a_{55}) \\
&\quad + (a_{66} + \mu)(a_{11}a_{22}(a_{44} + a_{33}(a_{44} + a_{55})) + a_{33}a_{44}a_{55}(a_{11} + a_{22})) \\
&\quad + (a_{44} + a_{55})(a_{11}a_{66}\mu(a_{22} + a_{33}) + a_{22}a_{33}a_{66}\mu) + a_{44}a_{55}a_{66}\mu(a_{22} + a_{33}) \\
&\quad + a_{11}a_{66}\mu(a_{44}a_{55} + a_{22}a_{33}) + a_{11}a_{22}a_{33}a_{44}a_{55}) \\
b_3 &= E^*F^*\beta_2^2((a_{66} + \mu)(a_{22} + a_{44} + a_{55}) + a_{66}\mu + a_{22}(a_{44} + a_{55})) \\
&\quad + E^*F^*\beta_1\beta_2\beta_3(a_{44} + a_{55} + a_{66} + \mu) + E^*I_S^*\beta_1^2((a_{66} + \mu)(a_{33} + a_{44} + a_{55}) \\
&\quad + a_{33}(a_{44} + a_{55}) + a_{66}\mu) + (a_{66} + \mu)((a_{44} + a_{55})(a_{11}(a_{22} + a_{33}) + a_{22}a_{33}) \\
&\quad + a_{44}a_{55}(a_{11} + a_{22} + a_{33}) + a_{11}a_{22}a_{33}) + (a_{44} + a_{55})(a_{66}\mu(a_{11} + a_{22} + a_{33}) + a_{11}a_{22}a_{33}) \\
&\quad + (a_{44}a_{55} + a_{66}\mu)(a_{11}(a_{22} + a_{33}) + a_{22}a_{33}) + a_{44}a_{55}a_{66}\mu \\
b_4 &= E^*(F^*\beta_2^2 a_{22} + I_S^*\beta_1^2 a_{33}) + E^*F^*\beta_1\beta_2\beta_3 + a_{11}a_{22}a_{33}) + a_{66}\mu(a_{22} + a_{33} + a_{44} + a_{55}) \\
&\quad + a_{55}(a_{66} + \mu)(a_{11} + a_{22} + a_{33} + a_{44}) + a_{44}(a_{55} + a_{66} + \mu)(a_{11} + a_{22} + a_{33}) \\
&\quad + (a_{44} + a_{55} + a_{66} + \mu)(E^*(F^*\beta_2^2 + I_S^*\beta_1^2) + a_{11}(a_{22} + a_{33}) + a_{22}a_{33}) \\
b_5 &= E^*(F^*\beta_2^2 + I_S^*\beta_1^2) + (a_{11} + \mu)(a_{22} + a_{33} + a_{44} + a_{55} + a_{66}) \\
&\quad + a_{55}a_{66} + (a_{55} + a_{66})(a_{22} + a_{33} + a_{44}) + a_{44}(a_{22} + a_{33}) + a_{22}a_{33} \\
b_6 &= a_{11} + a_{22} + a_{33} + a_{44} + a_{55} + a_{66} + \mu
\end{aligned}$$

Please cite this article in press as: N.H. Shah et al., Modelling COVID-19 transmission in the United States through interstate and foreign travels and evaluating impact of governmental public health interventions, J. Math. Anal. Appl. (2021), https://doi.org/10.1016/j.jmaa.2020.124896

Here, all the eigen values are negative if $a_{11} > 0, a_{22} > 0, a_{33} > 0, a_{44} > 0, a_{55} > 0, a_{66} > 0$ i.e. $E^* < \max\left\{\frac{\beta_4+\beta_5+\mu}{\beta_1}, \frac{\beta_3+\beta_6+\beta_7+\mu}{\beta_2}\right\}$. Then by Routh-Hurwitz criterion we say the endemic equilibrium point is locally asymptotically stable.

5. Optimal control

The novel corona virus is spread through human contact with infected individuals. Therefore, one can put control on respective situation to prevent its spreading.

Control description:

u_1: To prevent exposed foreign individuals in the interstate

u_2: Exposed interstate individuals should be quarantined

u_3: Exposed foreign individuals should be quarantined

u_4: Infected individuals should be quarantined

The objective function is,

$$J(c_i, \Lambda) = \int_0^T (A_1E^2 + A_2{I_S}^2 + A_3F^2 + A_4Q^2 + A_5I^2 + A_6H^2 + A_7R^2 + w_1{u_1}^2 + w_2{u_2}^2 + w_3{u_3}^2 + w_4{u_4}^2)dt$$

where, Λ denotes set of all compartmental variables, $A_1, A_2, A_3, A_4, A_5, A_6, A_7$ denote non-negative weight constants for compartments E, I_S, F, Q, I, H, R respectively. w_1, w_2, w_3 and w_4 are the weight constants for each control u_i where $i = 1, 2, 3, 4$ respectively.

Now, calculate every values of control variables from $t = 0$ to $t = T$ such that, $J(u_i(t)) = min\{J(u_i{}^*, \Lambda)/(u_i) \in \phi\}$, $i = 1, 2, 3, 4$ where, ϕ is a smooth function on the interval $[0, 1]$.

Related Langrangian function is given by,

$$\begin{aligned} L(\Lambda, A_i) = {} & A_1E^2 + A_2{I_S}^2 + A_3F^2 + A_4Q^2 + A_5I^2 + A_6H^2 + A_7R^2 + w_1{u_1}^2 + w_2{u_2}^2 + w_3{u_3}^2 + w_4{u_4}^2 \\ & + \lambda_1(B - \beta_1I_SE - \beta_2EF - \beta_{10}E - \mu E) + \lambda_2(\beta_1I_SE - \beta_4I_S - \beta_5I_S + \beta_3F - \mu S - (u_1 + u_2)I_S) \\ & + \lambda_3(\beta_2EF - (\beta_3 + \beta_6 + \beta_7 + \mu)F - \mu F + u_1I_S - u_3F) + \lambda_4(\beta_4I_S + \beta_6F - \beta_8Q - \mu Q + u_2I_S \\ & + u_4I + u_3F) + \lambda_5(\beta_7F + \beta_5I_s + \beta_8Q - \beta_9I - \mu I - u_4I) + \lambda_6(\beta_9I + \beta_{10}E - \beta_{11}H \\ & - (\mu + \mu_C)H) + \lambda_7(\beta_{11}H - \mu R) \end{aligned}$$

The adjoint equation variables, $\lambda_i = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7)$ for the system is calculated by taking partial derivatives of the Langrangian function with respect to each compartment variable.

$$\dot{\lambda_1} = -\frac{\partial L}{\partial E} = -2A_1E + (\lambda_1 - \lambda_2)\beta_1I_S + (\lambda_1 - \lambda_3)\beta_2F + (\lambda_1 - \lambda_6)\beta_{10} + \lambda_1\mu,$$

$$\dot{\lambda_2} = -\frac{\partial L}{\partial I_S} = -2A_2I_S + (\lambda_1 - \lambda_2)\beta_1E + (\lambda_2 - \lambda_4)(\beta_4 + u_2) + (\lambda_2 - \lambda_3)u_1 + (\lambda_1 - \lambda_2)\beta_5 + \lambda_2\mu,$$

$$\dot{\lambda_3} = -\frac{\partial L}{\partial F} = -2A_3F + (\lambda_1 - \lambda_3)\beta_2E + (\lambda_3 - \lambda_4)(\beta_6 + u_3) + (\lambda_3 - \lambda_2)\beta_3 + (\lambda_3 - \lambda_5)\beta_7 + \lambda_3\mu,$$

$$\dot{\lambda_4} = -\frac{\partial L}{\partial Q} = -2A_4Q + (\lambda_4 - \lambda_5)\beta_8 + \lambda_4\mu,$$

$$\dot{\lambda_5} = -\frac{\partial L}{\partial I} = -2A_5I + (\lambda_5 - \lambda_4)u_4 + (\lambda_5 - \lambda_6)\beta_9 + \lambda_5\mu,$$

$$\dot{\lambda_6} = -\frac{\partial L}{\partial H} = -2A_6H + (\lambda_6 - \lambda_7)\beta_{11} + (\mu + \mu_C)\lambda_6,$$

Please cite this article in press as: N.H. Shah et al., Modelling COVID-19 transmission in the United States through interstate and foreign travels and evaluating impact of governmental public health interventions, J. Math. Anal. Appl. (2021), https://doi.org/10.1016/j.jmaa.2020.124896

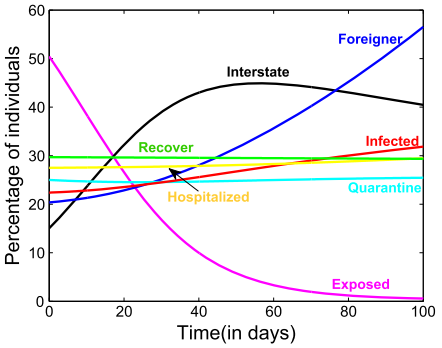


Fig. 2. Trajectories of each compartment showing flow of individuals in respective compartment.

$$\overset{\bullet}{\lambda_7} = -\frac{\partial L}{\partial R} = -2A_7R + \lambda_7\mu.$$

This calculation leads with resulting conditions as (Pontryagin, 1986) [8],

$$u_1{}^* = max\left(a_1, min\left(b_1, \frac{I_S(\lambda_2 - \lambda_3)}{2w_1}\right)\right),$$
$$u_2{}^* = max\left(a_2, min\left(b_2, \frac{I_S(\lambda_2 - \lambda_4)}{2w_2}\right)\right),$$
$$u_3{}^* = max\left(a_3, min\left(b_3, \frac{F(\lambda_3 - \lambda_4)}{2w_3}\right)\right),$$

and

$$u_4{}^* = max\left(a_4, min\left(b_4, \frac{I(\lambda_5 - \lambda_4)}{2w_4}\right)\right).$$

Based on analytical results, numerical simulation is given in next section.

6. Numerical simulation

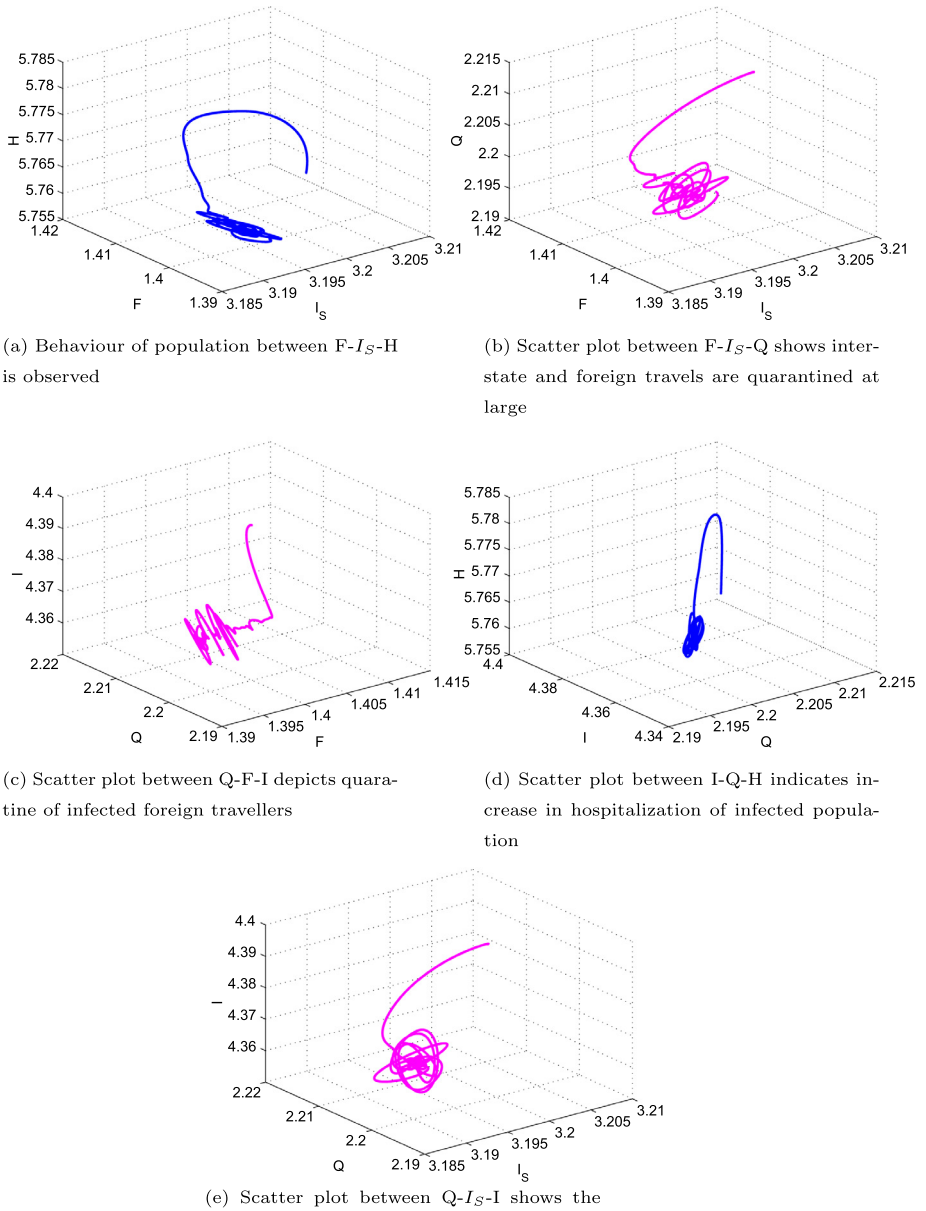
In this section we discuss the simulation performed for the system (1)

From Fig. 2, we observe 30% of interstate population is exposed to COVID-19 in 17.2 days. Whereas 24.55% of foreigner's population is exposed to COVID-19 in 21.8 days. 21% of foreigners come in contact with Interstate individuals in 7.1 days which increases the infectives of interstate to 22.78% in 9.2 days. Also 27.59% of interstate population gets hospitalized in 14.5 days.

Scatter plotting is shown in Fig. 3. Combined effect of group of three compartments is revealed in each plot. Fig. 3(a) depicts that; more infected interstate and foreign individuals will be hospitalised at higher rate of level. Fig. 3(b) shows that, individuals who travelled more will be quarantined. From Fig. 3(c), one can say that infected foreigner would be quarantined at higher rate. Infectedness in quarantined individuals increases which leads to the hospitalization of individuals as observed in Fig. 3(d). Fig. 3(e) describes that how interstate infected individuals are quarantined.

Fig. 4 shows the periodic nature of the interstate class exposed to the virus COVID-19. It indicates that interstate population is exposed again and again to the disease. It happens if the lockdown, social distancing is not followed as per government system. Which shows the importance of the government action taken against COVID-19 to protect the population. Fig. 5 shows the stability of the respective compartments at endemic equilibrium point. Since the government has decided to quarantine foreigners as soon as they arrive in their countries this makes the system stable as they are not exposed much to the COVID-19.

Please cite this article in press as: N.H. Shah et al., Modelling COVID-19 transmission in the United States through interstate and foreign travels and evaluating impact of governmental public health interventions, J. Math. Anal. Appl. (2021), https://doi.org/10.1016/j.jmaa.2020.124896



(a) Behaviour of population between F-I_S-H is observed

(b) Scatter plot between F-I_S-Q shows interstate and foreign travels are quarantined at large

(c) Scatter plot between Q-F-I depicts quarantine of infected foreign travellers

(d) Scatter plot between I-Q-H indicates increase in hospitalization of infected population

(e) Scatter plot between Q-I_S-I shows the behaviour of quarantined infected interstate population

Fig. 3. Scatter plotting between different compartment combination is observed showing the behaviour of respective combinations.

Fig. 6a, 6b shows the trajectory at the endemic equilibrium point for the system (1). Here we observe the importance of quarantine as the system is stable when interstate and foreigners are quarantined.

From Fig. 7, we observe foreigner are moving towards interstate population.

Fig. 8 and Fig. 9 illustrates the flow of interstate population and foreigners with the COVDI-19 infection. It shows that the interstate population gets the infection at a slower rate as compared to foreigners.

Fig. 10 shows the flow of interstate and foreigners towards hospitalization. Foreigners gets hospitalized at faster rate than interstate population.

Fig. 11 (a)-(g) show the oscillating behaviour of each compartment. As the epidemic nature of disease increases, this can oscillate the whole situation. In some intervals of data, exposed individuals increase (Fig. 11(a)) who are either interstate (Fig. 11(b)) or foreigner (Fig. 11(c)). If quarantined individuals do

Please cite this article in press as: N.H. Shah et al., Modelling COVID-19 transmission in the United States through interstate and foreign travels and evaluating impact of governmental public health interventions, J. Math. Anal. Appl. (2021), https://doi.org/10.1016/j.jmaa.2020.124896

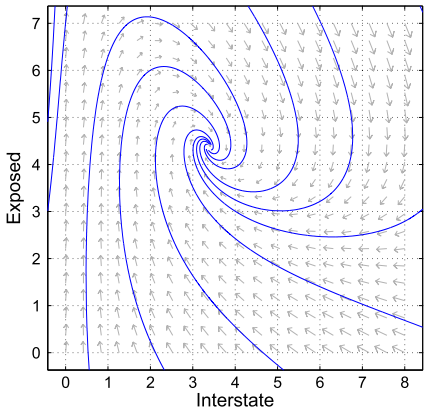


Fig. 4. We plot phase diagram of interstate class when exposed to COVID-19 observing again and again exposure of interstate population to COVID-19.

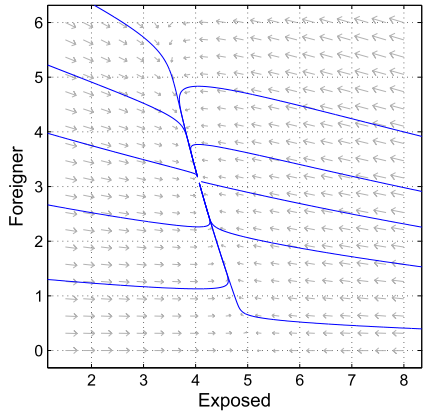
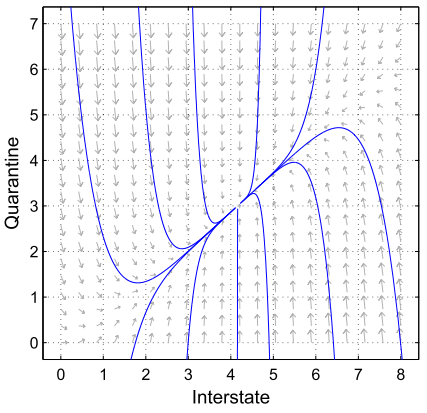
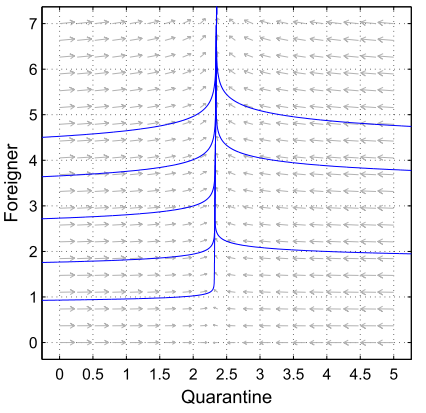


Fig. 5. This indicates phase plot of foreigner class when exposed to COVID-19. Here we observe convergent behaviour of respective classes making it stable.



(a) The convergence of interstate and quarantine population is observed making respective compartment stable



(b) The foreign travellers are quaratined and both the compartments converges

Fig. 6. Behaviour of interstate class with quarantine class (a) and Phase plot of quarantine with foreigners (b).

not follow quarantines rules which have been observed in Fig. 11(d). This leads to a greater number of infected individuals (Fig. 11(e)) hence they should be hospitalised (Fig. 11(f)) which effects on recovery rate (Fig. 11(f)).

Please cite this article in press as: N.H. Shah et al., Modelling COVID-19 transmission in the United States through interstate and foreign travels and evaluating impact of governmental public health interventions, J. Math. Anal. Appl. (2021), https://doi.org/10.1016/j.jmaa.2020.124896

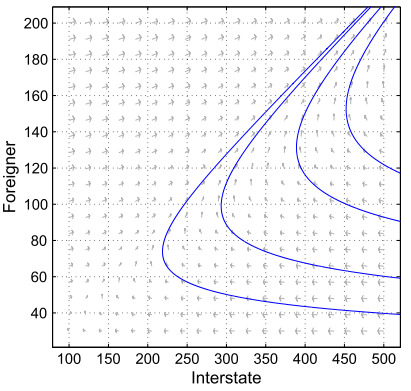


Fig. 7. Transition diagram of interstate and foreigner population. Here the movement of individuals between respective classes is observed.

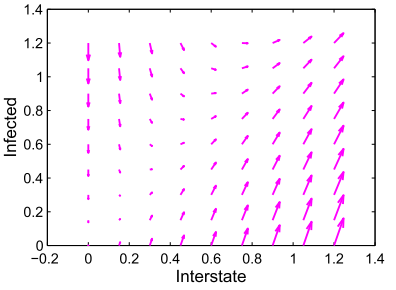


Fig. 8. Shows directional plot of Interstate population infected with COVID-19 indicating the infectiousness of interstate population.

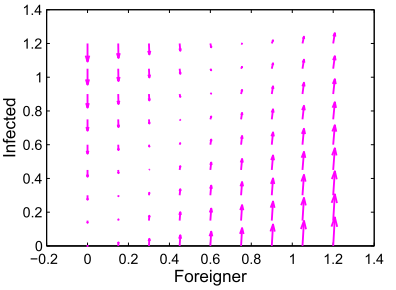
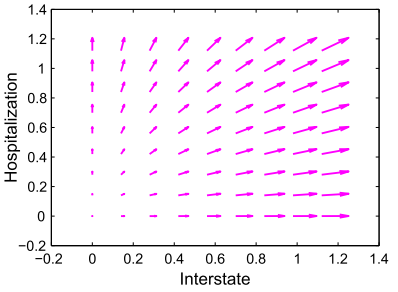
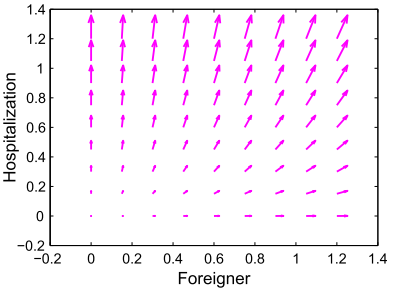


Fig. 9. Depicts behaviour of Foreigner population infected with COVID-19. Here we observe that foreign travellers are getting infection at large.



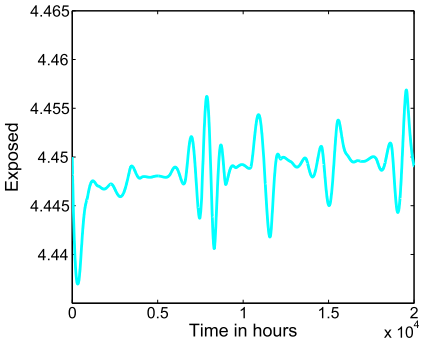
(a) Plot of directional diagram between the interstate and hospilzation classes indicates the direction of flow of respective population class



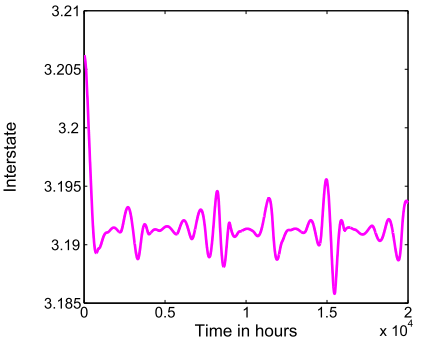
(b) The movement indicates that the foreigners are getting hospitalised

Fig. 10. Directional plot of hospitalization of interstate population (a) and foreigner (b).

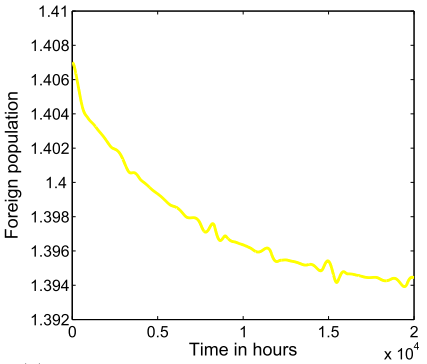
Please cite this article in press as: N.H. Shah et al., Modelling COVID-19 transmission in the United States through interstate and foreign travels and evaluating impact of governmental public health interventions, J. Math. Anal. Appl. (2021), https://doi.org/10.1016/j.jmaa.2020.124896



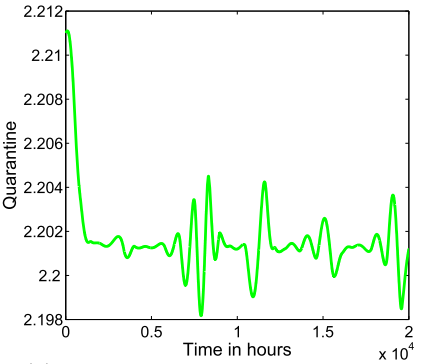
(a) The exposed class continous to oscillate with time



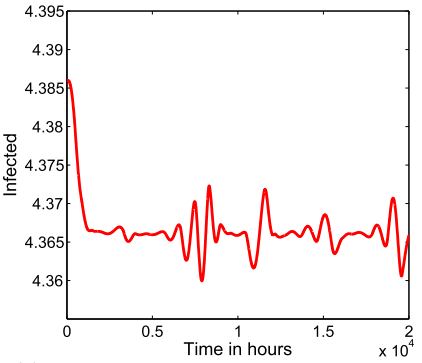
(b) Oscillation effect on interstate class



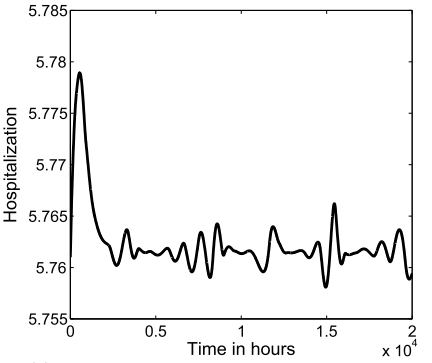
(c) Oscillation effect on foreigner class decreases with time



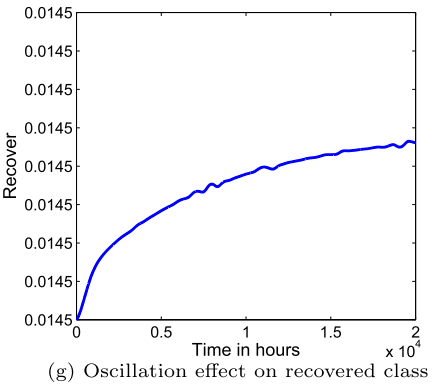
(d) We oberve oscillations in quarantine class



(e) This shows the existence of infected population remains in society



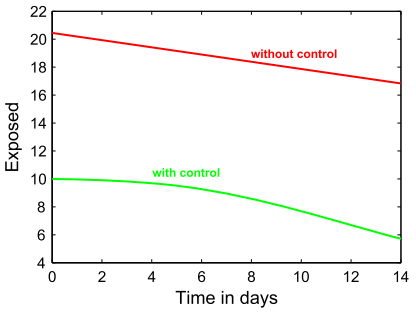
(f) Oscillation are oberved indicating continous existence of repective class



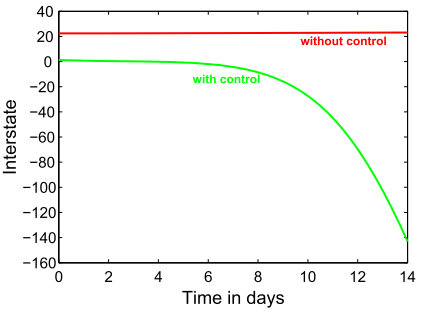
(g) Oscillation effect on recovered class

Fig. 11. Here we observe continuous fluctuation in all the compartments which very well depicts the scenario of COVID-19 among interstate and foreign travellers.

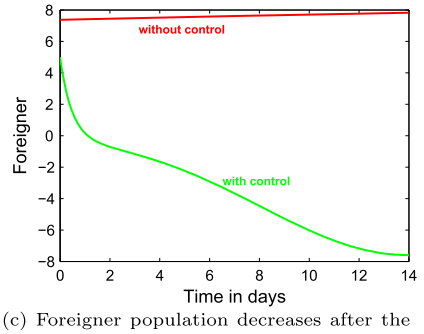
Please cite this article in press as: N.H. Shah et al., Modelling COVID-19 transmission in the United States through interstate and foreign travels and evaluating impact of governmental public health interventions, J. Math. Anal. Appl. (2021), https://doi.org/10.1016/j.jmaa.2020.124896



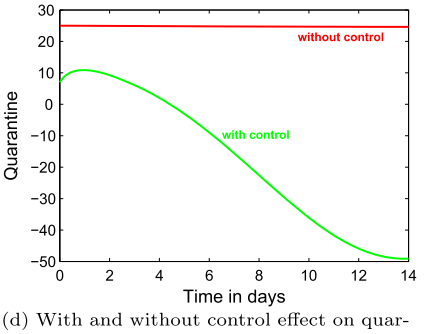
(a) Exposed population decreases after all the controls are applied



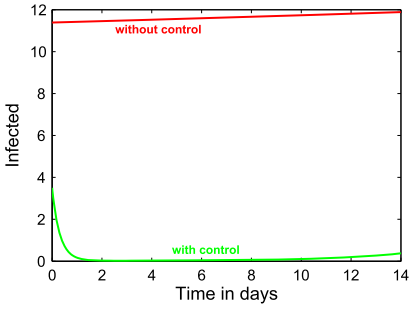
(b) Control effect on interstate class decreases the respective population



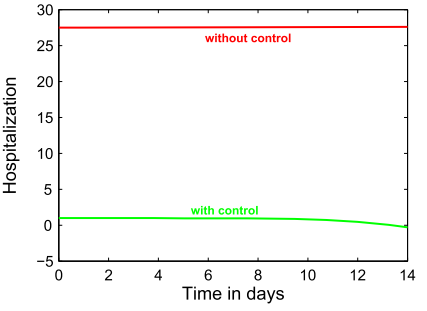
(c) Foreigner population decreases after the controls are applied



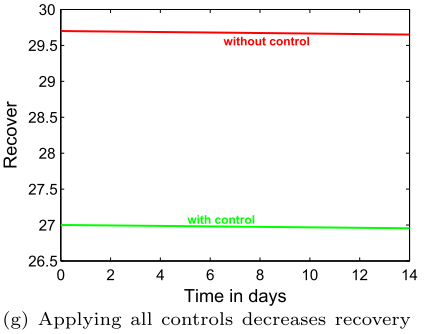
(d) With and without control effect on quarantine class is oberved



(e) Without and with controls effect on infected class can be observed



(f) Control effect on hospitalization class



(g) Applying all controls decreases recovery as infectiouness also decreases

Fig. 12. Effect of controls applied to the system (1) is observed on each compartment. Here it can be seen that after control is applied, population of each compartment decreases.

The above oscillating nature of the model is controlled by the Fig. 12(a)-(g). All the four controls are effective to our system (1). In the presence of all the controls we observe decrease in the number of exposed individuals. Quarantining interstate and foreign individuals also reduce the infection when controls are applied.

Please cite this article in press as: N.H. Shah et al., Modelling COVID-19 transmission in the United States through interstate and foreign travels and evaluating impact of governmental public health interventions, J. Math. Anal. Appl. (2021), https://doi.org/10.1016/j.jmaa.2020.124896

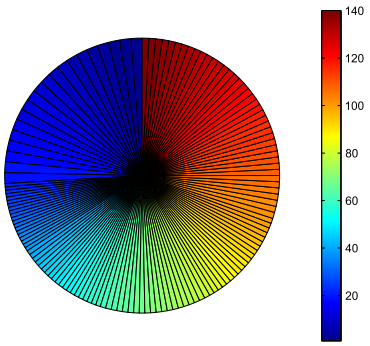
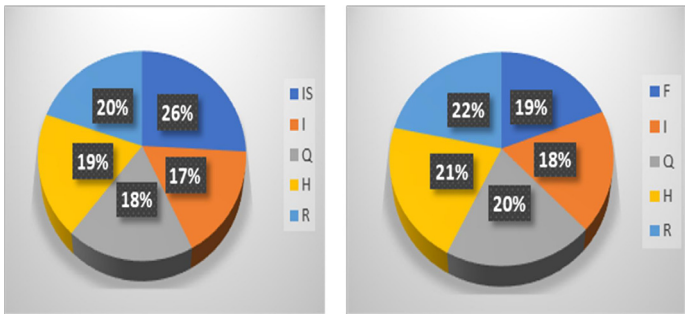


Fig. 13. Represents chaotic diagram showing mortality rate of 2019-nCoV.



(a) (b)

Fig. 14. Percentage wise distribution of interstate (a) and foreigner population in COVID-19 scenario (b). In Fig. 14 a, we observe out of 26% of interstate population 17% is infected and from Fig. 14 b, among 19% of foreigners we have 18% infected population.

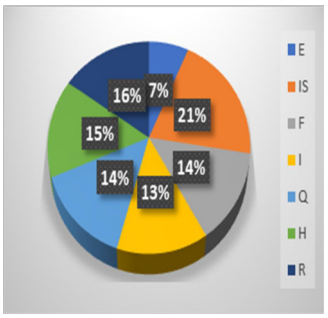


Fig. 15. Figure indicates percentage wise distribution of all the compartments of the system (1). Here we have 21% of interstate population, 14% of foreign population out of which 13% gets the infection with 15% getting hospitalised.

The Fig. 13, shows intensity of mortality due to COVID-19 among interstate and foreign travellers.

The Fig. 14 clearly shows the infected population of foreigner is more than that of interstate population which shows the importance of complete ban on air arrivals.

From the Fig. 15 it can be observed that 7% of population is exposed to COVID-19. Interstate population share the largest percentage. 14% of the population is quarantined including foreigners and interstate population. Similarly, the infection is 13%. The hospitalization is done at 15%. Of the total population recovery is 16%.

7. Discussion

Our model indicates that foreigners exhibit a larger infected population, hospitalization rate, and infection rate when compared to the interstate population. Moreover, as foreign individuals contact interstate

Please cite this article in press as: N.H. Shah et al., Modelling COVID-19 transmission in the United States through interstate and foreign travels and evaluating impact of governmental public health interventions, J. Math. Anal. Appl. (2021), https://doi.org/10.1016/j.jmaa.2020.124896

individuals, the rate of infection within the interstate population increases significantly. To the best of our knowledge, our study is the first to create a disjoint mutually exclusive compartmental model. Our model suggests that both foreign and interstate travel lead to increased risk of infection within the United States population. Consequently, we validate the effectiveness of quarantine as a public health intervention model by the US government and encourage implementation of efforts to mitigate interstate travel.

There are multiple reasons that foreigners have increased risk of obtaining COVID-19 when compared to interstate population. First, as people travel, they risk exposing themselves to a greater number of other individuals. The WHO indicates that transmission of COVID-19 occurs primarily through droplet transmission (2019). Most methods of international travel, including airways, railways, and waterways, crowd individuals in compact and enclosed spaces. Being in close contact with individuals with respiratory symptoms in an enclosed environment increases the risk of being exposed to infected mucosae [11]. Second, the guidelines and strategies for addressing the epidemic differ among countries. For example, while India has enforced total lockdown, the US government has not mandated enforced lockdown [3]. Consequently, when individuals from countries with different regulations arrive, they may be infected and increase the incidence of COVID-19. Finally, the vaccination standards differ among countries. In particular, BCG vaccine, believed to confer protective effects against COVID-19 is recommended in some countries, but not the US [7]. As a result, future research and modelling is necessary to determine the protective effects of the BCG vaccine, and its potential to reduce the incidence of COVID-19 within the United States.

Given that 2019-nCoV is no longer contained within Wuhan, we recommend the United States government close their borders to both foreign and interstate travel. We recommend significant public health interventions at both international and interstate levels otherwise large cities with close inter-transport systems could become outbreak epicentres. Finally, we recommend preparedness plans and mitigation interventions be readied for quick deployment on both a state and federal level. Based on our model, compliance with these recommendations will effectively reduce the transmission of COVID-19 as a result of foreign and interstate travel.

Acknowledgment

The first three authors thank DST-FIST file # MSI-097 for technical support to the department. Third author (ENJ) is funded by UGC granted National Fellowship for Other Backward Classes (NFO-2018-19-OBC-GUJ-71790).

References

[1] C. Castillo-Chavez, C.W. Castillo-Garsow, A.-A. Yakubu, Mathematical models of isolation and quarantine, J. Am. Med. Assoc. 290 (21) (2003) 2876–2877, https://doi.org/10.1001/jama.290.21.2876.
[2] Centers for Disease Control and Prevention, https://www.cdc.gov/coronavirus/2019-ncov/index.html, 2020.
[3] K. Chatterjee, K. Chatterjee, A. Kumar, S. Shankar, Healthcare impact of COVID-19 epidemic in India: a stochastic mathematical model, Med. J. Armed Forces India 76 (2) (2020) 147–155, https://doi.org/10.1016/j.mjafi.2020.03.022.
[4] J. Cohen, D. Normile, New SARS-like virus in China triggers alarm, https://doi.org/10.1126/science.367.6475.234, 2020.
[5] O. Diekmann, J. Heesterbeek, M.G. Roberts, The construction of next-generation matrices for compartmental epidemic models, J. R. Soc. Interface 7 (47) (2010) 873–885, https://doi.org/10.1098/rsif.2009.0386.
[6] M.L. Holshue, C. DeBolt, S. Lindquist, K.H. Lofy, J. Wiesman, H. Bruce, C. Spitters, K. Ericson, S. Wilkerson, A. Tural, et al., First case of 2019 novel coronavirus in the united states, N. Engl. J. Med. 10 (382) (2020) 929–936, https://doi.org/10.1056/NEJMoa2001191.
[7] A. Miller, M.J. Reandelar, K. Fasciglione, V. Roumenova, Y. Li, G.H. Otazu, Correlation between universal BCG vaccination policy and reduced morbidity and mortality for COVID-19: an epidemiological study, MedRxiv, https://doi.org/10.1101/2020.03.24.20042937.
[8] L.S. Pontryagin, Mathematical Theory of Optimal Processes, Routledge, 2018.
[9] E.J. Routh, A Treatise on the Stability of a Given State of Motion: Particularly Steady Motion, Macmillan and Company, 1877.
[10] B. Tang, X. Wang, Q. Li, N.L. Bragazzi, S. Tang, Y. Xiao, J. Wu, Estimation of the transmission risk of the 2019-nCOV and its implication for public health interventions, J. Clin. Med. 9 (2) (2020) 462, https://doi.org/10.3390/jcm9020462.

Please cite this article in press as: N.H. Shah et al., Modelling COVID-19 transmission in the United States through interstate and foreign travels and evaluating impact of governmental public health interventions, J. Math. Anal. Appl. (2021), https://doi.org/10.1016/j.jmaa.2020.124896

[11] A.J. Tatem, D.J. Rogers, S.I. Hay, Global transport networks and infectious disease spread, Adv. Parasitol. 62 (2006) 293–343, https://doi.org/10.1016/S0065-308X(05)62009-X.
[12] World health organization, https://www.who.int/emergencies/diseases/novel-coronavirus-2019, 2019.
[13] J.T. Wu, K. Leung, G.M. Leung, Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in Wuhan, China: a modelling study, Lancet 395 (10225) (2020) 689–697, https://doi.org/10.1016/S0140-6736(20)30260-9.

Please cite this article in press as: N.H. Shah et al., Modelling COVID-19 transmission in the United States through interstate and foreign travels and evaluating impact of governmental public health interventions, J. Math. Anal. Appl. (2021), https://doi.org/10.1016/j.jmaa.2020.124896

mathematics

MDPI

Article

Noncommutative Functional Calculus and Its Applications on Invariant Subspace and Chaos

Lvlin Luo [1,2,3]

[1] School of Mathematical Sciences, Fudan University, Shanghai 200433, China; luoll12@mails.jlu.edu.cn or luolvlin@fudan.edu.cn
[2] School of Mathematics, Jilin University, Changchun 130012, China
[3] School of Mathematics and Statistics, Xidian University, Xi'an 710071, China

Received: 5 August 2020; Accepted: 7 September 2020; Published: 9 September 2020

Abstract: Let $T : \mathbb{H} \to \mathbb{H}$ be a bounded linear operator on a separable Hilbert space $\mathbb{H}$. In this paper, we construct an isomorphism $F_{xx^*} : \mathcal{L}^2(\sigma(|T-a|), \mu_{|T-a|,\xi}) \to \mathcal{L}^2(\sigma(|(T-a)^*|), \mu_{|(T-a)^*|,F^{\mathbb{H}}_{xx^*}\xi})$ such that $(F_{xx^*})^2 = identity$ and $F^{\mathbb{H}}_{xx^*}$ is a unitary operator on $\mathbb{H}$ associated with F_{xx^*}. With this construction, we obtain a noncommutative functional calculus for the operator T and $F_{xx^*} = identity$ is the special case for normal operators, such that $S = R_{|(S-a)|,\xi}(M_{z\phi(z)} + a)R^{-1}_{|S-a|,\xi}$ is the noncommutative functional calculus of a normal operator S, where $a \in \rho(T)$, $R_{|T-a|,\xi} : \mathcal{L}^2(\sigma(|T-a|), \mu_{|T-a|,\xi}) \to \mathbb{H}$ is an isomorphism and $M_{z\phi(z)} + a$ is a multiplication operator on $\mathcal{L}^2(\sigma(|S-a|), \mu_{|S-a|,\xi})$. Moreover, by F_{xx*} we give a sufficient condition to the invariant subspace problem and we present the Lebesgue class $\mathcal{B}_{Leb}(\mathbb{H}) \subset \mathcal{B}(\mathbb{H})$ such that T is Li-Yorke chaotic if and only if T^{*-1} is for a Lebesgue operator T.

Keywords: chaos; invariant subspace; Lebesgue operator; noncommutative functional calculus

MSC: Primary 47A15, 47A16, 47A60, 47A65; Secondary 37D45

1. Introduction

1.1. Invariant Subspace

The invariant subspace problem has been stated by Beurling and von Neumann [1]. It can be formulated as follows.

Problem 1. *Does every bounded linear operator on a given linear space have a non-trivial invariant subspace?*

In 1966, Bernstein et al. [2] showed that if T is a bounded linear operator on a complex Hilbert space $\mathbb{H}$ and p is a nonzero polynomial such that $p(T)$ is compact, then T has non-trivial invariant subspace. Especially, when $p(t) = t$, which is, T itself is compact, the result was proved independently by von Neumann and N. Aronszajn, and in [3], this result was extended to compact operators on a Banach space.

Let T be a bounded linear operator on a Banach space. In 1973, Lomonosov [4] proved that if T is not a scalar multiple of the identity and commutes with a nonzero compact operator, then T has a non-trivial hyperinvariant subspace, which is, any bounded linear operator commuting with T has a non-trivial invariant subspace (other results see [5–7]).

In 1976, Enflo [8] was the first to construct an operator on a Banach space having no non-trivial invariant subspace and Nordgren et al. [9] proved that every operator has an invariant subspace if and only if every pair of idempotents has a common invariant subspace.

In 1983, Atzmon [10] constructed a nuclear Fréchet space $\mathbb{F}$ and a bounded linear operator, which has no non-trivial invariant subspace. Especially, in 1984, C. J. Read made an example, such that there is a bounded linear operator without non-trivial invariant subspace on ℓ_1 [11].

In 2011, Argyros et al. [12] constructed the first example of a Banach space for which every bounded linear operator on the space has the form $\lambda + K$ where λ is a real scalar and K is a compact operator, such that every bounded linear operator on the space has a non-trivial invariant subspace.

In 2013, Marcoux et al. [13] showed that, if a closed algebra of operators on a Hilbert space has a non-trivial almost-invariant subspace, then it has a non-trivial invariant subspace (more results see [14–18]).

In 2019, Tcaciuc [19] proved that, for any bounded operator T acting on an infinite-dimensional Banach space, there exists an operator F of rank at most one such that $T + F$ has an invariant subspace of infinite dimension and codimension.

For finite-dimensional vector spaces or nonseparable Hilbert spaces, the result is trivial. However, for infinite-dimensional separable Hilbert spaces, the problem is, after a long period of time, not yet completely solved.

1.2. Linear Dynamics

With the development of operator theory and dynamics progress, there are many papers about C^*-algebras and dynamics. Additionally, "the fundamental theorem of C^*-algebras [20]" is Gelfand-Naimark theorem [21]. Subsequently, in [22], Fujimoto said that this theorem eventually opened the gate to the subject of C^*-algebras. Hence, there are various attempts to generalize this theorem [23–27].

For the research on Problem 1 and with the development of chaos, Operator Dynamics or Linear Dynamics has aroused extensive attention as an important branch of functional analysis, which was probably born in 1982 with the Toronto Ph. D. thesis of C. Kitai [28]. More details of this subject can be found in [29–33].

If X is a metric space and T is a continuous self-map on X, then the pair (X, T) is called a topological dynamic systems, which is induced by the iteration

$$T^n = \underbrace{T \circ \cdots \circ T}_{n}, \qquad n \in \mathbb{N}, \text{ where } 0 \in \mathbb{N}.$$

Moreover, if T is a continuous invertible self-map on X, then (X, T) is called an invertible dynamic and if the metric space X and the continuous self-map T are both linear, then the topological dynamic systems (X, T) is called a linear dynamic.

For invertible dynamics, the relationship of Li-Yorke chaos between (X, f) and (X, f^{-1}) was raised by Stockman as an open question [34]. Additionally, in [35,36] and [37], the authors give counterexamples for this question in noncompact spaces and compact spaces, respectively. For an invertible bounded linear operator $T \in \mathcal{B}(\mathbb{H})$, the chaotic relationship between $(\mathbb{H}, T)$ and $(\mathbb{H}, T^{*-1})$ is also interesting.

Next, we give the following definition

Definition 1 (Li-Yorke chaos)**.** *Let $T \in \mathcal{B}(\mathbb{H})$. If there exists $x \in \mathbb{H}$, such that satisfies:*

$$(a)\ \overline{\lim_{n\to\infty}} \|T^n(x)\| > 0 \quad \textit{and}$$
$$(b)\ \underline{\lim_{n\to\infty}} \|T^n(x)\| = 0,$$

then the operator T is said to be Li-Yorke chaotic, and x is called a Li-Yorke chaotic point of T.

An example of an operator T that is Li-Yorke chaotic but T^{*-1} is not can be found in [38]. However, presently there is no general method to do this research. In fact, the C*-algebra $\mathcal{A}(T)$ generated by T cannot be used for that.

1.3. Motivation and Main Results

For an n-tuple $\mathcal{T}$ of not necessarily commuting operators, Colombo et al. [39] put to use the notion of slice monogenic functions [40] to define a new functional calculus, which is consistent with the Riesz–Dunford calculus in the case of a single operator and that allows the explicit construction of the eigenvalue equation for the n-tuple $\mathcal{T}$ based on a new notion of spectrum for $\mathcal{T}$ (more results, see [41–44]).

In 2010, for bounded operators defined on quaternionic Banach spaces, Colombo et al. [45] developed a noncommutative functional calculus that is based on the new notion of slice-regularity and that is based on the key tools of a new resolvent operator and a new eigenvalue problem, also, they extended this calculus to the unbounded case [46] (more results, see [47]).

In 2018, Monguzzi et al. [48] characterized the closed invariant subspaces for the $(^*-)$ multiplier operator of the quaternionic space of slice $\mathcal{L}^2$ functions, obtained the inner-outer factorization theorem for the quaternionic Hardy space on the unit ball and provided a characterization of quaternionic outer functions in terms of cyclicity.

In this paper, we give a noncommutative functional calculus for $T \in \mathcal{B}(\mathbb{H})$. Additionally, by this construction, we give some applications, such as its applications on the invariant subspace problem and chaos. The precise meaning of the multiplication operator $M_{z\psi(z)} = M_z M_{\psi(z)} = M_{\psi(z)} M_z = M_{\psi(z)z}$ will become clear in Theorem 3.

Let $\mathbb{H}$ be a separable Hilbert space over $\mathbb{C}$, $\mathcal{B}(\mathbb{H})$ be the set of all bounded linear operator on $\mathbb{H}$. For any given $T \in \mathcal{B}(\mathbb{H})$, we obtain a C*-algebra $\mathcal{A}(|T-a|)$ associated with the polar decomposition $T - a = U|T-a|$, where $a \in \rho(T)$ and ξ is a $\mathcal{A}(|T-a|)$-cyclic vector, such that $\mathbb{H} = \overline{\mathcal{A}(|T-a|)\xi}$. In this paper, we construct an isomorphism F_{xx^*}, such that the following diagram is valid.

$$\begin{array}{ccc} \mathcal{L}^2(\sigma(|T-a|), \mu_{|T-a|,\xi}) & \xrightarrow{\quad R_{|T-a|,\xi} \quad} & \mathbb{H} \\ F_{xx^*} \downarrow & & \downarrow F^{\mathbb{H}}_{xx^*} \\ \mathcal{L}^2(\sigma(|(T-a)^*|), \mu_{|(T-a)^*|,F^{\mathbb{H}}_{xx^*}\xi}) & \xrightarrow[\quad R_{|(T-a)^*|,F^{\mathbb{H}}_{xx^*}\xi} \quad]{} & \mathbb{H} \end{array}$$

where $F^{\mathbb{H}}_{xx^*}$ is the corresponding unitary operator associated with the isomorphism F_{xx^*} and $Fix(F^{\mathbb{H}}_{xx^*}) \neq \emptyset$, $(F^{\mathbb{H}}_{xx^*})^2 = identity$ and $(F_{xx^*})^2 = identity$. With this construction, we get a noncommutative functional calculus for the operator T such that

$$T - a = R_{|(T-a)^*|,F^{\mathbb{H}}_{xx^*}\xi} F_{xx^*} M_{z\psi(z)} R^{-1}_{|T-a|,\xi}\,.$$

Especially, $F_{xx^*} = identity$, which is the special case for normal operators, will become clear in Corollary 3, and, in this special case, we get that the noncommutative functional calculus of a normal operator S is just only $S = R_{|(S-a)|,\xi}(M_{z\phi(z)} + a)R^{-1}_{|S-a|,\xi}$, which is compatible with the classical normal operator functional calculus of [49]. Where $\psi(z) \in \mathcal{L}^\infty(\sigma(|T-a|), \mu_{|T-a|,\xi})$ and $\phi(z) \in \mathcal{L}^\infty(\sigma(|S-a|), \mu_{|S-a|,\xi})$.

Moreover, from F_{xx*}, we deduce a sufficient condition to Problem 1 on infinite-dimensional separable Hilbert spaces and present the Lebesgue class $\mathcal{B}_{Leb}(\mathbb{H}) \subset \mathcal{B}(\mathbb{H})$, such that, if T is a Lebesgue operator, then T is Li-Yorke chaotic if and only if T^{*-1} is.

In fact, we get that

$$\mathcal{B}_{Leb}(\mathbb{H}) \cap \mathcal{B}_{Nor}(\mathbb{H}) \neq \emptyset$$

and

$$\mathcal{B}_{Leb}(\mathbb{H}) \cap (\mathcal{B}(\mathbb{H}) \setminus \mathcal{B}_{Nor}(\mathbb{H})) \neq \emptyset,$$

where $\mathcal{B}_{Nor}(\mathbb{H})$ is the set of all normal operator on $\mathbb{H}$.

2. Decomposition and Isomorphic Representation

In this paper, $\bar{f}(\cdot)$ means the conjugate of the complex function $f(\cdot)$. Let X be a compact subset of $\mathbb{C}$, $\mathcal{C}(X)$ be the set of all continuous function on X, and $\mathcal{P}(x)$ be the set of all polynomial on X. For any given $T \in \mathcal{B}(\mathbb{H})$, let $\sigma(T)$ be its spectrum.

Following the polar decomposition theorem [50] (p. 15), we get that

$$T = U|T| \quad \text{and} \quad |T|^2 = T^*T.$$

Let $\mathcal{A}(|T|)$ be the complex C^*-algebra generated by $|T|$ and 1. Obviously, if T is invertible, then U is a unitary operator.

Lemma 1. *Let $X \subseteq \mathbb{C}$ be a compact subset not containing zero. If $\mathcal{P}(x)$ is dense in $\mathcal{C}(X)$, then $\mathcal{P}(\frac{1}{x})$ is also dense in $\mathcal{C}(X)$.*

Proof. By the properties of complex polynomials, we get that $\mathcal{P}(\frac{1}{x})$ is a subalgebra of $\mathcal{C}(X)$, which is closed under the standard algebraic operations. In addition, we have:
(1) $1 \in \mathcal{P}(\frac{1}{x})$;
(2) $\mathcal{P}(\frac{1}{x})$ separate the points of X;
(3) If $p(\frac{1}{x}) \in \mathcal{P}(\frac{1}{x})$, then $\bar{p}(\frac{1}{x}) \in \mathcal{P}(\frac{1}{x})$.
We get the conclusion from the Stone–Weierstrass theorem [49] (p. 145). □

For $X \subseteq \mathbb{R}_+$, there is $x \neq y \Longleftrightarrow x^2 \neq y^2$. With Lemma 1, we get the following result.

Lemma 2. *Let $X \subseteq \mathbb{R}_+$. If $\mathcal{P}(|x|)$ is dense in $\mathcal{C}(X)$, then $\mathcal{P}(|x|^2)$ is also dense in $\mathcal{C}(X)$.*

Using the GNS construction [49] (p. 250), for the C^*-algebra $\mathcal{A}(|T|)$, we have the following decomposition.

Lemma 3. *Let T be an invertible bounded linear operator on $\mathbb{H}$. Then there exists a sequence of nonzero $\mathcal{A}(|T|)$-invariant subspaces $\mathbb{H}_1, \mathbb{H}_2, \cdots, \mathbb{H}_i, \cdots$, such that:*
(1) $\mathbb{H} = \mathbb{H}_1 \oplus \mathbb{H}_2 \oplus \cdots \oplus \mathbb{H}_i \oplus \cdots$;
(2) For every $\mathbb{H}_i$, there is a $\mathcal{A}(|T|)$-cyclic vector ξ^i such that

$$\mathbb{H}_i = \overline{\mathcal{A}(|T|)\xi^i} = \overline{\mathcal{A}(|T|^{-1})\xi^i}$$

and

$$|T|\mathbb{H}_i = \mathbb{H}_i = |T|^{-1}\mathbb{H}_i \, .$$

Proof. The decomposition of (1) is obvious [51] (p. 54), Therefore,

$$|T|\mathbb{H}_i \subseteq \mathbb{H}_i \, ,$$

that is,

$$\mathbb{H}_i \subseteq |T|^{-1}\mathbb{H}_i \, .$$

From Lemma 1, we get that

$$\mathbb{H}_i = \overline{\mathcal{A}(|T|)\xi^i} = \overline{\mathcal{A}(|T|^{-1})\xi^i}$$

and

$$|T|^{-1}\mathbb{H}_i \subseteq \mathbb{H}_i \, .$$

Hence,

$$|T|\mathbb{H}_i = \mathbb{H}_i = |T|^{-1}\mathbb{H}_i\,.$$

□

Let $\xi \in \mathbb{H}$ be a $\mathcal{A}(|T|)$-cyclic vector, such that $\mathcal{A}(|T|)\xi$ is dense in $\mathbb{H}$. Because the spectrum is closed and $\sigma(|T|) \neq \emptyset$, on $\mathcal{C}(\sigma(|T|))$, we can define the nonzero linear functional

$$\rho_{|T|,\xi} : \rho_{|T|,\xi}(f) = \langle f(|T|)\xi, \xi\rangle\,, \quad \forall f \in \mathcal{C}(\sigma(|T|))\,.$$

It is easy to get that $\rho_{|T|,\xi}$ is a positive linear functional. By [51] (p. 54), and the Riesz–Markov theorem, on $\mathcal{C}(\sigma(|T|))$, we get that there is a uniquely finite positive Borel measure $\mu_{|T|,\xi}$, such that

$$\int\limits_{\sigma(|T|)} f(z)\mathrm{d}\mu_{|T|,\xi}(z) = \langle f(|T|)\xi, \xi\rangle\,, \quad \forall f \in \mathcal{C}(\sigma(|T|))\,.$$

Theorem 1. *Let T be an invertible bounded linear operator on $\mathbb{H}$, $\mathcal{A}(|T^n|)$ be the complex C^*-algebra generated by $|T^n|$ and 1 and let ξ_n be a $\mathcal{A}(|T^n|)$-cyclic vector, such that $\overline{\mathcal{A}(|T^n|)\xi_n} = \mathbb{H}$, where $n \in \mathbb{N}$. Subsequently:*

(1) there is a uniquely positive linear functional

$$\int\limits_{\sigma(|T^n|)} f(z)\mathrm{d}\mu_{|T^n|,\xi_n}(z) = \langle f(|T^n|)\xi_n, \xi_n\rangle\,, \quad \forall f \in \mathcal{L}^1(\sigma(|T^n|), \mu_{|T^n|,\xi_n})\,.$$

(2) there is a uniquely isomorphic representation $R_{|T^n|,\xi_n} : \mathcal{L}^2(\sigma(|T^n|), \mu_{|T^n|,\xi_n}) \to \mathbb{H}$ associated with the uniquely finite positive Borel measure $\mu_{|T^n|,\xi_n}$, which is complete.

Proof. (1) For $\mathcal{A}(|T^n|)$-cyclic vector ξ_n, we define the linear functional

$$\rho_{|T^n|,\xi_n}(f) = \langle f(|T^n|)\xi_n, \xi_n\rangle\,, \quad \forall f \in \mathcal{C}(\sigma(|T|)).$$

We get that, on $\mathcal{C}(\sigma(|T^n|))$, there is a uniquely finite positive Borel measure $\mu_{|T^n|,\xi_n}$, such that

$$\int\limits_{\sigma(|T^n|)} f(z)\mathrm{d}\mu_{|T^n|,\xi_n}(z) = \langle f(|T^n|)\xi_n, \xi_n\rangle\,, \quad \forall f \in \mathcal{C}(\sigma(|T^n|)).$$

Moreover, we can complete this Borel measure $\mu_{|T^n|,\xi_n}$ on $\sigma(|T^n|)$. For this completion, we keep the notation $\mu_{|T^n|,\xi_n}$. We know that this Borel measure is unique [52] .

For any $f \in \mathcal{L}^2(\sigma(|T^n|), \mu_{|T^n|,\xi_n})$, because of

$$\rho_{|T^n|,\xi_n}(|f|^2) = \rho_{|T^n|,\xi_n}(\bar{f}f) = \langle f(|T^n|)^* f(|T^n|)\xi_n, \xi_n\rangle = \|f(|T^n|)\xi_n\|^2_{\mathbb{H}} \geq 0,$$

we get that $\rho_{|T^n|,\xi_n}$ is a positive linear functional.

(2) We know that $\mathcal{C}(\sigma(|T^n|))$ is dense in $\mathcal{L}^2(\sigma(|T^n|), \mu_{|T^n|,\xi_n})$. For any $f, g \in \mathcal{C}(\sigma(|T^n|))$, we get

$$\begin{aligned}
&\langle f(|T^n|)\xi_n, g(|T^n|)\xi_n\rangle_{\mathbb{H}}\\
&= \langle g(|T^n|)^* f(|T^n|)\xi_n, \xi_n\rangle\\
&= \rho_{|T^n|,\xi_n}(\bar{g}f) = \int\limits_{\sigma(|T^n|)} f(z)\bar{g}(z)\mathrm{d}\mu_{|T^n|,\xi_n}(z)\\
&= \langle f, g\rangle_{\mathcal{L}^2(\sigma(|T^n|),\mu_{|T^n|,\xi_n})}\,.
\end{aligned}$$

Therefore,

$$R_{0,\xi_n} : \mathcal{C}(\sigma(|T^n|)) \to \mathbb{H}, \quad f(z) \to f(|T^n|)\xi_n$$

is a surjective isometry from $\mathcal{C}(\sigma(|T^n|))$ to $\mathcal{A}(|T^n|)\xi_n$.

Obviously, $\mathcal{C}(\sigma(|T^n|))$ and $\mathcal{A}(|T^n|)\xi_n$ are dense subspaces of $\mathcal{L}^2(\sigma(|T^n|), \mu_{|T^n|,\xi_n})$ and $\mathbb{H}$, respectively. Additionally, its closed extension

$$R_{|T^n|,\xi_n} : \mathcal{L}^2(\sigma(|T^n|), \mu_{|T^n|,\xi_n}) \to \mathbb{H}\,, \quad f(z) \to f(|T^n|)\xi_n$$

is an isomorphic operator.

Therefore, we get that $R_{|T^n|,\xi_n}$ is the uniquely isomorphic representation of $\mathbb{H}$ associated with the uniquely finite positive Borel measure $\mu_{|T^n|,\xi_n}$, which is complete. □

Let T be an invertible bounded linear operator on $\mathbb{H} = \mathbb{H}_{\xi^1} \oplus \mathbb{H}_{\xi^2} \oplus \cdots \oplus \mathbb{H}_{\xi^i} \oplus \cdots$ and ξ^i be a $\mathcal{A}(|T|^{-1})$-cyclic vector such that $\mathbb{H}_{\xi^i} = \overline{\mathcal{A}(|T|^{-1})\xi^i} = \overline{\mathcal{A}(|T|)\xi^i}$. If there exists a unitary operator $U_0 \in \mathcal{B}(\mathbb{H})$, such that $U_0\mathcal{P}(|T|^{-1}) = \mathcal{P}(|T^{-1}|)U_0$, then $\mathbb{H}_{U_0\xi^i} = \overline{\mathcal{A}(|T^{-1}|)U_0\xi^i} = U_0\mathbb{H}_{\xi^i}$ and we get two series of isomorphic representations

$$R_{|T|^{-1},\xi^i} : \mathcal{L}^2(\sigma(|T|^{-1}|_{\mathbb{H}_{\xi^i}}), \mu_{|T|^{-1},\xi^i}) \to \mathbb{H}_{\xi^i}\,, \qquad f(z) \to f(|T|^{-1})\xi^i$$

and

$$R_{|T^{-1}|,U_0\xi^i} : \mathcal{L}^2(\sigma(|T^{-1}||_{\mathbb{H}_{U_0\xi^i}}), \mu_{|T^{-1}|,U_0\xi^i}) \to \mathbb{H}_{U_0\xi^i}\,, \qquad g(y) \to g(|T^{-1}|)U_0\xi^i\,.$$

Let $\xi = \xi^1 \oplus \xi^2 \oplus \cdots \oplus \xi^i \oplus \cdots$. Subsequently, ξ is a $\mathcal{A}(|T|^{-1})$-cyclic vector, such that $\mathbb{H} = \overline{\mathcal{A}(|T|^{-1})\xi}$ and we get the following equation

$$R_{|T|^{-1},\xi} = R_{|T|^{-1},\xi^1} \oplus R_{|T|^{-1},\xi^2} \oplus \cdots \oplus R_{|T|^{-1},\xi^i} \oplus \cdots\,,$$

$$R_{|T^{-1}|,U_0\xi} = R_{|T^{-1}|,U_0\xi^1} \oplus R_{|T^{-1}|,U_0\xi^2} \oplus \cdots \oplus R_{|T^{-1}|,U_0\xi^i} \oplus \cdots\,,$$

$$\mathcal{L}^2(\sigma(|T|^{-1}), \mu_{|T|^{-1},\xi}) = \mathcal{L}^2(\sigma(|T|^{-1}|_{\mathbb{H}_{\xi^1}}), \mu_{|T|^{-1},\xi^1}) \oplus \cdots \oplus \mathcal{L}^2(\sigma(|T|^{-1}|_{\mathbb{H}_{\xi^i}}), \mu_{|T|^{-1},\xi^i}) \oplus \cdots\,,$$

and

$$\mathcal{L}^2(\sigma(|T^{-1}|), \mu_{|T^{-1}|,U_0\xi}) = \mathcal{L}^2(\sigma(|T^{-1}||_{\mathbb{H}_{U_0\xi^1}}), \mu_{|T^{-1}|,U_0\xi^1}) \oplus \cdots \oplus \mathcal{L}^2(\sigma(|T^{-1}||_{\mathbb{H}_{U_0\xi^i}}), \mu_{|T^{-1}|,U_0\xi^i}) \oplus \cdots\,.$$

3. Noncommutative Functional Calculus

We know that the spectral theory and functional calculus of normal operators [49] is very important in the study of operator theory and C^*-algebras [50]. Inspired by the Hua Loo-kang theorem on the automorphisms of a sfield [53], in this section, we give a useful construction from $\mathcal{L}^2(\sigma(|T^{-1}|), \mu_{|T^{-1}|,\eta})$ to $\mathcal{L}^2(\sigma(|T|^{-1}), \mu_{|T|^{-1},\xi})$ and with this construction, we give a noncommutative functional calculus for any given $T \in \mathcal{B}(\mathbb{H})$. However, there is valueless information just only from $R^{-1}_{|T|,\xi} \circ R_{|T^{-1}|,\eta}$ or $R^{-1}_{|T|^{-1},\xi} \circ R_{|T^{-1}|,\eta}$.

Lemma 4. *Let T be an invertible bounded linear operator on $\mathbb{H}$. Subsequently, we get*

$$\sigma(|T^{-1}|) = \sigma(|T|^{-1})\,.$$

Proof. Because of

$$\lambda \in \sigma(T^*T) \Longleftrightarrow \frac{1}{\lambda} \in \sigma(T^{*-1}T^{-1})\,,$$

we get

$$\lambda \in \sigma(|T|) \Longleftrightarrow \frac{1}{\lambda} \in \sigma(|T^{-1}|)\,.$$

That is, $\sigma(|T^{-1}|) = \sigma(|T|^{-1})$. □

Let T be an invertible bounded linear operator on $\mathbb{H}$, ξ be a $\mathcal{A}(|T|)$-cyclic vector, such that $\mathbb{H} = \overline{\mathcal{A}(|T|)\xi}$. On $\mathcal{P}(z)$, with $z \in \sigma(|T|)$, we define the mapping

$$F_{z^{-1}} : \mathcal{P}(z) \to \mathcal{P}(z^{-1}),\ F_{z^{-1}}(f(z)) = f(z^{-1})\ .$$

Following Lemma 3 and Theorem 1, $\mathcal{P}(z)$ and $\mathcal{P}(\frac{1}{z})$ are dense subspaces of $\mathcal{L}^2(\sigma(|T|), \mu_{|T|,\xi})$ and $\mathcal{L}^2(\sigma(|T|^{-1}), \mu_{|T|^{-1},\xi})$, respectively. Its closed extension

$$F_{z^{-1}} : \mathcal{L}^2(\sigma(|T|), \mu_{|T|,\xi}) \to \mathcal{L}^2(\sigma(|T|^{-1}), \mu_{|T|^{-1},\xi})\ ,\ \ F_{z^{-1}}(f(z)) = f(z^{-1})$$

is linear and for this closed extension we keep the notation $F_{z^{-1}}$.

Subsequently, we obtain that

$$\int\limits_{\sigma(|T|)} f(z^{-1})\mathrm{d}\mu_{|T|,\xi}(z) = \left\langle f(|T|^{-1})\xi, \xi \right\rangle = \int\limits_{\sigma(|T|^{-1})} f(z)\mathrm{d}\mu_{|T|^{-1},\xi}(z)$$

and

$$\mathrm{d}\mu_{|T|^{-1},\xi}(z) = |z|^2\mathrm{d}\mu_{|T|,\xi}(z)\ .$$

By a simple computation, we get that

$$\begin{aligned}
&\|F_{z^{-1}}(f(z))\|^2_{\mathcal{L}^2(\sigma(|T|^{-1}),\mu_{|T|^{-1},\xi})} \\
&= \int\limits_{\sigma(|T|^{-1})} F_{z^{-1}}(f(z))\bar{F}_{z^{-1}}(f(z))\mathrm{d}\mu_{|T|^{-1},\xi}(z) \\
&= \int\limits_{\sigma(|T|^{-1})} f(z^{-1})\bar{f}(z^{-1})\mathrm{d}\mu_{|T|^{-1},\xi}(z) \\
&= \int\limits_{\sigma(|T|)} |z|^2 f(z)\bar{f}(z)\mathrm{d}\mu_{|T|,\xi}(z) \\
&\leq \sup_{m\in\sigma(|T|)} m^2 \int\limits_{\sigma(|T|)} f(z)\bar{f}(z)\mathrm{d}\mu_{|T|,\xi}(z) \\
&\leq \sup_{m\in\sigma(|T|)} m^2 \|f(z)\|^2_{\mathcal{L}^2(\sigma(|T|),\mu_{|T|,\xi})}\ .
\end{aligned}$$

Hence, it follows that

$$\|F_{z^{-1}}\| \leq \sup_{m\in\sigma(|T|)} |m|\ .$$

By an application of the Banach inversion theorem [49] (p. 91), we get that $F_{z^{-1}}$ is an invertible bounded linear operator from $\mathcal{L}^2(\sigma(|T|), \mu_{|T|,\xi})$ to $\mathcal{L}^2(\sigma(|T|^{-1}), \mu_{|T|^{-1},\xi})$.

Next, we define the operator

$$F^{\mathbb{H}}_{z^{-1}} : \mathcal{A}(|T|)\xi \to \mathcal{A}(|T|^{-1})\xi,\quad F^{\mathbb{H}}_{z^{-1}}(f(|T|)\xi) = f(|T|^{-1})\xi\ .$$

By Lemma 1 and [51] (p. 55), we get that $F^{\mathbb{H}}_{z^{-1}}$ is an invertible bounded linear operator on the Hilbert space $\overline{\mathcal{A}(|T|)\xi} = \mathbb{H}$ and

$$\|F^{\mathbb{H}}_{z^{-1}}\| \leq \sup \sup_{m\in\sigma(|T|)} |m|\ .$$

Moreover, we obtain the following diagram.

$$\begin{array}{ccc} \mathbb{H} & \xrightarrow{\quad |T| \quad} & \mathbb{H} \\ F^{\mathbb{H}}_{z^{-1}} \downarrow & & \downarrow F^{\mathbb{H}}_{z^{-1}} \\ \mathbb{H} & \xrightarrow{\quad |T|^{-1} \quad} & \mathbb{H} \end{array}$$

By [53] and the isomorphic representations $R_{|T|^{-1},\xi}$ and $R_{|T^{-1}|,U_0\xi}$ of $\mathbb{H}$, also, by Lemma 2 and 3, naturally, we give the following definition.

Definition 2. *For invertible $T \in \mathcal{B}(\mathbb{H})$, let the symbol xx^* stand for $T^{-1}T^{*-1}$. Subsequently, we get that there is a linear algebraic isomorphism from $\mathcal{P}(xx^*)$ to $\mathcal{P}(x^*x)$, such that*

$$\mathcal{F}_{xx^*} : \mathcal{P}(xx^*) \to \mathcal{P}(x^*x), \quad p_n(xx^*) \to p_n(x^*x)\,.$$

Let ξ be a $\mathcal{A}(|T|^{-1})$-cyclic vector, such that $\overline{\mathcal{A}(|T|^{-1})\xi} = \mathbb{H}$ and $U_0 \in \mathcal{B}(\mathbb{H})$ be a unitary operator such that $U_0\mathcal{P}(|T|^{-1}) = \mathcal{P}(|T^{-1}|)U_0$. Subsequently, on $\sigma(|T|^{-1})$ we define

$$F_{xx^*} : R^{-1}_{|T|^{-1}} p_n(xx^*)\xi \to R^{-1}_{|T^{-1}|} p_n(x^*x)U_0\xi \ .$$

Obviously, $\mathcal{P}(|y|^2)$ is dense in $\mathcal{L}^2(\sigma(|T|^{-1}), \mu_{|T|^{-1},\xi})$ and $\mathcal{P}(|z|^2)$ is dense in $\mathcal{L}^2(\sigma(|T^{-1}|), \mu_{|T^{-1}|,U_0\xi})$. Then its closed extension is

$$F_{xx^*} : \mathcal{L}^2(\sigma(|T|^{-1}), \mu_{|T|^{-1},\xi})| \to \mathcal{L}^2(\sigma(|T^{-1}|), \mu_{|T^{-1}|,U_0\xi}) \ , \quad R^{-1}_{|T|^{-1}} f(xx^*)\xi \to R^{-1}_{|T^{-1}|} f(x^*x)U_0\xi \ .$$

For this closed extension, we keep the notation F_{xx^*}.

With the polar decomposition theorem [50] (p. 15), there is $T = U|T|$. For invertible $T \in \mathcal{B}(\mathbb{H})$, we get that

$$U^*T^*TU = TT^* \quad \text{and} \quad U^*|T|^{-2}U = |T^{-2}|\,.$$

In fact, when T is invertible, we can choose a special unitary operator, which shows that the operators $|T|^{-1}$ and $|T^{-1}|$ are unitary equivalent. This is explained in the following theorem.

Theorem 2. *Let T be an invertible bounded linear operator on $\mathbb{H}$ and $U_0 \in \mathcal{B}(\mathbb{H})$ be a unitary operator, such that $U_0\mathcal{P}(|T|^{-1}) = \mathcal{P}(|T^{-1}|)U_0$. Afterwards, there is a unitary operator $F^{\mathbb{H}}_{xx^*}$, such that*

$$F^{\mathbb{H}}_{xx^*}|T|^{-1} = |T^{-1}|F^{\mathbb{H}}_{xx^*}\,.$$

Moreover, $F^{\mathbb{H}}_{xx^*}$ is the corresponding unitary operator associated with the almost everywhere nonzero function $|\phi_{|T|}(z)|$, such that

$$\mathrm{d}\mu_{|T^{-1}|,U_0\xi} = |\phi_{|T|}(\frac{1}{z})|\mathrm{d}\mu_{|T|^{-1},\xi}\,,$$

where $|\phi_{|T|}(z)| \in \mathcal{L}^1(\sigma(|T|), \mu_{|T|,\xi})$ and ξ is a $\mathcal{A}(|T|)$-cyclic vector, such that $\overline{\mathcal{A}(|T|)\xi} = \mathbb{H}$.

Proof. By Lemma 3, let ξ be a $\mathcal{A}(|T|)$-cyclic vector, such that $\mathbb{H} = \overline{\mathcal{A}(|T|)\xi}$. By Definition 2, we have the linear operator $F_{xx^*} : \mathcal{L}^2(\sigma(|T|^{-1}), \mu_{|T|^{-1},\xi}) \to \mathcal{L}^2(\sigma(|T^{-1}|), \mu_{|T^{-1}|,U_0\xi})$.

This construction yields that F_{xx^*} is an invertible linear operator from $\mathcal{L}^2(\sigma(|T|^{-1}), \mu_{|T|^{-1},\xi})$ to $\mathcal{L}^2(\sigma(|T^{-1}|), \mu_{|T^{-1}|,U_0\xi})$. Hence, $F_{xx^*} \circ F_{z^{-1}}$ is an invertible linear operator from $\mathcal{L}^2(\sigma(|T|), \mu_{|T|,\xi})$ to $\mathcal{L}^2(\sigma(|T^{-1}|), \mu_{|T^{-1}|,U_0\xi})$.

By [53], we get that F_{xx^*} is a linear algebraic isomorphism from $\mathcal{P}(|y|^2)$ on $\sigma(|T|^{-1})$ to $\mathcal{P}(|z|^2)$ on $\sigma(|T^{-1}|)$. Additionally, by Lemma 2, $\mathcal{P}(|y|^2)$ is dense in $\mathcal{L}^2(\sigma(|T|^{-1}), \mu_{|T|^{-1},\xi})$ and $\mathcal{P}(|z|^2)$ is dense in $\mathcal{L}^2(\sigma(|T^{-1}|), \mu_{|T^{-1}|,U_0\xi})$.

Hence, we obtain

$$[\mathrm{d}\mu_{|T^{-1}|,U_0\xi}] = [\mathrm{d}\mu_{|T|^{-1},\xi}]\,,$$

that is, $\mathrm{d}\mu_{|T^{-1}|,U_0\xi}$ and $\mathrm{d}\mu_{|T|^{-1},\xi}$ are mutually absolutely continuous. Following [49], (IX Theorem 3.6) and the construction $F_{z^{-1}}$, we get that there exists $\phi_{|T|}(z) \in \mathcal{L}^1(\sigma(|T|), \mu_{|T|,\xi})$, where $|\phi_{|T|}(z)| \neq 0$, a.e., such that

$$\mathrm{d}\mu_{|T^{-1}|,U_0\xi} = |\phi_{|T|}(\frac{1}{z})|\mathrm{d}\mu_{|T|^{-1},\xi} = |z|^2|\phi_{|T|}(z)|\mathrm{d}\mu_{|T|,\xi}\ .$$

From Lemma 4, for any $p_n \in \mathcal{P}(\sigma(|T|^{-1})) \subseteq \mathcal{A}(\sigma(|T|^{-1}))$, because of

$$T^{*-1}p_n(|T|^{-1}) = p_n(|T^{-1}|)T^{*-1}\,,$$

with [50] (p. 60), we get that there is a unitary operator $U_0 \in \mathcal{B}(\mathbb{H})$, such that

$$U_0\mathcal{P}(|T|^{-1}) = \mathcal{P}(|T^{-1}|)U_0\ .$$

Hence, we conclude

$$U_0\mathcal{A}(|T|^{-1}) = \mathcal{A}(|T^{-1}|)U_0$$

and

$$\mathbb{H} = \overline{U_0\mathcal{A}(|T|^{-1})\xi} = \overline{\mathcal{A}(|T^{-1}|)U_0\xi}\ .$$

That is, $U_0\xi$ is a $\mathcal{A}(|T^{-1}|)$-cyclic vector. Additionally, with Theorem 1, we get

$$\int\limits_{\sigma(|T|^{-1})} f(z)\mathrm{d}\mu_{|T|^{-1},\xi}(z) = \left\langle f(|T|^{-1})\xi,\xi\right\rangle = \int\limits_{\sigma(|T|)} f(\frac{1}{z})\mathrm{d}\mu_{|T|,\xi}(z)$$

and

$$\int\limits_{\sigma(|T^{-1}|)} f(z)\mathrm{d}\mu_{|T^{-1}|,U_0\xi}(z) = \left\langle f(|T^{-1}|)U_0\xi, U_0\xi\right\rangle\ .$$

By a simple computation, we obtain that

$$\begin{aligned}
&\|F_{xx^*}\circ F_{z^{-1}}(f(z))\|^2_{\mathcal{L}^2(\sigma(|T^{-1}|),\mu_{|T^{-1}|,U_0\xi})}\\
&= \int\limits_{\sigma(|T^{-1}|)} F_{xx^*}\circ F_{z^{-1}}(f(z))\overline{F_{xx^*}\circ F_{z^{-1}}(f(z))}\mathrm{d}\mu_{|T^{-1}|,U_0\xi}(z)\\
&= \int\limits_{\sigma(|T^{-1}|)} F_{xx^*}(f(z^{-1}))\overline{F_{xx^*}(f(z^{-1}))}\mathrm{d}\mu_{|T^{-1}|,U_0\xi}(z)\\
&\triangleq \int\limits_{\sigma(|T|^{-1})} f(y^{-1})\bar{f}(y^{-1})\mathrm{d}\mu_{|T|^{-1},\xi}(y)\\
&= \int\limits_{\sigma(|T|^{-1})} F_{y^{-1}}(f(y))F_{y^{-1}}(\bar{f}(y))\mathrm{d}\mu_{|T|^{-1},\xi}(y)\\
&= \|F_{y^{-1}}(f(y))\|^2_{\mathcal{L}^2(\sigma(|T|^{-1}),\mu_{|T|^{-1},\xi})}
\end{aligned}$$

and $\triangleq$ is introduced by U_0.

Hence, F_{xx^*} is an isomorphism from $\mathcal{L}^2(\sigma(|T|^{-1}), \mu_{|T|^{-1},\xi})$ to $\mathcal{L}^2(\sigma(|T^{-1}|), \mu_{|T^{-1}|,U_0\xi})$.

With Theorem 1 and Definition 2, we have the operator

$$F^{\mathbb{H}}_{xx^*}:\mathbb{H}\to\mathbb{H},\quad F^{\mathbb{H}}_{xx^*}(R_{|T|^{-1},\xi}f(|T|^{-2}))=R_{|T^{-1}|,U_0\xi}(F_{xx^*}f(|T|^{-2})).$$

That is, $F^{\mathbb{H}}_{xx^*}=R_{|T^{-1}|,U_0\xi}F_{xx^*}R^{-1}_{|T|^{-1},\xi}$, such that the following diagram is valid.

$$\begin{array}{ccc}\mathcal{L}^2(\sigma(|T|^{-1}),\mu_{|T|^{-1},\xi}) & \xrightarrow{\quad R_{|T|^{-1},\xi}\quad} & \mathbb{H}\\ F_{xx^*}\downarrow & & \downarrow F^{\mathbb{H}}_{xx^*}\\ \mathcal{L}^2(\sigma(|T^{-1}|),\mu_{|T^{-1}|,U_0\xi}) & \xrightarrow[\quad R_{|T^{-1}|,U_0\xi}\quad]{} & \mathbb{H}\end{array}$$

Therefore, we said that the linear operator $F^{\mathbb{H}}_{xx^*}$ is associated with F_{xx^*} . Subsequently, we see that $F^{\mathbb{H}}_{xx^*}$ is a unitary operator and by Lemma 3, we obtain

$$\overline{\mathcal{A}(|T|^{-1})\xi}=\mathbb{H}=\overline{\mathcal{A}(|T^{-1}|)U_0\xi}\ .$$

Subsequently, we obtain

$$\begin{array}{ccc}\mathbb{H} & \xrightarrow{\quad |T|^{-1}\quad} & \mathbb{H}\\ F^{\mathbb{H}}_{xx^*}\downarrow & & \downarrow F^{\mathbb{H}}_{xx^*}\\ \mathbb{H} & \xrightarrow[\quad |T^{-1}|\quad]{} & \mathbb{H}\end{array}$$

Naturally, there is

$$F^{\mathbb{H}}_{xx^*}|T|^{-1}=|T^{-1}|F^{\mathbb{H}}_{xx^*}\ .$$

Afterwards, $F^{\mathbb{H}}_{xx^*}$ is the corresponding unitary operator associated with F_{xx^*} , which is, associated with the almost everywhere nonzero function $|\phi_{|T|}(\frac{1}{z})|$, such that

$$\mathrm{d}\mu_{|T^{-1}|,U_0\xi}=|\phi_{|T|}(\frac{1}{z})|\mathrm{d}\mu_{|T|^{-1},\xi}=|z|^2|\phi_{|T|}(z)|\mathrm{d}\mu_{|T|,\xi}\ ,$$

where $|\phi_{|T|}(z)|\in\mathcal{L}^1(\sigma(|T|),\mu_{|T|,\xi})$. □

We easily deduce $(F^{\mathbb{H}}_{xx^*})^*=F^{\mathbb{H}}_{xx^*}$ and the next results also readily follows. Let T be an invertible bounded linear operator on $\mathbb{H}$, ξ be a $\mathcal{A}(|T|)$-cyclic vector, such that $\mathbb{H}=\overline{\mathcal{A}(|T|)\xi}$ and let $U_0\in\mathcal{B}(\mathbb{H})$ be a unitary operator, such that $U_0\mathcal{P}(|T|^{-1})=\mathcal{P}(|T^{-1}|)U_0$. In the proof of Theorem 2, and with the isomorphic representations $R_{|T|^{-1},\xi}$ and $R_{|T^{-1}|,U_0\xi}$ of $\mathbb{H}$, we provide that

$$F_{xx^*}=R^{-1}_{|T^{-1}|,U_0\xi}\circ F^{\mathbb{H}}_{xx^*}\circ R_{|T|^{-1},\xi}\ .$$

Especially, let $U_0=F^{\mathbb{H}}_{xx^*}$. Subsequently,

$$F_{xx^*}=R^{-1}_{|T^{-1}|,F^{\mathbb{H}}_{xx^*}\xi}\circ F^{\mathbb{H}}_{xx^*}\circ R_{|T|^{-1},\xi}\ .$$

Corollary 1. *Let T be an invertible bounded linear operator on $\mathbb{H}$ and let ξ be a $\mathcal{A}(|T|)$-cyclic vector, such that $\mathbb{H}=\overline{\mathcal{A}(|T|)\xi}$. Then $\sigma(|T|)=\sigma(|T^*|)$ and the equality $F^{\mathbb{H}}_{xx^*}|T|=|T^*|F^{\mathbb{H}}_{xx^*}$ is valid. Moreover, $F^{\mathbb{H}}_{xx^*}$ is the corresponding unitary operator associated with F_{xx^*}, whih is, associated with the almost everywhere nonzero function $|\phi_{|T|}(z)|$, such that*

$$\mathrm{d}\mu_{|T^*|,F^{\mathbb{H}}_{xx^*}\xi}=|\phi_{|T|}(z)|\mathrm{d}\mu_{|T|,\xi}\ ,$$

where $|\phi_{|T|}(z)|\in\mathcal{L}^1(\sigma(|T|),\mu_{|T|,\xi})$.

Next, for any given $g(z) \in \mathcal{L}^\infty(\sigma(|T|), \mu_{|T|,\xi})$, we define

$$M_g : \mathcal{L}^2(\sigma(|T|), \mu_{|T|,\xi}) \to \mathcal{L}^2(\sigma(|T|), \mu_{|T|,\xi}), \quad M_g f(z) = g(z)f(z).$$

Theorem 3. *Let T be an invertible bounded linear operator on $\mathbb{H}$ and $T = U|T|$ be its Polar Decomposition. Let ξ be a $\mathcal{A}(|T|)$-cyclic vector, such that $\mathbb{H} = \overline{\mathcal{A}(|T|)\xi}$ and $|T| = R_{|T|,\xi} M_z R^{-1}_{|T|,\xi}$. Subsequently, there exists $\psi(z) \in \mathcal{L}^\infty(\sigma(|T|), \mu_{|T|,\xi})$, such that $U = R_{|T^*|,F^{\mathbb{H}}_{xx^*}\xi} F_{xx^*} M_{\psi(z)} R^{-1}_{|T|,\xi}$ and $T = R_{|T^*|,F^{\mathbb{H}}_{xx^*}\xi} F_{xx^*} M_{z\psi(z)} R^{-1}_{|T|,\xi}$. Here $M_{z\psi(z)} = M_z M_{\psi(z)} = M_{\psi(z)} M_z = M_{\psi(z)z}$.*

Proof. Let $U_0 = F^{\mathbb{H}}_{xx^*}$ in the proof of Theorem 2. Afterwards, we get

$$F_{xx^*} = R^{-1}_{|T^{-1}|,F^{\mathbb{H}}_{xx^*}\xi} \circ F^{\mathbb{H}}_{xx^*} \circ R_{|T|^{-1},\xi} \quad \text{and} \quad F^{\mathbb{H}}_{xx^*} = R_{|T^{-1}|,F^{\mathbb{H}}_{xx^*}\xi} \circ F_{xx^*} \circ R^{-1}_{|T|^{-1},\xi}.$$

By the polar decomposition theorem [50] (p. 15) we have $T = U|T|$. Hence, we get

$$T^*T = |T|^2 \quad \text{and} \quad TT^* = U|T|^2U^*.$$

By Corollary 1, we get

$$TT^* = F^{\mathbb{H}}_{xx^*}|T|^2(F^{\mathbb{H}}_{xx^*})^*,$$

that is,

$$F^{\mathbb{H}}_{xx^*}|T|^2(F^{\mathbb{H}}_{xx^*})^* = TT^* = U|T|^2U^*.$$

We see that

$$F^{\mathbb{H}}_{xx^*}U|T|^2 = |T|^2F^{\mathbb{H}}_{xx^*}U.$$

With the fact that $\{M_{\psi(z)} : \psi(z) \in \mathcal{L}^\infty(\sigma(|T|), \mu_{|T|,\xi})\}$ is a maximal abelian von Neumann algebra in $\mathcal{B}(\mathcal{L}^2(\sigma(|T|), \mu_{|T|,\xi}))$ and the Fuglede–Putnam theorem [49] (p. 279), we obtain that there exists $\psi(z) \in \mathcal{L}^\infty(\sigma(|T|), \mu_{|T|,\xi})$ such that

$$F^{\mathbb{H}}_{xx^*}U = R_{|T|,\xi} M_{\psi(z)} R^{-1}_{|T|,\xi} = R_{|T|^{-1},\xi} M_{\psi(\frac{1}{z})} R^{-1}_{|T|^{-1},\xi}.$$

Therefore, we get that

$$U = F^{\mathbb{H}}_{xx^*} R_{|T|^{-1},\xi} M_{\psi(\frac{1}{z})} R^{-1}_{|T|^{-1},\xi} = R_{|T^{-1}|,F^{\mathbb{H}}_{xx^*}\xi} \circ F_{xx^*} \circ R^{-1}_{|T|^{-1},\xi} R_{|T|^{-1},\xi} M_{\psi(\frac{1}{z})} R^{-1}_{|T|^{-1},\xi}.$$

That is,

$$U = R_{|T^{-1}|,F^{\mathbb{H}}_{xx^*}\xi} F_{xx^*} M_{\psi(\frac{1}{z})} R^{-1}_{|T|^{-1},\xi} = R_{|T^*|,F^{\mathbb{H}}_{xx^*}\xi} F_{xx^*} M_{\psi(z)} R^{-1}_{|T|,\xi}$$

and

$$T = U|T| = R_{|T^*|,F^{\mathbb{H}}_{xx^*}\xi} F_{xx^*} M_{\psi(z)} R^{-1}_{|T|,\xi} R_{|T|,\xi} M_z R^{-1}_{|T|,\xi} = R_{|T^*|,F^{\mathbb{H}}_{xx^*}\xi} F_{xx^*} M_{z\psi(z)} R^{-1}_{|T|,\xi}.$$

□

Corollary 2. *Let $T \in \mathcal{B}(\mathbb{H})$. Suppose $a \in \rho(T) = \mathbb{C} \setminus \sigma(T)$ and let ξ be a $\mathcal{A}(|T-a|)$-cyclic vector, such that $\mathbb{H} = \overline{\mathcal{A}(|T-a|)\xi}$. Subsequently, there exists a function $\psi(z) \in \mathcal{L}^\infty(\sigma(|T-a|), \mu_{|T-a|,\xi})$, such that*

$$T = R_{|(T-a)^*|,F^{\mathbb{H}}_{xx^*}\xi} F_{xx^*} M_{z\psi(z)} R^{-1}_{|T-a|,\xi} + a.$$

Proof. For $a \in \rho(T)$, $T - a$ is an invertible bounded linear operator on $\mathbb{H}$. By the proof of Theorem 3, we get that $T - a = R_{|(T-a)^*|,F^{\mathbb{H}}_{xx^*}\xi} F_{xx^*} M_{z\psi(z)} R^{-1}_{|T-a|,\xi}$, that is,

$$T = R_{|(T-a)^*|,F^{\mathbb{H}}_{xx^*}\xi} F_{xx^*} M_{z\psi(z)} R^{-1}_{|T-a|,\xi} + a.$$

□

The following definition is quite natural.

Definition 3. *For any given $T \in \mathcal{B}(\mathbb{H})$, we say that $R_{|(T-a)^*|,F^{\mathbb{H}}_{xx^*}\xi}F_{xx^*}M_{z\psi(z)}R^{-1}_{|T-a|,\xi}+a$ is the noncommutative functional calculus of T on $F_{xx^*}:\mathcal{L}^2(\sigma(|T|),\mu_{|T|,\xi})|\to\mathcal{L}^2(\sigma(|T^*|),\mu_{|T^*|,F^{\mathbb{H}}_{xx^*}\xi})$,, where ξ is a $\mathcal{A}(|T-a|)$-cyclic vector, such that $\mathbb{H}=\overline{\mathcal{A}(|T-a|)\xi}$, $\psi(z)\in\mathcal{L}^\infty(\sigma(|T-a|),\mu_{|T-a|,\xi})$ and $a\in\rho(T)$.*

In the final part of this section, we give some properties of normal operator through the noncommutative functional calculus.

Corollary 3. *For $T\in\mathcal{B}(\mathbb{H})$, if $TT^*=T^*T$, then there exists $\psi(z)\in\mathcal{L}^\infty(\sigma(|T-a|),\mu_{|T-a|,\xi})$ such that $R_{|(T-a)|,\xi}(M_{z\psi(z)}+a)R^{-1}_{|T-a|,\xi}$ is the noncommutative functional calculus of T on $\mathcal{L}^2(\sigma(|T-a|),\mu_{|T-a|,\xi})$. and we get $T\in\mathcal{A}^{'}(|T-a|)$. Where $a\in\rho(T)$, ξ is a $\mathcal{A}(|T-a|)$-cyclic vector, such that $\mathbb{H}=\overline{\mathcal{A}(|T-a|)\xi}$ and*

$$\mathcal{A}^{'}(|T-a|)=\{A\in\mathcal{B}(\mathbb{H}):AB=BA \text{ for every } B\in\mathcal{A}(|T-a|)\}.$$

Proof. For $TT^*=T^*T$ and $a\in\rho(T)$, we get that

$$F^{\mathbb{H}}_{xx^*}=identity \quad \text{and} \quad |T-a|=[(T-a)^*(T-a)]^{\frac{1}{2}}=[(T-a)(T-a)^*]^{\frac{1}{2}}=|(T-a)^*|.$$

Therefore, we see that

$$R^{-1}_{|(T-a)^*|,F^{\mathbb{H}}_{xx^*}\xi}=R^{-1}_{|T-a|,\xi}$$

and there exists $\psi(z)\in\mathcal{L}^\infty(\sigma(|T-a|),\mu_{|T-a|,\xi})$ such that

$$T-a=R_{|(T-a)|,\xi}M_{z\psi(z)}R^{-1}_{|T-a|,\xi}\,.$$

That is,

$$T=R_{|(T-a)|,\xi}(M_{z\psi(z)}+a)R^{-1}_{|T-a|,\xi}\,.$$

With the proof of Theorem 3, we get $T-a\in\mathcal{A}^{'}(|T-a|)$, which is, $T\in\mathcal{A}^{'}(|T-a|)$. □

Corollary 4. *Let $T\in\mathcal{B}(\mathbb{H})$. Subsequently, the operator T is normal if and only if T is unitary equivalent to $M_{\psi(z)}+a$ on $\mathcal{L}^2(\sigma(|T-a|),\mu_{|T-a|,\xi})$, and if and only if $T\in\mathcal{A}^{'}(|T-a|)$, where ξ is a $\mathcal{A}(|T-a|)$-cyclic vector, such that $\mathbb{H}=\overline{\mathcal{A}(|T-a|)\xi}$, $\psi(z)\in\mathcal{L}^\infty(\sigma(|T-a|),\mu_{|T-a|,\xi})$, and $a\in\rho(|T|)$.*

4. A Sufficient Condition

In this section, we study Problem 1 on infinite-dimensional separable Hilbert spaces. With the fact that the exist of non-trivial invariant subspace is unchanged by the similarity of bounded linear operators on Banach spaces [1], which is, for $R\in\mathcal{B}(\mathbb{B}_1)$ and $S\in\mathcal{B}(\mathbb{B}_2)$, if $T:\mathbb{B}_1\to\mathbb{B}_2$ is an invertible bounded linear operator and $S=TRT^{-1}$, then R has non-trivial invariant subspace if and only if S has, where $\mathbb{B}_1$ and $\mathbb{B}_2$ are Banach spaces. Therefore, for any given $T\in\mathcal{B}(\mathbb{H})$, using the construction of F_{xx^*}, we give a sufficient condition to Problem 1 on infinite-dimensional separable Hilbert spaces.

For convenience, we define $Fix(F^{\mathbb{H}}_{xx^*})=\{F^{\mathbb{H}}_{xx^*}(f)=f;f\in\mathbb{H}\}$. Obviously, $Fix(F^{\mathbb{H}}_{xx^*})$ is a closed subspace of $\mathbb{H}$.

Theorem 4. *Let $\dim\mathbb{H}>1$, $T\in\mathcal{B}(\mathbb{H})$ and $R_{|(T-a)^*|,F^{\mathbb{H}}_{xx^*}\xi}F_{xx^*}M_{z\psi(z)}R^{-1}_{|T-a|,\xi}$ be the noncommutative functional calculus of $T-a$ on $F_{xx^*}:\mathcal{L}^2(\sigma(|T-a|),\mu_{|T-a|,\xi})|\to\mathcal{L}^2(\sigma(|(T-a)^*|),\mu_{|(T-a)^*|,F^{\mathbb{H}}_{xx^*}\xi})$, where $a\in\rho(|T|)$, ξ is a $\mathcal{A}(|T-a|)$-cyclic vector, such that $\mathbb{H}=\overline{\mathcal{A}(|T-a|)\xi}$ and $\psi(z)\in\mathcal{L}^\infty(\sigma(|T-a|),\mu_{|T-a|,\xi})$. If $R_{|T-a|,\xi}M_{z\psi(z)}R^{-1}_{|T-a|,\xi}Fix(F^{\mathbb{H}}_{xx^*})\subseteq Fix(F^{\mathbb{H}}_{xx^*})$, then T has a non-trivial invariant subspace.*

Proof. It is enough to prove the result for infinite-dimensional separable complex Hilbert space $\mathbb{H}$. Obviously, if $A \subset \mathbb{H}$ is a non-trivial invariant subspace of T if and only if A is a non-trivial invariant subspace of $T-a$, where $a \in \mathbb{C}$.

Let $a \in \rho(T)$ and let ξ be a $\mathcal{A}(|T-a|)$-cyclic vector such that $\mathbb{H} = \overline{\mathcal{A}(|T-a|)\xi}$. Subsequently, following Corollary 2, we get that there exists $\psi(z) \in \mathcal{L}^\infty(\sigma(|T-a|), \mu_{|T-a|,\xi})$, such that $R_{|(T-a)^*|,F^{\mathbb{H}}_{xx^*}\xi}F_{xx^*}M_{z\psi(z)}R^{-1}_{|T-a|,\xi}$ is the noncommutative functional calculus of $T-a$ on

$$F_{xx^*} : \mathcal{L}^2(\sigma(|T-a|), \mu_{|T-a|,\xi})| \to \mathcal{L}^2(\sigma(|(T-a)^*|), \mu_{|(T-a)^*|,F^{\mathbb{H}}_{xx^*}\xi}) .$$

By the construction of $F^{\mathbb{H}}_{xx^*}$ in Theorem 2, we get $(F^{\mathbb{H}}_{xx^*})^2 = identity$ and $Fix(F^{\mathbb{H}}_{xx^*}) \neq \emptyset$.

(1) If $Fix(F^{\mathbb{H}}_{xx^*}) = \mathbb{H}$, that is $F^{\mathbb{H}}_{xx^*} = identity$, by the proof of Corollary 3, then T is unitary equivalent to $M_{z\psi(z)} + a$. Because $M_{z\psi(z)} + a$ is a normal operator, it possesses a non-trivial invariant subspace and, hence, the same is true for T. For details, see, e.g., [49].

(2) If $Fix(F^{\mathbb{H}}_{xx^*}) \neq \mathbb{H}$ and $R_{|T-a|,\xi}M_{z\psi(z)}R^{-1}_{|T-a|,\xi}Fix(F^{\mathbb{H}}_{xx^*}) \subseteq Fix(F^{\mathbb{H}}_{xx^*})$, then $Fix(F^{\mathbb{H}}_{xx^*})$ is a non-trivial invariant subspace of $F^{\mathbb{H}}_{xx^*}$ and we get

$$F^{\mathbb{H}}_{xx^*}R_{|T|,\xi}M_{z\psi(z)}R^{-1}_{|T|,\xi}Fix(F^{\mathbb{H}}_{xx^*}) \subseteq Fix(F^{\mathbb{H}}_{xx^*}).$$

Hence, $Fix(F^{\mathbb{H}}_{xx^*})$ is a non-trivial invariant subspace of $F^{\mathbb{H}}_{xx^*}R_{|T|,\xi}M_{z\psi(z)}R^{-1}_{|T|,\xi}$. With the proof of Theorem 3, we get that

$$F^{\mathbb{H}}_{xx^*}R_{|T|,\xi}M_{z\psi(z)}R^{-1}_{|T|,\xi} = R_{|(T-a)^*|,F^{\mathbb{H}}_{xx^*}\xi}F_{xx^*}M_{z\psi(z)}R^{-1}_{|T-a|,\xi} = T-a.$$

That is,

$$(T-a)Fix(F^{\mathbb{H}}_{xx^*}) \subseteq Fix(F^{\mathbb{H}}_{xx^*}).$$

□

5. Lebesgue Operator

In this section, we study chaos of an invertible bounded linear operator on an infinite-dimensional separable Hilbert space. For the example of integral calculus in mathematical analysis, we know that the convergence or the divergence of the weighted integral calculus of x and x^{-1} should be independent of each other; however, sometimes it happens that this indeed depends on a special choice of the weight function.

In the view of integral calculus, we define the Lebesgue class and prove that if T is a Lebesgue operator, then T is Li-Yorke chaotic if and only if T^{*-1} is. With the idea of the noncommutative functional calculus $R_{|(T-a)^*|,F^{\mathbb{H}}_{xx^*}\xi}F_{xx^*}M_{z\psi(z)}R^{-1}_{|T-a|,\xi}$, we give an example of a Lebesgue operator that is not a normal operator.

Let $\mathrm{d}x$ be the Lebesgue measure on $\mathcal{L}^2(\mathbb{R}_+)$. By Theorem 1, there exists a Borel measure $\mathrm{d}\mu_{|T^n|,\xi_n}$, which is complete, such that $\mathcal{L}^2(\sigma(|T^n|), \mathrm{d}\mu_{|T^n|,\xi_n})$ is a Hilbert space. If there exists $N > 0$, such that, for all $n \geq N$, the measure $\mathrm{d}\mu_{|T^n|,\xi_n}$ is absolutely continuity with respect to $\mathrm{d}x$, then using the Radon–Nikodym theorem [49] (p. 380), there exists $f_n \in \mathcal{L}^1(\mathbb{R}_+)$, such that $\mathrm{d}\mu_{|T^n|,\xi_n} = f_n(x)\mathrm{d}x$, where $n \in \mathbb{N}$, $n \geq N$ and $\mathbb{H} = \overline{\mathcal{A}(|T^n|)\xi_n}$.

Definition 4. *Let T be an invertible bounded linear operator on the separable Hilbert space $\mathbb{H}$ over $\mathbb{C}$. Suppose that the operator T satisfies the following conditions:*

(1) *There exists* $N \in \mathbb{N}$, *such that, for all* $n \geq N$

$$\begin{cases} \mathrm{d}\mu_{|T^n|,\xi_n} = f_n(x)\mathrm{d}x, \quad f_n \in \mathcal{L}^1(\mathbb{R}_+) \\ x^2 f_n(x) = f_n(x^{-1}), \qquad 0 < x < 1 \end{cases}$$

(2) *There exists* $N \in \mathbb{N}$, *such that for all* $n \geq N$ *and for any given nonzero* $x \in \mathbb{H}$, *there exists a nonzero function* $g_n(t) \in \mathcal{L}^2(\sigma(|T^n|), \mathrm{d}\mu_{|T^n|,\xi_n})$ *and a nonzero vector* $y \in \mathbb{H}$, *such that* $y = g_n(|T^n|^{-1})\xi_n$ *whenever* $x = g_n(|T^n|)\xi_n$.

Subsequently, the operator T *is said to be a Lebesgue operator, and the family of all Lebesgue operators on* $\mathbb{H}$ *is denoted by* $\mathcal{B}_{Leb}(\mathbb{H})$.

Theorem 5. *Let* T *be a Lebesgue operator on the separable Hilbert space* $\mathbb{H}$ *over* $\mathbb{C}$. *Subsequently,* T *is Li-Yorke chaotic if and only if* T^{*-1} *is.*

Proof. Let ξ_n be a $\mathcal{A}(|T^n|)$-cyclic vector such that

$$\overline{\mathcal{A}(|T^n|)\xi_n} = \mathbb{H}.$$

If x_0 is a Li-Yorke chaotic point of T, then by Definition 4, we see that, for $n \in \mathbb{N}$ large enough, there exist $g_n(x) \in \mathcal{L}^2(\sigma(|T^n|), \mathrm{d}\mu_{|T^n|,\xi_n})$, $f_n(x) \in \mathcal{L}^1(\mathbb{R}_+)$ and $y_0 \in \mathbb{H}$, such that $x_0 = g_n(|T^n|)\xi_n$, $y_0 = g_n(|T^n|^{-1})\xi_n$, and

$$\mathrm{d}\mu_{|T^n|,\xi_n} = f_n(x)\mathrm{d}x.$$

Therefore, we get the following

$$\begin{aligned}
\|T^n x_0\|_{\mathbb{H}}^2 &= \langle T^{n*}T^n x_0, x_0\rangle = \langle |T^n|^2 g_n(|T^n|)\xi_n, g_n(|T^n|)\xi_n\rangle = \langle g_n(|T^n|)^*|T^n|^2 g_n(|T^n|)\xi_n, \xi_n\rangle \\
&= \int\limits_{\sigma(|T^n|)} x^2 g_n(x)\bar{g}(x)\mathrm{d}\mu_{|T^n|,\xi_n}(x) = \int_0^{+\infty} x^2|g_n(x)|^2 f_n(x)\mathrm{d}x \\
&= \int_0^1 x^2|g_n(x)|^2 f_n(x)\mathrm{d}x + \int_1^{+\infty} x^2|g_n(x)|^2 f_n(x)\mathrm{d}x \\
&= \int_0^1 x^2|g_n(x)|^2 f_n(x)\mathrm{d}x + \int_0^1 x^{-4}|g_n(x^{-1})|^2 f_n(x^{-1})\mathrm{d}x \\
&\triangleq \int_0^1 |g_n(x)|^2 f_n(x^{-1})\mathrm{d}x + \int_0^1 x^{-2}|g_n(x^{-1})|^2 f_n(x)\mathrm{d}x \\
&= \int_1^{+\infty} x^{-2}|g_n(x^{-1})|^2 f_n(x)\mathrm{d}x + \int_0^1 x^{-2}|g_n(x^{-1})|^2 f_n(x)\mathrm{d}x \\
&= \int_0^{+\infty} x^{-2}|g_n(x^{-1})|^2 f_n(x)\mathrm{d}x = \int\limits_{\sigma(|T^n|)} x^{-2} g_n(x^{-1})\bar{g}_n(x^{-1})\mathrm{d}\mu_{|T^n|,\xi_n}(x) \\
&= \langle g_n(|T^n|^{-1})^*|T^n|^{-2} g_n(|T^n|^{-1})\xi_n, \xi_n\rangle = \langle |T^n|^{-2} g_n(|T^n|^{-1})\xi_n, g_n(|T^n|^{-1})\xi_n\rangle \\
&= \langle |T^n|^{-2} y_0, y_0\rangle = \langle T^{-n}T^{-n*}y_0, y_0\rangle \\
&= \|T^{*-n}y_0\|_{\mathbb{H}}^2
\end{aligned}$$

where $\triangleq$ is following Definition 4. By Definition 1, we get that T is Li-Yorke chaotic if and only if T^{*-1} is. □

Following [54], for $T \in \mathcal{B}(\mathbb{H})$, $x \in \mathbb{H}$ and $n \in \mathbb{N}$, we introduce the distributional function

$$F_x^n(\tau) = \frac{1}{n}\sharp\{0 \leq i \leq n : \|T^n(x)\| < \tau\}.$$

In addition, we denote

$$F_x(\tau) = \liminf_{n\to\infty} F_x^n(\tau), \qquad F_x^*(\tau) = \limsup_{n\to\infty} F_x^n(\tau),$$

and introduce the following definition.

Definition 5. *Let $T \in \mathcal{B}(\mathbb{H})$. If there exists $x \in \mathbb{H}$ and*

(1) If $F_x(\tau) = 0$, for some $\tau > 0$ and $F_x^*(\epsilon) = 1$ for $\forall \epsilon > 0$, then we say that T is distributionally chaotic or I-distributionally chaotic.

(2) If $F_x^*(\epsilon) > F_x(\tau)$ for $\forall \tau > 0$ and $F_x^*(\epsilon) = 1$ for $\forall \epsilon > 0$, then we say that T is II-distributionally chaotic.

(3) If $F_x^*(\epsilon) > F_x(\tau)$ for $\forall \tau > 0$, then we say that T is III-distributionally chaotic.

Corollary 5. *Let T be a Lebesgue operator on the separable Hilbert space $\mathbb{H}$ over $\mathbb{C}$. Then T is I-distributionally chaotic (or II-distributionally chaotic or III-distributionally chaotic) if and only if T^{*-1} is I-distributionally chaotic (or II-distributionally chaotic or III-distributionally chaotic).*

Theorem 6. *There exists an invertible bounded linear operator T on the separable Hilbert space $\mathbb{H}$ over $\mathbb{C}$, such that T is Lebesgue operator that is not a normal operator.*

Proof. Let $0 < a < b < +\infty$. Subsequently, $\mathcal{L}^2([a,b])$ is a separable Hilbert space over $\mathbb{C}$. Any separable Hilbert space over $\mathbb{R}$ can be expanded to a separable Hilbert space over $\mathbb{C}$. Without loss of generality, let $\mathcal{L}^2([a,b])$ be the separable Hilbert space over $\mathbb{R}$. We prove the conclusion by six parts:

(1) Let $0 < a < 1 < b = \frac{1}{a} < +\infty$. We construct a measure preserving transformation on $[a,b]$. Let $M = \{[a, \frac{b-a}{2}], [\frac{b-a}{2}, b]\}$. We get a Borel algebra $\xi(M)$ generated by M . We define $\Phi : [a,b] \to [a,b]$,

$$\Phi([a, \frac{b-a}{2}]) = [\frac{b-a}{2}, b], \qquad \Phi([\frac{b-a}{2}, b]) = [a, \frac{b-a}{2}].$$

Subsequently, Φ is an invertible measure preserving transformation on the Borel algebra $\xi(M)$. With [55] (p. 63), $U_\Phi \neq 1$ and U_Φ is a unitary operator associated with Φ, where U_Φ is the operation of composition

$$U_\Phi h = h \circ \Phi, \quad \forall h \in \mathcal{L}^2([a,b]).$$

(2) Define $M_x h = xh$ on $\mathcal{L}^2([a,b])$. Subsequently, M_x is an invertible positive operator.

(3) For $f(x) = \dfrac{|\ln x|}{x}$, $x > 0$, we define $\mathrm{d}\mu = f(x)\mathrm{d}x$. Afterwards, $f(x)$ is continuous and $f(x) > 0$, a.e., $x \in [a,b]$. Hence, $\mathrm{d}\mu$ that is absolutely continuous with respect to $\mathrm{d}x$ is a finite positive Borel measure that is complete. That is, $\mathcal{L}^2([a,b], \mathrm{d}\mu)$ is a separable Hilbert space over $\mathbb{R}$. Moreover, $\mathcal{L}^2([a,b])$ and $\mathcal{L}^2([a,b], \mathrm{d}\mu)$ are unitary equivalence.

(4) Let $T = U_\Phi M_x$. We get

$$T^*T = U_\Phi T T^* U_\Phi^* \qquad \text{and} \qquad U_\Phi \neq 1.$$

Because of

$$U_\Phi M_x \neq M_x U_\Phi \quad \text{and} \quad U_\Phi M_{x^2} \neq U_\Phi M_{x^2},$$

we get that T is not a normal operator and $\sigma(|T|) = [a,b]$.

(5) Let the operator $T = U_\Phi M_x$ on $\mathcal{L}^2([a,b])$ be corresponding to the operator T' on $\mathcal{L}^2([a,b], \mathrm{d}\mu)$. Subsequently, T' is an invertible bounded linear operator that is not a normal operator and $\sigma(|T'|) = [a,b]$.

(6) From

$$\int_a^b x^n f(x)\mathrm{d}x = \int_{a^n}^{b^n} t f(t^{\frac{1}{n}}) \frac{1}{n t^{\frac{n-1}{n}}} \mathrm{d}t \ .$$

Let

$$f_n(t) = \frac{1}{n} I_{[a^n, b^n]} f(t^{\frac{1}{n}}) \frac{1}{t^{\frac{n-1}{n}}} \ .$$

We get that $f_n(t)$ is continuous and almost everywhere positive. Hence, $f_n(t)\mathrm{d}t$ is a finite positive Borel measure that is complete.

For any $E \subseteq \mathbb{R}_+$, we define $I_E = 1$ when $x \in E$ else $I_E = 0$. Subsequently, I_E is the identity function on E. With a simple computing, we get that

$$f_n(t^{-1}) = \frac{1}{n} I_{[a^n,b^n]} f(t^{-\frac{1}{n}}) \frac{1}{t^{-\frac{n-1}{n}}} = \frac{1}{n} I_{[a^n,b^n]} \frac{|\ln t^{-\frac{1}{n}}|}{t^{-\frac{1}{n}}} \frac{1}{t^{-\frac{n-1}{n}}}$$
$$= \frac{1}{n} I_{[a^n,b^n]} t |\ln t^{\frac{1}{n}}|$$

and

$$t^2 f_n(t) = \frac{1}{n} I_{[a^n,b^n]} f(t^{\frac{1}{n}}) \frac{t^2}{t^{\frac{n-1}{n}}} = \frac{1}{n} I_{[a^n,b^n]} \frac{|\ln t^{\frac{1}{n}}|}{t^{\frac{1}{n}}} \frac{t^2}{t^{\frac{n-1}{n}}}$$
$$= \frac{1}{n} I_{[a^n,b^n]} t |\ln t^{\frac{1}{n}}| .$$

We see that $x^2 f_n(x) = f_n(x^{-1})$. From $\sigma(|T'^n|) = [a^n, b^n]$ and

$$\int_{a^n}^{b^n} t^2 f(t^{\frac{1}{n}}) \frac{1}{nt^{\frac{n-1}{n}}} \mathrm{d}t = \int_0^{+\infty} t^2 f_n(t) \mathrm{d}t ,$$

let $\mathrm{d}\mu_{|T'^n|} = f_n(t)\mathrm{d}t$.

Afterwards, $\mathrm{d}\mu_{|T'^n|}$ is a finite positive Borel measure that is complete. For any given nonzero $h(x) \in \mathcal{L}^2([a,b])$, we get the nonzero function $h(x^{-1}) \in \mathcal{L}^2([a,b])$.

Easily, we get that $I_{[a,b]}$ is a $\mathcal{A}(|M_x^n|)$-cyclic vector of the multiplication $M_x^n = M_{x^n}$ and $I_{[a^n,b^n]}$ is a $\mathcal{A}(|T'^n|)$-cyclic vector of $|T'^n|$. By Definition 4, we get that T' is Lebesgus operator, but is not a normal operator. □

Corollary 6. *There exists an invertible bounded linear operator T on the separable Hilbert space $\mathbb{H}$ over $\mathbb{C}$, such that T is a Lebesgue operator that is a positive operator.*

Corollary 7. *Let $\mathcal{B}_{Nor}(\mathbb{H})$ be the subspace of all normal bounded linear operator on an infinite-dimensional separable Hilbert space $\mathbb{H}$. Subsequently, the following families of linear operators are non-empty:*

$$\mathcal{B}_{Leb}(\mathbb{H}) \cap \mathcal{B}_{Nor}(\mathbb{H}) \text{ and } \mathcal{B}_{Leb}(\mathbb{H}) \cap (\mathcal{B}(\mathbb{H}) \setminus \mathcal{B}_{Nor}(\mathbb{H})).$$

In fact, both these families contain non-trivial members.

6. Conclusions

By the idea of the isomorphism construction F_{xx*} of this paper, we could study the operator using the integral calculus on $\mathbb{R}$. This way maybe neither change the properties of chaos nor the difficulty of computing, but with this we should find some operator class and study its properties, as we give the Lebesgue class in this section. Hence, if some properties of operators on $\mathbb{H}$ only depending the norm that is compatible with the inner product, then these properties only depend on the corresponding properties of elements in

$$\{M_{\psi_n(z)} + a_n : \psi_n(z) \in \mathcal{L}^\infty(\sigma(|T^n - a_n|), \mu_{|T^n - a_n|,\xi_n}), a_n \in \rho(T^n), n \in \mathbb{N}, \mathbb{H} = \overline{\mathcal{A}(|T^n - a_n|)\xi_n}\},$$

just keeping the noncommutative functional calculus in mind.

Funding: This research was funded by the National Nature Science Foundation of China (Grant No. 11801428).

Acknowledgments: This work is supported by the National Nature Science Foundation of China (Grant No. 11801428). I would like to thank the referee for his/her careful reading of the paper and helpful comments and

suggestions. Also, I shall extend my thanks to all those who have offered their help to me. Lastly, I sincerely appreciate the support and cultivation of Fudan University, Jilin University and Xidian University.

Conflicts of Interest: The author declare no conflict of interest.

References

1. Kaiser, J. Invariant Subspace Problem. Bachelor Thesis, Technische Universität Wien', Wien, NY, USA, 2016.
2. Bernstein, A.R.; Robinson, A. Solution of an invariant subspace problem of K. T. Smith and P. R. Halmos. *Pacific J. Math.* **1966**, *16*, 421–431. [CrossRef]
3. Aronszajn, N.; Smith, K.T. Invariant subspaces of completely continuous operators. *Ann. Math.* **1954**, *60*, 345–350. [CrossRef]
4. Lomonosov, V.I. Invariant subspaces of the family of operators that commute with a completely continuous operator. *Funkcional. Anal. Priložen.*, **1973**, *7*, 55–56. (In Russian) [CrossRef]
5. Fang, Q.; Xia, J. Invariant subspaces for certain finite-rank perturbations of diagonal operators. *J. Funct. Anal.* **2012**, *263*, 1356–1377. [CrossRef]
6. Kim, H. Hyperinvariant subspaces for operators having a normal part. *Oper. Matrices* **2011**, *5*, 487–494. [CrossRef]
7. Shi, L.; Wu, Y. Invariant and hyperinvariant subspaces for amenable operators. *J. Oper. Theory* **2013**, *69*, 87–100. [CrossRef]
8. Enflo, P.H. On the invariant subspace problem in Banach spaces. *Acta Math.* **1987**, *158*, 213–313. [CrossRef]
9. Nordgren, E.A.; Radjavi, H.; Rosenthal, P. A geometric equivalent of the invariant subspace problem. *Proc. Am. Math. Soc.* **1976**, *61*, 66–68. [CrossRef]
10. Atzmon, A. An operator without invariant subspaces on a nuclear Fréchet space. *Ann. Math.* **1983**, *117*, 669–694. [CrossRef]
11. Read, C.J. A solution to the invariant subspace problem. *Bull. Lond. Math. Soc.* **1984**, *16*, 337–401. [CrossRef]
12. Argyros, S.A.; Haydon, R.G. A hereditarily indecomposable $\mathscr{L}_\infty$-space that solves the scalar-plus-compact problem. *Acta Math.* **2011**, *206*, 1–54. [CrossRef]
13. Marcoux, L.W.; Popov, A.I.; Radjavi, H. On almost-invariant subspaces and approximate commutation. *J. Funct. Anal.* **2013**, *264*, 1088–1111. [CrossRef]
14. Elliott, G.A.; Yahaghi, B.R. Operators compatible with a chain of subspaces of a Banach space. *J. Ramanujan Math. Soc.* **2005**, *20*, 1–17.
15. Yahaghi, B.R. On injective or dense-range operators leaving a given chain of subspaces invariant. *Proc. Am. Math. Soc.* **2004** *132*, 1059–1066. [CrossRef]
16. Allan, G.R.; Zemánek, J. Invariant subspaces for pairs of projections. *J. Lond. Math. Soc.* **1998**, *57*, 449–468. [CrossRef]
17. Bernik, J.; Radjavi, H. Invariant and almost-invariant subspaces for pairs of idempotents. *Integral Equ. Oper. Theory* **2016**, *84*, 283–288. [CrossRef]
18. Popov, A.I.; Tcaciuc, A. Every operator has almost-invariant subspaces. *J. Funct. Anal.* **2013**, *265*, 257–265. [CrossRef]
19. Tcaciuc, A. The invariant subspace problem for rank-one perturbations. *Duke Math. J.* **2019**, *168*, 1539–1550. [CrossRef]
20. Doran, R.S. C^*-algebras: 1943–1993. A fifty year celebration. In *Proceedings of the AMS Special Session Held in San Antonio, Texas, 13–14 January 1993*; Contemporary Mathematics, 167; Doran, R.S., Ed.; American Mathmatical Society: New York, NY, USA, 1994.
21. Gelfand, I.M.; Naimark, M.A. On the embedding of normed linear rings into the ring of operators in Hilbert space. *Mat. Sbornik* **1943**, *12*, 197–213.
22. Fujimoto, I. A Gelfand-Naimark Theorem FOR C^*-Algebras. *Pac. J. Math.* **1998**, *184*, 95–119. [CrossRef]
23. Fell, J.M.G. The structure of algebras of operator fields. *Acta Math.* **1961**, *106*, 233–280. [CrossRef]
24. Fujimoto, I. A Gelfand representation for C^*-algebras. *Pac. J. Math.* **1971**, *39*, 1–11.
25. Kruszynski, P.; Woronowicz, S.L. A non-commutative Gelfand-Naimark theorem. *J. Oper. Theory* **1982**, *8*, 361–389.
26. Alfsen, E.M. On the Dirichlet problem of the Choquet boundary. *Acta Math.* **1968**, *120*, 149–159. [CrossRef]
27. Takesaki, M. A duality in the representation theory of C^*-algebras. *Ann. Math.* **1967**, *85*, 370–382. [CrossRef]

28. Kitai, C. Invariant Closed Sets for Linear Operators. Ph.D. Thesis, University of Toronto, Toronto, ON, Canada, 1982.
29. Bayart, F.; Matheron, E. *Dynamics of Linear Operators*; Cambridge University Press: Cambridge, UK, 2009.
30. Godefroy, G. Renorming of Banach spaces. In *Handbook of the Geometry of Banach Spaces*; North Holland: Amsterdam, The Netherlands, 2003; Volume 1, pp. 781–835.
31. Grosse-Erdmann, K.-G. Universal families and hypercyclic vectors. *Bull. Amerg. Math. Soc.* **1999**, *36*, 345–381. [CrossRef]
32. Grosse-Erdmann, K.-G.; Peris, A. *Linear Chaos*; Springer: London, UK, 2011.
33. Shapiro, J.H. Notes on the Dynamics of Linear Operators. 2001. Available online: http://joelshapiro.org/Pubvit/Downloads/LinDynamics/lindynamics.pdf (accessed on 5 May 2020).
34. Stockman, D. Li-Yorke Chaos in Models with Backward Dynamics. 2012. Available online: http://sites.udel.edu/econseminar/files/2012/03/lyc-backward.pdf (accessed on 5 May 2020).
35. Hou, B.; Luo, L. Li-Yorke chaos for invertible mappings on noncompact spaces. *Turk. J. Math.* **2016**, *40*, 411–416. [CrossRef]
36. Luo, L.; Hou, B. Some remarks on distributional chaos for bounded linear operators. *Turk. J. Math.* **2015**, *39*, 251–258. [CrossRef]
37. Luo, L.; Hou, B. Li-Yorke chaos for invertible mappings on compact metric spaces. *Arch. Math. (Basel)* **2017**, *108*, 65–69. [CrossRef]
38. Luo, L. Cowen-Douglas function and its application on chaos. *Ann. Funct. Anal.* **2020**, *11*, 897–913. [CrossRef]
39. Colombo, F.; Sabadini, I.; Struppa, D.C. A new functional calculus for noncommuting operators. *J. Funct. Anal.* **2008**, *254*, 2255–2274. [CrossRef]
40. Colombo, F.; Sabadini, I.; Struppa, D.C. Slice monogenic functions. *Isr. J. Math.* **2009**, *171*, 385–403. [CrossRef]
41. Alpay, D.; Colombo, F.; Qian, T.; Sabadini, I. The H^∞ functional calculus based on the S-spectrum for quaternionic operators and for n-tuples of noncommuting operators. *J. Funct. Anal.* **2016**, *271*, 1544–1584. [CrossRef]
42. Colombo, F.; Sabadini, I.; Struppa, D.C. A functional calculus for n-tuples of noncommuting operators. *Adv. Appl. Clifford Algebr.* **2009**, *19*, 225–236. [CrossRef]
43. Colombo, F.; Sabadini, I. The Cauchy formula with s-monogenic kernel and a functional calculus for noncommuting operators. *J. Math. Anal. Appl.* **2011**, *373*, 655–679. [CrossRef]
44. Colombo, F.; Gantner, J. An application of the S-functional calculus to fractional diffusion processes. *Milan J. Math.* **2018**, *86*, 225–303. [CrossRef]
45. Colombo, F.; Gentili, G.; Sabadini, I.; Struppa, D. C. Non commutative functional calculus: bounded operators. *Complex Anal. Oper. Theory* **2010**, *4*, 821–843. [CrossRef]
46. Colombo, F.; Gentili, G.; Sabadini, I.; Struppa, D.C. Non-commutative functional calculus: unbounded operators. *J. Geom. Phys.* **2010**, *60*, 251–259. [CrossRef]
47. Colombo, F.; Sabadini, I.; Struppa, D.C. *Noncommutative Functional Calculus: Theory and Applications of Slice Hyperholomorphic Functions*; Progress in Mathematics, 289; Birkhäuser/Springer Basel AG: Basel, Switzerland, 2011.
48. Monguzzi, A.; Sarfatti, G. Shift invariant subspaces of slice $\mathcal{L}^2$ functions. *Ann. Acad. Sci. Fenn. Math.* **2018**, *43*, 1045–1061. [CrossRef]
49. Conway, J.B. *A Course in Functional Analysis*, 2nd ed.; Springer-Verlag: New York, NY, USA, 1990.
50. Conway, J.B. *A Course in Operator Theory*; American Mathmatical Society: New York, NY, USA, 2000.
51. Arveson, W. *A Short Course on Spectral Theory*; Springer Science+Businee Media, LLC: New York, NY, USA, 2002.
52. Halmos, P.R. *Measure Theory*; Springer-Verlag: New York, NY, USA, 1974.
53. Hua, L. On the automorphisms of a sfield. *Proc. Nat. Acad. Sci. USA* **1949**, *35*, 386–389. [CrossRef]
54. Schweizer, B.; Smítal, J. Measures of chaos and a spectral decomposition of dynamical systems on the interval. *Trans. Amer. Math. Soc.* **1994**, *344*, 737–754. [CrossRef]
55. Walters, P. *An Introduction to Ergodic Theory*; Springer-Verlag: New York, NY, USA, 1982.

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 1-20, January 2020.

SINGULAR VALUES, EIGENVALUES AND DIAGONAL ELEMENTS OF THE COMMUTATOR OF 2×2 RANK ONE MATRICES*

CHE-MAN CHENG† AND YARU LIANG‡

Dedicated to Professor Yik-Hoi Au-Yeung

Abstract. The region of singular values of the commutator $XY - YX$ for 2×2 rank one complex matrices X and Y is determined. This answers in affirmative a conjecture raised in [D. Wenzel. A strange phenomenon for the singular values of commutators with rank one matrices. *Electron. J. Linear Algebra*, 30:649–669, 2015.] when 2×2 matrices are concerned. The approach and proofs also lead to a complete relation between the singular values, eigenvalues and diagonal elements of the commutator under consideration.

Key words. Commutator, Singular value, Eigenvalue, Diagonal element.

AMS subject classifications. 15A18.

1. Introduction.

1.1. Background and main results. Let $\mathbb{F}$ denote the set of real numbers $\mathbb{R}$ or the set of complex numbers $\mathbb{C}$, and let $\mathbf{i} = \sqrt{-1}$. We use column vectors for vectors in $\mathbb{F}^n$, and use row n-tuples for points in $\mathbb{F}^n$. The Euclidean inner product and norm on $\mathbb{F}^n$ are denoted by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$, respectively. Let $M_n(\mathbb{F})$ denote the set of $n \times n$ matrices with entries in $\mathbb{F}$. We use also $\| \cdot \|$ to denote the Frobenius norm on $M_n(\mathbb{F})$. For $X \in M_n(\mathbb{F})$, let $s_1(X) \geq \cdots \geq s_n(X)$ denote the singular values of X arranged in non-increasing order, and let $s(X) = (s_1(X), \ldots, s_n(X))^T$. Let $\|X\|_1 = s_1(X) + \cdots + s_n(X)$ denote the trace norm (also known as Schatten 1-norm and Ky-Fan n-norm) of X. Be aware that two norms are used in this paper. By a norm one matrix X it is always meant $\|X\| = 1$ unless otherwise stated. For $X, Y \in M_n(\mathbb{F})$, the commutator of X and Y is defined and denoted by

$$[X, Y] = XY - YX.$$

We assume $n > 1$ throughout the paper to avoid trivial situations.

Let

$$\Sigma_n(\mathbb{F}) = \{X : X \in M_n(\mathbb{F}), s(X) = (1, 0, \ldots, 0)^T\},$$

which is the set of rank one norm one matrices in $M_n(\mathbb{F})$. When $X, Y \in \Sigma_n(\mathbb{F})$, the rank of the commutator $[X, Y]$ is at most two. Let

$$\mathcal{S}_n^{\mathbb{F}} = \{(s_1([X, Y]), s_2([X, Y])) : X, Y \in \Sigma_n(\mathbb{F})\} \subset \mathbb{R}^2. \tag{1.1}$$

*Received by the editors on August 9, 2018. Accepted for publication on November 6, 2019. Handling Editor: Ilya Spitkovski. Corresponding Author: Che-Man Cheng. This research was supported by research grant MYRG2016-00121-FST from the University of Macau.

†Department of Mathematics, University of Macau, Macao, China (fstcmc@um.edu.mo).

‡Department of Mathematics, University of Macau, Macao, China (liangyaru1993@163.com).

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 1-20, January 2020.
http://repository.uwyo.edu/ela

It is proved in [7] that the set $\mathcal{S}_n^{\mathbb{R}}$ is the region $\mathcal{R}$ (see Figure 2.1) bounded by the segment joining $(0,0)$ and $(1,1)$, the segment joining $(0,0)$ and $(1,0)$, the segment joining $(1,0)$ and $\left(\frac{\sqrt{2}+1}{2}, \frac{\sqrt{2}-1}{2}\right)$, and the curve

$$\frac{4\sqrt{\cos\phi \sin\phi}}{1+2\cos\phi\sin\phi}(\cos\phi, \sin\phi), \quad \phi \in \left[\tan^{-1}\left(\tfrac{\sqrt{2}-1}{\sqrt{2}+1}\right), \frac{\pi}{4}\right]. \tag{1.2}$$

For an alternative characterization of $\mathcal{S}_n^{\mathbb{R}}$, see Theorem 1.5 below. It is also conjectured in [7, Conjecture 3.6] that $\mathcal{S}_n^{\mathbb{C}} = \mathcal{R}$. Numerical experiments highly suggest that this is true. Sadly, the approach used in [7] relies heavily on real numbers (in the form of angles) and cannot directly be adopted to the complex case.

When $X, Y \in \Sigma_n(\mathbb{F})$, we may assume $X = \mathbf{a}\mathbf{b}^*$ and $Y = \mathbf{c}\mathbf{d}^*$ where $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d} \in \mathbb{F}^n$ are unit vectors. It is shown in [7, Theorem 4.1] that $s_1([X,Y])$ and $s_2([X,Y])$ depend solely on

$$A = \langle \mathbf{a}, \mathbf{c}\rangle, \quad B = \langle \mathbf{b}, \mathbf{d}\rangle, \quad C = \langle \mathbf{c}, \mathbf{b}\rangle, \quad D = \langle \mathbf{d}, \mathbf{a}\rangle.$$

Based on these inner products, the result is proved. The main purpose of this paper is to prove in affirmative that the conjecture is true for 2×2 matrices. During our investigation, we found that there is a point in the proof in [7] that is not clear when 2×2 matrices are concerned. Let us first point out the difference between the cases $n = 2$ and $n \geq 3$.

It is trivial that $\mathcal{S}_2^{\mathbb{F}} \subseteq \mathcal{S}_3^{\mathbb{F}} \subseteq \mathcal{S}_4^{\mathbb{F}} \subseteq \cdots$. When $n > 4$ and $X, Y \in \Sigma_n(\mathbb{F})$, there exists a unitary (orthogonal if $\mathbb{F} = \mathbb{R}$) matrix $U \in M_n(\mathbb{F})$ such that $U^*XU, U^*YU \in M_4(\mathbb{F}) \oplus 0_{n-4}$. Consequently we know that $\mathcal{S}_k^{\mathbb{F}} = \mathcal{S}_4^{\mathbb{F}}$ for all $k > 4$. Using the following proposition, we can extend the result to 3×3 matrices to have $\mathcal{S}_k^{\mathbb{F}} = \mathcal{S}_3^{\mathbb{F}}$ for all $k > 3$.

PROPOSITION 1.1. *Suppose $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d} \in \mathbb{F}^4$ are unit vectors. Then there are unit vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4 \in \mathbb{F}^3$ such that*

$$\langle \mathbf{v}_1, \mathbf{v}_3\rangle = \langle \mathbf{a}, \mathbf{c}\rangle, \quad \langle \mathbf{v}_2, \mathbf{v}_4\rangle = \langle \mathbf{b}, \mathbf{d}\rangle, \quad \langle \mathbf{v}_3, \mathbf{v}_2\rangle = \langle \mathbf{c}, \mathbf{b}\rangle, \quad \langle \mathbf{v}_4, \mathbf{v}_1\rangle = \langle \mathbf{d}, \mathbf{a}\rangle.$$

Proof. By choosing a suitable unitary (orthogonal if $\mathbb{F} = \mathbb{R}$) matrix $U \in M_4(\mathbb{F})$ and considering $U\mathbf{x}$ for $\mathbf{x} \in \{\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}\}$, we may assume

$$\mathbf{a} = (a_1, 0, 0, 0)^T, \quad \mathbf{b} = (b_1, b_2, 0, 0)^T, \quad \mathbf{c} = (c_1, c_2, c_3, 0)^T, \quad \mathbf{d} = (d_1, d_2, d_3, d_4)^T.$$

The vectors

$$\mathbf{v}_1 = (a_1, 0, 0)^T, \quad \mathbf{v}_2 = (b_1, b_2, 0)^T, \quad \mathbf{v}_3 = (c_1, c_2, c_3)^T, \quad \mathbf{v}_4 = (d_1, d_2, \sqrt{|d_3|^2 + |d_4|^2})^T.$$

serve our purpose. ∎

The situation is different when $n = 2$. Suppose we choose

$$\mathbf{a} = (1, 0, 0, 0)^T, \quad \mathbf{b} = (0, 1, 0, 0)^T, \quad \mathbf{c} = (0, 0, 1, 0)^T, \quad \mathbf{d} = (1/\sqrt{2}, 0, 0, 1/\sqrt{2})^T.$$

Then

$$A = \langle \mathbf{a}, \mathbf{c}\rangle = 0, \quad B = \langle \mathbf{b}, \mathbf{d}\rangle = 0, \quad C = \langle \mathbf{c}, \mathbf{b}\rangle = 0, \quad D = \langle \mathbf{d}, \mathbf{a}\rangle = 1/\sqrt{2}.$$

However, for unit vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d} \in \mathbb{F}^2$,

$$A = \langle \mathbf{a}, \mathbf{c}\rangle = 0, \quad B = \langle \mathbf{b}, \mathbf{d}\rangle = 0, \quad C = \langle \mathbf{c}, \mathbf{b}\rangle = 0$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 1-20, January 2020.
http://repository.uwyo.edu/ela

imply $D = \langle \mathbf{d}, \mathbf{a} \rangle = 0$. Thus, we know that not all inner products (A, B, C, D) that can be achieved by vectors in $\mathbb{F}^4$ can be achieved by vectors in $\mathbb{F}^2$. It is then not clear that $\mathcal{S}_2^{\mathbb{R}}$ is not a proper subset of $\mathcal{S}_4^{\mathbb{R}}$ $(= \mathcal{R})$, although numerical experiments strongly suggest $\mathcal{S}_2^{\mathbb{R}} = \mathcal{R}$ and the boundary of $\mathcal{R}$ can be achieved by 2×2 real matrices (see the proof of [7, Proposition 3.3]).

We will first show in Section 3 that the smaller freedom in order 2 does not change the result.

THEOREM 1.2. $\mathcal{S}_2^{\mathbb{R}} = \mathcal{R}$.

This is not merely to give an alternative proof for 2×2 real matrices. The proof here also reveals that all the possible combinations of the singular values can be achieved by commutators having real eigenvalues and hence are orthogonally upper triangularizable. This fact is used in Section 4 for proving our main theorem.

THEOREM 1.3. $\mathcal{S}_2^{\mathbb{C}} = \mathcal{R}$.

Our approach and proofs also give immediately interesting results relating the singular values, eigenvalues and diagonal elements of the commutators under consideration. Before going to the lengthy proofs of Theorems 1.2 and 1.3, we include below a discussion on the results.

1.2. Singular values, eigenvalues and diagonal elements. Suppose $X, Y \in \Sigma_2(\mathbb{F})$ and $[X, Y] = \left[\begin{smallmatrix} \lambda & \delta \\ 0 & -\lambda \end{smallmatrix}\right]$ has singular values s_1 and s_2, and eigenvalues $\pm\lambda$. It follows readily from the Böttcher-Wenzel inequality (e.g. [2, 6]) that $|\lambda| \leq 1$ because

$$2|\lambda|^2 \leq \|[X, Y]\|^2 \leq 2\|X\|^2\|Y\|^2 = 2.$$

A simple proof of the inequality for 2×2 real matrix can be found in [1]. The proof there can easily be modified for 2×2 complex matrices. Our formulation leads us to consider the possible values of $|\delta|$ with $|\lambda|$ being fixed. The key result is that, for both the cases $\mathbb{F} = \mathbb{R}$ and $\mathbb{F} = \mathbb{C}$, $|\delta|$ can assume every value between 0 and a common maximum value $\delta_{|\lambda|}$ where $\delta_{|\lambda|}^2$ is given by

$$\delta_{|\lambda|}^2 = \begin{cases} 1 & \text{if } \ 0 \leq |\lambda| \leq 1/2, \\ 4|\lambda| - 4|\lambda|^2 & \text{if } \ 1/2 < |\lambda| \leq 1. \end{cases} \tag{1.3}$$

The graph of $\delta_{|\lambda|}^2$ is given below.

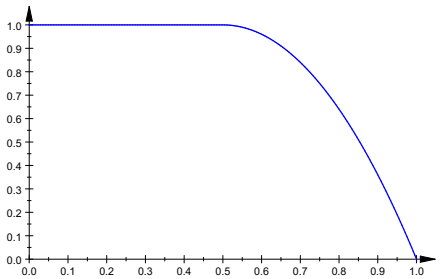


Figure 1.1. The graph of $\delta_{|\lambda|}^2$.

It is obvious that $\delta_{|\lambda|}^2$, and hence, $\delta_{|\lambda|}$ is non-increasing. This plain-looking fact will play a critical role in our later proof in Section 4.

When $X, Y \in \Sigma_2(\mathbb{C})$, $[X, Y]$ is unitarily triangularizable. Our key result asserts that when the complex commutator in triangular form has real eigenvalues and real δ, it can also be achieved by $X, Y \in \Sigma_2(\mathbb{R})$. On

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 1-20, January 2020.
http://repository.uwyo.edu/ela

the other hand, it is easy to deduce that

$$|\delta| = s_1 - s_2. \tag{1.4}$$

Consequently, together with the obvious condition $|\lambda|^2 = s_1 s_2$, we can easily deduce the following two theorems. The first one gives the relation between the eigenvalues and singular values of the commutators, and the second one gives a simple characterization on the singular values of the commutators.

THEOREM 1.4. *There exist $X, Y \in \Sigma_2(\mathbb{C})$ such that $[X, Y]$ has eigenvalues $\pm\lambda$ and singular values $s_1 \geq s_2$ if and only if $|\lambda| \leq 1$, $|\lambda|^2 = s_1 s_2$ and*

$$\begin{cases} s_1 - s_2 \leq 1 & \text{if } 0 \leq |\lambda| \leq 1/2, \\ s_1 + s_2 \leq 2\sqrt{|\lambda|} & \text{if } 1/2 < |\lambda| \leq 1. \end{cases}$$

Moreover, X and Y can be taken to be real if λ is real.

THEOREM 1.5. *There exist $X, Y \in \Sigma_2(\mathbb{C})$ such that $[X, Y]$ has singular values $s_1 \geq s_2$ if and only if $s_1 s_2 \leq 1$ and*

$$\begin{cases} s_1 - s_2 \leq 1 & \text{if } 0 \leq \sqrt{s_1 s_2} \leq 1/2, \\ s_1 + s_2 \leq 2(s_1 s_2)^{1/4} & \text{if } 1/2 < \sqrt{s_1 s_2} \leq 1. \end{cases}$$

Moreover, the singular values can always be attained by real matrices.

For $A \in M_n(\mathbb{C})$, the numerical range and numerical radius of A are defined respectively by

$$W(A) = \{x^* A x : x \in \mathbb{C}^n, \|x\| = 1\} \quad \text{and} \quad w(A) = \max\{|z| : z \in W(A)\}.$$

The study of the numerical range and numerical radius has a long history and is extensive. One may refer to [5, Chapter 1] for more information. For $[X, Y] = \left[\begin{smallmatrix} \lambda & \delta \\ 0 & -\lambda \end{smallmatrix}\right]$, the Elliptical Range Theorem (e.g., [5, Theorem 1.3.6]) tells us that $W([X, Y])$ is an elliptical disk with foci $\pm\lambda$ and minor axis $|\delta|$. Thus, from the above discussion, we have the following theorem.

THEOREM 1.6. *There exist $X, Y \in \Sigma_2(\mathbb{C})$ such that $W([X, Y])$ is an ellipse with foci $\pm\lambda$ ($\lambda \in \mathbb{C}$) and minor axis $\delta \geq 0$ if and only if*

$$0 \leq \delta \leq \begin{cases} 1 & \text{if } 0 \leq |\lambda| \leq 1/2, \\ 2\sqrt{|\lambda| - |\lambda|^2} & \text{if } 1/2 < |\lambda| \leq 1. \end{cases}$$

Moreover, X and Y can be taken to be real if λ is real.

From Theorem 1.6, we have

COROLLARY 1.7. *There exist $X, Y \in \Sigma_2(\mathbb{C})$ such that $[X, Y]$ has eigenvalues $\pm\lambda$ and $w([X, Y]) = r$ if and only if $0 \leq |\lambda| \leq 1$ and*

$$|\lambda| \leq r \leq \begin{cases} \sqrt{|\lambda|^2 + 1/4} & \text{if } 0 \leq |\lambda| \leq 1/2, \\ \sqrt{|\lambda|} & \text{if } 1/2 < |\lambda| \leq 1. \end{cases}$$

Moreover, X and Y can be taken to be real if λ is real.

The set $W(A)$ can be regarded as the collection of all values for the first diagonal entry of $U^* A U$ when U varies over all unitary matrices. From Corollary 1.7, and replacing $|\lambda|$ there by $\sqrt{s_1 s_2}$, we have the following relation between the singular values and diagonal elements.

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 1-20, January 2020.
http://repository.uwyo.edu/ela

COROLLARY 1.8. *There exist $X, Y \in \Sigma_2(\mathbb{C})$ such that $[X, Y]$ has singular values $s_1 \geq s_2$ and a diagonal element d if and only if $s_1 s_2 \leq 1$ and*

$$|d| \leq \begin{cases} \sqrt{s_1 s_2 + 1/4} & \text{if } 0 \leq \sqrt{s_1 s_2} \leq 1/2, \\ (s_1 s_2)^{1/4} & \text{if } 1/2 < \sqrt{s_1 s_2} \leq 1. \end{cases}$$

Moreover, X and Y can be taken to be real if d is real.

The elliptical disk with foci $\pm\lambda$ and minor axis $|\delta|$ is $\left\{z : |z-\lambda| + |z+\lambda| \leq 2\sqrt{|\lambda|^2 + |\delta|^2/4}\right\}$. From Theorem 1.4 and using (1.4), we can now have our ultimate result relating the singular values, eigenvalues and diagonal elements of the commutators under consideration.

THEOREM 1.9. *There exist $X, Y \in \Sigma_2(\mathbb{C})$ such that $[X, Y]$ has singular values $s_1 \geq s_2$, eigenvalues $\pm\lambda$ and diagonal elements $\pm d$ if and only if, in addition to the necessary conditions in Theorem 1.4,*

$$|d + \lambda| + |d - \lambda| \leq s_1 + s_2.$$

Moreover, X and Y can be taken to be real if λ and d are real.

Finally, we mention here another consequence of our study. There is a close relation between the region $\mathcal{S}_n^{\mathbb{C}}$ and the determination of the best (smallest) constant $C_{p,1,1}$ such that

$$\|XY - YX\|_p \leq C_{p,1,1}\|X\|_1\|Y\|_1, \quad X, Y \text{ are } n \times n \text{ complex matrices},$$

where $\|\cdot\|_p$ denotes the Schatten p-norm, $1 \leq p \leq \infty$. When $2 < p < \infty$, this is an unsolved situation of the general problem (see [8, 3]) of finding the best constant $C_{p,q,r}$ such that

$$\|XY - YX\|_p \leq C_{p,q,r}\|X\|_q\|Y\|_r, \quad X, Y \text{ are } n \times n \text{ complex matrices}.$$

For more information on commutator norm inequalities, see the surveys [2, 6]. In fact, we have $C_{p,1,1} = \max\{\|\mathbf{x}\|_p : \mathbf{x} \in \mathcal{S}_n^{\mathbb{C}}\}$ in which we also use $\|\cdot\|_p$ to denote the vector p-norm. In [4], the constant $C_{p,1,1}^{\mathbb{R}} = \max\{\|\mathbf{x}\|_p : \mathbf{x} \in \mathcal{S}_n^{\mathbb{R}}\}$ for real matrices is found via the determination of $C_{\infty,q,1}^{\mathbb{R}}$ for real matrices. Theorem 1.3 tells us that $\mathcal{S}_2^{\mathbb{C}} = \mathcal{S}_2^{\mathbb{R}}$ and consequently we can conclude that all the results obtained in [4] for real matrices are also true for 2×2 complex matrices.

2. Transforming the problem geometrically. Our approach is to consider, instead of the singular values $s_1([X, Y])$ and $s_2([X, Y])$ of the commutator $[X, Y]$, the characteristic polynomial of $[X, Y]^*[X, Y]$, i.e., the monic quadratic polynomial having $s_1^2([X, Y])$ and $s_2^2([X, Y])$ as roots. To this, we first consider

$$\{x^2 - (s_1^2 + s_2^2)x + s_1^2 s_2^2 : (s_1, s_2) \in \mathcal{R}\},$$

the set of monic quadratic polynomials having s_1^2 and s_2^2 as roots when (s_1, s_2) varies over $\mathcal{R}$. To describe the set, it is equivalent to consider the set of the varying coefficients given by

$$\mathcal{Q} = \{(s_1^2 + s_2^2, s_1^2 s_2^2) : (s_1, s_2) \in \mathcal{R}\} \subset \mathbb{R}^2$$

and we have the following characterization.

PROPOSITION 2.1. *The set $\mathcal{Q}$ (see Figure 2.2) is the region bounded by the segment joining $(0, 0)$ and $(1, 0)$, the curve $x = 2\sqrt{y}$ for $0 \leq y \leq 1$, the curve $x = 1 + 2y^{1/2}$ for $0 \leq y \leq 1/16$, and the curve $x = 4y^{1/4} - 2y^{1/2}$ for $1/16 \leq y \leq 1$.*

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 1-20, January 2020.
http://repository.uwyo.edu/ela

Proof. Let $F : \mathcal{R} \to \mathbb{R}^2$ be defined by $F(s_1, s_2) = (s_1^2 + s_2^2, s_1^2 s_2^2)$ which clearly is injective. Then $\mathcal{Q} = F(\mathcal{R})$. For $0 \le \beta \le 1$, let $C_\beta = \{(s_1, s_2) : (s_1, s_2) \in \mathcal{R}, s_1^2 s_2^2 = \beta\}$. When $\beta = 0$, C_0 is the line segment joining $(0, 0)$ and $(1, 0)$; when $0 < \beta \le 1$, C_β is the intersection of $\mathcal{R}$ and the hyperbola $s_1 s_2 = \sqrt{\beta}$, see Figure 2.1. [1]

Figure 2.1. The region $\mathcal{R}$ (green) and the curve $s_1 s_2 = \sqrt{\beta}$ (blue).

Figure 2.2. The region $\mathcal{Q}$ (green) and the segment $F(C_\beta)$ (blue).

Then

$$\bigcup_{0\le\beta\le1} C_\beta = \mathcal{R}, \quad \text{and hence,} \quad \mathcal{Q} = F(\mathcal{R}) = \bigcup_{0\le\beta\le1} F(C_\beta).$$

For each β, as C_β is closed and connected, $F(C_\beta)$ is a horizontal segment in $\mathcal{Q}$ with height β above the x-axis, see Figure 2.2. When β increases from 0 to 1, the curve C_β and the segment $F(C_\beta)$ sweep over the regions $\mathcal{R}$ and $\mathcal{Q}$, respectively. By clicking Figure 2.1 or 2.2, one can see the demonstration of the movement of the corresponding C_β and $F(C_\beta)$ when β increases.

Let $F(C_\beta) = \{(x, \beta) : x \in L_\beta\}$ where $L_\beta = \{s_1^2 + s_2^2 : (s_1, s_2) \in C_\beta\}$ is a closed interval. The result follows if we can show that

$$L_\beta = \begin{cases} [2\sqrt{\beta}, 1 + 2\sqrt{\beta}] & \text{if } 0 \le \beta \le 1/16, \\ [2\sqrt{\beta}, 4\beta^{1/4} - 2\sqrt{\beta}] & \text{if } 1/16 < \beta \le 1. \end{cases} \tag{2.1}$$

It remains to determine the two endpoints of L_β, i.e., to find the maximum and minimum of L_β.

When $\beta = 0$, $L_0 = [0, 1]$ obviously. We now suppose $\beta > 0$. When β is fixed and $s_1^2 s_2^2 = \beta$, as $s_1 \ge s_2$, we see that the bigger is s_1, the bigger is $s_1^2 + s_2^2$. Hence, the minimum of $s_1^2 + s_2^2$ occurs when $s_1 = s_2 = \beta^{1/4}$, and thus, the minimum of L_β is $2\sqrt{\beta}$. Similarly, the maximum of $s_1^2 + s_2^2$ occurs at a point (s_1^*, s_2^*) which is on the right-hand boundary of the region $\mathcal{R}$, i.e., on the segment joining $(1, 0)$ and $\left(\frac{\sqrt{2}+1}{2}, \frac{\sqrt{2}-1}{2}\right)$, or on the curve (1.2).

When $0 < \beta \le 1/16$, the point (s_1^*, s_2^*) is on the line segment joining $(1, 0)$ and $\left(\frac{\sqrt{2}+1}{2}, \frac{\sqrt{2}-1}{2}\right)$, i.e., $s_1^* - s_2^* = 1$, $s_1^* \in (1, (\sqrt{2}+1)/2]$. Hence, we know that $(s_1^*)^2 + (s_2^*)^2 = (s_1^* - s_2^*)^2 + 2s_1^* s_2^* = 1 + 2\sqrt{\beta}$.

When $1/16 \le \beta \le 1$, the point (s_1^*, s_2^*) is on the curve (1.2), say with $\phi = \phi^*$. Let $\alpha = (s_1^*)^2 + (s_2^*)^2$ be

[1] A sketch of the region $\mathcal{R}$ is given in [4].

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 1-20, January 2020.
http://repository.uwyo.edu/ela

the required maximum, and write $z = \cos\phi^* \sin\phi^*$. Easily,

$$\sqrt{\alpha} = \sqrt{(s_1^*)^2 + (s_2^*)^2} = \frac{4\sqrt{\cos\phi^* \sin\phi^*}}{1 + 2\cos\phi^* \sin\phi^*} = \frac{4\sqrt{z}}{1+2z} \tag{2.2}$$

and

$$\sqrt{\beta} = s_1^* s_2^* = \frac{16\cos\phi^* \sin\phi^*}{(1 + 2\cos\phi^* \sin\phi^*)^2} \cos\phi^* \sin\phi^* = \alpha z. \tag{2.3}$$

Multiplying (2.2) by $\sqrt{\alpha}$, we get $2\alpha z - 4\sqrt{\alpha z} + \alpha = 0$ and hence, by (2.3), $\alpha = 4\beta^{1/4} - 2\sqrt{\beta}$ as required. □

3. The real case. In this section, we give a proof of Theorem 1.2.

Proof of Theorem 1.2. For $X, Y \in \Sigma_2(\mathbb{R})$, $\|[X,Y]\|^2 = s_1^2([X,Y]) + s_2^2([X,Y])$ and $(\det[X,Y])^2 = s_1^2([X,Y]) s_2^2([X,Y])$. The set of characteristic polynomials of $[X,Y]^*[X,Y]$ is

$$\{x^2 - \|[X,Y]\|^2 x + (\det[X,Y])^2 : X, Y \in \Sigma_2(\mathbb{R})\}$$

and, as before, we consider the set of varying coefficients

$$\mathcal{T}(\mathbb{R}) = \{\big(\|[X,Y]\|^2, (\det[X,Y])^2\big) : X, Y \in \Sigma_2(\mathbb{R})\} \subset \mathbb{R}^2.$$

It is then clear that $\mathcal{S}_2^{\mathbb{R}} = \mathcal{R}$ if and only if $\mathcal{T}(\mathbb{R}) = \mathcal{Q}$ (defined in Section 2), and we now show that the latter is true. We note that for $X, Y \in \Sigma_2(\mathbb{R})$, one has $0 \le |\det[X,Y]| \le 1$. To prove the result, it suffices to show that for each $0 \le \beta \le 1$,

$$\big\{\|[X,Y]\|^2 : X, Y \in \Sigma_2(\mathbb{R}), (\det[X,Y])^2 = \beta\big\} \text{ is as in the right-hand side of (2.1).} \tag{3.1}$$

The proof is divided into two parts, depending on whether the eigenvalues of $[X,Y]$ are real or not.

3.1. Eigenvalues of $[\mathbf{X},\mathbf{Y}]$ are real. Suppose the eigenvalues of $[X,Y]$ are real (and opposite), i.e., $\det[X,Y] = -\sqrt{\beta} \le 0$. Under suitable simultaneous orthogonal similarity on X and Y, we may assume

$$[X,Y] = \begin{bmatrix} \lambda & \delta \\ 0 & -\lambda \end{bmatrix}, \tag{3.2}$$

where $\lambda \ge 0$ and $\delta \ge 0$. Of course $\lambda^2 = -\det[X,Y] = \sqrt{\beta}$, and

$$\|[X,Y]\|^2 = 2\lambda^2 + \delta^2 = 2\sqrt{\beta} + \delta^2.$$

Thus, to prove (3.1), we need to find the range of δ^2. For each $0 \le \lambda \le 1$, suppose the maximum value of δ is $\delta_\lambda \ge 0$. We have to show that δ_λ^2 is as given in (1.3) (note that as $\lambda \ge 0$ here, we drop the absolute value sign in $\delta_{|\lambda|}$) and that δ can attain every value between 0 and δ_λ. The proof is divided into several steps.

Step 1. We give an alternative form of (3.2). As X and Y are of rank one, suppose

$$X = \begin{bmatrix} \cos a \\ \sin a \end{bmatrix} \begin{bmatrix} \cos b & \sin b \end{bmatrix} = \begin{bmatrix} \cos a \cos b & \cos a \sin b \\ \sin a \cos b & \sin a \sin b \end{bmatrix} \tag{3.3}$$

and

$$Y = \begin{bmatrix} \cos h \\ \sin h \end{bmatrix} \begin{bmatrix} \cos k & \sin k \end{bmatrix} = \begin{bmatrix} \cos h \cos k & \cos h \sin k \\ \sin h \cos k & \sin h \sin k \end{bmatrix},$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 1-20, January 2020.
http://repository.uwyo.edu/ela

where $a, b, h, k \in \mathbb{R}$. From (3.2), we have

$$\begin{aligned}\cos a \sin b \sin h \cos k - \cos h \sin k \sin a \cos b &= \lambda,\\ \cos a \cos b \cos h \sin k + \cos a \sin b \sin h \sin k - \cos h \cos k \cos a \sin b - \cos h \sin k \sin a \sin b &= \delta,\\ \sin a \cos b \cos h \cos k + \sin a \sin b \sin h \cos k - \sin h \cos k \cos a \cos b - \sin h \sin k \sin a \cos b &= 0.\end{aligned}$$

The first equation can be rewritten as

$$\sin(a+b)\sin(h-k) - \sin(a-b)\sin(h+k) = 2\lambda,$$

while the second and third equations can be replaced by their sum and difference given by

$$\begin{aligned}\sin(a-b)\cos(h+k) - \sin(h-k)\cos(a+b) &= \delta,\\ -\sin(a+b)\cos(h+k) + \sin(h+k)\cos(a+b) &= \delta.\end{aligned}$$

Note that $a+b$ and $a-b$ can achieve any values independently, and so do $h+k$ and $h-k$. Thus, writing

$$a+b = A, \quad a-b = B, \quad h+k = H, \quad h-k = K, \tag{3.4}$$

the above three equations become, with independent variables A, B, H and K,

$$\sin A \sin K - \sin B \sin H = 2\lambda, \tag{3.5}$$

$$\sin B \cos H - \sin K \cos A = \delta, \tag{3.6}$$

$$-\sin A \cos H + \sin H \cos A = \delta, \tag{3.7}$$

respectively.

We first show that δ can be 0. Take $A = H = \pi/2$, and B and K satisfy $\sin B = -\lambda$ and $\sin K = \lambda$. Then (3.5)–(3.7) are satisfied with $\delta = 0$.

Step 2. We further transform the problem. We now assume $\delta > 0$. Equation (3.7) gives

$$\delta = \sin(H - A). \tag{3.8}$$

Equations (3.5) and (3.6) give

$$\begin{bmatrix} -\sin H & \sin A \\ \cos H & -\cos A \end{bmatrix} \begin{bmatrix} \sin B \\ \sin K \end{bmatrix} = \begin{bmatrix} 2\lambda \\ \delta \end{bmatrix},$$

and hence, with (3.8) and using Cramer's rule, we obtain

$$\sin B = \frac{-2\lambda \cos A - \delta \sin A}{\sin H \cos A - \sin A \cos H} = -\frac{2\lambda}{\delta}\cos A - \sin A$$

and

$$\sin K = \frac{-2\lambda \cos H - \delta \sin H}{\sin H \cos A - \sin A \cos H} = -\frac{2\lambda}{\delta}\cos H - \sin H.$$

Thus, equivalently, we need to find the range of δ subject to (3.8),

$$\left|\frac{2\lambda}{\delta}\cos A + \sin A\right| \le 1 \quad \text{and} \quad \left|\frac{2\lambda}{\delta}\cos H + \sin H\right| \le 1. \tag{3.9}$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 1-20, January 2020.
http://repository.uwyo.edu/ela

Step 3. Suppose $0 \le 2\lambda \le 1$ (i.e., $0 \le \beta \le 1/16$). For any $0 < \delta \le 1$, choose $H = \pi/2$ and A such that $\cos A = \delta$ and $\sin A = -\sqrt{1-\delta^2}$. Then $\sin(H - A) = \cos A = \delta$ and both inequalities in (3.9) are satisfied. Hence, δ can assume any value in $[0, 1]$ as required.

Step 4. We suppose $1 \le 2\lambda \le 2$ (i.e., $1/16 \le \beta \le 1$) and find the maximum value of δ. Geometrically, (3.9) means that the inner products of the vector $(2\lambda/\delta, 1)^T$ with the two unit vectors $(\cos A, \sin A)^T$ and $(\cos H, \sin H)^T$ have absolute values not bigger than one.

Suppose the maximum value of δ is $\delta_\lambda = \sin(H_0 - A_0) > 0$ where $0 < H_0 - A_0 < \pi$. If $\sin(H_0 - A_0) = 1$, $\{(\cos A_0, \sin A_0)^T, (\cos H_0, \sin H_0)^T\}$ is an orthonormal basis of $\mathbb{R}^2$. Then, (3.9) implies $\|(2\lambda/\delta_\lambda, 1)^T\| \le \sqrt{2}$. This gives a contradiction as $2\lambda/\delta_\lambda > 1$. So $\sin(H_0 - A_0) < 1$. We claim that for $\delta = \delta_\lambda$, both inequalities in (3.9) must hold in equality. If both of them are strict inequalities, we can purturb H_0 and A_0 a bit to have a bigger value of δ without violating (3.9), and this gives a contradiction. If exactly one of them is equality, we may consider replacing H_0 and A_0 by $H_0 + \epsilon$ and $A_0 + \epsilon$ for small suitable ϵ, resulting in both of them are strict inequalities and with $\delta = \sin((H_0 + \epsilon) - (A_0 + \epsilon)) = \delta_\lambda$. Thus, as in the previous case, we have a contradiction.

Now suppose both inequalities in (3.9) hold in equality. Geometrically, it is clear that there are 4 unit vectors $\mathbf{x} \in \mathbb{R}^2$ such that $|\langle \mathbf{x}, (2\lambda/\delta_\lambda, 1)^T \rangle| = 1$, namely, $\mathbf{u} = (0, 1)^T$, $\mathbf{v}$ and their negatives, where $\mathbf{v} = (\cos\theta, \sin\theta)^T$, $-\pi/2 < \theta < 0$, is the reflection of $\mathbf{u}$ across the vector $(2\lambda/\delta_\lambda, 1)^T$. See Figure 3.1 below.

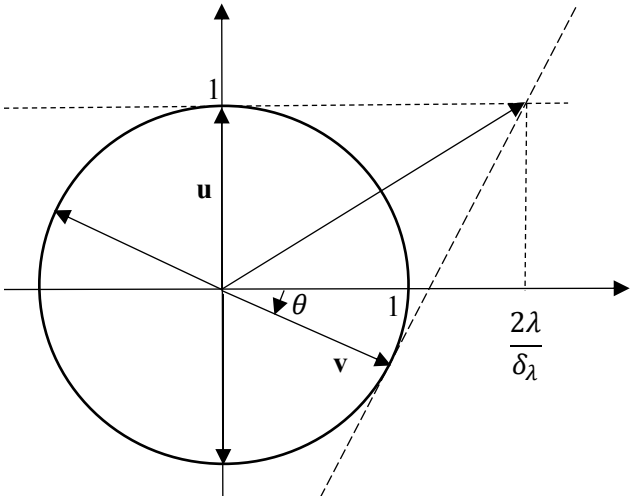


Figure 3.1. The vectors $\mathbf{u}$ and $\mathbf{v}$.

In other words, when restricting $-\pi < H_0, A_0 \le \pi$, we have $H_0, A_0 \in \{\pm\pi/2, \theta, \theta+\pi\}$. Since $\sin(H_0 - A_0) > 0$, the possible choices for (H_0, A_0) are $(\pi/2, \theta)$, $(\theta, -\pi/2)$, $(-\pi/2, \theta + \pi)$ and $(\theta + \pi, \pi/2)$.

We may take $(H_0, A_0) = (\pi/2, \theta)$. The other choices of (H_0, A_0) will always lead to this case. For example, if $(H_0, A_0) = (\theta, -\pi/2)$, (3.9) becomes

$$\left| \frac{2\lambda}{\sin(\theta - (-\pi/2))} \cos\left(-\frac{\pi}{2}\right) + \sin\left(-\frac{\pi}{2}\right) \right| = 1 \quad \text{and} \quad \left| \frac{2\lambda}{\sin(\theta - (-\pi/2))} \cos\theta + \sin\theta \right| = 1,$$

which is equivalent to

$$\left| \frac{2\lambda}{\sin(\pi/2 - \theta)} \cos\left(\frac{\pi}{2}\right) + \sin\left(\frac{\pi}{2}\right) \right| = 1 \quad \text{and} \quad \left| \frac{2\lambda}{\sin(\pi/2 - \theta)} \cos\theta + \sin\theta \right| = 1,$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 1-20, January 2020.
http://repository.uwyo.edu/ela

and these two new conditions exactly mean taking $(H_0, A_0) = (\pi/2, \theta)$.

So, fix now $(H_0, A_0) = (\pi/2, \theta)$. It is easy to check that the triangle with vertices $(0,0)$, $(0,1)$ and $(2\lambda/\delta_\lambda, 1)$ and the triangle with vertices $(0,0)$, $(\cos\theta, \sin\theta)$ and $(2\lambda/\delta_\lambda, 1)$ are congruent (see Figure 3.1). Consequently, in the triangle with vertices $(0,0)$, $(0,1)$ and $(2\lambda/\delta_\lambda, 1)$, the angle at $(0,0)$ is $\frac{\pi/2-\theta}{2}$ (remember $\theta < 0$). Hence,

$$\sqrt{1+\left(\frac{2\lambda}{\delta_\lambda}\right)^2}\cos\left(\frac{\pi/2-\theta}{2}\right) = 1,$$

which gives, with $\delta_\lambda = \sin(\pi/2-\theta)$,

$$\left(1+\frac{\lambda^2}{\sin^2((\pi/2-\theta)/2)\cos^2((\pi/2-\theta)/2)}\right)\cos^2\left(\frac{\pi/2-\theta}{2}\right) = 1.$$

Thus,

$$\lambda = \sin^2\left(\frac{\pi/2-\theta}{2}\right),$$

and hence,

$$\delta_\lambda^2 = \sin^2(\pi/2-\theta) = 4\sin^2\left(\frac{\pi/2-\theta}{2}\right)\cos^2\left(\frac{\pi/2-\theta}{2}\right) = 4\lambda(1-\lambda).$$

Step 5. Finally, we show that any value between 0 and δ_λ can be achieved by δ. For any $0 < \delta < \delta_\lambda$, take $H = \pi/2$ and A such that

$$(\cos A, \sin A) = \left(\delta, -\sqrt{1-\delta^2}\right). \tag{3.10}$$

Then $\delta = \cos A = \sin(H-A)$ and the second part of (3.9) is satisfied. It remains to show that the first part of (3.9) is also satisfied. With (3.10), it suffices to show $\left|2\lambda - \sqrt{1-\delta^2}\right| \le 1$ for all δ where $0 < \delta < \delta_\lambda$. Note that

$$\left|2\lambda - \sqrt{1-\delta^2}\right| \le 1 \quad \Leftrightarrow \quad 4\lambda^2 - \delta^2 \le 4\lambda\sqrt{1-\delta^2}.$$

If $4\lambda^2 - \delta^2 \le 0$, we are done. Now suppose $4\lambda^2 - \delta^2 > 0$. Then

$$4\lambda^2 - \delta^2 \le 4\lambda\sqrt{1-\delta^2} \quad \Leftrightarrow \quad 16\lambda^4 + 8\lambda^2\delta^2 + \delta^4 - 16\lambda^2 \le 0.$$

Since $0 < \delta < \delta_\lambda$, it suffices to show that $16\lambda^4 + 8\lambda^2\delta_\lambda^2 + \delta_\lambda^4 - 16\lambda^2 \le 0$. With $\delta_\lambda^2 = 4\lambda - 4\lambda^2$, the result follows from $16\lambda^4 + 8\lambda^2\delta_\lambda^2 + \delta_\lambda^4 - 16\lambda^2 = 0$.

3.2. Eigenvalues of $[\mathbf{X}, \mathbf{Y}]$ are purely imaginary. We now suppose the two eigenvalues of $[X, Y]$ are purely imaginary, i.e., $\det[X, Y] = \sqrt{\beta} > 0$. We claim that

$$\sqrt{\beta} \le 1/4. \tag{3.11}$$

We don't have the upper triangular form as in (3.2). Under suitable simultaneous orthogonal similarity on X and Y we may assume

$$X = \begin{bmatrix} p & q \\ 0 & 0 \end{bmatrix} \quad \text{where } p^2+q^2 = 1, \quad Y = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix}.$$

Then

$$[X, Y] = \begin{bmatrix} qy_{21} & py_{12} + q(y_{22}-y_{11}) \\ -py_{21} & -qy_{21} \end{bmatrix}.$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 1-20, January 2020.
http://repository.uwyo.edu/ela

Taking the determinant on both sides,

$$\sqrt{\beta} = -q^2 y_{21}^2 + p y_{21}[p y_{12} + q(y_{22} - y_{11})]. \tag{3.12}$$

As $y_{11}y_{22} = y_{21}y_{12}$ (i.e., $\det Y = 0$), we get

$$(q y_{21})^2 - p(y_{22} - y_{11})(q y_{21}) + (\sqrt{\beta} - p^2 y_{11} y_{22}) = 0.$$

Regarding this as a quadratic equation in qy_{21} with real coefficients, it has (one and hence) two real roots. Its discriminant must be non-negative, i.e.,

$$0 \le [p(y_{22} - y_{11})]^2 - 4(\sqrt{\beta} - p^2 y_{11} y_{22}) = [p(y_{22} + y_{11})]^2 - 4\sqrt{\beta}.$$

Thus, $\sqrt{\beta} \le p^2(y_{22} + y_{11})^2/4 \le 1/4$ as claimed.

To complete the proof, as $0 < \beta \le 1/16$, it suffices to show that $\|[X, Y]\|^2 \le 1 + 2\sqrt{\beta}$. Note that, using (3.12),

$$\begin{aligned}
&\|[X, Y]\|^2 \le 1 + 2\sqrt{\beta} \\
\Leftrightarrow\ & 2q^2 y_{21}^2 + 2\sqrt{\beta} + p^2 y_{21}^2 + [p y_{12} + q(y_{22} - y_{11})]^2 \le 1 + 2\sqrt{\beta} + 2\sqrt{\beta} \\
\Leftrightarrow\ & 2\{p y_{21}[p y_{12} + q(y_{22} - y_{11})]\} + p^2 y_{21}^2 + [p y_{12} + q(y_{22} - y_{11})]^2 \le 1 + 4\sqrt{\beta} \\
\Leftrightarrow\ & \{p y_{21} + [p y_{12} + q(y_{22} - y_{11})]\}^2 \le 1 + 4\sqrt{\beta}.
\end{aligned}$$

That is, subject to (3.12), we have to show that

$$[p(y_{21} + y_{12}) + q(y_{22} - y_{11})]^2 \le 1 + 4\sqrt{\beta}.$$

From $y_{11}^2 + y_{22}^2 + y_{12}^2 + y_{21}^2 = 1$ and $y_{11}y_{22} - y_{12}y_{21} = 0$, we get $(y_{21} + y_{12})^2 + (y_{22} - y_{11})^2 = 1$. The result is now clear as both $(y_{21} + y_{12}, y_{22} - y_{11})^T$ and $(p, q)^T$ are unit vectors. ∎

4. The complex case.

4.1. Complex vs. real. There are fundamental differences between the real and complex problems and we tried in vain to modify the proof of Theorem 1.2 to prove the complex case. As an illustration, suppose

$$X = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \begin{bmatrix} \bar{b}_1 & \bar{b}_2 \end{bmatrix} = \begin{bmatrix} a_1\bar{b}_1 & a_1\bar{b}_2 \\ a_2\bar{b}_1 & a_2\bar{b}_2 \end{bmatrix}$$

where $(a_1, a_2)^T$ and $(b_1, b_2)^T$ are unit vectors in $\mathbb{C}^2$. When the two vectors are real, in the proof of Theorem 1.2, we have

$$a_1\bar{b}_1 - a_2\bar{b}_2 = \cos a \cos b - \sin a \sin b = \cos(a + b) = \cos A$$

and

$$a_2\bar{b}_1 + a_1\bar{b}_2 = \sin a \cos b + \cos a \sin b = \sin(a + b) = \sin A.$$

In the real case, $|\cos A| \le 1$, $|\sin A| \le 1$ and $\|(\cos A, \sin A)^T\| = 1$. In the complex case, though we have

$$|a_1\bar{b}_1 - a_2\bar{b}_2| \le 1 \quad \text{and} \quad |a_2\bar{b}_1 + a_1\bar{b}_2| \le 1,$$

the norm of $(a_1\bar{b}_1 - a_2\bar{b}_2, a_2\bar{b}_1 + a_1\bar{b}_2)^T$ ranges from 0 to 2. For example, the matrices $\frac{1}{2}\begin{bmatrix} 1 & \mathbf{i} \\ -\mathbf{i} & 1 \end{bmatrix}$ and $\frac{1}{2}\begin{bmatrix} 1 & \mathbf{i} \\ \mathbf{i} & -1 \end{bmatrix}$ give the norms of the corresponding vectors 0 and 2, respectively. Consequently, there are several places in the proof of Theorem 1.2 where the geometric argument cannot be adopted directly to prove the complex problem.

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 1-20, January 2020.
http://repository.uwyo.edu/ela

4.2. Some lemmas. When $[X, Y]$ is not in triangular form, we may use $\|[X, Y]\|^2 - 2|\det[X, Y]|$ to represent δ^2 in our formulation. The following proposition tells us that if we can reduce one of the matrices X and Y to have zero trace then we are done.

PROPOSITION 4.1. *For $0 \le |\lambda| \le 1$ and $\delta^2_{|\lambda|}$ as given in (1.3),*

$$\max\{\|[X, Y]\|^2 - 2|\det[X, Y]| : X, Y \in \Sigma_2(\mathbb{C}), |\det[X, Y]| = |\lambda|^2, \operatorname{tr} X = 0\} \le \delta^2_{|\lambda|}.$$

Proof. Under suitable unitary similarity, we may assume $X = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$. With $Y = (y_{ij})$,

$$XY - YX = \begin{bmatrix} y_{21} & y_{22} - y_{11} \\ 0 & -y_{21} \end{bmatrix}$$

and so $|y_{21}| = \sqrt{|\det[X, Y]|} = |\lambda|$. Hence,

$$\|[X, Y]\|^2 = 2|\lambda|^2 + |y_{22} - y_{11}|^2 \le 2|\lambda|^2 + (|y_{22}| + |y_{11}|)^2 \tag{4.1}$$

$$\le 2|\lambda|^2 + (s_1(Y) + s_2(Y))^2 \tag{4.2}$$

$$= 2|\lambda|^2 + 1.$$

Inequality (4.2) follows from the relation between the singular values and diagonal elements of a matrix, e.g. see [5, (3.1.10a)]. Consequently, for $0 \le |\lambda| \le 1/2$, we get as desired the maximum to be bounded by $\delta^2_{|\lambda|} = 1$.

When $1/2 < |\lambda| \le 1$, we can have a smaller upper bound for $(|y_{22}| + |y_{11}|)^2$ instead of 1. The conditions $|y_{11}|^2 + |y_{22}|^2 + |\lambda|^2 + |y_{12}|^2 = 1$ (i.e., $\|Y\|^2 = 1$) and $|y_{11}||y_{22}| = |\lambda||y_{12}|$ (i.e., $\det Y = 0$) give

$$(|y_{11}| + |y_{22}|)^2 + (|\lambda| - |y_{12}|)^2 = 1.$$

Replacing $|y_{12}|$ by $|y_{11}||y_{22}|/|\lambda|$, and using $|y_{11}||y_{22}| \le \left(\frac{|y_{11}|+|y_{22}|}{2}\right)^2 \le 1/4 < |\lambda|^2$, we get

$$(|y_{11}| + |y_{22}|)^2 + \frac{\left(|\lambda|^2 - \left(\frac{|y_{11}|+|y_{22}|}{2}\right)^2\right)^2}{|\lambda|^2} \le 1$$

which, by direct calculation, gives

$$\left(|\lambda| + \frac{\left(\frac{|y_{11}|+|y_{22}|}{2}\right)^2}{|\lambda|}\right)^2 \le 1.$$

Consequently, after taking square root on both sides, we easily get

$$(|y_{11}| + |y_{22}|)^2 \le 4|\lambda| - 4|\lambda|^2.$$

Thus, from (4.1), the result follows. ∎

In the following lemma, we modify the proof of Theorem 1.2 to handle a particular case of the complex problem.

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 1-20, January 2020.
http://repository.uwyo.edu/ela

LEMMA 4.2. *Suppose* $X \in \Sigma_2(\mathbb{R})$ *and*

$$Y = \begin{bmatrix} c+d\mathbf{i} & y_{12} \\ y_{21} & -c+d\mathbf{i} \end{bmatrix} \in \Sigma_2(\mathbb{C}), \quad c, d, y_{12}, y_{21} \in \mathbb{R},\ d \neq 0,$$

such that $XY - YX = \begin{bmatrix} \lambda & \delta \\ 0 & -\lambda \end{bmatrix}$. *Then, with* $\delta_{|\lambda|}$ *as given in (1.3),*

(i) $|\delta| \leq \delta_{|\lambda|}$ *; or*

(ii) there exist $\tilde{X}, \tilde{Y} \in \Sigma_2(\mathbb{C})$ *such that* $\tilde{X}\tilde{Y} - \tilde{Y}\tilde{X} = \begin{bmatrix} \lambda & \tilde{\delta} \\ 0 & -\lambda \end{bmatrix}$ *with* $|\tilde{\delta}| > |\delta|$.

Proof. We remark that $\|Y\| = 1$ and $\det Y = 0$ read $2(c^2+d^2)+y_{12}^2+y_{21}^2 = 1$ and $-(c^2+d^2)-y_{12}y_{21} = 0$, respectively. So, $\begin{bmatrix} \sqrt{c^2+d^2} & y_{12} \\ y_{21} & -\sqrt{c^2+d^2} \end{bmatrix} \in \Sigma_2(\mathbb{R})$. Let

$$\begin{bmatrix} \sqrt{c^2+d^2} & y_{12} \\ y_{21} & -\sqrt{c^2+d^2} \end{bmatrix} = \begin{bmatrix} \cos h \\ \sin h \end{bmatrix} \begin{bmatrix} \cos k & \sin k \end{bmatrix} = \begin{bmatrix} \cos h \cos k & \cos h \sin k \\ \sin h \cos k & \sin h \sin k \end{bmatrix}.$$

The matrix on the left has zero trace and so the condition

$$0 = \cos h \cos k + \sin h \sin k = \cos(h-k)$$

grants $h - k \in \{\pi/2 + l\pi : l \text{ is an integer}\}$. Set $t = c/\sqrt{c^2+d^2}$, so that we can rewrite Y as

$$Y = \begin{bmatrix} t\cos h \cos k & \cos h \sin k \\ \sin h \cos k & t \sin h \sin k \end{bmatrix} + dI_2\mathbf{i}, \quad -1 < t < 1. \tag{4.3}$$

Suppose X is as in (3.3). We divide the proof into several steps.

Step 1. Parallel to Step 1 in Section 3.1, by replacing the terms $\cos h \cos k$ and $\sin h \sin k$ there by $t \cos h \cos k$ and $t \sin h \sin k$, we obtain (parallel to (3.5)–(3.7))

$$\sin K \sin A - \sin B \sin H = 2\lambda, \tag{4.4}$$

$$t \sin B \cos H - \sin K \cos A = \delta, \tag{4.5}$$

$$-t \sin A \cos H + \sin H \cos A = \delta, \tag{4.6}$$

where A, B, H and K (defined in (3.4)) are independent variables with $K \in \{\pi/2 + l\pi : l \text{ is an integer}\}$. From (4.6), in which the left-hand side can be regarded as the inner product of $(-\sin A, \cos A)^T$ and $(t\cos H, \sin H)^T$, we know that $|\delta| \leq 1$. Thus, we have (i) if $|\lambda| \leq 1/2$.

Step 2. Suppose $|\lambda| > 1/2$. Following the calculation in Step 2 in Section 3.1, we see that the solvability of (4.4)–(4.6) is equivalent to the solvability of (4.6),

$$\left| \frac{2\lambda}{\delta} \cos A + \sin A \right| = |\sin B| \leq 1 \quad \text{and} \quad \left| \frac{2\lambda}{\delta} t \cos H + \sin H \right| = |\sin K| = 1. \tag{4.7}$$

Note that as B and K are independent of the other variables, we may focus on t, δ A and H. If we want to show that there are matrices satisfying the assertion in (ii), it suffices to show that there are t_1, δ_1, A_1 and H_1 such that, with the terms $|\sin B|$ and $|\sin K|$ dropped, (4.6) and (4.7) are satisfied and $|\delta_1| > |\delta|$. The values of B and K can then be chosen suitably.

We use a perturbation argument, assuming that there are t, δ, A and H satisfying (4.6) and (4.7). Let us outline our steps first.

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 1-20, January 2020.
http://repository.uwyo.edu/ela

Step 2.1. Perturb t in (4.6) to t_1 to have δ_1 such that $|\delta_1| > |\delta|$. With the values t_1, δ_1 and H, the second part of (4.7) will probably be violated.
Step 2.2. Adjust H to H_1 so that t_1, δ_1 and H_1 satisfy the second part of (4.7). With the values t_1, δ_1 and H_1, (4.6) will probably be violated.
Step 2.3. Adjust A to A_1 so that t_1, δ_1, H_1 and A_1 satisfy (4.6) .

During the steps, we also have to ensure that the first part of (4.7) is always satisfied. Before we carry out our plan, we first note the following points.

Point 1. We now eliminate the situation that $\sin A \cos H = 0$, so that we can perturb t in (4.6) to have a bigger value of $|\delta|$. If $\sin A = 0$, we have $-\sin B \sin H = 2\lambda$ from (4.4) and this contradicts $2\lambda > 1$. If $\cos H = 0$, then (4.5) and (4.6) are independent of t. Take $t = 1$ in (4.3) to have

$$\hat{Y} = \begin{bmatrix} \cos h \cos k & \cos h \sin k \\ \sin h \cos k & \sin h \sin k \end{bmatrix} \in \Sigma_2(\mathbb{R}).$$

Readily, the pair $X, \hat{Y} \in \Sigma_2(\mathbb{R})$ satisfies (4.4)–(4.6). Thus, $|\delta| \le \delta_{|\lambda|}$ and we are done.

Point 2. We refer to the first part of (4.7). If equality holds, then $|\sin B| = 1$, and hence, $\operatorname{tr} X = \cos B = 0$. By Proposition 4.1, we have (i) and we are done. We now assume

$$\left| \frac{2\lambda}{\delta} \cos A + \sin A \right| < 1.$$

With this assumption, we know that the first part of (4.7) will not be violated if we perturb t, δ, H and A small enough. This ensures the first part of (4.7) will be satisfied throughout the perturbations.

Point 3. We show that

$$|\delta| < \sqrt{t^2 \cos H^2 + \sin^2 H}. \tag{4.8}$$

The main purpose of showing this is to guarantee that after we perturb t, δ and H to t_1, δ_1 and H_1, respectively, we still have

$$|\delta_1| < \sqrt{t_1^2 \cos H_1^2 + \sin^2 H_1}. \tag{4.9}$$

Consequently, we can perturb A to A_1 as required in Step 2.3. (Note: Geometrically, the left-hand side of (4.6) is the inner product of the vectors $(t \cos H, \sin H)^T$ and $(-\sin A, \cos A)^T$. To have A_1 in Step 2.3, we need $|\delta_1| < \|(t_1 \cos H_1, \sin H_1)^T\|$.)

We now prove (4.8). By the Cauchy-Schwarz inequality, we know from (4.6) that (4.8) is true when "$\le$" is written. If equality holds, then the two vectors $(-\sin A, \cos A)^T$ and $(t \cos H, \sin H)^T$ are linearly dependent and $\|(t \cos H, \sin H)^T\| = |\delta|$. So, we can rewrite the second part of (4.7) as

$$\left| \frac{2\lambda}{\delta}(-\sin A) + \cos A \right| = \frac{1}{|\delta|}. \tag{4.10}$$

The two vectors $(-\sin A, \cos A)^T$ and $(\cos A, \sin A)^T$ form an orthonormal basis of $\mathbb{R}^2$. Using (4.10) and the first part of (4.7) we get

$$\left(\frac{2\lambda}{|\delta|}\right)^2 + 1 = \left\| \left(\frac{2\lambda}{\delta}, 1\right)^T \right\|^2 = \left| \frac{2\lambda}{\delta}(-\sin A) + \cos A \right|^2 + \left| \frac{2\lambda}{\delta} \cos A + \sin A \right|^2 \le \frac{1}{|\delta|^2} + 1,$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 1-20, January 2020.
http://repository.uwyo.edu/ela

and this contradicts the assumption $2\lambda > 1$. Hence, we have (4.8).

We are ready to carry out the Steps 2.1–2.3.

- For Step 2.1, by Point 1, we may assume $\sin A \cos H \neq 0$. By a small perturbation of t to t_1 in (4.6), we get
$$-t_1 \sin A \cos H + \sin H \cos A = \delta_1, \qquad \text{where } |\delta_1| > |\delta|.$$
- For Step 2.2, with t_1 and δ_1 obtained, we adjust H to H_1 (with $|H - H_1|$ small) so that the second part of (4.7) is satisfied with t_1, δ_1 and H_1. This is possible because of the second part of (4.7), $\left\| \left(\frac{2\lambda t_1}{\delta_1}, 1\right)^T \right\| > 1$, and that t_1 and δ_1 are small perturbations of t and δ, respectively.
- For Step 2.3, with (4.9), we can adjust A suitably to A_1 (again with $|A_1 - A|$ small) so that t_1, δ_1, H_1 and A_1 satisfy (4.6).

Summing up, with reference to Point 2, we have found t_1, δ_1, H_1 and A_1 such that (4.6) and (4.7) are satisfied and $|\delta_1| > |\delta|$. Assertion (ii) follows. ∎

4.3. The main proof. We now give the proof of Theorem 1.3.

Proof of Theorem 1.3. Similar to $\mathcal{T}(\mathbb{R})$ in the proof of Theorem 1.2, let

$$\mathcal{T}(\mathbb{C}) = \{(\|[X,Y]\|^2, |\det[X,Y]|^2) : X, Y \in \Sigma_2(\mathbb{C})\} \subset \mathbb{R}^2.$$

To prove the theorem, it suffices to show that $\mathcal{T}(\mathbb{C}) = \mathcal{Q}$. As $\mathcal{Q} = \mathcal{T}(\mathbb{R}) \subseteq \mathcal{T}(\mathbb{C})$, it suffices to consider the right boundary of $\mathcal{T}(\mathbb{C})$ and show that

$$\max\left\{\|[X,Y]\|^2 : X, Y \in \Sigma_2(\mathbb{C}), |\det[X,Y]|^2 = \beta\right\} = \max\left\{\|[X,Y]\|^2 : X, Y \in \Sigma_2(\mathbb{R}), |\det[X,Y]|^2 = \beta\right\}.$$

The left boundary (which corresponds to diagonal $[X,Y]$) and the bottom boundary of $\mathcal{T}(\mathbb{R})$ and $\mathcal{T}(\mathbb{C})$ are obviously the same.

4.3.1. A transformation of the problem. Suppose

$$X = \tilde{X} + \frac{\operatorname{tr} X}{2} I_2$$

in which $\operatorname{tr} \tilde{X} = 0$, and similarily for $\tilde{Y}$. Then $XY - YX = \tilde{X}\tilde{Y} - \tilde{Y}\tilde{X}$ which allows us to work with zero trace matrices. As X is of rank one, its non-trivial eigenvalue is $\operatorname{tr} X$. On the other hand, suppose the eigenvalues of $\tilde{X}$ (which has zero trace) are $\pm\mu$. Then $\tilde{X} + \frac{\operatorname{tr} X}{2} I_2$ is of rank one if and only if $\frac{1}{2}\operatorname{tr} X = \pm\mu$. When $\mathbb{F} = \mathbb{R}$, the latter is possible only if $\tilde{X}$ has real eigenvalues, equivalently, $\det \tilde{X} \leq 0$. Note that

$$\|X\|^2 = \|\tilde{X}\|^2 + 2\left|\frac{1}{2}\operatorname{tr} X\right|^2 = \|\tilde{X}\|^2 + 2|\mu|^2 = \|\tilde{X}\|^2 + 2|\det \tilde{X}|.$$

Thus, instead of matrices from $\Sigma_2(\mathbb{F})$, we may assume, if $\mathbb{F} = \mathbb{R}$, the matrices are chosen from

$$\Phi(\mathbb{R}) = \{H : H \in M_2(\mathbb{R}), \operatorname{tr} H = 0, \|H\|^2 + 2|\det H| = 1, \det H \leq 0\}$$

that and, if $\mathbb{F} = \mathbb{C}$, the matrices are chosen from

$$\Phi(\mathbb{C}) = \{H : H \in M_2(\mathbb{C}), \operatorname{tr} H = 0, \|H\|^2 + 2|\det H| = 1\}.$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 1-20, January 2020.
http://repository.uwyo.edu/ela

We also note that

$$\|H\|^2 + 2|\det H| = s_1^2(H) + s_2^2(H) + 2s_1(H)s_2(H) = \big(s_1(H) + s_2(H)\big)^2 = \|H\|_1^2.$$

The condition $\|H\|^2 + 2|\det H| = 1$ in the definitions of $\Phi(\mathbb{F})$ above is equivalent to $\|H\|_1 = 1$.

We now work with matrices in $\Phi(\mathbb{F})$. We see from the proof of Theorem 1.2 that the region $\mathcal{T}(\mathbb{R})$ (i.e., $\mathcal{Q}$) can be fully filled by commutators that are orthogonally upper triangularizable. Thus, under simultaneous unitary (orthogonal if $\mathbb{F} = \mathbb{R}$) similarity, we may assume

$$H = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & -h_{11} \end{bmatrix},\ K = \begin{bmatrix} k_{11} & k_{12} \\ k_{21} & -k_{11} \end{bmatrix} \in \Phi(\mathbb{F})$$

are chosen such that

$$\begin{bmatrix} \lambda & \delta \\ 0 & -\lambda \end{bmatrix} = HK - KH = \begin{bmatrix} h_{12}k_{21} - k_{12}h_{21} & 2h_{11}k_{12} - 2h_{12}k_{11} \\ 2h_{21}k_{11} - 2h_{11}k_{21} & h_{21}k_{12} - k_{21}h_{12} \end{bmatrix}. \tag{4.11}$$

Though we may assume $\lambda, \delta \geq 0$ under diagonal unitary (orthogonal if $\mathbb{F} = \mathbb{R}$) similarity and multiplication with a unit scalar, we do not do so here. Such actions will be used later.

Without assuming $\delta, \lambda \geq 0$, we need to amend our problem. Our original formulation has $|\det([X, Y])|^2 = \beta$ with β being fixed and so $|\lambda|$ is fixed, and we need to determine the maximum of $|\delta|$. Thus, referring to (4.11), the equivalent problem is to find, for $0 \leq |\lambda| \leq 1$,

$$\delta_{|\lambda|}^{\mathbb{F}} = \max\{2|h_{11}k_{12} - k_{11}h_{12}| : |h_{12}k_{21} - k_{12}h_{21}| = |\lambda|, h_{11}k_{21} - k_{11}h_{21} = 0, H, K \in \Phi(\mathbb{F})\}. \tag{4.12}$$

The value of $\delta_{|\lambda|}^{\mathbb{R}}$ is exactly the $\delta_{|\lambda|}$ as given in (1.3). Here, we need to prove $\delta_{|\lambda|}^{\mathbb{R}} = \delta_{|\lambda|}^{\mathbb{C}}$.

For $\mathbb{F} = \mathbb{R}$, $\mathbb{C}$ and $0 \leq |\lambda| \leq 1$, we consider the following problem which has the constraint $|h_{12}k_{21} - k_{12}h_{21}| = |\lambda|$ in (4.12) relaxed:

$$\begin{aligned} \max \quad & F(H, K) = 2|h_{11}k_{12} - k_{11}h_{12}| \\ \text{subject to} \quad & |h_{12}k_{21} - k_{12}h_{21}| \geq |\lambda| && (4.13) \\ & h_{11}k_{21} - k_{11}h_{21} = 0 && (4.14) \\ & H, K \in \Phi(\mathbb{F}). && (4.15) \end{aligned}$$

Let us denote the maximum value of the above problem by $\Delta_{|\lambda|}^{\mathbb{F}}$. Obviously we have

$$\delta_{|\lambda|}^{\mathbb{R}} \leq \left\{ \begin{array}{c} \Delta_{|\lambda|}^{\mathbb{R}} \\ \\ \delta_{|\lambda|}^{\mathbb{C}} \end{array} \right\} \leq \Delta_{|\lambda|}^{\mathbb{C}}.$$

It is easy to see that $\Delta_{|\lambda|}^{\mathbb{F}} = \max\{\delta_t^{\mathbb{F}} : |\lambda| \leq t \leq 1\}$. From (1.3), we know that $\delta_{|\lambda|}^{\mathbb{R}}$ is non-increasing in $|\lambda|$ (see Figure 1.1 for $\big(\delta_{|\lambda|}^{\mathbb{R}}\big)^2$), and so

$$\delta_{|\lambda|}^{\mathbb{R}} = \Delta_{|\lambda|}^{\mathbb{R}}.$$

Hence, if we can show $\Delta_{|\lambda|}^{\mathbb{R}} = \Delta_{|\lambda|}^{\mathbb{C}}$, we get $\delta_{|\lambda|}^{\mathbb{R}} = \delta_{|\lambda|}^{\mathbb{C}}$ as required.

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 1-20, January 2020.
http://repository.uwyo.edu/ela

4.3.2. Proof of $\mathbf{\Delta}^{\mathbb{R}}_{|\lambda|} = \mathbf{\Delta}^{\mathbb{C}}_{|\lambda|}$. We now regard $\mathbb{F} = \mathbb{C}$. Suppose the maximum is attained with matrices H and K, i.e.,

$$0 < \Delta^{\mathbb{C}}_{|\lambda|} = F(H, K) = 2|h_{11}k_{12} - k_{11}h_{12}|.$$

We will show that, under different assumptions, either $\Delta^{\mathbb{C}}_{|\lambda|} \le \delta^{\mathbb{R}}_{|\lambda|}$ or else there is a contradiction.

Step 1. Firstly, we handle the situations that H or K has a zero entry. Note that if $|h_{12}| = 1$, then $H \in \Sigma_2(\mathbb{C})$ and has zero trace. By Proposition 4.1, we get $\Delta^{\mathbb{C}}_{|\lambda|} \le \delta^{\mathbb{R}}_{|\lambda|}$ and we are done. The same is true for K. We have the following three situations:

(I) $h_{11} = 0$. Then (4.14) implies $k_{11}h_{21} = 0$. If $k_{11} = 0$ then $\Delta^{\mathbb{C}}_{|\lambda|} = 0$ and we have a contradiction. If $h_{21} = 0$ then $|h_{12}| = 1$ and we are done.

(II) $h_{21} = 0$. Then $h_{11}k_{21} = 0$ by (4.14). If $h_{11} = 0$ we are back to (I). If $k_{21} = 0$ then $\lambda = 0$ by (4.13). Moreover, $\|H\|_1 = \|K\|_1 = 1$ implies $(2h_{11}, -h_{12})^T$ and $(\overline{k_{12}}, 2\overline{k_{11}})^T$ are unit vectors. Thus, $\Delta^{\mathbb{C}}_0 \le 1 = \delta^{\mathbb{R}}_0$ and we are done.

(III) $h_{12} = 0$. If $k_{12} = 0$ then $\Delta^{\mathbb{C}}_{|\lambda|} = 0$ and hence a contradiction. If $|k_{12}| = 1$, again, we are done. Suppose $0 < |k_{12}| < 1$. Then

$$|h_{21}| \ge |\lambda|/|k_{12}| \quad \text{and} \quad 2|h_{11}| = \Delta^{\mathbb{C}}_{|\lambda|}/|k_{12}|.$$

With $4|h_{11}|^2 + |h_{21}|^2 = \|H\|_1^2 = 1$, we get back to $\left(\Delta^{\mathbb{C}}_{|\lambda|}\right)^2 + |\lambda|^2 \le |k_{12}|^2 < 1$, which implies (as $1 + |\lambda| < 4|\lambda|$ for $|\lambda| > 1/3$)

$$\left(\Delta^{\mathbb{C}}_{|\lambda|}\right)^2 < 1 - |\lambda|^2 \le \left\{ \begin{array}{ll} 1 & \text{if } \ 0 \le |\lambda| \le 1/2 \\ 4|\lambda|(1-|\lambda|) & \text{if } \ 1/2 < |\lambda| \le 1 \end{array} \right\} = \left(\delta^{\mathbb{R}}_{|\lambda|}\right)^2,$$

which gives a contradiction.

Step 2. We derive some necessary conditions on H and K. From now on, we can assume all the entries of H and K nonzero. If $\det H = 0$, then $H \in \Sigma_2(\mathbb{C})$ with zero trace. By Proposition 4.1, we have $\Delta^{\mathbb{C}}_{|\lambda|} \le \delta^{\mathbb{R}}_{|\lambda|}$ and we are done. The same is true for K. We now further assume

$$\det H \neq 0 \quad \text{and} \quad \det K \neq 0. \tag{4.16}$$

Via multiplication by suitable unit scalars on H and K, we assume $h_{11} > 0$ and $k_{11} > 0$.

Write

$$H = \begin{bmatrix} h_{11} & |h_{12}|e^{\mathbf{i}\theta_{12}} \\ |h_{21}|e^{\mathbf{i}\theta_{21}} & -h_{11} \end{bmatrix} \quad \text{and} \quad K = \begin{bmatrix} k_{11} & |k_{12}|e^{\mathbf{i}\mu_{12}} \\ |k_{21}|e^{\mathbf{i}\mu_{21}} & -k_{11} \end{bmatrix},$$

where $\theta_{12}, \theta_{21}, \mu_{12}, \mu_{21} \in [0, 2\pi)$. Define

$$H_1(\theta) = \begin{bmatrix} h_{11} & h_{12}e^{\mathbf{i}\theta} \\ h_{21} & -h_{11} \end{bmatrix} \quad \text{and} \quad K_1(\theta) = \begin{bmatrix} k_{11} & k_{12}e^{\mathbf{i}\theta} \\ k_{21} & -k_{11} \end{bmatrix}, \quad \theta \in J,$$

where J is an open interval containing 0 such that $\det H_1(\theta)$ and $\det K_1(\theta)$ are nonzero on J. Such an interval exists because $H_1(0) = H$, $K_1(0) = K$ and (4.16). We see that for any $\theta \in J$, $H_1(\theta)$ and $K_1(\theta)$ satisfy (4.13) and (4.14), though they may not belong to $\Phi(\mathbb{C})$ because their trace norms may not be 1. If there exists a $\theta_0 \in J$ such that $\|H_1(\theta_0)\|_1^2 \cdot \|K_1(\theta_0)\|_1^2 < 1$, then for $\alpha = 1/\|H_1(\theta_0)\|_1$ and $\beta = 1/\|K_1(\theta_0)\|_1$, $\alpha\beta > 1$, $\|\alpha H_1(\theta_0)\|_1 = 1$ and $\|\beta K_1(\theta_0)\|_1 = 1$. It is easy to check that $\alpha H_1(\theta_0)$ and $\beta K_1(\theta_0)$ satisfy (4.13)–(4.15) and

$$F(\alpha H_1(\theta_0), \beta K_1(\theta_0)) = \left|\alpha\beta e^{\mathbf{i}\theta_0}(2h_{11}k_{12} - 2k_{11}h_{12})\right| = \alpha\beta\Delta^{\mathbb{C}}_{|\lambda|} > \Delta^{\mathbb{C}}_{|\lambda|}.$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 1-20, January 2020.
http://repository.uwyo.edu/ela

This gives a contradiction. Thus, the function $G(\theta) = \|H_1(\theta)\|_1^2 \cdot \|K_1(\theta)\|_1^2$ has a global minimum value 1 attained at $\theta = 0$ and, consequently, $G'(0) = 0$ and $G''(0) \geq 0$. As $\|H_1(0)\|_1 = \|K_1(0)\|_1 = 1$, we get

$$\left(\|H_1(\theta)\|_1^2\right)'\big|_{\theta=0} + \left(\|K_1(\theta)\|_1^2\right)'\big|_{\theta=0} = G'(0) = 0, \tag{4.17}$$

and

$$\left(\|H_1(\theta)\|_1^2\right)''\big|_{\theta=0} + 2\left(\|H_1(\theta)\|_1^2\right)'\big|_{\theta=0} \cdot \left(\|K_1(\theta)\|_1^2\right)'\big|_{\theta=0} + \left(\|K_1(\theta)\|_1^2\right)''\big|_{\theta=0} = G''(0) \geq 0. \tag{4.18}$$

From (4.17), we have $\left(\|H_1(\theta)\|_1^2\right)'\big|_{\theta=0} \cdot \left(\|K_1(\theta)\|_1^2\right)'\big|_{\theta=0} \leq 0$, and thus, (4.18) implies

$$\left(\|H_1(\theta)\|_1^2\right)''\big|_{\theta=0} + \left(\|K_1(\theta)\|_1^2\right)''\big|_{\theta=0} \geq 0. \tag{4.19}$$

We now obtain the explicit expressions for (4.17) and (4.19). From (4.14), since $h_{11}k_{21} \neq 0$, we have

$$\theta_{21} = \mu_{21}. \tag{4.20}$$

Then

$$\begin{aligned}\|H_1(\theta)\|_1^2 &= 2h_{11}^2 + |h_{12}|^2 + |h_{21}|^2 + 2\left|h_{11}^2 + |h_{12}||h_{21}|e^{\mathbf{i}(\theta_{12}+\theta_{21}+\theta)}\right| \\ &= 2h_{11}^2 + |h_{12}|^2 + |h_{21}|^2 + 2\sqrt{h_{11}^4 + 2h_{11}^2|h_{12}||h_{21}|\cos(\theta_{12}+\theta_{21}+\theta) + |h_{12}|^2|h_{21}|^2}.\end{aligned}$$

Thus,

$$\left(\|H_1(\theta)\|_1^2\right)' = \frac{-2h_{11}^2|h_{12}||h_{21}|\sin(\theta_{12}+\theta_{21}+\theta)}{\sqrt{h_{11}^4 + 2h_{11}^2|h_{12}||h_{21}|\cos(\theta_{12}+\theta_{21}+\theta) + |h_{12}|^2|h_{21}|^2}},$$

and hence,

$$\left(\|H_1(\theta)\|_1^2\right)'\big|_{\theta=0} = \frac{-2h_{11}^2|h_{12}||h_{21}|\sin(\theta_{12}+\theta_{21})}{|\det H|}. \tag{4.21}$$

With a similar expression for $\left(\|K_1(\theta)\|_1\right)'\big|_{\theta=0}$, condition (4.17) implies

$$\frac{h_{11}^2|h_{12}||h_{21}|\sin(\theta_{12}+\theta_{21})}{|\det H|} + \frac{k_{11}^2|k_{12}||k_{21}|\sin(\mu_{12}+\mu_{21})}{|\det K|} = 0. \tag{4.22}$$

Also, by direct calculation,

$$\left(\|H_1(\theta)\|_1^2\right)''\big|_{\theta=0} = \frac{-2|h_{11}|^2|h_{12}||h_{21}|\cos(\theta_{12}+\theta_{21})}{|\det H|} - \frac{2|h_{11}|^4|h_{12}|^2|h_{21}|^2\sin^2(\theta_{12}+\theta_{21})}{|\det H|^3}.$$

Thus, with a similar expression for $\left(\|K_1(\theta)\|_1\right)''\big|_{\theta=0}$, (4.19) implies

$$\begin{aligned}&\frac{h_{11}^2|h_{12}||h_{21}|\cos(\theta_{12}+\theta_{21})}{|\det H|} + \frac{h_{11}^4|h_{12}|^2|h_{21}|^2\sin^2(\theta_{12}+\theta_{21})}{|\det H|^3} \\ &+ \frac{k_{11}^2|k_{12}||k_{21}|\cos(\mu_{12}+\mu_{21})}{|\det K|} + \frac{k_{11}^4|k_{12}|^2|k_{21}|^2\sin^2(\mu_{12}+\mu_{21})}{|\det K|^3} \leq 0\end{aligned} \tag{4.23}$$

Step 3. We now come to the final argument. We refer to (4.22) and divide the proof into two cases, depending on whether $\sin(\theta_{12}+\theta_{21})$ is zero or not.

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 1-20, January 2020.
http://repository.uwyo.edu/ela

Case 1. $\sin(\theta_{12}+\theta_{21})=0$. By (4.22), $\sin(\mu_{12}+\mu_{21})=0$ and so both $\theta_{12}+\theta_{21}$ and $\mu_{12}+\mu_{21}$ are multiples of π. With (4.20), we can easily deduce that H and K are of the form

$$H=\begin{bmatrix} h_{11} & \tau_H|h_{12}|e^{-\theta_{21}\mathbf{i}} \\ |h_{21}|e^{\theta_{21}\mathbf{i}} & -h_{11}\end{bmatrix} \quad \text{and} \quad K=\begin{bmatrix} k_{11} & \tau_K|k_{12}|e^{-\theta_{21}\mathbf{i}} \\ |k_{21}|e^{\theta_{21}\mathbf{i}} & -k_{11}\end{bmatrix},$$

where $\tau_H,\tau_K\in\{1,-1\}$. Let $D=\operatorname{diag}(1,e^{\mathbf{i}\theta_{21}})$, which is unitary. Then D^*HD and D^*KD are real matrices. For notation simplicity, instead of using D^*HD and D^*KD, we now just assume H and K are real. We have three subcases.

Subcase 1.1. $\det H<0$ and $\det K<0$. Here, both H and K belong to $\Phi(\mathbb{R})$. Consequently, we have $\Delta^{\mathbb{R}}_{|\lambda|}\geq\Delta^{\mathbb{C}}_{|\lambda|}$ and the result follows.

Subcase 1.2. $\det H>0$ and $\det K>0$. The condition $\det H>0$ implies $h_{11}^2+h_{12}h_{21}<0$ and consequently the condition $\|H\|_1^2=1$ becomes $(h_{12}-h_{21})^2=1$, which is independent of h_{11}. It means that as long as the condition $h_{11}^2+h_{12}h_{21}<0$ is satisfied, we may vary h_{11} freely. The same is true for k_{11} when $\det K>0$. Thus, for $\epsilon>0$ but small enough, the pair

$$\hat{H}=\begin{bmatrix} (1+\epsilon)h_{11} & h_{12} \\ h_{21} & -(1+\epsilon)h_{11}\end{bmatrix} \quad \text{and} \quad \hat{K}=\begin{bmatrix} (1+\epsilon)k_{11} & h_{12} \\ k_{21} & -(1+\epsilon)k_{11}\end{bmatrix}$$

satisfies (4.13)–(4.15) and $F(\hat{H},\hat{K})=(1+\epsilon)F(H,K)>F(H,K)=\Delta^{\mathbb{C}}_{|\lambda|}$. This gives a contradiction.

Subcase 1.3. $\det H<0$ and $\det K>0$ (the other case $\det H>0$ and $\det K<0$ is the same). We check that $X=H+\sqrt{|\det H|}I_2\in\Sigma_2(\mathbb{R})$ and $Y=K+\sqrt{|\det K|}I_2\mathbf{i}\in\Sigma_2(\mathbb{C})$. By Lemma 4.2, we conclude that either $\Delta^{\mathbb{C}}_{|\lambda|}\leq\delta^{\mathbb{R}}_{|\lambda|}$ or there is another pair that gives a larger value of F. The latter contradicts the maximality of $F(H,K)$. The result follows.

Case 2. $\sin(\theta_{12}+\theta_{21})\neq 0$. Suppose $\sin(\theta_{12}+\theta_{21})>0$ (the case $\sin(\theta_{12}+\theta_{21})<0$ is similar). Then $\sin(\mu_{12}+\mu_{21})<0$ by (4.22) and we have from (4.21)

$$\big(\|H_1(\theta)\|_1^2\big)'\big|_{\theta=0}<0 \quad \text{and} \quad \big(\|K_1(\theta)\|_1^2\big)'\big|_{\theta=0}>0.$$

Subcase 2.1. $\cos((\theta_{12}+\theta_{21})-(\mu_{12}+\mu_{21}))\neq -1$. We can find an $\epsilon\in J$ (with $|\epsilon|$ small enough, to be determined later) such that

$$\cos((\theta_{12}+\theta_{21})-(\mu_{12}+\mu_{21}))>\cos((\theta_{12}+\theta_{21})-(\mu_{12}+\mu_{21})+\epsilon).$$

Then, using (4.20),

$$\begin{aligned}\Delta^{\mathbb{C}}_{|\lambda|}&=2\left|h_{11}|k_{12}|-k_{11}|h_{12}|e^{\mathbf{i}((\theta_{12}+\theta_{21})-(\mu_{12}+\mu_{21}))}\right| \\ &<2\sqrt{h_{11}^2|k_{12}|^2-2h_{11}k_{11}|h_{12}||k_{12}|\cos((\theta_{12}+\theta_{21})-(\mu_{12}+\mu_{21})+\epsilon)+k_{11}^2|h_{12}|^2}:=p\end{aligned}$$

and at the same time, again using (4.20),

$$\begin{aligned}|\lambda|\leq|h_{12}k_{21}-k_{12}h_{21}|&=\left||h_{12}||k_{21}|e^{\mathbf{i}(\theta_{12}+\mu_{21}-\mu_{12}-\theta_{21})}-|k_{12}||h_{21}|\right| \\ &<\sqrt{|h_{12}|^2|k_{21}|^2-2|h_{12}||k_{21}||h_{21}||k_{12}|\cos((\theta_{12}+\theta_{21})-(\mu_{12}+\mu_{21})+\epsilon)+|k_{12}|^2|h_{21}|^2}:=q.\end{aligned}$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 36, pp. 1-20, January 2020.
http://repository.uwyo.edu/ela

Suppose $\epsilon > 0$. Since $\left(\|H_1(\theta)\|_1^2\right)'\big|_{\theta=0} < 0$, we know that $\left(\|H_1(\theta)\|_1^2\right)'$, being continuous on J, is negative on a neighborhood N of 0. Thus, $\|H_1(\theta)\|_1^2$ is decreasing on N. We can assume ϵ small enough so that $\|H_1(\epsilon)\|_1^2 < \|H_1(0)\|_1^2 = 1$. Note that

$$H_1(\epsilon)K - KH_1(\epsilon) = \begin{bmatrix} h_{12}e^{\mathbf{i}\epsilon}k_{21} - k_{12}h_{21} & 2h_{11}k_{12} - 2k_{11}h_{12}e^{\mathbf{i}\epsilon} \\ 0 & -(h_{12}e^{\mathbf{i}\epsilon}k_{21} - k_{12}h_{21}) \end{bmatrix}$$

with $|2h_{11}k_{12} - 2k_{11}h_{12}e^{\mathbf{i}\epsilon}| = p > \Delta^{\mathbb{C}}_{|\lambda|}$ and $|h_{12}e^{\mathbf{i}\epsilon}k_{21} - k_{12}h_{21}| = q > |\lambda|$. We now have a contradiction because $H_1(\epsilon)/\|H_1(\epsilon)\|_1$ and K satisfy (4.13)–(4.15) and

$$F(H_1(\epsilon)/\|H_1(\epsilon)\|_1, K) = p/\|H_1(\epsilon)\|_1 > \Delta^{\mathbb{C}}_{|\lambda|}/\|H_1(\epsilon)\|_1 > \Delta^{\mathbb{C}}_{|\lambda|}.$$

If $\epsilon < 0$, we use $\left(\|K_1(\theta)\|_1^2\right)'\big|_{\theta=0} > 0$, and we have a contradiction similarly.

Subcase 2.2. $\cos((\theta_{12} + \theta_{21}) - (\mu_{12} + \mu_{21})) = -1$. We have

$$(\theta_{12} + \theta_{21}) = (\mu_{12} + \mu_{21}) + (2k+1)\pi \quad \text{for some integer } k.$$

This implies

$$(0 \neq)\ \sin(\theta_{12} + \theta_{21}) = -\sin(\mu_{12} + \mu_{21}) \quad \text{and} \quad \cos(\theta_{12} + \theta_{21}) = -\cos(\mu_{12} + \mu_{21}). \tag{4.24}$$

Then, (4.22) and the first part of (4.24) give

$$\frac{h_{11}^2|h_{12}||h_{21}|}{|\det H|} = \frac{k_{11}^2|k_{12}||k_{21}|}{|\det K|}. \tag{4.25}$$

We now refer to (4.23). Using (4.24), (4.25) and the assumption that all the entries of H and K are nonzero, we get a contradiction. □

Acknowledgments. The authors would like to thank Prof. D. Wenzel for his valuable discussion on this topic, and the referee for his/her constructive suggestions which greatly improved the readability of the paper.

REFERENCES

[1] A. Böttcher and D. Wenzel. How big can the commutator of two matrices be and how big is it typically? *Linear Algebra Appl.*, 403:216–228, 2005.
[2] C.-M. Cheng, X. Jin, and S. Vong. A survey on the Böttcher-Wenzel conjecture and related problems. *Oper. Matrices*, 9:659–673, 2015.
[3] C.-M. Cheng and C. Lei. On Schatten p-norms of commutators. *Linear Algebra Appl.*, 484:409–434, 2015.
[4] C.-M. Cheng and Y. Liang. Some sharp bounds for the commutator of real matrices. *Linear Algebra Appl.*, 521:263–282, 2017.
[5] R.A. Horn and C.R. Johnson. *Topics in Matrix Analysis.* Cambridge University Press, Cambridge, 1991.
[6] Z. Lu and D. Wenzel. Commutator estimates comprising the Frobenius norm – Looking back and forth. In: D. Bini, T. Ehrhardt, A. Karlovich, and I. Spitkovsky (editors), *Large Truncated Toeplitz Matrices, Toeplitz Operators, and Related Topics*, *Oper. Theory Adv. Appl.*, 259:533–559, 2017.
[7] D. Wenzel. A strange phenomenon for the singular values of commutators with rank one matrices. *Electron. J. Linear Algebra*, 30:649–669, 2015.
[8] D. Wenzel and K. Audenaert. Impressions of convexity – An illustration for commutator bounds. *Linear Algebra Appl.*, 433:1726–1759, 2010.

mathematics

MDPI

Article

Studying Bone Remodelling and Tumour Growth for Therapy Predictive Control

Raquel Miranda [1], Susana Vinga [1,2] and Duarte Valério [1,*]

[1] IDMEC, Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisboa, Portugal; raquellopesmiranda@tecnico.ulisboa.pt (R.M.); susanavinga@tecnico.ulisboa.pt (S.V.)

[2] INESC–ID, Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisboa, Portugal

* Correspondence: duarte.valerio@tecnico.ulisboa.pt; Tel.: +351-218417600

Received: 25 February 2020; Accepted: 28 April 2020; Published: 1 May 2020

Abstract: Bone remodelling consists of cycles of bone resorption and formation executed mainly by osteoclasts and osteoblasts. Healthy bone remodelling is disrupted by diseases such as Multiple Myeloma and bone metastatic diseases. In this paper, a simple mathematical model with differential equations, which takes into account the evolution of osteoclasts, osteoblasts, bone mass and bone metastasis growth, is improved with a pharmacokinetic and pharmacodynamic (PK/PD) scheme of the drugs denosumab, bisphosphonates, proteasome inhibitors and paclitaxel. The major novelty is the inclusion of drug resistance phenomena, which resulted in two variations of the model, corresponding to different paradigms of the origin and development of the tumourous cell resistance condition. These models are then used as basis for an optimization of the drug dose applied, paving the way for personalized medicine. A Nonlinear Model Predictive Control scheme is used, which takes advantage of the convenient properties of a suggested adaptive and democratic variant of Particle Swarm Optimization. Drug prescriptions obtained in this way provide useful insights into dose administration strategies. They also show how results may change depending on which of the two very different paradigms of drug resistance is used to model the behaviour of the tumour.

Keywords: bone remodelling; PK/PD; bone metastasis; model predictive control; particle swarm optimization

1. Introduction

Bone remodelling is a dynamic process that remains active throughout the entire life cycle. This mechanism depends on a complex control system which depends on innumerous hormones, cells, cytokines, among other endless variables. Cancer can be viewed as a loss of tissue homeostasis and it defines the diseases in which abnormal cells divide without control and can invade nearby tissues. Tumour presence provokes severe alterations in the bone remodelling regulation. Multiple myeloma (MM) is the most common cancer to involve bone.

The ongoing constant battle against cancer has been strengthened with groundbreaking discoveries over the years regarding experimental findings and mathematical modelling advances, fundamental for a better understanding of the relationship between experimental and theoretical approaches. Mathematical modelling efforts are crucial to identify the treatment schedules that maximally extend patient survival, perform qualitative and quantitative conclusions regarding certain physiological and biochemical counter-intuitive mechanisms or controlling drug-resistant sub populations within the tumour [1,2].

This paper develops models found in the literature of bone remodelling in the presence of tumours and cancer treatments, in the form of non-linear systems of differential equations, so as to include pharmacokinetic and pharmacodynamic (PK/PD) effects and the resistance to treatment that tumours

develop; such models are then used in simulations to find optimal treatment schedules and strategies, using Nonlinear Model Predictive Control.

The structure of the paper is as follows: Section 2 shortly reviews the mechanisms of bone remodelling and tumour growth; Section 3 presents the improved mathematical models; Section 4 presents the mathematical tools to optimise treatment schedules and Section 5 addresses their implementation; Section 6 shows and discusses simulation results, and Section 7 sums up conclusions.

2. Background—Bone Remodelling and Tumours

Bone remodelling is a coordinated, spatially heterogeneous and adaptive process. The tissue is continuous in a process of removal of old and damaged tissue by *osteoclasts* (OC) and subsequent reconstruction of the resorptive cavities with new material by *osteoblasts* (OB). This mechanism maintains the skeleton size, its structure and the mineral homoeostasis [3]. It serves to repair microdefects in the bone matrix and readjust the bone strength to meet new mechanical needs. In a healthy adult bone, the amount of bone that is absorbed is the same as the one formed afterwards, so that the bone mass remains approximately constant. The resorption and formation processes are strongly coupled through anatomic structures termed basic multicellular units (BMU).

OB, the bone forming cells, result from a differentiation pathway under the control of a defined series of transcription factors. It starts with mesenchymal stem cells (MSC), which differentiate into osteoprogenitors, which in turn give rise to preosteoblasts and finally transform into mature osteoblasts [4,5]. OC are multinucleated cells formed by the fusion of mononuclear progenitors of the monocyte/macrophage haematopoietic lineage [6]. Disturbances in osteoclasts and osteoblasts activity and coupling give rise to diseases such as osteoporosis or osteopetrosis.

2.1. Bone Remodelling Regulation

Osteoblastogenesis and osteoclastogenesis are tightly linked and regulated by several autocrine and paracrine signalling factors: proteins and hormones secreted by hemopoietic bone marrow cells or bone cells. Bone remodelling regulation is both systemic and local [7].

Osteoclast precursors express *Receptor Activator of Nuclear Factor kB* (RANK) and Macrophage Colony-stimulating Factor Receptor (c-fms). Osteoblasts produce and release Macrophage colony-stimulating factor (M-CSF) and express the receptor for activation of nuclear factor kappa B (NF-kB) ligand (RANKL), which are key regulators in osteoclasts differentiation and growth. As the protein RANKL and M-CFS bind to the preosteoclastic cells' receptors, RANK and c-fms respectively, signaling pathways are triggered, which promote the survival and differentiation into mature osteoclasts, leading to an increase of the resorption of the bone. Cells of the osteoblast lineage also segregate *osteoprotegerin* (OPG). This protein acts like a decoy receptor and binds with RANKL, keeping the latter to connect and activate its receptor in the surface of osteoclasts. OPG ihnibits their final differentiation and induces osteoclast apoptosis [6]. The RANKL/OPG ratio determines the degree of osteoclast differentiation, function and apoptosis.

Other factors responsible for bone remodelling regulation are within the following five groups [8]: Systemic hormones, Local cytokines and signals, vitamins and minerals, genetic factors and mechanical loading. Among innumerable agents, the *parathyroid hormone* (PTH), Insulin and Transforming Growth Factors (IGF and TGF), interleukins (IL) and Tumour Necrosis Factors (TNF) are highlighted for their relevance.

2.2. The Process of Bone Remodelling

Activation Phase: The cycle starts with the identification of a triggering signal, which can be a loss in mechanical loading, a disturbance in calcium homeostasis or a change in hormones/citokines concentrations. Osteocytes are cells that are trapped inside the bone matrix. These produce TGF-β, which inhibits osteoclastogenesis. Once osteocyte local apoptosis happen due to mechanical loading, the factor's local concentration decreases, allowing resorption to increase and the remodelling cycle to

begin [3]. Another concept suggests that osteoblastic cells receive the osteocytes signalling and activate the BMU [9]. When there is an hormone disturbance, such as a PTH increase, the cycle is also induced, since this hormone binds directly to OB and promotes OC differentiation and activation [4].

Reversal Phase: PTH interaction with the OB leads the latter to segregate monocyte chemoattractant protein-1 (MCP-1), which recruits the OC to the site. In response to the endocrine and mechanical activation signalling, osteoblasts also express matrix metalloproteinases (MMP). The bone surface is degraded by these proteins in order to facilitate osteoclast adhesion to the tissue. Osteoclasts attach to the bone, and after releasing acids, it absorbs th mineralized matrix. The remaining bone is degraded and removed by the enzyme cathepsin K. In this phase, osteoclast apoptosis occurs [10].

Formation Phase: The degradation of the bone matrix unleashes factors that attract the MSC.

Termination Phase: At this point, OC suffer apoptosis. Besides apoptosis, OB may also differentiate into bone lining cells or into osteocytes [5].

2.3. Tumour and Its Influence on Bone Remodelling

The origin and propagation of cancer appears to be caused by heritable changes in the genetic material of healthy cells—*mutations*. Cancer is commonly divided into categories according to its origin in the body (primary site). Cells from this primary site may spread to other parts of the host body through the bloodstream or lymphatic system, a process called metastasization. A bone metastasis is a part of the bone containing cancer cells and are the result of complex interactions between tumour cells, bone cells and their microenvironment [11]. They are responsible for deregulating the normal functioning of bone remodelling and are commonly classified in two extreme phenotypes according to the distortion of the coupling: *osteoblastic*, when bone formation is enhanced and *osteolytic*, when resorption is promoted [12]. This work is solely focused on the latter.

Tumour cells release several factors such as parathyroid hormone-related protein (PTHrP) and interleukins IL-6, IL-8 and IL-11 which promote osteolysis by stimulating osteoclast activity. A vicious cycle is established: factors that are trapped in the bone matrix and expressed during resorption, such as TGF-β, vascular endothelial growth factor (VEGF) and IGFs stimulate tumour cells' survival and proliferation and subsequently PTHrP production. [11,13–15].

These interactions are graphically depicted in Figure 1. For more details about bone remodelling see, e.g., [16–18]; as to the effects of of tumours and their treatments, see, e.g., [19–21].

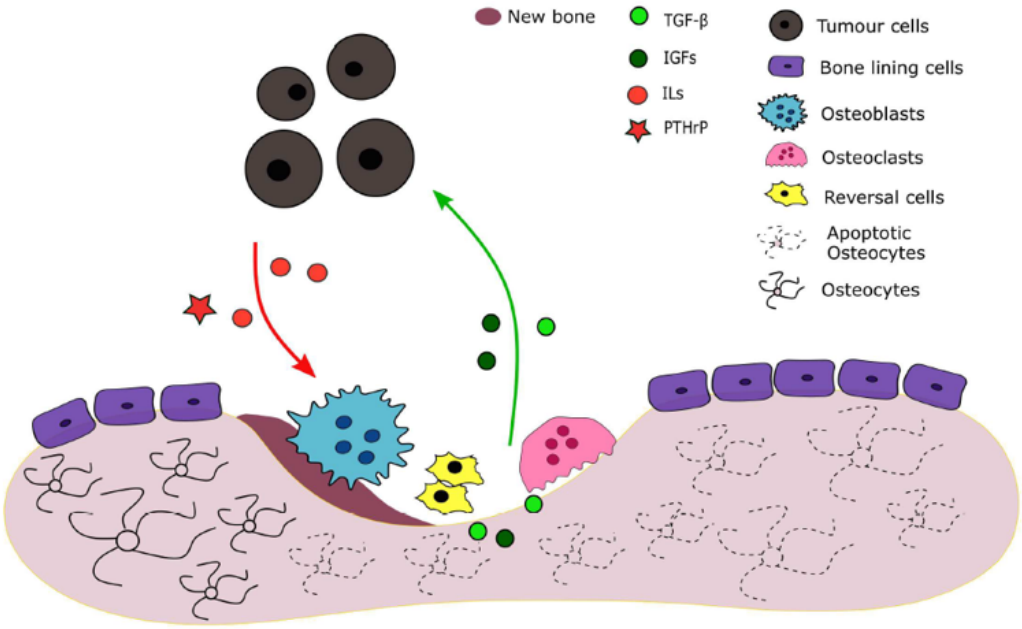


Figure 1. Tumour influence on bone remodelling.

3. Proposed Models

3.1. Pharmacodynamics and Pharmacokinetics

The proposed bone remodelling and tumour growth model is based on Ayati's [22] work. From that formulation, the therapy effects of four drugs were added: Denosumab (T_1), bisphosphonates or BP (T_2), Paclitaxel (T_3) and proteasome inhibitors or PI (T_4). The Pharmacokinectics, or PK (this designates the study of the time evolution of drug absorption, distribution, metabolism and excretion [23]), of each drug j is modelled as the two-compartmental model (1).

$$\frac{\mathrm{d}}{\mathrm{d}t}C_{g_j}(t) = -k_{a_j}C_{g_j}(t) \tag{1a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}C_{p_j}(t) = k_{a_j}C_{g_j}(t) - k_{e_j}C_{p_j}(t) \tag{1b}$$

The variable $C_g(t)$ is the concentration of the drug in a peripheral compartment, $C_p(t)$ is the effective concentration in plasma, k_a is the absorption rate and k_e, the elimination rate [24].

The Pharmacodynamics, or PD (this refers to the relationship between drug concentration at the site of action and the resulting effect [25]), may be modelled with a Hill function (Equation (2)), whose output $d_j(t)$ is the drug effect in the system and lies within the interval $[0, 1]$. The variable $C_{50}(t)$ is the concentration that achieves 50% of the drug's maximum effect.

$$d_j(t) = \frac{C_{p_j}(t)}{C_{50_j}(t) + C_{p_j}(t)} \tag{2}$$

3.2. Finding the Bone Mass

The model equations relative to the OC number $C(t)$, OB number $B(t)$, bone mass $z(t)$ in percentage of its steady-state value, and tumour burden $T(t)$ in percentage of bone mass at time t [days] are represented by (3). The tumour growth is described with a gompertzian curve. Behind this equation lies the idea that the *per capita* growth of the population decreases exponentially with time. Its sigmoidal shape is qualitatively conceivable; the growth rate derivative of small sized tumours should be increasing, since they easily adapt to the environment obstacles. As the tumour increases in size, it the proliferation becomes more difficult, considering that the host physiology is more degraded, and the resources start to lack.

$$\frac{\mathrm{d}}{\mathrm{d}t}C(t) = \alpha_1 C(t)^{G_{11}} B(t)^{G_{21}} - (\beta_1 + K_{d_2} d_2(t))C(t) \tag{3a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}B(t) = \alpha_2 C(t)^{G_{12}} B(t)^{G_{22}} - (\beta_2 - K_{d_4} d_4(t))B(t) \tag{3b}$$

$$G_{11} = g_{11}(1 + r_{11}\frac{T(t)}{L_T}) \tag{3c}$$

$$G_{21} = g_{21}(1 + r_{21}\frac{T(t)}{L_T}) - K_{d_1} d_1(t) \tag{3d}$$

$$G_{12} = g_{12}/(1 + r_{12}\frac{T(t)}{L_T}) \tag{3e}$$

$$G_{22} = g_{22}(1 - r_{22}\frac{T(t)}{L_T}) \tag{3f}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}z(t) = -k_1 \max[0, C(t) - \bar{C}] + k_2 \max[0, B(t) - \bar{B}] \tag{3g}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}T(t) = (\gamma_{tot})T(t)\log\left(\frac{L_T}{T(t)}\right) \tag{3h}$$

$$\gamma_{tot} = \gamma_T - K_{d_3}d_3(t) + K_T h(t) \tag{3i}$$

$$h(t) = \max[0, C(t) - \bar{C}] \tag{3j}$$

The parameters α_i and β_i are activities of cell production and removal, respectively. The index 1 refers to the OC population, and the index 2 to the OB. The autocrine and paracrine effects between OC and OB are not treated separately, their contributions are summed and expressed as the parameters g_{ij}, the net effectiveness of osteoclast or osteoblast-derived autocrine or paracrine factors. These can be positive (stimulatory) or negative (inhibitory).

The bone mass variation is attributed to the cell proliferation above the respective non-trivial steady-state levels, $\bar{C}$ (OC) and $\bar{B}$ (OB). The cells under this value are considered to be unable to resorpt or form bone, but they still participate in autocrine and paracrine signalling. The rates of bone formation and resorption are proportional to the number of osteoclasts and osteoblasts that exceed steady-state values, and k_i represents the normalized activity of bone resorption ($i = 1$) and formation ($i = 2$). The parameter L_T is an arbitrary value for the maximum size of $T(t)$ and γ_T is the respective growth constant. The values r translate the effect of the metastasis size in the autocrine and paracrine factors. The term $h(t)$, given by (3j), is introduced in this work and represents the influence of an excessive osteoclastic activity in tumour proliferation. The parameter K_T measures the effect of this influence.

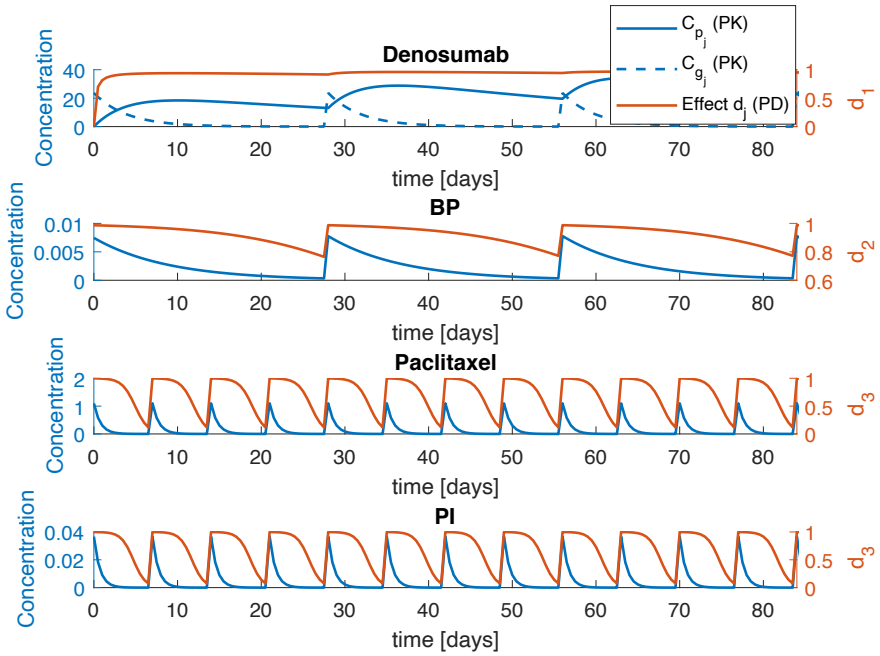


Figure 2. Pharmacokinetic/pharmacodynamic (PK/PD) behaviour of T_1, T_2, T_3 and T_4; concentrations in mg/L.

Table 1. PK/PD parameters.

Parameter	Units	T_1	T_2	T_3	T_4
C_{50}	mg/L	1.2	0.0001	0.002	0.00005
K_d		0.09	0.005	0.015	0.002
τ	day	4	4	1	1
D_0	mg	120	4	176	2.5
F		0.62	-	-	-
V_d	L	3.15	536.4	160.25	68.18
k_a	day^{-1}	0.2568	0	0	0
k_e	day^{-1}	0.0248	0.1139	1.2797	1.40

K_{d_j} is the maximum effect of the drug T_j. The system (3) is a general representation that comprises the four drugs. In reality, these drugs are grouped in three combined therapies: $T_1 + T_3$, $T_2 + T_3$ and $T_4 + T_3$. This means that there are only two inputs: the anti-cancer drug effect $d_3(t)$ and one of the three bone therapies effects, $d_1(t)$, $d_2(t)$ or $d_4(t)$. The PK/PD response is illustrated in Figure 2. The parameters of this simulation can be found in Table 1, including the periodicity of administration τ. The PK/PD initial conditions for a single administration are obtained with $C_g^0 = D_0 F / V_d$, for T_1 and $C_p^0 = D_0 / V_d$, for the remaining drugs. The parameters D_0, V_d and F correspond to the initial dose, volume distribution and bioavailability, respectively.

The PK/PD models and respective parameters regarding T_1, T_2 and T_3 were suggested by Coelho et al. [4]. The parameters regarding the PI model were estimated from a non-compartmental analysis in plasma of patients with advanced solid tumours, specifically the half life $t_{1/2}$ and estimated C_p^0 [26]. The remaining parameters of the model are fixed as: $\alpha_1 = 3\ \text{day}^{-1}$, $\alpha_2 = 4\ \text{day}^{-1}$, $\beta_1 = 0.2\ \text{day}^{-1}$, $\beta_2 = 0.02\ \text{day}^{-1}$, $k_1 = 7.48^{-2}\ \text{day}^{-1}$, $k_2 = 5.52^{-4}\ \text{day}^{-1}$, $g_{11} = 1.1$, $g_{12} = 1$, $g_{21} = -0.5$, $g_{22} = 0$, $L_T = 100\%$, $K_T = 4 \times 10^{-4}$, $\gamma_T = 5 \times 10^{-3}\ \text{day}^{-1}$, $r_{11} = 5 \times 10^{-3}$, $r_{12} = 0.0$, $r_{21} = 0.0$, $r_{22} = 0.2$, $K_r = 0.8$, $C_{50^{base}} = 2 \times 10^{-3}$ mg/L, $\lambda_1 = 10^{-6}$, $\lambda_2 = 10^{-6}$, $C(0) = 15$, $B(0) = 316$, $z(0) = 100\%$, $T(0) = 0.001\%$, $\bar{C} = 5$ and $\bar{B} = 316$.

3.3. Inclusion of Drug Resistance

Model (3) diverges into two variations that approach the resistance to paclitaxel in two different manners, the models M_{DR1} and M_{DR2}, each of them corresponding to a possible model of drug resistance found in the literature [24].

Model M_{DR1} considers that the resistance accumulation is caused by C_p levels of the drug below a certain threshold C_p^{th} [27]. The C_{50_3} is affected according to

$$\begin{aligned} C_{50_3}(t) &= f(t) C_{50}^{base} \\ f(t) &= 1 + K_r \int_0^t \max[0, C_p^{th} - C_p(\tau)] d\tau, \end{aligned} \tag{4}$$

where parameter K_r translates the capacity of the tumour cells to develop resistance and C_{50}^{base} is a constant which represents the initial value of C_{50_3}.

Model M_{DR2} is based on the Random Mutation Model (RMM) [28], a Darwinian theory that proposes the existence of two proliferative tumour cell populations: $S(t)$ is composed completely sensitive cells, and $R(t)$ by completely resistant ones. The combination between the RMM and the proposed model (3) results in the following tumour growth description:

$$T(t) = R(t) + S(t) \tag{5a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} S(t) = \gamma_S S(t) \log\left(\frac{L_T}{T(t)}\right) + \lambda_2 R(t) \log\left(\frac{L_T}{T(t)}\right) \tag{5b}$$

$$\gamma_S = \gamma_T - K_{d_3} d_3(t) - \lambda_1 + K_{T1} h(t); \tag{5c}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}R(t) = \gamma_R R(t)\log\left(\frac{L_T}{T(t)}\right) + \lambda_1 S(t)\log\left(\frac{L_T}{T(t)}\right) \tag{5d}$$

$$\gamma_R = \gamma_T - \lambda_2 + K_{T1}h(t), \tag{5e}$$

where λ_1 and λ_2 are the mutation and back-mutation rates between the S and R [29].

4. MPC–PSO scheme

4.1. Nonlinear Model Predictive Control

Model predictive control (MPC) refers to a class of control methods that make use of an an explicit process model to predict the future response of a system and obtain the control sequence over a certain horizon that minimizes a cost function. MPC performance is therefore highly dependent on the model performance. This also called receding horizon control does not allow the current time slot to be optimized, while keeping future time slots in account, which represents a major advantage; see Figure 3. The input is optimized for a finite prediction horizon N_p and subsequently the first entries of the optimized input sequence (control horizon N_c) are fed back. It represents an advantage over classical control since it has the ability to explicitly account for systems constraints, the constrained nonlinear optimization problem is easy to formulate, multivariable processes can be handled in a straightforward manner, and reference tracking can be improved if the references are known in advance. Besides, it easily handles nonlinear and time-varying plant dynamics, since the controller is a function of the system and can be modified in real time [30,31]. The proposed implementation of NMPC in this work will count on metaheuristics, specifically Particle Swarm Optimization, to solve the nonlinear problem at each step. Combining metaheuristics with MPC brings flexibility to design any type of cost function.

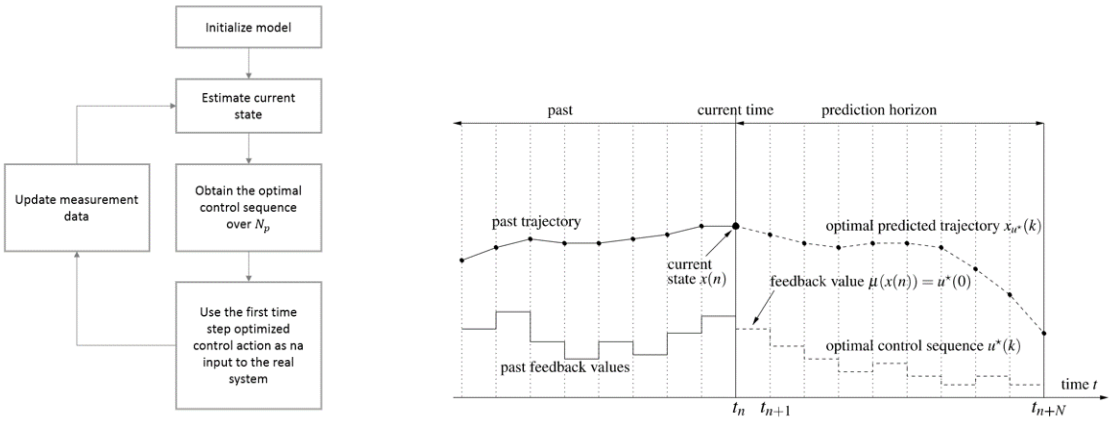


Figure 3. Flowchart of model predictive control (MPC) (**left**) and illustration of prediction horizon (**right**).

4.2. Proposed PSO Algorithm

Particle Swarm Optimization (PSO) is a collective, anarchic, nature-inspired population-based search algorithm [32,33]. It is inspired in the social behaviour of a bird flock. PSO algorithms are a common choice to solve the optimization problem involved in model predictive control schemes [34,35].

The swarm is composed by S particles wandering in a D-dimensional space. The position coordinates of each particle i are equivalent to a candidate solution. The particles' position and velocity are updated taking into account advantageous positions of the surrounding partners. This position adjustment depends on the difference between the particles' current position $\mathbf{x_i}$ and two others: $\mathbf{p}_i$

(the best position visited by particle i, and $\mathbf{p}_g$ (the best position visited by any particle of the swarm). At each iteration k, the particles' position and velocity vector $\mathbf{v_i}$ are updated according to

$$v_{ij}(k) = \chi\left[w_{ij}v_{ij}(k-1) + c_1(k)\text{rand}_1(p_{ij} - x_{ij}(k-1)) + c_2(k)\text{rand}_2(p_{g_j} - x_{ij}(k-1)) + c_3(k)\text{rand}_3 d_{ij}\right] \quad (6)$$

$$x_{ij}(k) = x_{ij}(k-1) + v_{ij}(k). \quad (7)$$

The acceleration/confidence parameters c_1, c_2 and c_3 are the cognitive, social and democratic coefficient, respectively. The particle's inertia is measured by the inertia weight w and rand is a random number in the interval [0,1]. The last term is not present in the traditional PSO. This democratic approach brings to the velocity update the opinion of all eligible particles of the swarm [36]. The vector $\mathbf{d_e}$ contains this swarm contribution and is obtained with

$$\mathbf{d_e}_i = \sum_{k=1}^{n} Q_{ik}(x_{kj} - x_{ij}) \quad (8a)$$

$$Q_{ik} = \frac{E_{ik}\frac{\text{cost}(x_{best})}{\text{cost}(x_k)}}{\sum_{j=1}^{n} E_{ij}\frac{\text{cost}(x_{best})}{\text{cost}(x_j)}} \quad (8b)$$

$$E_{ik} = \begin{cases} 1 & \text{if } \dfrac{\text{cost}(x_k) - \text{cost}(x_i)}{\text{cost}(x_{worst}) - \text{cost}(x_{best})} > \text{rand} \\ & \quad \wedge \text{cost}(x_k) < \text{cost}(x_i) \\ 0 & \text{otherwise} \end{cases}, \quad (8c)$$

where d_{ij} is the jth entry of vector $\mathbf{d_e}$ for each particle i, c_3 is the confidence coefficient which controls the weight of the the democratic quantities, and $\text{cost}(x)$ is the cost function chosen for the particular problem evaluated at x. This vector represents the democratic effect of the other particles in the movement of particle i. The weight of the kth particle is represented by Q_{ik}, which depends on the eligibility parameter E_{ik}. The best and worst particles of the swarm at each iteration are denoted x_{best} and x_{worst}, respectively.

The adaptive profile of the proposed algorithm, **AD-PSO** is translated in the inertia weight evolution with the iterations [37]. The inertia weight value regarding the jth variable of the particle i is updated with the Equations (9) and (10).

Value w_0 is a constant that defines the initial inertia weight. Parameter ϵ is a non-critical small and positive value that ensures a proper variation of the inertia weight. Value Λ is obtained with (10), where ϵ is a non-critical small and positive value that ensures a proper variation of the inertia weight.

$$w_{ij}(k+1) = \begin{cases} \min\left(1, w_{ij}(k) + (1-w_0)e^{\Lambda} + \epsilon\right) & \text{if } \delta_i(k) > 0 \wedge \delta_i(k-1) > 0 \\ \max\left(0.1, w_{ij}(k) - w_0\left(\left(1 - e^{\Lambda}\right) - \epsilon\right)\right) & \text{if } \delta_i(k) < 0 \wedge \delta_i(k-1) < 0 \\ w_{ij}(k) & \text{otherwise} \end{cases} \quad (9)$$

$$\Lambda = \frac{(x_{ij}(k+1) - p_{ij}(k))^2}{-2\sigma^2} \quad (10)$$

The values δ_i measure the success of particle i in the following manner:

$$\delta_i = \begin{cases} 1 & \text{if } \text{cost}(\mathbf{x}_1) < \text{cost}(\mathbf{p}_i) \\ -1 & \text{otherwise} \end{cases} \quad (11)$$

The **TVAC** algorithm (Time Varying Acceleration Coefficients) [38] dictates the dynamic behaviour of both c_1 and c_2 according to Equations (12).

$$c_1(k) = (c_{1_f} - c_{1_i})\frac{k}{n_{iter}} + c_{1_i} \tag{12a}$$

$$c_2(k) = (c_{2_f} - c_{2_i})\frac{k}{n_{iter}} + c_{2_i} \tag{12b}$$

The values c_{i_1} and c_{f1} are the initial and final values of the coefficient c_1 while c_{i2} and c_{f2} represent the initial and final values of c_2. These linearly increasing and decreasing behaviours are defined until the maximum number of iterations n_{iter}.

The minimum number of n_{iter} is set to $n_{iter} = 100$, however the algorithm stops when a state of convergence is achieved. The stopping criterion uses a counter θ which records the number of consecutive iterations with no improvement, after $k > n_{iter}$, according to (13). The algorithm stops when θ reaches a maximum value θ_{max}.

$$\theta(k) = \begin{cases} 0, & \text{if } \text{cost}(p_g(k)) < \text{cost}(p_g(k-1)) \vee k < n_{iter} \\ \theta(k-1) + 1, & \text{otherwise} \end{cases} \tag{13}$$

The dynamic parameters are maintained constant at $c_1(k) = 1.5$ and $c_2(k) = 2.5$ for $k > n_{iter}$.

The PSO parameters were fixed for this problem as: $S = 50$, $w_0 = 0.9$, $\epsilon = 0.005$, $c_{1_i} = c_{2_f} = 2.5$, $c_{2_i} = c_{1_f} = 1.5$, $\theta_{max} = 15$. The algorithm is depicted in Figure 4.

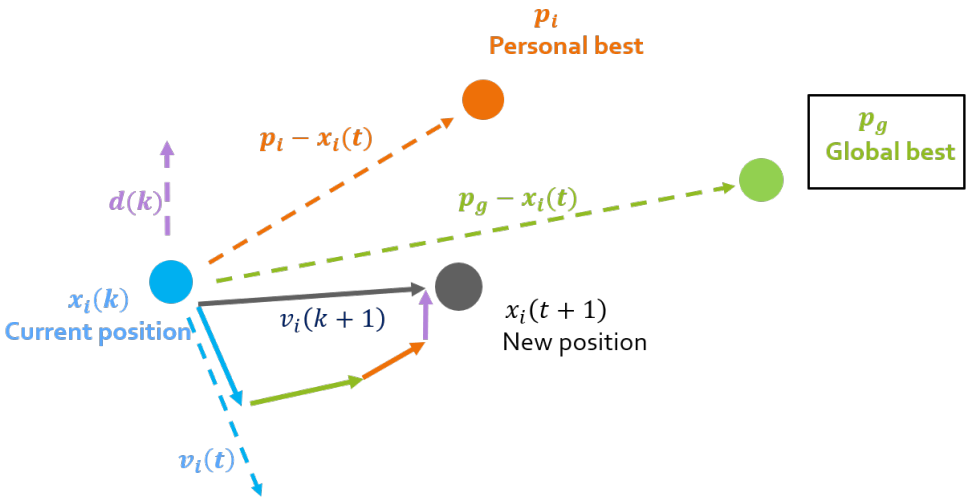


Figure 4. Particle Swarm Optimization (PSO).

5. Implementation

In this section, the NMPC-PSO scheme described in Section 4 is implemented with the objective of optimizing the prescription doses of the proposed therapies, when the drugs are administered with the fixed periodicity τ.

Only the therapies $T_1 + T_3$ and $T_4 + T_3$ are considered for this optimization. The BP therapy, although it results in a qualitatively viable therapy model, is not suitable for optimization. The rise of OC apoptosis due to BP decreases, although very slightly, the lower bound of the OC time response. The tumour $T(t)$ causes the opposite reaction: an increase of the mean value of $C(t)$, as well as its lower bound. When the tumour is proliferating, or has a substantial size, this lower bound increase cancels out the decrease caused by BP. When the tumour starts to be extinguished, the anti-resorptive effect

pushes the OC number to fall below zero. Although the negative values of OC are smaller orders than $\bar{C}$, it is enough to severely interfere with the dynamics, due to the appearance of complex numbers.

The standard regimen is defined as the prescription schedule which administrates constant values of C_p^0 and C_g^0. The decision variables are C_p^0, for the initial concentration of paclitaxel and PI and C_g^0 for the initial concentration for denosumab. These values for the initial concentration are fixed at $C_{g1}^0 = 23.62$, $C_{p4}^0 = 0.0367$ and $C_{p3}^0 = 1.0983$ (denosumab, PI and paclitaxel, respectively). The lower bound of the concentration is equal to zero for all drugs and the upper bound is considered to be three times the standard regimen doses, and therefore $C_{p1_{max}}^0 = 70.86$, $C_{p4_{max}}^0 = 0.11$ and $C_{p3_{max}}^0 = 3.29$. The maximum velocity of the particles, $\mathbf{v_{max}}$ is calculated offline as $\mathbf{v_{max}} = 0.2(\mathbf{x_{max}} - \mathbf{x_{min}})$ and the minimum as $\mathbf{v_{min}} = -\mathbf{v_{max}}$. The time of the diagnosis and therefore the beginning of the optimization is considered to be $t_{start} = 800$ days. The optimization is single-objective and the goal is to minimize the tumour size, the drug dosage and approximate the bone mass to a healthy level as much as possible. The objective function is defined as

$$\begin{aligned} J = w_1 \mathbf{D}_t^{-1}(1 - \frac{z(t)}{L_1})^2 + w_2 \mathbf{D}_t^{-0.8} \frac{T(t)}{L_2} \\ + w_3 \mathbf{D}_t^{-1} \frac{C_{p_{br}}(t)}{L_3} + w_4 \mathbf{D}_t^{-1} \frac{C_{p_3}(t)}{L_4}, \end{aligned} \tag{14}$$

where C_{p_3} and $C_{p_{br}}$ are the plasma concentration evolutions of paclitaxel and one of the two bone remodelling pharmaceuticals, C_{p_1} or C_{p_4}. The quantities L_1, L_2, L_3 and L_4 are maximum values of z, T, $C_{p_{br}}$ and C_{p3}, respectively and can be found in Table 2, as well as the weights w_1, w_2, w_3 and w_4. The functional operator D^{-1} is a Riemann integral, and D^α is a fractional derivative, defined here (according to Grünwald and Letnikoff) as [39]

$${}_cD_t^\alpha f(t) = \lim_{h \to 0^+} \frac{\sum_{k=0}^{[\frac{t-c}{h}]} (-1)^k \binom{\alpha}{k} f(t - kh)}{h^\alpha}, \tag{15}$$

where $\binom{a}{b}$, the combinations of a things, b at a time can be obtained with [40]

$$\binom{a}{b} = \begin{cases} \dfrac{\Gamma(a+1)}{\Gamma(b+1)\Gamma(a-b+1)} & \text{if } a, b, a-b \notin \mathbb{Z}^- \\ \dfrac{(-1)^b \Gamma(b-a)}{\Gamma(b+1)\Gamma(-a)} & \text{if } a \in Z^- \wedge b \in \mathbb{Z}_0^+ \\ 0 & \text{if } [(b \in \mathbb{Z}^- \vee b - a \in \mathbb{N}) \wedge a \notin \mathbb{Z}^-] \vee (a, b \in \mathbb{Z}^- \wedge a > b). \end{cases} \tag{16}$$

Γ is the gamma function, an extension of the factorial function.

A fractional derivative was used here because we are interested in attributing a weight to tumour level that increases with time; therefore, the functional order is set to $n = -0.8$, instead of order $n = -1$, which corresponds to a classical integrator and weights all past moments equally [40].

Table 2. Parameters of the objective function.

i	1	2	3		4
			Den.	**PI**	
L_i	100	100	3.29	0.11	70.86
w_i	4	15	0.5	1	1

6. Results

Figures 5 and 6 plot the system behaviour of models M_{DR1} and M_{DR2}, respectively, when each of the therapies are administrated. At first glance, both models under the therapies $T_1 + T_3$ an

$T_4 + T_3$ appear to have a similar qualitative behaviour. Comparing the bone resorption therapies (PI or denosumab), there are clear differences specially when it comes to the OC and OB oscillatory behaviour. The different type of action of both drugs justifies these discrepancies. One should not make comparison assumptions of these pharmaceuticals solely based on these simulations: the fact that $z(t)$ stabilization is more effective under a certain model may be due to the chosen system parameter values.

Regarding the different drug resistance models, the tumour growth behaviour is different depending on whether the phenomenon is modelled under the random mutation or the varying C_{50} model. Both models predict a decrease of T immediately after the therapy starts, more or less symmetric to the precedent increase for a few weeks. The tumour achieves then a low value that never reaches 0. In model $\mathbf{M}_{DR2}$, the drug resistance effects arise when the resistant population uncontrollable proliferation continues with the same strength as the initial cancer, after an apparent remission. As soon as the gradient turns positive, it is certain that the metastasis will grow abruptly to high values until death. On the other hand, model $\mathbf{M}_{DR1}$ allows a smoother accumulation of resistance. Even if the tumour does not reach values as low as with the last model, the cancer is maintained at a more constant value after the point when drug resistance is evident.

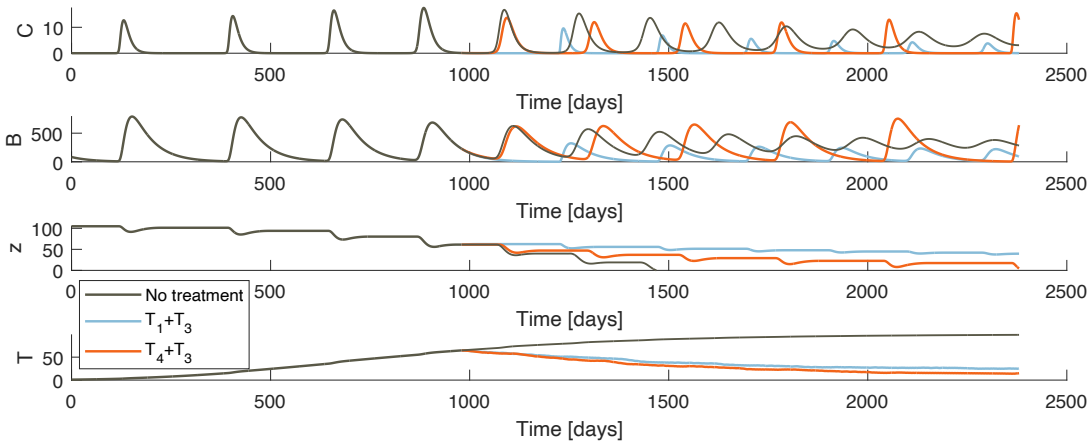


Figure 5. Model M_{DR1} behaviour when the tumour arises at $t = 0$ and the treatment begins at $t_{start} = 980$ days.

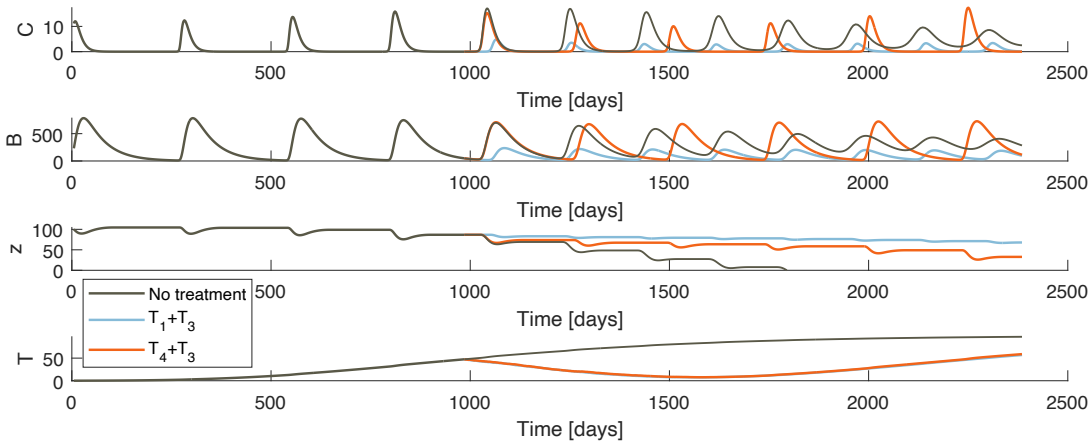


Figure 6. Model M_{DR2} behaviour when the tumour arises at $t = 0$ and the treatment begins at $t_{start} = 980$ days.

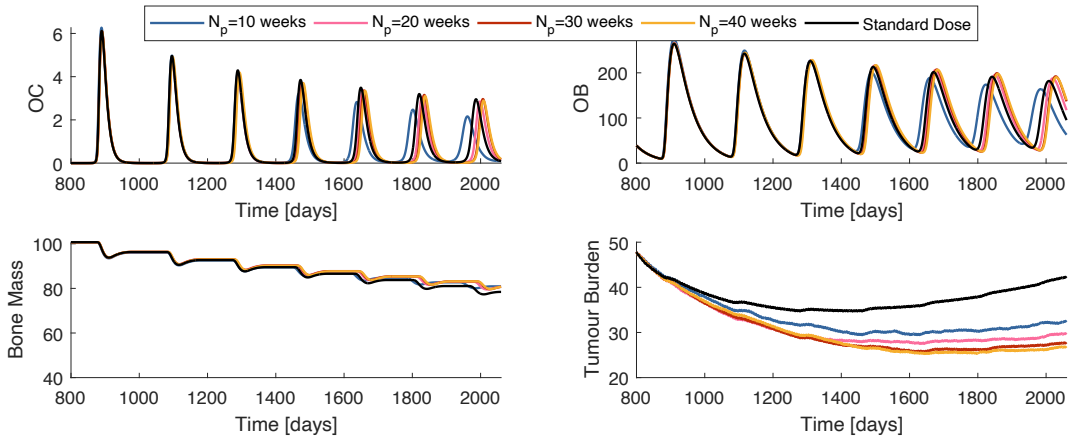


Figure 7. Effects in the system dynamics (treatment phase) when different N_p values are used (Model M_{DR1}).

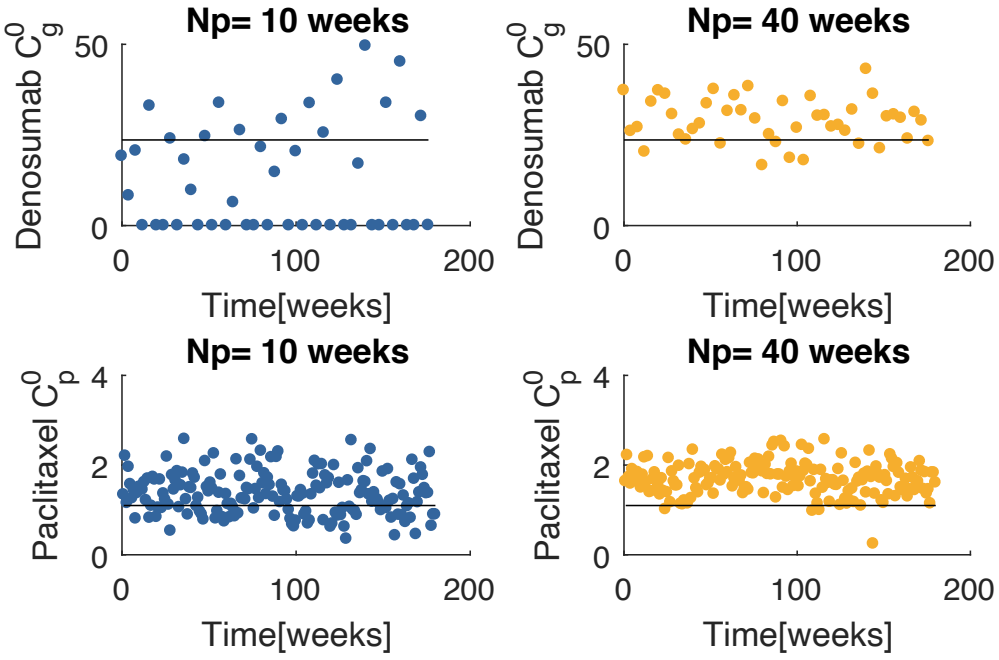


Figure 8. Best obtained prescriptions in mg with the model M_{DR1} (SIC).

6.1. Model M_{DR1}—Therapy $T_1 + T_3$

While the control horizon N_c was maintained constant, the prediction horizon N_p was varied between 10, 20, 30 and 40 weeks. The N_c is fixed to 4 weeks. The global best position $\mathbf{p}_g$ is initialized with the dose values of standard regimen, a strategy that from now on will be termed **SIC** (standard initial condition). Figure 7 is a comparison of the system reaction to the best obtained prescription when $N_p = 10$, $N_p = 20$, $N_p = 30$, $N_p = 40$ and when the prescription is standard. The cost decreases with the increase of N_p, when handling the model M_{DR1}. Nevertheless, all of the four prescriptions are more successful than the standard regimen.

Figure 8 contains the resultant prescription of denosumab and paclitaxel with the model M_{DR1}, when $N_p = 10$ and $N_p = 40$. The C_p^0 and C_g^0 mean values tend to coincide with a value higher than the respective standard dose, yet lower than the maximum values defined for this problem. Note that when $N_p = 10$, MPC obtains several null entries (22 out of 45 administrations), suggesting a higher τ for the denosumab. The increase in N_p tends to produce a drug concentration distribution less variant.

6.2. Model M_{DR2} — Therapy $T_1 + T_3$

The two populations model, $\mathbf{M}_{DR2}$, was subject to the same sensitivity analysis. The system dynamics when N_p is fixed to different values is compared in Figure 9. The almost indistinguishable curves translate the insensitivity of the model to N_p. This model allows a decrease of the tumour size to almost null values; however, when the resistance effects arise, the regrowth is extremely aggressive. The impossibility of tumour annihilation is associated with a resistant and proliferative population. Therefore, the resistance can never be defeated with a unique anti-cancer drug.

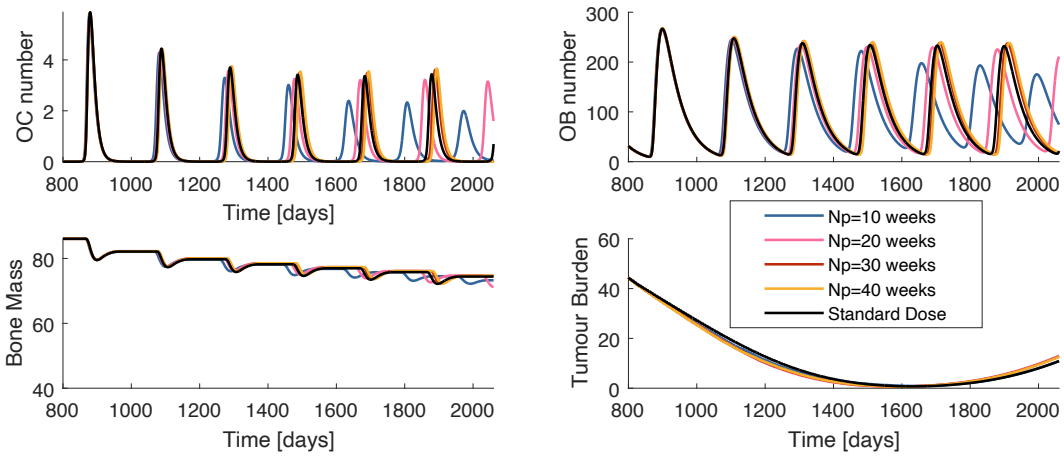


Figure 9. Effects in the system dynamics (treatment phase) when different N_p values are used (Model M_{DR2}).

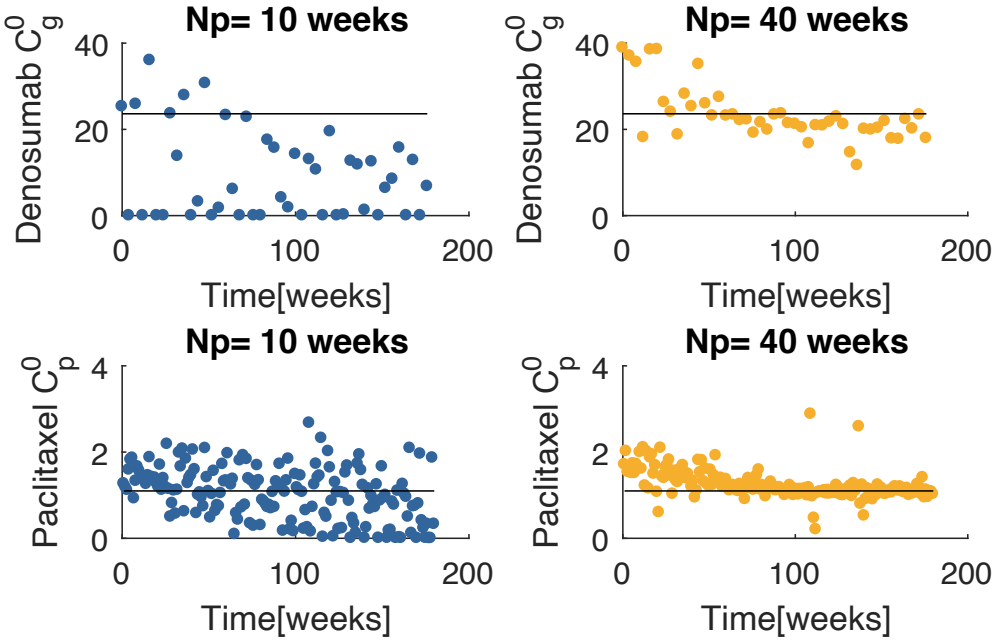


Figure 10. Best obtained prescriptions in mg with the model M_{DR2}.

The denosumab and paclitaxel prescription results are shown in Figure 10. From these plots it is verified that the paclitaxel is administrated even after the sensitive population is supposedly extinguished. One might suspect that this administration would be interrupted at some point, since it has no effect on the resistant population. In fact, the doses diminish over time, but they never reach 0.

This is due to the nature of the Gompertz equation: the sensitive population never actually reaches full extinction, just residual values. Besides, the S population keeps acquiring part of the resistance cells that suffer back-mutation. Therefore, there is a necessity of a paclitaxel continuity to keep the S population from regrowth. The outcome from the optimization when $N_p = 10$ suggests a higher τ_1 of denosumab, as it happened with M_{DR1}.

Although the algorithm accounts for one single cost, the proportions of each term of the objective function throughout the 45 months of treatment is displayed in Figure 11 (Model $\mathbf{M}_{DR1}$) and Figure 12 (Model $\mathbf{M}_{DR2}$). The left plots correspond to a patient who received the maximum dose allowed for this problem, while the right plots correspond to prescription resultant from optimization. As expected, the administration of the maximum allowable doses retrieve slightly better values for the tumour and bone mass associated cost; however, the inherent high administration cost does not allow this prescription to be a good option. Nevertheless, the optimization prescription outputs extremely close tumour and bone mass costs (in fact, almost indistinguishable) to those obtained with the most aggressive therapy. This represents a major advantage, since the patient is safeguarded from a therapy with higher drug exposure, but still obtaining very similar results.

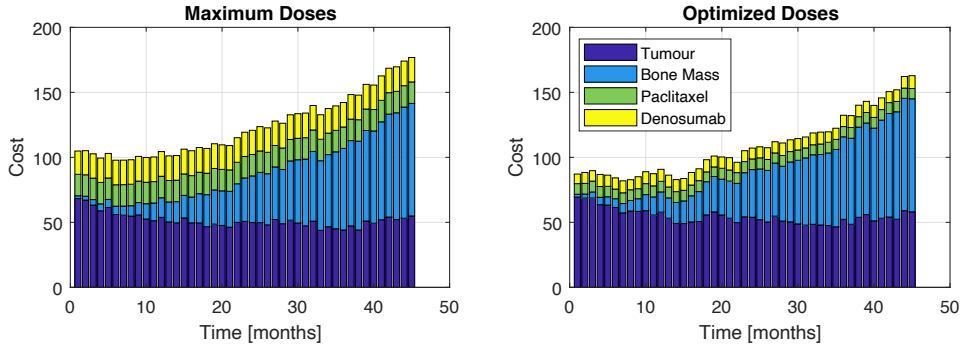


Figure 11. Proportions of the four cost contributions/terms in the objective function: Tumour, Bone Mass, Paclitaxel and Denosumab (Model $\mathbf{M}_{DR1}$, Therapy $T_1 + T_3$). The right plot represents the cost resultant from optimization, while the left represents the cost if the maximum dose was administrated.

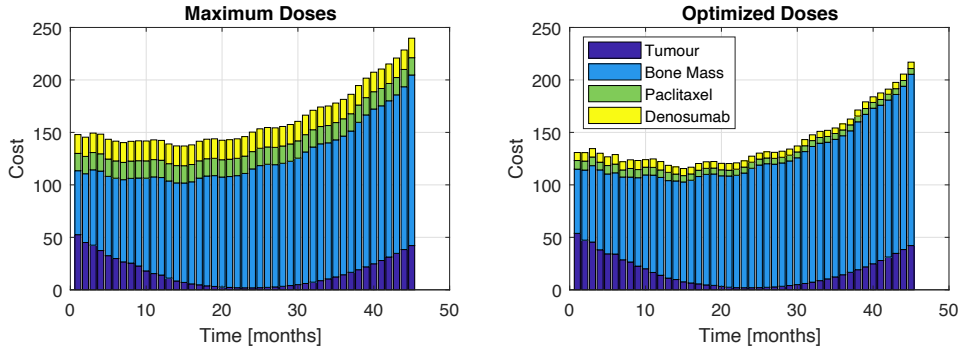


Figure 12. Proportions of the four cost contributions/terms in the objective function: Tumour, Bone Mass, Paclitaxel and denosumab (Model $\mathbf{M}_{DR2}$, Therapy $T_1 + T_3$). The right plot represents the cost resultant from optimization, while the left represents the cost if the maximum dose was administrated.

6.3. Therapy $T_4 + T_3$

When handling the therapy $T_4 + T_3$, the PSO deals with problems with higher dimensions, because T_4 has a more frequent administration than T_1. To avoid excessive computational effort, the optimization of this therapy was performed only once for each model, with $N_p = 40$ weeks. Figure 13 presents the best obtained prescriptions of PI and paclitaxel, for both models. As expected, the C_p^0 evolutions follow a similar pattern to that of the last therapy. When handling model M_{DR1}, a mobile mean value of both drugs is maintained almost constant, as well as the standard deviation.

When facing a two populations model, a decreasing tendency is evident for both dose value and standard deviation. In both therapies, $C^0_{p_3}$ converge to similar values.

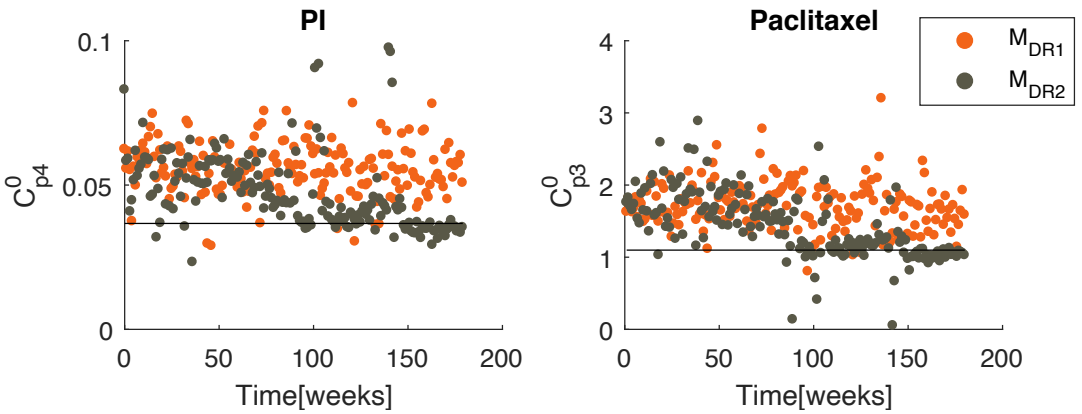


Figure 13. Best obtained PI prescription with both models ($N_p = 40$).

The system dynamics when the optimized regimen is applied is shown in Figure 14 for both models. The amplitude of oscillation of OC and OB is significantly higher when PI is used instead of denosumab, as the respective period.

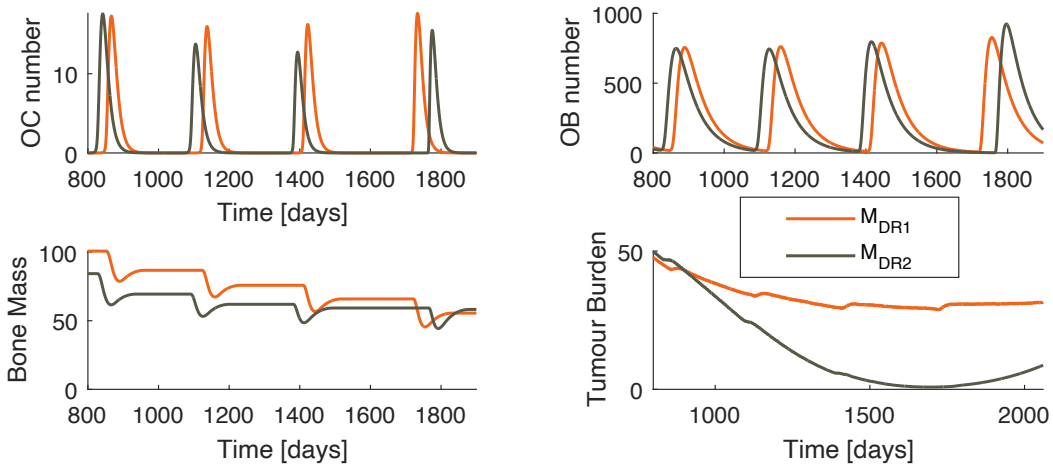


Figure 14. Dynamics of the systems M_{DR1} and M_{DR2} when gien an optimized prescription of paclitaxel and PI.

The cost proportions are once again represented, in Figures 15 and 16, for the models M_{DR1} and M_{DR2}, respectively. As before, the left plots refer to a patient who took maximum doses of both drugs and the right plots to a patient whose regimen was optimized. Figure 15 shows a similar increasing evolution to Figure 11, although the costs are considerably higher and there is a bigger discrepancy between the proportion of the weight related to the bone mass to the rest of the terms of J. Figures 14 and 16 reinforce the impossibility of avoiding a U-shaped tumour curve when dealing with M_{DR2}. The oscillating appearance of the curves of Figure 16 is due to the extremely high period and amplitudes that PIs provoke in the bone remodelling process.

6.4. Sensitivity to the Initial Global Best Position

The dependency of the optimization regarding the initial global best position is here analysed. The strategy SIC is replaced by the strategy LIC (low initial condition), which initializes $\mathbf{p_g}$ with concentration values that are ten times smaller than the standard regimen's. This section compares both strategies.

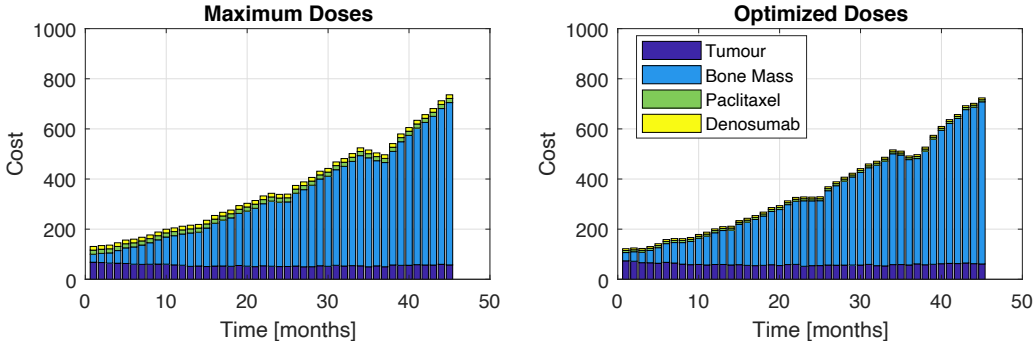


Figure 15. Proportions of the four cost contributions/terms in the objective function: Tumour, Bone Mass, Paclitaxel and PI (Model $\mathbf{M}_{DR1}$, Therapy $T_4 + T_3$). The right plot represents the cost resultant from optimization, while the left represents the cost if the maximum dose was administrated.

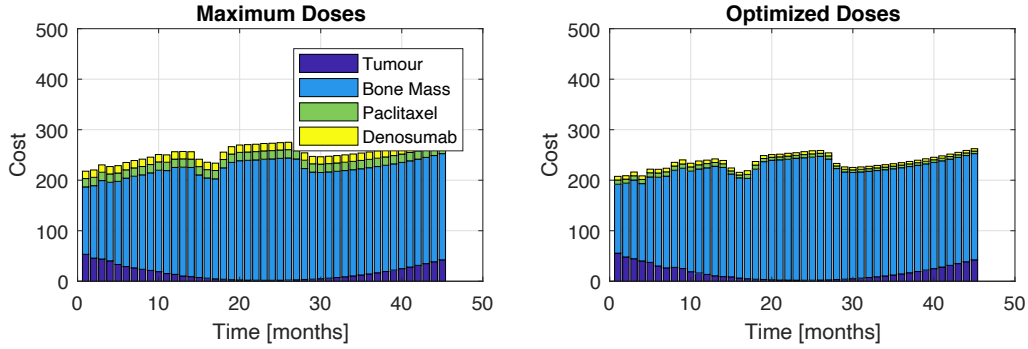


Figure 16. Proportions of the four cost contributions/terms in the objective function: Tumour, Bone Mass, Paclitaxel and PI (Model $\mathbf{M}_{DR2}$, Therapy $T_4 + T_3$). The right plot represents the cost resultant from optimization, while the left represents the cost if the maximum dose was administrated.

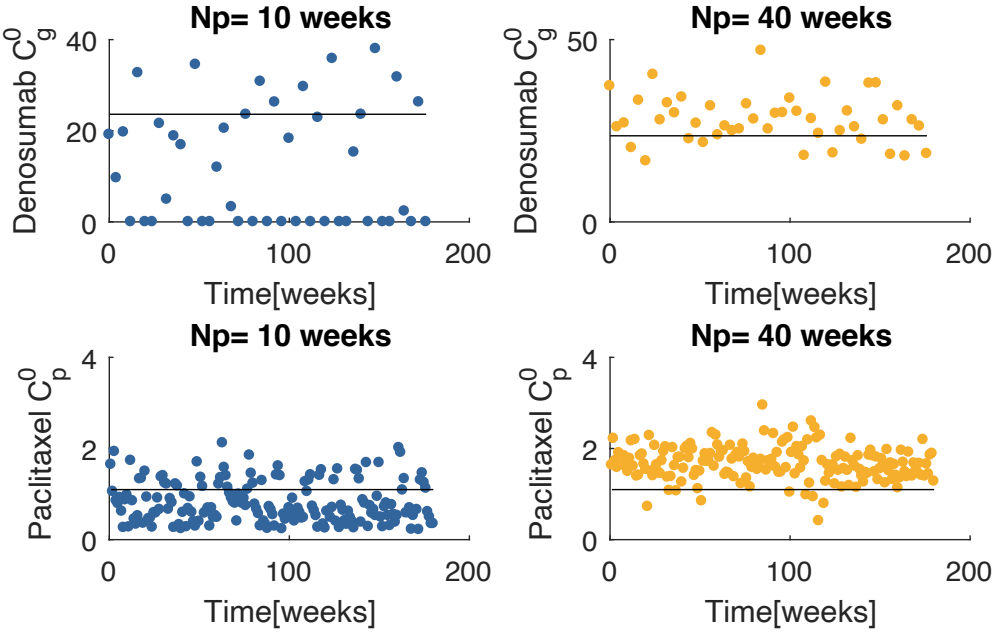
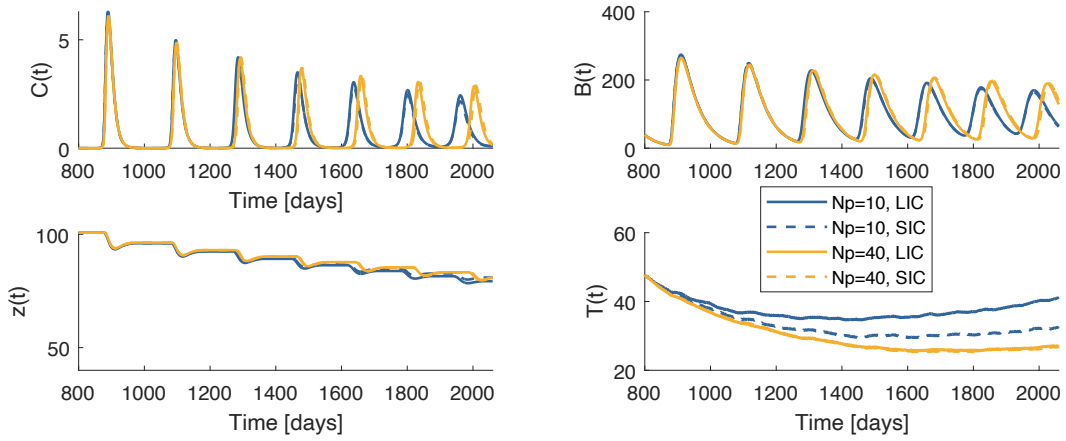
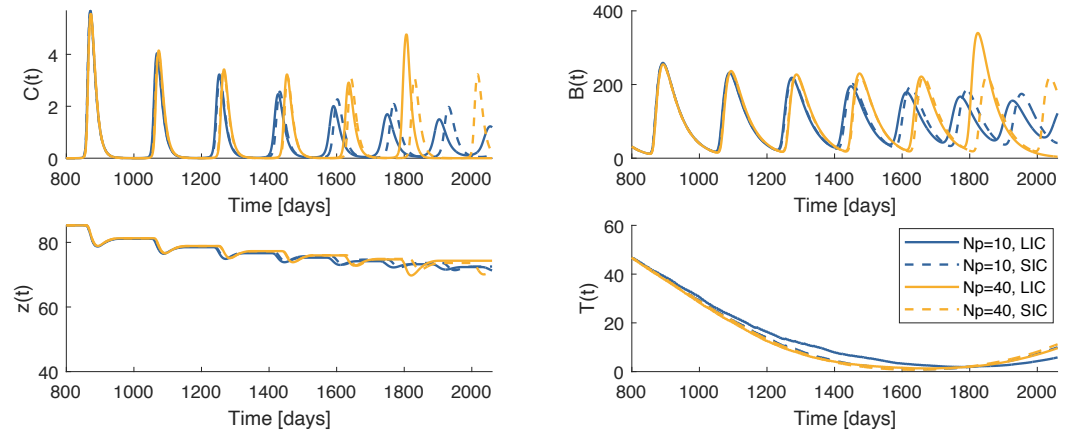


Figure 17. Best obtained prescriptions in mg with the model M_{DR1} (LIC).



(**a**) M_{DR1}



(**b**) M_{DR2}

Figure 18. Comparison between the system behaviour when the initial global best position of the PSO is the standard regimen (dashed lines) and a tenth of these values (solid lines). Legend: LIC—Low initial condition of the PSO, SIC—Standard initial condition of the PSO.

The prescription doses with LIC (Figure 17) appear not to have changed significantly, when compared with SIC (Figure 8). The only significant difference between these results is the paclitaxel administration when $N_p = 10$. The considerably lower doses resulted in poorer performance treating the tumour burden when compared to a standard initial condition, as one can verify in Figure 18a.

When handling M_{DR2}, the resultant prescriptions differences are much more pronounced (Figure 19). All of the dose values decrease with time, and this variation is significantly more accentuated with a low initial global best position, of both denosumab and paclitaxel. In fact, approximately in the last two thirds of the treatment period, almost all of the doses are far below the standard values. The input sequences obtained with SIC and LIC, although very disparate, produced quite similar results on the system behaviour (Figure 18b). The PSO fitness values evaluated when LIC is used are about 10% lower than those obtained with SIC when $N_p = 40$ but are almost equal when $N_p = 10$. The tumour decrease with the LIC prescription when $N_p = 10$ is not so fast, however the regrowth due to resistance is slightly postponed. Although this result translated in a higher cost associated to the tumour, the interpretation would be different if one attributed more importance to the late regrowth. In fact, this particular result supports the paradigm that says that the resistant population is delayed if the sensitive is not immediately eradicated.

One concludes that the optimization resultant input sequence with the model M_{DR2} is very sensitive to the initial global best position, contrarily to M_{DR1}. It seems more favourable to start the

global best to lower values, since the significant lower prescription doses are still enough to produce the same results as the SIC strategy. This analysis was not performed with the therapy $T_4 + T_3$ due to the excessive computational effort, but it is fair to generalize this conclusion.

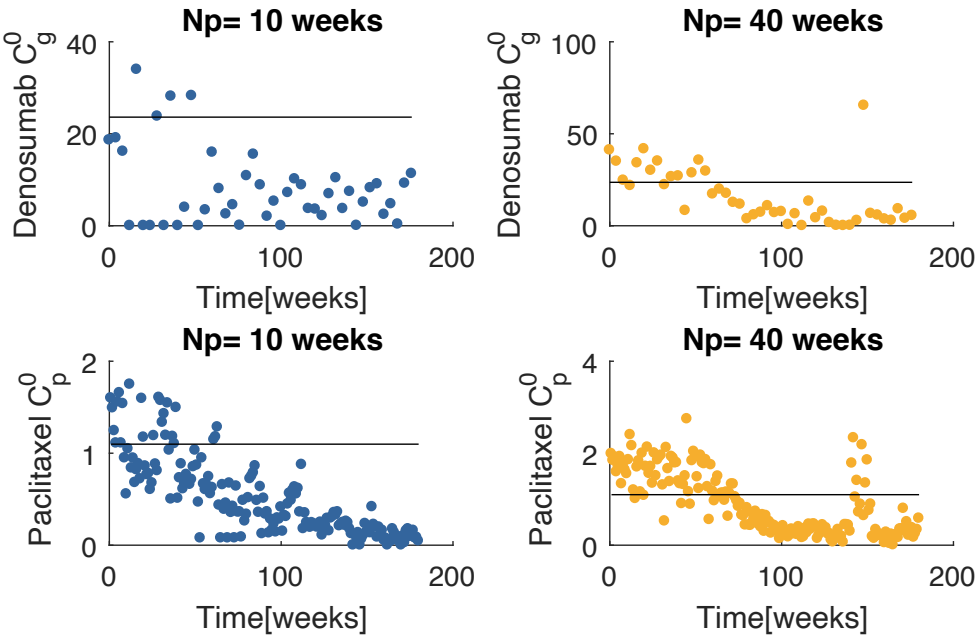


Figure 19. Best obtained prescriptions in mg with the model M_{DR2}.

7. Conclusions

The proposed model contains two major novelties: (1) the positive influence of the osteoclasts activity in the tumour proliferation, in order to recreate a vicious cycle between the two mechanisms, and (2) two drug resistance mechanisms regarding the immunity of tumour cells to paclitaxel, according to two different paradigms on the topic. The PK/PD of four drugs (Denosumab, BP, paclitaxel and PI) were included to simulate more accurately the system reaction to the therapy. Several factors regarding the OC and OB coupling and survival are yet to be considered and included in a more meticulous description of the system.

An NMPC with a PSO optimizer was the method to handle the problem, due to the severe nonlinearities of the system. Combining metaheuristics with MPC provided flexibility to design any type of cost function, ability to straightforwardly take the constraints into account and capability of solving the nonlinear problem. The developed adaptive algorithm (AD-PSO) is a novelty, and resulted from the combination of several existent variations of the PSO.

The drug dose optimization was performed on both resistance models and both proposed therapies. The healthy status was not achieved with any of the cases, due to the existence of drug resistance to the anti-cancer therapy. It is assumed that the M_{DR1} would require an impractically high and frequent dose to extinguish the tumour completely. The random mutation model was destined for therapy failure, due to the incontrovertible existence of a proliferative population, immune to paclitaxel. Once the sensitive cells number reduced to an apparent remission state, the administration of the anti-cancer therapy is decreased, while the resistant proliferates uncontrollably. The administration of a unique drug is insufficient to defeat the cancer burden.

When handling a two-populations model, it is desirable to initialize the PSO with low doses. The resultant input sequence tends to decrease to significantly lower values, although enough to keep the sensitive population from regrowth. When the paclitaxel dosage is too low, the performance of cancer treatment appears to be poorer based on cost. However, the tumour curve may be interpreted as favourable, if one attributes more importance to the fact that the regrowth is postponed. It was also

verified that the best obtained doses for both models safeguarded the patients from a high exposure to drugs, while outputting extremely close results to a regimen of intense maximal administration.

It is important to remark that the models are rough approximations of the reality. Several principles and variables are not taken into account, relatively to the type of patient and to the mechanisms involved. The difficulty of harvesting data for a wide range of conditions, specially regarding the origin of cancer (pre-diagnosis) and progression of untreated cancer (post-diagnosis), represents a barrier for the mathematical formulation and validation. Due to this lack of experimental data, the model was constructed and tuned based solely on the comparison between the qualitative behaviour and theoretical principles in the literature.

Future work includes adapting these methods of optimising cancer treatments for more accurate models, including mechanically induced bone remodelling [16], three-dimensional anomalous diffusion, modelled with fractional [41,42] and variable order derivatives [20,43,44], and additional biochemical interactions [4,17].

Author Contributions: Conceptualization, S.V. and D.V.; methodology, all authors; software, R.M.; validation, all authors; formal analysis, all authors; investigation, R.M.; resources, R.M.; data curation, R.M.; writing–original draft preparation, R.M.; writing–review and editing, all authors; visualization, R.M.; supervision, S.V. and D.V.; project administration, S.V.; funding acquisition, S.V. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by FCT, through IDMEC, under LAETA, project UID/EMS/50022/2020, through INESC–ID, project UIDB/50021/2020, and through project PERSEIDS, PTDC/EMS-SIS/0642/2014.

Acknowledgments: The authors would like to thank the collaboration of Hospital Santa Maria and IMM, and in particular Doctor Irina Alho for her help with the details of the biological processes addressed.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

BMU	basic multicellular unit
c-fms	Macrophage Colony-stimulating Factor Receptor
IGF	Insulin and Transforming Growth Factors
IL	interleukins
TGF	Transforming Growth Factor
M-CSF	Macrophage Colony-Stimulating Factor
MM	Multiple Myeloma
MPC	Model Predictive Control
MSC	Mesenchymal Stem Cells
NF-kB	nuclear factor kB
OB	Osteoblasts
OC	Osteoclasts
OPG	osteoprotegerin
PD	Pharmacodynamics
PK	Pharmacokinectics
PTH	parathyroid hormone
PTHrP	parathyroid hormone-related protein
PSO	Particle Swarm Optimization
RANK	Receptor Activator of Nuclear Factor kB
RANKL	NF-kB ligand
TNF	Tumour Necrosis Factors
VEGF	vascular endothelial growth factor

References

1. Araujo, R.P.; McElwain, D.S. A history of the study of solid tumour growth: The contribution of mathematical modelling. *Bull. Math. Biol.* **2004**, *66*, 1039–1091. [CrossRef] [PubMed]
2. Michor, F.; Beal, K. Improving cancer treatment via mathematical modeling: Surmounting the challenges is worth the effort. *Cell* **2015**, *163*, 1059–1063. [CrossRef] [PubMed]
3. Raggatt, L.J.; Partridge, N.C. Cellular and molecular mechanisms of bone remodeling. *J. Biol. Chem.* **2010**, *285*, 25103–25108. [CrossRef] [PubMed]
4. Coelho, R.M.; Lemos, J.M.; Alho, I.; Valério, D.; Ferreira, A.R.; Costa, L.; Vinga, S. Dynamic modeling of bone metastasis, microenvironment and therapy: Integrating parathyroid hormone (PTH) effect, anti-resorptive and anti-cancer therapy. *J. Theor. Biol.* **2016**, *391*, 1–12. [CrossRef]
5. Kular, J.; Tickner, J.; Chim, S.M.; Xu, J. An overview of the regulation of bone remodelling at the cellular level. *Clin. Biochem.* **2012**, *45*, 863–873. [CrossRef]
6. Teitelbaum, S.L. Bone resorption by osteoclasts. *Science* **2000**, *289*, 1504–1508. [CrossRef]
7. Hadjidakis, D.J.; Androulakis, I.I. Bone remodeling. *Ann. N. Y. Acad. Sci.* **2006**, *1092*, 385–396. [CrossRef]
8. Bartl, R.; Bartl, C. Control and Regulation of Bone Remodelling. In *Bone Disorders*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 31–38.
9. Martin, R. Toward a unifying theory of bone remodeling. *Bone* **2000**, *26*, 1–6. [CrossRef]
10. Khosla, S. Minireview: The OPG/RANKL/RANK system. *Endocrinology* **2001**, *142*, 5050–5055. [CrossRef]
11. Guise, T.A.; Mohammad, K.S.; Clines, G.; Stebbins, E.G.; Wong, D.H.; Higgins, L.S.; Vessella, R.; Corey, E.; Padalecki, S.; Suva, L.; et al. Basic mechanisms responsible for osteolytic and osteoblastic bone metastases. *Clin. Cancer Res.* **2006**, *12*, 6213s–6216s. [CrossRef]
12. Mundy, G.R. Metastasis: Metastasis to bone: Causes, consequences and therapeutic opportunities. *Nat. Rev. Cancer* **2002**, *2*, 584–593. [CrossRef] [PubMed]
13. Martin, T.J. Parathyroid hormone-related protein, its regulation of cartilage and bone development, and role in treating bone diseases. *Physiol. Rev.* **2016**, *96*, 831–871. [CrossRef] [PubMed]
14. Chen, Y.C.; Sosnoski, D.M.; Mastro, A.M. Breast cancer metastasis to the bone: Mechanisms of bone loss. *Breast Cancer Res.* **2010**, *12*, 215. [CrossRef] [PubMed]
15. Schmiedel, B.J.; Scheible, C.A.; Nuebling, T.; Kopp, H.G.; Wirths, S.; Azuma, M.; Schneider, P.; Jung, G.; Grosse-Hovest, L.; Salih, H.R. RANKL expression, function, and therapeutic targeting in multiple myeloma and chronic lymphocytic leukemia. *Cancer Res.* **2013**, *73*, 683–694. [CrossRef] [PubMed]
16. Liu, L.; Shi, Q.; Chen, Q.; Li, Z. Mathematical modeling of bone in-growth into undegradable porous periodic scaffolds under mechanical stimulus. *J. Tissue Eng.* **2019**, *10*, 2041731419827167. [CrossRef] [PubMed]
17. Coelho, R.M.; Neto, J.P.; Valério, D.; Vinga, S. Dynamic biochemical and cellular models of bone physiology: integrating remodelling processes, tumor growth and therapy. In *The Computational Mechanics of Bone Tissue*; Belinha, J., Manzanares-Céspedes, M.C., Completo, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2020. (In press)
18. Baldonedo, J.; Fernández, J.R.; Segade, A. Numerical Analysis of an Osseointegration Model. *Mathematics* **2020**, *8*, 87. [CrossRef]
19. Owen, R.; Reilly, G.C. In vitro Models of Bone Remodelling and Associated Disorders. *Front. Bioeng. Biotechnol.* **2018**, *6*, 134. [CrossRef]
20. Neto, J.P.; Valério, D.; Vinga, S. Variable order fractional derivatives and bone remodelling in the presence of metastases. In *Linear and Nonlinear Fractional Order Systems*; Azar, A.T., Radwan, A.G., Vaidyanathan, S., Eds.; Elsevier: Amsterdam, The Netherlands, 2018; Chapter 1, pp. 1–36.
21. Sieberath, A.; Bella, E.D.; Ferreira, A.M.; Gentile, P.; Eglin, D.; Dalgarno, K. A Comparison of Osteoblast and Osteoclast In Vitro Co-Culture Models and Their Translation for Preclinical Drug Testing Applications. *Int. J. Mol. Sci.* **2020**, *21*, 912. [CrossRef]
22. Ayati, B.P.; Edwards, C.M.; Webb, G.F.; Wikswo, J.P. A mathematical model of bone remodeling dynamics for normal bone cell populations and myeloma bone disease. *Biol. Direct* **2010**, *5*, 28. [CrossRef]
23. DiPiro, J.T. *Concepts in Clinical Pharmacokinetics*; ASH: Bethesda, MD, USA, 2010.

24. Miranda, R.; Valério, D.; Vinga, S. Bone Remodelling, Tumour Growth, and Fractional Order Therapy Predictive Control. In Proceedings of the International Conference on Fractional Differentiation and its Applications, Amman, Jordan, 16–18 July 2018. Available online: https://ssrn.com/abstract=3277347 (accessed on 25 February 2020).
25. Bassingthwaighte, J.B.; Butterworth, E.; Jardine, B.; Raymond, G.M. Compartmental modeling in the analysis of biological systems. In *Computational Toxicology*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 391–438.
26. Papandreou, C.N.; Daliani, D.D.; Nix, D.; Yang, H.; Madden, T.; Wang, X.; Pien, C.S.; Millikan, R.E.; Tu, S.M.; Pagliaro, L.; et al. Phase I trial of the proteasome inhibitor bortezomib in patients with advanced solid tumors with observations in androgen-independent prostate cancer. *J. Clin. Oncol.* **2004**, *22*, 2108–2121. [CrossRef]
27. Pinheiro, J.V.; Lemos, J.M.; Vinga, S. Nonlinear MPC of HIV-1 infection with periodic inputs. In Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC), Orlando, FL, USA, 12–15 December 2011; pp. 65–70.
28. Goldie, J.H.; Coldman, A.J. *Drug Resistance in Cancer: Mechanisms and Models*; Cambridge University Press: Cambridge, UK, 2009.
29. Monro, H.C.; Gaffney, E.A. Modelling chemotherapy resistance in palliation and failed cure. *J. Theor. Biol.* **2009**, *257*, 292–302. [CrossRef] [PubMed]
30. Camacho, E.F.; Alba, C.B. *Model Predictive Control*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
31. Grüne, L.; Pannek, J. Nonlinear model predictive control. In *Nonlinear Model Predictive Control*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 43–66.
32. Eberhart, R.; Kennedy, J. A new optimizer using particle swarm theory. In Proceedings of the Sixth International Symposium on Micro Machine and Human Science, Nagoya, Japan, 4–6 October 1995; pp. 39–43.
33. Kennedy, J. Particle swarm optimization. In *Encyclopedia of Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 760–766.
34. Coelho, J.; de Moura Oliveira, P.; Cunha, J.B. Greenhouse air temperature predictive control using the particle swarm optimisation algorithm. *Comput. Electron. Agric.* **2005**, *49*, 330–344. [CrossRef]
35. Mercieca, J.; Fabri, S.G. Particle swarm optimization for nonlinear model predictive control. In Proceedings of the Fifth International Conference on Advanced Engineering Computing and Applications in Science-ADVCOMP, Lisbon, Portugal, 20–21 November 2011; pp. 88–93.
36. Kaveh, A.; Zolghadr, A. Democratic PSO for truss layout and size optimization with frequency constraints. *Comput. Struct.* **2014**, *130*, 10–21. [CrossRef]
37. Taherkhani, M.; Safabakhsh, R. A novel stability-based adaptive inertia weight for particle swarm optimization. *Appl. Soft Comput.* **2016**, *38*, 281–295. [CrossRef]
38. Ratnaweera, A.; Halgamuge, S.K.; Watson, H.C. Self-organizing hierarchical particle swarm optimizer with time-varying acceleration coefficients. *IEEE Trans. Evol. Comput.* **2004**, *8*, 240–255. [CrossRef]
39. Valério, D.; Sá da Costa, J. Introduction to Single-Input, Single-Output Fractional Control. *IET Control Theory Appl.* **2011**, *5*, 1033–1057. [CrossRef]
40. Valério, D.; Sá da Costa, J. *An Introduction to Fractional Control*; Technical Report; IET: London, UK, 2013; ISBN 978-1-84919-545-4,
41. Christ, L.F.; Valério, D.; Coelho, R.M.; Vinga, S. Models of bone metastases and therapy using fractional derivatives. *J. Appl. Nonlinear Dyn.* **2018**, *7*, 81–94. [CrossRef]
42. Colli, P.; Gilardi, G.; Sprekels, J. A Distributed Control Problem for a Fractional Tumor Growth Model. *Mathematics* **2019**, *7*, 792. [CrossRef]
43. Neto, J.P.; Coelho, R.M.; Valério, D.; Vinga, S.; Sierociuk, D.; Malesza, W.; Macias, M.; Dzielinski, A. Simplifying biochemical tumorous bone remodeling models through variable order derivatives. *Comput. Math. Appl.* **2018**, *75*, 3147–3157. [CrossRef]
44. Valério, D.; Neto, J.; Vinga, S. Variable order 3D models of bone remodelling. *Bull. Pol. Acad. Sci. Tech. Sci.* **2019**, *67*, 501–508.

Computational and Applied Mathematics (2019) 38:177
https://doi.org/10.1007/s40314-019-0956-8

An iterative method for solving proximal split feasibility problems and fixed point problems

Wongvisarut Khuangsatung[1] · **Pachara Jailoka**[2] · **Suthep Suantai**[2]

Received: 4 February 2019 / Revised: 12 July 2019 / Accepted: 24 September 2019 /
Published online: 8 October 2019

Abstract
The purpose of this research is to introduce a regularized algorithm based on the viscosity method for solving the proximal split feasibility problem and the fixed point problem in Hilbert spaces. A strong convergence result of our proposed algorithm for finding a common solution of the proximal split feasibility problem and the fixed point problem for nonexpansive mappings is established. We also apply our main result to the split feasibility problem, and the fixed point problem of nonexpansive semigroups, respectively. Finally, we give numerical examples for supporting our main result.

Keywords Fixed point problems · Proximal split feasibility problems · Nonexpansive mappings

Mathematics Subject Classification 47H09 · 47H10

1 Introduction

Throughout this article, let H_1 and H_2 be two real Hilbert spaces. Let $f : H_1 \to \mathbb{R} \cup \{+\infty\}$ and $g : H_2 \to \mathbb{R} \cup \{+\infty\}$ be two proper and lower semicontinuous convex functions and $A : H_1 \to H_2$ be a bounded linear operator.

Communicated by Ernesto G. Birgin.

Suthep Suantai
suthep.s@cmu.ac.th

Wongvisarut Khuangsatung
wongvisarut_k@rmutt.ac.th

Pachara Jailoka
pachara.j4@gmail.com

[1] Department of Mathematics and Computer Science, Faculty of Science and Technology, Rajamangala University of Technology Thanyaburi (RMUTT), Thanyaburi, Pathumthani 12110, Thailand

[2] Data Science Research Center, Department of Mathematics, Faculty of Science, Chiang Mai University, Chiang Mai 50200, Thailand

In this paper, we focus our attention on the following proximal split feasibility problem (PSFP): find a minimizer x^* of f, such that Ax^* minimizes g, namely

$$x^* \in \operatorname{argmin} f \text{ such that } Ax^* \in \operatorname{argmin} g, \tag{1.1}$$

where $\operatorname{argmin} f := \{\bar{x} \in H_1 : f(\bar{x}) \leq f(x) \text{ for all } x \in H_1\}$ and $\operatorname{argmin} g := \{\bar{y} \in H_2 : g(\bar{y}) \leq g(y) \text{ for all } y \in H_2\}$. We assume that the problem (1.1) has a nonempty solution set $\Gamma := \operatorname{argmin} f \cap A^{-1}(\operatorname{argmin} g)$.

Censor and Elfving (1994) introduced the split feasibility problem (in short, SFP). The split feasibility problem (SFP) has been used for many applications in various fields of science and technology, such as in signal processing and image reconstruction, and especially applied in medical fields such as intensity-modulated radiation therapy (IMRT) (for details, see Censor et al. (2006) and the references therein). Let C and Q be nonempty, closed, and convex subsets of H_1 and H_2, respectively, and then, the SFP is to find a point:

$$x \in C \text{ such that } Ax \in Q, \tag{1.2}$$

where $A : H_1 \to H_2$ is a bounded linear operator. For solving the problem (1.2), Byrne (2002) introduced a popular algorithm which is called the CQ algorithm as follows:

$$x_{n+1} = P_C(x_n - \mu_n A^*(I - P_Q)Ax_n), \quad \forall n \geq 1,$$

where P_C and P_Q denote the metric projection onto the closed convex subsets C and Q, respectively, and A^* is the adjoint operator of A and $\mu_n \in (0, 2/\|A\|^2)$. Many research papers have increasingly investigated split feasibility problem [see, for instance (Lopez et al. 2012; Chang et al. 2014; Qu and Xiu 2005), and the references therein]. If $f = i_C$ [defined as $i_C(x) = 0$ if $x \in C$ and $i_C(x) = +\infty$ if $x \notin C$] and $g = i_Q$ are indicator functions of nonempty, closed, and convex sets C and Q of H_1 and H_2, respectively. Then, the proximal split feasibility problem (1.1) becomes the split feasibility problem (1.2).

Moudafi and Thakur (2014) introduced the split proximal algorithm with a way of selecting the step-sizes, such that its implementation does not need any prior information about the operator norm. Given an initial point $x_1 \in H_1$, assume that x_n has been constructed and $\|A^*(I - \operatorname{prox}_{\lambda g})Ax_n\|^2 + \|(I - \operatorname{prox}_{\lambda f})x_n\|^2 \neq 0$, and then compute x_{n+1} by the following iterative scheme:

$$x_{n+1} = \operatorname{prox}_{\lambda\mu_n f}(x_n - \mu_n A^*(I - \operatorname{prox}_{\lambda g})Ax_n), \quad \forall n \geq 1, \tag{1.3}$$

where the stepsize $\mu_n := \rho_n \dfrac{h(x_n) + l(x_n)}{\theta^2(x_n)}$ with $0 < \rho_n < 4$, $h(x) := \frac{1}{2}\|(I - \operatorname{prox}_{\lambda g})Ax\|^2$, $l(x) := \frac{1}{2}\|(I - \operatorname{prox}_{\lambda\mu_n f})x\|^2$ and $\theta^2(x) := \|A^*(I - \operatorname{prox}_{\lambda g})Ax\|^2 + \|(I - \operatorname{prox}_{\lambda\mu_n f})x\|^2$. If $\theta^2(x_n) = 0$, then x_n is a solution of (1.1) and the iterative process stops; otherwise, we set $n := n + 1$ and compute x_{n+1} using (1.3). They also proved the weak convergence of the sequence generated by Algorithm (1.3) to a solution of (1.1) under suitable conditions of parameter ρ_n where $\varepsilon \leq \rho_n \leq \dfrac{4h(x_n)}{h(x_n) + l(x_n)} - \varepsilon$ for some $\varepsilon > 0$.

Yao et al. (2014) gave the regularized algorithm for solving the proximal split feasibility problem (1.1) and proposed a strong convergence theorem under suitable conditions:

$$x_{n+1} = \operatorname{prox}_{\lambda\mu_n f}(\alpha_n u + (1 - \alpha_n)x_n - \mu_n A^*(I - \operatorname{prox}_{\lambda g})Ax_n), \quad \forall n \geq 1, \tag{1.4}$$

where the stepsize $\mu_n := \rho_n \dfrac{h(x_n) + l(x_n)}{\theta^2(x_n)}$ with $0 < \rho_n < 4$.

Shehu et al. (2015) introduced a viscosity-type algorithm for solving proximal split feasibility problems as follows:

$$\begin{cases} y_n = x_n - \mu_n A^*(I - \text{prox}_{\lambda g})Ax_n, \\ x_{n+1} = \alpha_n \psi(x_n) + (1-\alpha_n)\,\text{prox}_{\lambda\mu_n f}\, y_n, \quad \forall n \geq 1, \end{cases} \tag{1.5}$$

where $\psi : H_1 \to H_1$ is a contraction mapping. They also proved a strong convergence of the sequences generated by iterative schemes (1.5) in Hilbert spaces.

Recently, Shehu and Iyiola (2015) introduced the following algorithm for solving split proximal algorithms and fixed point problems for k-strictly pseudocontractive mappings in Hilbert spaces:

$$\begin{cases} u_n = (1-\alpha_n)x_n, \\ y_n = \text{prox}_{\lambda\gamma_n f}(u_n - \gamma_n A^*(I - \text{prox}_{\lambda g})Au_n), \\ x_{n+1} = (1-\beta_n)y_n + \beta_n T y_n, \quad \forall n \in \mathbb{N}, \end{cases} \tag{1.6}$$

where the stepsize $\gamma_n := \rho_n \dfrac{h(x_n) + l(x_n)}{\theta^2(x_n)}$ with $0 < \rho_n < 4$. They also showed that, under certain assumptions imposed on the parameters, the sequence $\{x_n\}$ generated by (1.6) converges strongly to $x^* \in F(S) \cap \Gamma$. Many researchers have proposed some methods to solve the proximal split feasibility problem [see, for instance (Shehu et al. 2015; Shehu and Iyiola 2017a, b, 2018; Abbas et al. 2018; Witthayarat et al. 2018), and the references therein].

We note that Algorithm (1.6) is the Halpern-type algorithm with $u \equiv 0$ fixed. However, a viscosity-type algorithm is more general and desirable than a Halpern-type algorithm, because a contraction which is used in the viscosity-type algorithm influences the convergence behavior of the algorithm.

In this paper, inspired and motivated by these studies, we are interested to study the proximal split feasibility problem and the fixed point problem in Hilbert spaces. In Sect. 3, we introduce a regularized algorithm based on the viscosity method for finding a common solution of the proximal split feasibility problem and the fixed point problem of nonexpansive mappings, and prove a strong convergence theorem under some suitable conditions. In Sects. 4 and 5, we apply our main result to the split feasibility problem, and the fixed point problem of nonexpansive semigroups, respectively. In the last section, we first give a numerical result in Euclidean spaces to demonstrate the convergence of our algorithm. We also show the number of iterations of our algorithm by choosing different contractions ψ. In this case, if we take $\psi = 0$ in our algorithm, then we obtain Algorithm (1.6) (Shehu and Iyiola 2015, Algorithm 1). Moreover, we give an example in the infinite-dimensional spaces for supporting our main theorem.

2 Preliminaries

Throughout this article, let H be a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$. Let C be a nonempty closed convex subset of H. Let $T : C \to C$ be a nonlinear mapping. A point $x \in C$ is called a fixed point of T if $Tx = x$. The set of fixed points of T is the set $F(T) := \{x \in C : Tx = x\}$.

Recall that A mapping T of C into itself is said to be

(i) nonexpansive if

$$\|Tx - Ty\| \leq \|x - y\|, \quad \forall x, y \in C.$$

(ii) contraction if there exists a constant $\delta \in [0, 1)$, such that

$$\|Tx - Ty\| \leq \delta \|x - y\|, \quad \forall x, y \in C.$$

Recall that the proximal operator $\text{prox}_{\lambda g} : H \to H$ is defined by:

$$\text{prox}_{\lambda g}\, x := \underset{u \in H}{\text{argmin}} \left\{ g(u) + \frac{1}{2\lambda} \|u - x\|^2 \right\}. \tag{2.1}$$

Moreover, the proximity operator of f is firmly nonexpansive, namely:

$$\left\langle \text{prox}_{\lambda g}(x) - \text{prox}_{\lambda g}(y), x - y \right\rangle \geq \| \text{prox}_{\lambda g}(x) - \text{prox}_{\lambda g}(y)\|^2; \tag{2.2}$$

for all $x, y \in H$, which is equivalent to

$$\| \text{prox}_{\lambda g}(x) - \text{prox}_{\lambda g}(y)\|^2 \leq \|x - y\|^2 - \|(I - \text{prox}_{\lambda g})(x) - (I - \text{prox}_{\lambda g})(y)\|^2. \tag{2.3}$$

for all $x, y \in H$. For general information on proximal operator, see Combettes and Pesquet (2011a).

In a real Hilbert space H, it is well known that:

(i) $\|\alpha x + (1 - \alpha)y\|^2 = \alpha \|x\|^2 + (1 - \alpha) \|y\|^2 - \alpha(1 - \alpha) \|x - y\|^2$, for all $x, y \in H$ and $\alpha \in [0, 1]$;
(ii) $\|x - y\|^2 = \|x\|^2 - 2\langle x, y\rangle + \|y\|^2$ for all $x, y \in H$;
(iii) $\|x + y\|^2 \leq \|x\|^2 + 2\langle y, x + y\rangle$ for all $x, y \in H$.

Recall that the (nearest-point) projection P_C from H onto C assigns to each $x \in H$ the unique point $P_C x \in C$ satisfying the property:

$$\|x - P_C x\| = \min_{y \in C} \|x - y\|.$$

Lemma 2.1 (Takahashi 2000) *Given $x \in H$ and $y \in C$. Then, $P_C x = y$ if and only if there holds the inequality:*

$$\langle x - y, y - z\rangle \geq 0, \quad \forall z \in C.$$

Lemma 2.2 (Xu 2003) *Let $\{s_n\}$ be a sequence of nonnegative real numbers satisfying:*

$$s_{n+1} = (1 - \alpha_n)s_n + \delta_n, \quad \forall n \geq 0,$$

where $\{\alpha_n\}$ is a sequence in $(0, 1)$ and $\{\delta_n\}$ is a sequence, such that

1. $\sum_{n=1}^{\infty} \alpha_n = \infty$;
2. $\limsup_{n \to \infty} \frac{\delta_n}{\alpha_n} \leq 0$ *or* $\sum_{n=1}^{\infty} |\delta_n| < \infty$.

Then, $\lim_{n \to \infty} s_n = 0$.

Definition 2.3 Let C be a nonempty closed convex subset of a real Hilbert space H. A mapping $S : C \to C$ is called demi-closed at zero if for any sequence $\{x_n\}$ which converges weakly to x, and if the sequence $\{Tx_n\}$ converges strongly to 0, then $Tx = 0$.

Lemma 2.4 (Browder 1976) *Let C be a nonempty closed convex subset of a real Hilbert space H. If $S : C \to C$ is a nonexpansive mapping, then $I - S$ is demi-closed at zero.*

Lemma 2.5 (Mainge 2008) *Let $\{\Gamma_n\}$ be a sequence of real numbers that does not decrease at infinity in the sense that there exists a subsequence $\{\Gamma_{n_i}\}$ of $\{\Gamma_n\}$ which satisfies $\Gamma_{n_i} < \Gamma_{n_i+1}$ for all $i \in \mathbb{N}$. Define the sequence $\{\tau(n)\}_{n\geq n_0}$ of integers as follows:*

$$\tau(n) = \max\{k \leq n : \Gamma_k < \Gamma_{k+1}\},$$

where $n_0 \in \mathbb{N}$, such that $\{k \leq n_0 : \Gamma_k < \Gamma_{k+1}\} \neq \emptyset$. Then, the following hold:

(i) $\tau(n_0) \leq \tau(n_0+1) \leq \cdots$ *and* $\tau(n) \longrightarrow \infty$;
(ii) $\Gamma_{\tau_n} \leq \Gamma_{\tau(n)+1}$ *and* $\Gamma_n \leq \Gamma_{\tau(n)+1}$, $\forall n \geq n_0$.

3 Main results

In this section, we introduce an algorithm and prove a strong convergence for solving a common element of the set of fixed points of a nonexpansive mapping and the set of solutions of proximal split feasibility problems (1.1). Let H_1 and H_2 be two real Hilbert spaces. Let $f : H_1 \to \mathbb{R} \cup \{+\infty\}$ and $g : H_2 \to \mathbb{R} \cup \{+\infty\}$ be two proper and lower semicontinuous convex functions and $A : H_1 \to H_2$ be a bounded linear operator. Let $S : H_1 \to H_1$ be a nonexpansive mapping and Let $\psi : H_1 \to H_1$ be a contraction mapping with $\delta \in (0, 1)$.

We introduce the modified split proximal algorithm as follows:

Algorithm 3.1 Given an initial point $x_1 \in H_1$. Assume that x_n has been constructed and $\|A^*(I - \text{prox}_{\lambda g})Ax_n\|^2 + \|(I - \text{prox}_{\lambda f})x_n\|^2 \neq 0$, then compute x_{n+1} by the following iterative scheme:

$$\begin{cases} y_n = \text{prox}_{\lambda\mu_n f}(\alpha_n\psi(x_n) + (1-\alpha_n)x_n - \mu_n A^*(I - \text{prox}_{\lambda g})Ax_n) \\ x_{n+1} = \beta_n y_n + (1-\beta_n)Sy_n, \quad \forall n \in \mathbb{N}, \end{cases} \tag{3.1}$$

where the stepsize $\mu_n := \rho_n \dfrac{\left(\frac{1}{2}\|(I - \text{prox}_{\lambda g})Ax_n\|^2\right) + \left(\frac{1}{2}\|(I - \text{prox}_{\lambda f})x_n\|^2\right)}{\|A^*(I - \text{prox}_{\lambda g})Ax_n\|^2 + \|(I - \text{prox}_{\lambda f})x_n\|^2}$ with $0 < \rho_n < 4$ and $\{\alpha_n\}, \{\beta_n\} \subset (0, 1)$.

We now prove our main theorem.

Theorem 3.2 *Let H_1 and H_2 be two real Hilbert spaces. Let $f : H_1 \to \mathbb{R} \cup \{+\infty\}$ and $g : H_2 \to \mathbb{R} \cup \{+\infty\}$be two proper and lower semicontinuous convex functions, and $A : H_1 \to H_2$ be a bounded linear operator. Let $\psi : H_1 \to H_1$ be a contraction mapping with $\delta \in [0, 1)$ and let $S : H_1 \to H_1$ be a nonexpansive mapping, such that $\Omega := F(S) \cap \Gamma \neq 0$. If the control sequences $\{\alpha_n\}$, $\{\beta_n\}$ and $\{\rho_n\}$ satisfy the following conditions:*

(C1) $\lim_{n\to\infty} \alpha_n = 0$ *and* $\sum_{n=1}^{\infty} \alpha_n = \infty$;
(C2) $0 < \liminf_{n\to\infty} \beta_n \leq \limsup_{n\to\infty} \beta_n < 1$;
(C3) $\varepsilon \leq \rho_n \leq \dfrac{4(1-\alpha_n)\left(\|(I - \text{prox}_{\lambda g})Ax_n\|^2\right)}{\left(\|(I - \text{prox}_{\lambda g})Ax_n\|^2\right) + \left(\|(I - \text{prox}_{\lambda f})x_n\|^2\right)} - \varepsilon$ *for some* $\varepsilon > 0$.

Then, the sequence $\{x_n\}$ defined by Algorithm 3.1 converges strongly to a point $x^* \in \Omega$ which also solves the variational inequality:

$$\langle(\psi - I)x^*, x - x^*\rangle \leq 0, \quad \forall x \in \Omega.$$

Proof Given any $\lambda > 0$ and $x \in H_1$, we define $h(x) := \frac{1}{2}\|(I - \text{prox}_{\lambda g})Ax\|^2$, $l(x) := \frac{1}{2}\|(I - \text{prox}_{\lambda f})x\|^2$, $\theta^2(x) := \|A^*(I - \text{prox}_{\lambda g})Ax\|^2 + \|(I - \text{prox}_{\lambda f})x\|^2$, and hence, $\mu_n = \rho_n \dfrac{h(x_n) + l(x_n)}{\theta^2(x_n)}$ where $0 < \rho_n < 4$.

By Banach fixed point theorem, there exists $x^* \in \Omega$ such that $x^* = P_\Omega \psi(x^*)$. Then, $x^* = \text{prox}_{\lambda\mu_n f}\, x^*$ and $Ax^* = \text{prox}_{\lambda g}\, Ax^*$. Since $\text{prox}_{\lambda g}$ is firmly nonexpansive, we have $I - \text{prox}_{\lambda g}$ is also firmly nonexpansive. Hence

$$\begin{aligned}\langle A^*(I - \text{prox}_{\lambda g})Ax_n, x_n - x^*\rangle &= \langle (I - \text{prox}_{\lambda g})Ax_n, Ax_n - Ax^*\rangle \\ &= \left\langle (I - \text{prox}_{\lambda g})Ax_n - (I - \text{prox}_{\lambda g})Ax^*, Ax_n - Ax^*\right\rangle \\ &\geq \|(I - \text{prox}_{\lambda g})Ax_n\|^2 = 2h(x_n). \end{aligned} \tag{3.2}$$

From the definition of y_n and the nonexpansivity of $\text{prox}_{\lambda\mu_n f}$, we have:

$$\begin{aligned}\|y_n - x^*\| &= \|\,\text{prox}_{\lambda\mu_n f}(\alpha_n\psi(x_n) + (1-\alpha_n)x_n - \mu_n A^*(I - \text{prox}_{\lambda g})Ax_n) - x^*\| \\ &\leq \|\alpha_n\psi(x_n) + (1-\alpha_n)x_n - \mu_n A^*(I - \text{prox}_{\lambda g})Ax_n - x^*\| \\ &\leq \alpha_n\|\psi(x_n) - x^*\| + (1-\alpha_n)\left\|x_n - \frac{\mu_n}{(1-\alpha_n)}A^*(I - \text{prox}_{\lambda g})Ax_n - x^*\right\|. \end{aligned} \tag{3.3}$$

From (3.2), we have:

$$\begin{aligned}&\left\|x_n - \frac{\mu_n}{(1-\alpha_n)}A^*(I - \text{prox}_{\lambda g})Ax_n - x^*\right\|^2 \\ &= \|x_n - x^*\|^2 + \frac{\mu_n^2}{(1-\alpha_n)^2}\|A^*(I - \text{prox}_{\lambda g})Ax_n\|^2 \\ &\quad - 2\frac{\mu_n}{(1-\alpha_n)}\langle A^*(I - \text{prox}_{\lambda g})Ax_n, x_n - x^*\rangle \\ &\leq \|x_n - x^*\|^2 + \frac{\mu_n^2}{(1-\alpha_n)^2}\|A^*(I - \text{prox}_{\lambda g})Ax_n\|^2 - 4\frac{\mu_n}{(1-\alpha_n)}h(x_n) \\ &= \|x_n - x^*\|^2 + \rho_n^2\frac{(h(x_n)+l(x_n))^2}{(1-\alpha_n)^2\theta^4(x_n)}\|A^*(I - \text{prox}_{\lambda g})Ax_n\|^2 - 4\rho_n\frac{(h(x_n)+l(x_n))}{(1-\alpha_n)\theta^2(x_n)}h(x_n) \\ &\leq \|x_n - x^*\|^2 + \rho_n^2\frac{(h(x_n)+l(x_n))^2}{(1-\alpha_n)^2\theta^2(x_n)} - 4\rho_n\frac{(h(x_n)+l(x_n))^2}{(1-\alpha_n)\theta^2(x_n)}\frac{h(x_n)}{(h(x_n)+l(x_n))} \\ &= \|x_n - x^*\|^2 - \rho_n\left(\frac{4h(x_n)}{(h(x_n)+l(x_n))} - \frac{\rho_n}{1-\alpha_n}\right)\left(\frac{(h(x_n)+l(x_n))^2}{(1-\alpha_n)\theta^2(x_n)}\right). \end{aligned} \tag{3.4}$$

By the condition (C3), we have $\dfrac{4h(x_n)}{(h(x_n)+l(x_n))} - \dfrac{\rho_n}{1-\alpha_n} \geq 0$ for all $n \geq 1$. From (3.3) and (3.4), we have:

$$\begin{aligned}\|y_n - x^*\| &\leq \alpha_n\|\psi(x_n) - x^*\| + (1-\alpha_n)\left\|x_n - \frac{\mu_n}{(1-\alpha_n)}A^*(I - \text{prox}_{\lambda g})Ax_n - x^*\right\| \\ &\leq \alpha_n\|\psi(x_n) - \psi(x^*)\| + \alpha_n\|\psi(x^*) - x^*\| + (1-\alpha_n)\left\|x_n - x^*\right\| \\ &\leq \alpha_n\delta\|x_n - x^*\| + \alpha_n\|\psi(x^*) - x^*\| + (1-\alpha_n)\left\|x_n - x^*\right\| \\ &= (1 - \alpha_n(1-\delta))\|x_n - x^*\| + \alpha_n\|\psi(x^*) - x^*\|. \end{aligned} \tag{3.5}$$

Since S is nonexpansive, by (3.1) and (3.5), we obtain:

$$\begin{aligned}\|x_{n+1}-x^*\| &= \|\beta_n y_n + (1-\beta_n)Sy_n - x^*\| \\ &\le \beta_n\|y_n - x*\| + (1-\beta_n)\|Sy_n - x^*\| \\ &\le \beta_n\|y_n - x*\| + (1-\beta_n)\|y_n - x^*\| \\ &= \|y_n - x*\| \\ &\le (1-\alpha_n(1-\delta))\|x_n - x^*\| + \alpha_n\|\psi(x^*) - x^*\| \\ &\le \max\left\{\|x_n - x^*\|, \frac{\|\psi(x^*) - x^*\|}{1-\delta}\right\}.\end{aligned}$$

By mathematical induction, we have:

$$\|x_n - x^*\| \le \max\left\{\|x_1 - x^*\|, \frac{\|\psi(x^*) - x^*\|}{1-\delta}\right\}, \quad \forall n \in \mathbb{N}.$$

Hence, $\{x_n\}$ is bounded and so are $\{\psi(x_n)\}$, $\{Sy_n\}$.

From the definition of y_n and (3.4), we have:

$$\begin{aligned}\|y_n - x^*\|^2 &= \|\operatorname{prox}_{\lambda\mu_n f}(\alpha_n\psi(x_n) + (1-\alpha_n)x_n - \mu_n A^*(I - \operatorname{prox}_{\lambda g})Ax_n) - x^*\|^2 \\ &\le \|\alpha_n\psi(x_n) + (1-\alpha_n)x_n - \mu_n A^*(I - \operatorname{prox}_{\lambda g})Ax_n - x^*\|^2, \\ &\le \alpha_n\|\psi(x_n) - x^*\|^2 + (1-\alpha_n)\left\|x_n - \frac{\mu_n}{(1-\alpha_n)}A^*(I - \operatorname{prox}_{\lambda g})Ax_n - x^*\right\|^2 \\ &\le \alpha_n\|\psi(x_n) - x^*\|^2 + (1-\alpha_n) \\ &\quad\times\left(\|x_n - x^*\|^2 - \rho_n\left(\frac{4h(x_n)}{(h(x_n)+l(x_n))} - \frac{\rho_n}{1-\alpha_n}\right)\left(\frac{(h(x_n)+l(x_n))^2}{(1-\alpha_n)\theta^2(x_n)}\right)\right) \\ &= \alpha_n\|\psi(x_n) - x^*\|^2 + (1-\alpha_n)\|x_n - x^*\|^2 \\ &\quad - \rho_n\left(\frac{4h(x_n)}{(h(x_n)+l(x_n))} - \frac{\rho_n}{1-\alpha_n}\right)\left(\frac{(h(x_n)+l(x_n))^2}{\theta^2(x_n)}\right). \qquad (3.6)\end{aligned}$$

From the definition of x_n and (3.6), we obtain:

$$\begin{aligned}\|x_{n+1}-x^*\|^2 &= \|\beta_n y_n + (1-\beta_n)Sy_n - x^*\|^2 \\ &\le \beta_n\|y_n - x*\|^2 + (1-\beta_n)\|Sy_n - x^*\|^2 \\ &\le \|y_n - x*\|^2 \\ &\le \alpha_n\|\psi(x_n) - x^*\|^2 + (1-\alpha_n)\|x_n - x^*\|^2 \\ &\quad - \rho_n\left(\frac{4h(x_n)}{(h(x_n)+l(x_n))} - \frac{\rho_n}{1-\alpha_n}\right)\left(\frac{(h(x_n)+l(x_n))^2}{\theta^2(x_n)}\right) \\ &\le \alpha_n\|\psi(x_n) - x^*\|^2 + \|x_n - x^*\|^2 \\ &\quad - \rho_n\left(\frac{4h(x_n)}{(h(x_n)+l(x_n))} - \frac{\rho_n}{1-\alpha_n}\right)\left(\frac{(h(x_n)+l(x_n))^2}{\theta^2(x_n)}\right).\end{aligned}$$

It implies that

$$\begin{aligned}\rho_n\left(\frac{4h(x_n)}{(h(x_n)+l(x_n))} - \frac{\rho_n}{1-\alpha_n}\right)\left(\frac{(h(x_n)+l(x_n))^2}{\theta^2(x_n)}\right) &\le \alpha_n\|\psi(x_n) - x^*\|^2 + \|x_n \\ &\quad - x^*\|^2 - \|x_{n+1} - x^*\|^2. \qquad (3.7)\end{aligned}$$

It follows from (3.6) that

$$\begin{aligned}\|x_{n+1}-x^*\|^2 &= \|\beta_n y_n + (1-\beta_n)Sy_n - x^*\|^2\\ &\leq \beta_n\|y_n - x*\|^2 + (1-\beta_n)\|Sy_n - x^*\|^2 - \beta_n(1-\beta_n)\|y_n - Sy_n\|^2\\ &\leq \|y_n - x*\|^2 - \beta_n(1-\beta_n)\|y_n - Sy_n\|^2\\ &\leq \alpha_n\|\psi(x_n) - x^*\|^2 + (1-\alpha_n)\|x_n - x^*\|^2 - \beta_n(1-\beta_n)\|y_n - Sy_n\|^2\\ &\leq \alpha_n\|\psi(x_n) - x^*\|^2 + \|x_n - x^*\|^2 - \beta_n(1-\beta_n)\|y_n - Sy_n\|^2,\end{aligned}$$

which implies that

$$\beta_n(1-\beta_n)\|y_n - Sy_n\|^2 \leq \alpha_n\|\psi(x_n) - x^*\|^2 + \|x_n - x^*\|^2 - \|x_{n+1} - x^*\|^2. \tag{3.8}$$

Now, we divide our proof into two cases.

Case 1 Suppose that there exists $n_0 \in \mathbb{N}$, such that $\{\|x_n - x^*\|\}_{n=1}^{\infty}$ is nonincreasing. Then, $\{\|x_n - x^*\|\}_{n=1}^{\infty}$ converges and $\|x_n - x^*\|^2 - \|x_{n+1} - x^*\|^2 \to 0$ as $n \to \infty$. From (3.7) and the condition (C1) and (C3), we obtain:

$$\rho_n\left(\frac{4h(x_n)}{(h(x_n)+l(x_n))} - \frac{\rho_n}{1-\alpha_n}\right)\left(\frac{(h(x_n)+l(x_n))^2}{\theta^2(x_n)}\right) \to 0 \text{ as } n \to \infty.$$

Hence, we have:

$$\frac{(h(x_n)+l(x_n))^2}{\theta^2(x_n)} \to 0 \text{ as } n \to \infty. \tag{3.9}$$

By the linearity and boundedness of A and the nonexpansivity of $\text{prox}_{\lambda g}$, we obtain that $\{\theta^2(x_n)\}$ is bounded.

It follows that

$$\lim_{n\to\infty}\left((h(x_n)+l(x_n))^2\right) = 0,$$

which implies that

$$\lim_{n\to\infty} h(x_n) = \lim_{n\to\infty} l(x_n) = 0.$$

Next, we show that $\limsup_{n\to\infty}\langle\psi(x^*) - x^*, x_n - x^*\rangle \leq 0$, where $x^* = P_\Omega\psi(x^*)$. Since $\{x_n\}$ is bounded, there exists a subsequence $\{x_{n_j}\}$ of $\{x_n\}$ satisfying $x_{n_j} \rightharpoonup \omega$ and

$$\limsup_{n\to\infty}\langle\psi(x^*) - x^*, x_n - x^*\rangle = \lim_{j\to\infty}\langle\psi(x^*) - x^*, x_{n_j} - x^*\rangle. \tag{3.10}$$

By the lower semicontinuity of h, we have:

$$0 \leq h(\omega) \leq \liminf_{j\to\infty} h(x_{n_j}) = \lim_{n\to\infty} h(x_n) = 0.$$

Therefore, $h(\omega) = \frac{1}{2}\|(I - \text{prox}_{\lambda g})A\omega\|^2 = 0$. Therefore, $A\omega$ is a fixed point of the proximal mapping of g or equivalently, $A\omega$ is a minimizer of g. Similarly, from the lower semicontinuity of l, we obtain:

$$0 \leq l(\omega) \leq \liminf_{j\to\infty} l(x_{n_j}) = \lim_{n\to\infty} l(x_n) = 0.$$

Therefore, $l(\omega) = \frac{1}{2}\|(I - \text{prox}_{\lambda\mu_n f})\omega\|^2 = 0$. That is $\omega \in F(\text{prox}_{\lambda\mu_n f})$. Then ω is a minimizer of f. Thus, $\omega \in \Gamma$. We observe that

$$0 < \mu_n < 4\frac{h(x_n)+l(x_n)}{\theta^2(x_n)} \to 0 \text{ as } n \to \infty,$$

and hence, $\mu_n \to 0$ as $n \to \infty$.

Next, we show that $\omega \in F(S)$. From (3.8) and the condition (C1), (C2), we have:

$$\|y_n - Sy_n\| \to 0 \text{ as } n \to \infty. \tag{3.11}$$

For each $n \geq 1$, let $u_n := \alpha_n \psi(x_n) + (1 - \alpha_n)x_n$. Then

$$\begin{aligned} \|u_n - x_n\| &= \|\alpha_n \psi(x_n) + (1 - \alpha_n)x_n - x_n\| \\ &= \alpha_n \|\psi(x_n) - x_n\|. \end{aligned}$$

From the condition (C1), we have:

$$\lim_{n\to\infty} \|u_n - x_n\| = 0. \tag{3.12}$$

Observe that

$$\|u_n - \text{prox}_{\lambda\mu_n f}\, x_n\| \leq \|u_n - x_n\| + \|(I - \text{prox}_{\lambda\mu_n f})x_n\|.$$

From $\lim_{n\to\infty} l(x_n) = \lim_{n\to\infty} \frac{1}{2}\|(I - \text{prox}_{\lambda\mu_n f})x_n\|^2 = 0$ and (3.12), we have:

$$\lim_{n\to\infty} \|u_n - \text{prox}_{\lambda\mu_n f}\, x_n\| = 0. \tag{3.13}$$

By the nonexpansiveness of $\text{prox}_{\lambda\mu_n f}$, we have:

$$\begin{aligned} \|y_n - \text{prox}_{\lambda\mu_n f}\, x_n\| &= \left\|\text{prox}_{\lambda\mu_n f}\left(u_n - \mu_n A^*\left(I - \text{prox}_{\lambda g}\right)Ax_n\right) - \text{prox}_{\lambda\mu_n f}\, x_n\right\| \\ &\leq \|u_n - \mu_n A^*(I - \text{prox}_{\lambda g})Ax_n - x_n\| \\ &\leq \|u_n - x_n\| + \mu_n \|A^*(I - \text{prox}_{\lambda g})Ax_n\|. \end{aligned}$$

From (3.13) and $\mu_n \to 0$ as $n \to \infty$, we have:

$$\lim_{n\to\infty} \|y_n - \text{prox}_{\lambda\mu_n f}\, x_n\| = 0. \tag{3.14}$$

Since

$$\|y_n - u_n\| \leq \|y_n - \text{prox}_{\lambda\mu_n f}\, x_n\| + \|u_n - \text{prox}_{\lambda\mu_n f}\, x_n\|,$$

from (3.13) and (3.14), we obtain:

$$\lim_{n\to\infty} \|y_n - u_n\| = 0. \tag{3.15}$$

From (3.12) and (3.15), we obtain

$$\lim_{n\to\infty} \|y_n - x_n\| = 0. \tag{3.16}$$

From

$$\|Sy_n - x_n\| \leq \|Sy_n - y_n\| + \|y_n - x_n\|,$$

by (3.11), (3.16), we get:

$$\lim_{n\to\infty} \|Sy_n - x_n\| = 0. \tag{3.17}$$

From the definition of x_n, we have:

$$\|x_{n+1} - x_n\| \leq \beta_n \|y_n - x_n\| + (1 - \beta_n)\|Sy_n - x_n\|.$$

This implies from (3.16), and (3.17) that

$$\lim_{n\to\infty} \|x_{n+1} - x_n\| = 0. \tag{3.18}$$

Using $x_{n_j} \rightharpoonup \omega \in H_1$ and (3.16), we obtain $y_{n_j} \rightharpoonup \omega \in H_1$. Since $y_{n_j} \rightharpoonup \omega \in H_1$, $\|y_n - Sy_n\| \to 0$ as $n \to \infty$, by Lemma 2.4, we have $\omega \in F(S)$. Hence, $\omega \in \mathcal{F} = F(S) \cap \Gamma$. Since $x_{n_j} \rightharpoonup \omega$ as $j \to \infty$ and $\omega \in \mathcal{F}$, by Lemma 2.1, we have:

$$\begin{aligned}\limsup_{n\to\infty} \langle \psi(x^*) - x^*, x_n - x^* \rangle &= \lim_{j\to\infty} \langle \psi(x^*) - x^*, x_{n_j} - x^* \rangle \\ &= \langle (\psi - I)x^*, \omega - x^* \rangle \\ &\leq 0. \end{aligned} \tag{3.19}$$

Now, by the nonexpansiveness of S and $\text{prox}_{\lambda\mu_n f}$, and from (3.1) and (3.4), we have:

$$\begin{aligned}\|x_{n+1} - x^*\|^2 &\leq \beta_n \|y_n - x^*\|^2 + (1-\beta_n)\|Sy_n - x^*\|^2 \leq \|y_n - x^*\|^2 \\ &\leq \|\alpha_n \psi(x_n) + (1-\alpha_n)x_n - \mu_n A^*(I - \text{prox}_{\lambda g})Ax_n - x^*\|^2 \\ &= (1-\alpha_n)^2 \|x_n - \frac{\mu_n}{(1-\alpha_n)} A^*(I - \text{prox}_{\lambda g})Ax_n - x^*\|^2 + \alpha_n^2 \|\psi(x_n) - x^*\|^2 \\ &\quad + 2\alpha_n(1-\alpha_n)\left\langle \psi(x_n) - x^*, x_n - \frac{\mu_n}{(1-\alpha_n)} A^*(I - \text{prox}_{\lambda g})Ax_n - x^* \right\rangle \\ &\leq (1-\alpha_n)^2 \|x_n - x^*\|^2 + \alpha_n^2 \|\psi(x_n) - x^*\|^2 \\ &\quad + 2\alpha_n(1-\alpha_n)\langle \psi(x_n) - x^*, x_n - x^* \rangle \\ &\quad - 2\alpha_n\mu_n \langle \psi(x_n) - x^*, A^*(I - \text{prox}_{\lambda g})Ax_n \rangle \\ &= (1-\alpha_n)^2 \|x_n - x^*\|^2 + \alpha_n^2 \|\psi(x_n) - x^*\|^2 \\ &\quad + 2\alpha_n(1-\alpha_n)\langle \psi(x_n) - \psi(x^*), x_n - x^* \rangle \\ &\quad + 2\alpha_n(1-\alpha_n)\langle \psi(x^*) - x^*, x_n - x^* \rangle \\ &\quad + 2\alpha_n\mu_n \langle x^* - \psi(x_n), A^*(I - \text{prox}_{\lambda g})Ax_n \rangle \\ &\leq (1 - 2\alpha_n + \alpha_n^2)\|x_n - x^*\|^2 + \alpha_n^2 \|\psi(x_n) - x^*\|^2 \\ &\quad + 2\alpha_n(1-\alpha_n)\delta \|x_n - x^*\|^2 \\ &\quad + 2\alpha_n(1-\alpha_n)\langle \psi(x^*) - x^*, x_n - x^* \rangle \\ &\quad + 2\alpha_n\mu_n \|\psi(x_n) - x^*\| \|A^*(I - \text{prox}_{\lambda g})Ax_n\| \\ &= (1 - 2\alpha_n + \alpha_n^2 + 2\alpha_n(1-\alpha_n)\delta)\|x_n - x^*\|^2 + \alpha_n^2 \|\psi(x_n) - x^*\|^2 \\ &\quad + 2\alpha_n(1-\alpha_n)\langle \psi(x^*) - x^*, x_n - x^* \rangle \\ &\quad + 2\alpha_n\mu_n \|\psi(x_n) - x^*\| \|A^*(I - \text{prox}_{\lambda g})Ax_n\| \\ &= (1-\epsilon_n)\|x_n - x^*\|^2 + \epsilon_n \xi_n, \end{aligned} \tag{3.20}$$

where $\epsilon_n = \alpha_n(2 - \alpha_n - 2(1-\alpha_n)\delta)$ and

$$\xi_n = \left[\frac{\alpha_n \|\psi(x_n) - x^*\|^2 + 2(1-\alpha_n)\langle \psi(x^*) - x^*, x_n - x^* \rangle + 2\mu_n \|A^*(I - \text{prox}_{\lambda g})Ax_n\| \|\psi(x_n) - x^*\|}{2 - \alpha_n - 2(1-\alpha_n)\delta} \right].$$

Note that $\mu_n \|A^*(I - \text{prox}_{\lambda g})Ax_n\| = \rho_n \frac{h(x_n) + l(x_n)}{\theta^2(x_n)} \|A^*(I - \text{prox}_{\lambda g})Ax_n\|$. Thus, $\mu_n \|A^*(I - \text{prox}_{\lambda g})Ax_n\| \to 0$ as $n \to \infty$. From the condition (C1), (3.19), (3.20) and Lemma 2.2, we can conclude that the sequence $\{x_n\}$ converges strongly to x^*.

Case 2 Assume that $\{\|x_n - x^*\|\}$ is not monotonically decreasing sequence. Then, there exists a subsequence n_l of n, such that $\|x_{n_l} - x^*\| < \|x_{n_l+1} - x^*\|$ for all $l \in \mathbb{N}$. Now, we define a positive integer sequence $\tau(n)$ by:

$$\tau(n) := \max\left\{k \in \mathbb{N} : k \leq n, \|x_{n_l} - x^*\| < \|x_{n_l+1} - x^*\|\right\}.$$

for all $n \geq n_0$ (for some n_0 large enough). By Lemma 2.5, we have τ which is a non-decreasing sequence, such that $\tau(n) \to \infty$ as $n \to \infty$ and

$$\|x_{\tau(n)} - x^*\|^2 - \|x_{\tau(n)+1} - x^*\|^2 \leq 0, \quad \forall n \geq n_0.$$

By a similar argument as that of case 1, we can show that

$$\rho_{\tau(n)}\left(\frac{4h(x_{\tau(n)})}{(h(x_{\tau(n)}) + l(x_{\tau(n)}))} - \frac{\rho_{\tau(n)}}{1-\alpha_{\tau(n)}}\right)\left(\frac{(h(x_{\tau(n)}) + l(x_{\tau(n)}))^2}{\theta^2(x_{\tau(n)})}\right) \to 0 \text{ as } n \to \infty.$$

Then, we have:

$$\frac{(h(x_{\tau(n)}) + l(x_{\tau(n)}))^2}{\theta^2(x_{\tau(n)})} \to 0 \text{ as } n \to \infty. \tag{3.21}$$

It follows that

$$\lim_{n\to\infty}\left((h(x_{\tau(n)}) + l(x_{\tau(n)}))^2\right) = 0,$$

which implies that

$$\lim_{n\to\infty} h(x_{\tau(n)}) = \lim_{n\to\infty} l(x_{\tau(n)}) = 0.$$

Moreover, we have

$$\limsup_{n\to\infty}\left\langle \psi(x^*) - x^*, x_{\tau(n)} - x^*\right\rangle \leq 0.$$

By the same computation as in Case 1, we have:

$$\|x_{\tau(n)+1} - x^*\|^2 \leq (1 - \epsilon_{\tau(n)})\|x_{\tau(n)} - x^*\|^2 + \epsilon_{\tau(n)}\xi_{\tau(n)}, \tag{3.22}$$

where $\epsilon_{\tau(n)} = \alpha_{\tau(n)}(2 - \alpha_{\tau(n)} - 2(1-\alpha_{\tau(n)})\delta)$ and

$$\xi_{\tau(n)} = \left[\frac{\alpha_{\tau(n)}\|\psi(x_{\tau(n)}) - x^*\|^2 + 2(1-\alpha_{\tau(n)})\langle\psi(x^*) - x^*, x_{\tau(n)} - x^*\rangle + 2\mu_{\tau(n)}\|A^*(I - \text{prox}_{\lambda g})Ax_{\tau(n)}\|\|\psi(x_{\tau(n)}) - x^*\|}{2 - \alpha_{\tau(n)} - 2(1-\alpha_{\tau(n)})\delta}\right].$$

Since $\|x_{\tau(n)} - x^*\|^2 \leq \|x_{\tau(n)+1} - x^*\|^2$, then by (3.22), we have:

$$\|x_{\tau(n)} - x^*\|^2 \leq \xi_{\tau(n)}.$$

We note that $\limsup_{n\to\infty} \xi_{\tau(n)} \leq 0$. Thus, it follows from above inequality that

$$\lim_{n\to\infty} \|x_{\tau(n)} - x^*\| = 0.$$

From (3.22), we also have:

$$\lim_{n\to\infty} \|x_{\tau(n)+1} - x^*\| = 0.$$

It follows from Lemma 2.5 that

$$0 \leq \|x_n - x^*\| \leq \|x_{\tau(n)+1} - x^*\| \to 0$$

as $n \to \infty$. Therefore, $\{x_n\}$ converges strongly to x^*. This completes the proof.

□

Taking $\psi(x) = u$ in Algorithm 3.1, we have the following Halpern-type algorithm.

Algorithm 3.3 Given an initial point $x_1 \in H_1$. Assume that x_n has been constructed and $\|A^*(I - \text{prox}_{\lambda g})Ax_n\|^2 + \|(I - \text{prox}_{\lambda f})x_n\|^2 \neq 0$, and then compute x_{n+1} by the following iterative scheme:

$$\begin{cases} y_n = \text{prox}_{\lambda\mu_n f}(\alpha_n u + (1-\alpha_n)x_n - \mu_n A^*(I - \text{prox}_{\lambda g})Ax_n) \\ x_{n+1} = \beta_n y_n + (1-\beta_n)Sy_n, \quad \forall n \in \mathbb{N}, \end{cases} \tag{3.23}$$

where the stepsize $\mu_n := \rho_n \dfrac{\left(\frac{1}{2}\|(I - \text{prox}_{\lambda g})Ax_n\|^2\right) + \left(\frac{1}{2}\|(I - \text{prox}_{\lambda f})x_n\|^2\right)}{\|A^*(I - \text{prox}_{\lambda g})Ax_n\|^2 + \|(I - \text{prox}_{\lambda f})x_n\|^2}$ with $0 < \rho_n < 4$ and $\{\alpha_n\}, \{\beta_n\} \subset [0, 1]$.

The following result is obtained directly by Theorem 3.2.

Corollary 3.4 *Let H_1 and H_2 be two real Hilbert spaces. Let $f : H_1 \to \mathbb{R} \cup \{+\infty\}$ and $g : H_2 \to \mathbb{R} \cup \{+\infty\}$be two proper and lower semicontinuous convex functions and $A : H_1 \to H_2$ be a bounded linear operator. Let $S : H_1 \to H_1$ be a nonexpansive mapping, such that $\Omega := F(S) \cap \Gamma \neq 0$. If the control sequences $\{\alpha_n\}$, $\{\beta_n\}$ and $\{\rho_n\}$ satisfy the following conditions:*

(C1) $\lim_{n\to\infty} \alpha_n = 0$ *and* $\sum_{n=1}^{\infty} \alpha_n = \infty$;

(C2) $0 < \liminf_{n\to\infty} \beta_n \leq \limsup_{n\to\infty} \beta_n < 1$;

(C3) $\varepsilon \leq \rho_n \leq \dfrac{4(1-\alpha_n)\left(\|(I - \text{prox}_{\lambda g})Ax_n\|^2\right)}{\left(\|(I - \text{prox}_{\lambda g})Ax_n\|^2\right) + \left(\|(I - \text{prox}_{\lambda f})x_n\|^2\right)} - \varepsilon$ *for some* $\varepsilon > 0$.

Then, the sequence $\{x_n\}$ defined by Algorithm 3.3 converges strongly to $z = P_\Omega u$.

4 Convergence theorem for split feasibility problems

In this section, we give an application of Theorem 3.2 to the split feasibility problem.

Algorithm 4.1 Given an initial point $x_1 \in H_1$. Assume that x_n has been constructed and $\|A^*(I - P_Q)Ax_n\|^2 + \|(I - P_C)x_n\|^2 \neq 0$, and then compute x_{n+1} by the following iterative scheme:

$$\begin{cases} y_n = P_C(\alpha_n \psi(x_n) + (1-\alpha_n)x_n - \mu_n A^*(I - P_Q)Ax_n) \\ x_{n+1} = \beta_n y_n + (1-\beta_n)Sy_n, \quad \forall n \in \mathbb{N}, \end{cases} \tag{4.1}$$

where the stepsize $\mu_n := \rho_n \dfrac{\left(\frac{1}{2}\|(I - P_Q)Ax_n\|^2\right) + \left(\frac{1}{2}\|(I - P_C)x_n\|^2\right)}{\|A^*(I - P_Q)Ax_n\|^2 + \|(I - P_C)x_n\|^2}$ with $0 < \rho_n < 4$ and $\{\alpha_n\}, \{\beta_n\} \subset (0, 1)$.

We now obtain a strong convergence theorem of Algorithm 4.1 for solving the split feasibility problem and the fixed point problem of nonexpansive mappings as follows:

Theorem 4.2 *Let H_1 and H_2 be two real Hilbert spaces, and let C and Q be nonempty, closed and convex subsets of H_1 and H_2, respectively. Let $A : H_1 \to H_2$ be a bounded linear operator. Let $\psi : H_1 \to H_1$ be a contraction mapping with $\delta \in [0, 1)$ and let $S : H_1 \to H_1$ be a nonexpansive mapping. Assume that $\Omega := F(S) \cap C \cap A^{-1}(Q) \neq \emptyset$. If the control sequences $\{\alpha_n\}$, $\{\beta_n\}$ and $\{\rho_n\}$ satisfy the following conditions:*

(C1) $\lim_{n\to\infty} \alpha_n = 0$ *and* $\sum_{n=1}^{\infty} \alpha_n = \infty$,

(C2) $0 < \liminf_{n\to\infty} \beta_n \leq \limsup_{n\to\infty} \beta_n < 1$;

(C3) $\varepsilon \leq \rho_n \leq \dfrac{4(1-\alpha_n)\left(\|(I-P_Q)Ax_n\|^2\right)}{\left(\|(I-P_Q)Ax_n\|^2\right)+\left(\|(I-P_C)x_n\|^2\right)} - \varepsilon$ *for some* $\varepsilon > 0$.

Then, the sequence $\{x_n\}$ *generated by Algorithm* 4.1 *converges strongly to* $z = P_\Omega \psi(z)$.

Proof Taking $f = i_C$ and $g = i_Q$ in Theorem 3.2 (i_C and i_Q are indicator functions of C and Q, respectively), we have $\text{prox}_{\lambda f} = P_C$ and $\text{prox}_{\lambda g} = P_Q$ for all λ. We also have $\text{argmin}\, f = C$ and $\text{argmin}\, g = Q$. Therefore, from Theorem 3.2, we obtain the desired result. □

5 Convergence theorem for nonexpansive semigroups

In this section, we prove a strong convergence theorem for finding a common solution of the proximal split feasibility problem and the fixed point problem of nonexpansive semigroups in Hilbert spaces.

Let C be a nonempty, closed, and convex subset of a real Banach space X. A one-parameter family $S = S(t) : t \geq 0 : C \to C$ is said to be a nonexpansive semigroup on C if it satisfies the following conditions:

(i) $S(0)x = x$ for all $x \in C$;
(ii) $S(s+t)x = S(s)S(t)x$ for all $t, s > 0$ and $x \in C$;
(iii) for each $x \in C$ the mapping $t \longmapsto S(t)x$ is continuous;
(iv) $\|S(t)x - S(t)y\| \leq \|x - y\|$ for all $x, y \in C$ and $t > 0$.

We use $F(S)$ to denote the common fixed point set of the semigroup S, i.e., $F(S) = \bigcap_{t>0} F(S(t)) = \{x \in C : x = S(t)x\}$. It is well known that $F(S)$ is closed and convex (see Browder 1956).

Definition 5.1 (Aleyner and Censor 2005) Let C be a nonempty, closed, and convex subset of a real Hilbert space H, $S = S(t) : t > 0$ be a continuous operator semigroup on C. Then, S is said to be uniformly asymptotically regular (in short, u.a.r.) on C if for all $h \geq 0$ and any bounded subset K of C, such that

$$\lim_{t\to\infty} \sup_{x\in K} \|S(h)(S(t)x) - S(t)x\| = 0.$$

Lemma 5.2 (Shimizu and Takahashi 1997) *Let C be a nonempty, closed, and convex subset of a real Hilbert space H, and let K be a bounded, closed, and convex subset of C. If we denote $S = S(t) : t > 0$ is a nonexpansive semigroup on C, such that $F(S) = \bigcap_{t>0} F(S(t)) \neq \emptyset$. For all $h > 0$, the set $\sigma_t(x) = \frac{1}{t}\int_0^t S(s)x\,\mathrm{d}s$, then*

$$\lim_{t\to\infty} \sup_{x\in K} \|\sigma_t(x) - S(h)\sigma_t(x)\| = 0.$$

Let H_1 and H_2 be two real Hilbert spaces. Let $f : H_1 \to \mathbb{R} \cup \{+\infty\}$ and $g : H_2 \to \mathbb{R} \cup \{+\infty\}$ be two proper and lower semicontinuous convex functions and $A : H_1 \to H_2$ be a bounded linear operator and let $\psi : H_1 \to H_1$ be a contraction mapping with $\delta \in [0, 1)$. Let $S := \{S(t) : t > 0\}$ be a u.a.r nonexpansive semigroup on H_1.

Algorithm 5.3 Given an initial point $x_1 \in H_1$. Assume that x_n has been constructed and $\|A^*(I - \text{prox}_{\lambda g})Ax_n\|^2 + \|(I - \text{prox}_{\lambda f})x_n\|^2 \neq 0$, and then compute x_{n+1} by the following iterative scheme:

$$\begin{cases} y_n = \text{prox}_{\lambda\mu_n f}(\alpha_n \psi(x_n) + (1-\alpha_n)x_n - \mu_n A^*(I - \text{prox}_{\lambda g})Ax_n) \\ x_{n+1} = \beta_n y_n + (1-\beta_n)S(t_n)y_n, \quad \forall n \in \mathbb{N}, \end{cases} \tag{5.1}$$

where the stepsize $\mu_n := \rho_n \dfrac{\left(\frac{1}{2}\|(I - \text{prox}_{\lambda g})Ax_n\|^2\right) + \left(\frac{1}{2}\|(I - \text{prox}_{\lambda f})x_n\|^2\right)}{\|A^*(I - \text{prox}_{\lambda g})Ax_n\|^2 + \|(I - \text{prox}_{\lambda f})x_n\|^2}$ with $0 < \rho_n < 4$, $\{\alpha_n\}, \{\beta_n\} \subset (0, 1)$ and $\{t_n\}$ is a positive real divergent sequence.

We now prove a strong convergence result for the problem (1.1) and the fixed point problem of nonexpansive semigroups as follows:

Theorem 5.4 *Suppose that $\bigcap_{t>0} F(S(t)) \cap \Gamma \neq 0$. If the control sequences $\{\alpha_n\}$, $\{\beta_n\}$ and $\{\rho_n\}$ satisfy the following conditions:*

(C1) $\lim_{n\to\infty} \alpha_n = 0$ *and* $\sum_{n=1}^{\infty} \alpha_n = \infty$;

(C2) $0 < \liminf_{n\to\infty} \beta_n \leq \limsup_{n\to\infty} \beta_n < 1$;

(C3) $\varepsilon \leq \rho_n \leq \dfrac{4(1-\alpha_n)\left(\|(I - \text{prox}_{\lambda g})Ax_n\|^2\right)}{\left(\|(I - \text{prox}_{\lambda g})Ax_n\|^2\right) + \left(\|(I - \text{prox}_{\lambda f})x_n\|^2\right)} - \varepsilon$ *for some* $\varepsilon > 0$.

Then, the sequence $\{x_n\}$ generated by Algorithm 5.3 converges strongly to a point $x^* \in \bigcap_{t>0} F(S(t)) \cap \Gamma$.

Proof By continuing in the same direction as in Theorem 3.2, we have that $\lim_{n\to\infty} \|y_n - S(t_n)y_n\| = 0$. Now, we only show that $\lim_{n\to\infty} \|y_n - S(h)y_n\| = 0$ for all $h \geq 0$. We observe that

$$\begin{aligned} \|y_n - S(h)y_n\| &\leq \|y_n - S(t_n)y_n\| + \|S(t_n)y_n - S(h)S(t_n)y_n\| + \|S(h)S(t_n)y_n - S(h)y_n\| \\ &\leq 2\|y_n - S(t_n)y_n\| + \sup_{x\in y_n} \|S(t_n)x - S(h)S(t_n)x\|. \end{aligned}$$

Since $\{S(t) : t \geq 0\}$ is a u.a.r. nonexpansive semigroup and $t_n \to \infty$ for all $h \geq 0$, we have:

$$\lim_{n\to\infty} \|y_n - S(h)y_n\| = 0,$$

for all $h \geq 0$. This completes the proof. □

6 Numerical examples

We first give a numerical example in Euclidean spaces to demonstrate the convergence of Algorithm (3.1).

Example 6.1 Let $H_1 = \mathbb{R}^2$ and $H_2 = \mathbb{R}^3$ with the usual norms. Define a mapping $S : \mathbb{R}^2 \to \mathbb{R}^2$ by:

$$S(a, b) := \frac{\sqrt{2}}{2}(a - b, a + b).$$

Table 1 The numerical experiment of Algorithm (6.1) by choosing $\delta = 0.1$

n	a_n	b_n	E_n
1	3.0000000	−2.0000000	–
2	0.1783143	−0.1100519	3.3961470
3	0.0082067	−0.0025830	0.2012117
4	0.0004998	0.0013948	0.0086729
5	0.0001562	0.0010892	0.0004598
6	0.0001076	0.0007884	0.0003047
7	0.0000801	0.0005827	0.0002075
8	0.0000608	0.0004388	0.0001452
9	0.0000467	0.0003353	0.0001045
10	0.0000363	0.0002591	0.0000769
⋮	⋮	⋮	⋮
28	0.0000008	0.0000055	0.0000012
29	0.0000007	0.0000046	0.00000098

One can show that S is nonexpansive. Define two functions $f:\mathbb{R}^2 \to (-\infty, \infty]$ and $g : \mathbb{R}^3 \to (-\infty, \infty]$ by $f := 0$, where 0 is a zero operator and

$$g(a, b, c) := \frac{|-3a + 7b - 2c|^2}{2}.$$

Then, the explicit forms of the proximity operators of f and g can be written by $\text{prox}_{\lambda f} = I$ and $\text{prox}_{1g} = B^{-1}$, where $B = \begin{pmatrix} 10 & -21 & 6 \\ -21 & 50 & -14 \\ 6 & -14 & 5 \end{pmatrix}$ (see Combettes and Pesquet 2011b). Let $A : \mathbb{R}^2 \to \mathbb{R}^3$ be defined by:

$$A := \begin{pmatrix} 2 & 1 \\ 7 & -3 \\ -5 & 4 \end{pmatrix},$$

and let $\Omega := F(S) \cap \text{argmin } f \cap A^{-1}(\text{argmin } g)$. Now, we rewrite Algorithm (3.1) in the form:

$$\begin{cases} y_n = \alpha_n \psi(x_n) + (1 - \alpha_n)x_n - \mu_n A^T(I - B^{-1})Ax_n \\ x_{n+1} = \beta_n y_n + (1 - \beta_n)Sy_n, \quad \forall n \in \mathbb{N}, \end{cases} \tag{6.1}$$

where

$$\mu_n = \frac{\rho_n}{2} \frac{\|(I - B^{-1})Ax_n\|^2}{\|A^T(I - B^{-1})Ax_n\|^2}.$$

Take $\alpha_n = \frac{1}{n+1}$, $\beta_n = \frac{1}{2}$, $\rho_n = \frac{2n}{n+1}$. Consider a contraction $\psi : \mathbb{R}^2 \to \mathbb{R}^2$ defined by $\psi(x) = \delta x$ for $0 \le \delta < 1$. We first start with the initial point $x_1 = (3, -2)$ and the stopping criterion for our testing process is set as: $E_n := \|x_n - x_{n-1}\| < 10^{-6}$, where $x_n = (a_n, b_n)$. In Table 1, we show the convergence behavior of Algorithm (6.1) by choosing $\delta = 0.1$. In Table 2, we also show the number of iterations of Algorithm (6.1) by choosing different constants δ. –

Table 2 The number of iterations of Algorithm (6.1) by choosing different constants δ

$\psi : \mathbb{R}^2 \to \mathbb{R}^2$, $\psi(x) = \delta x$			
Choices of δ	n (no. of iterations)	x_n	E_n
$\delta = 0$ (Shehu and Iyiola 2015, Algorithm 1)	42	$(-0.0000007, -0.0000048)$	0.00000098
$\delta = 0.05$	39	$(-0.0000007, -0.0000046)$	0.00000095
$\delta = 0.1$	29	$(-0.0000007, -0.0000046)$	0.00000098
$\delta = 0.2$	46	$(0.0000007, 0.0000050)$	0.00000099
$\delta = 0.5$	59	$(0.0000007, 0.0000052)$	0.00000097
$\delta = 0.9$	71	$(0.0000007, 0.0000049)$	0.00000088

Remark 6.2 In Example 6.1, by testing the convergence behavior of Algorithm (6.1), we observe that

(i) It converges to a solution, i.e., $x_n \to (0, 0) \in \Omega$.
(ii) The selection of a contraction ψ in our algorithm influences the number of iterations of the algorithm. We also note that if $\psi \equiv 0$ is zero, then our algorithm becomes Algorithm (1.6) (Shehu and Iyiola 2015, Algorithm 1).

Next, we give an example in the infinite-dimensional space L^2 as follows.

Example 6.3 Let $H_1 = L^2([0, 1]) = H_2$. Let $x \in L^2([0, 1])$. Define a bounded linear operator $A : L^2([0, 1]) \to L^2([0, 1])$ by:

$$(Ax)(t) := 3tx(t).$$

Define a mapping $S : L^2([0, 1]) \to L^2([0, 1])$ by:

$$(Sx)(t) := \sin(x(t)).$$

Then, S is nonexpansive. Let

$$C = \left\{x \in L^2([0, 1]) : \langle w, x\rangle \leq 0\right\},$$

where $w \in L^2([0, 1])$, such that $w(t) = 2t^3$, and let

$$Q = \left\{x \in L^2([0, 1]) : x \geq 0\right\}.$$

Define two functions $f, g : L^2([0, 1]) \to (-\infty, \infty]$ by $f := i_C$ and $g := i_Q$, where i_C and i_Q are indicator functions of C and Q, respectively. We can write the explicit forms of the proximity operators of f and g in the following forms:

$$\text{prox}_{\lambda f}\, x = P_C x = \begin{cases} x - \frac{\langle w, x\rangle}{\|w\|^2} w, & \text{if } x \notin C, \\ x, & \text{if } x \in C, \end{cases}$$

and $\text{prox}_{\lambda g}\, x = P_Q x = x_+$, where $x_+(t) = \max\{x(t), 0\}$ (see Cegielski 2012). Therefore, Algorithm (3.1) can be rewritten in the form:

$$\begin{cases} y_n = P_C(\alpha_n \psi(x_n) + (1 - \alpha_n)x_n - \mu_n A^*(I - P_Q)Ax_n) \\ x_{n+1} = \beta_n y_n + (1 - \beta_n)Sy_n, \quad \forall n \in \mathbb{N}; \end{cases} \tag{6.2}$$

$\mu_n = \rho_n \frac{\left(\frac{1}{2}\|(I - P_Q)Ax_n\|^2\right) + \left(\frac{1}{2}\|(I - P_C)x_n\|^2\right)}{\|A^*(I - P_Q)Ax_n\|^2 + \|(I - P_C)x_n\|^2}$, for finding a common element in the set $\Omega := F(S) \cap C \cap A^{-1}(Q)$. By choosing the control sequences $\{\alpha_n\}$, $\{\beta_n\}$ and $\{\rho_n\}$ satisfying the conditions (C1)–(C3) in Theorem 3.2, it can guarantee that the sequence $\{x_n\}$ generated by (6.2) converges strongly to $x^* = 0 \in \Omega$.

Acknowledgements The authors would like to thank the referees for valuable comments and suggestions for improving this work. W. Khuangsatung would like to thank Rajamangala University of Technology Thanyaburi and S. Suantai would like to thank Chiang Mai University for the financial support.

References

Abbas M, AlShahrani M, Ansari QH, Iyiola OS, Shehu Y (2018) Iterative methods for solving proximal split minimization problems. Numer Algorithms 78:193–215

Aleyner A, Censor Y (2005) Best approximation to common fixed points of a semigroup of nonexpansive operator. J Nonlinear Convex Anal 6(1):137–151

Browder FE (1956) Nonexpansive nonlinear operators in a Banach space. Proc Natl Acad Sci USA 54:1041–1044

Browder FE (1976) Nonlinear operators and nonlinear equations of evolution in Banach spaces. Proc. Symp. Pure Math. 18:78–81

Byrne C (2002) Iterative oblique projection onto convex sets and the split feasibility problem. Inverse Probl 18(2):441–453

Cegielski A (2012) Iterative methods for fixed point problems in Hilbert spaces. Lecture notes in mathematics, vol 2057. Springer, Heidelberg

Censor Y, Elfving T (1994) A multiprojection algorithm using Bregman projections in a product space. Numer Algorithms 8:221–239

Censor Y, Bortfeld T, Martin B, Trofimov A (2006) A unified approach for inversion problems in intensity-modulated radiation therapy. Phys Med Biol 51:2353–2365

Chang SS, Kim JK, Cho YJ, Sim J (2014) Weak and strong convergence theorems of solutions to split feasibility problem for nonspreading type mapping in Hilbert spaces. Fixed Point Theory Appl 2014:11

Combettes PL, Pesquet JC (2011a) Proximal splitting methods in signal processing. In: Fixed-point algorithms for inverse problems in science and engineering. Springer, New York, pp 185–212

Combettes PL, Pesquet JC (2011b) Proximal splitting methods in signal processing. Fixed Point Algorithms Inverse Probl Sci Eng 49:185–212

Lopez G, Martin-Marquez V, Wang F et al (2012) Solving the split feasibility problem without prior knowledge of matrix norms. Inverse Probl 28:085004

Mainge PE (2008) Strong convergence of projected subgradient methods for nonsmooth and nonstrictly convex minimization. Set Valued Anal 16:899–912

Moudafi A, Thakur BS (2014) Solving proximal split feasibility problems without prior knowledge of operator norms. Optim Lett 8:2099–2110

Qu B, Xiu N (2005) A note on the CQ algorithm for the split feasibility problem. Inverse Probl 21(5):1655–1665

Shehu Y, Iyiola OS (2015) Convergence analysis for proximal split feasibility problems and fixed point problems. J Appl Math Comput 48:221–239

Shehu Y, Iyiola OS (2017a) Convergence analysis for the proximal split feasibility problem using an inertial extrapolation term method. J Fixed Point Theory Appl 19(4):2483–2510

Shehu Y, Iyiola OS (2017b) Strong convergence result for proximal split feasibility problem in Hilbert spaces. Optimization 66(12):2275–2290

Shehu Y, Iyiola OS (2018) Accelerated hybrid viscosity and steepest-descent method for proximal split feasibility problems. Optimization 67(4):475–492

Shehu Y, Cai G, Iyiola OS (2015) Iterative approximation of solutions for proximal split feasibility problems. Fixed Point Theory Appl 2015:123

Shimizu T, Takahashi W (1997) Strong convergence to common fixed points of families of nonexpansive mappings. J Math Anal Appl 211(1):71–83

Takahashi W (2000) Nonlinear functional analysis. Yokohama Publishers, Yokohama

Witthayarat U, Cho YJ, Cholamjiak P (2018) On solving proximal split feasibility problems and applications. Ann Funct Anal 9(1):111–122

Xu HK (2003) An iterative approach to quadric optimization. J Optim Theory Appl 116:659–678

Yao Z, Cho SY, Kang SM et al (2014) A regularized algorithm for the proximal split feasibility problem. Abstr Appl Anal 6:894272

Khuangsatung and Sunthrayuth *Journal of Inequalities and Applications* (2018) 2018:167
https://doi.org/10.1186/s13660-018-1760-5

Journal of Inequalities and Applications
a SpringerOpen Journal

RESEARCH Open Access

The generalized viscosity explicit rules for a family of strictly pseudo-contractive mappings in a q-uniformly smooth Banach space

Wongvisarut Khuangsatung[1] and Pongsakorn Sunthrayuth[1*]

*Correspondence:
pongsakorn_su@rmutt.ac.th
[1]Department of Mathematics and Computer Science, Faculty of Science and Technology, Rajamangala University of Technology Thanyaburi (RMUTT), Pathumthani, Thailand

Abstract

In this paper, we construct an iterative method by a generalized viscosity explicit rule for a countable family of strictly pseudo-contractive mappings in a q-uniformly smooth Banach space. We prove strong convergence theorems of proposed algorithm under some mild assumption on control conditions. We apply our results to the common fixed point problem of convex combination of family of mappings and zeros of accretive operator in Banach spaces. Furthermore, we also give some numerical examples to support our main results.

Keywords: Strict pseudo-contractions; Banach space; Strong convergence; Fixed point problem; Iterative method

1 Introduction

In this paper, we assume that E is a real Banach space with dual space E^* and C is a nonempty subset of E. Let $q > 1$ be a real number. The *generalized duality mapping* $J_q : E \to 2^{E^*}$ is defined by

$$J_q(x) = \left\{\bar{x} \in E^* : \langle x, \bar{x}\rangle = \|x\|^q, \|\bar{x}\| = \|x\|^{q-1}\right\},$$

where $\langle \cdot, \cdot \rangle$ denotes the generalized duality pairing between elements of E and E^*. In particular, $J_q = J_2$ is called the *normalized duality mapping*. If E is smooth, then J_q is single-valued and denoted by j_q (see [1]). If $E := H$ is a real Hilbert space, then $J = I$, where I is the identity mapping. Further, we have the following properties of the generalized duality mapping J_q:

- $J_q(x) = \|x\|^{q-2} J_2(x)$ for all $x \in E$ with $x \neq 0$.
- $J(tx) = t^{q-1} J_q(x)$ for all $x \in E$ and $t \geq 0$.
- $J_q(-x) = -J_q(x)$ for all $x \in E$.

Let T be a self-mapping of C. We denote the fixed point set of the mapping T by $F(T) = \{x \in C : x = Tx\}$. A mapping $f : C \to C$ is said to be a *contraction* if there exists a constant $\rho \in (0, 1)$ satisfying

$$\left\|f(x) - f(y)\right\| \leq \rho\|x - y\|, \quad \forall x, y \in C.$$

We use Π_C to denote the collection of all contractions from C into itself. Recall that a mapping $T : C \to C$ is said to be *nonexpansive* if

$$\|Tx - Ty\| \leq \|x - y\|, \quad \forall x, y \in C.$$

A mapping $T : C \to C$ is said to be λ-*strict pseudo-contraction* if for all $x, y \in C$, there exist $\lambda > 0$ and $j_q(x - y) \in J_q(x - y)$ such that

$$\langle Tx - Ty, j_q(x - y)\rangle \leq \|x - y\|^q - \lambda \|(I - T)x - (I - T)y\|^q, \quad \forall x, y \in C. \tag{1}$$

It is not hard to show that (1) equivalent to the following inequality:

$$\langle (I - T)x - (I - T)y, j_q(x - y)\rangle \geq \lambda \|(I - T)x - (I - T)y\|^q, \quad \forall x, y \in C. \tag{2}$$

If $E := H$ is a Hilbert space, then (1) (and so (2)) is equivalent to the following inequality:

$$\|Tx - Ty\|^2 \leq \|x - y\|^2 + k\|(I - T)x - (I - T)y\|^2, \quad \forall x, y \in C, \tag{3}$$

where $k = 1 - 2\lambda < 1$. We assume that $k \geq 0$, so that $k \in [0, 1)$. Note that the class of strictly pseudo-contractive mappings include the class of nonexpansive mappings as a particular case in Hilbert spaces. Clearly, T is nonexpansive if and only if T is a 0-strict pseudo-contraction. Strict pseudo-contractions were first introduced by Browder and Petryshyn [2] in 1967. They have more powerful applications than nonexpansive mappings do in solving inverse problems (see, e.g., [3]). Therefore it is more interesting to study the theory of iterative methods for strictly pseudo-contractive mappings. Several researchers studied the class of strictly pseudo-contractive mappings in Hilbert and Banach spaces (see, e.g., [4–9] and the references therein).

Now, we give some examples of λ-strictly pseudo-contractive mappings.

Example 1.1 ([8]) Let $E = \mathbb{R}$ with the usual norm, and let $C = (0, \infty)$. Let $T : C \to C$ be defined by

$$Tx = \frac{x^2}{1 + x}, \quad x \in C.$$

Then, T is a 1-strict pseudo-contraction.

Example 1.2 ([8]) Let $E = \mathbb{R}$ with the usual norm, and let $C = [-1, 1]$. Let $T : C \to C$ be defined by

$$Tx = \begin{cases} x, & x \in [-1, 0], \\ x - x^2, & x \in [0, 1]. \end{cases}$$

Then, T is a λ-strict pseudo-contraction with constant $\lambda > 0$.

Over the last several years, the implicit midpoint rule (IMR) has become a powerful numerical method for numerically solving time-dependent differential equations (in particular, stiff equations) (see [10–15]) and differential algebraic equations (see [16]). Consider

the following initial value problem:

$$x'(t) = f\big(x(t)\big), \quad x(t_0) = x_0, \tag{4}$$

where $f : \mathbb{R}^M \to \mathbb{R}^M$ is a continuous function. The IMR is an implicit method given by the following finite difference scheme [17]:

$$\begin{cases} y_0 = x_0, \\ y_{n+1} = y_n + hf(\frac{y_n+y_{n+1}}{2}), \quad n \geq 0, \end{cases} \tag{5}$$

where $h > 0$ is a time step. It is known that if $f : \mathbb{R}^M \to \mathbb{R}^M$ is Lipschitz continuous and sufficiently smooth, then the sequence $\{y_n\}$ converges to the exact solution of (4) as $h \to 0$ uniformly over $t \in [t_0, t^*]$ for any fixed $t^* > 0$. If the function f is written as $f(x) = x - g(x)$, then (5) becomes

$$\begin{cases} y_0 = x_0, \\ y_{n+1} = y_n + h[\frac{y_n+y_{n+1}}{2} - g(\frac{y_n+y_{n+1}}{2})], \quad n \geq 0, \end{cases} \tag{6}$$

and the critical points of (4) are the fixed points of the problem $x = g(x)$.

Based on IMR (5), Alghamdi et al. [18] introduced the following two algorithms for the solution of the fixed point problem $x = Tx$, where T is a nonexpansive mapping in a Hilbert space H:

$$x_{n+1} = x_n - t_n\left[\frac{x_n + x_{n+1}}{2} - T\left(\frac{x_n + x_{n+1}}{2}\right)\right], \quad n \geq 0, \tag{7}$$

$$x_{n+1} = (1 - t_n)x_n + t_n T\left(\frac{x_n + x_{n+1}}{2}\right), \quad n \geq 0, \tag{8}$$

for $x_0 \in H$, with $\{t_n\}_{n=1}^{\infty} \subset (0, 1)$. They proved that these two schemes converge weakly to a point in $F(T)$.

To obtain strong convergence, Xu et al. [19] applied the viscosity approximation method introduced by Moudafi [20] to the IMR for a nonexpansive mapping T and proposed the following *viscosity implicit midpoint rule* in Hilbert spaces H as follows:

$$x_{n+1} = \alpha_n f(x_n) + (1 - \alpha_n)T\left(\frac{x_n + x_{n+1}}{2}\right), \quad n \geq 1, \tag{9}$$

where $\{\alpha_n\}$ is a real control condition in $(0, 1)$. They also proved that the sequence $\{x_n\}$ generated by (9) converges strongly to a point $x^* \in F(T)$, which solves the variational inequality

$$\langle (f - I)x^*, z - x^* \rangle \leq 0, \quad z \in F(T). \tag{10}$$

Later, Ke and Ma [21] improved the viscosity implicit midpoint rule by replacing the midpoint by any point of the interval $[x_n, x_{n+1}]$. They introduced the so-called *generalized viscosity implicit rules* to approximating the fixed point of a nonexpansive mapping T in

Hilbert spaces H as follows:

$$x_{n+1} = \alpha_n f(x_n) + (1-\alpha_n)T\big(s_n x_n + (1-s_n)x_{n+1}\big), \quad n \geq 1. \tag{11}$$

They also proved that the sequence $\{x_n\}$ generated by (11) converges strongly to a point $x^* \in F(T)$ that solves the variational inequality (10).

In numerical analysis, it is clear that the computation by the IMR is not an easy work in practice. Because the IMR need to compute at every time steps, it can be much harder to implement. To overcome this difficulty, for solving (4), we consider the helpful method, the so-called *explicit midpoint method* (EMR), given by the following finite difference scheme [22, 23]:

$$\begin{cases} y_0 = x_0, \\ \bar{y}_{n+1} = y_n + hf(y_n), \\ y_{n+1} = y_n + hf(\frac{y_n+\bar{y}_{n+1}}{2}), \quad n \geq 0. \end{cases} \tag{12}$$

Note that the EMR (12) calculates the system status at a future time from the currently known system status, whereas IMR (5) calculates the system status involving both the current state of the system and the later one (see [23, 24]).

In 2017, Marino et al. [25] combined the generalized viscosity implicit midpoint rules (11) with the EMR (12) for a quasi-nonexpansive mapping T and introduced the following so-called *generalized viscosity explicit midpoint rule* in Hilbert spaces H as follows:

$$\begin{cases} \bar{x}_{n+1} = \beta_n x_n + (1-\beta_n)Tx_n, \\ x_{n+1} = \alpha_n f(x_n) + (1-\alpha_n)T(s_n x_n + (1-s_n)\bar{x}_{n+1}), \quad n \geq 1. \end{cases} \tag{13}$$

They also showed that, under certain assumptions imposed on the parameters, the sequence $\{x_n\}$ generated by (13) converges strongly to a point $x^* \in F(T)$, which solves the variational inequality (10).

The above results naturally bring us to the following questions.

Question 1 Can we extend the generalized viscosity explicit midpoint rule (13) to higher spaces other than Hilbert spaces? Such as a 2-uniformly smooth Banach space or, more generally, in a q-uniformly smooth Banach space.

Question 2 Can we obtain a strong convergence result of generalized viscosity explicit midpoint rule (13) for finding the set of common fixed points of a family of mappings? Such as a countable family of strict pseudo-contractions.

The purpose of this paper is to give some affirmative answers to the questions raised. We introduce an iterative algorithm for finding the set of common fixed points of a countable family of strict pseudo-contractions by a generalized viscosity explicit rule in a q-uniformly smooth Banach space. We prove the strong convergence of the proposed algorithm under some mild assumption on control conditions. We apply our results to the common fixed point problem of a convex combination of a family of mappings and zeros of an accretive operator in Banach spaces. Furthermore, we also give some numerical examples to support our main results.

2 Preliminaries

Let E be a real Banach space with norm $\|\cdot\|$ and dual space E^* of E. The symbol $\langle x, x^*\rangle$ denotes the pairing between E and E^*, that is, $\langle x, x^*\rangle = x^*(x)$, the value of x^* at x. The *modulus of convexity* of E is the function $\delta : (0,2] \to [0,1]$ defined by

$$\delta(\epsilon) = \inf\left\{1 - \frac{\|x+y\|}{2} : x, y \in E, \|x\| = \|y\| = 1, \|x-y\| \geq \epsilon\right\}.$$

A Banach space E is said to be *uniformly convex* if $\delta_E(\epsilon) > 0$ for all $\epsilon \in (0,2]$. For $p > 1$, we say that E is said to be *p-uniformly convex* if there is $c_p > 0$ such that $\delta_E(\epsilon) \geq c_p \epsilon^p$ for all $\epsilon \in (0,2]$.

The *modulus of smoothness* of E is the function $\rho_E : \mathbb{R}^+ := [0,\infty) \to \mathbb{R}^+$ defined by

$$\rho_E(\tau) = \sup\left\{\frac{\|x+\tau y\| + \|x-\tau y\|}{2} - 1 : \|x\|, \|y\| \leq 1\right\}.$$

A Banach space E is said to be *uniformly smooth* if $\frac{\rho_E(\tau)}{\tau} \to 0$ as $\tau \to 0$. For $q > 1$, a Banach space E is said to be *q-uniformly smooth* if there exists $c_q > 0$ such that $\rho_E(\tau) \leq c_q \tau^q$ for all $\tau > 0$. If E is q-uniformly smooth, then $q \leq 2$, and E is also uniformly smooth. Further, E is p-uniformly convex (q-uniformly smooth) if and only if E^* is q-uniformly smooth (p-uniformly convex), where $p \geq 2$ and $1 < q \leq 2$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$. It is well known that Hilbert spaces L_p and l_p ($p > 1$) are uniformly smooth (see [26]). More precisely, the spaces L_p and l_p are $\min\{p,2\}$-uniformly smooth for every $p > 1$.

Definition 2.1 Let C a be nonempty closed convex subsets of E, and let Q be a mapping of E onto C. Then Q is said to be:

- *sunny* if $Q(Qx + t(x - Qx)) = Qx$ for all $x \in C$ and $t \geq 0$.
- *retraction* if $Qx = x$ for all $x \in C$.
- a sunny nonexpansive retraction if Q is sunny, nonexpansive, and a retraction from E onto C.

It is known that if $E := H$ is a real Hilbert space, then a sunny nonexpansive retraction Q coincides with the metric projection from E onto C. Moreover, if E is uniformly smooth and T is a nonexpansive mapping of C into itself with $F(T) \neq \emptyset$, then $F(T)$ is a sunny nonexpansive retraction from E onto C (see [27]). We know that in a uniformly smooth Banach space, a retraction $Q : C \to E$ is sunny and nonexpansive if and only if $\langle x - Qx, j_q(y - Qx)\rangle \leq 0$ for all $x \in E$ and $y \in C$ (see [28]).

Lemma 2.2 ([29]) *Let C be a nonempty closed convex subset of a uniformly smooth Banach space E. Let $S : C \to C$ be a nonexpansive self-mapping such that $F(S) \neq \emptyset$ and $f \in \Pi_C$. Let $\{z_t\}$ be the net sequence defined by*

$$z_t = tf(z_t) + (1-t)Sz_t, \quad t \in (0,1).$$

Then:

(i) *$\{x_t\}$ converges strongly as $t \to 0$ to a point $Q(f) \in F(S)$, which solves the variational inequality*

$$\left\langle (I-f)Q(f), j_q\big(Q(f) - z\big)\right\rangle \leq 0, \quad z \in F(S).$$

(ii) *Suppose that* $\{x_n\}$ *is a bounded sequence such that* $\lim_{n\to\infty}\|x_n - Sx_n\| = 0$. *If* $Q(f) := \lim_{t\to 0} x_t$ *exists, then*

$$\limsup_{n\to\infty}\langle (f - I)Q(f), j_q(x_n - Q(f))\rangle \leq 0.$$

Lemma 2.3 ([30]) *Let* C *be a nonempty closed convex subset of a real* q*-uniformly smooth Banach space* E. *Let* $T : C \to C$ *be a* λ*-strict pseudo-contraction. For all* $x \in C$, *we define* $T_\theta x := (1-\theta)x + \theta Tx$. *Then, as* $\theta \in (0, \delta]$, $\delta = \min\{1, (\frac{q\lambda}{\kappa_q})^{\frac{1}{q-1}}\}$, *where* κ_q *is the* q*-uniform smoothness constant, and* $T_\theta : C \to C$ *is nonexpansive such that* $F(T_\theta) = F(T)$.

Using the concept of subdifferentials, we have the following inequality.

Lemma 2.4 ([31]) *Let* $q > 1$, *and let* E *be a real normed space with the generalized duality mapping* J_q. *Then, for any* $x, y \in E$, *we have*

$$\|x + y\|^q \leq \|x\|^q + q\langle y, j_q(x + y)\rangle, \tag{14}$$

where $j_q(x + y) \in J_q(x + y)$.

Lemma 2.5 ([32]) *Let* $p > 1$ *and* $r > 0$ *be two fixed real numbers, and let* E *be a uniformly convex Banach space. Then, for all* $x, y \in B_r$ *and* $t \in [0, 1]$,

$$\|tx + (1-t)y\|^p \leq t\|x\|^p + (1-t)\|y\|^p - t(1-t)c\|x - y\|^p,$$

where $c > 0$.

Lemma 2.6 ([33]) *Suppose that* $q > 1$. *Then*

$$ab \leq \frac{1}{q}a^q + \left(\frac{q-1}{q}\right)b^{\frac{q}{q-1}}$$

for positive real numbers a, b.

Lemma 2.7 ([34]) *Let* $\{a_n\}$ *be a sequence of nonnegative real numbers,* $\{\gamma_n\}$ *be a sequence of* $(0, 1)$ *with* $\sum_{n=1}^{\infty}\gamma_n = \infty$, $\{c_n\}$ *be a sequence of nonnegative real number with* $\sum_{n=1}^{\infty} c_n < \infty$, *and let* $\{b_n\}$ *be a sequence of real numbers with* $\limsup_{n\to\infty} b_n \leq 0$. *Suppose that*

$$a_{n+1} = (1 - \gamma_n)a_n + \gamma_n b_n + c_n$$

for all $n \in \mathbb{N}$. *Then,* $\lim_{n\to\infty} a_n = 0$.

Lemma 2.8 ([35]) *Let* $\{s_n\}$ *be sequences of real numbers such that there exists a subsequence* $\{n_i\}$ *of* $\{n\}$ *such that* $s_{n_i} < s_{n_i+1}$ *for all* $i \in \mathbb{N}$. *Then there exists an increasing sequence* $\{m_k\} \subset \mathbb{N}$ *such that* $\lim_{k\to\infty} m_k = \infty$ *and the following properties are satisfied by all sufficiently large numbers* $k \in \mathbb{N}$:

$$s_{m_k} \leq s_{m_k+1} \quad \text{and} \quad s_k \leq s_{m_k+1}.$$

In fact, $m_k := \max\{j \leq k : s_j \leq s_{j+1}\}$.

Definition 2.9 ([34]) Let C be a nonempty closed convex subset of a real Banach space E. Let $\{T_n\}_{n=1}^{\infty}$ be a family of mappings of C into itself. We say that $\{T_n\}_{n=1}^{\infty}$ satisfies the $AKTT$-condition if

$$\sum_{n=1}^{\infty} \sup_{w\in C} \|T_{n+1}w - T_n w\| < \infty. \tag{15}$$

Lemma 2.10 ([34]) *Let C be a nonempty closed convex subset of a real Banach space E. Suppose that $\{T_n\}_{n=1}^{\infty}$ satisfies the AKTT-condition. Then, for each $x \in C$, $\{T_n x\}$ converges strongly to some point of C. Moreover, let T be the mapping of C into itself defined by $Tx = \lim_{n\to\infty} T_n x$ for all $x \in C$. Then, $\lim_{n\to\infty} \sup_{w\in C} \|Tw - T_n w\| = 0$.*

In the following, we will write that $(\{T_n\}, T)$ satisfies the $AKTT$-condition if $\{T_n\}$ satisfies the $AKTT$-condition and T is defined by Lemma 2.10 with $F(T) = \bigcap_{n=1}^{\infty} F(T_n)$.

3 Main results

Theorem 3.1 *Let C be a nonempty closed convex subset of a real uniformly convex and q-uniformly smooth Banach space E. Let $f \in \Pi_C$ with coefficient $\rho \in (0,1)$, and let $\{T_n\}_{n=1}^{\infty} : C \to C$ be a family of λ-strict pseudo-contractions such that $\Omega := \bigcap_{n=1}^{\infty} F(T_n) \neq \emptyset$. For all $x \in C$, define the mapping $S_n x = (1-\theta_n)x + \theta_n T_n x$, where $0 < \theta_n \leq \delta$, $\delta = \min\{1, (\frac{q\lambda}{\kappa_q})^{\frac{1}{q-1}}\}$, and $\sum_{n=1}^{\infty} |\theta_{n+1} - \theta_n| < \infty$. For given $x_1 \in C$, let $\{x_n\}$ be a sequence generated by*

$$\begin{cases} \bar{x}_{n+1} = \beta_n x_n + (1-\beta_n) S_n x_n, \\ x_{n+1} = \alpha_n f(x_n) + (1-\alpha_n) S_n (t_n x_n + (1-t_n)\bar{x}_{n+1}), \quad n \geq 1, \end{cases} \tag{16}$$

where $\{\alpha_n\}$, $\{\beta_n\}$, and $\{t_n\}$ are sequences in $(0,1)$ satisfying the following conditions:

(C1) $\lim_{n\to\infty} \alpha_n = 0$, $\sum_{n=1}^{\infty} \alpha_n = \infty$;

(C2) $\liminf_{n\to\infty} \beta_n(1-\beta_n)(1-t_n) > 0$.

Suppose in addition that $(\{T_n\}_{n=1}^{\infty}, T)$ satisfies the AKTT-condition. Then, $\{x_n\}$ defined by (16) converges strongly to $x^* = Q(f) \in \Omega$, which solves the variational inequality

$$\big\langle (I-f)Q(f), j_q\big(Q(f) - z\big)\big\rangle \leq 0, \quad z \in \Omega, \tag{17}$$

where Q is a sunny nonexpansive retraction of C onto Ω.

Proof First, we show that $\{x_n\}$ is bounded. From Lemma 2.3 we have that S_n is nonexpansive such that $F(S_n) = F(T_n)$ for all $n \geq 1$. Put $z_n := t_n x_n + (1-t_n)\bar{x}_{n+1}$. For each $z \in \Omega := \bigcap_{n=1}^{\infty} F(T_n)$, we have

$$\begin{aligned} \|z_n - z\| &= \big\|t_n(x_n - z) + (1-t_n)(\bar{x}_{n+1} - z)\big\| \\ &\leq t_n\|x_n - z\| + (1-t_n)\|\bar{x}_{n+1} - z\| \\ &\leq t_n\|x_n - z\| + (1-t_n)\big(\beta_n\|x_n - z\| + (1-\beta_n)\|S_n x_n - z\|\big) \\ &\leq t_n\|x_n - z\| + (1-t_n)\beta_n\|x_n - z\| + (1-t_n)(1-\beta_n)\|x_n - z\| \\ &= \|x_n - z\|. \end{aligned} \tag{18}$$

It follows that

$$\begin{aligned}
\|x_{n+1}-z\| &= \big\|\alpha_n f(x_n)+(1-\alpha_n)S_n z_n - z\big\| \\
&= \big\|\alpha_n\big(f(x_n)-f(z)\big)+\alpha_n\big(f(z)-z\big)+(1-\alpha_n)(S_n z_n - z)\big\| \\
&\le \alpha_n\big\|f(x_n)-f(z)\big\|+\alpha_n\big\|f(z)-z\big\|+(1-\alpha_n)\|S_n z_n - z\| \\
&\le \big(1-(1-\rho)\alpha_n\big)\|x_n-z\|+(1-\rho)\alpha_n\frac{\|f(z)-z\|}{1-\rho} \\
&\le \max\left\{\|x_n-z\|,\frac{\|f(z)-z\|}{1-\rho}\right\}.
\end{aligned}$$

By induction we have

$$\|x_n-z\|\le\max\left\{\|x_1-z\|,\frac{\|f(z)-z\|}{1-\rho}\right\},\quad n\ge 1.$$

Hence $\{x_n\}$ is bounded. Consequently, we deduce immediately that $\{f(x_n)\}$ and $\{S_n(t_n x_n+(1-t_n)\bar{x}_{n+1})\}$ are bonded. Let $x^*=Q(f)$. By the convexity of $\|\cdot\|^q$ and Lemma 2.5 we have

$$\begin{aligned}
\big\|S_n z_n - x^*\big\|^q &\le \big\|z_n-x^*\big\|^q \\
&= \big\|t_n\big(x_n-x^*\big)+(1-t_n)\big(\bar{x}_{n+1}-x^*\big)\big\|^q \\
&\le t_n\big\|x_n-x^*\big\|^q+(1-t_n)\big\|\bar{x}_{n+1}-x^*\big\|^q \\
&= t_n\big\|x_n-x^*\big\|^q+(1-t_n)\big\|\beta_n\big(x_n-x^*\big)+(1-\beta_n)\big(S_n x_n-x^*\big)\big\|^q \\
&\le t_n\big\|x_n-x^*\big\|^q+(1-t_n)\big[\beta_n\big\|x_n-x^*\big\|^q+(1-\beta_n)\big\|S_n x_n-x^*\big\|^q \\
&\quad -\beta_n(1-\beta_n)c\|x_n-S_n x_n\|^q\big] \\
&\le \big\|x_n-x^*\big\|^q-\beta_n(1-\beta_n)(1-t_n)c\|x_n-S_n x_n\|^q. \qquad (19)
\end{aligned}$$

It follows from Lemma 2.4 and (19) that

$$\begin{aligned}
&\big\|x_{n+1}-x^*\big\|^q \\
&= \big\|\alpha_n\big(f(x_n)-x^*\big)+(1-\alpha_n)\big(S_n z_n-x^*\big)\big\|^q \\
&= \big\|\alpha_n\big(f(x_n)-f\big(x^*\big)\big)+\alpha_n\big(f\big(x^*\big)-x^*\big)+(1-\alpha_n)\big(S_n z_n-x^*\big)\big\|^q \\
&\le \big\|\alpha_n\big(f(x_n)-f\big(x^*\big)\big)+(1-\alpha_n)\big(S_n z_n-x^*\big)\big\|^q+q\alpha_n\big\langle f\big(x^*\big)-x^*,j_q\big(x_{n+1}-x^*\big)\big\rangle \\
&\le \alpha_n\big\|f(x_n)-f\big(x^*\big)\big\|^q+(1-\alpha_n)\big\|S_n z_n-x^*\big\|^q+q\alpha_n\big\langle f\big(x^*\big)-x^*,j_q\big(x_{n+1}-x^*\big)\big\rangle \\
&\le \alpha_n\big\|f(x_n)-f\big(x^*\big)\big\|^q+(1-\alpha_n)\big[\big\|x_n-x^*\big\|^q-\beta_n(1-\beta_n)(1-t_n)c\|x_n-S_n x_n\|^q\big] \\
&\quad +q\alpha_n\big\langle f\big(x^*\big)-x^*,j_q\big(x_{n+1}-x^*\big)\big\rangle \\
&\le \big(1-(1-\rho)\alpha_n\big)\big\|x_n-x^*\big\|^q-(1-\alpha_n)\beta_n(1-\beta_n)(1-t_n)c\|x_n-S_n x_n\|^q \\
&\quad +q\alpha_n\big\langle f\big(x^*\big)-x^*,j_q\big(x_{n+1}-x^*\big)\big\rangle. \qquad (20)
\end{aligned}$$

The rest of the proof will be divided into two cases:

Case 1. Suppose that there exists $n_0 \in \mathbb{N}$ such that $\{\|x_n - x^*\|\}_{n=n_0}^{\infty}$ is nonincreasing. This implies that $\{\|x_n - x^*\|\}_{n=1}^{\infty}$ is convergent. From (20) we see that

$$(1-\alpha_n)\beta_n(1-\beta_n)(1-s_n)c\|x_n - S_n x_n\|^q \le \|x_n - x^*\|^q - \|x_{n+1} - x^*\|^q + \alpha_n M,$$

where $c > 0$ and $M = \sup_{n\ge 1}\{q\|f(x^*) - x^*\|\|x_{n+1} - x^*\|^{q-1}, (1-\rho)\|x_n - x^*\|^q\} < \infty$. From $(C1)$ and $(C2)$ we get that

$$\lim_{n\to\infty} \|x_n - S_n x_n\| = 0. \tag{21}$$

We observe that

$$\begin{aligned}
&\sup_{x\in\{x_n\}} \|S_{n+1}x - S_n x\| \\
&\quad = \sup_{x\in\{x_n\}} \|(1-\theta_{n+1})x + \theta_{n+1}T_{n+1}x - (1-\theta_n)x - \theta_n T_n x\| \\
&\quad \le |\theta_{n+1} - \theta_n| \sup_{x\in\{x_n\}} \|x\| + \theta_{n+1} \sup_{x\in\{x_n\}} \|T_{n+1}x - T_n x\| + |\theta_{n+1} - \theta_n| \sup_{x\in\{x_n\}} \|T_n x\| \\
&\quad \le |\theta_{n+1} - \theta_n|\Big(\sup_{x\in\{x_n\}} \|x\| + \sup_{x\in\{x_n\}} \|T_n x\|\Big) + \sup_{x\in\{x_n\}} \|T_{n+1}x - T_n x\|.
\end{aligned}$$

Since $\{T_n\}_{n=1}^{\infty}$ satisfies the $AKTT$-condition and $\sum_{n=1}^{\infty} |\theta_{n+1} - \theta_n| < \infty$, we have

$$\sum_{n=1}^{\infty} \sup_{x\in\{x_n\}} \|S_{n+1}x - S_n x\| < \infty,$$

that is, $\{S_n\}_{n=1}^{\infty}$ satisfies the $AKTT$-condition. From this we can define the nonexpansive mapping $S : C \to C$ by $Sx = \lim_{n\to\infty} S_n x$ for all $x \in C$. Since $\{\theta_n\}$ is bounded, there exists a subsequence $\{\theta_{n_i}\}$ of $\{\theta_n\}$ such that $\theta_{n_i} \to \theta$ as $i \to \infty$. It follows that

$$Sx = \lim_{i\to\infty} S_{n_i}x = \lim_{i\to\infty}\big[(1-\theta_{n_i})x + \theta_{n_i}T_{n_i}x\big] = (1-\theta)x + \theta Tx, \quad x \in C.$$

This shows that $F(S) = F(T) = \bigcap_{n=1}^{\infty} F(T_n) := \Omega$. By (21) and Lemma 2.10 we have

$$\begin{aligned}
\|x_n - Sx_n\| &\le \|x_n - S_n x_n\| + \|S_n x_n - Sx_n\| \\
&\le \|x_n - S_n x_n\| + \sup_{x\in\{x_n\}} \|S_n x - Sx\| \to 0 \quad \text{as } n \to \infty.
\end{aligned} \tag{22}$$

Let $\{z_t\}$ be a sequence defined by

$$z_t = f(z_t) + (1-t)Sz_t, \quad t \in (0,1).$$

From Lemma 2.2(i) we know that $\{x_t\}$ converges strongly to $x^* = Q(f)$, which solves the variational inequalities

$$\big\langle (I-f)Q(f), j_q\big(Q(f) - z\big)\big\rangle \le 0, \quad z \in \Omega.$$

Moreover, we obtain that

$$\limsup_{n\to\infty}\left\langle f\left(x^*\right)-x^*,j_q\left(x_n-x^*\right)\right\rangle\leq 0. \tag{23}$$

Note that

$$\begin{aligned}\|S_nz_n-x_n\| &\leq \|S_nz_n-S_nx_n\|+\|S_nx_n-x_n\|\\ &\leq \|z_n-x_n\|+\|S_nx_n-x_n\|\\ &= (1-s_n)(1-\beta_n)\|S_nx_n-x_n\|+\|S_nx_n-x_n\|\\ &\leq 2\|x_n-S_nx_n\|.\end{aligned}$$

From (21), we get that

$$\lim_{n\to\infty}\|S_nz_n-x_n\|=0. \tag{24}$$

It follows that

$$\begin{aligned}&\|x_{n+1}-x_n\|\\ &\quad\leq \left\|\alpha_n\left(f(x_n)-x_n\right)+(1-\alpha_n)(S_nz_n-x_n)\right\|\\ &\quad\leq \alpha_n\left\|f(x_n)-x_n\right\|+(1-\alpha_n)\|S_nz_n-x_n\|\to 0 \quad \text{as } n\to\infty.\end{aligned} \tag{25}$$

We also have

$$\limsup_{n\to\infty}\left\langle f\left(x^*\right)-x^*,j_q\left(x_{n+1}-x^*\right)\right\rangle\leq 0. \tag{26}$$

Again from (20), we have

$$\left\|x_{n+1}-x^*\right\|^q \tag{27}$$

$$\leq \left(1-(1-\rho)\alpha_n\right)\left\|x_n-x^*\right\|^q+q\alpha_n\left\langle f\left(x^*\right)-x^*,j_q\left(x_{n+1}-x^*\right)\right\rangle. \tag{28}$$

Apply Lemma 2.7 and (26) to (27), we obtain that $x_n\to x^*$ as $n\to\infty$.

Case 2. There exists a subsequence $\{n_i\}$ of $\{n\}$ such that

$$\left\|x_{n_i}-x^*\right\|\leq\left\|x_{n_{i+1}}-x^*\right\|$$

for all $i\in\mathbb{N}$. By Lemma 2.8, there exists a nondecreasing sequence $\{m_k\}\subset\mathbb{N}$ such that $m_k\to\infty$ as $k\to\infty$ and

$$\left\|x_{m_k}-x^*\right\|\leq\left\|x_{m_k+1}-x^*\right\| \quad\text{and}\quad \left\|x_k-x^*\right\|\leq\left\|x_{m_k+1}-x^*\right\| \tag{29}$$

for all $k\in\mathbb{N}$. From (20) we have

$$\begin{aligned}&(1-\alpha_{m_k})\beta_{m_k}(1-\beta_{m_k})(1-s_{m_k})c\|x_{m_k}-S_{m_k}x_{m_k}\|^q\\ &\quad\leq \left\|x_{m_k}-x^*\right\|^q-\left\|x_{m_k+1}-x^*\right\|^q+\alpha_{m_k}M\\ &\quad\leq \alpha_{m_k}M,\end{aligned}$$

where $c > 0$ and $M < \infty$. This implies by $(C1)$ and $(C2)$ that

$$\|x_{m_k} - S_{m_k}x_{m_k}\| \to 0 \quad \text{as } k \to \infty. \tag{30}$$

Since

$$\begin{aligned}
&\sup_{x\in\{x_{m_k}\}} \|S_{m_k+1}x - S_{m_k}x\| \\
&\quad = \sup_{x\in\{x_{m_k}\}} \big\|(1-\theta_{m_k+1})x + \theta_{m_k+1}T_{m_k+1}x - (1-\theta_{m_k})x - \theta_{m_k}T_{m_k}x\big\| \\
&\quad \le |\theta_{m_k+1} - \theta_{m_k}| \sup_{x\in\{x_{m_k}\}} \|x\| + \theta_{m_k+1} \sup_{x\in\{x_{m_k}\}} \|T_{m_k+1}x - T_{m_k}x\| \\
&\qquad + |\theta_{m_k+1} - \theta_{m_k}| \sup_{x\in\{x_{m_k}\}} \|T_{m_k}x\| \\
&\quad \le |\theta_{m_k+1} - \theta_{m_k}| \Big(\sup_{x\in\{x_{m_k}\}} \|x\| + \sup_{x\in\{x_{m_k}\}} \|T_{m_k}x\| \Big) + \sup_{x\in\{x_{m_k}\}} \|T_{m_k+1}x - T_{m_k}x\| < \infty,
\end{aligned}$$

that is, $\{S_{m_k}\}_{k=1}^{\infty}$ satisfies the $AKTT$-condition. Then, by (30) and Lemma 2.10, we get that

$$\begin{aligned}
&\|x_{m_k} - Sx_{m_k}\| \\
&\quad \le \|x_{m_k} - S_{m_k}x_{m_k}\| + \|S_{m_k}x_{m_k} - Sx_{m_k}\| \\
&\quad \le \|x_{m_k} - S_{m_k}x_{m_k}\| + \sup_{x\in\{x_{m_k}\}} \|S_{m_k}x - Sx\| \to 0 \quad \text{as } k \to \infty.
\end{aligned} \tag{31}$$

By the same argument as in Case 1, we can show that

$$\limsup_{k\to\infty} \big\langle f\big(x^*\big) - x^*, j\big(x_{m_k} - x^*\big)\big\rangle \le 0. \tag{32}$$

It follows from (31) that

$$\begin{aligned}
\|S_{m_k}z_{m_k} - x_{m_k}\| &\le \|S_{m_k}z_{m_k} - S_{m_k}x_{m_k}\| + \|S_{m_k}x_{m_k} - x_{m_k}\| \\
&\le \|z_{m_k} - x_{m_k}\| + \|S_{m_k}x_{m_k} - x_{m_k}\| \\
&= (1-s_{m_k})(1-\beta_{m_k})\|S_{m_k}x_{m_k} - x_{m_k}\| + \|S_{m_k}x_{m_k} - x_{m_k}\| \\
&\le 2\|x_{m_k} - S_{m_k}x_{m_k}\| \to 0 \quad \text{as } k \to \infty,
\end{aligned}$$

and hence

$$\begin{aligned}
\|x_{m_k+1} - x_{m_k}\| &\le \big\|\alpha_{m_k}\big(f(x_{m_k}) - x_{m_k}\big) + (1-\alpha_{m_k})(S_{m_k}z_{m_k} - x_{m_k})\big\| \\
&\le \alpha_{m_k}\big\|f(x_{m_k}) - x_{m_k}\big\| + (1-\alpha_{m_k})\|S_{m_k}z_{m_k} - x_{m_k}\| \to 0 \quad \text{as } k \to \infty.
\end{aligned}$$

Then, we also have

$$\limsup_{k\to\infty} \big\langle f\big(x^*\big) - x^*, j_q\big(x_{m_k+1} - x^*\big)\big\rangle \le 0. \tag{33}$$

Again from (27) we have

$$\begin{aligned}&\left\|x_{m_k+1}-x^*\right\|^q\\&\quad\le\left(1-(1-\rho)\alpha_{m_k}\right)\left\|x_{m_k}-x^*\right\|^q+q\alpha_{m_k}\left\langle f\left(x^*\right)-x^*,j_q\left(x_{m_k+1}-x^*\right)\right\rangle,\end{aligned}\tag{34}$$

which implies that

$$\begin{aligned}(1-\rho)\alpha_{m_k}\left\|x_{m_k}-x^*\right\|^q&\le\left\|x_{m_k}-x^*\right\|^q-\left\|x_{m_k+1}-x^*\right\|^q\\&\quad+q\alpha_{m_k}\left\langle f\left(x^*\right)-x^*,j_q\left(x_{m_k+1}-x^*\right)\right\rangle\\&\le q\alpha_{m_k}\left\langle f\left(x^*\right)-x^*,j_q\left(x_{m_k+1}-x^*\right)\right\rangle.\end{aligned}\tag{35}$$

Since $\alpha_{m_k}>0$, we get $\lim_{k\to\infty}\|x_{m_k}-x^*\|=0$. So, we have

$$\begin{aligned}\left\|x_k-x^*\right\|&\le\left\|x_{m_k+1}-x^*\right\|\\&=\left\|x_{m_k}-x^*\right\|+\left\|x_{m_k+1}-x^*\right\|-\left\|x_{m_k}-x^*\right\|\\&\le\left\|x_{m_k}-x^*\right\|+\|x_{m_k+1}-x_{m_k}\|\to 0\quad\text{as } k\to\infty,\end{aligned}$$

which implies that $x_k\to x^*$ as $k\to\infty$. This completes the proof. □

Applying Theorem 3.1 to a 2-uniformly smooth Banach space, we obtain the following result.

Corollary 3.2 *Let C be a nonempty closed convex subset of a real uniformly convex and 2-uniformly smooth Banach space E. Let $f\in\Pi_C$ with coefficient $\rho\in(0,1)$, and let $\{T_n\}_{n=1}^{\infty}:C\to C$ be a family of λ-strict pseudo-contractions such that $\Omega:=\bigcap_{n=1}^{\infty}F(T_n)\neq\emptyset$. For all $x\in C$, define the mapping $S_nx=(1-\theta)x+\theta T_nx$, where $0<\theta\le\delta$, $\delta=\min\{1,\frac{\lambda}{K^2}\}$, and $\sum_{n=1}^{\infty}|\theta_{n+1}-\theta_n|<\infty$. For given $x_1\in C$, let $\{x_n\}$ be a sequence generated by*

$$\begin{cases}\bar{x}_{n+1}=\beta_nx_n+(1-\beta_n)S_nx_n,\\x_{n+1}=\alpha_nf(x_n)+(1-\alpha_n)S_n(t_nx_n+(1-t_n)\bar{x}_{n+1}),\quad n\ge1,\end{cases}\tag{36}$$

where $\{\alpha_n\}$, $\{\beta_n\}$, and $\{t_n\}$ are sequences in $(0,1)$ satisfying the conditions $(C1)$ and $(C2)$ of Theorem 3.1. Suppose in addition that $(\{T_n\}_{n=1}^{\infty},T)$ satisfies the AKTT-condition. Then $\{x_n\}$ converges strongly to $x^*=Q(f)\in\Omega$, which solves the variational inequality

$$\left\langle(I-f)Q(f),j\left(Q(f)-z\right)\right\rangle\le0,\quad\forall z\in\Omega,\tag{37}$$

where Q is a sunny nonexpansive retraction of C onto Ω.

Utilizing the fact that a Hilbert space H is uniformly convex and 2-uniformly smooth with the best smooth constant $\kappa_2=1$, we obtain the following result.

Corollary 3.3 *Let C be a nonempty closed convex subset of a Hilbert space H. Let $f\in\Pi_C$ with coefficient $\rho\in(0,1)$, and let $\{T_n\}_{n=1}^{\infty}:C\to C$ be a family of λ-strict pseudo-contractions with $\lambda\in[0,1)$ such that $\Omega:=\bigcap_{n=1}^{\infty}F(T_n)\neq\emptyset$. For all $x\in C$, define the mapping*

$S_n x = (1-\theta_n)x + \theta_n T_n x$, *where* $0 < \theta_n \leq \delta$, $\delta = \min\{1, 2\lambda\}$, *and* $\sum_{n=1}^{\infty} |\theta_{n+1} - \theta_n| < \infty$. *For given* $x_1 \in C$, *let* $\{x_n\}$ *be a sequence generated by*

$$\begin{cases} \bar{x}_{n+1} = \beta_n x_n + (1-\beta_n) S_n x_n, \\ x_{n+1} = \alpha_n f(x_n) + (1-\alpha_n) S_n (t_n x_n + (1-t_n)\bar{x}_{n+1}), \quad n \geq 1, \end{cases} \tag{38}$$

where $\{\alpha_n\}$, $\{\beta_n\}$, *and* $\{t_n\}$ *are sequences in* $(0,1)$ *satisfying conditions* $(C1)$ *and* $(C2)$ *of Theorem* 3.1. *Suppose, in addition, that* $(\{T_n\}_{n=1}^{\infty}, T)$ *satisfies the AKTT-condition. Then* $\{x_n\}$ *converges strongly to* $x^* = P(f) \in \Omega$, *which solves the variational inequality*

$$\langle (I-f)P(f), P(f) - z \rangle \leq 0, \quad z \in \Omega, \tag{39}$$

where P *is a metric projection of* C *onto* Ω.

4 Application

4.1 The generalized viscosity explicit rules for convex combination of family of mappings

In this subsection, we apply our main result to convex combination of a countable family of strict pseudo-contractions. The following lemmas can be found in [36, 37].

Lemma 4.1 ([36, 37]) *Let* C *be a closed convex subset of a smooth Banach space* E. *Suppose that* $\{T_n\}_{n=1}^{\infty} : C \to C$ *is a family of* λ*-strictly pseudo-contractive mappings with* $\bigcap_{n=1}^{\infty} F(T_n) \neq \emptyset$ *and* $\{\mu_n\}_{n=1}^{\infty}$ *is a real sequence in* $(0,1)$ *such that* $\sum_{n=1}^{\infty} \mu_n = 1$. *Then the following conclusions hold*:

(i) *A mapping* $G : C \to E$ *defined by* $G := \sum_{n=1}^{\infty} \mu_n T_n$ *is a* λ*-strictly pseudocontractive mapping.*

(ii) $F(G) = \bigcap_{n=1}^{\infty} F(T_n)$.

Lemma 4.2 ([37]) *Let* C *be a closed convex subset of a smooth Banach space* E. *Suppose that* $\{T_k\}_{k=1}^{\infty} : C \to C$ *is a countable family of* λ*-strictly pseudocontractive mappings with* $\bigcap_{k=1}^{\infty} F(S_k) \neq \emptyset$. *For all* $n \in \mathbb{N}$, *define* $S_n : C \to C$ *by* $S_n x := \sum_{k=1}^{n} \mu_n^k T_k x$ *for all* $x \in C$, *where* $\{\mu_n^k\}$ *is a family of nonnegative numbers satisfying the following conditions*:

(i) $\sum_{k=1}^{n} \mu_n^k = 1$ *for all* $n \in \mathbb{N}$;

(ii) $\mu^k := \lim_{n\to\infty} \mu_n^k > 0$ *for all* $k \in \mathbb{N}$;

(iii) $\sum_{n=1}^{\infty} \sum_{k=1}^{n} |\mu_{n+1}^k - \mu_n^k| < \infty$.

Then:

(1) *Each* T_n *is a* λ*-strictly pseudocontractive mapping.*

(2) $\{T_n\}$ *satisfies the AKTT-condition.*

(3) *If* $T : C \to C$ *is defined by* $Tx = \sum_{k=1}^{\infty} \mu^k S_k x$ *for all* $x \in C$,

then, $Tx = \lim_{n\to\infty} T_n x$ *and* $F(T) = \bigcap_{n=1}^{\infty} F(T_n) = \bigcap_{k=1}^{\infty} F(S_k)$.

Using Theorem 3.1 and Lemmas 4.1 and 4.2, we obtain the following result.

Theorem 4.3 *Let* C *be a nonempty closed convex subset of a real uniformly convex and* q*-uniformly smooth Banach space* E. *Let* $f \in \Pi_C$ *with coefficient* $\rho \in (0,1)$, *and let* $\{T_k\}_{k=1}^{\infty} : C \to C$ *be a countable family of* λ_k*-strict pseudo-contractions with* $\inf\{\lambda_k : k \in \mathbb{N}\} = \lambda > 0$.

For all $x \in C$, define a mapping $S_n x := (1-\theta_n)x + \theta_n \sum_{k=1}^{n} \mu_n^k T_k x$ such that $\Omega := \bigcap_{k=1}^{\infty} F(T_k) \neq \emptyset$, where $0 < \theta_n \leq \delta$, $\delta = \min\{1, (\frac{q\lambda}{\kappa_q})^{\frac{1}{q-1}}\}$, and $\sum_{n=1}^{\infty} |\theta_{n+1} - \theta_n| < \infty$. For given $x_1 \in C$, let $\{x_n\}$ be a sequence generated by

$$\begin{cases} \bar{x}_{n+1} = \beta_n x_n + (1-\beta_n) S_n x_n, \\ x_{n+1} = \alpha_n f(x_n) + (1-\alpha_n) S_n (t_n x_n + (1-t_n)\bar{x}_{n+1}), \quad n \geq 1, \end{cases} \tag{40}$$

where $\{\alpha_n\}$, $\{\beta_n\}$, and $\{t_n\}$ are sequences in $(0,1)$ satisfy conditions $(C1)$ and $(C2)$ of Theorem 3.1, and $\{\mu_n^k\}$ is a real sequence satisfying (i)–(iii) *of Lemma 4.2. Then $\{x_n\}$ converges strongly to a $x^* \in \Omega$.*

4.2 The generalized viscosity explicit rules for zeros of accretive operators

In this subsection, we apply our main result to problem of finding a zero of an accretive operator. An operator $A \subset E \times E$ is said to be accretive if for all (x_1, y_1) and $(x_2, y_2) \in A$, there exists $j_q \in J_q(x_1 - x_2)$ such that $\langle y_1 - y_2, j_q \rangle \geq 0$. An operator A is said to satisfy the range condition if $\overline{D(A)} = R(I + \lambda A)$ for all $\lambda > 0$, where $D(A)$ is the domain of A, $R(I + \lambda A)$ is the range of $I + \lambda A$, and $\overline{D(A)}$ is the closure of $D(A)$. If A is an accretive operator that satisfies the range condition, then we can defined a single-valued mapping $J_\lambda^A : R(I + \lambda A) \to D(A)$ by $J_\lambda = (I + \lambda A)^{-1}$, which is called the *resolvent* of A. We denote $A^{-1}0$ by the set of zeros of A, that is, $A^{-1}0 = \{x \in D(A) : 0 \in Ax\}$. It is well known that J_λ is nonexpansive and $F(J_\lambda) = A^{-1}0$ (see [38]). We also know the following [39]: For all $\lambda, \mu > 0$ and $x \in R(I + \lambda A) \cap R(I + \mu A)$, we have

$$\|J_\lambda x - J_\mu x\| \leq \frac{|\lambda - \mu|}{\lambda} \|x - J_\lambda x\|.$$

Lemma 4.4 ([34]) *Let C be a nonempty closed convex subset of a Banach space E. Let $A \subset E \times E$ be an accretive operator such that $A^{-1}0 \neq \emptyset$, which satisfies the condition $\overline{D(A)} \subset C \subset \bigcap_{\lambda>0} R(I + \lambda A)$. Suppose that $\{\lambda_n\} \subset (0, \infty)$ such that $\inf\{\lambda_n : n \in \mathbb{N}\} > 0$ and $\sum_{n=1}^{\infty} |\theta_{n+1} - \theta_n| < \infty$. Then, $\{J_{\lambda_n}\}$ satisfies the AKTT-condition. Consequently, for each $x \in C$, $\{J_{\lambda_n} x\}$ converges strongly to some point of C. Moreover, let $J_\lambda : C \to C$ be defined by $J_\lambda x = \lim_{n\to\infty} J_{\lambda_n} x$ for all $x \in C$ and $F(J_\lambda) = \bigcap_{n=1}^{\infty} F(J_{\lambda_n})$, where $\lambda_n \to \lambda$ as $n \to \infty$. Then, $\lim_{n\to\infty} \sup_{x\in C} \|J_\lambda x - J_{\lambda_n} x\| = 0$.*

Utilizing Theorem 3.1 and and Lemma 4.4, we obtain the following result.

Theorem 4.5 *Let C be a nonempty closed convex subset of a q-uniformly smooth Banach space E. Let $f \in \Pi_C$ with coefficient $\rho \in (0,1)$ and let $A \subset E \times E$ be an accretive operator such that $A^{-1}0 \neq \emptyset$ which satisfies the condition $\overline{D(A)} \subset C \subset \bigcap_{\lambda>0} R(I + \lambda A)$. Suppose that $\{\lambda_n\} \subset (0, \infty)$ is such that $\inf\{\lambda_n : n \in \mathbb{N}\} > 0$ and $\sum_{n=1}^{\infty} |\lambda_{n+1} - \lambda_n| < \infty$. For given $x_1 \in C$, let $\{x_n\}$ be the sequence generated by*

$$\begin{cases} \bar{x}_{n+1} = \beta_n x_n + (1-\beta_n) J_{\lambda_n} x_n, \\ x_{n+1} = \alpha_n f(x_n) + (1-\alpha_n) J_{\lambda_n} (t_n x_n + (1-t_n)\bar{x}_{n+1}), \quad n \geq 1, \end{cases} \tag{41}$$

where $\{\alpha_n\}$, $\{\beta_n\}$, and $\{t_n\}$ are sequences in $(0,1)$ satisfying conditions $(C1)$ and $(C2)$ of Theorem 3.1. Then $\{x_n\}$ converges strongly to $x^* \in A^{-1}0$.

4.3 The generalized viscosity explicit rules with weak contraction

In this subsection, we apply our main result to the viscosity approximation method with weak contraction.

Definition 4.6 ([40–42]) Let C be a closed and convex subset of a real Banach space E. A mapping $g : C \to C$ is said to be *weakly contractive* if there exists a continuous strictly increasing function $\psi : [0,\infty) \to [0,\infty)$ with $\psi(0) = 0$ and $\lim_{t\to\infty} \psi(t) = \infty$ such that

$$\big\|g(x) - g(y)\big\| \le \|x - y\| - \psi\big(\|x - y\|\big), \quad x, y \in C.$$

As a particular case, if $\psi(t) = (1-\rho)t$ for all $t \ge 0$, where $\rho \in (0,1)$, then the weakly contractive mapping is contraction with coefficient ρ.

In 2001, Rhoades [42] first proved Banach's contraction principle for the weakly contractive mapping in complete metric space.

Lemma 4.7 ([42]) *Let (E,d) be a complete metric space, and let g be a weakly contractive mapping on E. Then g has a unique fixed point in E.*

Lemma 4.8 ([43]) *Assume that $\{a_n\}$ and $\{b_n\}$ are sequences of nonnegative real number, and $\{\lambda_n\}$ is a sequence of a positive real number satisfying the conditions $\sum_{n=1}^{\infty} \lambda_n = \infty$ and $\lim_{n\to\infty} \frac{b_n}{\lambda_n} = 0$. Suppose that*

$$a_{n+1} \le a_n - \lambda_n \psi(a_n) + b_n, \quad n \ge 1,$$

where $\psi(t)$ is a continuous strictly increasing function on $\mathbb{R}$ with $\psi(0) = 0$. Then, $\lim_{n\to\infty} a_n = 0$.

Utilizing Theorem 3.1, we obtain the following result.

Theorem 4.9 *Let C be a nonempty closed convex subset of a real uniformly convex and q-uniformly smooth Banach space E. Let $g : C \to C$ be a weak contraction, and let $\{T_n\}_{n=1}^{\infty} : C \to C$ be a family of λ-strict pseudo-contractions such that $\Omega := \bigcap_{n=1}^{\infty} F(T_n) \neq \emptyset$. For all $x \in C$, define the mapping $S_n x = (1-\theta_n)x + \theta_n T_n x$, where $0 < \theta_n \le \delta$, $\delta = \min\{1, (\frac{q\lambda}{\kappa_q})^{\frac{1}{q-1}}\}$, and $\sum_{n=1}^{\infty} |\theta_{n+1} - \theta_n| < \infty$. For given $x_1 \in C$, let $\{x_n\}$ be the sequence generated by*

$$\begin{cases} \bar{x}_{n+1} = \beta_n x_n + (1-\beta_n) S_n x_n, \\ x_{n+1} = \alpha_n g(x_n) + (1-\alpha_n) S_n(t_n x_n + (1-t_n)\bar{x}_{n+1}), \quad n \ge 1, \end{cases} \tag{42}$$

where $\{\alpha_n\}$, $\{\beta_n\}$, and $\{t_n\}$ are sequences in $(0,1)$ satisfy conditions $(C1)$ and $(C2)$ of Theorem 3.1. Suppose in addition that $(\{T_n\}_{n=1}^{\infty}, T)$ satisfies the AKTT-condition. Then $\{x_n\}$ converges strongly to $x^* \in \Omega$.

Proof By the smoothness of E there exists a sunny nonexpansive retraction Q from C onto Ω. Moreover, $Q(g)$ is a weakly contractive mapping of C into itself. For all $x, y \in C$, we have

$$\big\|Q\big(g(x)\big) - Q\big(g(y)\big)\big\| \le \big\|g(x) - g(y)\big\| \le \|x - y\| - \psi\big(\|x - y\|\big).$$

Lemma 4.7 guarantees that $Q(g)$ has a unique fixed point $x^* \in C$ such that $x^* = Q(g)$. Now, we define a sequence $\{y_n\}$ and $y_1 \in C$ as follows:

$$\begin{cases} \bar{y}_{n+1} = \beta_n y_n + (1-\beta_n) S_n y_n, \\ y_{n+1} = \alpha_n g(y_n) + (1-\alpha_n) S_n (t_n y_n + (1-t_n)\bar{y}_{n+1}), \quad n \geq 1. \end{cases}$$

Then, by Theorem 3.1 with a constant $f = g(x^*)$, we have that $\{y_n\}$ converges strongly to $x^* = Q(g)) \in \Omega$. Next, we show that $x_n \to x^*$ as $n \to \infty$. Since

$$\|\bar{x}_{n+1} - \bar{y}_{n+1}\| \leq \beta_n \|x_n - y_n\| + (1-\beta_n)\|S_n x_n - S_n y_n\| \leq \|x_n - y_n\|,$$

it follows that

$$\begin{aligned} &\|x_{n+1} - y_{n+1}\| \\ &\quad = \left\|\alpha_n\big(g(x_n) - g\big(x^*\big)\big) + (1-\alpha_n)\big(S_n\big(t_n x_n + (1-t_n)\bar{x}_{n+1}\big) - S_n\big(t_n y_n + (1-t_n)\bar{y}_{n+1}\big)\big)\right\| \\ &\quad \leq \alpha_n \left\|g(x_n) - g\big(x^*\big)\right\| + (1-\alpha_n)\left\|S_n\big(t_n x_n + (1-t_n)\bar{x}_{n+1}\big) - S_n\big(t_n y_n + (1-t_n)\bar{y}_{n+1}\big)\right\| \\ &\quad \leq \alpha_n \left\|g(x_n) - g(y_n)\right\| + \alpha_n \left\|g(y_n) - g\big(x^*\big)\right\| \\ &\qquad + (1-\alpha_n)\big(t_n\|x_n - y_n\| + (1-t_n)\|\bar{x}_{n+1} - \bar{y}_{n+1}\|\big) \\ &\quad \leq \alpha_n \|x_n - y_n\| - \alpha_n \psi\big(\|x_n - y_n\|\big) + \alpha_n \left\|y_n - x^*\right\| \\ &\qquad - \alpha_n \psi\big(\left\|y_n - x^*\right\|\big) + (1-\alpha_n)\|x_n - y_n\| \\ &\quad \leq \|x_n - y_n\| - \alpha_n \psi\big(\|x_n - y_n\|\big) + \alpha_n \left\|y_n - x^*\right\|. \end{aligned} \tag{43}$$

Since $\{y_n\}$ converges strongly to x^*, applying Lemma 4.8 to (43), we obtain that $\lim_{n\to\infty} \|x_n - y_n\| = 0$. Therefore $x_n \to x^*$. This completes the proof. □

5 Numerical examples

In this section, we present a numerical example of our main result.

Example 5.1 Let $E = \ell_4$ and $C = \{\mathbf{x} = (x_1, x_2, x_3, x_4, \ldots) \in \ell_4 : x_i \in \mathbb{R} \text{ for } i = 1, 2, 3, \ldots\}$ with norm $\|\mathbf{x}\|_{\ell_4} = (\sum_{i=1}^{\infty} |x_i|^4)^{1/4}$. Let $f : C \to C$ be the contraction defined by $f(\mathbf{x}) = \frac{1}{3}\mathbf{x}$. Let $\{T_n\}_{n=1}^{\infty} : C \to C$ be the strictly pseudo-contractive mapping defined by

$$T_n \mathbf{x} = \begin{cases} \frac{1}{n}(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, 0, 0, 0, \ldots) - 2\mathbf{x} & \text{if } \mathbf{x} \neq \mathbf{0}, \\ \mathbf{0} & \text{if } \mathbf{x} = \mathbf{0}, \end{cases}$$

where $\mathbf{0} = (0, 0, 0, 0, 0, 0, 0, \ldots)$ is the null vector on ℓ_4.

- We show that T_n is strictly pseudo-contractive. For each $n \geq 1$, if $\mathbf{x}, \mathbf{y} \neq \mathbf{0}$, then

$$\begin{aligned} \big\langle (I - T_n)\mathbf{x} - (I - T_n)\mathbf{y}, j_2(\mathbf{x} - \mathbf{y})\big\rangle &= \big\langle 3\mathbf{x} - 3\mathbf{y}, j_2(\mathbf{x} - \mathbf{y})\big\rangle \\ &= 3\|\mathbf{x} - \mathbf{y}\|_{\ell_4}^2 \\ &= \frac{1}{3}\|3\mathbf{x} - 3\mathbf{y}\|_{\ell_4}^2 \\ &\geq \lambda \left\|(I - T_n)\mathbf{x} - (I - T_n)\mathbf{y}\right\|_{\ell_4}^2 \end{aligned}$$

for $\lambda \leq \frac{1}{3}$. Then, we can choose $\lambda = \frac{1}{3}$. Thus, T_n is $\frac{1}{3}$-strictly pseudo-contractive with $\bigcap_{n=1}^{\infty} F(T_n) = \{\mathbf{0}\}$. Further, we observe that T_n is not nonexpansive.

- We show that $(\{T_n\}_{n=1}^{\infty}, T)$ satisfies the AKTT-condition. Since

$$\begin{aligned}
&\sup_{\mathbf{x}\in\ell_4} \|T_{n+1}\mathbf{x} - T_n\mathbf{x}\|_{\ell_4} \\
&\quad = \sup_{\mathbf{x}\in\ell_4} \left\| \frac{1}{n+1}\left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, 0, 0, 0, \ldots\right) - 2\mathbf{x} - \frac{1}{n}\left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, 0, 0, 0, \ldots\right) + 2\mathbf{x} \right\|_{\ell_4} \\
&\quad = \left\| \frac{1}{n+1}\left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, 0, 0, 0, \ldots\right) - \frac{1}{n}\left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, 0, 0, 0, \ldots\right) \right\|_{\ell_4} \\
&\quad = \left(\frac{1}{n} - \frac{1}{n+1}\right) \left\| \left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, 0, 0, 0, \ldots\right) \right\|_{\ell_4}.
\end{aligned}$$

So we have

$$\begin{aligned}
\sum_{n=1}^{\infty} \sup_{\mathbf{x}\in\ell_4} \|T_{n+1}\mathbf{x} - T_n\mathbf{x}\|_{\ell_4} &= \lim_{n\to\infty} \sum_{k=1}^{n} \sup_{\mathbf{x}\in\ell_4} \|T_{k+1}\mathbf{x} - T_k\mathbf{x}\|_{\ell_4} \\
&= \left\| \left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, 0, 0, 0, \ldots\right) \right\|_{\ell_4} < \infty,
\end{aligned}$$

that is, $(\{T_n\}_{n=1}^{\infty}, T)$ satisfies the AKTT-condition, where $T : C \to C$ is defined by

$$T\mathbf{x} = \lim_{n\to\infty} T_n\mathbf{x} = -2\mathbf{x}, \quad \mathbf{x} \in C.$$

Since in ℓ_4, $q = 2$ and $\kappa_2 = 3$, we can choose $\theta_n = \frac{1}{9n} + \frac{1}{9}$. Define the mapping $\{S_n\}_{n=1}^{\infty} : C \to C$ by

$$S_n\mathbf{x} = \begin{cases} (\frac{2}{3} - \frac{1}{3n})\mathbf{x} + (\frac{1}{9n^2} + \frac{1}{9n})(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, 0, 0, 0, \ldots) & \text{if } \mathbf{x} \neq \mathbf{0}, \\ \mathbf{0} & \text{if } \mathbf{x} = \mathbf{0}. \end{cases}$$

Since $(\{T_n\}_{n=1}^{\infty}, T)$ satisfies the AKTT condition, we also have that $(\{S_n\}_{n=1}^{\infty}, S)$ satisfies the AKTT condition, where $S : C \to C$ is defined by

$$S\mathbf{x} = \lim_{n\to\infty} S_n\mathbf{x} = \frac{2}{3}\mathbf{x}, \quad \mathbf{x} \in C.$$

Then, we have $F(S) = F(T) = \bigcap_{n=1}^{\infty} F(T_n) = \{\mathbf{0}\}$. Let $\alpha_n = \frac{1}{32n+1}$, $\beta_n = \frac{1}{100n+3} + 0.32$, and $t_n = \frac{n}{2n+1}$. So our algorithm (16) has the following form:

$$\begin{cases} \bar{\mathbf{x}}_{n+1} = (\frac{1}{100n+3} + 0.32)\mathbf{x}_n + (0.68 - \frac{1}{100n+3})S_n\mathbf{x}_n, \\ \mathbf{x}_{n+1} = \frac{1}{32n+2} f(\mathbf{x}_n) + \frac{32n}{32n+1} S_n(\frac{n}{2n+1}\mathbf{x}_n + \frac{n+1}{2n+1}\bar{\mathbf{x}}_{n+1}), \quad n \geq 1. \end{cases} \tag{44}$$

Let $\mathbf{x_1} = (1, -0.25, 1.46, 1.85, 0, 0, 0, \ldots)$ be the initial point. Then, we obtain numerical results in Table 1 and Fig. 1.

Table 1 The values of the sequences $\{\mathbf{x}_n\}$

n	$\mathbf{x}_n$	$\|\mathbf{x}_{n+1} - \mathbf{x}_n\|_{\ell_4}$
1	(1.000000, −0.250000, 1.460000, 1.850000, 0, 0, 0,…)	1.459e+00
50	(0.007006, 0.003503, 0.002335, 0.001751, 0, 0, 0,…)	1.471e−04
100	(0.003416, 0.001708, 0.001139, 0.000854, 0, 0, 0,…)	3.531e−05
150	(0.002258, 0.001129, 0.000753, 0.000565, 0, 0, 0,…)	1.549e−05
200	(0.001687, 0.000843, 0.000562, 0.000422, 0, 0, 0,…)	8.657e−06
⋮	⋮	⋮
400	(0.000838, 0.000419, 0.000279, 0.000210, 0, 0, 0,…)	2.143e−06
450	(0.000745, 0.000372, 0.000248, 0.000186, 0, 0, 0,…)	1.692e−06
500	(0.000670, 0.000335, 0.000223, 0.000167, 0, 0, 0,…)	1.369e−06

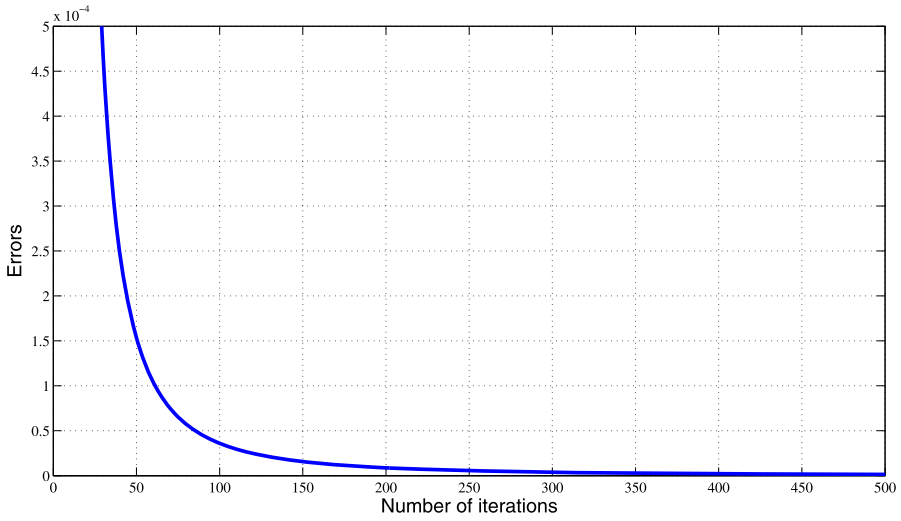


Figure 1 The behavior of errors

6 Conclusion

In this work, we introduce an algorithm by a generalized viscosity explicit rule for finding a common fixed point of a countable family of strictly pseudo-contractive mappings in a q-uniformly smooth Banach space. We obtain some strong convergence theorem for the sequence generated by the proposed algorithm under suitable conditions. However, we should like remark the following:

(1) We extend the results of Ke and Ma [21] and Marino et al. [25] from a one nonexpansive mapping in Hilbert spaces to a countable family of strictly pseudo-contractive mappings in a q-uniformly smooth Banach space.

(2) Our result is proved with a new assumption on the control conditions $\{\beta_n\}$ and $\{t_n\}$.

(3) The method of proof of our result is simpler in comparison with the results of [19, 21, 44, 45]). Moreover, we remove the conditions $\sum_{n=1}^{\infty} |\alpha_{n+1} - \alpha_n| < \infty$ and $0 < \epsilon \leq s_n \leq s_{n+1} < 1$ in Theorem 3.1 of [21].

(4) We give a numerical example that shows the efficiency and implementation of our main result in the space ℓ_4, which is a uniformly convex and 2-uniformly smooth Banach space but not a Hilbert space.

Acknowledgements
The authors would like to thank the Rajamangala University of Technology Thanyaburi for financial support.

Funding
P. Sunthrayuth was supported by RMUTT research foundation scholarship of the Rajamangala University of Technology Thanyaburi under Grant No. NRF04066005.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
Both authors contributed equally to the writing of this paper. Both authors read and approved the final manuscript.

Publisher's Note

Received: 18 February 2018 Accepted: 29 May 2018 Published online: 11 July 2018

References

1. Takahashi, W.: Nonlinear Functional Analysis. Yokohama Publishers, Yokohama (2000)
2. Browder, F.E., Petryshyn, W.V.: Construction of fixed points nonlinear mappings in Hilbert space. J. Math. Anal. Appl. **20**, 197–228 (1967)
3. Scherzer, O.: Convergence criteria of iterative methods based on Landweber iteration for solving nonlinear problems. J. Math. Anal. Appl. **194**, 911–933 (1991)
4. Cai, G.: Viscosity iterative algorithm for variational inequality problems and fixed point problems in a real q-uniformly smooth Banach space. Fixed Point Theory Appl. **2015**, 67 (2015)
5. Zhang, H., Su, Y.: Convergence theorems for strict pseudo-contractions in q-uniformly smooth Banach spaces. Nonlinear Anal. **71**, 4572–4580 (2009)
6. Zhou, H.: Convergence theorems of fixed points for κ-strict pseudo-contractions in Hilbert spaces. Nonlinear Anal. **69**, 456–462 (2008)
7. Jung, J.S.: Strong convergence of iterative methods for κ-strictly pseudo-contractive mappings in Hilbert spaces. Appl. Math. Comput. **215**, 3746–3753 (2010)
8. Sahu, D.R., Petruşel, A.: Strong convergence of iterative methods by strictly pseudocontractive mappings in Banach spaces. Nonlinear Anal. **74**, 6012–6023 (2011)
9. Cholamjiak, P., Suantai, S.: Strong convergence for a countable family of strict pseudocontractions in q-uniformly smooth Banach spaces. Comput. Math. Appl. **62**, 787–796 (2011)
10. Auzinger, W., Frank, R.: Asymptotic error expansions for stiff equations: an analysis for the implicit midpoint and trapezoidal rules in the strongly stiff case. Numer. Math. **56**, 469–499 (1989)
11. Bader, G., Deuflhard, P.: A semi-implicit mid-point rule for stiff systems of ordinary differential equations. Numer. Math. **41**, 373–398 (1983)
12. Deuflhard, P.: Recent progress in extrapolation methods for ordinary differential equations. SIAM Rev. **27**(4), 505–535 (1985)
13. Schneider, C.: Analysis of the linearly implicit mid-point rule for differential-algebra equations. Electron. Trans. Numer. Anal. **1**, 1–10 (1993)
14. Somalia, S.: Implicit midpoint rule to the nonlinear degenerate boundary value problems. Int. J. Comput. Math. **79**(3), 327–332 (2002)
15. van Veldhuxzen, M.: Asymptotic expansions of the global error for the implicit midpoint rule (stiff case). Computing **33**, 185–192 (1984)
16. Schneider, C.: Analysis of the linearly implicit mid-point rule for differential-algebraic equations. Electron. Trans. Numer. Anal. **1**, 1–10 (1993)
17. Hairer, E., Nørsett, S.P., Wanner, G.: Solving Ordinary Differential Equations I: Nonstiff Problems, 2nd edn. Springer Series in Computational Mathematics. Springer, Berlin (1993)
18. Alghamdi, M.A., Alghamdi, M.A., Shahzad, N., Xu, H.-K.: The implicit midpoint rule for nonexpansive mappings. Fixed Point Theory Appl. **2014**, 96 (2014)
19. Xu, H.-K., Alghamdi, M.A., Shahzad, N.: The viscosity technique for the implicit midpoint rule of nonexpansive mappings in Hilbert spaces. Fixed Point Theory Appl. **2015**, 41 (2015)
20. Moudafi, A.: Viscosity approximation methods for fixed point problems. J. Math. Anal. Appl. **241**, 46–55 (2000)
21. Ke, Y., Ma, C.: The generalized viscosity implicit rules of nonexpansive mappings in Hilbert spaces. Fixed Point Theory Appl. **2015**, 190 (2015)
22. Palais, R.S., Palais, R.A.: Differential Equations, Mechanics, and Computation. American Mathematical Soc., Providence (2009)
23. Hoffman, J.D.: Numerical Methods for Engineers and Scientists, 2nd edn. Dekker, New York (2001)
24. Moaveni, S.: Finite Element Analysis Theory and Application with ANSYS, 3 edn. Pearson Education, Upper Saddle River (2008)
25. Marino, G., Scardamaglia, B., Zaccone, R.: A general viscosity explicit midpoint rule for quasi-nonexpansive mappings. J. Nonlinear Convex Anal. **18**(1), 137–148 (2017)
26. Xu, Z.B., Roach, G.F.: Characteristic inequalities of uniformly smooth Banach spaces. J. Math. Anal. Appl. **157**, 189–210 (1991)
27. Reich, S.: Strong convergence theorems for resolvents of accretive operators in Banach spaces. J. Math. Anal. Appl. **75**, 287–292 (1980)

28. Song, Y., Ceng, L.: A general iteration scheme for variational inequality problem and common fixed point problems of nonexpansive mappings in q-uniformly smooth Banach spaces. J. Glob. Optim. **57**, 1327–1348 (2013)
29. Cai, G., Bu, S.: An iterative algorithm for a general system of variational inequalities and fixed point problems in q-uniformly smooth Banach spaces. Optim. Lett. **7**, 267–287 (2013)
30. Zhang, H., Su, Y.: Convergence theorems for strict pseudo-contractions in q-uniformly smooth Banach spaces. Nonlinear Anal. **71**, 4572–4580 (2009)
31. Chidume, C.: Geometric Properties of Banach Spaces and Nonlinear Iterations. Springer, Berlin (2009)
32. Xu, H.K.: In: Inequalities in Banach Spaces with Applications, Nonlinear Analysis: Theory, Methods & Applications, vol. 16, pp. 1127–1138 (1991)
33. Mitrinović, D.S.: Analytic Inequalities. Springer, New York (1970)
34. Aoyama, K., Kimura, Y., Takahashi, W., Toyoda, M.: Approximation of common fixed points of a countable family of nonexpansive mapping in a Banach space. Nonlinear Anal. **67**, 2350–2360 (2007)
35. Maingé, P.E.: Strong convergence of projected subgradient methods for nonsmooth and nonstrictly convex minimization. Set-Valued Anal. **16**, 899–912 (2008)
36. Boonchari, D., Saejung, S.: Weak and strong convergence theorems of an implicit iteration for a countable family of continuous pseudocontractive mappings. J. Comput. Appl. Math. **233**, 1108–1116 (2009)
37. Boonchari, D., Saejung, S.: Construction of common fixed points of a countable family of λ-demicontractive mappings in arbitrary Banach spaces. Appl. Math. Comput. **216**, 173–178 (2010)
38. Takahashi, W.: Nonlinear Functional Analysis. Yokohama Publishers, Yokohama (2000)
39. Eshita, K., Takahashi, W.: Approximating zero points of accretive operators in general Banach spaces. JP J. Fixed Point Theory Appl. **2**, 105–116 (2007)
40. Alber, Ya.I., Guerre-Delabriere, S.: Principle of weakly contractive maps in Hilbert spaces. Oper. Theory, Adv. Appl. **98**, 7–22 (1997)
41. Alber, Ya.I., Guerre-Delabriere, S., Zelenko, L.: Principle of weakly contractive maps in metric spaces. Commun. Appl. Nonlinear Anal. **5**(1), 45–68 (1998)
42. Rhoades, B.E.: Some theorems on weakly contractive maps. Nonlinear Anal. **47**, 2683–2693 (2001)
43. Alber, Ya.I., Iusem, A.N.: Extension of subgradient techniques for nonsmooth optimization in Banach spaces. Set-Valued Anal. **9**, 315–335 (2001)
44. Yao, Y., Shahzad, N., Liou, Y.-C.: Modified semi-implicit midpoint rule for nonexpansive mappings. Fixed Point Theory Appl. **2015**, 166 (2015)
45. Luo, P., Cai, G., Shehu, Y.: The viscosity iterative algorithms for the implicit midpoint rule of nonexpansive mappings in uniformly smooth Banach spaces. J. Inequal. Appl. **2017**, 154 (2017)

KYUNGPOOK Math. J. 59(2019), 83-99
https://doi.org/10.5666/KMJ.2019.59.1.83
pISSN 1225-6951 eISSN 0454-8124

Weak and Strong Convergence of Hybrid Subgradient Method for Pseudomonotone Equilibrium Problems and Nonspreading-Type Mappings in Hilbert Spaces

Wanna Sriprad* and Somnuk Srisawat
Department of Mathematics and Computer Scicence, Faculty of Science and Technology, Rajamangala University of Technology Thanyaburi, Pathum Thani 12110, Thailand
e-mail : wanna_sriprad@rmutt.ac.th *and* somnuk_s@rmutt.ac.th

Abstract. In this paper, we introduce a hybrid subgradient method for finding an element common to both the solution set of a class of pseudomonotone equilibrium problems, and the set of fixed points of a finite family of κ-strictly presudononspreading mappings in a real Hilbert space. We establish some weak and strong convergence theorems of the sequences generated by our iterative method under some suitable conditions. These convergence theorems are investigated without the Lipschitz condition for bifunctions. Our results complement many known recent results in the literature.

1. Introduction

Let H be a real Hilbert space in which the inner product and norm are denoted by $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$, respectively. Let C be a nonempty closed convex subset of H. Let $T : C \to C$ be a mapping. A point $x \in C$ is called a *fixed point* of T if $Tx = x$ and we denote the set of fixed points of T by $F(T)$. Recall that a mapping $T : C \to C$ is said to be *nonexpansive* if

$$\|Tx - Ty\| \leq \|x - y\|, \text{ for all } x, y \in C,$$

and it is said to be *quasi-nonexpansive* if $F(T) \neq \emptyset$ and

$$\|Tx - Ty\| \leq \|x - y\|, \text{ for all } x \in C, \text{ and } y \in F(T).$$

* Corresponding Author.
Received October 16, 2016; revised January 22, 2019; accepted January 28, 2019.
2010 Mathematics Subject Classification: 47H05, 47H09, 47H10.
Key words and phrases: pseudomonotone equilibrium problem, κ-strictly presudononspreading mapping, nonspreading mapping, hybrid subgradient method, fixed point.
This work was supported by Faculty of Science and Technology, Rajamangala University of Technology Thanyaburi (RMUTT) Thailand.

A mapping $T : C \to C$ is said to be *a strict pseudocontraction* if there exists a constant $k \in [0, 1)$ such that

$$\|Tx - Ty\|^2 \leq \|x - y\|^2 + k\|(I - T)x - (I - T)y\|^2, \forall x, y \in C,$$

where I is the identity mapping on H. If $k = 0$, then T is nonexpansive on C.

In 2008, Kohsaka and Takahashi [15] defined a mapping T in a in Hilbert spaces H to be *nonspreading* if

$$2\|Tx - Ty\|^2 \leq \|Tx - y\|^2 + \|Ty - x\|^2, \text{ for all } x, y \in C.$$

Following the terminology of Browder-Petryshyn [10], Osilike and Isiogugu [17] called a mapping T of C into itself *κ-strictly pseudononspreading* if there exists $\kappa \in [0, 1)$ such that

$$\|Tx-Ty\|^2 \leq \|x-y\|^2+2\langle x-Tx, y-Ty\rangle+\kappa\|x-Tx-(y-Ty)\|^2, \text{ for all } x, y \in C.$$

Clearly, every nonspreading mapping is κ-strictly pseudononspreading but the converse is not true; see [17]. We note that the class of strict pseudocontraction mappings and the class of κ-strictly pseudononspreading mappings are independent.

In 2010, Kurokawa and Takahashi [16] obtained a weak mean ergodic theorem of Baillon's type [7] for nonspreading mappings in Hilbert spaces. Furthermore, using the idea of mean convergence in Hilbert spaces, they also proved a strong convergence theorem of Halpern's type [12] for this class of mappings. After that, in 2011, Osilike and Isiogugu [17] introduced the concept of κ-strictly pseudononspreading mappings and they proved a weak mean convergence theorem of Baillon's type similar to [16]. They further proved a strong convergence theorem using the idea of mean convergence. This theorem extended and improved the main theorems of [16] and gave an affirmative answer to an open problem posed by Kurokawa and Takahashi [16] for the case when the mapping T is averaged. In 2013 Kangtunyakarn [14] proposed a new technique, using the projection method, for κ-strictly pseudononspreading mappings. He obtained a strong convergence theorem for finding the common element of the set of solutions of a variational inequality, and the set of fixed points of κ-strictly pseudononspreading mappings in a real Hilbert space.

On the other hand, let F be a bifunction of $C \times C$ into $\mathbb{R}$, where $\mathbb{R}$ is the set of real numbers. The equilibrium problem for $F : C \times C \to \mathbb{R}$ is to find $x \in C$ such that

$$F(x, y) \geq 0 \text{ for all } y \in C. \tag{1.1}$$

The set of solutions of (1.1) is denoted by $EP(F, C)$. It is well known that there are several problems, such as complementarity problems, minimax problems, the Nash equilibrium problem in noncooperative games, fixed point problems, optimization problems, that can be written in the form of an EP. In other words, the EP is a unifying model for several problems arising in physics, engineering, science, optimization, economics, etc.; see [6, 8, 11] and the references therein.

In recent years the problem of finding an element common to the set of solutions of a equilibrium problems, and the set of fixed points of nonlinear mappings, has become a fascinating subject, and various methods have been developed by many authors for solving this problem (see [1, 4, 5, 20]). Most of all the existing algorithms for this problem are based on applying the proximal point method to the equilibrium problem $EP(F, C)$, and using a Mann's iteration to the fixed point problems of nonexpansive mappings. The convergence analysis has been considered when the bifunction F is monotone. This is because the proximal point method is not valid when the underlying operator F is pseudomonotone.

Recently, Anh [2] introduced a new hybrid extragradient iteration method for finding a element common to the set of fixed points of a nonexpansive mapping and the set of solutions of equilibrium problems for a pseudomonotone bifunctions. In this algorithm the equilibrium bifunction is not required to satisfy any monotonicity property, but it must satisfy a Lipschitz-type continuous bifunction i.e. there are two Lipschitz constants $c_1 > 0$ and $c_2 > 0$ such that

$$f(x, y) + f(y, z) \geq f(x, z) - c_1\|x - y\|^2 - c_2\|y - z\|^2, \ \forall x, y, z \in C. \tag{1.2}$$

They obtained strongly convergent theorems for the sequences generated by these processes in a real Hilbert space.

Anh and Muu [3] reiterated that the Lipschitz-type condition (1.2) is not in general satisfied, and if it is, that finding the constants c_1 and c_2 is not easy. They further observed that solving strongly convex programs is also difficult except in special cases when C has a simple structure. They introduced and studied a new algorithm, which is called a hybrid subgradient algorithm for finding a common point in the set of fixed points of nonexpansive mappings and the solution set of a class of pseudomonotone equilibrium problems in a real Hilbert space. The proposed algorithm is a combination of the well-known Mann's iterative scheme for fixed point and the projection method for equilibrium problems. Furthermore, the proposed algorithm uses only one projection and does not require any Lipschitz condition for the bifunctions. To be more precise, they proposed the following iterative method:

$$\begin{cases} x_0 \in C, \\ w_n \in \partial_{\epsilon_n} F(x_n, \cdot)x_n, \\ u_n = P_C(x_n - \gamma_n w_n), \ \gamma_n = \frac{\beta_n}{\max\{\sigma_n, \|w_n\|\}}, \\ x_{n+1} = \alpha_n x_n + (1 - \alpha_n)Tu_n, \ \text{for each } n = 1, 2, 3, ..., \end{cases} \tag{1.3}$$

where $\partial_\epsilon F(x, \cdot)(x)$ stands for ϵ-subdifferential of the convex function $F(x, \cdot)$ at x and $\{\epsilon_n\}$, $\{\gamma_n\}, \{\beta_n\}, \{\sigma_n\}$, and $\{\alpha_n\}$ were chosen appropriately. Under certain conditions, they prove that $\{x_n\}$ converges strongly to a common point in the set of a class of pseudomonotone equilibrium problems and the set of fixed points of nonexpansive mapping. Using the idea of Anh and Muu [3], Thailert et al. [21] proposed a new algorithm for finding a common point in the solution set of a class of pseudomonotone equilibrium problems and the set of common fixed points of a

family of strict pseudocontraction mappings in a real Hilbert space. Then Thailert et al. [22] introduced new general iterative methods for finding a common element in the solution set of pseudomonotone equilibrium problems and the set of fixed points of nonexpansive mappings which is a solution of a certain optimization problem related to a strongly positive linear operator. Under suitable control conditions, They proved the strong convergence theorems of such iterative schemes in a real Hilbert space.

In this paper, motivated by Anh and Muu [3], Kangtunyakarn [14], and other research going on in this direction, we proposed a hybrid subgradient method for the pseudomonotone equilibrium problem and the finite family of κ-strictly pseudononspreading mapping in a real Hilbert space. The weak and strong convergence of the proposed methods is investigated under certain assumptions. Our results improve and extend many recent results in the literature.

2. Preliminaries

Let H be a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$, respectively. It is well-known that for all $x, y, z \in H$ and α, β, $\gamma \in [0, 1]$, with $\alpha+\beta+\gamma = 1$ there holds

$$\|x - y\|^2 = \|x\|^2 - \|y\|^2 - 2\langle x - y, y\rangle, \tag{2.1}$$

and

$$\| \alpha x + \beta y + \gamma z \|^2 = \alpha \| x \|^2 + \beta \| y \|^2 + \gamma \|z\|^2 - \alpha\beta \| x - y \|^2 - \beta\gamma \|y - z\|^2. \tag{2.2}$$

Let C be a nonempty closed convex subset of H. Then, for any $x \in H$, there exists a unique nearest point of C, denoted by $P_C x$, such that $\| x - P_C x \| \leq \| x - y \|$ for all $y \in C$. Such a P_C is called the metric projection from H into C. We know that P_C is nonexpansive. It is also known that, $P_C x \in C$ and

$$\langle x - P_C x, P_C x - z\rangle \geq 0, \quad \text{for all } x \in H \text{ and } z \in C. \tag{2.3}$$

It is easy to see that (2.3) equivalent to

$$\|x - z\|^2 \geq \|x - P_C x\|^2 + \|z - P_C x\|^2, \quad \text{for all } x \in H \text{ and } z \in C. \tag{2.4}$$

Lemma 2.1.([19]) *Let H be a real Hilbert space, let C be a nonempty closed convex subset of H and let A be a mapping of C into H. Let $u \in C$. Then for $\lambda > 0$,*

$$u \in VI(C, A) \Leftrightarrow u = P_C(I - \lambda A)u,$$

where P_C is the metric projection of H onto C.

Recall that a bifunction $F : C \times C \to \mathbb{R}$ is said to be

(i) *η-strongly monotone* if there exists a number $\eta > 0$ such that

$$F(x, y) + F(y, x) \leq -\eta\|x - y\|^2, \text{ for all } x, y \in C,$$

(ii) *monotone* on C if

$$F(x, y) + F(y, x) \leq 0, \text{ for all } x, y \in C,$$

(iii) *pseudomonotone* on C with respect to $x \in C$ if

$$F(x, y) \geq 0 \text{ implies } F(y, x) \leq 0, \text{ for all } y \in C.$$

It is clear that (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii), for every $x \in C$. Moreover, F is said to be *pseudomonotone* on C with respect to $A \subseteq C$, if it is pseudomonotone on C with respect to every $x \in A$. When $A \equiv C$, F is called pseudomonotone on C.

The following example, taken from [18], shows that a bifunction may not be pseudomonotone on C, but yet is pseudomonotone on C with respect to the solution set of the equilibrium problem defined by F and C:

$$F(x, y) := 2y|x|(y - x) + xy|y - x|, \text{ for all } x, y \in \mathbb{R}, C := [-1, 1].$$

Clearly, $EP(F) = \{0\}$. Since $F(y, 0) = 0$ for every $y \in C$, this bifunction is pseudomonotone on C with respect to the solution $x^* = 0$, However, F is not pseudomonotone on C. In fact, both $F(-0.5, 0.5) = 0.25 > 0$ and $F(0.5, -0.5) = 0.25 > 0$.

For solving the equilibrium problem (1.1), let us assume that Δ is an open convex set containing C and the bifunction $F : \Delta \times \Delta \to \mathbb{R}$ satisfies the following assumptions:

(A1) $F(x, x) = 0$ for all $x \in C$ and $F(x, \cdot)$ is convex and lower semicontinuous on C;

(A2) for each $y \in C$, $F(\cdot, y)$ is weakly upper semicontinuous on the open set Δ;

(A3) F is pseudomonotone on C with respect to $EP(F, C)$ and satisfies the strict paramonotonicity property, i.e., $F(y, x) = 0$ for $x \in EP(F, C)$ and $y \in C$ implies $y \in EP(F, C)$;

(A4) if $\{x_n\} \subseteq C$ is bounded and $\epsilon_n \to 0$ as $n \to \infty$, then the sequence $\{w_n\}$ with $w_n \in \partial_n F(x_n, \cdot)x_n$ is bounded, where $\partial_\epsilon F(x, \cdot)x$ stands for the ϵ-subdifferential of the convex function $F(x, \cdot)$ at x.

The following idea of the ϵ-subdierential of convex functions can be found in the work of Bronsted and Rockafellar [9] but the theory of ϵ-subdierential calculus was given by Hiriart-Urruty [13].

Definition 2.2. Consider a proper convex function $\phi : C \to \overline{\mathbb{R}}$. For a given $\epsilon > 0$, the ϵ-subdierential of ϕ at $x_0 \in Dom\phi$ is given by

$$\partial_\epsilon \phi(x_0) = \{x \in C : \phi(y) - \phi(x_0) \geq \langle x, y - x_0\rangle - \epsilon, \ \forall y \in C\}.$$

Remark 2.3. It is known that if the function ϕ is proper lower semicontinuous convex, then for every $x \in Dom\phi$, the ϵ-subdierential $\partial_\epsilon\phi(x)$ is a nonempty closed convex set (see [13]).

Next, throughout this paper, weak and strong convergence of a sequence $\{x_n\}$ in H to x are denoted by $x_n \rightharpoonup x$ and $x_n \to x$, respectively. In order to prove our main results, we need the following lemmas.

Lemma 2.4.([17]) *Let C be a nonempty closed convex subset of a real Hilbert space H, and let $T : C \to C$ be a κ-strictly pseudonospreading mapping. If $F(T) \neq \emptyset$, then it is closed and convex.*

Remark 2.5. If $T : C \to C$ is a κ-strictly pseudononspreading mapping with $F(T) \neq \emptyset$, then from Lemma 2.8 in [14] and Lemma 2.1, we have $F(T) = VI(C, (I - T)) = F(P_C(I - \lambda(I - T)))$, for all $\lambda > 0$.

Lemma 2.6. *Let H be a real Hilbert space and C be a nonempty closed convex subset of H. For every $i = 1, 2, ..., N$, let $T_i : C \to C$ be a finite family of κ_i-strictly pseudononspreading mappings with $\bigcap_{i=1}^N F(T_i) \neq \emptyset$. Let $\{a_1, a_2, ..., a_n\} \subset (0, 1)$ with $\Sigma_{i=1}^N a_i = 1$, let $\bar{\kappa} = max\{\kappa_1, \kappa_2, ..., \kappa_N\}$ and let $\lambda \in (0, 1 - \bar{\kappa})$. Then*

(i) $\bigcap_{i=1}^N F(T_i) = F(\Sigma_{i=1}^N a_i P_C(I - \lambda(I - T_i)))$.

(ii) $\|\Sigma_{i=1}^N a_i P_C(I - \lambda(I - T_i))x - y\|^2 \leq \|x - y\|^2$, *for all $x \in C$ and $y \in \bigcap_{i=1}^N F(T_i)$, i.e. $\Sigma_{i=1}^N a_i P_C(I - \lambda(I - T_i))$ is quasi-nonexpansive.*

Proof. (i) It easy to see that $\bigcap_{i=1}^N F(T_i) \subseteq F(\Sigma_{i=1}^N a_i P_C(I - \lambda(I - T_i)))$. Let $x \in F(\Sigma_{i=1}^N a_i P_C(I - \lambda(I - T_i)))$ and let $x^* \in \bigcap_{i=1}^N F(T_i) \subseteq F(\Sigma_{i=1}^N a_i P_C(I - \lambda(I - T_i)))$. Note that for every $i = 1, 2, 3, ..., N$ we have

$$\begin{aligned} \|P_C(I - \lambda(I - T_i))x - x^*\|^2 &\leq \|x - x^* - \lambda(I - T_i)\|^2 \\ &= \|x - x^*\|^2 - 2\lambda\langle x - x^*, (I - T_i)x\rangle \\ &\quad + \lambda^2\|(I - T_i)x\|^2. \end{aligned} \tag{2.5}$$

Put $A_i = I - T_i$, for all $i = 1, 2, ..., N$, we have $T_i = I - A_i$ and

$$\begin{aligned} \|T_i x - T_i x^*\|^2 &= \|(I - A_i)x - (I - A_i)x^*\|^2 \\ &= \|(x - x^*) - A_i x\|^2 \\ &= \|x - x^*\|^2 - 2\langle x - x^*, A_i x\rangle + \|A_i x\|^2 \\ &\leq \|x - x^*\|^2 + \kappa_i\|(I - T_i)x - (I - T_i)x^*\|^2 + 2\langle x - T_i x, x^* - T_i x^*\rangle \\ &= \|x - x^*\|^2 + \kappa_i\|(I - T_i)x\|^2, \end{aligned} \tag{2.6}$$

which implies that

$$(1 - \kappa_i)\|(I - T_i)x\|^2 \leq 2\langle x - x^*, A_i x\rangle, \quad \text{for all } i = 1, 2, 3, ..., N$$

From (2.5) and (2.6), we have

$$\begin{aligned}\|P_C(I-\lambda(I-T_i))x-x^*\|^2 &\leq \|x-x^*\|^2-2\lambda\langle x-x^*,(I-T_i)x\rangle\\ &\quad+\lambda^2\|(I-T_i)x\|^2\\ &\leq \|x-x^*\|^2-\lambda(1-\kappa_i)\|(I-T_i)x\|^2\\ &\quad+\lambda^2\|(I-T_i)x\|^2\\ &= \|x-x^*\|^2-\lambda[(1-\kappa_i)-\lambda]\|(I-T_i)x\|^2\\ &\leq \|x-x^*\|^2, \qquad (2.7)\end{aligned}$$

for all $i=1,2,3,\ldots,N$.
From the definition of x and (2.7), we have

$$\begin{aligned}\|x-x^*\|^2 &= \|\Sigma_{i=1}^N a_iP_C(I-\lambda(I-T_i))x-x^*\|^2\\ &= a_1\|P_C(I-\lambda(I-T_1))x-x^*\|^2+a_2\|P_C(I-\lambda(I-T_2))x-x^*\|^2+\cdots\\ &\quad+a_N\|P_C(I-\lambda(I-T_N))x-x^*\|^2-a_1a_2\|P_C(I-\lambda(I-T_1))x\\ &\quad-P_C(I-\lambda(I-T_2))x\|^2-a_2a_3\|P_C(I-\lambda(I-T_2))x-\\ &\quad P_C(I-\lambda(I-T_3))x\|^2-\cdots-a_{N-1}a_N\|P_C(I-\lambda(I-T_{N-1}))x-\\ &\quad P_C(I-\lambda(I-T_N))x\|^2\\ &\leq \|x-x^*\|^2-a_1a_2\|P_C(I-\lambda(I-T_1))x-P_C(I-\lambda(I-T_2))x\|^2\\ &\quad-a_2a_3\|P_C(I-\lambda(I-T_2))x-P_C(I-\lambda(I-T_3))x\|^2-\cdots\\ &\quad-a_{N-1}a_N\|P_C(I-\lambda(I-T_{N-1}))x-P_C(I-\lambda(I-T_N))x\|^2.\end{aligned}$$

This implies that

$$P_C(I-\lambda(I-T_1))x=P_C(I-\lambda(I-T_2))x=\cdots=P_C(I-\lambda(I-T_N))x$$

Since $x\in F(\Sigma_{i=1}^N a_iP_C(I-\lambda(I-T_i)))$, we get that $x=P_C(I-\lambda(I-T_i))x$, for all $i=1,2,3,...,N$ From Remark 2.5, we have $x\in F(T_i)$, for all $i=1,2,3,...,N$. That is $x\in\bigcap_{i=1}^N F(T_i)$. Hence $F(\Sigma_{i=1}^N a_iP_C(I-\lambda(I-T_i)))\subseteq\bigcap_{i=1}^N F(T_i)$.
(ii) Let $x\in C$ and $y\in\bigcap_{i=1}^N F(T_i)=F(\Sigma_{i=1}^N a_iP_C(I-\lambda(I-T_i)))$
As the same argument as in (i), we can show that

$$\|P_C(I-\lambda(I-T_i))x-y\|^2\leq\|x-y\|^2, \qquad (2.8)$$

for all $i=1,2,3,...,N$. Thus

$$\begin{aligned}\|\Sigma_{i=1}^N a_iP_C(I-\lambda(I-T_i))x-y\|^2 &\leq a_1\|P_C(I-\lambda(I-T_1))x-y\|^2\\ &\quad+a_2\|P_C(I-\lambda(I-T_2))x-y\|^2+\cdots\\ &\quad+a_N\|P_C(I-\lambda(I-T_N))x-y\|^2\\ &\leq \Sigma_{i=1}^N a_i\|x-y\|^2=\|x-y\|^2. \qquad (2.9)\end{aligned}$$

□

Lemma 2.7.([23]) *Let $\{a_n\}$ and $\{b_n\}$ be two sequences of nonnegative real numbers such that*

$$a_{n+1} \leq a_n + b_n, \ \ n \geq 1,$$

where $\sum_{n=0}^{\infty} b_n < \infty$. Then the sequence $\{a_n\}$ is convergent.

3. Weak Convergence Theorem

In this section, we prove weak convergence theorem for finding a common element in the solution set of a class of pseudomonotone equilibrium problems and the set of fixed points of a finite family of κ-strictly presudononspreading mappings in a real Hilbert space.

Theorem 3.1. *Let C be a closed convex subset of a real Hilbert space H and $F : C \times C \to \mathbb{R}$ be a bifunction satisfying (A1)–(A4). Let $\{\kappa_1, \kappa_2, ..., \kappa_N\} \subset [0, 1)$ and $\{T_i\}_{i=1}^N$ be a finite family of κ_i-strictly pseudononspreading mappings of C into itself such that $\Omega := \bigcap_{i=1}^N F(T_i) \cap EP(F, C) \neq \emptyset$. Let $x_0 \in C$ and $\{x_n\}$ be a sequence generated by*

$$\text{(3.1)} \quad \begin{cases} x_0 \in C, \\ w_n \in \partial_{\epsilon_n} F(x_n, \cdot)x_n, \\ u_n = P_C(x_n - \rho_n w_n), \ \rho_n = \frac{\delta_n}{max\{\sigma_n, \|w_n\|\}}, \\ x_{n+1} = \alpha_n x_n + \beta_n \Sigma_{i=1}^N a_i P_C(I - \lambda_n^i(I - T_i))x_n + \gamma_n u_n, \ \forall n \in \mathbb{N}, \end{cases}$$

where $a, b, c, d, \lambda \in \mathbb{R}$, $a_i \in (0, 1)$, for all $i = 1, 2, ..., N$ with $\Sigma_{i=1}^N a_i = 1$, $\{\alpha_n\}$, $\{\beta_n\}$, $\{\gamma_n\} \subset [0, 1]$ with $\alpha_n + \beta_n + \gamma_n = 1$ and $\{\delta_n\}$, $\{\epsilon_n\}$,$\{\lambda_n^i\} \subset (0, \infty)$ satisfying the following conditions:

(i) $0 < \lambda \leq \lambda_n^i \leq \min\{1 - \kappa_1, 1 - \kappa_2, ..., 1 - \kappa_N\}$ *and* $\Sigma_{n=1}^{\infty} \lambda_n^i < \infty$ *for all* $i = 1, 2, ..., N$;

(ii) $0 < a < \alpha_n, \beta_n, \gamma_n < b < 1$;

(iii) $\sum_{n=0}^{\infty} \delta_n = \infty$, $\sum_{n=0}^{\infty} \delta_n^2 < \infty$, *and* $\sum_{n=0}^{\infty} \delta_n \epsilon_n < \infty$.

Then the sequence $\{x_n\}$ converges weakly to $\bar{x} \in \Omega$.

Proof. First, we will show that $\{x_n\}$ is bounded. Let $p \in \Omega$. Then we have

$$\begin{aligned} \|u_n - p\|^2 &= \|x_n - p\|^2 - \|u_n - x_n\|^2 + 2\langle x_n - u_n, p - u_n\rangle \\ &\leq \|x_n - p\|^2 + 2\langle x_n - u_n, p - u_n\rangle. \end{aligned} \quad \text{(3.2)}$$

Since $u_n = P_C(x_n - \rho_n w_n)$ and $p \in C$, we get that

$$\langle x_n - u_n, p - u_n\rangle \leq \rho_n \langle w_n, p - u_n\rangle. \quad \text{(3.3)}$$

Substuting (3.3) into (3.2), we have

$$\begin{aligned}
\|u_n - p\|^2 &\leq \|x_n - p\|^2 + 2\rho_n\langle w_n, p - u_n\rangle \\
&= \|x_n - p\|^2 + 2\rho_n\langle w_n, p - x_n\rangle + 2\rho_n\langle w_n, x_n - u_n\rangle \\
&\leq \|x_n - p\|^2 + 2\rho_n\langle w_n, p - x_n\rangle + 2\rho_n\|w_n\|\|x_n - u_n\| \\
&\leq \|x_n - p\|^2 + 2\rho_n\langle w_n, p - x_n\rangle + 2\delta_n\|x_n - u_n\|. \qquad (3.4)
\end{aligned}$$

By using $u_n = P_C(x_n - \rho_n w_n)$ and $x_n \in C$ again, we get

$$\begin{aligned}
\|x_n - u_n\|^2 &= \langle x_n - u_n, x_n - u_n\rangle \\
&\leq \rho_n\langle w_n, x_n - u_n\rangle \\
&\leq \rho_n\|w_n\|\|x_n - u_n\| \\
&\leq \delta_n\|x_n - u_n\|, \qquad (3.5)
\end{aligned}$$

which implies that

$$\|x_n - u_n\| \leq \delta_n. \qquad (3.6)$$

By condition (iii), we have

$$\lim_{n\to\infty}\|x_n - u_n\| = 0. \qquad (3.7)$$

Combining (3.4) and (3.6), we obtain

$$\|u_n - p\|^2 \leq \|x_n - p\|^2 + 2\rho_n\langle w_n, p - x_n\rangle + 2\delta_n^2. \qquad (3.8)$$

Since $w_n \in \partial_{\epsilon_n} F(x_n, \cdot)x_n$, $p \in C$ and $F(x, x) = 0$ for each $x \in C$, we obtain that

$$\begin{aligned}
\langle w_n, p - x_n\rangle &\leq F(x_n, p) - F(x_n, x_n) + \epsilon_n \\
&= F(x_n, p) + \epsilon_n. \qquad (3.9)
\end{aligned}$$

Thus, it follows from (3.8) and (3.9) that

$$\|u_n - p\|^2 \leq \|x_n - p\|^2 + 2\rho_n F(x_n, p) + 2\rho_n\epsilon_n + 2\delta_n^2. \qquad (3.10)$$

Form Lemma 2.6 (ii), we have

$$\|\Sigma_{i=1}^N a_i P_C(I - \lambda_n^i(I - T_i))x_n - p\|^2 \leq \|x_n - p\|^2. \qquad (3.11)$$

From (3.1), (3.10) and (3.11), we have

$$\begin{aligned}
\|x_{n+1} - p\|^2 &= \|\alpha_n x_n + \beta_n\Sigma_{i=1}^N a_i P_C(I - \lambda_n^i(I - T_i))x_n + \gamma_n u_n - p\|^2 \\
&\leq \alpha_n\|x_n - p\|^2 + \beta_n\|\Sigma_{i=1}^N a_i P_C(I - \lambda_n^i(I - T_i))x_n - p\|^2 \\
&\quad +\gamma_n\|u_n - p\|^2 - \alpha_n\beta_n\|x_n - \Sigma_{i=1}^N a_i P_C(I - \lambda_n^i(I - T_i))x_n\|^2 \\
&\leq \alpha_n\|x_n - p\|^2 + \beta_n\|x_n - p\|^2 + \gamma_n\Big(\|x_n - p\|^2 + 2\rho_n F(x_n, p) \\
&\quad +2\rho_n\epsilon_n + 2\delta_n^2\Big) - \alpha_n\beta_n\|x_n - \Sigma_{i=1}^N a_i P_C(I - \lambda_n^i(I - T_i))x_n\|^2 \\
&= \|x_n - p\|^2 + 2\gamma_n\rho_n F(x_n, p) + 2\gamma_n\rho_n\epsilon_n + 2\gamma_n\delta_n^2 \\
&\quad -\alpha_n\beta_n\|x_n - \Sigma_{i=1}^N a_i P_C(I - \lambda_n^i(I - T_i))x_n\|^2. \qquad (3.12)
\end{aligned}$$

Since $p \in EP(F,C)$ and F is pseudomonotone on F with respect to p, we get that $F(x_n,p) \leq 0$ for all $n \in \mathbb{N}$. Then from (3.12) it follows that

$$\begin{aligned} \|x_{n+1}-p\|^2 &\leq \|x_n-p\|^2 + 2\gamma_n\rho_n\epsilon_n + 2\gamma_n\delta_n^2 \\ &\quad -\alpha_n\beta_n\|x_n - \Sigma_{i=1}^N a_i P_C(I-\lambda_n^i(I-T_i))x_n\|^2 \\ &\leq \|x_n-p\|^2 + 2\gamma_n\rho_n\epsilon_n + 2\gamma_n\delta_n^2. \end{aligned} \tag{3.13}$$

Let $\eta_n = 2\gamma_n\rho_n\epsilon_n + 2\gamma_n\delta_n^2$ for all $n \geq 0$. From condition (ii) and (iii), we get that

$$\Sigma_{n=0}^{\infty}\eta_n = \Sigma_{n=0}^{\infty}(2\gamma_n\rho_n\epsilon_n + 2\gamma_n\delta_n^2) \leq 2b\Sigma_{n=0}^{\infty}\rho_n\epsilon_n + 2b\Sigma_{n=0}^{\infty}\delta_n^2 < +\infty$$

Now applying Lemma 2.7 to (3.13), we obtain that the $\lim_{n\to\infty}\|x_n-p\|$ exists, i.e. $\lim_{n\to\infty}\|x_n-p\| = \bar{a}$ for some $\bar{a} \in C$. Thus $\{x_n\}$ is bounded. Also, it easy to verify that $\{u_n\}$ and $\{\Sigma_{i=1}^N a_i P_C(I-\lambda_n^i(I-T_i))x_n\}$ are also bounded.

Next, we will show that $\limsup_{n\to\infty} F(x_n,p) = 0$ for any $p \in \Omega$. Since F is pseudomonotone on C and $F(p,x_n) \geq 0$, we have $-F(x_n,p) \geq 0$. From (3.12) and condition (ii), we have

$$\begin{aligned} 2\gamma_n\rho_n[-F(x_n,p)] &\leq \|x_n-p\|^2 - \|x_{n+1}-p\|^2 \\ &\quad +2\gamma_n\rho_n\epsilon_n + 2\gamma_n\delta_n^2 \\ &\leq \|x_n-p\|^2 - \|x_{n+1}-p\|^2 + 2b\rho_n\epsilon_n + 2b\delta_n^2. \end{aligned} \tag{3.14}$$

Summing up (3.14) for every n, we obtain

$$\begin{aligned} 0 &\leq 2\sum_{n=0}^{\infty}\gamma_n\rho_n[-F(x_n,p)] \\ &\leq \|x_0-p\|^2 + 2b\sum_{n=0}^{\infty}\rho_n\epsilon_n + 2b\sum_{n=0}^{\infty}\delta_n^2 < +\infty. \end{aligned} \tag{3.15}$$

By the assumption (A_4), we can find a real number w such that $\|w_n\| \leq w$ for every n. Setting $\Gamma := \max\{\sigma, w\}$, where σ is a real number such that $0 < \sigma_n < \sigma$ for every n, it follows from (ii) that

$$0 \leq \frac{2a}{\Gamma}\sum_{n=0}^{\infty}\delta_n[-F(x_n,p)] \tag{3.16}$$

$$\leq 2\sum_{n=0}^{\infty}\gamma_n\rho_n[-F(x_n,p)] < +\infty, \tag{3.17}$$

which implies that

$$0 \leq \sum_{n=0}^{\infty}\delta_n[-F(x_n,p)] < +\infty. \tag{3.18}$$

Combining with $-F(x_n, p) \geq 0$ and $\sum_{n=0}^{\infty} \delta_n = \infty$, we can deduced that $\limsup_{n\to\infty} F(x_n, p) = 0$ as desired.

Next, we will show that $\omega_\omega(x_n) \subset \Omega$, where $\omega_\omega(x_n) = \{x \in H : x_{n_i} \rightharpoonup x$ for some subsequence $\{x_{n_i}\}$ of $\{x_n\}\}$. In deed since $\{x_n\}$ is bounded and H is reflexive, $\omega_\omega(x_n)$ is nonempty. Let $\bar{x} \in \omega_\omega(x_n)$. Then there exists subsequence $\{x_{n_i}\}$ of $\{x_n\}$. converging weakly to $\bar{x}$, that is $x_{n_i} \rightharpoonup \bar{x}$ as $i \to \infty$. By the convexity, C is weakly closed and hence $\bar{x} \in C$. Since $F(\cdot, p)$ is weakly upper semicontinuous for $p \in \Omega$, we obtain

$$\begin{aligned} F(\bar{x}, p) &\geq \limsup_{i\to\infty} F(x_n, p) \\ &= \lim_{i\to\infty} F(x_{n_i}, p) \\ &= \limsup_{n\to\infty} F(x_n, p) \\ &= 0. \end{aligned} \tag{3.19}$$

Since F is pseudomontone with respect to p and $F(p, \bar{x}) \geq 0$, we obtain $F(\bar{x}, p) \leq 0$. Thus $F(\bar{x}, p) = 0$. Furthermore, by assumption (A_3), we get that $\bar{x} \in EP(F, C)$. On the other hand, from (3.13) and conditions (ii)–(iii), we have

$$\begin{aligned} &\alpha_n\beta_n\|x_n - \Sigma_{i=1}^N a_i P_C(I-\lambda_n^i(I-T_i))x_n\|^2 \\ &\qquad \leq \|x_n - p\|^2 - \|x_{n+1} - p\|^2 + 2\gamma_n\rho_n\epsilon_n + 2\gamma_n\delta_n^2 \\ &\qquad \leq \|x_n - p\|^2 - \|x_{n+1} - p\|^2 + 2b\rho_n\epsilon_n + 2b\delta_n^2 \end{aligned} \tag{3.20}$$

taking the limit as $n \to \infty$ yields

$$\lim_{n\to\infty} \|x_n - \Sigma_{i=1}^N a_i P_C(I - \lambda_n^i(I - T_i))x_n\| = 0. \tag{3.21}$$

Now, we will show that $\bar{x} \in \bigcap_{i=1}^N F(T_i)$. Assume that $\bar{x} \notin \bigcap_{i=1}^N F(T_i)$. By Lemma 2.6, we have $\bar{x} \notin F(\Sigma_{i=1}^N a_i P_C(I - \lambda_n(I - T_i)))$. From the Opial's condition, (3.21) and condition (i), we can write

$$\begin{aligned} \liminf_{i\to\infty} \|x_{n_i} - \bar{x}\| &< \liminf_{i\to\infty} \|x_{n_i} - \Sigma_{i=1}^N a_i P_C(I - \lambda_n^i(I - T_i))\bar{x}\| \\ &\leq \liminf_{i\to\infty} \Big(\|x_{n_i} - \Sigma_{i=1}^N a_i P_C(I - \lambda_n^i(I - T_i))x_{n_i}\| \\ &\qquad + \|\Sigma_{i=1}^N a_i P_C(I - \lambda_n^i(I - T_i))x_{n_i} - \Sigma_{i=1}^N a_i P_C(I - \lambda_n^i(I - T_i))\bar{x}\| \Big) \\ &\leq \liminf_{i\to\infty} \Big(\|x_{n_i} - \bar{x}\| + \Sigma_{i=1}^N a_i \lambda_n^i \|(I - T_i)x_{n_i} - (I - T_i)\bar{x}\| \Big) \\ &\leq \liminf_{i\to\infty} \|x_{n_i} - \bar{x}\|. \end{aligned}$$

This is a contradiction. Then $\bar{x} \in \bigcap_{i=1}^N F(T_i)$. Thus $\bar{x} \in EP(F, C) \cap F(T) = \Omega$ and so $\omega_\omega(x_n) \subset \Omega$.

Finally, we prove that $\{x_n\}$ converge weakly to an element of Ω. It's sufficient to show that $\omega_\omega(x_n)$ is a single point set. Taking $z_1, z_2 \in \omega_\omega(x_n)$ arbitrarily, and let $\{x_{n_k}\}$ and $\{x_{n_m}\}$ be subsequence of $\{x_n\}$ such that $x_{n_k} \rightharpoonup z_1$ and $x_{n_m} \rightharpoonup z_2$ respectively. Since $\lim_{n\to\infty} \|x_n - p\|$ exists for all $p \in \Omega$ and $z_1, z_2 \in \Omega$, we get that $\lim_{n\to\infty} \|x_n - z_1\|$ and $\lim_{n\to\infty} \|x_n - z_2\|$ exist. Now, assume that $z_1 \neq z_2$, then by the Opial's condition,

$$\begin{aligned}
\lim_{n\to\infty} \|x_n - z_1\| &= \lim_{k\to\infty} \|x_{n_k} - z_1\| \\
&< \lim_{k\to\infty} \|x_{n_k} - z_2\| \\
&= \lim_{n\to\infty} \|x_n - z_2\| \\
&= \lim_{m\to\infty} \|x_{n_m} - z_2\| \\
&< \lim_{m\to\infty} \|x_{n_m} - z_1\| \\
&= \lim_{n\to\infty} \|x_n - z_1\|, \qquad (3.22)
\end{aligned}$$

which is a contradiction. Thus $z_1 = z_2$. This show that $\omega_\omega(x_n)$ is single point set. i.e. $x_n \rightharpoonup \bar{x}$. This completes the proof. □

If we set $\kappa_i = 0$ for all $i = 1, 2, ..., N$ then we get the following Corollary.

Corollary 3.2. *Let C be a closed convex subset of a real Hilbert space H and $F : C \times C \to \mathbb{R}$ be a bifunction satisfying (A1)–(A4). Let $\{T_i\}_{i=1}^N$ be a finite family of nonspreading mappings of C into itself such that $\Omega := \bigcap_{i=1}^N F(T_i) \cap EP(F, C) \neq \emptyset$. Let $x_0 \in C$ and $\{x_n\}$ be a sequence generated by*

$$(3.23) \quad \begin{cases} x_0 \in C, \\ w_n \in \partial_{\epsilon_n} F(x_n, \cdot)x_n, \\ u_n = P_C(x_n - \rho_n w_n), \ \rho_n = \frac{\delta_n}{max\{\sigma_n, \|w_n\|\}}, \\ x_{n+1} = \alpha_n x_n + \beta_n \Sigma_{i=1}^N a_i P_C(I - \lambda_n^i(I - T_i))x_n + \gamma_n u_n, \ \forall n \in \mathbb{N}, \end{cases}$$

where $a, b, c, d, \lambda \in \mathbb{R}$, $a_i \in (0, 1)$, for all $i = 1, 2, ..., N$ with $\Sigma_{i=1}^N a_i = 1$, $\{\alpha_n\}$, $\{\beta_n\}$, $\{\gamma_n\} \subset [0, 1]$ with $\alpha_n + \beta_n + \gamma_n = 1$ and $\{\delta_n\}$, $\{\epsilon_n\}$,$\{\lambda_n^i\} \subset (0, \infty)$ satisfying the following conditions:

(i) $0 < \lambda \leq \lambda_n^i < 1$ *and* $\Sigma_{n=1}^\infty \lambda_n^i < \infty$ *for all* $i = 1, 2, ..., N$;

(ii) $0 < a < \alpha_n, \beta_n, \gamma_n < b < 1$;

(ii) $\sum_{n=0}^\infty \delta_n = \infty$, $\sum_{n=0}^\infty \delta_n^2 < \infty$, *and* $\sum_{n=0}^\infty \delta_n \epsilon_n < \infty$.

Then the sequence $\{x_n\}$ converges weakly to $\bar{x} \in \Omega$.

4. Strong Convergence Theorem

In this section, to obtain strong convergence result, we add the control condition $\lim_{n\to\infty} \alpha_n = \frac{1}{2}$, and then we get the strong convergence theorem for finding a common element in the solution set of a class of pseudomonotone equilibrium problems and the set of fixed points of a finite family of κ-strictly presudononspreading mappings in a real Hilbert space.

Theorem 4.1. *Let C be a closed convex subset of a real Hilbert space H and $F : C \times C \to \mathbb{R}$ be a bifunction satisfying (A1)–(A4). Let $\{\kappa_1, \kappa_2, ..., \kappa_N\} \subset [0, 1)$ and $\{T_i\}_{i=1}^N$ be a finite family of κ_i-strictly pseudononspreading mappings of C into itself such that $\Omega := \bigcap_{i=1}^N F(T_i) \cap EP(F, C) \neq \emptyset$. Let $x_0 \in C$ and $\{x_n\}$ be a sequence generated by*

$$(4.1) \quad \begin{cases} x_0 \in C, \\ w_n \in \partial_{\epsilon_n} F(x_n, \cdot)x_n, \\ u_n = P_C(x_n - \rho_n w_n),\ \rho_n = \frac{\delta_n}{max\{\sigma_n, \|w_n\|\}}, \\ x_{n+1} = \alpha_n x_n + \beta_n \Sigma_{i=1}^N a_i P_C(I - \lambda_n^i(I - T_i))x_n + \gamma_n u_n,\ \forall n \in \mathbb{N}, \end{cases}$$

where $a, b, c, d, \lambda \in \mathbb{R}$, $a_i \in (0, 1)$, for all $i = 1, 2, ..., N$ with $\Sigma_{i=1}^N a_i = 1$, $\{\alpha_n\}$, $\{\beta_n\}$, $\{\gamma_n\} \subset [0, 1]$ with $\alpha_n + \beta_n + \gamma_n = 1$ and $\{\delta_n\}$, $\{\epsilon_n\}$,$\{\lambda_n^i\} \subset (0, \infty)$ satisfying the following conditions:

(i) $0 < \lambda \leq \lambda_n^i \leq \min\{1 - \kappa_1, 1 - \kappa_2, ..., 1 - \kappa_N\}$ *and* $\Sigma_{n=1}^\infty \lambda_n^i < \infty$ *for all* $i = 1, 2, ..., N$;

(ii) $0 < a < \alpha_n, \beta_n, \gamma_n < b < 1$ *and* $\lim_{n\to\infty} \alpha_n = \frac{1}{2}$;

(iii) $\sum_{n=0}^\infty \delta_n = \infty$, $\sum_{n=0}^\infty \delta_n^2 < \infty$, *and* $\sum_{n=0}^\infty \delta_n \epsilon_n < \infty$.

Then the sequence $\{x_n\}$ converges strongly to $\bar{x} \in \Omega$.

Proof. By a similar argument to the proof of Theorem 3.1 and (2.4), we have

$$\|\Sigma_{i=1}^N a_i P_C(I - \lambda_n^i(I - T_i))x_n - P_\Omega(x_n)\|^2 \leq \|\Sigma_{i=1}^N a_i P_C(I - \lambda_n^i(I - T_i))x_n - x_n\|^2 - \|x_n - P_\Omega(x_n)\|^2$$

and

$$(4.2) \qquad \|u_n - P_\Omega(x_n)\|^2 \leq \|u_n - x_n\|^2 - \|x_n - P_\Omega(x_n)\|^2.$$

It follows from (4.2) and condition (ii) that

$$\begin{aligned}
&\|x_{n+1} - P_\Omega(x_{n+1})\|^2 \\
&\quad\leq \|\alpha_n x_n + \beta_n \Sigma_{i=1}^N a_i P_C(I - \lambda_n^i(I - T_i))x_n + \gamma_n u_n - P_\Omega(x_n)\|^2 \\
&\quad\leq \alpha_n\|x_n - P_\Omega(x_n)\|^2 + \beta_n\|\Sigma_{i=1}^N a_i P_C(I - \lambda_n^i(I - T_i))x_n - P_\Omega(x_n))\|^2 \\
&\qquad +\gamma_n\|u_n - P_\Omega(x_n)\|^2 \\
&\quad\leq \alpha_n\|x_n - P_\Omega(x_n)\|^2 + \beta_n\Big(\|\Sigma_{i=1}^N a_i P_C(I - \lambda_n^i(I - T_i))x_n - x_n\|^2 \\
&\qquad -\|x_n - P_\Omega(x_n)\|^2\Big) + \gamma_n\Big(\|u_n - x_n\|^2 - \|x_n - P_\Omega(x_n)\|^2\Big) \\
&\quad= (\alpha_n - (\beta_n + \gamma_n))\|x_n - P_\Omega(x_n)\|^2 + \beta_n\|\Sigma_{i=1}^N a_i P_C(I - \lambda_n^i(I - T_i))x_n - x_n\|^2 \\
&\qquad +\gamma_n\|u_n - x_n\|^2. \\
&\quad\leq (2\alpha_n - 1)\|x_n - P_\Omega(x_n)\|^2 + b\|\Sigma_{i=1}^N a_i P_C(I - \lambda_n^i(I - T_i))x_n - x_n\|^2 \\
&\qquad +b\|u_n - x_n\|^2.
\end{aligned}$$

Combining (3.7), (3.21), conditions (ii)–(iii), and the boundedness of the sequence $\{x_n - P_\Omega(x_n)\}$, we obtain

$$\lim_{n\to\infty}\|x_{n+1} - P_\Omega(x_{n+1})\| = 0 \tag{4.3}$$

Since Ω is convex, for all $m > n$, we have $\frac{1}{2}(P_\Omega(x_m) + P_\Omega(x_n)) \in \Omega$, and therefore

$$\begin{aligned}
\|P_\Omega(x_m) - P_\Omega(x_n)\|^2 &= 2\|x_m - P_\Omega(x_m)\|^2 + 2\|x_m - P_\Omega(x_n)\|^2 \\
&\quad -4\|x_m - \frac{1}{2}(P_\Omega(x_m) + P_\Omega(x_n))\|^2 \\
&\leq 2\|x_m - P_\Omega(x_m)\|^2 + 2\|x_m - P_\Omega(x_n)\|^2 \\
&\quad -4\|x_m - P_\Omega(x_m)\|^2 \\
&= 2\|x_m - P_\Omega(x_n)\|^2 - 2\|x_m - P_\Omega(x_m)\|^2. \qquad (4.4)
\end{aligned}$$

Using (3.13) with $p = P_\Omega(x_n)$, we have

$$\begin{aligned}
\|x_m - P_\Omega(x_n)\|^2 &\leq \|x_{m-1} - P_\Omega(x_n)\|^2 + \eta_{m-1} \\
&\leq \|x_{m-2} - P_\Omega(x_n)\|^2 + \eta_{m-1} + \eta_{m-2} \\
&\leq \ldots \\
&\leq \|x_n - P_\Omega(x_n)\|^2 + \sum_{j=n}^{m-1}\eta_j, \qquad (4.5)
\end{aligned}$$

where $\eta_j = 2\gamma_j\rho_j\epsilon_j + 2\gamma_j\delta_j^2$. It follows from (4.4) and (4.5) that

$$\|P_\Omega(x_m) - P_\Omega(x_n)\|^2 \leq 2\|x_n - P_\Omega(x_n)\|^2 + 2\sum_{j=n}^{m-1}\eta_j - 2\|x_m - P_\Omega(x_m)\|^2. \tag{4.6}$$

Together with (4.3) and $\sum_{j=0}^{\infty}\eta_j < +\infty$, this implies that $\{P_\Omega(x_n)\}$ is a Cauchy sequence, Hence $\{P_\Omega(x_n)\}$ strongly converges to some point $x^* \in \Omega$. Moreover, we obtain

$$x^* = \lim_{i\to\infty} P_\Omega(x_{n_i}) = P_\Omega(\bar{x}) = \bar{x}, \tag{4.7}$$

which implies that $P_\Omega(x_i) \to x^* = \bar{x} \in \Omega$. Then from (4.3) and (4.7), we can conclude that $x_n \to \bar{x}$. This completes the proof. □

If we set $\kappa_i = 0$ for all $i = 1, 2, ..., N$ then we get the following Corollary.

Corollary 4.2. *Let C be a closed convex subset of a real Hilbert space H and $F : C \times C \to \mathbb{R}$ be a bifunction satisfying (A1)–(A4). Let $\{T_i\}_{i=1}^N$ be a finite family of nonspreading mappings of C into itself such that $\Omega := \bigcap_{i=1}^N F(T_i) \cap EP(F, C) \neq \emptyset$. Let $x_0 \in C$ and $\{x_n\}$ be a sequence generated by*

$$\begin{cases} x_0 \in C, \\ w_n \in \partial_{\epsilon_n} F(x_n, \cdot)x_n, \\ u_n = P_C(x_n - \rho_n w_n),\ \rho_n = \frac{\delta_n}{max\{\sigma_n, \|w_n\|\}}, \\ x_{n+1} = \alpha_n x_n + \beta_n \Sigma_{i=1}^N a_i P_C(I - \lambda_n^i(I - T_i))x_n + \gamma_n u_n,\ \forall n \in \mathbb{N}, \end{cases} \tag{4.8}$$

where $a, b, c, d, \lambda \in \mathbb{R}$, $a_i \in (0, 1)$, for all $i = 1, 2, ..., N$ with $\Sigma_{i=1}^N a_i = 1$, $\{\alpha_n\}$, $\{\beta_n\}$, $\{\gamma_n\} \subset [0, 1]$ with $\alpha_n + \beta_n + \gamma_n = 1$ and $\{\delta_n\}$, $\{\epsilon_n\}$,$\{\lambda_n^i\} \subset (0, \infty)$ satisfying the following conditions:

(i) $0 < \lambda \leq \lambda_n^i < 1$ *and* $\Sigma_{n=1}^{\infty}\lambda_n^i < \infty$ *for all* $i = 1, 2, ..., N$;

(ii) $0 < a < \alpha_n, \beta_n, \gamma_n < b < 1$ *and* $\lim_{n\to\infty} \alpha_n = \frac{1}{2}$;

(iii) $\sum_{n=0}^{\infty}\delta_n = \infty$, $\sum_{n=0}^{\infty}\delta_n^2 < \infty$, *and* $\sum_{n=0}^{\infty}\delta_n\epsilon_n < \infty$.

Then the sequence $\{x_n\}$ converges weakly to $\bar{x} \in \Omega$.

Acknowledgements. The authors would like to thank the faculty of science and technology, Rajamangala University of Technology Thanyaburi (RMUTT), Thailand for the financial support. Moreover, the authors would like to thank the referees for their valuable suggestions and comments which helped to improve the quality and readability of the paper.

References

[1] P. N. Anh, *Strong convergence theorems for nonexpansive mappings and Ky Fan inequalities*, J. Optim. Theory Appl., **154**(2012), 303–320.

[2] P. N. Anh, *A hybrid extragradient method extended to fixed point problems and equilibrium problems*, Optimization, **62(2)**(2013), 271–283.

[3] P. N. Anh and L. D. Muu, *A hybrid subgradient algorithm for nonexpansive mappings and equilibrium problems*, Optim. Lett., **8**(2014), 727–738.

[4] P. N. Anh and D. X. Son, *A new method for a finite family of pseudocontractions and equilibrium problems*, J. Appl. Math. Inform., **29**(2011), 1179–1191.

[5] K. Aoyama, Y. Kimura, W. Takahashi and M. Toyoda, *Approximation of common fixed points of a countable family of nonexpansive mappings in a Banach space*, Nonlinear Anal., **67**(2007), 2350–2360.

[6] A. Auslender, M. Teboulle and S. Ben-Tiba, *A logarithmic quadratic proximal method for variational inequalities*,Comput. Optim. Appl., **12**(1999), 31–40.

[7] J. Baillon, *Un theoreme de type ergodique pour les contractions nonlineaires dans un espace de Hilbert*, C. R. Acad. Sci. Paris Ser. A-B, **280**(1975), A1511–A1514.

[8] E. Blum and W. Oettli ,*From optimization and variational inequalities to equilibrium problems*, Math. Student, **63**(1994), 123–145.

[9] A. Brøndsted and R. T. Rockafellar, *On the subdifferentiability of convex functions*, Proc. Am. Math. Soc., **16**(1965), 605–611.

[10] F. E. Browder and W. V. Petryshyn, *Construction of fixed points of nonlinear mappings in Hilbert space*, J. Math. Anal. Appl., **20**(1967), 197–228.

[11] P. L. Combettes and S. A. Hirstoaga, *Equilibrium programming in Hilbert spaces*, J. Nonlinear Convex Anal., **6**(2005), 117–136.

[12] B. Halpern, *Fixed points of nonexpanding maps*, Bull. Amer. Math. Soc., **73**(1967), 957–961.

[13] J.-B. Hiriart-Urruty, *ε-Subdifferential calculus*, Convex Analysis and Optimization, Res. Notes in Math., **57**(1982), 43–92.

[14] A. Kangtunyakarn, *The methods for variational inequality problems and fixed point of κ-strictly pseudononspreading mapping*, Fixed Point Theory Appl., **2013:171**(2013), 15 pp.

[15] F. Kohsaka and W. Takahashi, *Fixed point theorems for a class of nonlinear mappings related to maximal monotone operators in Banach spaces*, Arch. Math. (Basel), **91**(2008), 166–177.

[16] Y. Kurokawa and W. Takahashi, *Weak and strong convergence theorems for nonspreading mappings in Hilbert spaces*, Nonlinear Anal., **73(6)**(2010), 1562–1568.

[17] M. O. Osilike and F. O. Isiogugu, *Weak and strong convergence theorems for nonspreading-type mappings in Hilbert spaces*, Nonlinear Anal., **74(5)**(2011), 1814–1822.

[18] P. Santos and S. Scheimberg, *An inexact subgradient algorithm for equilibrium problems*, Comput. Appl. Math., **30**(2011), 91–107.

[19] W. Takahashi, *Introduction to nonlinear and Convex Analysis*, Yokohama Publishers, Yokohama, 2009.

[20] S. Takahashi and W. Takahashi, *Viscosity approximation methods for equilibrium problems and fixed point problems in Hilbert spaces*, J. Math. Anal. Appl., **331(1)**(2007), 506–515.

[21] E. Thailert, R. Wangkeeree and C. Khantree, *A Hybrid subgradient algorithm for finding a common solution of an equilibrium problem and a family of strict pseudo-contraction mappings,* J. Appl. Math., (2014), Art. ID 142671, 8 pp.

[22] E. Thailert, R. Wangkeeree and P. Preechasilp, *A new general iterative methods for solving the equilibrium problems, variational inequality problems and fixed point problems of nonexpansive mappings,* Thai J. Math., **14(1)**(2016), 53–67.

[23] H.-K. Xu, *Viscosity approximation methods for nonexpansive mappings*, J. Math. Anal. Appl., **298**(2004), 279–291.

Abdullah et al. *Advances in Difference Equations* (2018) 2018:66
https://doi.org/10.1186/s13662-018-1516-z

Advances in Difference Equations
a SpringerOpen Journal

RESEARCH Open Access

New mathematical model of vertical transmission and cure of vector-borne diseases and its numerical simulation

Abdullah[1], Aly Seadawy[2,3*] and Wang Jun[1]

*Correspondence:
Aly742001@yahoo.com
[2]Mathematics Department, Faculty of Science, Taibah University, Al-Madinah Al-Munawarah, Saudi Arabia
[3]Mathematics Department, Faculty of Science, Beni-Suef University, Beni-Suef, Egypt
Full list of author information is available at the end of the article

Abstract

In this research article, a new mathematical model for the transmission dynamics of vector-borne diseases with vertical transmission and cure is developed. The non-negative solutions of the model are shown. To understand the dynamical behavior of the epidemic model, the theory of basic reproduction number is used. As this number increases, the disease invades the population and vice versa. The effect of vertical transmission and cure rate on the basic reproduction number is shown. The disease-free and endemic equilibria of the model are found and both their local and global stabilities are presented. Finally, numerical simulations are carried out graphically to show the dynamical behaviors. These results show that vertical transmission and cure have a valuable effect on the transmission dynamics of the disease.

Keywords: Vector-borne disease; Vertical transmission; Cure; Stability; Numerical simulation

1 Introduction

Vector-borne diseases are infectious diseases transmitted to humans and animals by blood-feeding arthropods. Some common vector-borne diseases are West Nile virus, dengue fever, Rift Valley fever, malaria, and viral encephalitis caused by pathogens such as bacteria, viruses, and parasites. The arthropods are blood sucking insects and arachnids such as ticks, mosquitoes, biting flies, and lice called vectors [1]. The vectors receive pathogens from an infected host and transmit them to a human host, as humans are the major host, or animals. However, direct transmissions, such as transplantation related transmission, transfusion related transmission, and needle-stick-related transmission, are also possible [2]. In case of some diseases such as AIDS and Hepatitis B, it is possible for the offspring of infected parents to be born infected. This type of transmission is called vertical transmission. Now it is found that vector-borne diseases can also be transmitted vertically [3, 4]. Also new research shows that virus is transmitted from female mosquitos to their eggs at a high rate [5], which causes vertical transmission of the disease.

Vector-borne diseases are prevalent in hot areas, such as tropics and subtropics, and are relatively rare in temperate zones. Vector-borne infectious diseases remain amongst the

most important cause of global health illness and are major killers, particularly of children. The World Health Organization reports the numbers of deaths in different regions of the world annually. Nearly 700 million people get mosquito-borne illnesses that cause about one million deaths each year. Worldwide, malaria is the leading cause of premature mortality, particularly in children under the age of five. Nearly half of the world's population is at risk of malaria, and every year 198 million cases (uncertainty range: 124–283 million) and 584,000 deaths (range: 367,000–755,000) occur according to the World Malaria Report 2014 [6]. According to WHO, an estimated of 3.3 billion people in 97 countries are at risk of malaria. Currently, dengue threatens up to 40% of the world's population, and there may be 50–100 million infections annually [7]. More than 2.5 billion people over 40% of the world's population are now at risk of dengue.

From the above discussion it is clear that it is necessary to control such epidemic diseases. Control measures for vector-borne diseases are important because most are zoonoses. For the control measure, it is necessary to understand the dynamical features of diseases and treat the infected hosts. Therefore, deciphering the mechanisms and modeling of such diseases are of great interest. Our paper involves such an epidemic model for the transmission dynamics of vector-borne diseases that incorporates both horizontal and vertical transmission in the vector–host population.

Up to date, many mathematical models have been investigated to understand the mechanism of real world phenomena. Researchers investigate different methods to solve these models both analytically and numerically (e.g., see [8–21]). Several models of infectious diseases have been developed in the literature [22–27]. The model first proposed by Ross [28] and subsequently modified by Macdonald [29] has influenced both the modeling and the application of control strategies to a vector-borne disease. The model presented in [30] studied the analysis of a simple vector–host epidemic model with horizontal transmission. We extend their model by including vertical transmission in both vector and host populations, and treatment class in the host population with different interaction rates.

The structure of this paper is as follows: Section 1 represents the introductory remarks with a brief history. Section 2 is about the derivation of SITR epidemic model and shows the non-negative solutions of the proposed model. In Section 3, we find the disease-free and endemic equilibria and prove their local stability. In Section 4, we use mathematical analysis to establish global stability results for the proposed model. We use Lyapunov function theory to show global stability of both disease-free and endemic equilibria. Parameter estimation and numerical results are discussed in Section 5. Finally, we give conclusion.

2 Model framework

The total population sizes at time t for human hosts and vectors are denoted by $N_1(t)$ and $N_2(t)$, respectively. The population of size $N_1(t)$ is divided into four distinct classes: the susceptible population of size $S(t)$, the infectious population of size $I(t)$, the population under treatment of size $T(t)$, and the recovered population of size $R(t)$. Thus $N_1(t) = S(t) + I(t) + T(t) + R(t)$. The vector population $N_2(t)$ has the subclasses denoted by $V(t)$ and $W(t)$ for the susceptible and infected classes, respectively. Thus, $N_2(t) = V(t) + W(t)$. The mathematical model can be represented by the following nonlinear system of ordinary

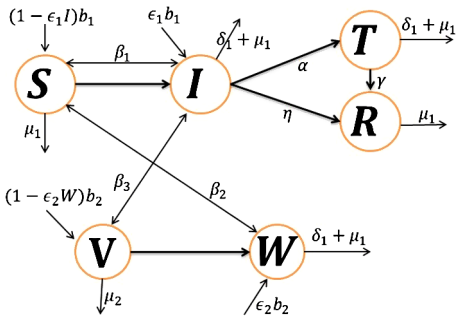


Figure 1 The flow chart represents the interactions and transfer of a vector-borne disease in both human and vector populations

differential equations:

$$\begin{aligned}
\frac{dS}{dt} &= (1-\epsilon_1 I)b_1 - \beta_1 SI - \beta_2 SW - \mu_1 S,\\
\frac{dI}{dt} &= \epsilon_1 b_1 I + \beta_1 SI + \beta_2 SW - \alpha I - \eta I - \delta_1 I - \mu_1 I,\\
\frac{dT}{dt} &= \alpha I - \gamma T - \delta_1 T - \mu_1 T,\\
\frac{dR}{dt} &= \eta I + \gamma T - \mu_1 R,\\
\frac{dV}{dt} &= (1-\epsilon_2 W)b_2 - \beta_3 VI - \mu_2 V,\\
\frac{dW}{dt} &= \epsilon_2 b_2 W + \beta_3 VI - \delta_2 W - \mu_2 W,
\end{aligned} \tag{1}$$

with the initial conditions

$$S(0)\geq 0, \qquad I(0)\geq 0, \qquad T(0)\geq 0, \qquad R(0)\geq 0, \qquad V(0)\geq 0, \qquad W(0)\geq 0. \tag{2}$$

The human host population is recruited at a constant birth rate b_1 in which a fraction ϵ_1 were born infected from their infected parents. β_1 is the rate of direct transmission of the disease, β_2 is the vector mediated transmission rate, μ_1 is the natural mortality rate of a human. Infectious humans are treated at a rate α, recover naturally at a rate η, and suffer disease-induced death at a rate δ_1. Treated humans recover at a rate γ. It is assumed that recovered individuals acquire lifelong immunity against re-infection. Similarly, b_3 is the constant recruitment rate of vector population in which the ratio ϵ_2 are infected by birth from their infected parents. Susceptible mosquitoes become infected by biting infected human at a rate β_3, μ_2 is the natural mortality rate of vector population. Infectious vectors die due to disease at a rate δ_2. The complete dynamics of the proposed model is represented by the flow chart in Figure 1.

2.1 Properties of solutions

The proposed model (1) is a system of nonlinear ordinary differential equations with the initial conditions (2). To be epidemiologically and mathematically meaningful, it is important to prove that all the solutions with the given initial conditions will remain nonnegative and bounded for all finite time. The model shall be analyzed in a biologically meaningful feasible region governed by a positive invariant set.

Theorem 2.1 *There exists a unique and bounded solution of the system of equations* (1), *in a positively invariant set, that remains for all finite time* $t \geq 0$.

Proof The right-hand side of each equation is continuous in the convex domain $E = (t, S(t), I(t), T(t), R(t), V(t), W(t))$ of $(6+1)$-dimensional space R_+^{6+1} with continuous partial derivatives. So problem (1) has a unique solution in R_+^6 which exists for a given finite time $t \in [0, \infty)$ and initial conditions (2).

As the total population sizes are $N_1 = S + I + T + R$ and $N_2 = V + W$, so from (1) we get

$$\frac{dN_1}{dt} = b_1 - \mu_1 N_1 - \delta_1(I + T) \quad \text{and} \quad \frac{dN_2}{dt} = b_2 - \mu_2 N_v - \delta_2 I_v. \tag{3}$$

Then

$$\begin{aligned} &\frac{dN_1}{dt} \leq b_1 - \mu_1 N_1 \quad \text{and} \quad \frac{dN_2}{dt} \leq b_2 - \mu_2 N_v. \\ &\Rightarrow \quad N_1 \leq N_1(0)e^{-\mu_1(t)} + \frac{b_1}{\mu_1}\left(1 - e^{-\mu_1(t)}\right) \quad \text{and} \\ &\qquad N_2 \leq N_2(0)e^{-\mu_2(t)} + \frac{b_2}{\mu_2}\left(1 - e^{-\mu_2(t)}\right), \end{aligned}$$

which shows that

$$\lim_{t\to\infty} \sup N_1 \leq \frac{b_1}{\mu_1} \quad \text{and} \quad \lim_{t\to\infty} \sup N_2 \leq \frac{b_2}{\mu_2}. \tag{4}$$

The given initial conditions (2) make sure that $N_1(0) \geq 0$ and $N_2(0) \geq 0$. Thus the feasible region for system (1) is

$$\Phi = \left\{(S, I, T, R, V, W) \in R_+^6, N_1 \leq \frac{b_1}{\mu_1}, N_2 \leq \frac{b_2}{\mu_2}\right\}.$$

Thus the total populations and each population class remain bounded for all finite time $t \geq 0$. □

The above theorem shows that model (1) is well posed epidemiologically and mathematically in a positively invariant set Φ. We shall study the dynamics of this basic model in Φ, so, all the solutions of system (1) start and remain in Φ for all $t \geq 0$. All the parameters and state variables for the model should be non-negative for all time because they represent the number of the population sizes of humans and vectors.

3 Equilibrium points

3.1 Disease-free equilibrium

The ability to invade a population is an important concern of an infectious disease. The steady state solutions of an epidemiological model at which the population remains in the absence of disease is called disease-free equilibrium point. In order to find the disease-free equilibrium of the proposed model (1), we set the right-hand side of all equations equal to zero and set $I = T = 0$ and $W = 0$. Also there is no infected recruitment in the populations, so we put the parameters $\epsilon_1 = \epsilon_2 = 0$, which implies that $(1 - \epsilon_1)b_1 = b_1$ and $(1 - \epsilon_2)b_2 = b_2$

mean that the total recruited population is only susceptible. By direct calculations, we get the disease-free equilibrium point E_1 in the feasible region Φ, which is given by

$$E_1 = (S_1, I_1, T_1, R_1, V_1, W_1) = \left(\frac{b_1}{\mu_1}, 0, 0, 0, \frac{b_2}{\mu_2}, 0\right).$$

The dynamics of model (1) is analyzed by a dimensionless number called basic reproduction number denoted by R_0, defined as "The expected number of secondary cases produced by a typical infected individual during its entire period of infectiousness in a completely susceptible population" [31]. Mathematically, R_0 is defined as

$$R_0 \propto \left(\frac{\text{infection}}{\text{contact}}\right) \cdot \left(\frac{\text{contact}}{\text{time}}\right) \cdot \left(\frac{\text{time}}{\text{infection}}\right).$$

More precisely,

$$R_0 \propto T \cdot C \cdot D,$$

where T is the transmissibility (i.e., probability of infection given contact between a susceptible individual and an infected one), C is the average rate of contact between susceptible and infected individuals, and D is the duration of infectiousness. This quantity serves as a threshold parameter that predicts whether a disease will spread in a community or will simply die out. It can be calculated by the method of next generation matrix given in [32]. In the vector–host model (1), infected states are I, T, and W and uninfected states are S, R, and V. The matrices $\mathcal{F}$ and $\mathcal{V}$ are the rate of production of new infections and the transition rates between states, respectively, which are given by

$$\mathcal{F} = \begin{pmatrix} \epsilon_1 b_1 I + \beta_1 SI + \beta_2 SW \\ 0 \\ 0 \end{pmatrix}, \qquad \mathcal{V} = \begin{pmatrix} (\alpha + \eta + \delta_1 + \mu_1)I \\ -\alpha I + (\gamma + \delta_1 + \mu_1)T \\ -\epsilon_2 b_2 W - \beta_3 VI + (\delta_2 + \mu_2)W \end{pmatrix}.$$

At the disease-free equilibrium $S = N_1 = \frac{b_1}{\mu_1}$, $I = T = 0$, $V = N_2 = \frac{b_2}{\mu_2}$, and $W = 0$. The Jacobian matrices at the disease-free equilibrium of $\mathcal{F}$ and $\mathcal{V}$ are F and V, respectively, where

$$F = \begin{pmatrix} \epsilon_1 b_1 + \beta_1 N_1 & 0 & \beta_2 N_1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$V = \begin{pmatrix} \alpha + \eta + \delta_1 + \mu_1 & 0 & 0 \\ -\alpha & \gamma + \delta_1 + \mu_1 & 0 \\ -\beta_3 N_2 & 0 & -\epsilon_2 b_2 + \delta_2 + \mu_2 \end{pmatrix}.$$

F and V are the rates for new infections and transitions near the equilibrium. We used MATLAB(R2010A) to find V^{-1} and FV^{-1}, which gives the times spent in each state and the total production of new infections over the course of an infection, respectively. The

largest eigenvalue of FV^{-1} is the basic reproduction number R_0, given by

$$R_0 = \frac{\epsilon_1 b_1 + \beta_1 N_1}{k} + \frac{\beta_2 \beta_3 N_1 N_2}{mk},$$

where $k = \alpha + \delta_1 + \mu_1 + \eta$ and $m = \delta_2 + \mu_2 - \epsilon_2 b_2$. When there is no vertical transmission, $\epsilon_1 = \epsilon_2 = 0$, then R_0 is the basic reproductive number for the model with only horizontal transmission. Geometrically it means that the number of new infections comes from both direct and indirect transmission. In the presence of vertical transmission, $\epsilon_1, \epsilon_2 > 0$, R_0 increases as these vertical transmission parameters increase, because vertical transmission directly increases the number of infectious populations. Also we can see the inverse relation of treatment strategies with R_0 and the direct relation with new infections and total population.

The basic reproduction number R_0 has a significant effect on the dynamics of infection. As we can see from the first and second equations of model (1),

$$\frac{dS}{dt} = b_1 - kR_0 I - \mu_1 S, \qquad \frac{dI}{dt} = k(R_0 - 1)I. \tag{5}$$

When $R_0 < 1$, it means that each infected individual infects less than one other individual averagely by ever kind of transmission, then the change in the number of infected population is negative, so the disease simply dies out. On the other hand, when $R_0 > 1$, it means that each infected individual infects more than one other individual, then the change is positive and invasion is always possible (see the survey paper by Hethcote [33]). For $R_0 = 1$, it means that each infectious individual infects one other individual as a whole, then there is no change in the infected population, so the infection constantly remains in the population. Also the effect of R_0 on the susceptible population is shown in the first equation of (5). All these facts are shown in Figures 2 and 3.

Theorem 3.1 *The disease-free equilibrium point E_1 is locally asymptotically stable if $R_0 < 1$, otherwise unstable.*

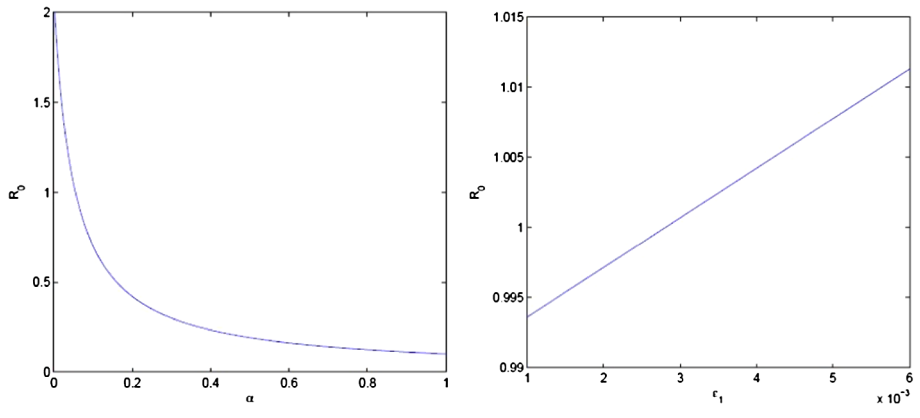


Figure 2 The first figure shows that R_0 decreases with increasing cure rate. The second figure shows that R_0 increases as vertical transmission increases

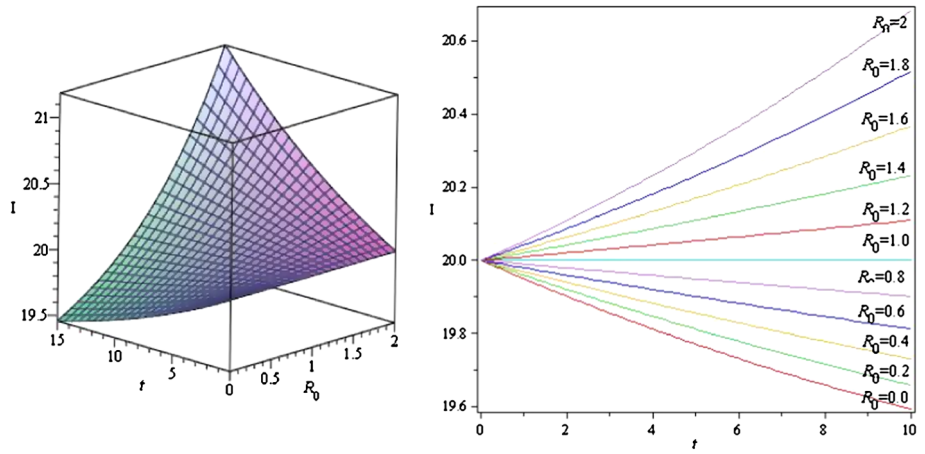


Figure 3 The figures show the threshold behavior of R_0 and its critical value $R_0 = 1$

Proof This can be proved by linearizing system (1) around E_1, which gives the following Jacobian matrix:

$$J_1 = \begin{bmatrix} -\mu_1 & -\epsilon_1 b_1 - \beta_1 \frac{b_1}{\mu_1} & 0 & 0 & 0 & -\beta_2 \frac{b_1}{\mu_1} \\ 0 & \epsilon_1 b_1 + \beta_1 \frac{b_1}{\mu_1} - k & 0 & 0 & 0 & \beta_2 \frac{b_1}{\mu_1} \\ 0 & \alpha & -l & 0 & 0 & 0 \\ 0 & \eta & \gamma & -\mu_1 & 0 & 0 \\ 0 & -\beta_3 \frac{b_2}{\mu_2} & 0 & 0 & -\mu_2 & -\epsilon_2 b_2 \\ 0 & \beta_3 \frac{b_2}{\mu_2} & 0 & 0 & 0 & -m \end{bmatrix},$$

where $l = \gamma + \delta_1 + \mu_1$.

The characteristic equation of J_1 is

$$(x + \mu_1)(x + \mu_1)(x + \mu_2)(x + l)\left(c_0 x^2 + c_1 x + c_2\right) = 0, \tag{6}$$

where

$$c_0 = \mu_1 \mu_2,$$

$$c_1 = k\mu_1\mu_2 + m\mu_1\mu_2 - \beta_1 b_1 \mu_2 - b_1 \epsilon_1 \mu_1 \mu_2,$$

$$c_2 = km\mu_1\mu_2(1 - R_0).$$

Four eigenvalues $-\mu_1$, $-\mu_1$, $-\mu_2$, and $-l$ out of six have a negative real part. The remaining two eigenvalues are the roots of the equation $c_0 x^2 + c_1 x + c_2 = 0$. For $R_0 < 1$ and $k + m > \beta_1 N_1 + b_1 \epsilon_1$, we have $c_1 > 0$ and $c_1 c_2 > 0$. So, according to the Routh–Hurwitz criteria [34], these two eigenvalues have a negative real part.

Since each eigenvalue of the characteristic equation (6) has a negative real part when $R_0 < 1$, according to the Routh–Hurwitz method [34], system (1) is locally asymptotically stable at the disease-free equilibrium point E_2 and unstable when $R_0 > 1$. The dynamical behaviors of the model at disease-free equilibrium are shown in Figure 4. □

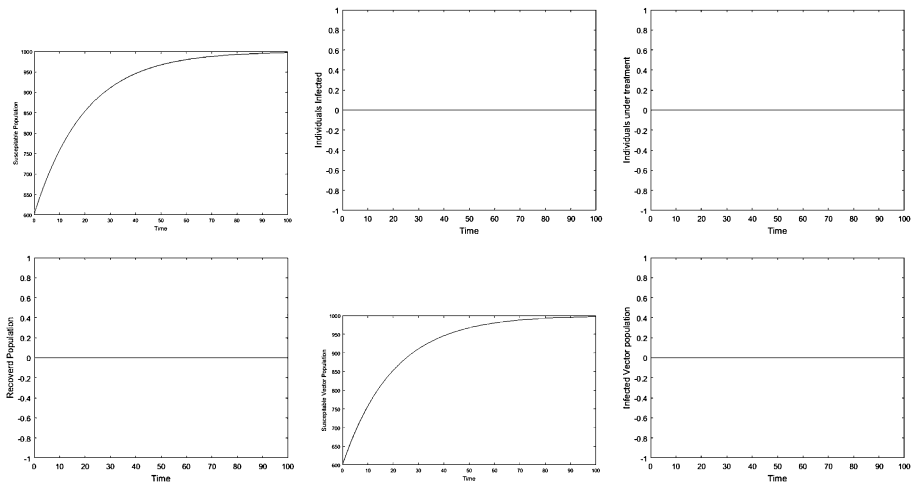


Figure 4 The plots show the dynamical behavior of the model at disease-free equilibrium

3.2 Endemic equilibrium

The constant presence of a disease or an infectious agent within a given geographic area is called endemic. The endemic equilibrium state is the state where the disease cannot be totally eradicated but remains in the population. In order to find positive solutions of system (1), let $E_2 = (S_2, I_2, T_2, R_2, V_2, W_2)$ represent any arbitrary endemic equilibrium. Setting left-hand side equal to zero and solving the equations simultaneously at steady state, we obtain

$$S_2 = \frac{b_1 - kI_2}{\mu_1}, \qquad T_2 = \frac{\alpha I_2}{l}, \qquad R_2 = \frac{(l\eta + \gamma\alpha)I_2}{\mu_1 l},$$
$$V_2 = \frac{mW_2}{\beta_3 I_2}, \qquad W_2 = \frac{\mu_2\beta_3 N_2 I_2}{\beta_3(\delta_2 + \mu_2)I_2 + \mu_2 m}.$$

Theorem 3.2 *The endemic equilibrium point E_2 is locally asymptotically stable if $R_0 > 1$, otherwise unstable.*

Proof To show these results, we linearize system (1) around E_2, which gives the following Jacobian matrix:

$$J_2 = \begin{bmatrix} -Q & -T & 0 & 0 & 0 & -\beta_2 S_2 \\ Q - \mu_1 & T - K & 0 & 0 & 0 & \beta_2 S_2 \\ 0 & \alpha & -l & 0 & 0 & 0 \\ 0 & \eta & \gamma & -\mu_1 & 0 & 0 \\ 0 & -\beta_3 V_2 & 0 & 0 & -\beta_3 I_2 - \mu_2 & -\epsilon_2 b_2 \\ 0 & \beta_3 V_2 & 0 & 0 & \beta_3 I_2 & -m \end{bmatrix},$$

where

$$Q = \beta_1 I_2 + \beta_2 W_2 + \mu_1, \qquad T = \epsilon_1 b_1 + \beta_1 S_2.$$

Two of the eigenvalues are $-\mu_1$ and $-l$. The remaining eigenvalues are the eigenvalues of the following matrix:

$$J_2^* = \begin{bmatrix} -Q & -T & 0 & -\beta_2 S_2 \\ Q-\mu_1 & T-K & 0 & \beta_2 S_2 \\ 0 & -\beta_3 V_2 & -\beta_3 I_2 - \mu_2 & -\epsilon_2 b_2 \\ 0 & \beta_3 V_2 & \beta_3 I_2 & -m \end{bmatrix}.$$

We make an elementary row operation for the Jacobian matrix J_2^* to obtain the following matrix:

$$J_2^* = \begin{bmatrix} -Q & -T & 0 & -\beta_2 S_2 \\ 0 & \frac{\mu_1 T}{Q} - K & 0 & \frac{\mu_1}{Q}\beta_2 S_2 \\ 0 & 0 & -\mu_2 & -\epsilon_2 b_2 - m \\ 0 & 0 & 0 & -M \end{bmatrix},$$

where

$$M = m + L + \frac{\beta_3 I_2}{\mu_2}(m + \epsilon_2 b_2) \quad \text{and} \quad L = m + \frac{\mu_1 \beta_2 \beta_3 S_2 V_2}{\mu_1 T - KQ}.$$

J_2^* is a lower triangular matrix and its eigenvalues are the elements of the main diagonal which are given by $-Q$, $\frac{\mu_1 T}{Q} - K$, $-\mu_2$, and $-M$. Three of the eigenvalues have a negative real part. The second eigenvalue $\frac{\mu_1 T}{Q} - K$ has a negative real part if and only if $\frac{\mu_1 T}{Q} - K < 0$. Using the value of Q and T, we can rewrite this equation by rearranging it as follows:

$$-2\beta_1\beta_3 K(\delta_2 + \mu_2) I_2^2 + \left[\beta_3(\delta_2 + \mu_2)\mu_1 K(1 - R_0)\right] I_2 + \mu_2 m \mu_1 K(1 - R_0). \tag{7}$$

All the coefficients of this equation are negative if $R_0 > 1$. Thus all the eigenvalues have negative real parts, which shows that the endemic equilibrium point E_2 is locally asymptotically stable iff $R_0 > 1$. □

4 Global stability analysis

In this section, we study the global analysis of the disease-free and endemic equilibria using the direct Lyapunov method which requires the construction of a function with specific properties. In order to do this, we derive the following results.

Theorem 4.1 *When $R_0 < 1$, then the disease-free equilibrium E_1 of system* (1) *is globally asymptotically stable on* Φ.

Proof To show the global stability of the disease-free equilibrium E_1, we construct the following Lyapunov function, following the method used in [35]:

$$U(t) = I + \frac{\beta_2 b_1}{m\mu_1} W, \quad \text{with time derivative } U'(t) = \dot{I} + \frac{\beta_2 b_1}{m\mu_1}\dot{W}. \tag{8}$$

Then U is C^1 on the interior of Φ, E_1 is the global minimum of U on Φ, and $U(t) = 0$ at E_1. Putting the values from model (1), we obtain

$$\begin{aligned} U'(t) &= \epsilon_1 b_1 I + \beta_1 SI + \beta_2 SW - \alpha I - \eta I - \delta_1 I - \mu_1 I \\ &\quad + \frac{\beta_2 b_1}{m\mu_1}(\epsilon_2 b_2 W + \beta_3 VI - \delta_2 W - \mu_2 W), \\ &\leq \epsilon_1 b_1 I + \beta_1 N_1 I + \beta_2 N_1 W - kI \\ &\quad + \frac{\beta_2 b_1}{m\mu_1}(\beta_3 N_2 I - mW), \quad \text{since } S \leq N_1, \text{ and } V \leq N_2 \\ &= (R_0 - 1)I. \end{aligned} \tag{9}$$

Equation (9) shows that $U'(t)$ is negative if $R_0 < 1$. Also $U'(t) = 0$ at E_1. Substituting $I = T = R = W = 0$ in the equations for $S(t)$ and $V(t)$ of model (1) shows that $S(t) \to \frac{b_1}{\mu_1}$ and $V(t) \to \frac{b_2}{\mu_2}$ as $t \to \infty$. Similarly, substituting in the equations for $T(t)$ and $R(t)$ shows that $(T(t), R(t)) \to (0, 0)$ as $t \to \infty$. Therefore the largest compact invariant set in $\{(S_h, E_h, I_h, N_h, S_v, E_v, I_v) \in \Phi : U'(t) = 0\}$ is the singleton disease-free equilibrium point $\{E_f\}$. Therefore, from LaSalle's principle [36], the disease-free equilibrium E_f is globally asymptotically stable in Φ. □

Theorem 4.2 *For $R_0 > 1$, the endemic equilibrium E_2 is globally asymptotically stable.*

Proof For the global stability of the endemic equilibria, we construct the following Lyapunov function:

$$Y(t) = \frac{1}{\beta_1 S_2}(S - S_2 \log S) + \frac{1}{\beta_3 V_2}(V - V_2 \log V) + \frac{1}{\beta_1 S_2} I + \frac{1}{\beta_3 V_2} W. \tag{10}$$

Taking the time derivative of W, we get

$$\begin{aligned} Y'(t) &= \frac{1}{\beta_1 S_2}(S - S_2)\left[\frac{b_1}{S} - \frac{\epsilon_1 b_1 I}{S} - \beta_1 I - \beta_2 W - \mu_1\right] \\ &\quad + \frac{1}{\beta_3 V_2}(V - V_2)\left[\frac{b_2}{V} - \frac{\epsilon_2 b_2 W}{V} - \beta_3 I - \mu_2\right] \\ &\quad + \frac{1}{\beta_1 S_2}[\beta_1 SI + \beta_2 SW - K_1 I], \end{aligned} \tag{11}$$

where $K_1 = \alpha + \delta_1 + \mu_1 + \eta - \epsilon_1 b_1$. Let us consider

$$\begin{aligned} &\mu_1 = \frac{b_1}{S_2} \quad \Rightarrow \quad b_1 = \mu_1 S_2, \qquad \mu_2 = \frac{b_2}{V_2} \quad \Rightarrow \quad b_2 = \mu_2 V_2, \\ &K_1 = 2\beta_1 S_2, \quad \text{and} \quad m = \frac{\beta_2 \beta_3 V_2}{\beta_1}. \end{aligned} \tag{12}$$

Rearranging equation (11), we get

$$Y'(t) = -\frac{\mu_1}{\beta_1}\left(\frac{S}{S_2} + \frac{S_2}{S} - 2\right) - \frac{\mu_2}{\beta_3}\left(\frac{V}{V_2} + \frac{V_2}{V} - 2\right). \tag{13}$$

Since

$$\frac{S}{S_2} + \frac{S_2}{S} \geq 2 \quad \text{and} \quad \frac{V}{V_2} + \frac{V_2}{V} \geq 2, \tag{14}$$

because the arithmetic mean is greater than or equal to the geometric mean. Thus $Y'(t) \leq 0$ for all $(S, I, T, R, V, W) \in \Phi$ and the equality $(Y'(t) = 0)$ holds for E_2. The proof is completed as in the proof of Theorem (4.1). □

5 Numerical simulation and graphs

We collect data from different sources and use the Runge–Kutta fourth order scheme to solve the model. Some of the parameter values are based on reality, for example, the death rate of humans by nature, corresponding to life expectancy of a 70-year-old human, is $\mu_1 = 0.000039$ per day, and the death rate of mosquitoes is $\mu_2 = 0.1$ per day corresponding

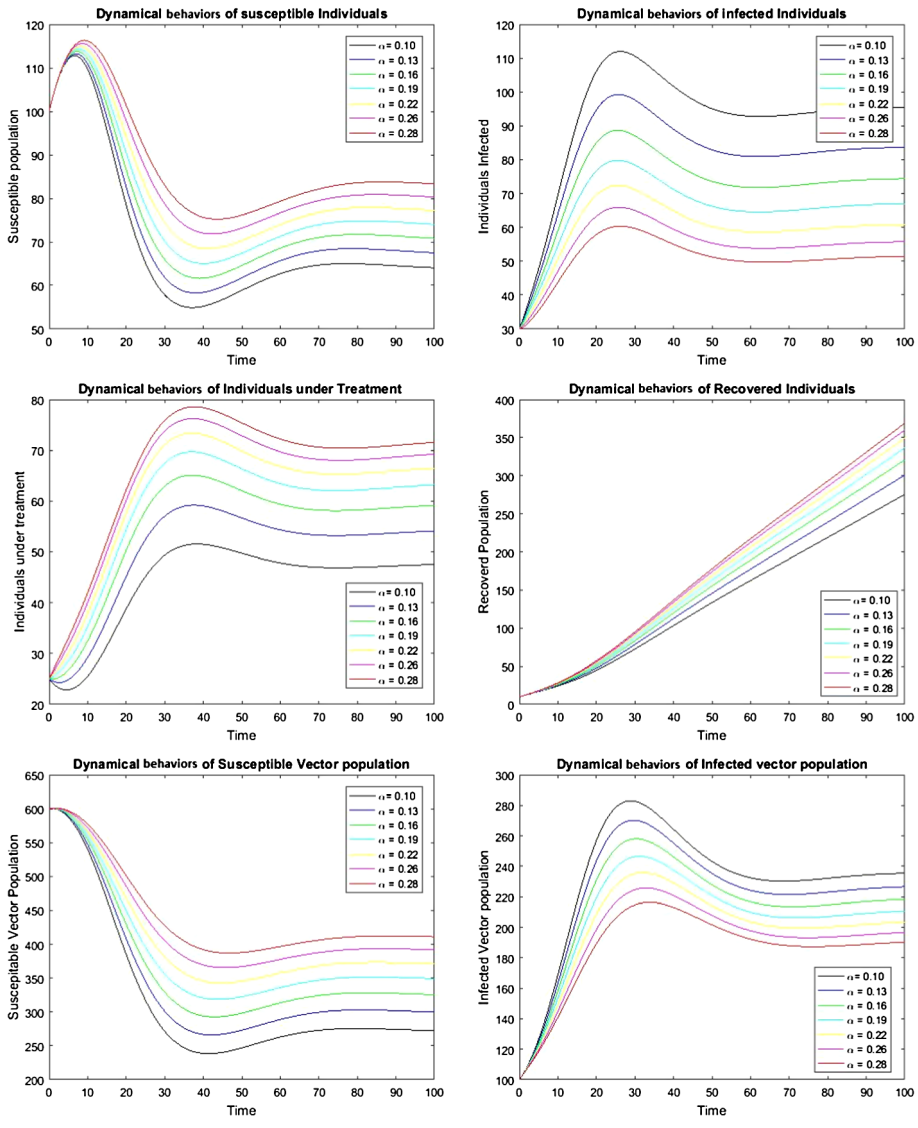


Figure 5 The plots show the dynamical behaviors of population sizes with increasing cure rate

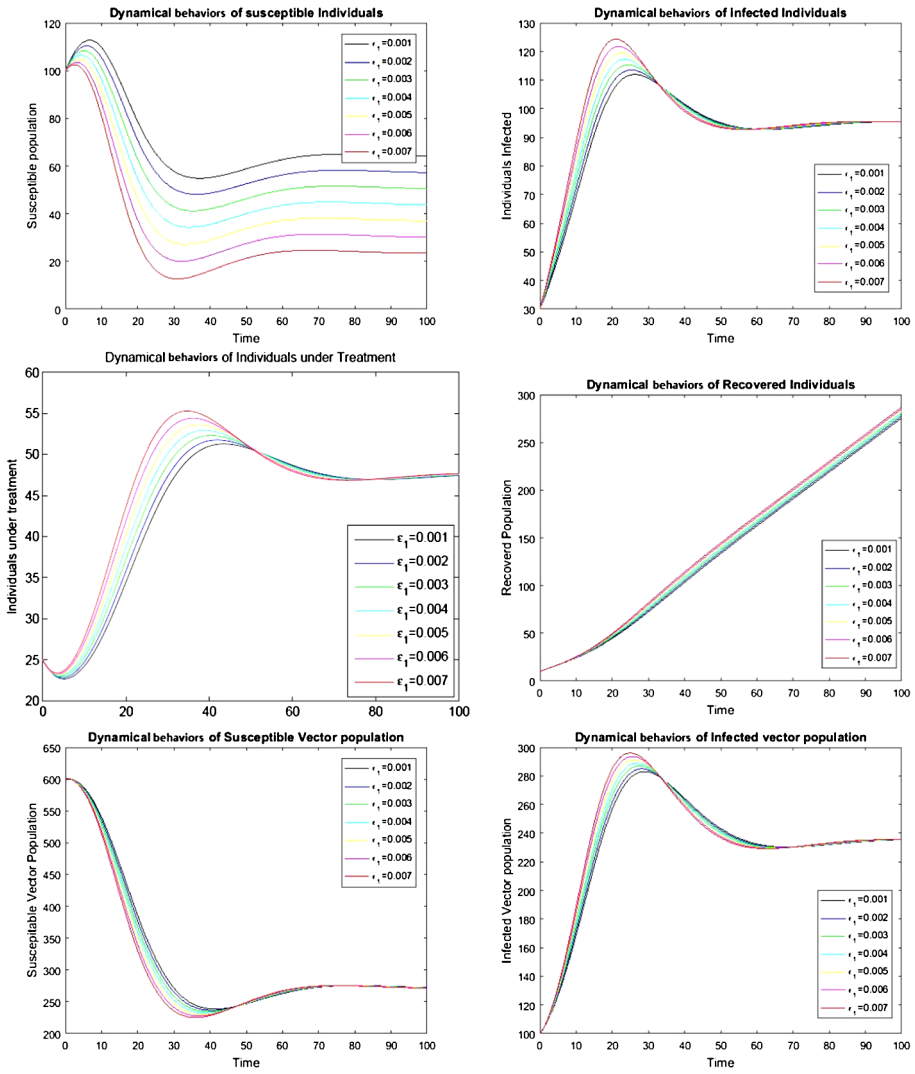


Figure 6 The plots show the dynamical behavior of population sizes with increasing vertical transmission rate

to mosquito's average life span of 10 days. Some of the parameter values are chosen from [25, 35]. The human's and vector's recruitment rates are $b_1 = 20$ and $b_2 = 100$ per day, respectively. The disease-induced death rates of humans and mosquitoes are $\delta_1 = 0.01$ and $\delta_2 = 0.21$, respectively. $\beta_1 = 0.00001$ and $\beta_2 = 0.0012$ are the transmission probabilities of dengue from human to human and vector to human population, respectively, $\beta_3 = 0.001$ is the transmission probability of dengue from human to vector population. Given different values to the treatment parameter $0 \leq \alpha \leq 1$ to check the treatment effects. The natural recovery rate is $\eta = 0.01$, and the recovery rate due to treatment is $\gamma = 0.4$. We suppose the values of ϵ_1, ϵ_2 and the initial population sizes. In rare cases the new offspring of infected parents are infected so take $\epsilon_1 = 0.001$ and the vertical transmission rate for mosquitos is $\epsilon_2 = 0.002$. For initial values, let $S(0) = 100$, $I(0) = 30$, $T(0) = 25$, $R(0) = 10$, $V(0) = 600$, and $W(0) = 100$. After solving we draw the results graphically and show the effect of cure rate and vertical transmission. Figure 5 shows the effect of cure rate on each population class, and Figure 6 shows the effect of vertical transmission. Figures 7 and 8 show the

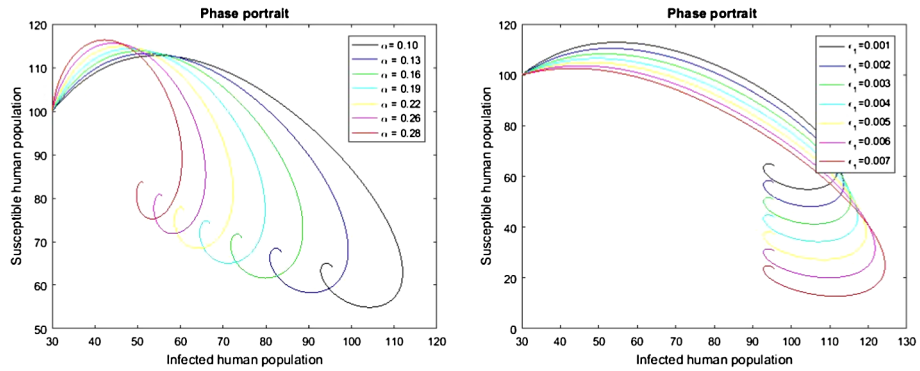


Figure 7 The plots show the phase portrait of the susceptible human population versus the infected human population

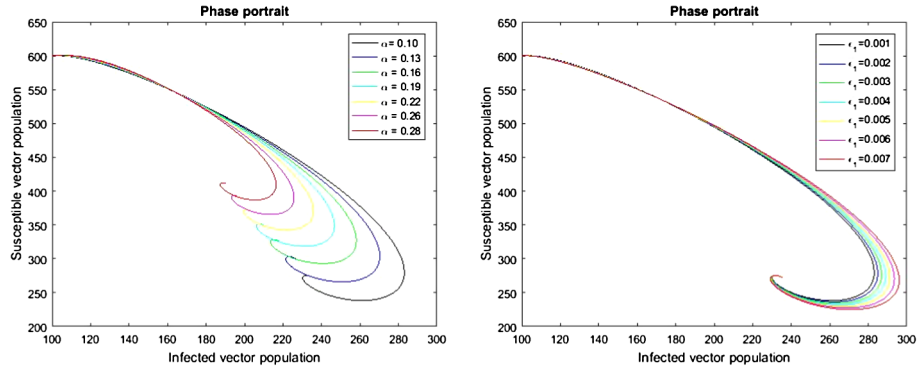


Figure 8 The plots show the phase portrait of the susceptible vector population versus the infected vector population

phase portraits of susceptible population versus infected population of human and vector populations, respectively.

6 Conclusion

The spread of different infectious diseases causes very high mortality rates in a population. Vector-borne diseases are infectious diseases transmitted to humans and animals through vectors. These diseases propagate from the infected to the susceptible population in different ways. This paper formulated an epidemic model for the transmission dynamics of vector-borne diseases with both vertical and horizontal transmissions with treatment strategy. The equilibrium points and the basic reproduction of the model are found. The basic reproduction number, which is a threshold quantity, has an important role in the epidemiology of the disease. As this number increases the disease invades the population, and as it decreases the disease simply dies out. Figure 2 shows that R_0 decreases as treatment strategies increase and increases as vertical transmission increases. Figure 3 shows the threshold behavior of R_0 and the critical value $R_0 = 1$. As R_0 increases, the infected population increases with time. For $R_0 < 1$, the number of infected population decreases; for $R_0 = 1$, the infected population remains constant; and for $R_0 > 1$, the number of infected population increases. It is also shown that when $R_0 < 1$ the disease-free equilibrium is lo-

cally and globally asymptotically stable; and for $R_0 > 1$, the positive endemic equilibrium is locally and globally asymptotically stable.

Numerical simulations are carried out graphically to show the dynamical behavior of the diseases. Figure 5 shows the effect of cure rate on the transmission dynamics of the disease. As treatment strategy increases, the susceptible population and the recovered human population increase while the infected population decreases. Figure 6 shows the effect of vertical transmission. As vertical transmission increases, the susceptible population decreases and the infected population increases. Finally, Figures 7 and 8 show the phase portraits of the susceptible populations versus the infected populations which move towards the stable points.

Acknowledgements
This work was supported by NNSFC (Grants 11571140, 11671077), Fellowship of Outstanding Young Scholars of Jiangsu Province (BK20160063), the Six Big Talent Peaks Project in Jiangsu Province (XYDXX-015), and NSF of Jiangsu Province (BK20150478).

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
All authors read and approved the final manuscript.

Author details
[1]Department of Mathematics, Faculty of Science, Jiangsu University, Zhenjiang, P.R. China. [2]Mathematics Department, Faculty of Science, Taibah University, Al-Madinah Al-Munawarah, Saudi Arabia. [3]Mathematics Department, Faculty of Science, Beni-Suef University, Beni-Suef, Egypt.

Publisher's Note

Received: 29 November 2017 Accepted: 6 February 2018 Published online: 21 February 2018

References

1. Service, M.W.: Blood-Sucking Insects: Vector of Disease. Arnold, Victoriya (1986)
2. Wiwanitkit, V.: Unusual mode of transmission of dengue. J. Infect. Dev. Ctries. **30**, 51–54 (2009)
3. Antonis, A.F.G., Kortekaas, J., Kant, J., Vloet, R.P.M., Vogel-Brink, A., Stockhofe, N., Moormann, R.J.M.: Vertical transmission of rift valley fever virus without detectable maternal viremia. Vector-Borne Zoonotic Dis. **13**, 601–607 (2013)
4. Turchetti, A.P., Souza, T.D., Paixão, T.A., Santos, R.L.: Sexual and vertical transmission of visceral leishmaniasis. J. Infect. Dev. Ctries. **8**(4), 403–407 (2014)
5. Buckner, E.A., Alto, B.W., Lounibos, L.P.: Vertical transmission of Key West Dengue-1 virus by Aedes aegypti and Aedes albopictus (Diptera: Culicidae) mosquitoes from Florida. J. Med. Entomol. **50**(6), 1291–1297 (2013)
6. The World Health Report 2014: Changing history, WHO, Geneva, Switzerland
7. World Health Organization: Dengue and dengue haemorrhagic fever, Geneva, Report (2002)
8. Yang, X.J., Gao, F.: A new technology for solving diffusion and heat equations. Therm. Sci. **21**, 133–140 (2017)
9. Yang, X.J.: A new integral transform with an application in heat-transfer problem. Therm. Sci. **20**, S677–S681 (2016)
10. Abdullah, Seadawy, A.R., Jun, W.: Mathematical methods and solitary wave solutions of three-dimensional Zakharov–Kuznetsov–Burgers equation in dusty plasma and its applications. Results Phys. **7**, 4269–4277 (2017)
11. Yang, X.J.: A new integral transform operator for solving the heat-diffusion problem. Appl. Math. Lett. **64**, 193–197 (2017)
12. Yang, X.J.: New integral transforms for solving a steady heat transfer problem. Therm. Sci. **21**(1), S79–S87 (2017)
13. Seadawy, A.R.: Two-dimensional interaction of a shear flow with a free surface in a stratified fluid and its solitary-wave solutions via mathematical methods. Eur. Phys. J. Plus **132**, 518 (2017)
14. Lu, D., Seadawy, A.R., Arshad, M.: Bright–dark solitary wave and elliptic function solutions of unstable nonlinear Schrödinger equation and their applications. Opt. Quantum Electron. **50**, 23 (2018)
15. Kumar, D., Seadawy, A.R., Joardar, A.K.: Modified Kudryashov method via new exact solutions for some conformable fractional differential equations arising in mathematical biology. Chin. J. Phys. **56**, 75–85 (2018)
16. Seadawy, A.R., El-Rashidy, K.: Traveling wave solutions for some coupled nonlinear evolution equations. Math. Comput. Model. **57**, 1371–1379 (2013)
17. Seadawy, A.R.: Stability analysis solutions for nonlinear three-dimensional modified Korteweg–de Vries–Zakharov–Kuznetsov equation in a magnetized electron-positron plasma. Physica A **455**, 44–51 (2016)
18. Lu, D., Seadawy, A.R., Arshad, M.: Applications of extended simple equation method on unstable nonlinear Schrödinger equations. Optik **140**, 136–144 (2017)
19. Seadawy, A.R.: Solitary wave solutions of two-dimensional nonlinear Kadomtsev–Petviashvili dynamic equation in dust-acoustic plasmas. Pramana J. Phys. **89**, 49 (2017)

20. Seadawy, A.R.: The generalized nonlinear higher order of KdV equations from the higher order nonlinear Schrödinger equation and its solutions. Optik **139**, 31–43 (2017)
21. Seadawy, A.R.: Three-dimensional nonlinear modified Zakharov–Kuznetsov equation of ion-acoustic waves in a magnetized plasma. Comput. Math. Appl. **71**, 201–212 (2016)
22. Blayneh, K.W., Jang, S.R.: A discrete SIS-model for a vector-transmitted disease. Appl. Anal. **85**, 1271–1284 (2006)
23. Bowman, C., Gumel, A.B., Driessche, P.V.D., Wu, J., Zhu, H.: A mathematical model for assessing control strategies against West Nile virus. Bull. Math. Biol. **67**, 1107–1133 (2005)
24. Ali, N., Zaman, G., Abdullah, Alqahtani, A.M., Alshomrani, A.S.: The effects of time lag and cure rate on the global dynamics of HIV-1 model. BioMed Res. Int. **2017**, Article ID 8094947 (2017)
25. Lashari, A.A., Zaman, G.: Global dynamics of vector borne disease with horizontal transmission in host population. Comput. Math. Appl. **61**, 745–754 (2011)
26. Khan, T., Zamana, G., Chohan, M.I.: The transmission dynamic and optimal control of acute and chronic hepatitis B. J. Biol. Dyn. **11**, 172–189 (2016)
27. Khan, T., Jung, I.H., Khan, A., Zaman, G.: Classification and sensitivity analysis of the transmission dynamic of hepatitis B, pp. 14–22 (2017)
28. Ross, R.: The Prevention of Malaria, 2nd edn. Murray, London (1911)
29. Macdonald, G.: The analysis of equilibrium in malaria. Trop. Dis. Bull. **49**, 813–828 (1952)
30. Lashari, A.A., Hattaf, K., Zaman, G., Li, X.-Z.: Backward bifurcation and optimal control of a vector borne disease. Appl. Math. Inf. Sci. **7**, 301–309 (2013)
31. Diekmann, O., Heesterbeek, J.A.P., Metz, J.A.J.: On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. J. Math. Biol. **28**, 365 (1990)
32. Driessche, P.V.D., Watmough, J.: Reproduction number and sub-threshold endemic equilibria for compartmental models of disease transmission. Math. Biosci. **180**, 29–48 (2002)
33. Hethcote, H.W.: The mathematics of infectious diseases. SIAM Rev. **42**, 599 (2000)
34. Rao, V.S.H., Rao, P.R.S.: Dynamic Models and Control of Biological Systems. Springer, Dordrecht (2009)
35. Garbab, S.M., Safi, M.A., Gumel, A.B.: Cross-immunity-induced backward bifurcation for a model of transmission dynamics of two strains of influenza. Nonlinear Anal., Real World Appl. **14**, 1384–1403 (2013)
36. LaSalle, J.P.: The Stability of Dynamical Systems. SIAM, Philadelphia (1976)

J. Math. Biol. (2018) 77:857–933
https://doi.org/10.1007/s00285-018-1229-7

Mathematical Biology

Mathematical modeling of climate change and malaria transmission dynamics: a historical review

Steffen E. Eikenberry[1,2] · **Abba B. Gumel**[3]

Received: 7 June 2017 / Revised: 16 March 2018 / Published online: 24 April 2018

Abstract Malaria, one of the greatest historical killers of mankind, continues to claim around half a million lives annually, with almost all deaths occurring in children under the age of five living in tropical Africa. The range of this disease is limited by climate to the warmer regions of the globe, and so anthropogenic global warming (and climate change more broadly) now threatens to alter the geographic area for potential malaria transmission, as both the *Plasmodium* malaria parasite and *Anopheles* mosquito vector have highly temperature-dependent lifecycles, while the aquatic immature *Anopheles* habitats are also strongly dependent upon rainfall and local hydrodynamics. A wide variety of process-based (or mechanistic) mathematical models have thus been proposed for the complex, highly nonlinear weather-driven *Anopheles* lifecycle and malaria transmission dynamics, but have reached somewhat disparate conclusions as to optimum temperatures for transmission, and the possible effect of increasing temperatures upon (potential) malaria distribution, with some projecting a large *increase* in the area at risk for malaria, but others predicting primarily a *shift* in the disease's geographic range. More generally, both global and local environmental changes drove the initial emergence of *P. falciparum* as a major human pathogen in tropical Africa some 10,000 years ago, and the disease has a long and deep history through the present. It is the goal of this paper to review major aspects of malaria biology, methods for

Steffen E. Eikenberry
seikenbe@asu.edu

Abba B. Gumel
agumel@asu.edu

1 Global Security Initiative, Arizona State University, Tempe, AZ, USA

2 Present Address: School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ, USA

3 School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ, USA

formalizing these into mathematical forms, uncertainties and controversies in proper modeling methodology, and to provide a timeline of some major modeling efforts from the classical works of Sir Ronald Ross and George Macdonald through recent climate-focused modeling studies. Finally, we attempt to place such mathematical work within a broader historical context for the "million-murdering Death" of malaria.

Keywords Malaria · Climate change · Ross–Macdonald · Thermal-response

Mathematics Subject Classification 01-02 · 92-02 · 92B05

1 Introduction

Malaria, a potentially deadly disease caused by protozoan parasites known as *Plasmodium* that infect and replicate within human blood cells, is spread between humans via the bite of the infected female adult *Anopheles* mosquito, and is one of the greatest infectious maladies to beset mankind. There are five (previously four) *Plasmodium* species that commonly infect humans, namely *P. falciparum*, *P. vivax*, *P. ovale*, *P. malariae*, and, very recently, *P. knowlesi* (Antinori et al. 2012). Of these, *P. vivax* and *P. falciparum* are preeminent by far, responsible for nearly all malaria deaths in 2015 (estimated at 438,000 by the World Health Organization (WHO) (WHO 2015), although another major estimate is appreciably higher, at 631,000 deaths (Gething et al. 2016), and the confidence intervals for both estimates are broad). Over 90% of all malarial mortality is attributable to *P. falciparum* in sub-Saharan Africa, where children under the age of five are chiefly burdened (WHO 2015), and Fig. 1 demonstrates the concentration of malaria risk in this region. Consequently, the focus in this paper is almost exclusively on *P. falciparum* malaria in Africa.

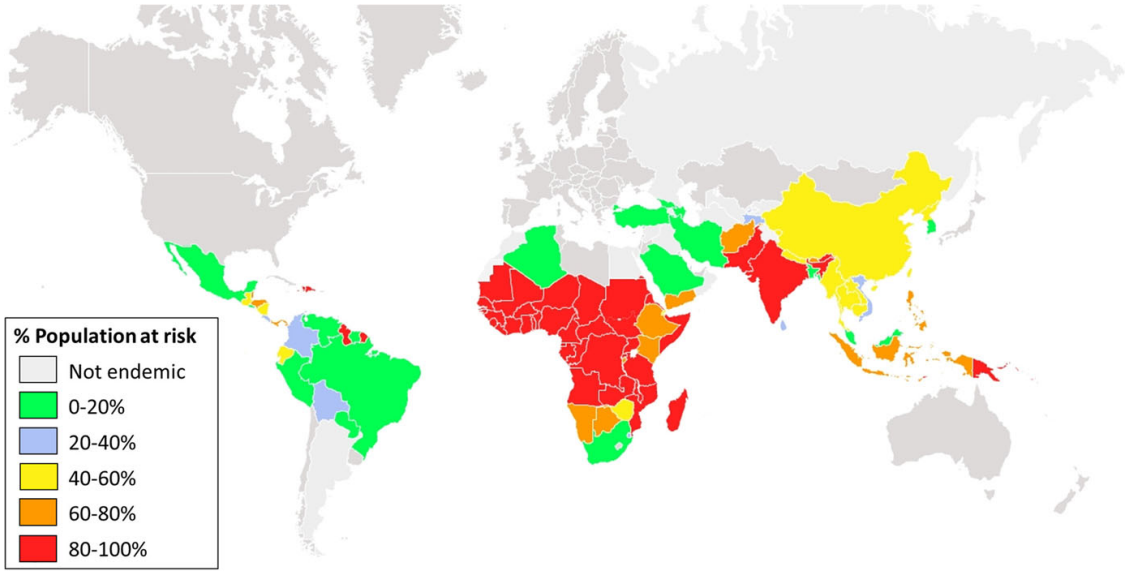


Fig. 1 Global populations at risk of malaria, in 2013. Tropical Africa is at highest risk, with many countries having 100% of their populations at risk; mortality is also strongly concentrated in this region. Map generated by the World Health Organization's Malaria Mapper (http://www.worldmalariareport.org/node/68), based on the World Malaria Report, 2015 (WHO 2015)

The emergence of *P. falciparum* as a major human disease, likely dating back to the acquisition of *P. falciparum* from a gorilla in Africa some 10,000 years ago (Loy et al. 2017; Carter and Mendis 2002), was directly linked to environmental changes, namely, the end of the last ice age leading to an era of global warming and the subsequent birth of human agricultural civilization, which, via land-use changes and the concentration of human settlement, allowed malaria and its mosquito vectors to thrive (Carter and Mendis 2002; Webb 2014; Packard 2007). The parasites and vectors, having temperature- and rainfall-dependent lifecycles, are restrained by climate to the globe's warmer latitude and altitude ranges (Patz et al. 1996). Thus, in the modern era, anthropogenic global warming, driven principally by fossil fuel combustion, but secondarily by global land-use changes (IPCC 2013) (primarily deforestation (IPCC 2013), which is in turn driven mainly by agriculture (Rudel et al. 2009; McKinley et al. 2011)), threatens to expand the potential range, and possibly the overall burden as well, of malarial disease. Aside from global restraints, malaria incidence follows altitude in multiple countries, such as Zimbabwe and Kenya (Patz et al. 1996), and the recent expansion of disease into some upland areas, notably the highlands of western Kenya, may be at least partly attributable to warmer temperatures (Pascual et al. 2006; Pascual and Bouma 2009).

However, the ultimate effect of climate change on malaria is far from certain, as a wide milieu of social, biotic, and abiotic factors influence the disease in non-linear ways, and the global burden of malaria contracted enormously over the twentieth century in the face of modest warming (Carter and Mendis 2002; Gething et al. 2010) (although this pattern generalizes poorly to malaria in Africa (Carter and Mendis 2002)). Over the last few decades, a number of mathematical models, typically statistical (using data and statistical approaches to correlate some climate variables with malaria incidence) or mechanistic (accounting for the detailed dynamic nonlinear processes involved in disease transmission, also sometimes referred to as "process-based"), have been employed to assess the likely impact of anthropogenic climate change on malaria transmission dynamics and control. These models have reached divergent conclusions, with some predicting a large *expansion* in the continental land area suitable for transmission (Martens et al. 1999; Caminade et al. 2014; Tanser et al. 2003) and in the number of people at risk of malaria (Martens et al. 1999; Patz et al. 1996; Pascual et al. 2006), while others predict only modest poleward (and altitudinal) *shifts* in the burden of disease, with little net effect (Gething et al. 2010; Rogers and Randolph 2000; Hay et al. 2002), and the issue remains unresolved thus far. The goal in this paper is not to ultimately resolve this issue (laudable as it is), but to attempt to lay a foundation to aid such a resolution.

Malaria was one of the first human diseases to be subject to mathematical inquiry. Sir Ronald Ross, who first elucidated how *Plasmodia* spread via the intermediary mosquito (Cox 2010), proposed a mechanistic transmission model including the human host and the mosquito in the early 1900s, although it did not address the all-important mosquito lifecycle (beyond infection of a constant population) (Smith et al. 2012). Following in Ross's footsteps, the highly influential malariologist George Macdonald reformulated the basic model in the early 1950s (Macdonald 1952, 1956a, b, 1957) (presented in detail in Sect. 4.2), and derived an expression for the *basic reproduction number*, $\mathcal{R}_0$, defined as the average number of secondary cases a single

initial case will generate in a completely susceptible (uninfected and non-immune) population.[1] Macdonald showed that $\mathcal{R}_0$ is most sensitive to changes in adult mosquito survival probability, thereby providing a theoretical rationale for insecticide spraying as the foundation of malaria eradication efforts during the 1960s (Macdonald 1956b; Nájera et al. 2011). The "Ross–Macdonald" model has been extremely influential: Reiner et al. (2013), in a systematic review of 388 models of mosquito-borne pathogens found in the literature between 1970 and 2010, determined most to be similar to the Ross–Macdonald framework, and Macdonald's expression for $\mathcal{R}_0$ has also been used in many climate-focused (or climate-driven) mechanistic modeling studies.

Moving back to the effect of climate change, there has been significant controversy as to its likely effect on human malaria (and other diseases) and this has been informed by modeling studies. Several works in the 1990s by Martens, Lindsay and colleagues (Martens et al. 1995a, b; Lindsay and Martens 1998; Martens et al. 1997, 1999), using Macdonald's $\mathcal{R}_0$, and drawing from several sources quantifying how parasite and vector lifecycle parameters might change with temperature, predicted that a significantly expanded area of the globe could become vulnerable to epidemic malaria under climate change. However, these conclusions did not go unchallenged.

Rogers and Randolph (2000) were critical of these process-based methods, and instead employed a statistical method, whereby they inferred climatologic limits to malaria based on temperature, rainfall, and saturation vapor pressure, and current malaria distribution, and then projected how malaria suitability would change in the future, under global climate model (or general circulation model, GCM) projections, finding a decrease in some mainly equatorial areas and a modest poleward increase, with little net change overall; this agrees in principle with a historical time-series analysis by Small et al. (2003). Other authors have similarly argued that climate change is more likely to induce a geographic *shift* in the burden of disease, with little net increase (Lafferty 2009). Even if true, Pascual and Bouma (2009) have pointed out that a geographically balanced shift does not equal a population-balanced shift: highland regions in eastern Africa, most likely to become vulnerable to malaria under a warming climate, are also far more populous than nearby lowland areas that could see a decline in malaria burden.

Gething et al. (2010) also argued, essentially, that because the global burden of malarial disease decreased dramatically from 1900 to 2007 while global mean air temperatures increased (the average temperature increase from the 1850–1900 period to the 1986–2005 period was 0.61 °C (IPCC 2014)), then non-climactic factors must be of vastly overriding importance and that climate change will affect malaria but little in the future. They bolstered this argument by estimating how $\mathcal{R}_0$ must have changed overall since 1900 and in response to different interventions, based on a Ross–Macdonald-style model for $\mathcal{R}_0$, and concluded that projected mean increases

[1] For autonomous mechanistic models (i.e., models that do not incorporate explicit dependency of time or climate variables on the right-hand sides of the equations of the model), $\mathcal{R}_0$ is typically computed using standard linearization or the *next generation operator* method (Diekmann et al. 1990; Diekmann and Heesterbeek 2000; Van den Driessche and Watmough 2002). For non-autonomous mechanistic models, such as the weather-driven models given in Sect. 6 by Eqs. (52)–(58) and (61)–(66), $\mathcal{R}_0$ is numerically approximated using the *next infection operator* approach, as formulated in Bacaër (2007) and Wang and Zhao (2008).

in $\mathcal{R}_0$ under future warming (Martens et al. 1997, 1999; Lindsay and Martens 1998) are one to two orders of magnitude smaller (and, thus, likely trivial). Agricultural practices and land-use, in particular, have been considered a human factor of greater importance than climate change (Lafferty 2009).

A variety of newer process-based models for the transmission cycle were developed in the decade after Martens, varying in their basic construction and hypotheses for the effects of rainfall and temperature on both vector and parasite (see, for instance, (Hoshen and Morse 2004; Bomblies et al. 2008; Parham and Michael 2010; Alonso et al. 2011; Ermert et al. 2011a, b; Parham et al. 2012; White et al. 2011)), but generally concluded that increasing temperatures favor malaria transmission. For example, Parham and Michael (2010) concluded in 2010 that transmission is optimized in the 32–33 °C temperature range. Caminade et al. (2014) published projections for the population at risk of malaria using five malaria models from this period, suggesting a net increase in the global population at risk of malaria, but with high uncertainty.

Mordecai and colleagues (Mordecai et al. 2013), in an influential paper published in 2013, used a set of *unimodal* functions (i.e., hump-shaped) for the relationship between temperature and vector parameters (such as larval development rate, larval survival, adult survival, biting rate, fecundity, and vector competence), as well as the parasite development rate, in contrast to many prior works, which had used *monotonic* relationships for some or all of these (temperature-dependent) parameters. Using a newer expression for $\mathcal{R}_0$, based on the model by Parham and Michael (2010), Mordecai and colleagues concluded that malaria transmission is optimized at a significantly lower temperature range, 25–28 °C, and found this to better match field measurements of the entomological inoculation rate (EIR) (the infectious biting rate).

Subsequently, Ryan et al. (2015b) used the Mordecai et al. (2013) thermal-response curves to develop a series of maps for malaria transmission potential across Africa from the year 2000 to 2080 under a mid-range emissions scenario (SRES A1B). Broadly speaking, this work predicted a modest increase in the total land area at risk for *any* malaria transmission, while the net area suitable for intense, year-round transmission would decrease (especially in western coastal Africa). Furthermore, these authors predicted increased malaria potential in the cooler southern and eastern regions of Africa, but a decrease in the hotter western and central African regions (especially the Democratic Republic of Congo) by 2080, and an southeasterly shift over time in the populations most at risk of malaria, with notable increases in the Lake Victoria region (near the Kenyan highlands) and eastern highland Madagascar. This work is especially laudable in its nuanced approach to malaria transmission potential, differentiating between year-round and seasonal potential, and consideration of the *populations*, not just geographic areas, at risk.

Despite its virtues, the work of Ryan et al. (2015b) did not explicitly consider rainfall or hydrodynamics, but applied a mask that limited transmission only to those regions with enough vegetation to be considered wet enough to support anopheline habitat. Earlier (process-based) malaria potential maps based on temperature, e.g. that of Craig et al. (1999), somewhat similarly restricted transmission to areas with grossly sufficient rainfall. Indeed, most of the works reviewed thus far have focused primarily on ambient temperature as an explanatory variable, with rainfall often a secondary, and variously modeled, factor. Given the absolute necessity of appropriate aquatic habi-

tat to the vector lifecycle, hydrodynamics and habitat modeling at both the regional and micro-scale represent a relatively (but not entirely) neglected factor. A variety of relatively simple relations between rainfall and immature mosquito survival and carrying capacity have been employed (Yé et al. 2009; White et al. 2011; Hoshen and Morse 2004), while several more complex efforts (Paaijmans et al. 2008a, b; Parham et al. 2012; Asare et al. 2016a, b) have physically modeled the heat and water balance within *Anopheles* microhabitats, as reviewed in Sects. 5.2.5 and 5.2.6. Several authors have additionally modeled regional hydrodynamics, e.g. (Bomblies et al. 2009; Bomblies 2012; Tompkins and Ermert 2013; Asare et al. 2016c). Of especial note, Bomblies and colleagues have considered detailed hydrodynamics at the village scale (Bomblies et al. 2008, 2009; Bomblies 2012), and concluded that such detailed modeling is necessary to explain both interseasonal variation (Bomblies 2012) and intervillage variation in vector abundance (Bomblies et al. 2009), and this modeling formed the basis for a recent comprehensive study suggesting little effect of climate change on malaria incidence in western Africa (Yamana et al. 2016).

While much controversy has centered on the appropriate functions relating vector and parasite parameters to temperature (and secondarily, to rainfall) and how variations in these drive climate-related predictions, more basic modeling choices also affect model predictions. In particular, the population biology of the *Anopheles* vectors is crucial to understanding many aspects of the disease, as well as assessing control strategies and projecting future outcomes. Malaria models that do not incorporate the dynamics of the juvenile stages of the mosquito are known to give results that do not generally match observed epidemiology (Okuneye and Gumel 2017; Beck-Johnson et al. 2013), and the vector lifecycle *per se* is the focus of several models (Beck-Johnson et al. 2013), most recently by Abdelrazec and Gumel (2017), who studied the effect of both temperature and rainfall on the population biology of mosquitoes. Another fundamental issue is that most vector and parasite lifecycle process times (e.g., larval development time) are non-exponentially distributed, yet most differential equations-based disease transmission models implicitly assume exponentially-distributed waiting times, an assumption found to affect model dynamics unfavorably by Christiansen-Jucht et al. (2015) and Lunde et al. (2013b).

Addressing this deeper problem of model construction, Gumel and colleagues (Agusto et al. 2015; Okuneye and Gumel 2017), have recently developed and analyzed several complex weather-driven mechanistic models that extend the prior studies by incorporating a broader array of biological, ecological and epidemiological factors, such as the dynamics of immature mosquitoes, host age-structure (Okuneye and Gumel 2017) and host immunity-boosting due to repeated exposure to malaria infection (Agusto et al. 2015). In particular, Agusto et al. (2015), adopting the thermal-response functions of Mordecai et al. (2013), and using a 14-dimensional mechanistic model and weather data for numerous locations within Africa, predicted that malaria infection generally increases in the 16–28 °C range, but decreases beginning at temperature values between 25 and 28 °C, depending on the African region (these results are comparable to those of Mordecai et al. (2013), but more nuanced). Yamana et al. (2013) also extended a prior agent-based model by Bomblies et al. (2008) to include partial immunity induced by repeated infection, and predicted that immunity can damp both the spatial and temporal variation in clinical disease in response to environmental variabil-

ity (Yamana et al. 2013, 2017). It should be emphasized that many prior weather-driven malaria modeling studies do not include immunity (or use only very simple representations of immunity), even though it is known that the unique malaria immune response is fundamental to malaria epidemiology and pathogenesis, and is itself the focus of a long modeling tradition; see, for instance, (Dietz et al. 1974; Aron 1988; Gupta and Day 1994; Gupta et al. 1999a, b; Filipe et al. 2007; Griffin et al. 2010, 2015) and Sect. 7.1.

Another recent effort is that of Okuneye and Gumel (2017), who additionally incorporated age-structure (as stated earlier, age-structure is crucially important because children under the age of five suffer the majority of the malaria burden in endemic areas) into a mechanistic temperature- and rainfall-dependent model, finding transmission to be maximized in the 21–25 °C temperature and 95–125 mm rainfall ranges in the Kwa-Zulu Natal province of South Africa.

Yet another basic issue that must be mentioned is that of diurnal temperature variation, and the (time-varying) disparity between ambient air and water temperature. Paaijmans et al. (2010) demonstrated empirically that the magnitude of temperature fluctuation affects *Anopheles* development and survival in a manner not captured by mean temperature alone. Average diurnal temperature range varies on a continental scale (Paaijmans et al. 2010), and this therefore may be an under-appreciated parameter in malaria potential projections. Diurnal temperature variations have not been considered in most models, although there are some recent exceptions, e.g. (Agusto et al. 2015; Beck-Johnson et al. 2017). Furthermore, the water temperature in immature mosquito habitats generally differs from ambient air; this disparity may be captured by physical hydrodynamic modeling, although a simple linear offset is sometimes assumed (Agusto et al. 2015). Finally, adult anophelines are also exposed to multiple microenvironments with varying temperatures, and often prefer to feed and/or rest indoors, where temperatures are typically warmer on average, but also less variable than out-of-doors (Afrane et al. 2005; Blanford et al. 2013; Singh et al. 2016).

While many malaria modeling studies have focused on the global scale (i.e., the potential global malaria range due to climate change), studies more limited in scale may provide better insight (Pascual and Bouma 2009; Alonso et al. 2011). In particular, a model region is the highlands of East Africa, where malaria burden was previously rare but has become more common since the 1970s; this increase may be at least partially attributable to global warming (Pascual and Bouma 2009). Human activity in the Kenyan highlands is recapitulating, in some sense, the early social and climatic changes that first gave birth to *P. falciparum* some 10,000 years ago. Temperatures are increasing (Pascual et al. 2006), the rain forests have recently been mostly cleared for crops, cattle grazing, logging, and housing construction (Minakawa et al. 1999), and the region is subject to intense population growth and human migration. Several researchers have made this area their focus (e.g., Githeko and Ndegwa 2001; Hay et al. 2002; Zhou et al. 2004; Pascual et al. 2006, 2008; Chaves and Koenraadt 2010; Alonso et al. 2011; Snow et al. 2015), and we suggest that a more limited geographic scope of study may better elucidate the competing effects of treatment, land use, migration, and climate on malaria. Also of note, malaria is highly endemic in hotter western Africa, an area which is also the focus of several studies, and the effect of climate change in this region could, conversely, be to slightly reduce malaria potential (Ryan et al. 2015b; Yamana et al. 2016).

In summary, although there is general agreement that climate change will increase the *potential* for malaria transmission at more northerly and southerly latitudes (and at higher altitudes), it is unclear if this represents a *shift* in malaria distribution with little net increase (or even decrease) in malaria burden, or an *expansion* in burden. The most likely scenario may be a hybrid result, with net expansion in malaria range, but shifts in the intensity of transmission within that range, especially towards southern and eastern Africa and highland areas. Further, the magnitude of the climate effect, and how it compares to other anthropogenic and abiotic factors, remains uncertain.

Malaria is a complex disease, with a complex history, and the controversy just outlined cannot be fully addressed without a broad background. The goal of this paper is to provide the reader with at least some of the requisite background needed for effective modeling of the disease dynamics, and to provide sufficient resources to help the reader in beginning their own investigations. Finally, it should be noted that mathematical models can, broadly speaking, be divided into the classifications of non-parametric and parametric (with parametric models also referred to as "process-based" or "mechanistic"), where the former attempt to inferentially draw conclusions directly from (usually time-series) data without positing any particular mechanistic system, while the latter posit some particular hypothesis for a system's workings (expressed mathematically). In this paper, our focus shall be on the latter.

This paper is organized as follows. We begin with an overview of the malaria lifecycles, immunology, and epidemiologic principles to establish a basis for later sections. To properly appreciate the role of mathematical modeling in providing deeper qualitative and quantitative insight on the transmission dynamics and control of malaria, some familiarity with the historical development of modeling frameworks and concepts is invaluable. To this end, we first present a historical overview of the disease in general, and move on to quantitative malariology through the early twentieth century, focusing on the early but deeply influential work of Ross and Macdonald. We also touch on some important later extensions by authors including Garrett-Jones, Dietz, and Molineaux. We then shift focus to climate, beginning with an extensive discussion of the anopheline and parasite lifecycles and their relation to weather (mainly temperature and rainfall), since these are fundamental to any predictions we care to make. In Sect. 6, we subsequently present a partial genealogy of recent mathematical works addressing weather and malaria transmission, and close with a brief discussion of multi-patch meta-population modeling, which may be of especial importance in understanding the spread of malaria between lowland and highland regions of Kenya. Finally, we briefly discuss other aspects of the disease that are pertinent to a fully comprehensive quantitative modeling framework (such as malaria immunity).

2 Introduction to malaria lifecycles, immunology and clinical disease

2.1 Parasite lifecycle

Plasmodium spp., the causative agent in malaria, are sexually-reproducing eukaryotic protozoans that undergo a complex lifecycle that requires switching between evolutionarily-distant vertebrate and invertebrate dipterian hosts. The basic evolution-

ary logic follows. Pre-*Plasmodium* parasites likely evolved from free-living sexual protozoans to live *extracellularly* in the midgut of aquatic invertebrates (Carter and Mendis 2002). Proliferative potential was then increased with the evolution of a second parasitic *intracellular* asexual reproductive stage, known as *schizogony*, by which a single cell may produce vast numbers of daughter cells, or spores. A minority of these daughter spores differentiate into male and female forms, which then recombine in a form of extracellular sexual reproduction known as *sporogony* (Antinori et al. 2012). The *Plasmodia*'s evolutionary innovation is to spatially separate the schizogonic cycle and sporogonic cycle into two separate hosts, with sporogony occurring in the mosquito.

Let us consider the particulars of human *Plasmodia*, where schizogony (asexual clonal expansion of many daughter spores) occurs in the human host, and in two phases: first in liver hepatocytes and then within red blood cells (RBCs, or erythrocytes). Sporogony then occurs in the mosquito midgut following a blood meal, to ultimately yield parasitic forms infectious to humans (Antinori et al. 2012). We may consider the cycle to begin with the bite of an infectious mosquito, who probes the dermis and injects saliva containing no more than 10–100 highly motile asexual *sporozoites* (Antinori et al. 2012). Sporozoites penetrate into blood vessels within minutes, travel to the liver and establish infection in hepatocytes within 30 min of biting (Guilbride et al. 2012). While the skin has traditionally been thought of as a passive waypoint in the infection cycle, more recent data indicates that a small number of sporozoites remaining in skin may exploit the inherently immunoregulatory nature of this environment to suppress anti-*Plasmodium* immunity and induce tolerance (Crompton et al. 2014), with important implications for vaccine development (Guilbride et al. 2012).

Shifting focus, sporozoites within hepatocytes initiate the first round of shizogony, so-called "pre-erythrocyte" shizogony, proliferating asexually to produce, in the case of *P. falciparum*, up to 30,000–40,000 asexual *merozoites* (Antinori et al. 2012; Crompton et al. 2014) contained within a "tissue schizont". Once mature, the tissue schizont, along with the parent hepatocyte, ruptures to spill the merozoites into the bloodstream, where they actively infect red blood cells, initiating the erythrocyte cycle of schizogony (Antinori et al. 2012), whereby merozoites expand, via several intermediate stages, within the erythrocyte and rupture it every 24–72 h (48 h for *P. falciparum*), freeing more merozoites to repeat the cycle (Antinori et al. 2012). It should be noted that while this is the end of the hepatic stage for *P. falciparum*, *P. vivax* and *P. ovale* have a dormant liver form known as the hypnozoite (Greek "sleeping animal") that can cause reinfection years later (Carter and Mendis 2002).

Erythrocytes, lacking a nucleus and most typical eukaryotic organelles, are essentially masses of hemoglobin, an iron-containing oxygen-carrying molecule, wrapped in plasma membrane and suited only for passive O_2 and CO_2 transport. *Plasmodia*, on the other hand, are "fully realized" eukaryotes, that hijack completely the erythrocytes they invade (Tilley et al. 2011). An invading merozoite passes through an immature "ring" stage to become a *trophozoite*, a feeding form that consumes 70% of the erythrocyte hemoglobin, converting it to the toxic byproduct hematin, which is then detoxified to hemozoin (Baton and Ranford-Cartwright 2005; Tilley et al. 2011). Notably, quinine antimalarial drugs act by preventing the detoxification of hematin (Tilley et al. 2011), and artemisinin, the most effective antimalarial, is also likely

involved in hemoglobin digestion (Tilley et al. 2011). The nourished trophozoite then becomes a "blood schizont," dividing asexually into 6–36 (20 on average) daughter merozoites that are released with the host cell's rupture (Tilley et al. 2011).

The erythrocyte cycle can continue essentially indefinitely, but it is ultimately a reproductive dead end: the merozoite must die with the man. To escape the human host and live on, a subset of blood schizonts commit their merozoite offspring to becoming *gametocytes*, sexually differentiated male and female parasite forms; all progeny of a schizont become either male, female, or asexual (the most typical fate). Sexually committed merozoites proceed as others, by invading an erythrocyte to become a feeding trophozoite, but then form either a single macrogametocyte (female) or single microgametocyte (male) (Baton and Ranford-Cartwright 2005). Committing to gametocytogenesis is risky, for terminally differentiated gametocytes cannot reproduce further in the blood of man. Gametocytogenesis in *Plasmodia* does not occur after a fixed number of erythrocyte cycles, as it does in some related parasites, and the decision to commit to gametocytogenesis remains poorly understood (Baton and Ranford-Cartwright 2005).

When taken up in a blood meal, the gametocytes rapidly initiate the sporogonic cycle. Upon arrival at the mosquito midgut the macrogametocyte dissolves its erythrocyte host within minutes and becomes spherical and immotile. The microgametocyte, on the other hand, undergoes the dramatic process of exflagellation, whereby eight daughter genomes are produced that attach to long writhing flagella, and break free to become highly motile wrigglers that find and fuse with a macrogametocyte to form the zygote (Baton and Ranford-Cartwright 2005; Antinori et al. 2012). The zygote in turn transforms into a banana-shaped ookinete (Greek "moving egg"), which penetrates both through the peritrophic matrix, a chitinous matrix extruded by the mosquito gut to sequester the blood meal, and then through the epithelial cells lining the gut wall. Next, it transforms into an oocyst, producing thousands of daughter sporozoites by nuclear division (Antinori et al. 2012). Eventually, the mature oocyst ruptures into the mosquito's hemocoelic cavity (Baton and Ranford-Cartwright 2005), and sporozoites travel through the hemolymph to infect the salivary glands where, after about one day, they are reprogrammed to be highly infectious to humans (Antinori et al. 2012), and the cycle can begin again. The cycle is depicted in its entirety in Fig. 2.

The complex within-host dynamics of human *Plasmodium* infection, how these affect the efficacy of treatment and control measures, and their interaction with the immune response, have been the focus of multiple modeling works, e.g. (Teboh-Ewungkem et al. 2010; Li et al. 2011; Gurarie et al. 2012; Eckhoff 2012; Demasse and Ducrot 2013; Childs and Buckee 2015; Childs and Prosper 2017; Tabo et al. 2017). However, to our knowledge no climate-focused models have focused deeply upon these within-host dynamics, although it is likely that such work is needed to fully elucidate how climate change might affect malaria epidemiology and control efforts in the future (see also Sect. 7.3).

2.2 Vector characteristics and lifecycle

Malaria is transmitted by adult female *Anopheles* mosquitoes, yet of the more than 450 known anopheline species, only about 60 can serve as actual vectors (Cohuet et al.

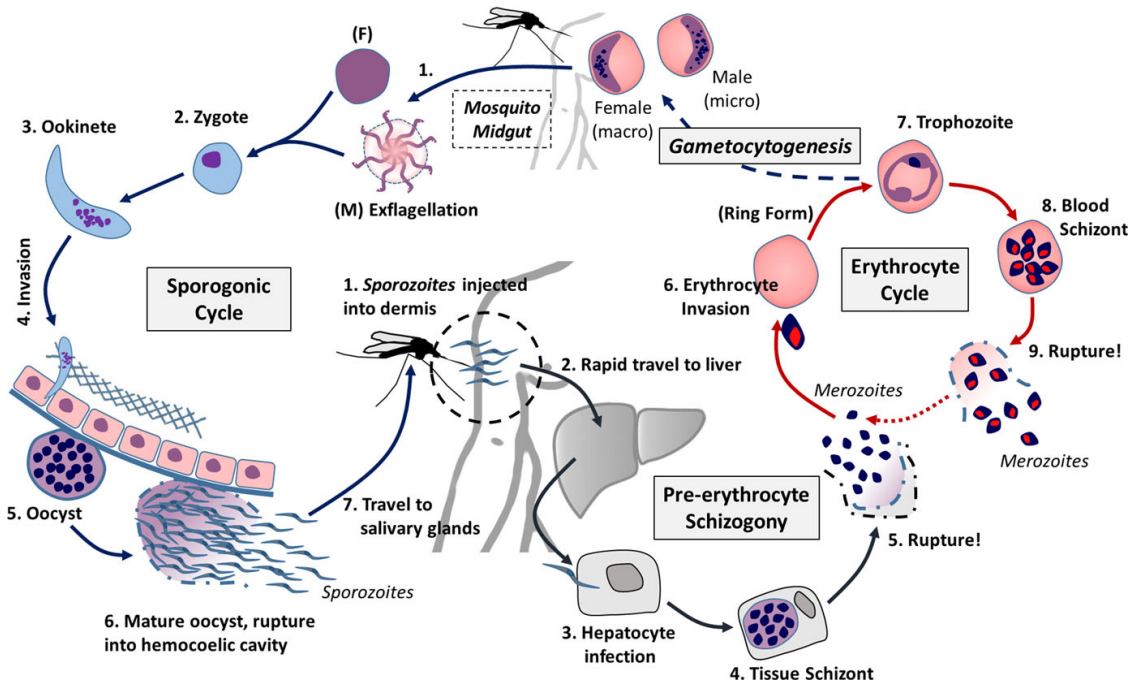


Fig. 2 The *Plasmodium* lifecycle. The right side depicts schizogony in man, where sporozoites from an infectious bite invade hepatocytes in the liver, undergo a round of replication, and then enter, as merozoites, into the erythrocyte cycle in blood. A minority of blood trophozoites differentiate to male and female gametocytes that are taken up by a biting mosquito to initiate sporogony, as depicted on the left side, whereby ookinetes penetrate the peritrophic matrix and gut epithelium to form oocysts, eventually rupturing to yield sporozoites that travel to the salivary glands. Further details are given in the text

2010), with 41 considered major vectors (Sinka et al. 2010), and most of these are rather inefficient at transmitting the disease. To effectively transmit disease, the mosquito must be susceptible to infection (many are completely refractory to *Plasmodium*), must habitually bite man (many mosquitoes strongly prefer other animals), and must live long enough for the sporogenic cycle to reach completion (Cohuet et al. 2010).

In Africa, three anopheline species are preeminent, namely *A. gambiae*, *A. arabiensis*, and *A. funestus*, with *A. gambiae* likely the single most important species (Sinka et al. 2010), and the focus of most modeling studies. A point of terminology to avoid confusion in the literature is in order here: the *A. gambiae* complex is a collection of seven morphologically indistinguishable species later recognized to be distinct, and includes *A. gambiae sensu stricto* (Latin "in the strict sense") which is the species referred to by the unqualified term *A. gambiae*, and *A. arabiensis* (Sinka et al. 2010). *A. gambiae sensu lato* (Latin "in the general sense") refers to the species complex. The existence of multiple distinct species, including the important vector *A. arabiensis*, within the *A. gambiae* complex clearly complicates matters, from both a modeling and malaria control standpoint, and these vectors vary, for example, in their susceptibility to insecticide-treated bednets (Kitau et al. 2012).

Briefly, the lifecycle of the *Anopheles* mosquito, while simpler than that of its *Plasmodium* parasite, is not trivial, with mosquitoes passing through three immature, aquatic stages (egg, larva, pupa), and a final adult stage. The adult female mosquito lifecycle is centered around the *gonotrophic cycle*: the taking of a blood meal to fuel egg development, which takes several days and is highly temperature dependent

(with higher temperatures decreasing the time for larval development), followed by oviposition of eggs in a suitable aquatic habitat, only to repeat until inevitable death (Detinova 1962). Typical blood meal size is 2–3 μL, and *A. gambaie* may lay anywhere between about 10 and 150 eggs per gonotrophic cycle (this is the "fecundity") (Takken et al. 1998; Afrane et al. 2005), but more typically between about 40 and 85 under field conditions (Afrane et al. 2005). Most eggs hatch within 2–3 days (Yaro et al. 2006), but time to hatching is modestly temperature dependent (Bayoh and Lindsay 2003). Eggs hatch to become larvae and actively feed upon algae and bacteria, growing through four moltings, and are thus divided into four stages known as instars (conceptually, first- and second-instars are lumped as "early," with third- and fourth- "late-instars"); *Anopheles* larvae also lie parallel to the water surface to obtain oxygen. Finally, fourth instar larvae become nonfeeding pupae that undergo metamorphosis to adult mosquitoes. Immature stage development rate and survival are both strongly temperature-dependent (Bayoh and Lindsay 2003), and the complete lifecycle is given in schema in Fig. 3.

Anophelines have varying habitat preferences, and are widely adapted to different environmental niches (Sinka et al. 2010), but the *A. gambiae* complex tends, unsurprisingly, to prefer conditions associated with anthropogenic alteration of the environment. Specifically, *A. gambiae* and *A. arabiensis* larvae prefer small, temporary, sunlit pools, with little vegetation (Minakawa et al. 1999, 2004), the kind created by deforestation, construction, and livestock, e.g. hoofprints. These pools are warmer, support more algae (the major larval food source), and have fewer predators than natural water bodies (Minakawa et al. 1999). A series of studies by Afrane and colleagues (Afrane et al. 2005, 2007, 2008) confirm that deforestation in Kenyan highlands creates habitat that strongly supports *A. gambiae* proliferation. *A. funestus*, the other major African vector,

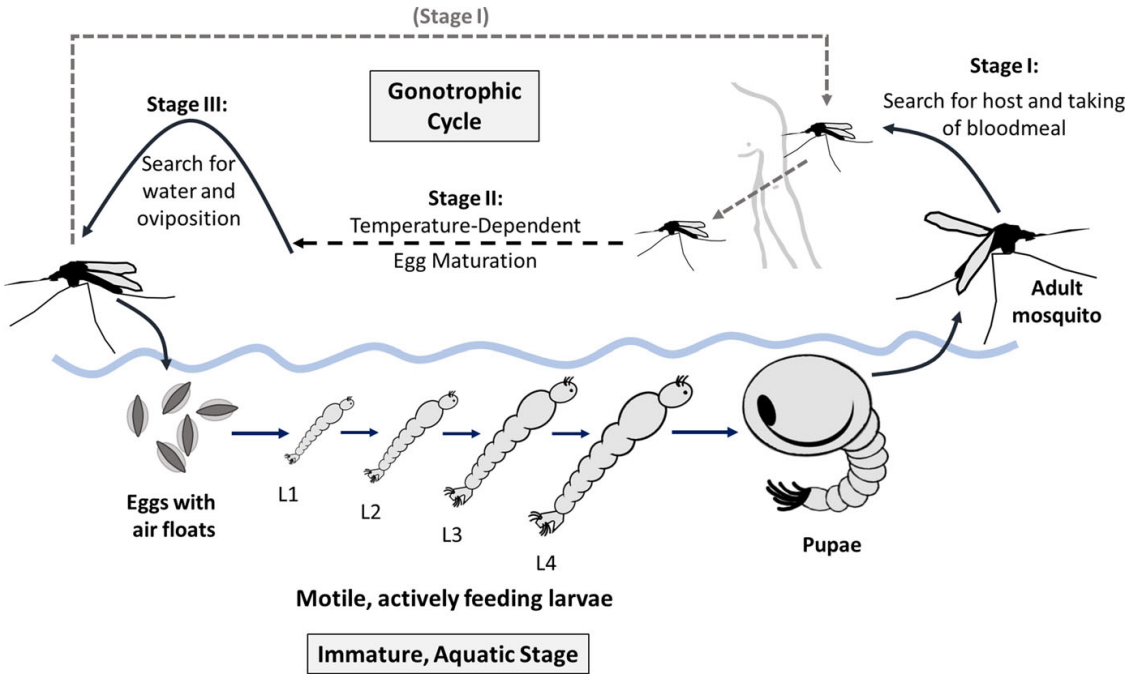


Fig. 3 *Anopheles* mosquito lifecycle. Immature mosquitoes pass through aquatic egg, larvae, and pupae stages, with the actively feeding larvae divided into four instar stages. Adult female mosquitoes pass through the gonotrophic cycle, by which bloodmeals nourish the development of new eggs, with further details in the text

is also aided by deforestation, but tends to prefer larger permanent or semipermanent habitats with established vegetation (Minakawa et al. 2005).

Anophelines can further be characterized along several axes that affect transmission potential and the efficacy of various control efforts (Sinka et al. 2010): (1) anthropophilia versus zoophilia, or the preference for taking blood meals from humans or non-human animals, respectively, with the anthropophilic index defined as the fraction of blood meals taken from man, (2) endophagic versus exophagic, referring to a preference for feeding indoors or out-of-doors, respectively, and (3) endophilic versus exophilic, meaning the favored location for resting between blood meals (this may differ pre-feeding and post-feeding). The highly efficient malaria vectors, such as *A. gambiae*, tend to be highly anthropophilic with anthropophilic indices approaching one, but as reviewed by Sinka et al. (2010), even these vectors are likely very opportunistic, and apparent anthropophilia may simply be a (partial) result of host availability; preferences along the other axes may also have been overstated in past studies, and anophelines are quite adaptable in general (Sinka et al. 2010).

Briefly, *A. gambiae* feeds late at night, has typically been reported as endophagic and endophilic, but this likely varies, and Odiere et al. (2007), working in western Kenya, found no preference. The closely related *A. arabiensis* also feeds at night, and may show an exophagic and exophilic preference in comparison to *A. gambiae* (Sinka et al. 2010). Behaviorally, the adult *A. funestus* is extremely similar to *A. gambiae* (Sinka et al. 2010).

Finally, not only do the innate characteristics of certain anophelines favor malaria spread, but there is even some evidence that mosquito behavior may also be modulated by *Plasmodium* infection to enhance transmission, a notion termed the "manipulation hypothesis" by Cator et al. (2012). When carrying the infectious sporozoite parasite stage, various *Anopheles* may take more frequent bloodmeals with more probing attempts per meal, may be more likely to feed from multiple hosts, and bloodmeal volume may be smaller. In contrast, when burdened by the non-infectious oocyst stage, mosquitoes seem less attracted to hosts and less persistent in bloodmeal attempts, a behavioral response that could decrease pre-infectious mortality by avoiding risky biting attempts (Cator et al. 2012; Nguyen et al. 2017), and the overall effect of such manipulations on malaria transmission could be quite significant, as suggested by mathematical analysis by Cator et al. (2014). However, most evidence for such manipulation comes from lab studies using a variety of vector-host combinations (Cator et al. 2012), and it is also unclear if such behavioral changes represent specific parasitic manipulations or more generic responses to infection (Cator et al. 2013). Moreover, several recent studies using field isolates of *P. falciparum* and anthropophilic *Anopheles* found no evidence that infection altered host-seeking behavior (Vantaux et al. 2015; Nguyen et al. 2017).

2.3 Immunity and clinical disease

In areas of intense *P. falciparum* transmission, young children are exposed to hundreds of infectious bites per year, and yet, unlike many viral diseases where a single exposure can be sufficient to imbue robust, lifelong immunity, immunity to malaria is

gained only slowly and incompletely over the course of years (Crompton et al. 2014). Characteristically, children under the age of five are susceptible to the most severe, life-threatening forms of the disease, such as severe malarial anemia and cerebral malaria, the disease transitions to an uncomplicated febrile disease through adolescence, and by adulthood it only rarely manifests clinically, with asymptomatic disease common (Crompton et al. 2014) (and a possible *Plasmodium* reservoir complicating eradication efforts). This hard-won immunity is short-lived: when adults from endemic areas move, they become vulnerable to severe disease within a few years, although they may retain protection against the worst manifestations of disease (Filipe et al. 2007). This dynamic is especially salient to malaria elimination efforts, and there is very real danger when the disease is eliminated locally but may be reintroduced to now non-immune populations (Webb 2014; Snow 2015), and even control measures, such as bednets or intermittent preventive (drug) therapy, while initially beneficial, have the potential to increase disease burden later in time, as they induce a decrease in population-level immunity (Ghani et al. 2009).

Thus, it is clear that there is a distinct disparity between clinical immunity against *P. falciparum* malaria (protection against clinical disease and severe symptoms) and infectious immunity (protection against infection, *per se*, by blood-stage parasites). Immunity to the most severe forms of disease may be also differ fundamentally from immunity to uncomplicated disease, with perhaps only several infections (and possibly as few as one) needed to confer long-lasting protection (Gupta et al. 1999a, b). The pathogenesis of clinical disease is primarily related to (1) sequestration of parasitized erythrocytes in organs such as the brain (this sequestration allows parasites to avoid the spleen, where they could be destroyed by macrophages), and (2) the systemic inflammatory response (Crompton et al. 2014). With respect to the former, the *P. falciparum* erythrocyte membrane protein-1 (*Pf*EMP1), expressed on the surface of infected cells, facilitates sequestration. It is also encoded on the *var* gene, of which there are about 60 distinct versions, each clonally expressed and encoding an antigenically distinct *Pf*EMP1. This antigenic variation, and the extreme genetic diversity of *P. falciparum* in general, help to explain why effective immunity requires so many exposures (Crompton et al. 2014).

It is worth noting that all actively clinical disease takes place during the erythrocyte stage of infection, with the skin and hepatocyte stages clinically silent. This may be at least partly related to the very different orders of magnitude involved at the different stages. Generally, fewer than 100 sporozoites infect the skin, and only tens of hepatocytes are infected. These numbers may simply be too small to initiate immunity, or, they may even induce immune tolerance, especially in the skin (Guilbride et al. 2012). In severe infections, on the other hand, total body trophozoite burden may number in the hundreds of billions (Trape et al. 1994).

2.4 Epidemiologic classification

P. falciparum transmission intensity in endemic zones varies across orders of magnitude, from one infectious bite per person per year, to more than one per day in many holoendemic areas (Rodriguez-Barraquer et al. 2016), and partly as a consequence of

its unique immunology, different malaria transmission intensities give differing age-distributions of parasitemia, clinical disease, and serious disease (Aron 1988; Snow 2015). It must be emphasized that *end*emic and *epi*demic malaria are very different beasts (Snow 2015; Hay et al. 2008): endemic (from Greek meaning "in the people") disease is constantly present in a population, whereas an epidemic (Greek "upon the people") is a temporary disease flare out of proportion to the past. Populations living with endemic malaria have varying degrees of immunity, but epidemic malaria can be calamitous when it tears through previously unexposed groups, or more perniciously, groups transiently protected by malaria control programs that lapse, leaving the people newly vulnerable after the waning of prior immunity (Snow 2015; Webb 2014).

The most common classification for endemicity is now based upon the fraction of the population that has parasites detectable in their peripheral blood, the so-called "parasite rate" (PR), and furthermore uses the parasite rate in the 2–10 year age group, $Pf\mathrm{PR}_{2-10}$, with zones classified as *holoendemic* ($Pf\mathrm{PR}_{2-10} > 75\%$), *hyperendemic* ($Pf\mathrm{PR}_{2-10}$ 50–75%), *mesoendemic* ($Pf\mathrm{PR}_{10}$ 10–50%), and *hypoendemic* ($Pf\mathrm{PR}_{10}$ 1–10%) (Snow 2015). In the hypoendemic and extreme hypoendemic ($< 1\%$) range, transmission becomes unstable (Snow 2015), and populations with very low burdens of malaria are vulnerable to epidemics of severe disease. Indeed, the venerable Macdonald considered the stable/unstable classification axis to be the more legitimate on a fundamental basis (Snow 2015).

The age-distribution of clinical disease varies across endemic zones. In holoendemic zones, most severe disease occurs in the first few years of life, rapidly tapering off by adolescence (Aron 1988; Gupta and Day 1994; Filipe et al. 2007; Crompton et al. 2014; Snow 2015), with the burden of severe disease dropping in absolute terms and shifting towards older age groups as the level of endemicity decreases, as demonstrated in Fig. 4. In holoendemic areas, the PR peaks later than does clinical disease (Trape et al. 1994; Rodriguez-Barraquer et al. 2016), and remains relatively high even into middle and old age, when clinical disease is rare. However, although the PR remains high, the parasite burden continues to decline with age (Trape et al. 1994), as also shown in Fig. 4. These observations have motivated many mathematical models attempting to elucidate the dynamics of immunity acquisition.

3 General historical background

3.1 Overview

The history of malaria, and its emergence as a major human pathogen over the last several 10,000 years, is intimately linked to the evolution of human agricultural civilization and the profound changes in both human populations and the environment that this engendered. This was directly coupled to global climate, as climate change following the end of the last ice age and the onset of the holocene era was fundamental to agriculture, and also allowed the wider spread of mosquito vectors in a warmer world (Carter and Mendis 2002). For the interested reader, scholarly histories of the disease include those by Webb (2014) and Packard (2007).

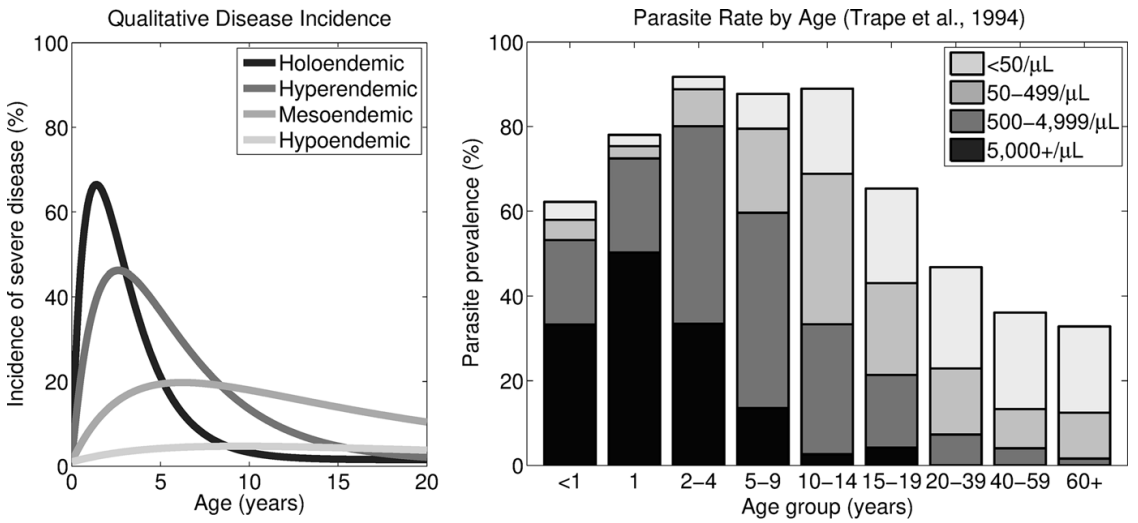


Fig. 4 The left panel gives the *qualitative* shape of severe disease incidence by age through adolescence under holo-, hyper-, meso-, and hypoendemic transmission conditions, based on Aron (1988), Snow (2015). The right panel shows overall parasite rate in the holoendemic village of Dielmo, Senegal (Trape et al. 1994), subdivided by the actual density of trophozoites in the peripheral blood. While PR in the youngest is only about twice that of those over 40, children under four suffered clinical malaria attacks at a rate 40-fold higher (Trape et al. 1994). Note that the broad plateau in PR from roughly age two to 15 in the face of dramatically falling serious disease incidence has been observed elsewhere (Gupta and Day 1994)

The clearing of forests for agriculture creates myriad microenvironments for anopheline mosquitoes, and concentrated human settlements are capable of supporting the virulent *P. falciparum*, which only emerged within the last 10,000 years, while a warmer climate supports its mosquito vectors (Webb 2014; Packard 2007). Malaria has also helped shape human biological evolution: in pre-agricultural Africa Duffy antigen (an erythrocyte membrane chemokine) negativity spread through the African heart to confer complete resistance to *P. vivax*, at no apparent cost, while the more recent advent of *P. falciparum* selected for a variety of far less benign genetic anemias, the best known being the sickle cell trait, which protects against severe disease in the heterozygous form, but *causes* crippling sickle cell disease in homozygotes (Carter and Mendis 2002). Following its earlier evolution in Africa, malaria, especially *P. vivax*, escaped that continent and into much of the rest of the world, its spread strongly associated with agricultural expansion and population movements (Packard 2007).

It was not until the end of the nineteenth century that the microbiological basis of the disease was discerned. This coincided with the onset of the colonial era, or "Scramble for Africa" spanning roughly 1879 through 1914, and during which various European powers conquered and carved up the African continent (Webb 2014). Thus, early "scientific" malaria control efforts in Africa were inescapably linked to colonial medicine, a primary focus of which was protecting Europeans and preserving the productivity of subservient African laborers, with less regard for the general African populace (Webb 2014). Lasting from 1955–1969, the World Health Organization's Malaria Eradication Programme saw significant mixed successes, but ultimately failed to eliminate the disease. In Africa, widespread chloroquine treatment during the 1970s was a primary cause of lowering malaria burden, but the spread of chloroquine resistance across the continent in subsequent years, the HIV/AIDS epidemic, agricul-

tural expansion, and devastating wars among many newly independent African states fueled a malaria resurgence. New global efforts since about 2000, largely centered on insecticide-treated bednets (ITNs) and treatment with the new artimisinin compounds have seen significant success (Bhatt et al. 2015), but it remains to be seen whether these gains will continue or even be maintained (Webb 2014).

In the following sections, we discuss more extensively the early origins of malaria, its link to agriculture and human activity, and then review in greater depth the era since the late nineteenth century. Our focus in the latter is on tropical Africa and *P. falciparum*, and a working theme is that *P. falciparum* differs qualitatively from the other human *Plasmodia*, representing a unique burden on African populations.

3.2 Origins and evolution

The *Plasmodia* are ancient parasites belonging to the order haemosporidia—single-celled parasites which alternate between a wide variety of vertebrate hosts and blood-sucking arthropods—and mammal-specific *Plasmodia* have coexisted with mammals for much if not all their evolutionary history, with one recent estimate dating their origin between 64 and 120 million years ago (Silva et al. 2015). Haemosporidia burdened animals even earlier, likely since almost the first appearance of Diptera insects (flies and mosquitoes) 150–200 million years ago (Carter and Mendis 2002).

Early studies found *P. falciparum* and a very closely related chimpanzee *Plasmodium*, *P. reichenowi*, to differ substantially in morphology and lifecycle from *P. malariae*, *P. ovale*, and *P. vivax*, and hence the former were categorized as a separate subgenus, *Laverania* (Loy et al. 2017). Later molecular studies confirmed that the *Laverania* diverged from the other mammalian *Plasmodia* on the order of 100 million years ago (Carter and Mendis 2002; Silva et al. 2015).

Moving forward in time, the evolutionary origins of modern human *Plasmodia* within the last 100,000 years, mainly *P. vivax* and *P. falciparum*, have been of some controversy, but the weight of the evidence supports, in our view, an out-of-Africa origin for all modern human malaria (see Loy et al. (2017) for a recent review). Under pre-agricultural conditions, scattered mobile populations of low density were unlikely to support intense transmission rates, and the overall malaria burden was probably low. Under these poor transmission conditions, *P. malariae*, which can cause a chronic low-grade infection lasting decades, and *P. vivax* and *P. ovale*, both of which have a dormant liver stage that can lead to reinfection and transmission years after initial infection, are much more competitive than the highly virulent and short-lived *P. falciparum* (Carter and Mendis 2002).

A powerful piece of circumstantial evidence supports the existence of relatively longstanding *P. vivax* infection in pre-agricultural Africa: Duffy antigen negativity. The Duffy antigen is a chemokine expressed on RBC membranes, and also happens to be an essential receptor for *P. vivax* merozoite entry into RBCs (Carter and Mendis 2002). Homozygotes for Duffy negativity are thus completely immune to *P. vivax*, and moreover appear to suffer no ill health-effects. In native populations, Duffy negativity prevalence is almost 100% in most west and central Africa (Culleton and Carter 2012), likely the ancestral seat of malaria and the areas of the most intense malaria

transmission today, while Duffy negativity is highly prevalent throughout the rest of the continent. Since homozygosity is required for significant benefit, one may that expect tens of thousands of years are necessary for Duffy negativity to become fixed in a population (Carter and Mendis 2002), and Hamblin and Rienzo (2000) estimated a selective sweep may have occurred 33,000 years ago (95% CI 65,000–97,200 years ago). Thanks to Duffy negativity, *P. vivax* was likely driven nearly to extinction in Africa, but escaped into Asia and the larger world (Liu et al. 2014), perhaps around 10,000 years ago, where populations have had insufficient time to evolve Duffy negativity (Culleton and Carter 2012).

About 10,000 years ago, African proto-agriculture led to more sedentary, larger human settlements that could sustain more virulent, short-lived infections (Carter and Mendis 2002). It was around this time that *P. falciparum* in gorillas may have crossed over into humans, according to a recent hypothesis (Loy et al. 2017). Even if the gorilla hypothesis is false, it is clear that *P. falciparum* did not become a significant human disease until between 5000 and 10,000 years ago, and that its rise was related to that of agriculture (Carter and Mendis 2002; Webb 2014). Malaria, both *P. vivax* and *P. falciparum*, likely spread through most of the inhabited world during early historical times (i.e. before the common era), although *P. vivax* mainly affected the more northerly regions, given its dormant phase and better cold tolerance versus *P. falciparum* (Packard 2007); malaria was rapidly introduced to the New World following its discovery by Europeans.

In the nineteenth century, malaria reached its global zenith, with most of the globe's population at risk (Carter and Mendis 2002), but then declined into extinction in most of Europe and the Americas by the mid-twentieth century, its retreat primarily caused by agricultural modernization and changing living conditions that discouraged transmission, and aided by later eradication programs (Packard 2007; Carter and Mendis 2002). This, however, was not the experience of tropical Africa, and from here out we will restrict our attention to this continent.

3.3 The colonial era, Africa, and modern malariology

> With tears and toiling breath I find thy cunning seeds, O million-murdering Death. I know this little thing a myriad men will save.
>
> Ronald Ross, fragment from *In Exile, Reply - What Ails the Solitude*

In the late 1800s, spurred largely by the discoveries of Koch and Pasteur, the search was on for bacterial causes of many diseases, and in 1880, Charles Laveran, an obscure French army officer stationed in Algeria (a French colony at the time, having been subdued in a bloody war of conquest spanning 1830–1847, and that killed as much as a third of the native population (Kiernan 2007)), observed a variety of strange writhing forms within the erythrocytes of malaria victims, which he would come to identify as a protozoan parasite that he named *Oscillaria malariae*. It was the first protozoan

discovered to infect man, and Laveran would receive the Nobel Prize in 1907 for this discovery (Cox 2010).

It fell chiefly to Ronald Ross, a British physician, to elucidate the vector by which the malarious protozoan was transmitted, the female anopheline mosquito, which he showed in birds in 1897, and in humans in Freetown, Sierra Leone, in 1899 (Cox 2010). Note that while to modern ears, the idea of a mosquito transmitting a disease is entirely natural, at that time it was a truly novel notion (Cox 2010). Sierra Leone had been established as a British colony in 1787, and due to the high malaria burden came to be known as the "White Man's Grave" (Bockarie et al. 1999), and indeed, malaria has been credited by some historians as protecting the interior of Africa from the depredations of European colonialism during the slave era (Webb 2014). The discovery of infected anopheline vectors *A. gambiae* and *A. funestus*, along with their breeding sites in myriad small pools, by Ross and his colleagues during their 1899 expedition led to vector control measures including bednets, window screens, and larval control via oiling of pools (Bockarie et al. 1999; Webb 2014).

Ross also recommended segregating European and African populations to protect the Europeans (Bockarie et al. 1999); this too would be a feature, although varying in degree by time and place, of colonial malaria control efforts (Webb 2014).

In Sierra Leone and elsewhere, subsequent efforts included draining or oiling pools, and removing household receptacles that could support breeding. Other anti-larval efforts, of which Ross was a champion, included treating pools with a highly toxic copper-based compound known as Paris Green, stocking with larvivorous fish, and, by World War II, treating with oils containing the pesticide DDT (dichloro-diphenyl-trichloroethane) (Bockarie et al. 1999; Webb 2014). While sometimes effective, anti-larval measures required ongoing action, were labor-intensive, and dependent upon funding. A common pattern was concentrating malaria control efforts in urban areas, and in commercial areas where European interests desired a healthy indigenous workforce, but with European health as a priority. There was also legitimate concern that measures decreasing malaria prevalence among native populations could reduce immunity, rendering them vulnerable to epidemic malaria (Webb 2014).

DDT, first used against malaria by the US Army in World War II, has a long-lasting residual effect, such that a dwelling need be sprayed only infrequently to have a toxic effect on mosquitoes. Thus, the 1940s and 50s ushered in the pesticide era, with indoor residual spraying (IRS) increasingly used by national control programmes (Nájera et al. 2011). Macdonald's mathematical model (Macdonald 1957) (discussed in Sect. 4.2) provided a powerful theoretical basis for increasing adult mortality as the linchpin of control (Nájera et al. 2011), and furthermore, the strategy of spraying was viewed as general and inexpensive, compared to expensive quinine treatment or labor- and capital-intensive environmental engineering and larvaciding (Webb 2014). Against this background, the WHO launched its Global Malaria Eradication Programme (GMEP) in 1955, based on spraying with DDT and related compounds supplemented by mass drug treatment, and efforts were geared toward *eradication* over *control*, with malaria control viewed with contempt by the program's architects (Nájera et al. 2011). The GMEP coordinated with national control programmes, and launched a number of pilot projects, the most famous and well-done being the Garki Project, which motivated the Garki mathematical model, by Dietz et al. (1974) (Sect. 4.4).

A deeply unfortunate effect of the GMEP was the undermining of the specialized field of malariology (why study malaria if all one need do is spray DDT, regardless of the particular vector, geography, socially, biology, etc.?) (Nájera et al. 2011; Webb 2014), as well as the (temporary) abandonment of many control measures other than IRS (Nájera et al. 2011). Despite very meaningful successes, including the eradication of malaria from many countries (especially outside of Africa), the GMEP was beset by setbacks, and in 1969 it was acknowledged that eradication was not a realistic short-term goal in many regions, marking the end of the programme (Nájera et al. 2011).

Beginning in the 1960s, the synthetic anti-malarial chloroquine became widely and inexpensively available on the African continent, with dramatically positive health consequences, being chiefly responsible for marked reductions in child and malaria-specific mortality through the 1960s and 70s (Carter and Mendis 2002; Webb 2014). However, malaria began a resurgence throughout Africa beginning in the late 1970s and 1980s, largely attributable to the evolution and spread of chloroquine-resistant *P. falciparum* (Carter and Mendis 2002). However, financial- and debt-crises borne by the recently independent African states, reductions in public health spending, widespread and large-scale political violence and chaos, the HIV/AIDS epidemic, and the widespread expansion of rural agriculture into west and central African rainforests, where deforestation created new habitat for anopheline vectors, all played roles (Webb 2014). It was also revealed in this era that prior concerns that "protecting" populations in endemic zones where elimination was infeasible could dangerously undermine immunity were well-founded, as deadly epidemics swept through many such regions, most notably in the highlands of Madagascar in 1986 (Carter and Mendis 2002; Webb 2014).

In the face of devastating infectious disease across the Global South, the WHO and several other organizations founded the Roll Back Malaria Partnership in 1998, while the WHO's "Global Fund to Fight AIDS, Tuberculosis and Malaria" was established in 2002, and in 2007, the Bill and Melinda Gates Foundation announced a campaign to eradicate malaria (Webb 2014). New tools became available, mainly insecticide-treated bednets (ITNs), and the burden of chloroquine resistance was relieved with newer artemisinin-based combination therapy. These campaigns have enjoyed apparent success, with a 57% decrease in African malaria mortality (per 10,000) from 2000 to 2015 (Gething et al. 2016); Bhatt et al. (2015) estimated that 68% of avoided malaria cases in Africa (from 2000 to 2015) could be attributed to ITNs.

Whether these gains will be maintained has yet to be seen. Malaria eradication and control programs have historically seen their greatest success in the first few years (Webb 2014), not all countries have experienced similar improvements under similar control programs (Snow et al. 2015), malaria incidence has recently increased locally in some areas, e.g. coastal Kenya (Snow et al. 2015), *Plasmodium* artemisin resistance has emerged in southeast Asia (Webb 2014), and perhaps even more worrisome, widespread pyrethroid resistance (the insecticide in ITNs) is evolving in vectors across Africa (Hemingway et al. 2016), although the impact of these developments in resistance is yet to be proven. Furthermore, a general dynamic of increased control of a childhood illness is an early drop in disease transmission, and a consequent shift in disease burden from younger to older ages that generates a rebound increase in

incidence after a few years (Griffin et al. 2016). In the context of malaria, an in-depth modeling study by Griffin et al. (2016) suggested that merely sustaining current control efforts, even absent new vector or parasite resistance, will lead to increases in malaria incidence and mortality by 2020. And finally, climate change continues in its insidious trajectory, with uncertain consequences.

4 Early mathematical models of malaria

4.1 Sir Ronald Ross, a pioneer of quantitative epidemiology

Ross proposed several mathematical models for malaria transmission in the early 1900s that were analyzed and modified by others, including Alfred J. Lotka (Smith et al. 2012). These were extended by Macdonald and other investigators in the 1950s, and this work, which focused heavily on $\mathcal{R}_0$ and mosquito eradication for malaria control, would prove to be very influential in guiding the ultimately failed GMEP (1955–1969). Smith and colleagues, who expertly reviewed the early history of Ross and Macdonald (Smith et al. 2012), have argued that there is no single or canonical "Ross–Macdonald" model, and that it is more instructive to understand this as a family of models characterized by a set of broadly shared assumptions and key entomological and epidemiologic parameters, whose estimation was historically motivated by quantitative models. Note, however, there is a clear "Macdonald" model, as presented in Macdonald (1957).

Ross's original 1908 model is of purely historical interest, so we will skip to the 1911 model (which was solved and extensively analyzed by Lotka) given as (Smith et al. 2012)

$$\frac{dX}{dt} = mabz(H - X) - rX, \tag{1}$$

$$\frac{dZ}{dt} = acx(M - Z) - gZ, \tag{2}$$

where H is the total human population with X the infected component, M and Z are similarly the total and infected mosquito populations, m is M/H (mosquitoes/man), a is the mosquito biting rate (bites/mosquito/day), z is the infectious mosquito fraction (Z/M), r is the human recovery rate (day^{-1}), b is the probability of human infection after an infectious bite (omitted and implicitly 1 in Ross's original model, we include it for clarity), c is the probability of a human infecting a mosquito upon biting, $x = X/H$ is the parasite rate, and g is the mosquito death rate (day^{-1}).

Sharpe and Lotka extended this model, as reviewed by Smith et al. (2012), to include latency between inoculation and infectivity in both man and mosquito, but because their model failed to consider mosquito mortality during latency, biological conclusions were flawed.

4.2 Ross–Macdonald

In a series of works in the early 1950s, Macdonald formulated a highly influential model, based on Ross's basic model. The major mathematical innovations introduced by Macdonald over Ross were accounting for the delay to infectiousness in mosquitoes, and superinfection, where multiple malarial strains can coinfect a host and are independently cleared, thus altering the recovery rate from the infected to recovered/susceptible category. Unfortunately, this concept was incorrectly translated into mathematics by Macdonald, apparently due to a miscommunication (Smith et al. 2012). The correct form was described by Walton in 1947 (per Dietz et al. (1974)), and was incorporated into the influential "Garki" model devised by Dietz et al. in 1974 (Dietz et al. 1974). We have, using $x(t)$ as the proportion of infected humans (the "parasite rate"), Macdonald's model (Macdonald 1957; MacDonald et al. 1968)

$$\frac{dx}{dt} = h(1-x) - \rho(r,h)x, \tag{3}$$

where h is the inoculation rate and r is the first-order rate of recovery from each infecting malarial strain, each of which is assumed to be cleared independently; the overall rate of recovery, $\rho(r,h)$, is a function of the inoculation rate and strain-specific recovery rate. Now, inoculation is given as

$$h = \frac{ma^2bcp^n x}{ax - \ln(p)} = \frac{ma^2bcx}{ax+g}\exp(-ng), \tag{4}$$

where n is the duration of the sporogonic cycle, p is the daily probability of survival, implying $p = \exp(-g)$ and that $\exp(-ng)$ is the fraction of mosquitoes surviving from the time of exposure to infectivity; other parameters are as in the Ross model (c was assumed to be unity by Macdonald (1957), but we have included it for generality). In his 1957 book (Macdonald 1957), Macdonald derives this expression as

$$h = mabcs, \tag{5}$$

where s is the sporozoite rate (i.e. the fraction of mosquitoes with sporozoites in their salivary glands) which in turn is derived as follows. We have, from the exponential distribution, that the expected (mean) lifetime of any mosquito is

$$\frac{1}{g} = \frac{1}{-\ln(p)}. \tag{6}$$

We then have that the total mosquito-days spent in a potentially infectious state, i.e. they have survived at least n days (duration of the sporogonic cycle), is

$$\frac{p^n}{-\ln(p)}. \tag{7}$$

To determine the sporozoite rate, we need know what fraction of such potentially infectious mosquitoes actually are infectious. We have that the average number of infectious feeds in a day is ax, and so the probability of taking no infectious feeds in a day is $\exp(-ax)$, and the chances of both surviving and taking no infectious feeds is $p\exp(-ax)$. It follows (again from the exponential distribution) that the expected life taking no infectious feeds is

$$\frac{1}{-\ln\left(p\exp(-ax)\right)} = \frac{1}{ax - \ln(p)}, \tag{8}$$

and the total mosquito-days spent in the potentially but non-infectious state is

$$\frac{p^n}{ax - \ln(p)}. \tag{9}$$

We finally arrive at the sporozoite rate (i.e. fraction of mosquitoes in an infectious state) as

$$s = \frac{(7) - (9)}{(6)} = \frac{p^n ax}{ax - \ln(p)}. \tag{10}$$

This expression can also be derived by applying a quasi-steady-state assumption to a delay-differential version of Ross's model,

$$\frac{dx(t)}{dt} = mabz(t)(1 - x(t)) - \rho(r, h)x(t), \tag{11}$$

$$\frac{dz(t)}{dt} = acx(t - n)(1 - z(t - n))\exp(-ng) - gz(t). \tag{12}$$

That is, setting $dz/dt = 0$ (assuming $x(t - n) = x(t)$) and solving for z. Moving on, the recovery rate, $\rho(r, h)$, takes the form

$$\rho(r, h) = \begin{cases} r - h, & h < r \\ 0, & h \geq r \end{cases}, \tag{13}$$

but this implies no recovery ever occurs when inoculation exceeds the *strain-specific* recovery rate (clearly an error). The correct form, given by Dietz et al. (1974), is

$$\rho(r, h) = \frac{h}{\exp\left(\frac{h}{r}\right) - 1}. \tag{14}$$

Finally, from Macdonald's model, we can derive the following expression for $\mathcal{R}_0$:

$$\mathcal{R}_0 = \frac{ma^2bcp^n}{-r\ln(p)} = \frac{ma^2bc\exp(-ng)}{rg}. \tag{15}$$

The key conclusion from this expression is that daily mosquito survival, p, appears in both the numerator and the denominator, such that decreasing it lowers $\mathcal{R}_0$ logarithmically (Macdonald's $\mathcal{R}_0$, as a function of several different parameters of the first

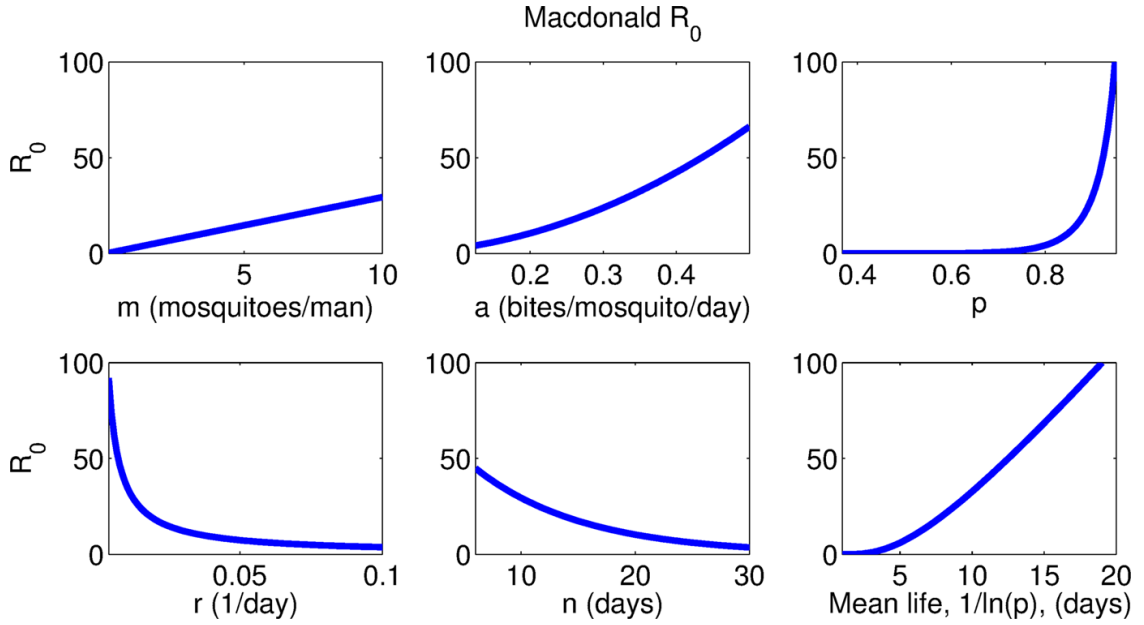


Fig. 5 Change in Macdonald's $\mathcal{R}_0$ as a function of each major parameter (except b and c, which have a straightforward linear effect). Daily mosquito survival, p, is the most sensitive parameter, and $\mathcal{R}_0$ is also given as a function of expected mosquito life, $1/\ln(p)$

equality in Eq. (15), is depicted in Fig. 5). This suggests targeting the adult mosquito vector as the most efficacious strategy for malaria control, and indeed, this was the basic approach of the GMEP, which relied principally upon indoor residual spraying with DDT (and other pesticides) to achieve this end, as discussed already in Sect. 3.3.

While $\mathcal{R}_0$ in Eq. (15) suggests targeting mosquito daily survival, p, over mosquito density, m, it is obvious from the basic ecology that these are *not* independent parameters. Nor, indeed, is a, the biting rate, since biting provides blood needed to nourish the mosquito's eggs. Both Ross and Macdonald were pioneering thinkers, but it seems clear that a more robust model framework that more fully accounts for the vector lifecycle is necessary for us to be confident in any conclusions. We shall explore some of these issues in detail in Sect. 6.

4.3 Vectorial capacity

In 1964, Garrett-Jones (1964) proposed an alternative metric, contra $\mathcal{R}_0$, for assessing and motivating vector control, namely the *vectorial capacity* (VC). It was defined qualitatively, for a vector population, as "the average number of inoculations, ..., originating from one case of malaria in unit time [typically in days], that the [vector] population would distribute to the human host if all female adult mosquitoes biting the human host became infected" (Garrett-Jones 1964). In other words, it is the number of new malaria cases (i.e. infectious bites, assuming all such bites result in infection) originating from a single case in a single day. Garrett-Jones (1964) formally defined it as the product of (1) the *man-biting rate* (total bites/person/day), which is the total number of mosquitoes infected from a single case in a single day, (2) the *expectation of infective life*, and (3) the *man-biting habit*, the number of bites on the human host

per day per individual mosquito. Using the Ross–Macdonald parameters, the vectorial capacity is given by Garrett-Jones and Shidrawi (1969)

$$\mathrm{VC} = \frac{ma^2cp^n}{-\ln(p)} = \frac{ma^2c\exp(-ng)}{g}. \tag{16}$$

Note that the original form implicitly assumed $c = 1$, but we have included it for generality. The vectorial capacity is very similar to the concept of $\mathcal{R}_0$, but is simply considering the number of new cases (assuming 100% infectiousness of bites) that result in the first unit of time from the original case, rather than over the lifetime of the first case. Where $\mathcal{R}_0$ has units of *cases*, VC has units *cases/day*, and the two terms relate (again, under Macdonald's model) as

$$\mathcal{R}_0 = \mathrm{VC} \times \underbrace{b}_{\text{probability that man is infected by a bite}} \times \underbrace{\frac{1}{r}}_{\text{Expected time that first case is infectious}}. \tag{17}$$

The concept of vectorial capacity was used in a number of subsequent quantitative studies, such as Garrett-Jones and Shidrawi (1969), Dietz et al. (1974), Molineaux et al. (1978), and it is also noteworthy that VC is the component of $\mathcal{R}_0$ that is most directly affected by weather parameters (Craig et al. 1999).

4.4 Developments post-Ross–Macdonald

The next major mathematical modeling contribution to malaria transmission dynamics was by Dietz et al. (1974) and entailed the inclusion, into the basic Ross–Macdonald framework, of a kind of slowly-acquired immunity that results in a non-infectious parasitemia following inoculation that is cleared relatively rapidly. Non-immune hosts are assumed to manifest infectious clinical disease that transitions to a non-infectious parasitemia that is cleared slowly. The model was fit to data for two villages in the Garki district of Nigeria, where data on the parasite and sporozoite rate had been collected by age. A follow-up work by Molineaux et al. (1978) in 1978 compared the model against several other datasets, where vectorial capacity was estimated from entomologic parameters and observed host-biting rates.

Many of the major modeling contributions following this work concern the proper or realistic modeling of immunity, especially the distinction between anti-disease (resistance against the harmful clinical manifestations of parasite infection, such as fever, anemia, etc.) and anti-parasite immunity (resistance against the actual *Plasmodium* infection), and how these are induced with infection and lost with time. Since climate, and not immunity, *per se*, is our primary focus, we defer a brief discussion of this model genealogy to Sect. 7.1. Moreover, most climate-focused models have only included fairly rudimentary descriptions of immunity, if it is addressed at all (but see (Yamana et al. 2013, 2017) for exceptions), and the hybridization of these two modeling traditions is a major future challenge.

5 Quantifying the relationships between weather and the parasite and vector lifecycles

Understanding the quantitative relationships between weather, primarily temperature and rainfall (and to a lesser degree, relative humidity), and the malaria parasite and vector lifecycles is critical to a realistic and meaningful assessment of the impact of current and projected climate change on malaria transmission dynamics. We review some of the most widely used quantitative relationships and the data they are based here, but it should be understood that these data are drawn from a variety of *Plasmodium* and *Anopheles* species, and the widely used formula of Moshkovsky for sporogonic and gonotrophic durations, for example, is based on data from the 1930s obtained in the European vector *A. maculipennis* (Detinova 1962). Moreover, there is some suggestion that thermal sensitivities may vary between laboratory and field strains of *P. falciparum*, which may be more adapted to local conditions, although Lyons et al. (2012) observed similar thermal tolerances between laboratory and wild strains *A. arabiensis* and *A. funestus*. While still poorly understood in general, short- and long-term adaptations in both vector and parasite to local temperature ranges and shifts could limit the ultimate generalizability of model inferences made assuming thermal response functions uniform throughout space and time (Sternberg and Thomas 2014).

Furthermore, while only a single thermal response function for a given process is typically considered in models, major vectors may differ importantly in how they respond to both mean and fluctuating temperatures, with Lyons et al. (2013) observing survival and development in the three major African vectors, *A. gambiae*, *A. arabiensis*, and *A. funestus* to vary both in response to mean temperature and temperature fluctuations, with *A. funestus* in particular much more sensitive to temperature fluctuations than *A. arabiensis*. Moreover, most models use thermal response functions drawn from multiple species, and thus how explicit consideration of the varying responses to weather between relevant vectors may affect model conclusions remains an open question.

5.1 Parasite

The sporogonic cycle of *Plasmodia*, i.e. infection and sexual reproduction in the mosquito midgut to ultimately yield infectious saliva sporozoites, is very clearly directly influenced by climate, with warmer temperatures (at least to a point), leading to more rapid parasite development (decreasing n, under the Ross–Macdonald framework), and this has been the focus of most mathematical works. We also note, however, that temperature may affect infectivity to both mosquito (c, per Ross–Macdonald) and man (b). Temperatures above about 30 °C may decrease *P. falciparum* survival in the mosquito midgut, and hence decrease c (Eling et al. 2001; Okech et al. 2004a), and Paaijmans et al. (2012) (in a rodent model) found increasing temperatures to decrease the prevalence of sporozoites in infected mosquitoes, and hence decrease b. However, these factors are less frequently accounted for in models, and we restrict further attention to the sporogonic cycle and n.

5.1.1 Sporogonic cycle and temperature

The classical formula of Moshkovsky It has long been recognized that the duration of the sporogonic (extrinsic) cycle in mosquito, denoted by n, is hyperbolically related to temperature. That is, given a constant D, measured in degree-days (the "sum of heat," as elaborated below), a minimum temperature, T_{min}, and mean ambient temperature, $T > T_{min}$ (in °C), we have

$$n = \frac{D}{T - T_{min}}. \tag{18}$$

This relation is based on a now obscure 1935 work in Russian by Nikolaev (1935), who, as related by Detinova (see pp. 122–150 of Detinova (1962)), gathered data on the duration of sporogony in the *A. maculipennis* mosquito, the historic vector of malaria in Europe and the Middle East (Djadid et al. 2007), and one that has recently been found to be expanding its range northeasterly into Eurasia, likely as a consequence of global warming (Novikov and Vaulin 2014). Nikolaev (1935) determined that *P. falciparum* sporogony ceases below 18 °C (16 °C for *P. vivax*), and his data was used by later authors (Moshkovsky in particular) to estimate the natural length of the sporogonic cycle.

Briefly, the Moshkovsky method (as given in Detinova (1962)) assumes the "sum of heat" hypothesis that a fixed amount of heat—which can be measured in degree-days calculated *beyond the minimum temperature required for sporogony to progress at all*—is required to complete the cycle. This "sum of effective temperatures" was estimated at 105 °C-days for *P. vivax*, 111 °C-days for *P. falciparum*, and 144 °C-days for *P. malariae*. Given daily mean diurnal temperature, the time to completion of sporogony can be straightforwardly calculated for any day of the year. The formula of Moshkovsky has been used in many more modern works (Molineaux et al. 1978; Craig et al. 1999; Hoshen and Morse 2004; Parham and Michael 2010). It should, however, be emphasized that, even in the early days, it was observed that temperatures above 32 °C were lethal to developing parasites, blocking sporogony (Macdonald 1957). Sporogonic duration as a function of constant temperature is given in Fig. 6.

Recent developments More recent studies are consistent with lower temperatures prolonging sporogony, but several laboratory works suggested (constant) temperatures of only 30 °C could significantly impede early *P. falciparum* sporogony in *A. stephensi* (a vector commonly found in the Indian subcontinent) (Noden et al. 1995; Eling et al. 2001). This finding's generality was challenged, however, by work by Okech and colleagues (Okech et al. 2004a, b) who used wild *P. falciparum* strains in western Kenya. They found that naturally fluctuating field temperatures, up to 33 °C maximum, did not interrupt parasite development in *A. gambiae* (Okech et al. 2004b), and under laboratory conditions constant exposure to 32 °C decreased, but did not eliminate, early *P. falciparum* survival in mosquito midgut and infectivity (Okech et al. 2004a). These results indicate that wild *P. falciparum* typically exposed to hotter conditions is more thermally adapted than lab-raised strains, and thus it seems likely that the Moshkovsky formula amended by sporogonic arrest around 32–34 °C is a reasonable description for wild *P. falciparum*, although mosquito infectivity may decrease at 30–32 °C (Okech et al. 2004a). Alternatively, several authors (Paaijmans et al. 2009; Mordecai et al.

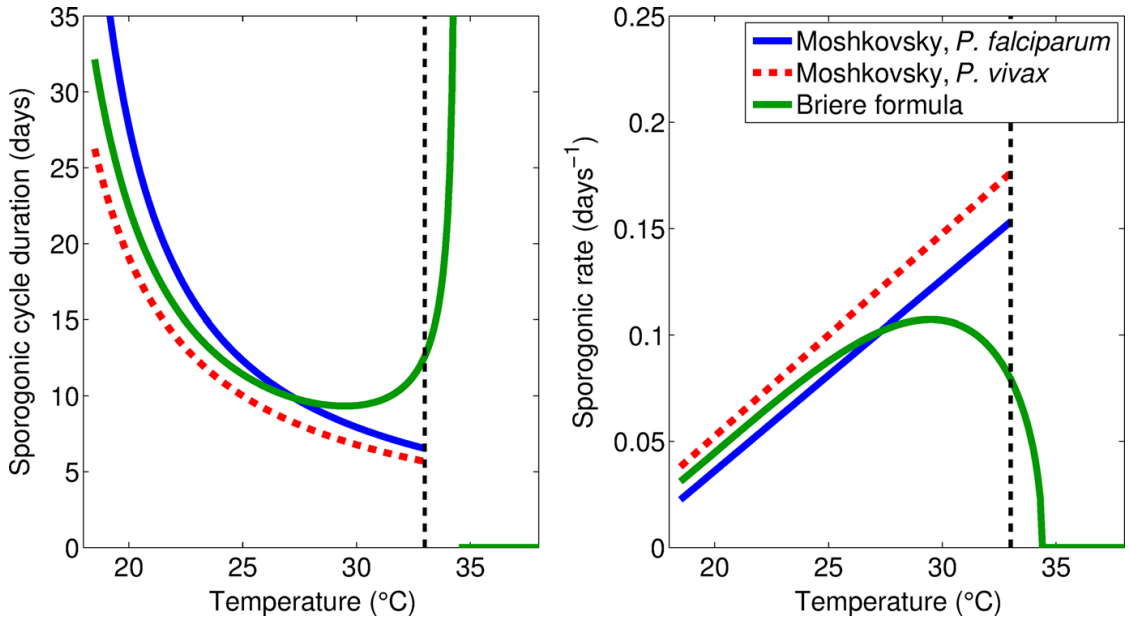


Fig. 6 Duration (left panel) and rate (right panel) of the sporogonic cycle as a function of constant temperature, per Moshkovsky's formula for *P. falciparum* ($D = 111$, $T_{min} = 16$) and *P. vivax* ($D = 105$, $T_{min} = 14.5$) (Detinova 1962), and assuming sporogony ceases above about 33 °C. Curves calculated from Briere's formula, using parameter values from Mordecai et al. (2013), are also given for comparison

2013) have more recently used a general unimodal functional form given by Briere et al. (1999),

$$r(T) = cT(T - T_0)(T_m - T)^{\frac{1}{2}}, \tag{19}$$

with Mordecai et al. (2013), for example, using c $= 0.000111$, $T_m = 34.4$ °C, and $T_0 = 14.7$ °C. Such a formulation gives developmental rates qualitatively similar to those under amended versions of Moshovsky's formula, as shown in Fig. 6.

5.2 Vector

5.2.1 Gonotrophic cycle, oviposition, and biting rates

Adult female *Anopheles* mosquitoes take blood meals from human hosts almost exclusively to provide energy and nutrients for their eggs, and the gonotrophic cycle (defined as the time between blood meals) is classically divided into the following three stages (Detinova 1962):

1. Search for host and attack.
2. Digestion of blood meal and egg maturation. This stage is highly temperature-dependent.
3. Search for body of water and oviposition. This stage is dependent upon the availability of suitable standing waters, itself a function of recent rainfall (except for those vectors that breed in more permanent bodies of water).

The terminology around gonotrophy is sometimes abused, with either stage II (blood meal digestion) or stages II and III (blood meal digestion and oviposition) together

sometimes conflated with the gonotrophic cycle as a whole. Gonotrophic cycle duration is important because, firstly, it partially determines mosquito population levels. Second, it has been traditionally assumed that mosquitoes take only a single blood meal between ovipositions (Scott and Takken 2012), implying that the duration of the gonotrophic cycle *is identical to the biting interval*. There is a biochemical basis for this assumption, with mosquito host-seeking behavior inhibited following a blood meal via two mechanisms that act in sequence (Takken et al. 2001). First, gastric distention in the period immediately following feeding activates inhibitory mechanoreceptors. Second, blood digestion and adequate nutritional status seem to stimulate neuropeptides that decrease the sensitivity of olfactory receptors in the antennae (Takken et al. 2001). Evolutionarily, avoiding unnecessary blood meals has obvious potential benefits, as biting requires energy to seek out hosts who, as we might expect, are then rather hostile to blood-sucking insects (Klowden and Briegel 1994), although as a nocturnal feeder biting somnolent hosts, *A. gambiae* may have faced less such selective pressure (Klowden and Briegel 1994).

Multiple blood feeding The assumption of one blood meal per gonotrophic cycle is clearly violated for at least a subset of *A. gambiae* mosquitoes, namely newly-emerged adult females with insufficient nutritional reserves to fuel a gonotrophic cycle from a single blood meal (Scott and Takken 2012). These mosquitoes, whose ovaries do not progress beyond Christophers' stage II (an early stage in ovary development) after their first blood meal are termed "pre-gravid," and typically take a second blood meal within 24 h to stimulate egg development (Scott and Takken 2012). Unsurprisingly, it is smaller mosquitoes that appear to be more nutritionally deficient, requiring two blood meals to complete their first gonotrophic cycle, with the first blood meal apparently devoted mainly to shoring up nutritional reserves, while larger individuals can successfully complete oogenesis with a single blood meal (Takken et al. 1998). After this early transient phase, surviving small and large adult mosquitoes may behave similarly, but with smaller mosquitoes now more likely to be infected by *Plasmodium* (Takken et al. 1998). Emerging adult mosquito size is dependent on larval conditions, with higher densities, higher temperatures (which increase the rate of larval development), and poor nutrition favoring smaller adult sizes (Smith et al. 2012). It should then be noted that temperature could have a subtle second-order effect on malaria transmission by increasing the proportion of smaller adult female mosquitoes more likely to engage in multiple biting behavior. Further, most "extra" bites would occur early in life before the completion of sporogony.

Whether multiple blood feeding is common among anophelines outside of this context is unclear. Klowden and Briegel (1994) observed laboratory *A. gambiae* to strongly seek human hosts every 24 h following a blood meal, corresponding to regular nocturnal feeding behavior and suggesting no inhibition of host-seeking. In direct contrast, however, in laboratory-raised *A. gambiae* of uniform size (thus avoiding contamination by smaller pre-gravid mosquitoes), Takken et al. (2001) observed blood meals to strongly inhibit host seeking up to 72 h following a blood meal (at 27 °C), with this interval corresponding to egg maturation or oviposition in most individuals. Despite its potential importance to malaria epidemiology, the pre-gravid anopheline is typically neglected in mathematical models, although Depinay et al. (2004) took the

probability of taking multiple bloodmeals to be a function of both mosquito weight and parity.

Quantifying the gonotrophic cycle Leaving the complications of possible multiple blood feedings aside, let us now review quantifications of the gonotrophic cycle duration, or (typically) equivalently, biting rate. The classic works of Macdonald (Macdonald 1956b, 1957) and other derived works (e.g., (Molineaux et al. 1978)) have generally assumed a constant biting interval of about two days for *A. gambiae* (Molineaux et al. (1978) chose two days on the basis of the distribution of abdominal stages in sprayed mosquitoes collected in the Garki Project), although some older field studies suggested three days as more likely (Takken et al. 2001). However, it has been well-understood since at least the 1950s that stage II varies greatly in inverse proportion to temperature (Macdonald 1957).

Duration of stage II Detinova (1962), in his influential work, presented data on the duration of stage II, which also followed the basic form of Moshkovsky's degree-day formula, but the constants varied with relative humidity (RH), such that this stage is quicker under more humid conditions; curves and formula constants are given in Fig. 7. More recent studies include (Ra et al. 2005; Afrane et al. 2005; Lardeux et al. 2008; Mala et al. 2014). Probably most notable is a 2008 laboratory study (Lardeux et al. 2008) on *A. pseudopunctipennis*, a major South American vector, who studied the duration of stage II/III (i.e. blood meal to oviposition) under a range of temperatures, and used the following general function, adopted from Lactin et al. (1995) (based in turn on Logan et al. (1976)) for development rate, $r(T)$, assumed to be the reciprocal of the mean time to oviposition (i.e., it is assumed that gonotrophic duration is exponentially distributed):

$$r(T) = \exp(\rho T) - \exp\left(\rho T_m - \frac{(T_m - T)}{\Delta}\right) + \lambda, \tag{20}$$

where T_m is the "thermal maximum," or lethal temperature, Δ is the temperature range over which the development rate begins to fall from a maximum to zero ("temperature boundary layer"), ρ determines the increase in development rate with temperature and, per Logan et al. (1976), can be "interpreted as a composite value for critical enzyme-catalysed biochemical reactions," and lastly, λ gives $r(T)$ when $T = T_m$. Development time reached its nadir at 31 °C, and there was no oviposition at 37 °C, as all mosquitoes suffered mortality before they could oviposit.

Now, a subtlety here is that a single development rate does not fully describe the data, and because mosquitoes lay their eggs only at night, the actual time to oviposition clusters at intervals of 24 h. For example, the overall mean time to oviposition at 35 °C was 2.3 days, but roughly 70% of mosquitoes oviposited on night two, and 30% on night three. The lower the temperature, the longer the mean time, and oviposition events are spread over more nights; this is shown in Fig. 7, where simulated cohorts of ovipositing mosquitoes are generated using parameters fit to Eq. (20) as reported in Lardeux et al. (2008). It is obvious from this data that time to oviposition is not

exponentially distributed, and the epidemiological implications could be addressed with some future model.

Duration of stages I and III Work by Detinova (1962) from 1953 on *A. maculipennis* in the USSR, in which the contraction of ovarioles (which begins following oviposition) was determined in dissected mosquitoes, suggested that a majority of mosquitoes take a blood meal within 8 h of oviposition, and over two-thirds of mosquitoes have taken a blood meal within 24 h of oviposition. With respect to the third stage, this author also determined that oviposition usually occurs within 24 h of egg maturation. Thus, it is reasonable to assume that the third and first phase take, together, around 24 h (but occasionally up to 48 h), while the second phase is highly temperature-dependent. Note, however, that low availability of oviposition sites (related, say, to low rainfall) could significantly prolong search time, as studied in a simple model by Gu et al. (2006).

Overall duration The overall duration of the gonotrophic cycle can be reasonably estimated as 24 h for stage I and III combined, plus a temperature dependent term for stage II, either Moshkovsky's formula, or a relation derived from other data, such as that of Lardeux et al. (2008). These two options (and using several relative humidities for Moshkovsky's formula) are compared in the right panel of Fig. 7.

Time to first bloodmeal We finally note that the there is a small delay, on the order of 1–3 days, between emergence from pupal stage to the first bloodmeal ("pre-bloodmeal period") and the start of the gonotrophic cycle proper (Paaijmans et al.

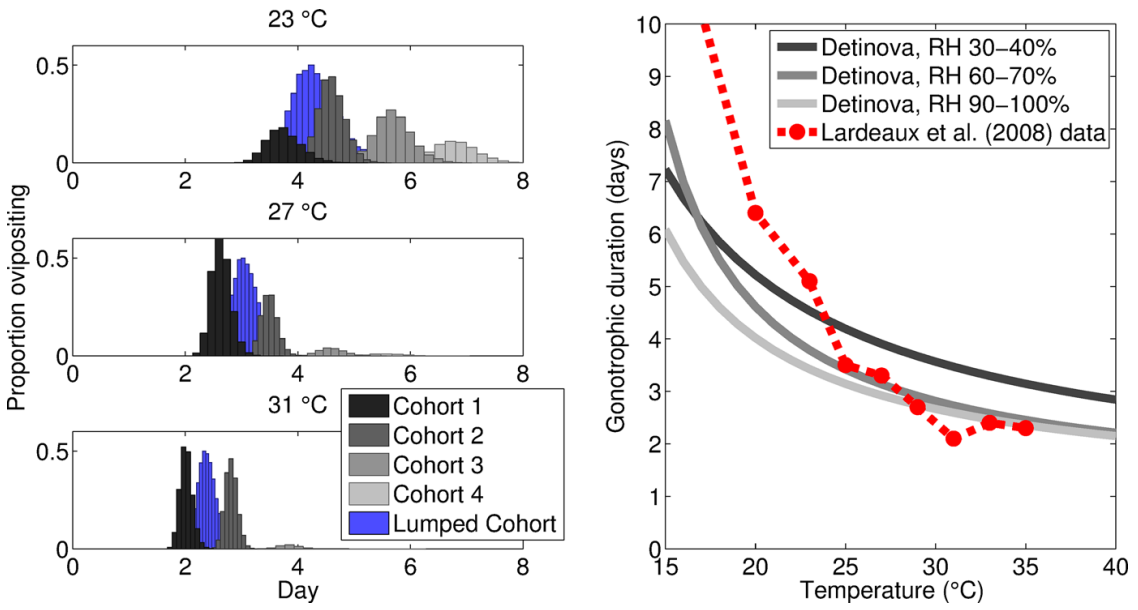


Fig. 7 The left panel shows simulated proportions of mosquitoes that oviposit, as a function of time, at different ambient temperatures based on the experiments of Lardeux et al. (2008). Oviposition is spread over several nights, with the number of nights increasing as mean time to oviposition increases. The blue histograms show simulated time to oviposition if all events occur on a single night (curve is scaled down for clarity). The right panel compares the duration of the gonotrophic cycle as measured by Lardeux et al. (2008), and as calculated by Moshkovsky's formula plus an additional day for stages I and III, and using, per Detinova (1962), $D = 65.4$, $T_{min} = 4.5$; $D = 36.5$, $T_{min} = 9.9$; and $D = 37.1$, $T_{min} = 7.7$, for relative humidities (RH) of 30–40%, 60–70%, and 90–100%, respectively

2013a). Furthermore, Paaijmans et al. (2013a) found this pre-bloodmeal period to be temperature-dependent in *An stephensi*, being about three days at 18 °C but just one day at either 26 or 32 °C.

Thus, we have the pre-bloodmeal period as an additional adult stage that is typically disregarded in models, but acts to delay time to first infection and infectivity in a temperature-dependent manner. This phenonemon may slightly narrow the temperature range over which malaria is effectively transmitted (Paaijmans et al. 2013a), and moreover, will shift infectivity to older mosquito age classes, which could in turn interact with age-dependent temperature-mediated mortality (see Sect. 5.2.2).

5.2.2 Daily mosquito survival

Age, *Plasmodium* infection, and temperature all influence mosquito survival. Classically, it has been assumed that random mortality events such as predation, etc. account for most mosquitoes deaths, such that senescence can be disregarded and death modeled as a Poisson process, i.e. survival times are exponentially distributed (this is also the implicit assumption of those mathematical models that employ first-order death kinetics in differential equations), and daily survival probability is typically assumed to be around 0.90–0.95 (Macdonald 1956a; MacDonald et al. 1968). Nevertheless, mosquito death hazard clearly increases with age (Clements and Paterson 1981; Okech et al. 2003; Dawes et al. 2009; Christiansen-Jucht et al. 2014), with this most convincingly demonstrated in laboratory studies (e.g., Bayoh 2001; Christiansen-Jucht et al. 2014), although Clements and Paterson (1981) found evidence of age-dependent mortality in wild populations as early as 1981, and recent analysis by Ryan et al. (2015a) also suggested senescence occurs in wild mosquitos, despite high extrinsic mortality rates. Survival is often described by the Gompertzian distribution (Clements and Paterson 1981; Bayoh 2001; Christiansen-Jucht et al. 2014; Ryan et al. 2015a), given as

$$S(t;\lambda,\theta) = \exp\left(\frac{\lambda}{\theta}\left(1-\exp(\theta t)\right)\right), \tag{21}$$

with the model essentially positing exponentially increasing mortality with age, although other mathematical descriptions (e.g., double exponential model, quadratic model) can also describe age-dependent mortality increases (Clements and Paterson 1981), and one especially useful description is the gamma distribution, which may be straightforwardly implemented in an ordinary differential equations (ODE) setting, as discussed further in Sect. 6.3.5.

Laboratory mortality is straightforward to measure; wild mosquito survival can be estimated by means of mark-release-recapture (MRR) experiments. Survival time is, in this case, often fit to an exponential model, as in, e.g. (Midega et al. 2007; Olayemi and Ande 2008), but see Ryan et al. (2015a) for a recent example of applying the Gompertz distribution to wild populations.

Temperature and humidity dependence Temperature strongly affects mosquito survival, with (age-dependent) death rates increasing above about 23 °C, and death via thermal stress occurs rapidly by 40 °C (*A. gambiae* survival also tends to increase at higher relative humidity). Several works have relied on a simple survival curve-fit by

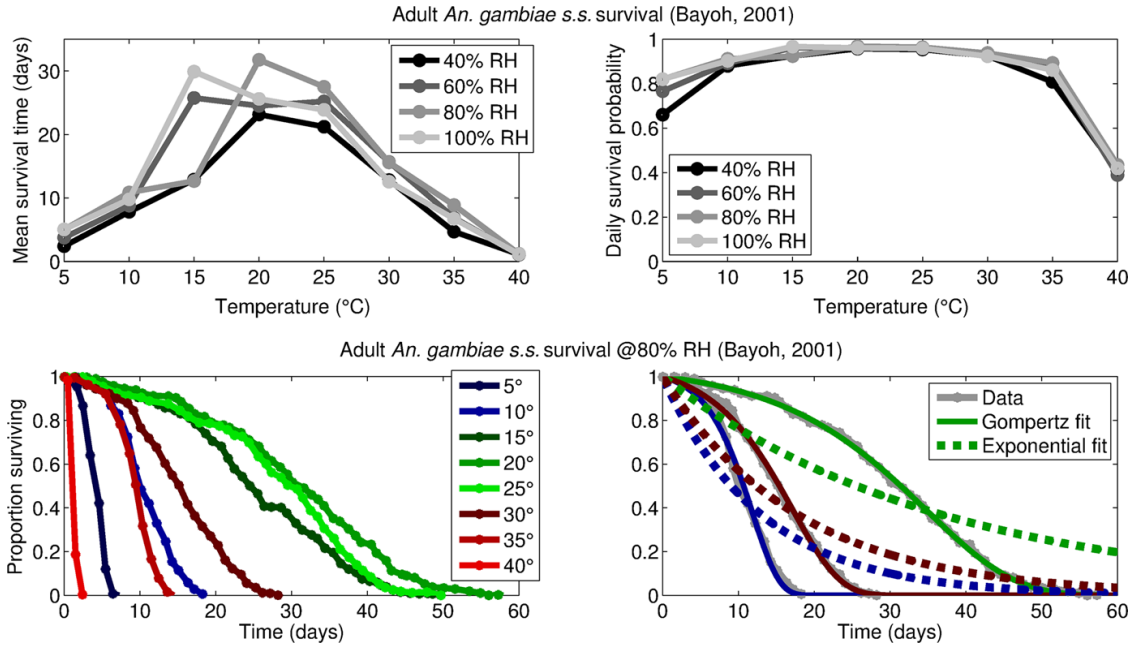


Fig. 8 Adult *A. gambiae* survival data adapted from Bayoh (2001). The top left panel gives reported mean female survival as a function of temperature and relative humidity (adapted from Table 6.1 of Bayoh (2001)), while the top right shows this transformed into daily survival probability, assuming exponentially distributed survival times (this was in turn used to inform daily survival in Mordecai et al. (2013)); note that these curves are not greatly dissimilar to Martens' survival fit (Fig. 14). The bottom left shows survival curves at 80% RH, adapted from Figure 6.4 of Bayoh (2001). The bottom right shows that while a Gompertzian survival function gives an excellent fit to the data (curves shown for 10, 20, and 30 °C), an exponential distribution fits rather poorly

Martens et al. (1995b) to three data points from 1949 (see Fig. 14), but better data is now available. Bayoh, in a 2001 dissertation (Bayoh 2001), presented data on adult *A. gambiae* survival in conditions between 5 and 45 °C, at relative humidities of 40, 60, 80, and 100%, as given in Fig. 8. This data has been used in a number of models, beginning with Mordecai et al. (2013), but these generally assume exponentially distributed survival, which is not actually consistent with the data, as also seen in Fig. 8.

Most recently, Christiansen-Jucht et al. (2014) reared *A. gambiae* larvae under four different temperatures (23, 27, 31, and 35 °C), recorded their survival, and then examined the survival curves for adults from these cohorts under the same temperature regimes (no larvae survived at 35 °C, so this was excluded from adult experiments). Higher temperatures reduced survival of either mosquito stage in general, and adult attrition was exacerbated by a disconnect between larval and adult temperatures. For example, mosquitoes subjected to 31 °C as adults were protected somewhat by higher larval temperatures, while adult mosquitoes kept at 23 or 27 °C suffered noticeably higher attrition if they emerged from 31 °C conditions. Overall, survival was reasonably well-described by a Gompertz distribution, with Gompertz parameters as a function of larval (T_L) and adult temperatures (T_A), given in Fig. 9.

Blood meal dependence Seeking a host and taking a blood meal are much riskier endeavors than resting to digest blood, and therefore Lindsay and Birley (1996) suggested that survival may be relatively constant *per* gonotrophic cycle (at about 50% per cycle), regardless of cycle length, and Dawes et al. (2009) also observed a spike in

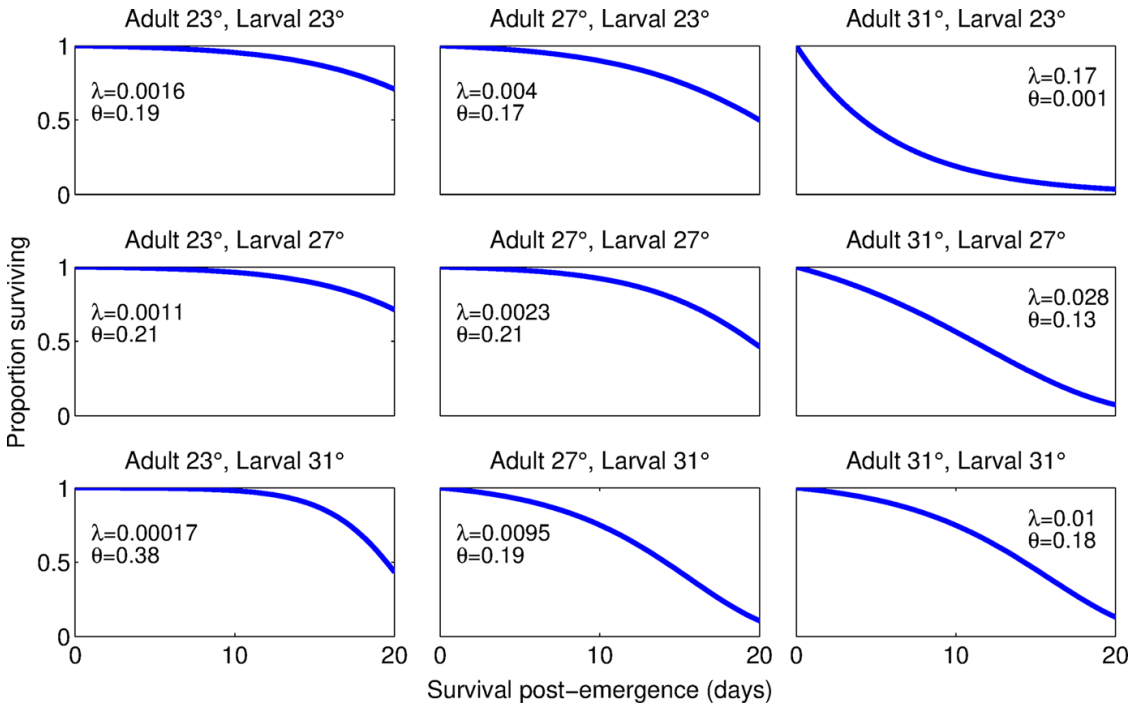


Fig. 9 Gompertzian survival distributions fit from Christiansen-Jucht et al. (2014), who recorded adult *A. gambiae* survival at different temperatures following larval maturation at various temperatures

mortality after feeding *A. stephensi* even under experimental conditions. The notion of constant mortality per blood meal was incorporated into a vector lifecycle model by Hoshen and Morse (2004). As a late-night feeding species, *A. gambiae* may be less likely to suffer attrition when feeding than many other mosquitoes (Klowden and Briegel 1994).

Infection dependence Dawes et al. (2009) observed the survival of *A. stephensi*-fed *Plasmodium*-infected blood to decrease with increasing parasite load (as ookinetes), and as reviewed, infection with a higher number of oocytes also seems to incur a survival cost. An older meta-analysis by Ferguson and Read (2002) also concluded that *Plasmodium* infection decreases mosquito survival, and intriguingly, Pollitt et al. (2013) more recently observed that infection with higher oocyst densities both decreased vector survival and, perhaps via increased competition among parasites for nutrients or a more robust immune response, also decreased the number of infectious sporozoites resulting from infection. In sum, infection is not benign in the adult mosquito, but this has been rarely, if ever, incorporated into malaria transmission models.

5.2.3 Temperature and immature development and survival

Both development time and mortality of aquatic stage anophelines are clearly and nonlinearly related to temperature, although some models (e.g., Craig et al. 1999; Hoshen and Morse 2004; Parham and Michael 2010) have only considered development time as a decreasing function of temperature with mortality temperature-independent. Several works (Craig et al. 1999; Hoshen and Morse 2004; Parham and Michael 2010;

Alonso et al. 2011) have used a relation derived from a 1947 work by Jepson et al. (1947), who reported the development rate of *A. gambiae* as a function of mean temperature at 11 natural breeding sites in Mauritius, with the larval duration time (in days), $l(T)$, reported by Craig et al. (1999) as

$$l(T) = \frac{1}{0.00554T - 0.06737}. \tag{22}$$

More recently, Bayoh and Lindsay (2003, 2004) determined *A. gambiae* larval development and survival as a function of temperature under laboratory conditions, at constant temperatures between 10 and 40 °C. Now, in Bayoh and Lindsay (2003), development time from egg to adult generally decreased with temperature from 18 to 26 °C, while leveling off at about 10 days from 26 to 32 °C, as shown in Fig. 10. Below 18 °C and above 32 °C, no eggs survived to adulthood, presumably because of high attrition rates, rather than altered development time. Overall development rate, $r(T)$, was described by the authors using the function

$$r(T) = a + bT + ce^{T} + de^{-T}. \tag{23}$$

In a second work, Bayoh and Lindsay (2004) describe temperature-dependent larval survival time and the fraction surviving to adulthood, as shown in Fig. 10. It is important to note that survival did *not* follow an exponential distribution. That is, life expectancy was highly correlated with age, with the age-life expectancy curve shifted by ambient temperature, such that for a given temperature, most larvae live a similar lifespan. Therefore, as with adult survival (Sect. 5.2.2), a simple differential equation assuming first-order death kinetics is not a true representation of the biology.

5.2.4 Larval density and immature development and survival

Larval population density negatively affects anopheline larval survival and other life history traits, primarily development time and adult size, as demonstrated in several laboratory and semi-natural artificial breeding site experiments (Lyimo et al. 1992; Schneider et al. 2000; Gimnig et al. 2002; Jannat and Roitberg 2013; Muriu et al. 2013), although the exact relationships between density and life history vary somewhat with experimental setting. Rainfall and other climate variables strongly determine the size and availability of aquatic habitats, especially the small temporary pools preferred by *A. gambiae* (Minakawa et al. 1999), and thus, understanding larval density-dependence is necessary for a full accounting of weather and the anopheline lifecycle.

For various mosquitoes, it had generally been found that crowding leads to longer development times, lower survival, and smaller adults, but this was not tested in anophelines until 1992, when Lyimo et al. (1992) raised *A. gambiae* in the lab in plastic trays at either 24, 27, or 30 °C, at densities of 0.5, 1.0, or 2.0 larvae/cm^2. Yet, their results were rather curious, suggesting increased mortality with increasing density at 24 or 30 °C, but the opposite at 27 °C, while higher densities actually seemed to slightly reduce development time. Schneider et al. (2000) performed similar experiments upon populations of *A. gambiae* and *A. arabiensis* at 27 °C, again at densities

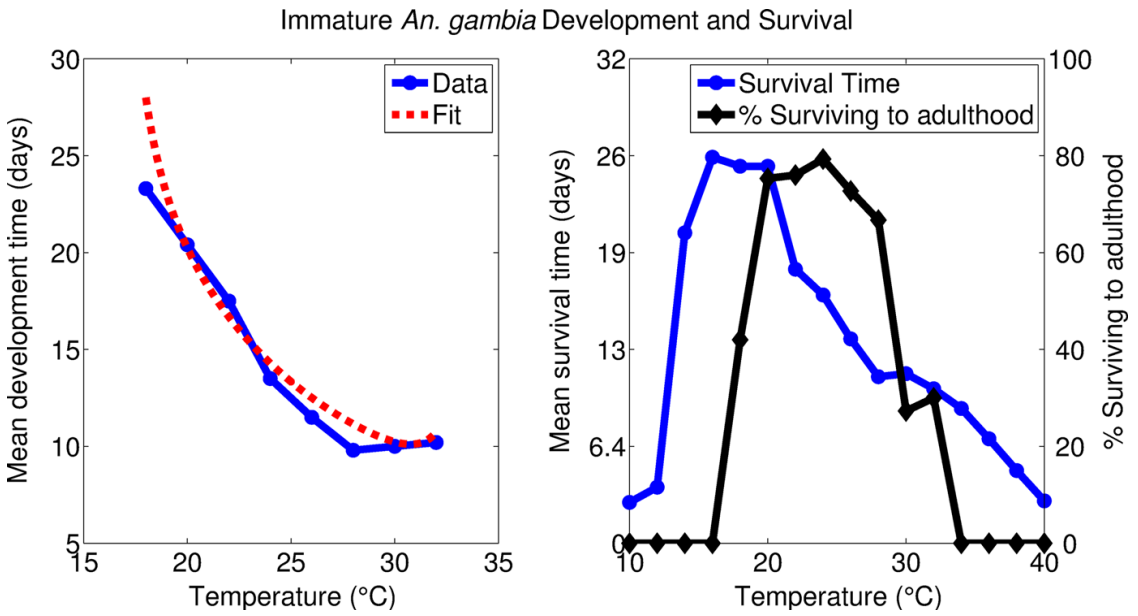


Fig. 10 The left panel gives immature *A. gambiae* development time as a function of temperature, from Bayoh and Lindsay (2003), and the fit to Eq. (23) ($a = -.05$, b $= 0.005$, c $= -2.139 \times 10^{-16}$; d $= -281357.656$), while the right panel gives overall survival time and the percentage of immature mosquitoes surviving to adulthood (from Bayoh and Lindsay (2004)). Note that the peaks of the survival duration and survival to adulthood curves are offset as a consequence of temperature-dependent development. That is, while there is a penalty to absolute survival duration within the 20–30 °C range, immature development time also falls over this temperature range. Thus, up to about 25 °C, the faster development time outweighs the mortality cost, such that overall survival to adulthood is maximized at higher temperature values than absolute survival time

of 0.5, 1.0, or 2.0 larvae/cm^2, and observed the highest density to reduce survival, while the density effect on development time was equivocal.

In sharp contradistinction, three more recent studies (Gimnig et al. 2002; Muriu et al. 2013; Jannat and Roitberg 2013) that used lower larval density ranges showed a very clear negative relationship between density and both development time and adult mosquito weight. Gimnig et al. (2002) raised *A. gambiae* larvae in outdoor artificial habitats mimicking typical field conditions in west Kenya (essentially dried mud pits), each containing about 1 L of water and 600 cm^2 of water surface area, and examined life history parameters across a range of densities, from 0.0333–0.333 larvae/cm^2: the development and weight trends were very clear. Furthermore, while survival did decrease with density, this was not significant under statistical analysis. Muriu et al. (2013) performed similar outdoor experiments in coastal Kenya (density range 0.0333–0.5322 larvae/cm^2), published density-dependent survival and development curves, and found development rate, survival, and weight to uniformly decrease with density.

Finally, Jannat and Roitberg (2013) most recently attempted to separate the effects of competition for food from crowding *per se* in *A. gambiae*, by raising larvae at different densities with food at either a fixed per capita level or at a fixed total level (and at 30 ± 2 °C, 75–80 %RH). Even with adequate food per larvae, crowding led to higher mortality, smaller adults, and a skewed male:female ratio favoring females. A possible mechanism for the latter is that male larvae are smaller and thus may be more vulnerable to crowding stresses. Under fixed total food resources, time to adulthood

was additionally prolonged, and larvae similarly suffered increased mortality and smaller size. Note that crowding, by itself, did not lead to prolonged development time, which may therefore be a consequence solely of nutritional stress.

Given the divergent results of the lower (Gimnig et al. 2002; Muriu et al. 2013; Jannat and Roitberg 2013) and higher density (Lyimo et al. 1992; Schneider et al. 2000) studies, we must then ask, what larval densities are typically encountered in the field? Early field studies performed and reviewed by Service (1971), using 100 mL dippers with a 9.5 cm diameter, gave 0.5–2 fourth-instar larvae per dip. Since about 5% of all larvae were fourth-instar in these studies, this suggests an overall density of 0.035–0.14 larvae/cm^2, and is congruent with mark-release-recapture experiments in five pools (Service 1971) that suggested 0.04–0.10 larvae/cm^2 (again assuming 5% of larvae are fourth-instar).

Much more recently, Kweka et al. (2012) sampled 51 aquatic habitats in western Kenya over 85 weeks, and found about 6 *A. gambaie* larvae per 350 mL dipper in hoofprints and swamps. Supposing, say, a circular right angle geometry for hoofprints with a generous depth of 10–25 cm, then this translates into no more than 0.05–0.15 larvae/cm^2.

Considering all the aforementioned collectively, it seems likely that, at low to moderately high densities that are within the range typically encountered in the field, the dominant effect of increasing density is nutrient competition, in turn resulting in delayed development, smaller adult sizes, and increased mortality. Developmental effects may plateau around 0.5 larvae/cm^2, and at very high densities crowding may directly stress larvae to reduce survival. Survival seems to be affected across density ranges, but the exact relationship varies significantly among studies. Development time as a function of larval density from the studies reviewed above are compared graphically in Fig. 11, along with survival curves adapted from Muriu et al. (2013). In conclusion, those lab/semi-field studies employing lower densities are much more realistic and relevant, and thus we may expect density to deleteriously affect both survival and development rates, although often only the former is considered in models, but see, e.g., Lunde et al. (2013b) for an exception.

5.2.5 Rainfall: simple models for carrying capacity, oviposition, and survival

Modeling of rainfall's effect on the vector lifecycle is more variable than that of temperature in published works, and it is more uncertain. By determining the quantity of habitat available, rainfall affects both oviposition and density-dependent larval development and survival. Excessive rainfall also can increase immature mosquito attrition via washout of habitats, and this has been modeled as well. We review several of these phenomena and basic modeling approaches here, while we explore more complex physical modeling of the anopheline microhabitat in the next section.

Simple carrying capacity models Perhaps the most straightforward way to model this notion is to have the carrying capacity of larval habitats, K, be a function of recent rainfall. Yé et al. (2009) modeled growth of the adult mosquito population very simply using a logistic growth term, with K linearly proportional to the prior week's summed rainfall (under the assumption that several days' worth of rain contribute to breeding

pools); the larval compartment was not modeled. Several other recent works used a 'quasi-logistic" term to describe the larval death rate (White et al. 2011; Christiansen-Jucht et al. 2015). This was incorporated into a simple mosquito lifecycle model by White et al. (2011) as

$$\frac{dE}{dt} = \beta M - \frac{1}{d_E}E - \mu_{E_0}E\left(1 + \frac{E+L}{K(t)}\right), \tag{24}$$

$$\frac{dL}{dt} = \frac{1}{d_E}E - \frac{1}{d_L}L - \mu_{L_0}L\left(1 + \gamma\frac{E+L}{K(t)}\right), \tag{25}$$

$$\frac{dP}{dt} = \frac{1}{d_L}L - \frac{1}{d_P}P - \mu_P P, \tag{26}$$

$$\frac{dM}{dt} = \frac{1}{2}\frac{P}{d_P} - \mu_M M, \tag{27}$$

where E and L are the number of early and late larval instars, respectively, P is pupae count, and M is the adult female mosquito count; d_i is the mean duration of stage i, and the $1/2$ factor in Eq. (27) accounts for the emerging adult male:female sex ratio. The larval "carrying capacity", $K(t)$ (in units larvae), was modeled as a convolution of recent rainfall with some weighting function, either a constant, linearly decreasing, or exponentially decreasing function; the latter is given as

$$K(t) = \lambda\frac{1}{\tau\left(1 - \exp\left(\frac{-t}{\tau}\right)\right)}\int_0^t \exp\left(\frac{-(t-\acute{t})}{\tau}\right)\text{rain}(\acute{t})d\acute{t}, \tag{28}$$

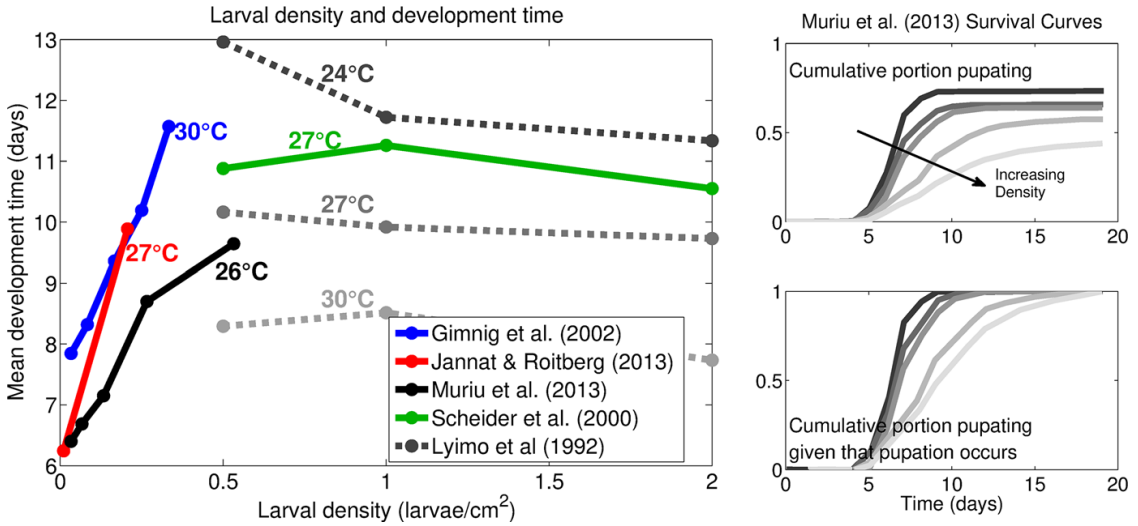


Fig. 11 The left panel gives the relationship between larval density (per unit surface area) and mean development time, with data extracted from five studies (Gimnig et al. 2002; Muriu et al. 2013; Jannat and Roitberg 2013; Lyimo et al. 1992; Schneider et al. 2000). Curves are labeled with the approximate experimental mean temperature, and there are three such curves from Lyimo et al. (1992). Note that mean development time for Muriu et al. (2013) was extracted from the survival curves given on the right of the figure, where the top gives cumulative portion of larvae surviving to pupation, and the bottom shows time to pupation for those that survived to pupation

where 2τ is the mean of the exponential distribution, rain(t) is daily rainfall, and λ is a free scalar. In other terminology, $K(t)$ is given by passing rainfall through a low-pass filter, with one option (the exponential filter) a leaky integrator.

It is important to point out that, under this formulation, $K(t)$ is not actually a carrying capacity. It is, rather, the inverse of a density-dependent death term, as expanding Eq. (25), for example, gives

$$\frac{dE}{dt} = \beta M - \frac{1}{d_E} E - \mu_{E_0} E - \mu_{E_0} \frac{E^2}{K(t)} - \mu_{E_0} E \frac{L}{K(t)}. \tag{29}$$

The actual carrying capacity is a complex function of $K(t)$, β, and the other model parameters. One can see this should be true by analogy to the simpler Verhulst equation (Vogels et al. 1975) for population growth, which imagines death to increase with the square of population as

$$\frac{dy}{dt} = ay - by^2, \tag{30}$$

which can be rearranged to the "ecological" logistic equation as

$$\frac{dy}{dt} = ry\left(1 - \frac{y}{K}\right), \tag{31}$$

with $r = a$ and $K = a/b$ the "carrying capacity". Thus, carrying capacity here is not independent, but is a function of other more fundamental model parameters. It is important that claims concerning the biological meaning of mathematical terms are internally consistent, otherwise serious errors could be introduced, say, when parameterizing a "carrying capacity" from experimental data or in the interpretation of model output. Several other models (such as Agusto et al. (2015), Okuneye and Gumel (2017)) have used the logistic term to model the birth rate of immature mosquitoes (although without explicit dependence of the carrying capacity on rainfall, but see Depinay et al. (2004) for an exception), a setting where it more properly imposes a well-defined limit to larval population size.

Oviposition Oviposition is affected by both the raw availability of breeding habitat, and the density of larvae within potential positing sites (Sumba et al. 2008). Hoshen and Morse (2004) modeled this very simply, by assuming that each oviposition event yields γR_d eggs, where R_d is the dekadal (i.e., sum of the prior ten days) rainfall in mm, and γ was 1 egg/mm. It is reasonable that rainfall over the recent past should be integrated, as it is the sum of standing water that provides habitat, but this simple linear relation is probably not supported.

Sumba et al. (2008) studied experimentally how the presence of larvae in aquatic habitat either encouraged or discouraged *A. gambiae* oviposition. Interestingly, they found that while pre-existing larvae uniformly discouraged oviposition when distilled water was used, when natural anophelene pool water was used, low densities of larvae actually encouraged oviposition with a shift to deterrence only at high density. Additionally, the larger late instars were more of a deterrent, while the presence of one-day old eggs had no effect either way. As illustrated in Fig. 12, the experimental

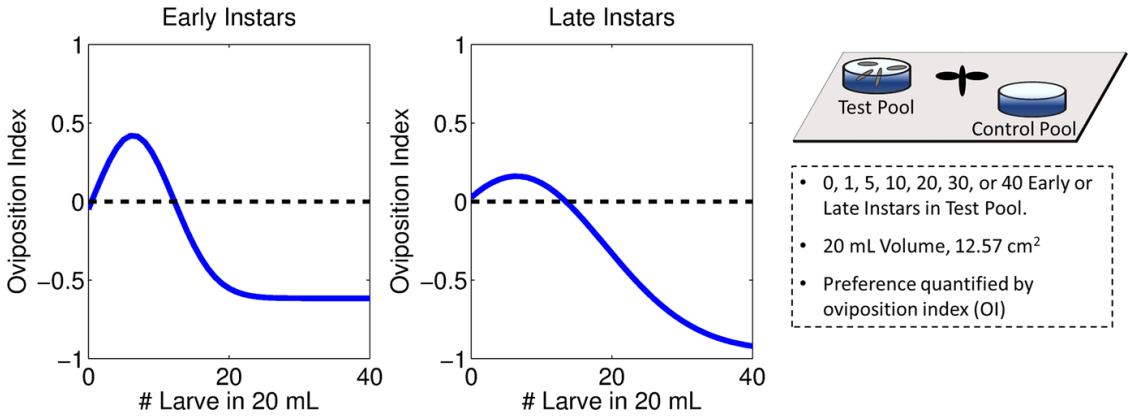


Fig. 12 Oviposition index in *A. gambiae* as a function of larval count (in 20 mL of liquid) in the test pool, when larvae are either early (left panel) or late (right panel) instars, as determined by the experiments of Sumba et al. (2008) where water was taken for natural anophelene pools (the presence of larvae uniformly deterred oviposition when distilled water was used). Curves are fits to $\mathrm{OI} = a\exp(-b(x-c)^2) - d$, where x is the number of larvae, and for early instars $a = 1.037$, $b = 0.015$, $c = 6.34$, $d = 0.616$; for late instars $a = 1.1136$, $b = 0.00315$, $c = 6.42$, $d = 0.9524$

assay offered gravid mosquitoes a choice between two pools (2 cm deep, 4 cm diameter, 12.57 cm^2 SA, 20 mL liquid volume), one containing between 1 and 40 early or late instars (test), the other empty (control). Pool preference was quantified by the oviposition index (OI), defined as

$$\mathrm{OI} = \frac{N_t - N_s}{N_t + N_s}, \tag{32}$$

where N_t is the number of eggs laid in the test pool and N_s the number in the control pool. The OI ranges from -1 to $+1$, with positive values indicating a preference for the test pool, and Sumba et al. (2008) fit oviposition data using the functional form

$$\mathrm{OI} = a\exp\left(-b(x-c)^2\right) - d, \tag{33}$$

where x is the number of larvae and a, b, and c are free parameters; results for early and late instars are given in Fig. 12. Now, these experiments used extraordinarily high larval densities, as field studies reviewed previously (Service 1971; Kweka et al. 2012) suggest less than 0.02 larvae/mL is more common, and therefore, within plausible field densities larval density likely has no or a slightly positive effect on oviposition.

These results were incorporated into a model by Parham et al. (2012), who recast Sumba et al.'s curve fits for OI to depend upon density, not larval count, with density calculated from total habitat volume, itself determined according to a hydrodynamics model (presented in part below), and then assumed that the fraction of eggs a gravid female lays, f_t, is

$$f_t = \frac{N_t}{N_s + N_t} = \frac{1}{2}(\mathrm{OI} + 1). \tag{34}$$

While this relation has an empirical basis, it is unclear if only half of eggs should be laid at an OI of 0, which simply implies no preference for empty over already

occupied pools, or, more generally, that a preference rank for unoccupied pools in a simple forced choice experiment translates linearly to overall oviposition likelihood in a general environment.

Rainfall-dependent mortality A widely-cited 2007 experiment by Paaijmans et al. (2007) placed either first- or fourth-instar *A. gambiae* larvae in outdoor artificial basin habitats in western Kenya over the course of the rainy season, and observed significantly increased rates of both flushing losses and larvae mortality during rainy nights, with the younger first-instar larvae's suffering greater, in absolute terms. On this basis, several models have posited different functional forms for increased mortality from rainfall, with Parham and Michael (2010), for example, giving a unimodal relationship between egg survival and rainfall (see Sect. 6.2.2), and Tompkins and Ermert (2013) took larval mortality to increase with precipitation.

5.2.6 Rainfall, hydrodynamics, and the microhabitat

Several more complex, but more realistic, physical descriptions of the microhabitat geometry and heat and water balance have been proposed (Paaijmans et al. 2008a, b; Parham et al. 2012; Asare et al. 2016a, b, c); models for regional-scale hydrology have also been applied to estimating malaria transmission (Bomblies et al. 2008, 2009; Bomblies 2012; Tompkins and Ermert 2013; Asare et al. 2016c), but we restrict our attention here to microscale dynamics. Detailed microscale models have multiple advantages over the simpler approaches discussed above. First, rainfall can directly inform habitat volume and surface area, yielding a time-dependent immature carrying capacity (broadly defined) from basic physical principles and geometric parameters. Second, such models relate water temperature to ambient air temperature in a physically realistic and *non-constant* manner. Finally, one may generate predictions on how local variations in habitat geometry, such as shading, interact with more global parameters, such as ambient temperature. In this section, we review the basic construction of a comprehensive microhabitat model. Figure 13 summarizes the key parameters describing habitat geometry, and the major mechanisms for heat and water volume loss/gain.

Before continuing, we also note that for any hydrodynamic model describing habitat volume and/or surface area, such metrics must be translated into some kind of carrying capacity or density-dependent death term, etc. in the mosquito population dynamics model, and several authors have assumed a biomass carrying capacity for anopheline ponds to be about 300 mg m^{-2}, with fourth stage instars weighing 0.45 mg (Depinay et al. 2004; Bomblies et al. 2008; Tompkins and Ermert 2013). A separate method is that of Lunde et al. (2013b), who calculated an immature anopheline carrying capacity at a relatively high spatial scale as a composite function of soil moisture and *potential* river length, with the latter determined from the HydroSHEDS database, which in turn gives water accumulation potential based upon the Earth's topology.

Basic geometry To model an *Anopheles* microhabitat, one must prescribe some volume (V)-area (A)-depth (h) relationship, with one popular option a simple set of equations developed by Hayashi and colleagues (Hayashi and Van der Kamp 2000; Brooks and Hayashi 2002), that describe small topographic depressions in terms of three

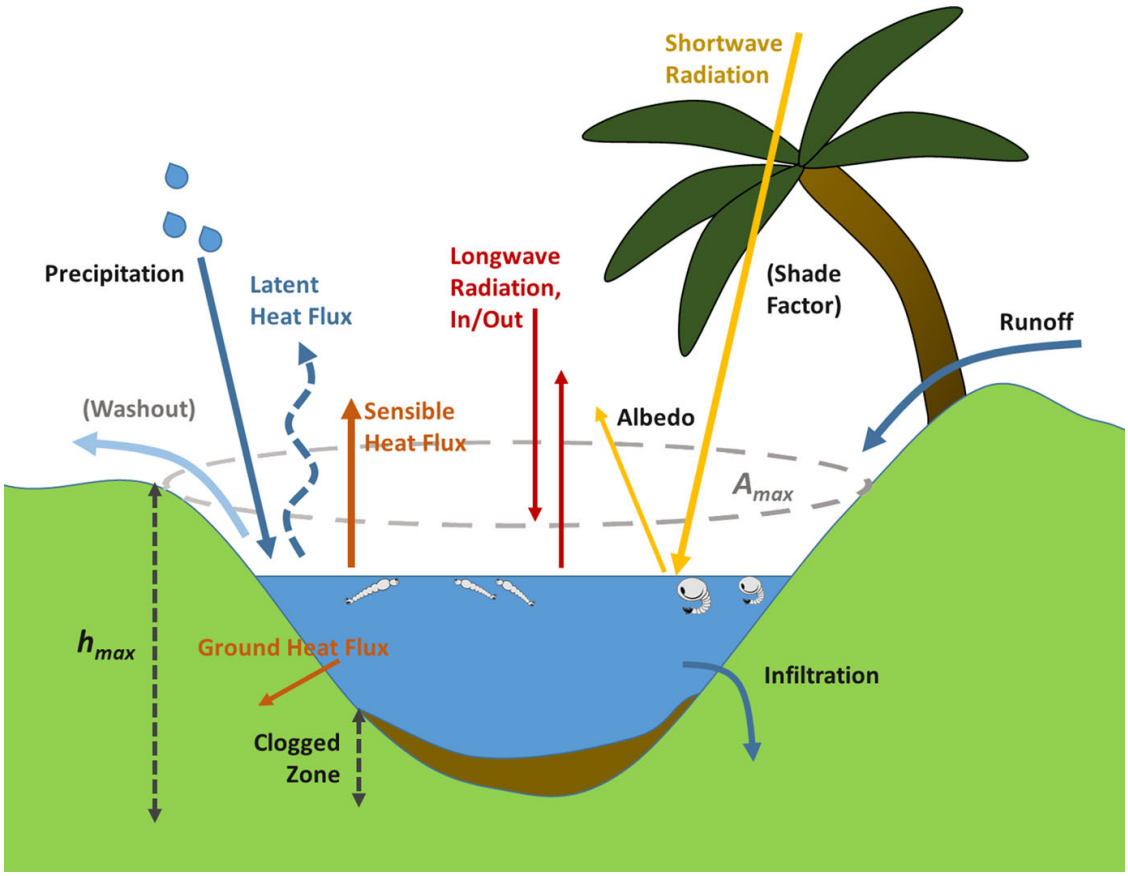


Fig. 13 General schematic for a single *Anopheles* breeding site microhabitat, and the coupled mechanisms determining the overall heat and water balance

empirical parameters: maximum surface area A_{max}, maximum depth, h_{max}, and a dimensionless shape parameter p, such that $p < 1$ and $p > 1$ indicate concave and convex geometries, respectively. If depth, h, is known, A and V are determined as

$$A = A_{max}\left(\frac{h}{h_{max}}\right)^{\frac{2}{p}} \tag{35}$$

$$V = \frac{A_{max}h_{max}}{1+\frac{2}{p}}\left(\frac{h}{h_{max}}\right)^{1+\frac{2}{p}}. \tag{36}$$

If volume, V, is prescribed instead, it is straightforward to rearrange the equations to solve for h and A. Alternatively, Parham et al. (2012) employed a simple right-angle cone as a microhabitat geometry.

Heat-balance The heat and water volume balance within a microhabitat are linked, and, generally speaking, we have the change in heat, dQ/dt (in W), as

$$\frac{dQ}{dt} = A(R_n - \lambda E - H - G) + P_Q - I_Q \tag{37}$$

where R_n is net radiation per unit area (W m^{-2}), λE is latent heat flux (W m^{-2}), H is sensible heat flux (W m^{-2}), G is heat flux through the surrounding soil (W m^{-2})

(Allen et al. 1998; Asare et al. 2016b), while heat is also contained in the water gained via precipitation and runoff, P_Q, and lost via infiltration, I_Q. We briefly examine each component of Eq. (37). First, net radiation, R_n, decomposes as the sum of incoming shortwave radiation, R_s, and incoming and outgoing longwave radiation, L_{in} and L_{out}, respectively,

$$R_n = R_s - L_{out} + L_{in}, \tag{38}$$

Shortwave radiation is the fraction of the global horizontal irradiance (GHI) that is not reflected or blocked by shade,

$$R_s = \text{GHI}(1-a)(1-SF), \tag{39}$$

where a is the albedo of water and SF is a shade factor. The longwave radiation balance can be determined from the relatively simple relations,

$$L_{out} = \epsilon_w \sigma T_{w(K)}^4, \tag{40}$$

$$L_{in} = \epsilon_a \sigma T_{a(K)}^4 (1-SF), \tag{41}$$

where $\epsilon_w \approx 0.98$ the emissivity of water (Paaijmans et al. 2008b), $\sigma = 5.67 \times 10^8$ W m^{-2} K^{-4} the Stefan-Boltzmann constant, $T_{w(K)}$ and $T_{a(K)}$ are water and air temperatures in Kelvin, and finally, ϵ_a is either the clear-sky emissivity, or the cloud-corrected atmospheric emissivity. A variety of algorithms exist to estimate ϵ_a, as reviewed by Flerchinger et al. (2009), with one simple option for clear-sky emissivity due to Ångström (Flerchinger et al. 2009) given as

$$\epsilon_a = \left(.83 - .18 \times 10^{-.067 e_a}\right), \tag{42}$$

where e_a is saturation pressure (kPa).

Latent heat flux, λE, determined from the mass of water evaporated per unit time, E (kg day^{-1}), and the latent heat of evaporation, λ, equal to 2.45 MJ kg^{-1} at 20 °C (Allen et al. 1998), is a relatively complex phenomenon, as it involves both storage of heat into the latent form of water vapor at the water surface, and the removal of this vapor from the surface. In general, with lower vapor pressure at the water surface and faster wind speed, both processes (water vapor formation and removal) are accelerated. These qualitative notions can be simply formalized to model evaporation (i.e. water mass loss) as a bulk transfer process (Sene et al. 1991; Paaijmans et al. 2008a; Asare et al. 2016b)

$$E = Cu(e_{sw} - e_a), \tag{43}$$

where C is the mass transfer coefficient, u (m s^{-1}) is wind speed at reference height, e_{sw} is the vapor pressure at saturation for the water surface temperature, and e_a is the atmospheric vapor pressure at reference height. An alternative method is to use the Penman-Monteith equation, or a variation thereof, derived from the simultaneous solution of equations for energy balance and mass transfer as (Finch and Hall 2001)

$$E = \frac{1}{\lambda}\left(\frac{\Delta\Lambda + \frac{\rho c_p(e_s - e_a)}{r_a}}{\Delta + \gamma\left(1 + \frac{r_s}{r_a}\right)}\right) \tag{44}$$

where Λ is the available energy (typically taken as $R_n - G$ (Allen et al. 1998)), Δ is slope of the vapor pressure curve (kPa $^{\circ}$C $^{-1}$), e_s is the saturation vapor pressure, e_a is the actual vapor pressure, ρ_a is the density of air, c_p is the specific heat of air, γ is the psychrometric constant, defined as

$$\gamma = \frac{c_p P}{\epsilon\lambda}, \tag{45}$$

where P is the atmospheric pressure and $\epsilon = 0.622$ is the ratio of the molecular weight of water vapor to dry air. Furthermore, in this formulation, water vapor mass transfer is governed by two resistances in series, first a "bulk" surface resistance, r_s, which is a sum measure of the resistance to vapor flow from soil and vegetation and may be set to zero over open water, and a second aerodynamic resistance, r_a, which is inversely proportional to wind speed (see Finch and Hall (2001) or Allen et al. (1998)). Parham et al. (2012) employed the FAO, or modified, Penman-Monteith equation for transevaporative flux from a vegetated surface to describe water loss from anopheline habitat; further details may be found in Parham et al. (2012) and Allen et al. (1998).

Similar to Eq. (43), sensible heat flux, H, may be given as (Asare et al. 2016b)

$$H = \rho_a c_p C u (T_w - T_a), \tag{46}$$

and G is typically taken as some small fraction, f, of R_n, perhaps 0.15 (Asare et al. 2016b; Paaijmans et al. 2008a), i.e.

$$G = f R_n. \tag{47}$$

Finally, the heat contained in precipitation and infiltration water (P_Q and I_Q respectively) is simply determined from the density and specific heat of water, and using the volume-balance principles presented next.

Volume-balance The change in water volume, dV/dt, may be approximated as a function of an imposed precipitation time-series, $P(t)$ (m day^{-1} or m s^{-1} if appropriate), and loss to evaporation (E) and infiltration (I), as

$$\frac{dV}{dt} = P(t)(A + R_{frac}(A_{catch} - A)) - A(E + I), \tag{48}$$

where A_{catch} is the catchment area for precipitation runoff, R_{frac} is that fraction of runoff water within the catchment area that makes it to the habitat, and Eq. (48) is also subject to the constraint that V not exceed V_{max}. Evaporation, E, is determined as above. Infiltration of water in sandy pools in the Sahel region proceeds in a roughly biphasic manner, where water is initially lost rapidly to the porous sandy soil, but low permeability clay that collects at the bottom creates a "clogged" zone, in which

infiltration slows dramatically (Desconnets et al. 1997; Porphyre et al. 2005). This phenomenon was simply modeled by Asare et al. (2016a), who gave I as

$$I = I_{max}\left(\frac{A}{A_{max}}\right), \tag{49}$$

where I_{max} (m day^{-1}) is the maximum infiltration rate. A more comprehensive model could also incorporate overflow and washout of immature anophelines during heavy rains.

Water versus air temperature Water and air temperatures in open anopheline habitats are likely to vary by 3–6 °C or more (Parham et al. 2012), and such a disparity could strongly affect optimal ambient temperatures for malaria transmission. Paaijmans et al. (2008b) observed mean water temperatures to be several degrees Celsius above the surrounding air in artificial habitats, with the difference greatest during the day, when solar gain into pools is high. The difference between maximum air and water temperatures could exceed 10 °C, and, unsurprisingly, there was greater thermal stability in larger pools (see also Paaijmans et al. (2008a)). Such an air-water temperature disparity is also observed in simulations of the theoretical microhabitat framework presented above. Thus, while many works assume a constant air-water temperature difference (often zero), this is probably overly simplistic.

6 A partial genealogy of recent weather-driven malaria models

It is well beyond our scope to review all malaria mathematical models that incorporate weather, and we must omit any discussion of many excellent works (a partial list of works not considered further here includes (Depinay et al. 2004; Lou and Zhao 2010; Cailly et al. 2012; Dembele et al. 2009; Bomblies et al. 2009; Eckhoff 2011; White et al. 2011; Lunde et al. 2013a, b; Tompkins and Ermert 2013; Nikolov et al. 2016)). We restrict our attention to process-based, mechanistic models, and among these focus on several influential lines of work from the past two decades, beginning with widely cited works from the 1990s by Martens and colleagues which suggested a significantly increased malaria range with global warming, and moving through a 2013 paper by Mordecai et al. (2013), who concluded prior works had significantly overstated the optimum temperature range for transmission. Beyond that work, we highlight several recent efforts with slightly different focuses, namely immunity (Agusto et al. 2015; Yamana et al. 2013, 2017), host age-structure (Okuneye and Gumel 2017), and age-dependent vector survival (Christiansen-Jucht et al. 2015). Recently, the importance of daily and seasonal temperature fluctuations in determining malaria potential has recognized, and we discuss several recent works which make this their focus (Paaijmans et al. 2009, 2010, 2013b; Blanford et al. 2013; Beck-Johnson et al. 2017). All models are informed by the empirical vector/parasite-weather relations covered in the prior section, and many directly employ the Ross–Macdonald model framework. We close this section with a discussion of host mobility in multi-patch geometries, which heretofore has not been incorporated in weather-driven malaria models, but is

of potentially great value in studying the possible combined roles of climate change and human mobility in the spread of malaria into highland Kenya.

6.1 Epidemic potential models (1995–1999)

Several works by Martens and others (Martens et al. 1995a, b, 1997, 1999), published in the late 1990s, used the notion of "epidemic potential" (EP), defined to be the inverse of the critical vector density, m_c (units mosquitoes/man), in turn derived from the Macdonald model by setting $\mathcal{R}_0 = 1$ (see Eq. (15)) and solving for m, giving

$$m_c = k_1 \frac{-\ln(p)}{a^2 p^n}, \text{EP} = \frac{1}{m_c}, \tag{50}$$

where k_1 is a constant, and equal to $r/(bc)$ for Macdonald's model. The parameters n (sporogonic duration), a (daily biting rate), and p (daily mosquito survival) were determined as a function of temperature, with n determined via Moshkovsky's formula ($D = 111$ and $T_{min} = 16 - 19$), a assumed to be proportional to the length of the gonotrophic cycle and similarly determined from Moshkovsky's formula (Eq. (18), with $D = 36.5$ and $T_{min} = 9.9$), and p determined via a trinomial fit to three data points from a 1949 work, such that

$$p(T) = \exp\left(\frac{-1}{-4.4 + 1.31T - .03T^2}\right). \tag{51}$$

These parameters, and the resulting EP, are graphed as functions of temperature in Fig. 14. Using geographical data from GCMs, these authors then concluded that global warming could cause overall epidemic potential to increase by 12–27% in 2050 (Martens et al. 1997). The idea of EP is primarily applicable to areas not currently endemic for malaria: it is meant to elucidate what regions may *become* vulnerable in the future. However, Rogers and Randolph (2000) pointed out that, in areas where $\mathcal{R}_0 < 1$ (and, thus, not vulnerable to epidemics), an increase in EP does not mean that $\mathcal{R}_0$ increases above 1. Thus, the EP notion can overestimate the effect of temperature changes. Even considering that critical mosquito densities may decrease with warming, this is a very incomplete picture, as it merely gives a threshold mosquito density necessary to spark an epidemic, while failing to predict how mosquito densities will change under climate change. This is key: we cannot assume that altering major anopheline lifecycle parameters will affect disease transmission but not the *Anopheles* population itself.

The relations reported by Martens et al. (1995b) helped inform a widely cited effort by Craig et al. (1999) (see also Snow et al. (1999)) that used "fuzzy logic" to derive maps of climatic suitability for malaria transmission in Africa. These authors determined that, when rainfall is not limiting, yearly mean temperatures above 22 °C lead to perennial infection, 18 °C is too cold for stable transmission but does allow epidemics in warmer years, while 15 °C is prohibitory.

A later 2011 work by Gething et al. (2011) also used similar temperature-dependent terms and the notion of vector capacity to map global temperature constraints on *P.falciparum* and *P.vivax* transmission.

6.2 More complex models through Mordecai et al. (2013)

6.2.1 Hoshen and Morse model (2004), reformulation and extensions (2011)

In 2004, Hoshen and Morse (2004) developed a more comprehensive model of the mosquito lifecycle, explicitly including temperature dependent progression through the immature (egg, larvae, pupae) mosquito stages (considered as one), coupled with the temperature-dependent gonotrophic cycle of adult female mosquitoes, by which a blood meal is taken to yield oviposition of new eggs (the number of eggs laid is dependent upon the sum of the prior ten day's rainfall), along with a sporogonic cycle and a basic susceptible-exposed-infected-susceptible (SEIS) model for human infection. Note that the sporogonic cycle in infected mosquitoes advances independently of the gonotrophic cycle, and the overall model architecture is schematized in Fig. 15. Immature mosquitoes progress through *physiologic* time (as opposed to chronological time) at a rate, $m(T)$ (1/day), determined as the inverse of the sum of the duration of all larval stages (note that the immature mosquito class is still considered as one by Hoshen and Morse (2004)), with these durations determined from Jepson et al. (1947) (Eq. (22) of Sect. 5.2.3), which gives a hyperbolic relationship between total time in the immature stage and temperature. Death occurs daily, independent of tem-

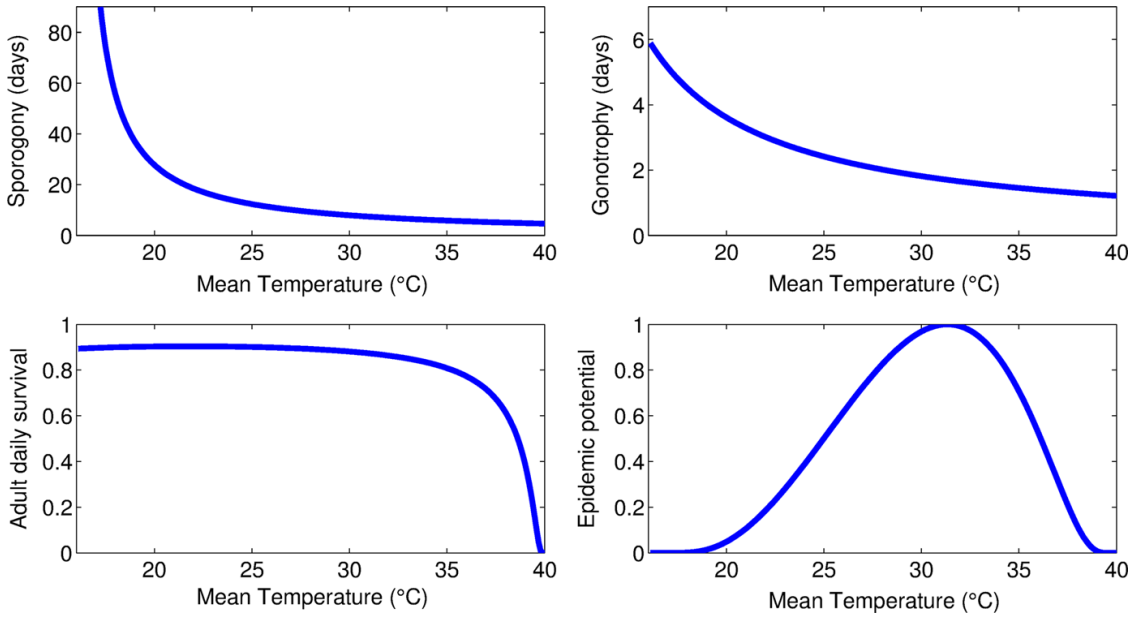


Fig. 14 Temperature dependence of vector and parasite parameters as per Martens et al. (1995b), Martens et al. (1997). The top panels show the duration of the sporogonic and gonotrophic cycles as functions of mean temperature, according to the formula of Moshkovsky using $D = 111$ and $T_{min} = 16$ and $D = 36.5$ and $T_{min} = 9.9$, respectively. The bottom shows adult mosquito survival per the relation derived by Martens et al. (1995b), and the normalized epidemic potential (see Eq. (50)), assuming the inverse of the biting rate $(1/a)$ is equal to the gonotrophic duration

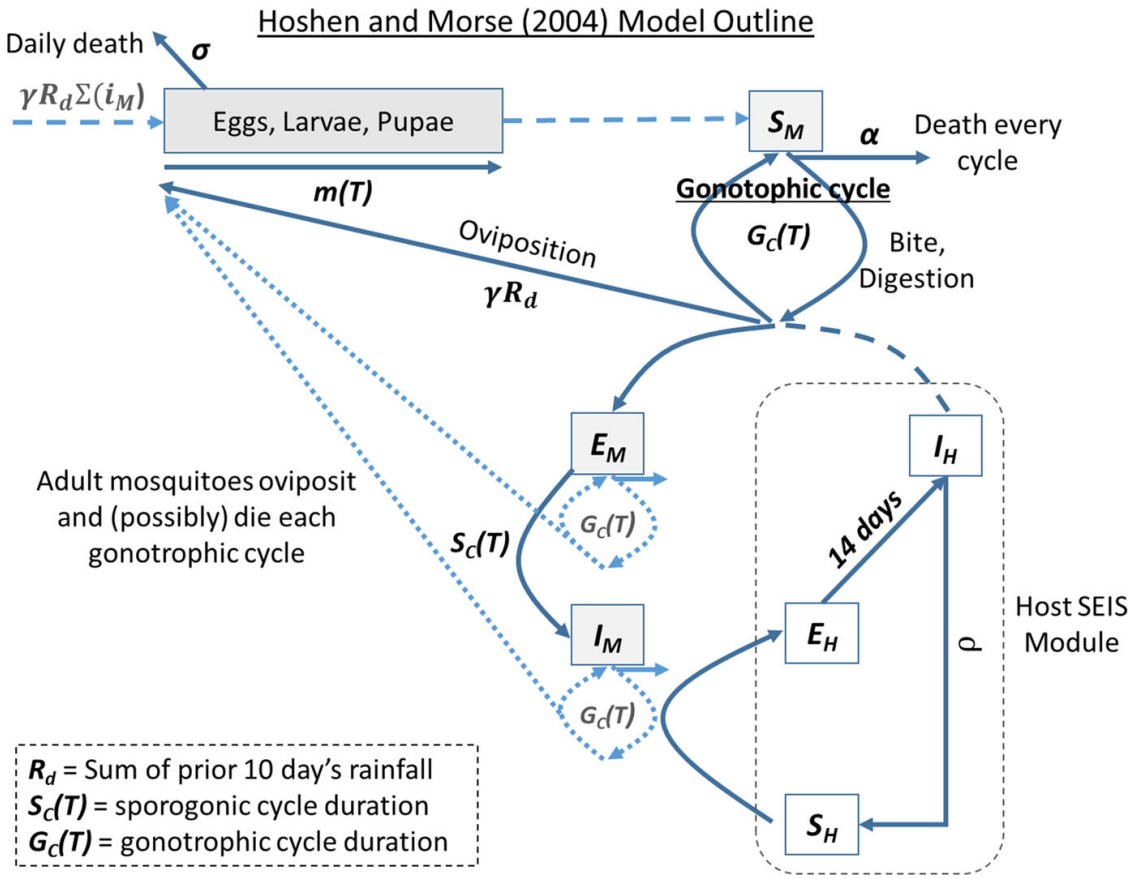


Fig. 15 A schematic for the essential elements of the Hoshen and Morse model. Note that this schematic represents our particular interpretation of the original difference equation-based model

perature, with an assumed daily survival of 90%. The model follows previous work in using Moshkovsky's hyperbolic formula for sporogonic, S_C, and gonotrophic, G_C, durations (see Eq. (18) and Sects. 5.1.1 and 5.2.1).

At the end of each gonotrophic cycle, the (adult female) mosquito takes a blood meal, surviving the risky adventure with probability $\alpha \approx 0.5$, and thus daily survival probability is α^{1/G_C}, giving an overall per-capita death rate of $-\ln(\alpha^{1/G_C})$ (denoted μ in the model equations below). Note that adult survival is therefore indirectly coupled to temperature, and no other mechanism for death is considered. After each gonotrophic cycle, mosquitoes lay $\gamma\ R_d$ eggs, where R_d is the sum of the prior ten days of rainfall in mm, and γ was 1 egg/mm, thus coupling oviposition to habitat created by rain. The fraction of blood meals taken on humans is B, and mosquitoes contract malaria from infected humans with probability χ. If a mosquito is infected, then the sporogonic cycle initiates, taking 111 degree days.

Finally, the human component is a simple susceptible-exposed-infectious-susceptible (SEIS) model with a 14 day delay from exposure to infectivity, with no superinfection or immunity included. A small influx of infected individuals is included, presumably from migration. A small, constant influx of mosquitoes is also included in the model. The model was formulated by Hoshen and Morse as a fairly extensive difference equation, which we do not repeat here, but the general model framework can be translated into continuous time, a more mathematically popular setting. Now, there are three

major continuous-time forms common among malaria models: ordinary-differential equations (ODEs), delay-differential equations (DDEs), and age-structured advection partial-differential equations (PDEs), where the latter describes the movement of populations through infinite-dimensional time and age (age is analogous to space in a traditional advection equation). The simpler ODE version, which is *not* a "direct" translation (it is not generally possible to directly translate from a continuous age-structured model to an ODE model of finite dimension) can be written as

$$\frac{dL}{dt} = \gamma R_d \frac{1}{G_C(T)}(S_M + E_M + I_M) - m(T)L - \sigma L, \tag{52}$$

$$\frac{dS_M}{dt} = m(T)L + \lambda_M - \chi B \frac{1}{G_C(T)} S_M \frac{I_H}{N} - \mu S_M, \tag{53}$$

$$\frac{dE_M}{dt} = \chi B \frac{1}{G_C(T)} S_M \frac{I_H}{N} - \frac{1}{S_C(T)} E_M - \mu E_M \tag{54}$$

$$\frac{dI_M}{dt} = \frac{1}{S_C(T)} E_M - \mu I_M, \tag{55}$$

$$\frac{dS_H}{dt} = -\beta B I_M \frac{1}{G_C(T)} \frac{S_H}{N} + \rho I_H, \tag{56}$$

$$\frac{dE_H}{dt} = \beta B I_M \frac{1}{G_C(T)} \frac{S_H}{N} - \frac{1}{14} E_H, \tag{57}$$

$$\frac{dI_H}{dt} = \frac{1}{14} E_H - \rho I_H + \lambda_H, \tag{58}$$

where

$$\mu = -\ln\left(\alpha^{1/G_C(T)}\right), \tag{59}$$

and where L is the lumped larval stage; S_M, E_M, and I_M are the susceptible, exposed, and infectious adult mosquito populations; S_H, E_H, and I_H are similarly susceptible, exposed, and infectious host populations,with $N = I_H + E_H + I_H$. Latency from exposure to infection is a fixed 14 days in humans, β is the probably an infectious bite infects, ρ is the rate at which infected recover (with no immunity), λ_M and λ_H are respectively the influxes of adult mosquitoes and infected humans from migration, and it assumed that there is no mortality in the human population. Other parameters are as above. Now, to make this translation from Hoshen and Morse's discretely age-structured model, we have assumed exponentially waiting times in every stage, but as intimated throughout Sect. 5, and addressed more directly with a model by Christiansen-Jucht et al. (2015), this is not generally a valid assumption.

Staying truer to the original work, it is also possible to write the model as a set of either age-structured advection equations, or, equivalently, a delay-differential equation, with that caveat that it necessary to cast the model in physiologic, rather than chronological, time. The translation from ODE to either of these settings is relatively straightforward, but requires care with the boundary conditions.

At this point, with the Hoshen and Morse work, we have a fairly realistic model for the mosquito lifecycle, explicitly including dependence upon blood meals for reproduction, and monotonically increasing relationships between temperature and larval

development, sporogony, gonotrophy (and, hence, biting rate), while daily mosquito survival decreases monotonically (as a consequence of mortality occuring during biting, which increases with temperature).

The model of Hoshen and Morse, also referred to as the Liverpool Malaria Model (LMM), was revised modestly by Ermert and colleagues (Ermert et al. 2011a, b), who also performed a more thorough literature review for parameter values, and also applied the model to West African field data (Ermert et al. 2011b). While Hoshen and Morse present their model as being valid only for predicting zones of epidemic malaria (e.g., mesoendemic or hypoendemic area), given that it does not account for immunity, Ermert et al. (2011a) argue that their updated version can apply to endemic zones as well, but the original authors' claim seems biologically well-founded. In any case, it is extremely likely that without explicitly incorporating the effects of immunity in a robust manner, predictions concerning climate change or other interventions on malaria in endemic zones will be suspect, given the fundamental importance of the immune dynamic to the history and epidemiology of the disease (see Sects. 7.1 and 3.3).

6.2.2 Parham and Michael model (2010)

The next major effort we discuss is that of Parham and Michael (2010), a delay differential equation that couples an SEI model for mosquito dynamics with an SIR model for the human population, and directly adopts the temperature-dependent relations for adult mosquito survival, sporogonic duration, and gonotrophic duration from Martens et al. (1995b) (Sect. 6.1), the larval maturity relation from Jepson et al. (1947) (Eq. (22)), and finally adds an assumed nonlinear relation between rainfall and daily egg survival probability, $p_E(R)$,

$$p_E(R) = \frac{4p_{ME}}{R_{LE}^2} R(R_{LE} - R), \tag{60}$$

where R_{LE} is the threshold beyond which no eggs survive due to washout, and p_{ME} is the maximum daily survival fraction. The model governing equations for susceptible, S_M, exposed, E_M, and infectious, I_M mosquito populations, and similarly named human populations (including recovered, R_H, and total human population, N), with time-dependence suppressed except for delay-terms, is given by

$$\frac{dS_M}{dt} = \lambda(R,T) - a(T)b_1 S_M \frac{I_H}{N} - \mu(T)S_M, \tag{61}$$

$$\frac{dE_M}{dt} = a(T)b_1 S_M \frac{I_H}{N} - \mu(T)E_M - a(T)b_1 S_M(t-\tau_M(T))l_M(T)\frac{I_H(t-\tau_M(T))}{N}, \tag{62}$$

$$\frac{dI_M}{dt} = a(T)b_1 S_M(t-\tau_M(T))l_M(T)\frac{I_H(t-\tau_M(T))}{N} - \mu(T)I_M, \tag{63}$$

$$\frac{dS_H}{dt} = -a(T)b_2 I_M \frac{S_H}{N}, \tag{64}$$

$$\frac{dI_H}{dt} = a(T)b_2 I_M \frac{S_H}{N} - \gamma I_H, \tag{65}$$

$$\frac{dR_H}{dt} = \gamma I_H. \tag{66}$$

The adult mosquito influx term, $\lambda(R, T)$, is given as

$$\lambda(R, T) = \frac{Bp_E(R)p_L(T)p_P(R)}{\tau_E + \tau_L(T) + \tau_P}, \tag{67}$$

with B here the number of eggs per oviposition per adult, p_i the daily survival of immature stage i ($i = E, L$, and P for eggs, larvae, and pupae, respectively), and τ_i the development time for stage i (note that while $\lambda(R, T)$ is independent of the total adult mosquito population in this model, these parameters should more generally be coupled for maximum fidelity to the underlying biology); also note that the governing equation for E_M may technically be omitted, as it is uncoupled from the rest of the model. Rainfall affects egg survival, $p_E(R)$, per Eq. (60), and larval development hyperbolically per Eq. (22) (Jepson et al. 1947). As above, other temperature-dependent parameters assume the relations of Martens et al. (1995b) given in Sect. 6.1.

The overall model framework is thus similar to the ODE version of Hoshen and Morse's model, as related in Eqs. (52)–(58), but it incorporates rainfall differently, does not directly couple the gonotrophic cycle to oviposition or death, and does not explicitly consider the immature mosquito life stage. Note that, while biting is temperature-dependent, oviposition is not coupled to biting directly.

Parham and Michael (2010) also take recovery refractory to further infection in humans to be an absorbing state, which is unrealistic for malaria in general, but reasonable as an approximation for a single epidemic in a previously unexposed population. Thus, the model is an extension of the ideas contained in the simpler works of Martens et al. and Craig et al. (1999) to design a continuous, delay differential equation SEIR framework that is more amenable to explicit (rigorous) analysis than the original difference equation formulation of Hoshen and Morse (2004), but one that has some minor mathematical infidelities to the underlying biology. Parham and Michael (2010) derived several complex expressions for $\mathcal{R}_0$ under different simplifying assumptions, and arrived at their key prediction: *malaria transmission is maximized, both in endemic and epidemic areas, in the 32–33* °C *temperature range.*

These authors and several colleagues also published a 2012 work (Parham et al. 2012) that considered the vector lifecycle in detail, developing a hydrodynamics model relating rainfall to habitat volume. This allows a calculation of immature mosquito density, and hence several detailed expressions for density-dependent oviposition and density-dependent larval mortality.

6.2.3 Alonso et al. model (2011)

Alonso et al. (2011) developed another ordinary differential equations based model for a highland tea plantation in Kenya. Larval maturity rate and daily mosquito survival follow the monotonic Jepson (Eq. (22)) and Martens et al. relations (Sect. 6.1), respectively, as in the prior models, with sporogonic duration also hyperbolically

related to temperature. The model was novel compared to earlier works for including symptomatic and asymptomatic carriers of infection, superinfection of asymptomatic carriers to the symptomatic state, and clinical treatment of symptomatic sufferers. These authors also introduce a new aquatic stage death term, given as the sum of a constant background, temperature-dependent, and rainfall dependent term:

$$\delta_L = \delta_0 + \delta_L(T) + \delta_L(R). \tag{68}$$

For temperature dependence, $\delta_L(T)$, Alonso et al. (2011) fit a rather complex piecewise defined function to much more modern data than previously used, namely laboratory *A. gambiae* survival under different temperatures, by Bayoh and Lindsay (2004). However, a fourth-order polynomial also provides a satisfactory fit, and this data is reviewed in Sect. 5.2.3. Rainfall dependence is given by the assumed function

$$\delta_L(R) = \delta_R \Theta(R - \langle R \rangle_{12}), \tag{69}$$

where δ_R is a constant, $\Theta(x)$ is x if $x > 0$ and 0 else, and $\langle R \rangle_{12}$ is a 12-month rainfall moving average. This function is intended to capture washout mortality from heavy rains. As in other works, the biting rate is assumed to be equal to the inverse of the gonotrophic duration, which, in a departure from the Moshkovsky formula (Eq. (18)), is determined from data collected by Afrane et al. (2005) as

$$t_a = \frac{1}{0.091678T - 1.7982}, \tag{70}$$

although this preserves the same basic hyperbolic relationship between temperature and gonotrophy. Finally, it is assumed that each oviposition event yields 66 eggs on average.

The application of this model is rather novel, compared to much of the literature, in that it has as its relatively narrow focus a malarious tea plantation under fairly constant conditions over multiple decades, which experienced rising temperatures since the 1970s, and for which detailed time-series data of clinical malaria cases was available dating from that time. The model was trained using data up to 1985, and was then used to generate counterfactual time-series for more recent malaria burden (with confidence intervals), one where the observed warming did occur, and another where temperature in the 1990s remained "similar" to the 1970s. The model clearly predicted that actually observed cases (which were in the range projected with warming), would have been very unlikely without warming, supporting a clear role for temperature in increasing malaria burden in at least one *specific* area.

6.3 Some recent works (2013–2017)

6.3.1 Beck-Johnson et al. (2013) model

Beck-Johnson et al. (2013) also developed a delay-differential model for immature mosquito development through egg, larva, and pupa, with development time inversely

proportional to temperature (according to a power law, based on Lardeux et al. (2008)), but with survival at each immature (and adult) stage, determined according to a Gaussian function of temperature, such that survival is maximized at intermediate temperatures. The mathematical difficulties of non-constant delays in development of each stage were resolved by re-scaling the model to physiologic, instead of chronological, time, and density-dependent larval mortality was also included.

Anticipating Mordecai et al. (2013) somewhat, the model suggested that adult mosquito prevalence is maximized around 20–30 °C, while the *potentially infectious* mosquito population (i.e. that fraction of the adult population that survives long enough for at least one temperature-dependent sporogonic cycle to elapse), is maximized at a slightly higher range (24–30 °C), but both populations abruptly decline beyond 32 °C.

6.3.2 The Mordecai et al. response (2013)

In 2013, Mordecai et al. (2013) aggregated an updated collection of data to derive functions relating vector and parasite parameters to temperature that were uniformly *unimodal* (i.e. with a peak at some intermediate temperate), rather than *monotonic*, as in most of the previous works discussed. Based on this, and using a formula for $\mathcal{R}_0$ derived partly from Parham and Michael (2010),

$$\mathcal{R}_0 = \left(\frac{a(T)^2 bc(T) \exp(-\mu(T)/PDR(T)) EFD(T) p_{EA}(T) MDR(T)}{Nr\mu(T)^3} \right)^{\frac{1}{2}}, \quad (71)$$

where N is human density, $\mu(T)$ is the adult mosquito death rate, $PDR(T)$ is the sporogonic rate, $MDR(T)$ is the larval development rate, $p_{EA}(T)$ is probability of survival to adulthood, $a(T)$ is the biting rate, and $bc(T)$ is vector competence, these authors concluded that previous works had dramatically overestimated the optimum temperature range for malaria transmission, concluding that transmission is maximized at 25 °C.

It is important to examine in some detail the thermal response functions used by Mordecai et al. (2013) and their sources. All are graphically illustrated in Fig. 16, with data sources summarized in the caption. Functions are given as either a quadratic polynomial, or according to a three-parameter (c, T_m, T_0) unimodal function developed by Briere et al. (1999) to describe arthropod development rates:

$$r(T) = cT(T - T_0)(T_m - T)^{\frac{1}{2}}. \quad (72)$$

One problem with collating these thermal-response curves from different data sources is that the parameters considered are not necessarily independent. For example, larval development is supposed to cease by 34 °C, yet this may represent a conflation of mortality with development: obviously no larvae complete development if temperatures prohibit survival, but the development rate *per se* does not necessarily go to zero. As survival and larval development compound multiplicatively in Mordecai et al.'s expression for $\mathcal{R}_0$, this could bias the optimal temperature range downward. A

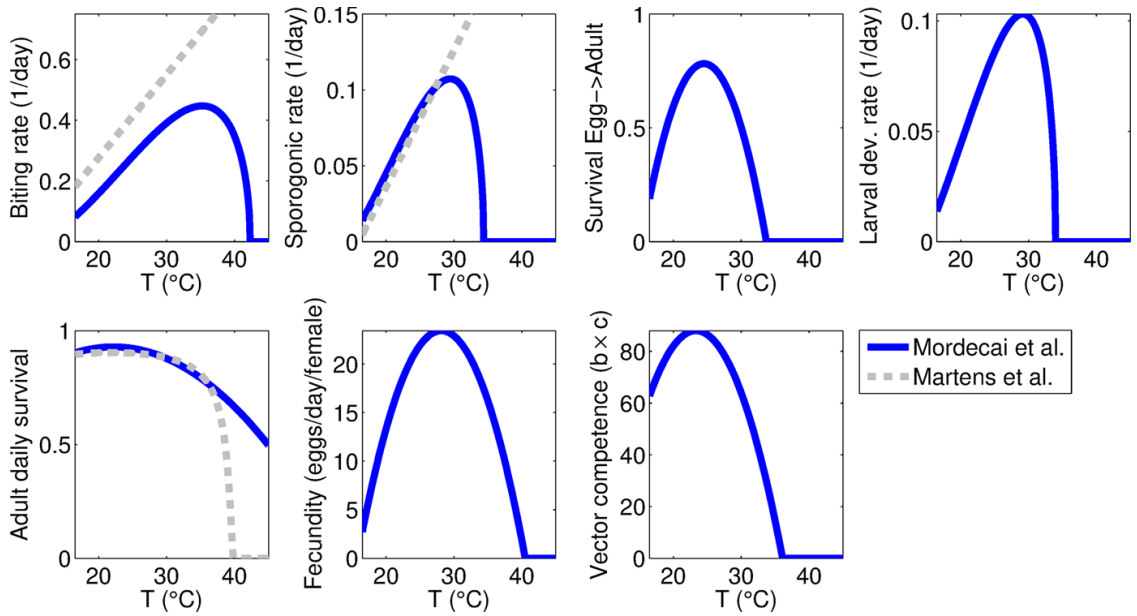


Fig. 16 Unimodal thermal-response curves for vector and parasite parameters used by Mordecai et al. (2013); parameters also used by Martens et al. (1997) are also shown in gray. Egg-to-adult survivorship and development rates were fit to Bayoh and Lindsay (2003), adult mortality was estimated from Bayoh (2001), assuming exponentially distributed survival times, and bite rate was estimated from gonotrophic cycle duration data by Lardeux et al. (2008). Vector competence, $b \times c$, (which *should* be < 1) was estimated from a 1940 work. The rate of the sporogonic cycle was also estimated from several older works, and a newer work by Eling et al. (2001). Fecundity (eggs/female/day) was determined from a work on *Aedes albopictus* (Delatte et al. 2009), a dengue vector, and is obviously not coupled to the biting rate

similar argument applies to sporogonic cycle duration, and several other parameters, such as biting rate and eggs laid per day, are also not independent. This basic problem is not unique to the 2013 Mordecai et al. work, and applies deeply to the foundational Ross–Macdonald models.

Lunde et al. (2013a), using an ODE model for malaria transmission, compared six different temperature-dependent adult mosquito death rates, including that of Martens et al. (1997), Ermert et al. (2011a), Parham et al. (2012), Mordecai et al. (2013), and a separate work by Lunde et al. (2013b), all of which, except those due to Martens et al., were determined from data by Bayoh (2001), and all, except an early polynomial also due to Martens et al., similarly suggested transmission to be optimal in the 24–27 °C range.

6.3.3 Ryan et al. (2015): Malaria mapping under Mordecai's curves

In an interesting follow-up, Ryan et al. (2015b) developed a series of malaria potential maps across Africa under current conditions and under projected warming through 2080 (under a mid-range emissions scenario, SRES A1), using the thermal-response functions and $\mathcal{R}_0$ expression of Mordecai et al. (2013) (Equation 71). This work was novel in that it considered both year-round and seasonal malaria potential, and considered the populations, not just land area, at risk. Furthermore, although this work focused on temperature only, while many prior similar mapping efforts have filtered results geographically by applying a minimum rainfall threshold for malaria transmission (e.g. Craig et al. (1999)), Ryan et al. (2015b) applied an aridity mask

on the basis of the normalized Normalized Difference Vegetation Index (NDVI), an index determined from satellite measurements of the radiation reflected by a surface, and calculated from the difference in reflectance in the visible light and near-infrared spectrum bands. Values above 0.2 quantify vegetation greenness (Baeza et al. 2011), and at mid- and low latitudes, NDVI correlates with wetness (Suzuki et al. 2006); NDVI has an advantage over rainfall in that it may capture areas of low rainfall still suitable for anophelines due to irrigation, rivers, or other permanent water sources (Baeza et al. 2011; Ryan et al. 2015b).

Ryan et al. (2015b) predicted that, as the globe warms, the areas most suitable for intense, year-round transmission will shift southeasterly from western coastal Africa, with the new peak in transmission potential importantly centered around heavily populated areas such as western Uganda, northern Tanzania, the Lake Victoria region near the Kenyan highlands, and highland Madagascar. They also concluded that, while the total area at any risk for malaria may increase slightly, the area at the highest risk will drop. Of note, this work did not consider daily temperature variations (see Gething et al. (2011) for an example of a mapping effort which does this), and did not consider more detailed hydrodynamics.

6.3.4 Synthesis models with partial immunity and age-structure

Recently, Agusto et al. (2015) proposed a fairly comprehensive ODE model that includes the larval vector stage, with density-limited growth (via a logistic growth term), adapts the temperature-dependent lifecycle parameters from Mordecai et al. (2013), and extended a basic SEIR module for human infection to include three additional recovered states, each representing a boost in immunity (which is then lost to lower recovered echelons and then to the base susceptible compartment via first-order kinetics), based on Niger and Gumel (2008). Additionally, the model considered seasonally varying temperature profiles for different regions of sub-Saharan Africa. They predicted malaria burden (as measured in terms of the total number of new cases of infection) to increase with temperature in the range 16–28 °C, but to decrease for temperature values above 28°C in West Africa, 27 °C in Central Africa, 26 °C in East Africa and 25 °C in South Africa. They also found that omitting either immature mosquito dynamics or temperature variability could significantly affect predictions, but the immunity-boosting module had little effect on infection incidence.

A similar effort by Okuneye and Gumel (2017) included partial resistance after infection, but also subdivided the human population into those under five (which bear, by far, the brunt of disease) and those over, and considered rainfall (adopted from Parham and Michael (2010)) in addition to temperature. This work found, for the Kwa-Zulu Natal province of South Africa, an increase in malaria burden with increasing mean monthly temperature and rainfall in the ranges 17–25 °C and 32–110 mm, respectively, and that malaria transmission is maximized for mean monthly temperature and rainfall in the ranges 21–25 °C and 95–125 mm. This model demonstrated dynamics only marginally affected by the inclusion of immunity and age-structure, in terms of infection incidence. While the notion that infection *per se* is minimally affected by age-structure and immunity is reasonably congruent with epidemiologic data, e.g. Trape et al. (1994), elucidating the effect of climate on serious clinical

disease incidence (i.e., that disease which is severe enough to present clinically, as opposed to simply the acquisition of new infection, which may be asymptomatic among immunes) should be a future modeling goal. It should also be noted that, as in Mordecai et al. (2013), these works treat some interdependent processes as independent, e.g. the temperature-dependent biting and oviposition rates.

A parallel series of works by Yamana, Bomblies, and colleagues (Bomblies et al. 2008; Yamana et al. 2013, 2016), founded upon the agent-based HYDREMATS model developed by Bomblies et al. (2008) incorporated detailed hydrodynamic modeling at the village scale, and in a 2013 work (Yamana et al. 2013), the model framework was extended to include the gradual acquisition of partial immunity in humans, whereby repeated infections both reduced the probability of infection and increased the infection clearing rate, and concluded that, via immunity, large differences in infectious biting rate did not, in two highly malarious villages, translate into comparable differences in infection. This framework was also applied in a later (Yamana et al. 2016) model-driven effort that concluded climate change is unlikely to appreciably affect malaria burden in Western Africa.

6.3.5 Christiansen-Jucht model (2015) and age-dependent survival

It has frequently been observed that neither larval nor adult survival times are exponentially distributed (see Sects. 5.2.3 and 5.2.2), and this fact was incorporated into an age-structured vector lifecycle model by Lunde et al. (2013b), but the formulation was rather complex; note that delay-differential models and those with multiple age compartments (e.g., for larvae) also yield at least some non-exponential survival times.

We focus here on a more recent and easily digestible effort by Christiansen-Jucht et al. (2015) that included age-dependent survival at both the larval and adult stages (in addition to temperature-dependent survival), with data and model linked via the gamma distribution. The model entails an ODE framework with larval and adult populations divided into multiple age categories, with first-order transitions through categories and the final category terminating in death, as depicted in Fig. 17. We subdivide into four model combinations: (1) Baseline model without age-dependent survival, (2) Baseline + Larval age-dependent survival, (3) Baseline + Adult age-dependent survival, and (4) Larval and adult age-dependence.

Fortuitously, mosquito (and other age-dependent) survival data can be reasonably well-described by the two-parameter gamma distribution, with α the shape parameter and β the rate parameter, and it so happens that an ODE with α age compartments that are traversed at rate β yields an overall gamma-distributed survival time (Wearing et al. 2005) (obviously α must be an integer for this to hold); this basic fact also facilitates ODE modeling of other non-Poisson time-dependent processes, such as infection clearance (Wearing et al. 2005). Note that if $\alpha = 1$, then we reduce to exponentially distributed survival times. If $y_i(t)$ is the number of mosquitoes in compartment i, we have simply

$$\frac{dy_1}{dt} = -\beta y_1, \tag{73}$$

$$\frac{dy_i}{dt} = \beta y_{i-1} - \beta y_i, \tag{74}$$

and the total number of surviving mosquitoes at any time is

$$\sum_{i=1}^{\alpha} y_i(t). \tag{75}$$

An example of a single cohort of adult mosquitoes is given in Fig. 18, using $\alpha = 7$ and β as fitted to adult survival data at 20 and 30 °C extracted from Bayoh (2001).

Now, using the framework given in Fig. 17, α and temperature-dependent β parameters were fit for larvae and adults ($\alpha = 7$ for larvae, $\alpha = 3$ for adults), other temperature-dependent survival and larval development rates were adapted from Parham et al. (2012), and larval density-dependent death with carrying capacity a function of rainfall as adapted from White et al. (2011), Christiansen-Jucht et al. (2015) compared the ability of the four model permutations to fit mosquito abundance data from The Gambia, finding that Models 3 and 4 (adult age-dependent survival with and without larval age-dependent survival, respectively) gave nearly identical fits that were superior to either Model 1 or 2 (baseline or larval age-dependence only), suggesting that non-exponential death rates in adult mosquitoes are important for model fidelity to data.

6.3.6 Temperature variability and extreme weather

While most modeling works have considered ambient temperature as a single constant, in the last few years there has been increasing experimental and theoretical interest in the role of temperature variability, especially diurnal variation, on vector survival

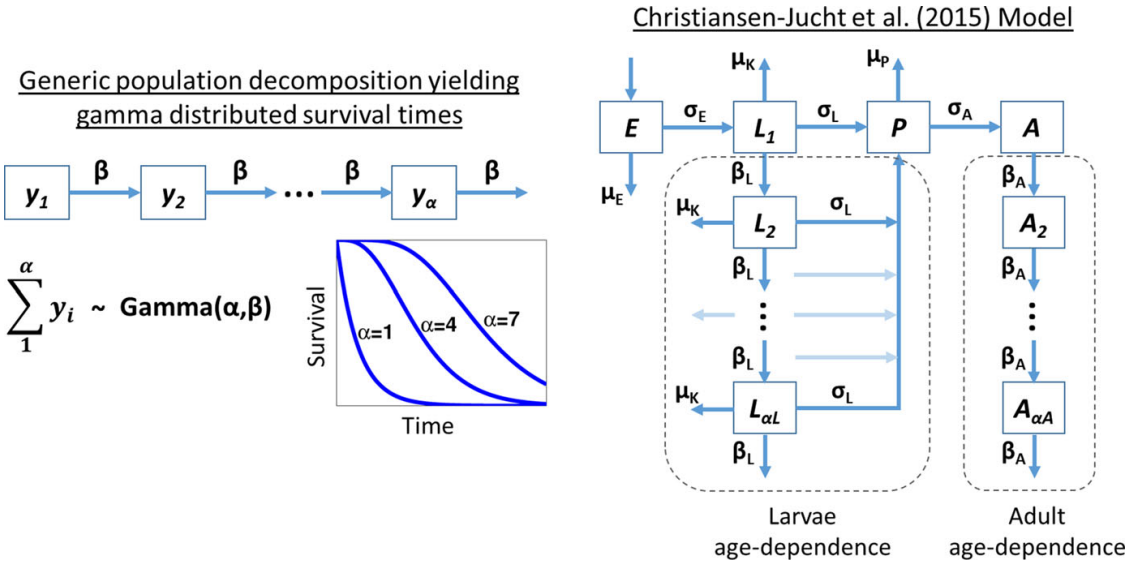


Fig. 17 Generic technique for modeling age-dependent events in any population (left) and 2015 Christiansen-Jucht et al. model framework (right) employing this technique for the *A. gambiae* lifecycle, with age-dependent survival in both larval and adult populations

and development and malaria transmission potential (Paaijmans et al. 2009, 2010; Gething et al. 2011; Paaijmans et al. 2013b; Blanford et al. 2013; Lyons et al. 2013; Murdock et al. 2016; Beck-Johnson et al. 2017). For any given mean temperature point, survival and development rates do not generally change symmetrically with temperature excursions in either direction (given the non-linear nature of the related thermal-response functions), and thus we may expect exposure to fluctuating temperatures (as experienced in the field) to have a fundamentally different effect on vectors and parasites than exposure to a constant temperature.

This was first addressed in a theoretical context by Paaijmans et al. (2009), who showed that, using a unimodal Briere function (see Eq. (72)) for the *Plasmodium* sporogonic duration, fluctuations about relatively low temperatures enhanced development and hence malaria potential, relative to a constant temperature, while fluctuations at higher temperatures had the opposite effect, suggesting that most existing theoretical works may have systematically under- and overestimated malaria potential at cool and high temperatures, respectively. This pattern was subsequently observed experimentally in a rodent malaria model (Paaijmans et al. 2010), and Blanford et al. (2013) scaled these results up geographically to Kenya, predicting the cooler highlands to be relatively more vulnerable to malaria once diurnal temperature variation is accounted for; similar predictions were made at a continental scale as well. Paaijmans et al. (2013b) also found temperature fluctuations to enhance and inhibit *A. stephensi* development and survival at low and high temperatures, respectively, and similar results have been obtained for other ectotherms, e.g. Bayu et al. (2017).

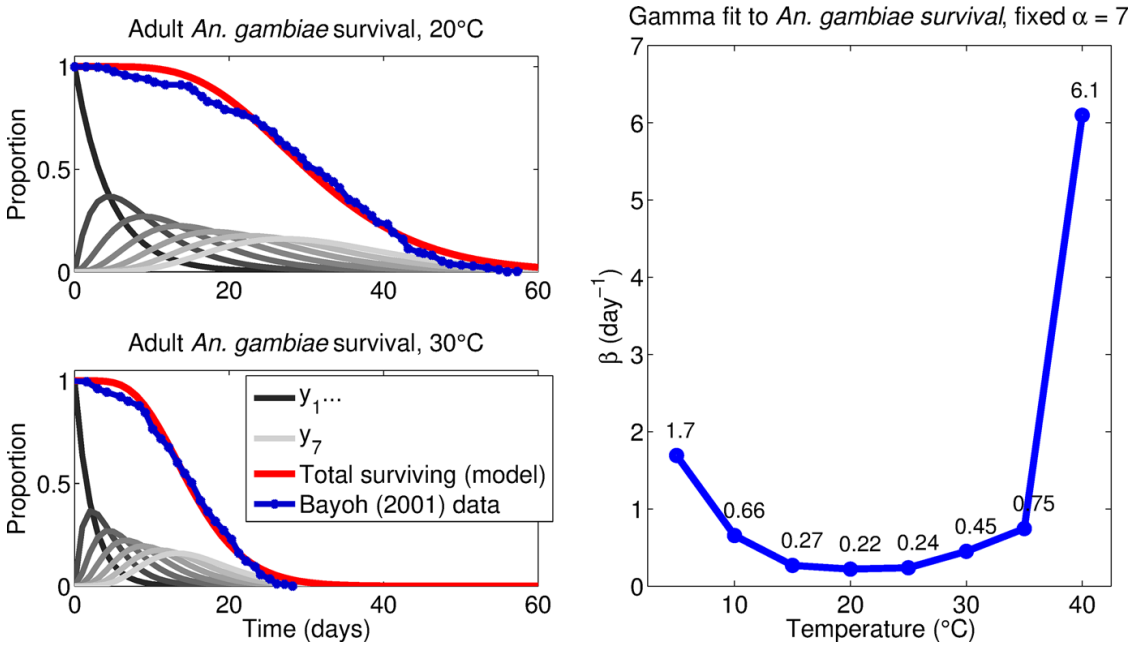


Fig. 18 Age-dependent death modeled using a multi-compartment model with α compartments and transition rate β, yielding a Gamma (α,β) distributed survival curve. A seven-compartment model, i.e. $\alpha = 7$, with β fit to adult *A. gambiae* survival data from Bayoh (2001) at different temperatures. The left panels show how the distribution among age-compartments shifts over time, co-plotted with overall model survival and data, using β for 20 and 30 °C. The right panel gives β as a function of temperature across the full data's full temperature range

The aforementioned works focused either on the sporgonic cycle (Paaijmans et al. 2009, 2010; Blanford et al. 2013) or *Anopheles* development (Paaijmans et al. 2013b) in isolation, but very recently, Beck-Johnson et al. (2017) explored daily and annual temperature variations under their previously discussed 2013 model (Beck-Johnson et al. 2013) (Sect. 6.3.1). This work suggests more subtle effects of temperature variation on both total adult mosquito and potentially infectious adult mosquito populations. Of particular interest, increasing the daily temperature range narrowed the mean temperature range over which both such populations could exist, by decreasing mosquito abundance at *both* higher and lower temperatures. That is, while the works above, which focused on a single aspect of the malaria transmission cycle suggested that temperature fluctuations asymmetrically favor transmission at low temperatures, the more complete model of Beck-Johnson et al. (2017) contradicts this notion to some degree. However, further complicating the picture, greater *annual* temperature variation tended to increase both the range of the infectious mosquito population (i.e. there were more times with both fewer and greater numbers of mosquitoes) and the mean at lower temperatures, while also decreasing the infectious population at the high temperature range. Thus, temperature fluctuations at different scales may affect malaria transmission in a variety of subtle ways that are not predictable from isolated components of the malaria lifecycle.

Several other models have incorporated seasonal temperature variability, e.g. Agusto et al. (2015), although this was not their primary focus. Also of note, Gething et al. (2011) using a Ross–Macdonald-style expression for vector potential, developed global maps for malaria suitability by temperature that was novel in that it superimposed diurnal sinusoidal temperature variations onto monthly temperature trends drawn from the WorldClim database, while a similar effort by Garske et al. (2013) inferred air temperatures from land temperature data and also imposed diurnal temperature variation on a Ross–Macdonald-like formulation for malaria potential. We suggest that comparing the predicted malaria maps using more complex models with and without such temperature variations would be a valuable contribution to the literature.

It should be mentioned that the works reviewed here have generally considered ambient air temperature only. Yet, as already discussed, the temperature in aquatic anopheline habitats may differ appreciably from the air, and, depending upon the habitat size and thermal stability, temperature variations in water may be smaller or greater than in air, and how this complexity might further alter the predicted role of temperature fluctuations, both current and as anticipated under climate change, is an open question. Notably, the work by Blanford et al. (2013) did consider indoor temperatures, which tend to be slightly higher but less variable than ambient (Singh et al. 2016; Blanford et al. 2013), and the indoor/outdoor temperature difference has been almost uniformly neglected in modeling works (but see Singh et al. (2016) for a recent exception), despite the indoor preference seen in many anophelines.

Finally, climate change is likely to increase weather extremes, including drought, extreme rainfall events, and heat waves, which may be expected to affect the lifecycles of a range of ectotherms, independent of average weather conditions (Ma et al. 2015). Southern Africa, especially, may be vulnerable to greater frequency of extreme rainfall events (Engelbrecht et al. 2013). While a potential increase in immature *Anopheles* mortality due to heavy rainfall is sometimes accounted for (Paaijmans et al. 2007)

to our knowledge no mathematical work has more explicitly examined how weather extremes under climate change, and especially extreme high temperatures, might affect malaria epidemiology.

6.4 Towards a meta-population model for the Kenyan highlands

Recall that the Kenyan highlands have seen malaria incidence increase since the 1970s in conjunction with increasing temperatures and broad changes in the populace, including rapid population growth and deforestation, making this a model region for studying the impact of current and projected climate change on malaria transmission (Minakawa et al. 1999; Pascual et al. 2006; Chaves and Koenraadt 2010). Human mobility can link regions with varying local transmission dynamics (e.g. as a result of climate), and while no climate-focused models have accounted for this thus far, we present basic mathematical frameworks by which this might be studied in the future. Any geography may be conceptually divided into multiple *patches*, each with its own sub-model describing local disease dynamics, and with movement of hosts and/or vectors between patches occurring at prescribed rates; thus movement of *Plasmodium* reservoirs from one patch to another may spread disease. A natural division is to consider highland and lowland areas as two separate patches, but any two-patch model easily generalizes to an n-patch model.

It is necessary to distinguish between two forms of mobility: (1) migration, representing permanent resettling in a new area, and (2) visitation, or transient excursions with no permanent change of address (Mandal et al. 2011). Both are salient, as large-scale migration into the Kenyan highlands has occurred over the past few decades (Minakawa et al. 1999), while small-scale and circulatory movements between communities are also common. Indeed, transient mobility also interacts with age because, in many parts of rural Africa, mothers from rural areas tend to take their infants and young children (not of school-going age) on a short trip (usually for a day or two) to conduct businesses at open markets in neighbouring urban communities, thereby exposing them to malaria infection, especially if this trip is from highland areas of low endemicity to lowland areas of high malaria burden. In a multi-patch model, permanent migration is represented, in analogy to fluid dynamics, by the classical Eulerian approach, whereas transient mobility with a home patch is captured by the Lagrangian approach (Castillo-Chavez et al. 2016).

Mathematically, Eulerian migration resembles a diffusion process, and we define k_{ji} to be the first-order rate-constant for movement from patch j to i. A very simple Ross-style model with Eulerian migration among human hosts, with H_i the total human population and X_i the infected population in patch i (and M_i and Z_i the total and infected mosquito populations), is

$$\frac{dH_i}{dt} = \sum_{j=1, j\neq i}^{\Phi} k_{ji} H_j - \sum_{j=1, j\neq i}^{\Phi} k_{ij} H_i, \tag{76}$$

$$\frac{dX_i}{dt} = ab\left(\frac{Z_i}{H_i}\right)(H_i - X_i) - rX_i + \sum_{j=1, j\neq i}^{\Phi} k_{ji}X_j - \sum_{j=1, j\neq i}^{\Phi} k_{ij}X_i, \quad (77)$$

$$\frac{dZ_i}{dt} = ac\left(\frac{X_i}{H_i}\right)(M_i - Z_i) - gZ_i, \quad (78)$$

where Φ is the total number of patches. In other words, it is simply the standard model augmented by migration terms among the hosts. The Lagrangian formulation is slightly less obvious. We have p_{ij} as the fraction of time that individuals from patch i spend in patch j, subject to the constraint that $\sum_{j=1}^{\Phi} p_{ij} = 1$. We also have that the *effective* population of patch i is $\sum_{k=1}^{\Phi} p_{ki}H_k$. A Ross model with Lagrangian motility, between an arbitrary number of patches, is therefore given by

$$\frac{dX_i}{dt} = ab\sum_{j=1}^{\Phi}\left(\frac{p_{ij}Z_j}{\sum_{k=1}^{\Phi}(p_{kj}H_k)}\right)(H_i - X_i) - rX_i, \quad (79)$$

$$\frac{dZ_i}{dt} = ac\sum_{j=1}^{\Phi}\left(\frac{p_{ji}X_j}{p_{ji}H_j}\right)(M_i - Z_i) - gZ_i. \quad (80)$$

Note that hosts do not move between home patches, but the effective population of each patch is modulated by the time fraction every other patch population spends within it. The two mobility modes can easily be combined into a single model, essentially by augmenting Eqs. (79)–(80) with the Eulerian mobility terms.

Multiple authors have studied multi-patch Ross–Macdonald style models (Torres-Sorando and Rodríguez 1997; Auger et al. 2008; Cosner et al. 2009; Prosper et al. 2012; Ruktanonchai et al. 2016), beginning with Eulerian mobility in Torres-Sorando and Rodríguez (1997) and Lagrangan mobility in Cosner et al. (2009); see also Agusto (2014) for a brief review. Agusto (2014) recently studied a more complex (but weather-independent) multi-patch malaria model that also included the vector lifecycle, and moreover, focused on the spread of drug-sensitive and drug-resistance *Plasmodium* strains under Eulerian migration; to our knowledge, mobility has not been incorporated into a weather-driven model using the patch framework, and we suggest it as an important extension of current models. Additionally, we offer our speculation that a multi-patch model at very local scales with mobility among vectors, in addition to (or in lieu of) hosts, could help elucidate how the presence of varied microenvironments across a fine spatial scale might alter transmission dynamics.

7 Other modeling challenges: immunity, treatment, within-host disease, and other abiotic factors

In this section, we briefly touch on some other malaria modeling challenges and traditions, including immunity and within-host dynamics, treatment, resistance, and socioeconomic factors. While there is insufficient space do any of these topics justice, we believe that incorporating many of these aspects of disease in climate-driven model frameworks is a fundamental challenge for the future, and hope to at least make the reader aware of these issues and some useful references.

7.1 Malaria immunity

Numerous studies have attempted to quantify the acquisition of protection against clinical malaria with repeated infection, beginning as early as the Garki model of Dietz et al. (1974) (Sect. 4.4), and include (Aron 1983, 1988; Gupta and Day 1994; Gupta et al. 1999a, b; Filipe et al. 2007; Griffin et al. 2010, 2015). Filipe et al. (2007), for example, concluded that a short-term, primarily clinical, immunity strongly coupled to cumulative exposures and a longer-term, primarily anti-parasitic, immunity more weakly coupled to exposure are both necessary to best fit epidemiologic data. These authors modeled infection using an age-structured model, where exposed persons either manifest clinical disease or asymptomatic disease, with the probability of clinical disease modulated by an underlying level of immunity, I_s, which increases in direct proportion to EIR, and decays exponentially with a decay half-life of about 5 years. The probability of clinical (versus asymptomatic) disease, ϕ, decreases according to a sigmoid function of I_s. Additionally, asymptomatic infection is assumed to clear at an exponential rate that increases with age, given that some exposure has occurred, but otherwise independent of cumulative exposure; the half-life of this parasitic immunity was determined to be at least 20 years. A similar approach was also used recently by Griffin et al. (2010) and Griffin et al. (2015).

This model-based conclusions is concordance with clinical data per Rodriguez-Barraquer et al. (2016), who performed an analysis of detailed longitudinal data of 93 children over the first 5 years of life in a holoendemic region of Uganda, who underwent surveillance for both clinical disease and (microscopy-detected) asymptomatic parasitemia. Infection was modeled essentially as a Poisson process, with individual infection hazard varying according to a Gamma distribution.

This study suggested a binary effect of age, with increasing age a risk factor for infection, but also independently protective against malaria, given infection (regardless of infection history). The models suggested a 6% decrease in malaria given infection per year of age, and a 2% decrease in malaria per infection. Note, however, that there were 5.33 episodes of malaria per person-year (and 0.588 asymptomatic parasitemias per person-year), suggesting that infectious history was dominant over age in conferring protection. A further finding was that clinical immunity developed faster in children treated with artemether-lumefantrine (AL) versus dihydroartemisinin-piperaquine (DP). The latter drug confers longer-lasting post-treatment protection against infection.

As already alluded to multiple times, given its apparent historical and epidemiologic importance, further mathematical investigations incorporating immunity with climate factors are essential. Towards that end, several recent climate-focused works have included relatively simple representations of partial immunity (Agusto et al. 2015; Yamana et al. 2013, 2016, 2017), although they have not generally considered the disparity between anti-disease and anti-parasite immunity. In particular, work by Yamana et al. (2013, 2017) represents important steps towards understanding the interplay between environment and immunity.

7.2 Treatment, control measures, and resistance

There was a large drop in the African malaria burden in the 1960s and 70s, largely attributable to widespread pharmacologic treatment with chloroquine, but this was undermined by the evolution and spread of chloroquine resistance *P. falciparum* strains (Carter and Mendis 2002). Artemisinin compounds are now highly effective, but resistance has already been detected in southeast Asia, and it seems that the basic biology dictates that it is only a matter of time before resistance reaches Africa (Webb 2014). A wide variety of mathematical models have addressed the evolution of drug resistance in both infectious disease and cancer. Some works focused on modeling drug resistance in malaria include (Hastings 1997, 2003; Yeung et al. 2004; Pongtavornpinyo et al. 2009; Saralamba et al. 2011; Agusto 2014; Forouzannia and Gumel 2015), and Hastings and Watkins (2005) provide an excellent review of the main concepts involved in modeling resistance. These are only a few examples from the literature, but any deeper review of these works is unfortunately beyond our scope.

We also note that numerous models have focused on other control efforts, such as ITNs, (most classically we have Macdonald's prediction that adult mosquito survival should be targeted, motivating insecticide-based control), with some recent efforts including Smith et al. (2009), Griffin et al. (2010) and Nikolov et al. (2016), and finally, a recent and broad review of the evolutionary principles underlying resistance is given by Huijben and Paaijmans (2017).

7.3 Within-host disease dynamics

Coupling an explicit model for the within-host dynamics of malaria (principally the within-host immune response) to its epidemiology is a fundamental challenge. Indeed, even describing the within-host dynamics mathematically in a way the reproduces the qualitative dynamics of long-term infection has proven most difficult, and this issue has its own extensive literature that it is beyond the scope of this paper, although a very partial reference list includes Teboh-Ewungkem et al. (2010), Li et al. (2011), Saralamba et al. (2011), Gurarie et al. (2012), Eckhoff (2012), Demasse and Ducrot (2013), Childs and Buckee (2015), Childs and Prosper (2017), Tabo et al. (2017), and a recent work by Childs and Buckee (2015) highlights the problems of accurately modeling within-host dynamics in some depth. Since these dynamics may interact in unexpected ways with malaria epidemiology (Childs and Buckee 2015), a complete

understanding of climate and malaria will likely necessitate a deeper consideration of the in-host stages than has heretofore been attempted.

7.4 Other abiotic factors

Broad changes in socioeconomic conditions, e.g. widespread urbanization and increasing material standards of living, and agricultural modernization played fundamental roles in the retreat of malaria from most of the world outside tropical Africa (Webb 2014; Packard 2007). Socioeconomic conditions also determine access to medical treatment and effective antimalarials (Webb 2014), and rural populations suffer a higher malarial burden than do urban populations (Rodriguez-Barraquer et al. 2016), partly because urban land-use patterns support fewer vectors and partly because control efforts have historically focused on urban over rural areas (Webb 2014). Land-use changes, driven by social or economic imperatives, strongly affect anopheline habitat, e.g. deforestation in the Kenyan highlands (Afrane et al. 2005). Incorporating social factors directly into models is a challenge, but can probably be expected to affect biophysical parameters in somewhat predictable downstream ways; see also Mandal et al. (2011) for a brief review of models relating socioeconomics to malaria.

8 Conclusions

Of the historical human *Plasmodia*, *P. falciparum* is evolutionarily distant from the others (Silva et al. 2015), and uniquely virulent. While in the West, the so-called diseases of civilization are generally thought of as those cardiovascular ailments, diabetes, and obesity that tend to accompany high-energy diets and sedentary lifestyles, *P. falciparum* malaria was perhaps the first true disease of civilization, ushered in by global warming, anthropogenic alteration of the environment, and concentrated human settlement. Malaria has been subject, more than most diseases, to mathematical investigations, and these efforts, mainly the pioneering works of Ross and Macdonald, had real influence on malaria control efforts through the twentieth century. It is likely, then, that the mathematician can play a central role in informing an understanding and mitigation of global warming's future impact of malaria. Malaria, however, is a complex disease with its own particular history, and so we take the view that mathematical efforts are best informed by history, and to that end have attempted a reasonably thorough historical review of the disease.

Globally, malaria retreated dramatically over the course of the twentieth century, most markedly in post-World War II Southeast Asia (Carter and Mendis 2002). This drop occurred in the face of very modest global warming (about 0.6 °C) (IPCC 2013; Gething et al. 2010), and is almost certainly attributable to a variety of non-climatic changes, especially economic and agricultural modernization, urbanization, and broad increases in population health and health services (Chaves and Koenraadt 2010; Webb 2014). While this would seem to suggest climate change as a minor, at best, factor in future transmission scenarios, the consensus view of the IPCC is that risks from climate change are not likely to be strongly felt until the global temperature anomaly exceeds at least 1 °C, with impacts increasing dramatically above 2 °C (IPCC 2014);

the temperature anomaly is virtually guaranteed to reach 1.5 °C by the end of the twenty-first century, and may well exceed 5 °C (IPCC 2014).

Tropical Africa never saw the deep reductions in malaria mortality that other regions did and, while it enjoyed a transient drop in deaths in the 1960s and 70s, malaria resurged in the later 1970s, an era corresponding with widespread chloroquine resistance, other broad socioeconomic changes (Carter and Mendis 2002; Webb 2014), and the acceleration of global warming (IPCC 2014). While recent control efforts have been reasonably successful, with malaria mortality in Africa now at an historical nadir, their sustainability is threatened on multiple fronts (see Sect. 3.3), and future temperature increases in Africa are likely to exceed the global mean (Niang et al. 2014), where global warming may negatively affect agricultural output, food security, economic development and overall health, and may displace populations (Niang et al. 2014), all developments likely to increase populations' vulnerability to malaria. Thus, it seems likely that future warming will interact with malaria (and other infectious disease) in a nonlinear manner. Therefore, we must be cautious in extrapolating from past climate and malaria trends to the future, and a mechanistic framework for the disease and climate, based on the many excellent works reviewed in this paper, may help to guide us.

There are perhaps five major challenges to mechanistically modeling the relationship between climate change and malaria: (1) thermal-response functions linking temperature and vector/lifecycle parameters; (2) the relationship between weather (mainly rainfall) and anopheline habitat availability, and further, the effect of habitat size on the vector; (3) temperature variability, at both a diurnal and seasonal scale, and in the different microenvironments involved in the *Anopheles* lifecycle; (4) the incorporation of essential non-climatic factors, especially malaria immunity, but also treatment and other control interventions, host mobility across zones of varying endemicities, resistance, and broad socioeconomic factors; and finally (5) basic model construction, i.e. the general choice of biologic actors and their interactions. Most controversy has more explicitly centered on the first two, especially thermal-response functions, and to that end we have presented a detailed summary of experimental data to inform the modeler (Sect. 5), as well as a partial genealogy of the work informing this controversy. Box 1 provides a summary of some of the most important, at least in our view, modeling challenges moving forward.

Box 1: Major climate-related future modeling challenges

We suggest the following as the most pressing future modeling issues with respect to climate and malaria:

- Fully defining the effects of different thermal response functions on vector and parasite ecology and malaria epidemiology under climate change, at both the gross qualitative scale (e.g. monotonic vs. unimodal) and under smaller quantitative variations representing local variations due to adaption or short-term evolution.
- Refining (and enhancing the realism of) the mathematical description of habitat availability, rainfall, and the effects of these factors upon immature anophelines.
- A full accounting of temperature variability, both at a diurnal and seasonal scale, and across the microenvironments to which anophelines are regularly exposed (aquatic, outdoor, and indoor environments). Accounting for disparities between air and water temperatures may be of particular import.
- Assessing the combined impact of host (and vector) mobility and climate change (as in the East African highlands), as well as the interaction of climate with large-scale local population growth (both urban and rural).
- A hybridization of two modeling traditions, namely the climate-focused literature reviewed here, and the rich tradition focusing upon the unique immunology of malaria.
- Finally, and most broadly, an exploration of the interaction between climate and the myriad other factors influencing this disease, including treatment and other control interventions, resistance, and changing socioeconomic conditions.

While the earlier works of Martens and colleagues (e.g. Martens et al. (1999)) suggested a significant global increase in potential malaria burden with global warming, these efforts were informed by thermal-response functions that fail to capture the deleterious effects of high temperatures on both vector and parasite. The unimodal thermal-response functions of Mordecai et al. (2013) suggest a lower temperature range for optimal transmission (25–28 vs. 32–33 °C), but also view some interconnected components of the vector lifecycle as independent, e.g. biting rate and oviposition, and death prior to larval development may be conflated with arrested development. Mechanistic models have properly captured these lifecycle interdependencies to lesser and greater degrees (e.g. Hoshen and Morse 2004; Parham and Michael 2010), and while several later works employing the Mordecai et al. relations in more complex mechanistic models (Agusto et al. 2015; Okuneye and Gumel 2017), have reached conclusions generally congruent with Mordecai et al. (2013), these too view certain dependent processes as independent. Therefore, more careful inclusion of thermal-response functions in mechanistic frameworks should be a goal of future work.

Relating rainfall to the vector lifecycle is less prominent in this modeling tradition, and when it has not been ignored entirely, its presumed effect is more variable and

ad hoc across models. Nevertheless, precipitation patterns are likely to change across Africa with climate change, and rainfall patterns often drive interannual malaria variability (Pascual et al. 2008). More realistic hydrodynamic models are likely to be informative, and we suggest the hydrodynamic models of, for example, Bomblies et al. (2008), Parham et al. (2012) and Asare et al. (2016a) as a starting point; some of these models' basic features and construction are reviewed in Sect. 5.2.5.

Basic model construction, including weather-independent components, is clearly fundamental, with published models encompassing a wide range of biological detail and realism. Mechanistic works have demonstrated that greater mathematical fidelity to the details of the vector lifecycle, e.g. via the inclusion of immature larval stages (Agusto et al. 2015; Beck-Johnson et al. 2013) or age-dependent survival (Christiansen-Jucht et al. 2015), significantly improves model predictions in relation to data. Thus, it is important to elucidate how such deeper model construction choices affect predictions, in conjunction with various thermal- and rainfall-response functions. Furthermore, malaria immunity is fundamental to its epidemiology, and while largely neglected in climate-focused works, there is a large body of immune-focused models, and several recent works have demonstrated that immunity interacts importantly with environment (Yamana et al. 2013, 2017).

Human mobility, both via long-term migration and shorter-term circulation of individuals between areas of low and high malaria burden, may affect malaria incidence, but it is unknown as yet how this mobility dynamic interacts with climate, and we suggest that a multi-patch model that takes into account the detailed lifecycles of vector and parasite, temperature variability and the physics of anopheline habitat, along with the host phenomena of immunity, superinfection, and asymptomatic infection would entail a somewhat unified framework for studying malaria spread from lowland to highland regions. We further suggest that this more limited geographic scope may better elucidate the effect of climate change on malaria transmission than global scale models. Nevertheless, the development of a new family of malaria potential maps, employing a variety of more recent malaria models, temperature variability, and IPCC climate projections, may be highly instructive in determining the robustness of past authors' conclusions and sensitivity to modeling choices. Towards such an end, the recent works of Ryan et al. (2015b) and Yamana et al. (2016) can serve as excellent guides.

Additionally, making the primary focus of malaria mapping upon populations at risk, both current and projected, rather than land area, is likely of import, especially since most population in Africa is currently concentrated in relatively warm Western coastal Africa, where malaria is highly endemic and the effect of climate change upon malaria may be equivocal, and in cooler areas of Eastern Africa, mainly Ethiopia and the region surrounding Lake Victoria, where global warming may be more likely to increase disease potential. Dramatic population growth is projected throughout sub-Saharan Africa, and Nigeria, already the most populous African nation and one with a high malaria burden, may see its population more than double by 2050 (United Nations 2017). Urban environments also are less malarious than the countryside, and so the urban versus rural divide is also important to consider (as done in Ryan et al. (2015b)).

There are multiple modeling traditions within the field of mathematical malariology, and we have been restricted by space to focus primarily on only one; we have very briefly touched on these traditions in Sect. 7, and we also refer the reader to other useful reviews, including Mandal et al. (2011), Smith et al. (2012) and Reiner et al. (2013). Malaria immunity, abiotic factors including land use, and control efforts have all proven fundamental to the epidemiology of the disease, and unifying weather-driven models with these biologic and social phenomena is a fundamental task for the future.

Acknowledgements This work is supported, in part, by the Global Security Initiative of Arizona State University. One of the authors (ABG) is grateful to National Institute for Mathematical and Biological Synthesis (NIMBioS) for funding the Working Group on Climate Change and Vector-borne Diseases (VBDs). NIMBioS is an Institute sponsored by the National Science Foundation, the U.S. Department of Homeland Security, and the U.S. Department of Agriculture through NSF Award #EF-0832858, with additional support from The University of Tennessee, Knoxville. The authors are grateful to the two anonymous reviewers for their very constructive comments, which have significantly enhanced the clarity of the paper. Author SEE is also grateful to Lindsey Van Sambeek for her assistance with Fig. 3.

References

Abdelrazec A, Gumel AB (2017) Mathematical assessment of the role of temperature and rainfall on mosquito population dynamics. J Math Biol 74(6):1351–1395

Afrane YA, Lawson BW, Githeko AK, Yan G (2005) Effects of microclimatic changes caused by land use and land cover on duration of gonotrophic cycles of *Anopheles gambiae* (Diptera: Culicidae) in western Kenya highlands. J Med Entomol 42(6):974–980

Afrane YA, Zhou G, Lawson BW, Githeko AK, Yan G (2007) Life-table analysis of Anopheles arabiensis in western Kenya highlands: effects of land covers on larval and adult survivorship. Am J Trop Med Hyg 77(4):660–666

Afrane YA, Little TJ, Lawson BW, Githeko AK, Yan G (2008) Deforestation and vectorial capacity of *Anopheles gambiae* Giles mosquitoes in malaria transmission, Kenya. Emerg Infect Dis 14(10):1533–1538

Agusto FB (2014) Malaria drug resistance: the impact of human movement and spatial heterogeneity. Bull Math Biol 76(7):1607–1641

Agusto FB, Gumel AB, Parham PE (2015) Qualitative assessment of the role of temperature variations on malaria transmission dynamics. J Biol Syst 23(4):597–630

Allen RG, Pereira LS, Raes D, Smith M (1998) Food and Agriculture Organization of the United Nations. FAO Irrigation and drainage paper No. 56: Crop evapotranspiration (Guidelines for computing crop water requirements). Rome, Italy

Alonso D, Bouma MJ, Pascual M (2011) Epidemic malaria and warmer temperatures in recent decades in an East African highland. Proc R Soc B 278(1712):1661–1669

Antinori S, Galimberti L, Milazzo L, Corbellino M (2012) Biology of human malaria plasmodia including *Plasmodium knowlesi*. Mediterr J Hematol Infect Dis 4(1):2012013

Aron JL (1983) Dynamics of acquired immunity boosted by exposure to infection. Math Biosci 64(2):249–259

Aron JL (1988) Mathematical modelling of immunity to malaria. Math Biosci 90(1):385–396

Asare EO, Tompkins AM, Amekudzi LK, Ermert V (2016) A breeding site model for regional, dynamical malaria simulations evaluated using in situ temporary ponds observations. Geospat Health 11(1s):390

Asare EO, Tompkins AM, Amekudzi LK, Ermert V, Redl R (2016) Mosquito breeding site water temperature observations and simulations towards improved vector-borne disease models for Africa. Geospat Health 11(s1):391

Asare EO, Tompkins AM, Bomblies A (2016) A regional model for malaria vector developmental habitats evaluated using explicit, pond-resolving surface hydrology simulations. PLoS ONE 11(3):e0150626

Auger P, Kouokam E, Sallet G, Tchuente M, Tsanou B (2008) The RossMacdonald model in a patchy environment. Math Biosci 216(2):123–131

Bacaër N (2007) Approximation of the basic reproduction number R_0 for vector-borne diseases with a periodic vector population. Bull Math Biol 69(3):1067–1091
Baeza A, Bouma MJ, Dobson AP, Dhiman R, Srivastava HC, Pascual M (2011) Climate forcing and desert malaria: the effect of irrigation. Malar J 10(1):190
Baton LA, Ranford-Cartwright LC (2005) Spreading the seeds of million-murdering death: metamorphoses of malaria in the mosquito. Trends Parasitol 21(12):573–580
Bayoh MN (2001) Studies on the development and survival of anopheles gambiae sensu stricto at various temperatures and relative humidities. (Doctoral dissertation). Durham theses, Durham University. http://etheses.dur.ac.uk/4952/
Bayoh MN, Lindsay SW (2003) Effect of temperature on the development of the aquatic stages of *Anopheles gambiae* sensu stricto (Diptera: Culicidae). Bull Entomol Res 93(05):375–381
Bayoh MN, Lindsay SW (2004) Temperaturerelated duration of aquatic stages of the Afrotropical malaria vector mosquito *Anopheles gambiae* in the laboratory. Med Vet Entomol 18(2):174–179
Bayu MS, Ullah MS, Takano Y, Gotoh T (2017) Impact of constant versus fluctuating temperatures on the development and life history parameters of *Tetranychus urticae* (Acari: Tetranychidae). Exp Appl Acarol 72(3):205–227
Beck-Johnson LM, Nelson WA, Paaijmans KP, Read AF, Thomas MB, Bjørnstad ON (2013) The effect of temperature on Anopheles mosquito population dynamics and the potential for malaria transmission. PLoS ONE 8(11):e79276
Beck-Johnson LM, Nelson WA, Paaijmans KP, Read AF, Thomas MB, Bjørnstad ON (2017) The importance of temperature fluctuations in understanding mosquito population dynamics and malaria risk. R Soc Open Sci 4(3):160969
Bhatt S, Weiss DJ, Cameron E, Bisanzio D, Mappin B, Dalrymple U et al (2015) The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. Nature 526(7572):207–211
Blanford JI, Blanford S, Crane RG, Mann ME, Paaijmans KP, Schreiber KV, Thomas MB (2013) Implications of temperature variation for malaria parasite development across Africa. Sci Rep 3:1300
Bockarie MJ, Gbakima AA, Barnish G (1999) It all began with Ronald Ross: 100 years of malaria research and control in Sierra Leone (1899–1999). Ann Trop Med Parasitol 93(3):213–224
Bomblies A (2012) Modeling the role of rainfall patterns in seasonal malaria transmission. Clim Change 112(3–4):673–685
Bomblies A, Duchemin JB, Eltahir EA (2008) Hydrology of malaria: model development and application to a Sahelian village. Water Resour Res 44:W12445
Bomblies A, Duchemin JB, Eltahir EA (2009) A mechanistic approach for accurate simulation of village scale malaria transmission. Malar J 8(1):223
Briere JF, Pracros P, Le Roux AY, Pierre JS (1999) A novel rate model of temperature-dependent development for arthropods. Environ Entomol 28(1):22–29
Brooks RT, Hayashi M (2002) Depth-area-volume and hydroperiod relationships of ephemeral (vernal) forest pools in southern New England. Wetlands 22(2):247–255
Cailly P, Tran A, Balenghien T, LAmbert G, Toty C, Ezanno P (2012) A climate-driven abundance model to assess mosquito control strategies. Ecol Modell 227:7–17
Caminade C, Kovats S, Rocklov J, Tompkins AM, Morse AP, Coln-Gonzlez FJ et al (2014) Impact of climate change on global malaria distribution. Proc Natl Acad Sci USA 111(9):3286–3291
Carter R, Mendis KN (2002) Evolutionary and historical aspects of the burden of malaria. Clin Microbiol Rev 15(4):564–594
Castillo-Chavez C, Bichara D, Morin BR (2016) Perspectives on the role of mobility, behavior, and time scales in the spread of diseases. Proc Natl Acad Sci USA 113(51):14582–14588
Cator LJ, Lynch PA, Read AF, Thomas MB (2012) Do malaria parasites manipulate mosquitoes? Trends Parasitol 28(11):466–470
Cator LJ, George J, Blanford S, Murdock CC, Baker TC, Read AF, Thomas MB (2013) 'Manipulation' without the parasite: altered feeding behaviour of mosquitoes is not dependent on infection with malaria parasites. Proc R Soc B 280:20130711
Cator LJ, Lynch PA, Thomas MB, Read AF (2014) Alterations in mosquito behaviour by malaria parasites: potential impact on force of infection. Malar J 13(1):164
Chaves LF, Koenraadt CJ (2010) Climate change and highland malaria: fresh air for a hot debate. Q Rev Biol 85(1):27–55
Childs LM, Buckee CO (2015) Dissecting the determinants of malaria chronicity: why within-host models struggle to reproduce infection dynamics. J R Soc Interface 12(104):20141379

Childs LM, Prosper OF (2017) Simulating within-vector generation of the malaria parasite diversity. PLoS ONE 12(5):e0177941

Christiansen-Jucht C, Parham PE, Saddler A, Koella JC, Basez MG (2014) Temperature during larval development and adult maintenance influences the survival of *Anopheles gambiae* ss. Parasites Vectors 7:489

Christiansen-Jucht C, Erguler K, Shek CY, Basez MG, Parham PE (2015) Modelling *Anopheles gambiae* ss population dynamics with temperature-and age-dependent survival. Int J Environ Res Public Health 12(6):5975–6005

Clements AN, Paterson GD (1981) The analysis of mortality and survival rates in wild populations of mosquitoes. J Appl Ecol 18(2):373–399

Cohuet A, Harris C, Robert V, Fontenille D (2010) Evolutionary forces on Anopheles: what makes a malaria vector? Trends Parasitol 26(3):130–136

Cosner C, Beier JC, Cantrell RS, Impoinvil D, Kapitanski L, Potts MD et al (2009) The effects of human movement on the persistence of vector-borne diseases. J Theor Biol 258(4):550–560

Cox FE (2010) History of the discovery of the malaria parasites and their vectors. Parasites Vectors 3:5

Craig MH, Snow RW, Le Sueur D (1999) A climate-based distribution model of malaria transmission in sub-Saharan Africa. Parasitol Today 15(3):105–111

Crompton PD, Moebius J, Portugal S, Waisberg M, Hart G, Garver LS et al (2014) Malaria immunity in man and mosquito: insights into unsolved mysteries of a deadly infectious disease. Annu Rev Immunol 32:157–187

Culleton R, Carter R (2012) African Plasmodium vivax: distribution and origins. Int J Parasitol 42(12):1091–1097

Dawes EJ, Churcher TS, Zhuang S, Sinden RE, Basez MG (2009) Anopheles mortality is both age-and Plasmodium-density dependent: implications for malaria transmission. Malar J 8:228

Delatte H, Gimonneau G, Triboire A, Fontenille D (2009) Influence of temperature on immature development, survival, longevity, fecundity, and gonotrophic cycles of *Aedes albopictus*, vector of chikungunya and dengue in the Indian Ocean. J Med Entomol 46(1):33–41

Demasse RD, Ducrot A (2013) An age-structured within-host model for multistrain malaria infections. SIAM J Appl Math 73(1):572–593

Dembele B, Friedman A, Yakubu AA (2009) Malaria model with periodic mosquito birth and death rates. J Biol Dyn 3(4):430–445

Depinay JMO, Mbogo CM, Killeen G, Knols B, Beier J, Carlson J et al (2004) A simulation model of African Anopheles ecology and population dynamics for the analysis of malaria transmission. Malar J 3:29

Desconnets JC, Taupin JD, Lebel T, Leduc C (1997) Hydrology of the HAPEX-Sahel Central Super-Site: surface water drainage and aquifer recharge through the pool systems. J Hydrol 188:155–178

Detinova TS (1962) Age grouping methods in Diptera of medical importance with special reference to some vectors of malaria. Age grouping methods in diptera of medical importance with special reference to some vectors of malaria

Diekmann O, Heesterbeek JAP (2000) Mathematical epidemiology of infectious diseases: model building, analysis, and interpretation. Wiley, Chichester

Diekmann O, Heesterbeek JAP, Metz JA (1990) On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. J Math Biol 28(4):365–382

Dietz K, Molineaux L, Thomas A (1974) A malaria model tested in the African savannah. Bull WHO 50:347–357

Djadid ND, Gholizadeh S, Tafsiri E, Romi R, Gordeev M, Zakeri S (2007) Molecular identification of Palearctic members of Anopheles maculipennis in northern Iran. Malar J 6:6

Eckhoff PA (2011) A malaria transmission-directed model of mosquito life cycle and ecology. Malar J 10:303

Eckhoff P (2012) *P. falciparum* infection durations and infectiousness are shaped by antigenic variation and innate and adaptive host immunity in a mathematical model. PLoS ONE 7(9):e44950

Eling W, Hooghof J, van de Vegte-Bolmer M, Sauerwein R, Van Gemert GJ (2001) Tropical temperatures can inhibit development of the human malaria parasite *Plasmodium falciparum* in the mosquito. Proc Sect Exp Appl Entomol Neth Entomol Soc 12:151–156

Engelbrecht CJ, Engelbrecht FA, Dyson LL (2013) Highresolution modelprojected changes in midtropospheric closedlows and extreme rainfall events over southern Africa. Int J Climatol 33(1):173–187

Ermert V, Fink AH, Jones AE, Morse AP (2011) Development of a new version of the Liverpool Malaria Model. I. Refining the parameter settings and mathematical formulation of basic processes based on a literature review. Malar J 10:35

Ermert V, Fink AH, Jones AE, Morse AP (2011) Development of a new version of the Liverpool Malaria Model. II. Calibration and validation for West Africa. Malar J 10:62

Ferguson HM, Read AF (2002) Why is the effect of malaria parasites on mosquito survival still unresolved? Trends Parasitol 18(6):256–261

Filipe JA, Riley EM, Drakeley CJ, Sutherland CJ, Ghani AC (2007) Determination of the processes driving the acquisition of immunity to malaria using a mathematical transmission model. PLoS Comput Biol 3(12):e255

Finch JW, Hall RL (2001) Environmental Agency R&D Technical Report W6-043/TR: estimation of open water evaporation: a review of methods

Flerchinger GN, Xaio W, Marks D, Sauer TJ, Yu Q (2009) Comparison of algorithms for incoming atmospheric longwave radiation. Water Resour Res 45:W03423

Forouzannia F, Gumel A (2015) Dynamics of an age-structured two-strain model for malaria transmission. Appl Math Comput 250:860–886

Garrett-Jones C (1964) Prognosis for interruption of malaria transmission through assessment of the mosquito's vectorial capacity. Nature 204:1173–1175

Garrett-Jones C, Shidrawi GR (1969) Malaria vectorial capacity of a population of *Anopheles gambiae*: an exercise in epidemiological entomology. Bull WHO 40(4):531–545

Garske T, Ferguson NM, Ghani AC (2013) Estimating air temperature and its influence on malaria transmission across Africa. PLoS ONE 8(2):e56487

Gething PW, Smith DL, Patil AP, Tatem AJ, Snow RW, Hay SI (2010) Climate change and the global malaria recession. Nature 465(7296):342–345

Gething PW, Van Boeckel TP, Smith DL, Guerra CA, Patil AP, Snow RW, Hay SI (2011) Modelling the global constraints of temperature on transmission of *Plasmodium falciparum* and *P. vivax*. Parasites Vectors 4(1):92

Gething PW, Casey DC, Weiss DJ, Bisanzio D, Bhatt S, Cameron E et al (2016) Mapping *Plasmodium falciparum* mortality in Africa between 1990 and 2015. N Engl J Med 375(25):2435–2445

Ghani AC, Sutherland CJ, Riley EM, Drakeley CJ, Griffin JT, Gosling RD, Filipe JA (2009) Loss of population levels of immunity to malaria as a result of exposure-reducing interventions: consequences for interpretation of disease trends. PLoS ONE 4(2):e4383

Gimnig JE, Ombok M, Otieno S, Kaufman MG, Vulule JM, Walker ED (2002) Density-dependent development of *Anopheles gambiae* (Diptera: Culicidae) larvae in artificial habitats. J Med Entomol 39(1):162–172

Githeko AK, Ndegwa W (2001) Predicting malaria epidemics in the Kenyan highlands using climate data: a tool for decision makers. Glob Change Hum Health 2(1):54–63

Griffin JT, Hollingsworth TD, Okell LC, Churcher TS, White M, Hinsley W et al (2010) Reducing *Plasmodium falciparum* malaria transmission in Africa: a model-based evaluation of intervention strategies. PLoS Med 7(8):e1000324

Griffin JT, Hollingsworth TD, Reyburn H, Drakeley CJ, Riley EM, Ghani AC (2015) Gradual acquisition of immunity to severe malaria with increasing exposure. Proc R Soc B 282:20142657

Griffin JT, Bhatt S, Sinka ME, Gething PW, Lynch M, Patouillard E et al (2016) Potential for reduction of burden and local elimination of malaria by reducing *Plasmodium falciparum* malaria transmission: a mathematical modelling study. Lancet Infect Dis 16(4):465–472

Gu W, Regens JL, Beier JC, Novak RJ (2006) Source reduction of mosquito larval habitats has unexpected consequences on malaria transmission. Proc Natl Acad Sci USA 103(46):17560–17563

Guilbride DL, Guilbride PD, Gawlinski P (2012) Malaria's deadly secret: a skin stage. Trends Parasitol 28(4):142–150

Gupta S, Day KP (1994) A theoretical framework for the immunoepidemiology of *Plasmodium falciparum* malaria. Parasite Immunol 16(7):361–370

Gupta S, Snow RW, Donnelly CA, Marsh K, Newbold C (1999) Acquired immunity and postnatal clinical protection in childhood cerebral malaria. Proc R Soc B 266(1414):33–38

Gupta S, Snow RW, Donnelly CA, Marsh K, Newbold C (1999) Immunity to non-cerebral severe malaria is acquired after one or two infections. Nat Med 5(3):340–343

Gurarie D, Karl S, Zimmerman PA, King CH, Pierre TG, Davis TM (2012) Mathematical modeling of malaria infection with innate and adaptive immunity in individuals and agent-based communities. PLoS ONE 7(3):e34040

Hamblin MT, Di Rienzo A (2000) Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet 66(5):1669–1679

Hastings IM (1997) A model for the origins and spread of drug-resistant malaria. Parasitology 115(2):133–141

Hastings IM (2003) Malaria control and the evolution of drug resistance: an intriguing link. Trends Parasitol 19(2):70–73

Hastings IM, Watkins WM (2005) Intensity of malaria transmission and the evolution of drug resistance. Acta Trop 94(3):218–229

Hay SI, Cox J, Rogers DJ, Randolph SE, Stern DI, Shanks GD, Myers MF, Snow RW (2002) Climate change and the resurgence of malaria in the East African highlands. Nature 415:905–909

Hay SI, Smith DL, Snow RW (2008) Measuring malaria endemicity from intense to interrupted transmission. Lancet Infect Dis 8(6):369–378

Hayashi M, Van der Kamp G (2000) Simple equations to represent the volumeareadepth relations of shallow wetlands in small topographic depressions. J Hydrol 237(1):74–85

Hemingway J, Ranson H, Magill A, Kolaczinski J, Fornadel C, Gimnig J et al (2016) Averting a malaria disaster: Will insecticide resistance derail malaria control? Lancet 387(10029):1785–1788

Hoshen MB, Morse AP (2004) A weather-driven model of malaria transmission. Malar J 3:32

Huijben S, Paaijmans KP (2017) Putting evolution in elimination: winning our ongoing battle with evolving malaria mosquitoes and parasites. Evol Appl 11(4):415–430

IPCC (2013) Summary for policymakers. In: Stocker TF, Qin D, Plattner GK, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex Y, Midgley PM (eds) Climate change 2013: the physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge

IPCC (2014) Summary for policymakers. In: Field CB, Barros VR, Dokken DJ, Mach KJ, Mastrandrea MD, Bilir TE, Chatterjee M, Ebi KL, Estrada YO, Genova RC, Girma B, Kissel ES, Levy AN, MacCracken S, Mastrandrea PR, White LL (eds) Climate change 2014: impacts, adaptation, and vulnerability. Part A: global and sectoral aspects. Contribution of working group II to the fifth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge, pp 1–32

Jannat KNE, Roitberg BD (2013) Effects of larval density and feeding rates on larval life history traits in *Anopheles gambiae* ss (Diptera: Culicidae). J Vector Ecol 38(1):120–126

Jepson WF, Moutia A, Courtois C (1947) The malaria problem in Mauritius: the bionomics of Mauritian anophelines. Bull Entomol Res 38(01):177–208

Kiernan B (2007) Blood and soil: a world history of genocide and extermination from sparta to darfur. Yale University Press, Harrisburg

Kitau J, Oxborough RM, Tungu PK, Matowo J, Malima RC, Magesa SM, Bruce J, Mosha FW, Rowland MW (2012) Species shifts in the *Anopheles gambiae* complex: do LLINs successfully control *Anopheles arabiensis*? PLoS ONE 7(3):e31481

Klowden MJ, Briegel H (1994) Mosquito gonotrophic cycle and multiple feeding potential: contrasts between Anopheles and Aedes (Diptera: Culicidae). J Med Entomol 31(4):618–622

Kweka EJ, Zhou G, Munga S, Lee MC, Atieli HE, Nyindo M et al (2012) Anopheline larval habitats seasonality and species distribution: a prerequisite for effective targeted larval habitats control programmes. PLoS ONE 7(12):e52084

Lactin DJ, Holliday NJ, Johnson DL, Craigen R (1995) Improved rate model of temperature-dependent development by arthropods. Environ Entomol 24(1):68–75

Lafferty KD (2009) The ecology of climate change and infectious diseases. Ecology 90(4):888–900

Lardeux FJ, Tejerina RH, Quispe V, Chavez TK (2008) A physiological time analysis of the duration of the gonotrophic cycle of Anopheles pseudopunctipennis and its implications for malaria transmission in Bolivia. Malar J 7:141

Li Y, Ruan S, Xiao D (2011) The within-host dynamics of malaria infection with immune response. Math Biosci Eng 8(4):999–1018

Lindsay SW, Birley MH (1996) Climate change and malaria transmission. Ann Trop Med Parasitol 90(6):573–588

Lindsay SW, Martens WJ (1998) Malaria in the African highlands: past, present and future. Bull WHO 76(1):33–45

Liu W, Li Y, Shaw KS, Learn GH, Plenderleith LJ, Malenke JA et al (2014) African origin of the malaria parasite *Plasmodium vivax*. Nat Commun 5:3346

Logan JA, Wollkind DJ, Hoyt SC, Tanigoshi LK (1976) An analytic model for description of temperature dependent rate phenomena in arthropods. Environ Entomol 5(6):1133–1140

Lou Y, Zhao XQ (2010) A climate-based malaria transmission model with structured vector population. SIAM J Appl Math 70(6):2023–2044

Loy DE, Liu W, Li Y, Learn GH, Plenderleith LJ, Sundararaman SA, Sharp PM, Hahn BH (2017) Out of Africa: origins and evolution of the human malaria parasites *Plasmodium falciparum* and *Plasmodium vivax*. In J Parasitol 47(2–3):87–97

Lunde TM, Bayoh MN, Lindtjørn B (2013) How malaria models relate temperature to malaria transmission. Parasites Vectors 6:20

Lunde TM, Korecha D, Loha E, Sorteberg A, Lindtjørn B (2013) A dynamic model of some malaria-transmitting anopheline mosquitoes of the Afrotropical region. I. Model description and sensitivity analysis. Malar J 12:28

Lyimo EO, Takken W, Koella JC (1992) Effect of rearing temperature and larval density on larval survival, age at pupation and adult size of *Anopheles gambiae*. Entomol Exp Appl 63(3):265–271

Lyons CL, Coetzee M, Terblanche JS, Chown SL (2012) Thermal limits of wild and laboratory strains of two African malaria vector species, *Anopheles arabiensis* and *Anopheles funestus*. Malar J 11(1):226

Lyons CL, Coetzee M, Chown SL (2013) Stable and fluctuating temperature effects on the development rate and survival of two malaria vectors, *Anopheles arabiensis* and *Anopheles funestus*. Parasit Vectors 6(1):104

Ma G, Hoffmann AA, Ma CS (2015) Daily temperature extremes play an important role in predicting thermal effects. J Exp Biol 218(14):2289–2296

Macdonald G (1952) The analysis of equilibrium in malaria. Trop Dis Bull 49(9):813–829

Macdonald G (1956) Epidemiological basis of malaria control. Bull WHO 15:613–626

Macdonald G (1956) Theory of the eradication of malaria. Bull WHO 15:369–387

Macdonald G (1957) The epidemiology and control of malaria. Oxford University Press, Oxford

MacDonald G, Cuellar CB, Foll CV (1968) The dynamics of malaria. Bull WHO 38(5):743–755

Mala AO, Irungu LW, Mitaki EK, Shililu JI, Mbogo CM, Njagi JK, Githure JI (2014) Gonotrophic cycle duration, fecundity and parity of *Anopheles gambiae* complex mosquitoes during an extended period of dry weather in a semi arid area in Baringo County, Kenya. Int J Mosq Res 1(2):28–34

Mandal S, Sarkar RR, Sinha S (2011) Mathematical models of malaria—a review. Malar J 10:202

Martens WJM, Jetten TH, Rotmans J, Niessen LW (1995) Climate change and vector-borne diseases: a global modelling perspective. Glob Environ Change 5(3):195–209

Martens WJ, Niessen LW, Rotmans J, Jetten TH, McMichael AJ (1995) Potential impact of global climate change on malaria risk. Environ Health Perspect 103(5):458–464

Martens WJ, Jetten TH, Focks DA (1997) Sensitivity of malaria, schistosomiasis and dengue to global warming. Clim Change 35(2):145–156

Martens P, Kovats RS, Nijhof S, De Vries P, Livermore MTJ, Bradley DJ et al (1999) Climate change and future populations at risk of malaria. Glob Environ Change 9(S1):S89–S107

McKinley DC, Ryan MG, Birdsey RA, Giardina CP, Harmon ME, Heath LS et al (2011) A synthesis of current knowledge on forests and carbon storage in the United States. Ecol Appl 21(6):1902–1924

Midega JT, Mbogo CM, Mwambi H, Wilson MD, Ojwang G, Mwangangi JM et al (2007) Estimating dispersal and survival of *Anopheles gambiae* and *Anopheles funestus* along the Kenyan coast by using markreleaserecapture methods. J Med Entomol 44(6):923–929

Minakawa N, Mutero CM, Githure JI, Beier JC, Yan G (1999) Spatial distribution and habitat characterization of anopheline mosquito larvae in Western Kenya. Am J Trop Med Hyg 61(6):1010–1016

Minakawa N, Sonye G, Mogi M, Yan G (2004) Habitat characteristics of *Anopheles gambiae* ss larvae in a Kenyan highland. Med Vet Entomol 18(3):301–305

Minakawa N, Munga S, Atieli F, Mushinzimana E, Zhou G, Githeko AK, Yan G (2005) Spatial distribution of anopheline larval habitats in Western Kenyan highlands: effects of land cover types and topography. Am J Trop Med Hyg 73(1):157–165

Molineaux L, Dietz K, Thomas A (1978) Further epidemiological evaluation of a malaria model. Bull WHO 56(4):565–571

Mordecai EA, Paaijmans KP, Johnson LR, Balzer C, BenHorin T, Moor E et al (2013) Optimal temperature for malaria transmission is dramatically lower than previously predicted. Ecol Lett 16(1):22–30

Murdock CC, Sternberg ED, Thomas MB (2016) Malaria transmission potential could be reduced with current and future climate change. Sci Rep 6:27771

Muriu SM, Coulson T, Mbogo CM, Godfray HCJ (2013) Larval density dependence in *Anopheles gambiae* ss, the major African vector of malaria. J Anim Ecol 82(1):166–174

Nájera JA, González-Silva M, Alonso PL (2011) Some lessons for the future from the Global Malaria Eradication Programme (1955–1969). PLoS Med 8(1):e1000412

Nguyen PL, Vantaux A, Hien DF, Dabiré KR, Yameogo BK, Gouagna LC, Fontenille D, Renaud F, Simard F, Costantini C, Thomas F (2017) No evidence for manipulation of *Anopheles gambiae*, An. coluzzii and An. arabiensis host preference by *Plasmodium falciparum*. Sci Rep 7(1):9415

Niang I, Ruppel OC, Abdrabo MA, Essel A, Lennard C, Padgham J, Urquhart P (2014) Africa. In: Barros VR, Field CB, Dokken DJ, Mastrandrea MD, Mach KJ, Bilir TE, Chatterjee M, Ebi KL, Estrada YO, Genova RC, Girma B, Kissel ES, Levy AN, MacCracken S, Mastrandrea PR, White LL (eds) Climate change 2014: impacts, adaptation, and vulnerability. Part B: regional aspects. Contribution of working group II to the fifth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge, pp 1199–1265

Niger AM, Gumel AB (2008) Mathematical analysis of the role of repeated exposure on malaria transmission dynamics. Differ Equ Dyn Syst 16(3):251–287

Nikolaev BP (1935) The influence of temperature on the development of the malaria parasite in the mosquito. Tr Paster Inst Epidem Bakt (Leningr) 2:108

Nikolov M, Bever CA, Upfill-Brown A, Hamainza B, Miller JM, Eckhoff PA, Wenger EA, Gerardin J (2016) Malaria elimination campaigns in the Lake Kariba region of Zambia: a spatial dynamical model. PLoS Comput Biol 12(11):e1005192

Noden BH, Kent MD, Beier JC (1995) The impact of variations in temperature on early *Plasmodium falciparum* development in *Anopheles stephensi*. Parasitology 111(05):539–545

Novikov YM, Vaulin OV (2014) Expansion of *Anopheles maculipennis* ss (Diptera: Culicidae) to north-eastern Europe and northwestern Asia: causes and consequences. Parasites Vectors 7:389

Odiere M, Bayoh MN, Gimnig J, Vulule J, Irungu L, Walker E (2007) Sampling outdoor, resting *Anopheles gambiae* and other mosquitoes (Diptera: Culicidae) in western Kenya with clay pots. J Med Entomol 44(1):14–22

Okech BA, Gouagna LC, Killeen GF, Knols BG, Kabiru EW, Beier JC et al (2003) Influence of sugar availability and indoor microclimate on survival of *Anopheles gambiae* (Diptera: Culicidae) under semifield conditions in western Kenya. J Med Entomol 40(5):657–663

Okech BA, Gouagna LC, Walczak E, Kabiru EW, Beier JC, Yan G, Githure JI (2004) The development of *Plasmodium falciparum* in experimentally infected *Anopheles gambiae* (Diptera: Culicidae) under ambient microhabitat temperature in western Kenya. Acta Trop 92(2):99–108

Okech BA, Gouagna LC, Kabiru EW, Walczak E, Beier JC, Yan G, Githure JI (2004) Resistance of early midgut stages of natural *Plasmodium falciparum* parasites to high temperatures in experimentally infected *Anopheles gambiae* (Diptera: Culicidae). J Parasitol 90(4):764–768

Okuneye K, Gumel AB (2017) Analysis of a temperature-and rainfall-dependent model for malaria transmission dynamics. Math Biosci 287:72–92

Olayemi IK, Ande AT (2008) Survivorship of *Anopheles gambiae* in relation to malaria transmission in Ilorin, Nigeria. Online J Health Allied Sci 7(3):1

Paaijmans KP, Wandago MO, Githeko AK, Takken W (2007) Unexpected high losses of *Anopheles gambiae* larvae due to rainfall. PLoS ONE 2(11):e1146

Paaijmans KP, Heusinkveld BG, Jacobs AF (2008) A simplified model to predict diurnal water temperature dynamics in a shallow tropical water pool. Int J Biometeorol 52(8):797–803

Paaijmans KP, Jacobs AFG, Takken W, Heusinkveld BG, Githeko AK, Dicke M, Holtslag AAM (2008) Observations and model estimates of diurnal water temperature dynamics in mosquito breeding sites in western Kenya. Hydrol Processes 22(24):4789–4801

Paaijmans KP, Read AF, Thomas MB (2009) Understanding the link between malaria risk and climate. Proc Natl Acad Sci USA 106(33):13844–13849

Paaijmans KP, Blanford S, Bell AS, Blanford JI, Read AF, Thomas MB (2010) Influence of climate on malaria transmission depends on daily temperature variation. Proc Natl Acad Sci USA 107(34):15135–15139

Paaijmans KP, Blanford S, Chan BH, Thomas MB (2012) Warmer temperatures reduce the vectorial capacity of malaria mosquitoes. Biol Lett 8(3):465–468

Paaijmans KP, Cator LJ, Thomas MB (2013) Temperature-dependent pre-bloodmeal period and temperature-driven asynchrony between parasite development and mosquito biting rate reduce malaria transmission intensity. PLoS ONE 8(1):e55777

Paaijmans KP, Heinig RL, Seliga RA, Blanford JI, Blanford S, Murdock CC, Thomas MB (2013) Temperature variation makes ectotherms more sensitive to climate change. Glob Change Biol 19(8):2373–2380

Packard RM (2007) The making of a tropical disease: a short history of malaria. Johns Hopkins University Press, Baltimore

Parham PE, Michael E (2010) Modeling the effects of weather and climate change on malaria transmission. Environ Health Perspect 118(5):620–626

Parham PE, Pople D, Christiansen-Jucht C, Lindsay S, Hinsley W, Michael E (2012) Modeling the role of environmental variables on the population dynamics of the malaria vector *Anopheles gambiae* sensu stricto. Malar J 11:271

Pascual M, Bouma MJ (2009) Do rising temperatures matter. Ecology 90(4):906–912

Pascual M, Ahumada JA, Chaves LF, Rodo X, Bouma M (2006) Malaria resurgence in the East African highlands: temperature trends revisited. Proc Natl Acad Sci USA 103(15):5829–5834

Pascual M, Cazelles B, Bouma MJ, Chaves LF, Koelle K (2008) Shifting patterns: malaria dynamics and rainfall variability in an African highland. Proc R Soc B 275(1631):123–132

Patz JA, Epstein PR, Burke TA, Balbus JM (1996) Global climate change and emerging infectious diseases. JAMA 275(3):217–223

Pollitt L, Churcher TS, Dawes EJ, Khan SM, Sajid M, Basáñez MG, Colegrave N, Reece SE (2013) Costs of crowding for the transmission of malaria parasites. Evol Appl 6(4):617–629

Pongtavornpinyo W, Hastings IM, Dondorp A, White LJ, Maude RJ, Saralamba S, Day NP, White NJ, Boni MF (2009) Probability of emergence of antimalarial resistance in different stages of the parasite life cycle. Evol Appl 2(1):52–61

Porphyre T, Bicout DJ, Sabatier P (2005) Modelling the abundance of mosquito vectors versus flooding dynamics. Ecol Model 183(2):173–181

Prosper O, Ruktanonchai N, Martcheva M (2012) Assessing the role of spatial heterogeneity and human movement in malaria dynamics and control. J Theor Biol 303:1–14

Ra GL, Quiones ML, Vlez ID, Zuluaga JS, Rojas W, Poveda G, Ruiz D (2005) Laboratory estimation of the effects of increasing temperatures on the duration of gonotrophic cycle of *Anopheles albimanus* (Diptera: Culicidae). Mem Inst Oswaldo Cruz 100(5):515–520

Reiner RC, Perkins TA, Barker CM, Niu T, Chaves LF, Ellis AM et al (2013) A systematic review of mathematical models of mosquito-borne pathogen transmission: 1970–2010. J R Soc Interface 10:20120921

Rodriguez-Barraquer I, Arinaitwe E, Jagannathan P, Boyle MJ, Tappero J, Muhindo M et al (2016) Quantifying heterogeneous malaria exposure and clinical protection in a cohort of Ugandan children. J Infect Dis 214(7):1072–1080

Rogers DJ, Randolph SE (2000) The global spread of malaria in a future, warmer world. Science 289(5485):1763–1766

Rudel TK, Defries R, Asner GP, Laurance WF (2009) Changing drivers of deforestation and new opportunities for conservation. Conserv Biol 23(6):1396–1405

Ruktanonchai NW, Smith DL, De Leenheer P (2016) Parasite sources and sinks in a patched Ross–Macdonald malaria model with human and mosquito movement: implications for control. Math Biosci 279:90–101

Ryan SJ, Ben-Horin T, Johnson LR (2015) Malaria control and senescence: the importance of accounting for the pace and shape of aging in wild mosquitoes. Ecosphere 6(9):170

Ryan SJ, McNally A, Johnson LR, Mordecai EA, Ben-Horin T, Paaijmans K, Lafferty KD (2015) Mapping physiological suitability limits for malaria in Africa under climate change. Vector Borne Zoonotic Dis 15(12):718–725

Saralamba S, Pan-Ngum W, Maude RJ, Lee SJ, Tarning J, Lindegårdh N et al (2011) Intrahost modeling of artemisinin resistance in *Plasmodium falciparum*. Proc Natl Acad Sci USA 108(1):397–402

Schneider P, Takken W, McCall PJ (2000) Interspecific competition between sibling species larvae of *Anopheles arabiensis* and *A. gambiae*. Med Vet Entomol 14(2):165–170

Scott TW, Takken W (2012) Feeding strategies of anthropophilic mosquitoes result in increased risk of pathogen transmission. Trends Parasitol 28(3):114–121

Sene KJ, Gash JH, McNeil DD (1991) Evaporation from a tropical lake: comparison of theory with direct measurements. J Hydrol 127(1–4):193–217

Service MW (1971) Studies on sampling larval populations of the *Anopheles gambiae* complex. Bull WHO 45(2):169–180

Silva JC, Egan A, Arze C, Spouge JL, Harris DG (2015) A new method for estimating species age supports the co-existence of malaria parasites and their mammalian hosts. Mol Biol Evol 32(5):1354–1364

Singh P, Yadav Y, Saraswat S, Dhiman RC (2016) Intricacies of using temperature of different niches for assessing impact on malaria transmission. Indian J Med Res 144(1):67–75

Sinka ME, Bangs MJ, Manguin S, Coetzee M, Mbogo CM, Hemingway J et al (2010) The dominant Anopheles vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic prcis. Parasites Vectors 3:117

Small J, Goetz SJ, Hay SI (2003) Climatic suitability for malaria transmission in Africa, 1911–1995. Proc Natl Acad Sci USA 100(26):15341–15345

Smith DL, Hay SI, Noor AM, Snow RW (2009) Predicting changing malaria risk after expanded insecticide-treated net coverage in Africa. Trends Parasitol 25(11):511–516

Smith DL, Battle KE, Hay SI, Barker CM, Scott TW, McKenzie FE (2012) Ross, Macdonald, and a theory for the dynamics and control of mosquito-transmitted pathogens. PLoS Pathog 8(4):e1002588

Snow RW (2015) Global malaria eradication and the importance of *Plasmodium falciparum* epidemiology in Africa. BMC Med 13:23

Snow RW, Craig MH, Deichmann U, Le Sueur D (1999) A preliminary continental risk map for malaria mortality among African children. Parasitol Today 15(3):99–104

Snow RW, Kibuchi E, Karuri SW, Sang G, Gitonga CW, Mwandawiro C et al (2015) Changing malaria prevalence on the Kenyan coast since 1974: climate, drugs and vector control. PLoS ONE 10(6):e0128792

Sternberg ED, Thomas MB (2014) Local adaptation to temperature and the implications for vector-borne diseases. Trends Parasitol 30(3):115–122

Sumba LA, Ogbunugafor CB, Deng AL, Hassanali A (2008) Regulation of oviposition in *Anopheles gambiae* ss: role of inter-and intra-specific signals. J Chem Ecol 34(11):1430–1436

Suzuki R, Xu J, Motoya K (2006) Global analyses of satellitederived vegetation index related to climatological wetness and warmth. Int J Climatol 26(4):425–438

Tabo Z, Luboobi LS, Ssebuliba J (2017) Mathematical modelling of the in-host dynamics of malaria and the effects of treatment. J Math Comput Sci 17(1):1–21

Takken W, Klowden MJ, Chambers GM (1998) Articles: effect of body size on host seeking and blood meal utilization in *Anopheles gambiae* sensu stricto (Diptera: Culicidae): the disadvantage of being small. J Med Entomol 35(5):639–645

Takken W, van Loon JJ, Adam W (2001) Inhibition of host-seeking response and olfactory responsiveness in *Anopheles gambiae* following blood feeding. J Insect Physiol 47(3):303–310

Tanser FC, Sharp B, le Sueur D (2003) Potential effect of climate change on malaria transmission in Africa. Lancet 362(9398):1792–1798

Teboh-Ewungkem MI, Podder CN, Gumel AB (2010) Mathematical study of the role of gametocytes and an imperfect vaccine on malaria transmission dynamics. Bull Math Biol 72(1):63–93

Tilley L, Dixon MW, Kirk K (2011) The *Plasmodium falciparum*-infected red blood cell. Int J Biochem Cell Biol 43(6):839–842

Tompkins AM, Ermert V (2013) A regional-scale, high resolution dynamical malaria model that accounts for population density, climate and surface hydrology. Malar J 12(1):65

Torres-Sorando L, Rodríguez DJ (1997) Models of spatio-temporal dynamics in malaria. Ecol Modell 104(2):231–240

Trape JF, Rogier C, Konate L, Diagne N, Bouganali H, Canque B et al (1994) The Dielmo project: a longitudinal study of natural malaria infection and the mechanisms of protective immunity in a community living in a holoendemic area of Senegal. Am J Trop Med Hyg 51(2):123–137

United Nations, Department of Economic and Social Affairs, Population Division (2017) World population prospects: the 2017 revision, key findings and advance tables. Working paper no. ESA/P/WP/248

Van den Driessche P, Watmough J (2002) Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. Math Biosci 180(1):29–48

Vantaux A, de Sales Hien DF, Yameogo B, Dabiré KR, Thomas F, Cohuet A, Lefèvre T (2015) Host-seeking behaviors of mosquitoes experimentally infected with sympatric field isolates of the human malaria parasite *Plasmodium falciparum*: no evidence for host manipulation. Front Ecol Evol 3:86

Vogels M, Zoeckler R, Stasiw DM, Cerny LC (1975) PF Verhulsts notice sur la loi que la populations suit dans son accroissement from correspondence mathematique et physique. Ghent, vol. X, 1838. J Biol Phys 3(4):183–192

Vogels M, Zoeckler R, Stasiw DM, Cerny LC (1975) PF Verhulsts notice sur la loi que la populations suit dans son accroissement from correspondence mathematique et physique. Ghent, vol. X, 1838. J Biol Phys 3(4):183–192

Wearing HJ, Rohani P, Keeling MJ (2005) Appropriate models for the management of infectious diseases. PLoS Med 2(7):e174

Webb JLA Jr (2014) The long struggle against malaria in tropical Africa. Cambridge University Press, New York

White MT, Griffin JT, Churcher TS, Ferguson NM, Basez MG, Ghani AC (2011) Modelling the impact of vector control interventions on *Anopheles gambiae* population dynamics. Parasites Vectors 4:153

White MT, Griffin JT, Churcher TS, Ferguson NM, Basez MG, Ghani AC (2011) Modelling the impact of vector control interventions on *Anopheles gambiae* population dynamics. Parasites Vectors 4:153

Yamana TK, Bomblies A, Laminou IM, Duchemin JB, Eltahir EA (2013) Linking environmental variability to village-scale malaria transmission using a simple immunity model. Parasit Vectors 6(1):226

Yamana TK, Bomblies A, Eltahir EA (2016) Climate change unlikely to increase malaria burden in West Africa. Nat Clim Change 6(11):1009

Yamana TK, Qiu X, Eltahir EA (2017) Hysteresis in simulations of malaria transmission. Adv Water Resour 108:416–422

Yaro AS, Dao A, Adamou A, Crawford JE, Ribeiro JM, Gwadz R et al (2006) The distribution of hatching time in *Anopheles gambiae*. Malar J 5:19

Yé Y, Hoshen M, Kyobutungi C, Louis VR, Sauerborn R (2009) Local scale prediction of *Plasmodium falciparum* malaria transmission in an endemic region using temperature and rainfall. Global Health Action 2(s1):1923

Yeung S, Pongtavornpinyo W, Hastings IM, Mills AJ, White NJ (2004) Antimalarial drug resistance, artemisinin-based combination therapy, and the contribution of modeling to elucidating policy choices. Am J Trop Med Hyg 71(2S):179–186

Zhou G, Minakawa N, Githeko AK, Yan G (2004) Association between climate variability and malaria epidemics in the East African highlands. Proc Natl Acad Sci USA 101(8):2375–2380

ISSN 1995-0802, Lobachevskii Journal of Mathematics, 2018, Vol. 39, No. 7, pp. 985–991. © Pleiades Publishing, Ltd., 2018.

Polynomial Greatest Common Divisor as a Solution of System of Linear Equations

D. A. Dolgov*

(Submitted by F. M. Ablayev)

Department of System Analysis and Information Technologies,
Institute of Computational Mathematics and Information Technologies,
Kazan (Volga Region) Federal University, ul. Kremlevskaya 18, Kazan, Tatarstan, 420008 Russia

Received December 6, 2017

Abstract—In this article we present a new algebraic approach to the greatest common divisor (GCD) computation of two polynomials based on Bezout's identity. This approach is based on the solution of system of linear equations. Also we introduce the dmod operation for polynomials. This operation on polynomials f, g is used to reduce the degree of the larger polynomial f in a finite field F_p. This operation saves $GCD(f, g)$. Also we present some ideas how to reduce spurious factors that arise at the procedure.

DOI: 10.1134/S1995080218070090

Keywords and phrases: *Polynomial GCD, Euclidean algorithm, system of linear equations, generalized Schur algorithm.*

1. INTRODUCTION

Polynomials have been studying for a very long time. A whole series of objects is connected with polynomials: zero, negative, complex numbers, the emergence of the theory of groups as a section of mathematics and the allocation of classes of special functions in analysis. More applications are found for polynomials of one variable.

Computation of greatest common divisor (GCD) of polynomials of one variable can be implemented like as the GCD computation for integer numbers by the Euclid GCD algorithm using operation of division at long integers. The polynomial GCD has specific properties that make it a fundamental notion in various areas of algebra. Often the roots of GCD of two polynomials are common roots of the two polynomials, and this allows us to get information of the roots without computing them. Some results concerning theory of polynomials can be found in [1–3, 6]. Computation of GCD of two polynomials can be used in cryptography (public key cryptography by the means of elliptic curves), finite fields, computer algebra, coding theory (cyclic redundancy codes and BCH codes). In particular it can be used in polynomial factorisation problem.

In this article we present a new algebraic approach to the GCD computation of two polynomials of one variable based on Bezout's identity. This approach is based on the solution of system of linear equations. So, we turn from the problem of GCD computation to the problem of solving a system of linear equations. Also we present the dmod operation on polynomials f, g, which is used to reduce the degree of the larger polynomial f in a finite field F_p. Also we present some ideas how to reduce spurious factors that arise at the procedure.

*E-mail: Dolgov.kfu@gmail.com

2. GCD OF TWO POLYNOMIALS

In this section we present results about GCD of two polynomials and rank of the Sylvester matrix, for example theorem 4 from [3].

We consider a field k. The set of all polynomials with coefficients in k is denoted by $k[x]$. Here and below we consider polynomials in one variable x. We have two polynomials $f(x),\ g(x) \in k[x]$:

$$f(x) = f_0x^n + f_1x^{n-1} + ... + f_n = \sum_{i=0}^{n} f_ix^{n-i}, \quad f_0, ..., f_n \in k,$$

$$g(x) = g_0x^m + g_1x^{m-1} + ... + g_m = \sum_{z=0}^{m} g_zx^{m-z}, \quad g_0, ..., g_m \in k.$$

If polynomial f has leading coefficient 1 then f is called monic.

Definition 1. *Let f and g be two polynomials in $k[x]$ with one of them non-zero. The greatest common divisor of f and g is the unique polynomial $d = GCD(f, g)$ with the following properties: 1) d divides f and g; 2) c divides f and g implies c divides d.*

In this article we consider polynomials with integer coefficients. We have Bezout's identity [1]: there exist two polynomials u and v such that $GCD(f, g) = d = uf + vg$, where $deg(u) = s,\ deg(v) = t,\ d(x) = d_0x^r + d_1x^{r-1} + ... + d_r$ is polynomial of the smallest degree, which can be represented in the this form.

Theorem 1. $deg(u) < deg(g) - deg(d),\ deg(v) < deg(f) - deg(d)$.

Proof. We divide Bezout's identity by $d(x)$. So, we have

$$(f'_0x^{n-r} + ... + f'_{n-r})(u_0x^s + ... + u_s) + (g'_0x^{m-r} + ... + g'_{m-r})(v_0x^t + ...) = 1,$$

$f'_i,\ g'_i$ are coefficients. They are not zeros. We multiply the brackets. Degrees at new polynomials must be the same. So,

$$n - r + s = m - r + t \equiv n - r - t + s = m - r, \tag{1}$$

$f(x) \neq g(x),\ r < \min\{deg(f(x)), deg(g(x)\} = \min\{n, m\}, m - r > 0,\ n - r > 0$. We seek to find the polynomial with smallest coefficients. So, we choose $n - r - t > 0 \equiv t < n - r$ and $m - r - s > 0 \equiv s < m - r$. □

The maximum possible degree for s is $m - r - 1$. It's for t is $n - r - 1$. It's possible for (1) identity. For Bezout's identity we take in abundance $s = m - 1$ and $t = n - 1$. As a result we have the same degrees after multiplication of the brackets. After that we equate degrees and obtain system of equations; $q = n + m - 1, f''_1 = q - f'_1, f''_2 = q - f'_2, g''_1 = q - g'_1, g''_2 = q - g'_2$:

$$\sum_{f'_1=0}^{r} a_{f'_1}x^{f''_1} + \sum_{f'_2=r+1}^{n+m-1} a_{f'_2}x^{f''_2} + \sum_{g'_1=0}^{r} a_{g'_1}x^{g''_1} + \sum_{g'_2=r+1}^{n+m-1} a_{g'_2}x^{g''_2} = \sum_{d=0}^{r} a_dx^{q-d}. \tag{2}$$

System can be present in the matrix representation: $Syl_{n,m}(f, g)UV^T = d$. Let $Syl_{n,m}(f, g)$ is a coefficient matrix (called Sylvester matrix), $deg(f) = n,\ deg(g) = m,\ n > m$. Size of matrix A is $(n + m) * (n + m)$. Each line consists of coefficients of corresponding degree. $d_k(l)$ is denoted the coefficient of degree x^l. UV is a vector, which consist of coefficients of the u and v polynomials. We define it as a system

$$\begin{pmatrix} f_0 & 0 & \cdots & 0 & g_0 & 0 & \cdots & 0 \\ f_1 & f_0 & \cdots & 0 & g_1 & g_0 & \cdots & 0 \\ f_2 & f_1 & \cdots & 0 & g_2 & g_1 & \cdots & 0 \\ \vdots & \vdots & & & \vdots & \vdots & & \vdots \\ f_m & f_{m-1} & \cdots & f_1 & g_m & g_{m-1} & \cdots & 0 \\ \vdots & & & \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & f_n & 0 & 0 & \cdots & g_m \end{pmatrix} \begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_{m-1} \\ v_0 \\ \vdots \\ v_{n-1} \end{pmatrix} = \begin{pmatrix} d_0(n + m - 1) \\ d_1(n + m - 2) \\ \vdots \\ d_{m-1}(n) \\ d_m(n - 1) \\ \vdots \\ d_{n+m-1}(0) \end{pmatrix}. \tag{3}$$

Theorem 2. *If* $d(x) = 1$, then $rank(Syl_{n,m}(f,g)) = deg(f) + deg(g)$, where $f \neq g$.

Proof. Assume the contrary, that it's false. Let $n = deg(f), m = deg(g)$, $n > m$. At first we choose 1 or $m+1$ column (let it be $m+1$ column) and express another columns. As a result we have remainder of dividing one polynomial into another and first row has one 1 and $n+m-1$ zeros. At first step we choose second column, because due to the Euclidean GCD we have $GCD(f,g) = GCD(g, f \ mod \ g)$. So at first column we have new remainder. Also at the $m+2$ column we have remainder $f \ mod \ g$, which was at first step in the first column. We repeat this procedure until n step. □

Theorem 3. *If* $d(x) \neq 1$, *then* $\max\{deg(f), deg(g)\} < rank(Syl_{n,m}(f,g)) < deg(f) + deg(g)$, *where* $f \neq g$.

Proof. The second part of the inequality follows from Theorem 1. We prove the first part. Assume the contrary, that it's false. Let $n = deg(f)$, $m = deg(g)$, $n > m$, $a_{i,j}$ is a value of the matrix $Syl_{n,m}(f,g)$ at i-th row and j-th column. We consider k-th row, where $m < k < n$, because $a_{m,0}$, ..., $a_{m,m-1}$ are zeros.

Let $a_{k,2m} = \sum_{i=m}^{k-1} \alpha_i a_{i,2m}$, $a_{k,m} = \sum_{i=m}^{k-1} \alpha'_i a_{i,m}$. If $a_{k,m} = a_{k,2m}$, then $\alpha_i = \alpha'_i$. We consider next elements: $a_{k,m-1}, a_{k,2m-1}$. At j-th step we have $a_{k,2m-j} = \sum_{i=m-j}^{k-1} \alpha_i a_{i,2m-j}$, $a_{k,m-j} = \sum_{i=m-j}^{k-1} \alpha'_i a_{i,m-j}$, $a_{k,m-j} \neq a_{k,2m-j}$, so $\alpha_i \neq \alpha'_i$. Contradiction. □

Also determinant of the $Syl_{n,m}(f,g)$ is equals to 0, if $GCD(f,g) \neq 1$.

Theorem 4. $deg(d) = deg(f) + deg(g) - rank(Syl_{n,m}(f,g))$.

Proof. Arithmetic operations over the rows, row interchange does not change the corank of the matrix. In particular, we may replace f by the remainder r obtained when dividing f by g. Let $f(x) = q(x)g(x) + r(x)$. So, $rank(Syl_{n,m}(f,g)) = rank(Syl_{n,m}(f - qg, g)) = rank(Syl_{n,m}(r,g))$. Then $deg(r) < m \leq n$ and the first row of $Syl_{n,m}(f,g)$ has a single non-zero element g_0 in row $m+1$. We may thus delete the first row and the $(m+1)$-th column without changing the corank, and this yields $Syl_{n-1,m}(f,g)$. Repeating, we see that if $r \neq 0$ and $k = deg(r)$, then $Syl_{n,m}(f,g)$ has the same corank as $Syl_{k,m}(r,g)$ and $Syl_{m,k}(g,r)$. We repeat from the start, by dividing g by r and so on; this yields the Euclidean algorithm for finding the GCD d, and we finally end up with the Sylvester matrix $Syl_{k,l}(0,d)$, for some $k > 0$ and $l = deg(d)$, which evidently has corank l since the first l columns are 0 and the last k are independent. At the end we have $k * k$ minor, which is triangular. So, $corank(Syl_{n,m}(f,g)) = corank(Syl_{k,l}(0,d)) = deg(d) \Rightarrow deg(d) = n + m - rank(Syl_{n,m}(f,g))$. □

Corollary 1. *The number of linearly independent rows includes all first zero rows and the first row containing the coefficient of* $d(x)$.

Proof. Let it's false. So, number of linearly independent rows does not contain a row with the first coefficient of $d(x)$. So, we have 2 ways. If it contains more rows with coefficients of $d(x)$, then $n + m - rank((Syl_{n,m}(f,g)) <$ deg(d). If it doesn't contain any row with coefficient of $d(x)$, then $n + m - rank((Syl_{n,m}(f,g)) >$ deg(d). Contradiction. □

We can use $rank$ operation to coprimality testing of two polynomials, where one of them is normalized: if $rank((Syl_{n,m}(f,g)) = n + m$, so two polynomials are mutually simple.

3. DMOD OPERATION FOR POLYNOMIALS

In this section we generalize the dmod operation to polynomials. Before now the dmod operation was defined only for numbers. Operation was used in the main loop of k-ary GCD to adjust input numbers u and v so that they are roughly the same size when the k-ary reduction is performed [4, 5].

The dmod operation for polynomials can used to create new polynomial, which contains GCD of two polynomials. Polynomials over a finite field are considered.

If polynomial in a ring $Z[x]$ then we can't find multiplicative inversion from any polynomial. In $Z[x]$ we have 2 invertible constant polynomials: $1, -1$.

Let F_p is a field of p elements, where p is a prime number. Here and below of this chapter, we consider polynomials with coefficients in the finite field F_p. $F_p[x]$ is a polynomial ring over the field F_p. Invertible elements of $F_p[x]$ are nonzero constant polynomials (polynomials of degree 0), that is, polynomials $1, 2, p-1 \in F_p[x]$. The reducibility and irreducibility of polynomials depends on whether

in which ring or over which field we are considering it. Here and below of this chapter we consider polynomials $f(x),\ g(x)$. For brevity, we denote them $f,\ g$.

Definition 2. *Difference of degrees of polynomials is function, which equals* $\rho(f,g) = deg(f) - deg(g) + 1$.

Here and below we denote irreducible polynomial of degree $\rho(f,g)$ in the F_p as $irr(\rho(f,g))$.

Definition 3. *The dmod operation is defined as* $d \bmod(f,g) \stackrel{\text{def}}{=} \frac{f-(f/g \bmod irr(\rho(f,g)))g}{irr(\rho(f,g))}$.

Theorem 5. *The dmod operation is unsolvable in the* $Z[x]$.

Proof. $Z[x]$ is a ring of all polynomials with coefficients in Z. In general $deg(g) > 0$, so we can't find multiplcative inverse of g polynomial in numerator. So, we can not integer divide f polynomial by $irr(\rho(f,g))$ and reduce the order of original polynomial. □

Theorem 6. *The dmod operation preserves GCD.*

Proof. Numerator can be presented as $f - qg$, where $q \in F_p[x]/irr(\rho(f,g))$, $irr(\rho(f,g))$ is an irreducible polynomial in a $F_p[x]$ with prime p. □

Theorem 7. *In the dmod operation the remainder from the division the numerator by the denominator is equals to* 0.

Proof. Let it's false. So, $f - (f/g \bmod irr(\rho(f,g)))g = q(x)irr(\rho(f,g)) + c(x)$ for some $k(x), c(x)$ with coefficients in F_p. Let's consider the numerator by the modulo $irr(\rho(f,g))$:

$$f \bmod irr(\rho(f,g)) = (f/g \bmod irr(\rho(f,g)))(g \bmod irr(\rho(f,g)) = f \bmod irr(\rho(f,g)).$$

Contradiction. □

Corollary 2. *The dmod operation gives polynomial, which consists* $d(x) = GCD(f,g)$, *if* $\max\{deg(f), deg(g)\} - \rho(f,g) = deg(d)$.

Proof. It's true due to Theorems 6, 7 and condition of the theorem. □

Is it possible to single-valued compare arithmetic properties of sum or product of polynomials in the dmod operation compared with sum or multiplication of dmod terms? No. Since we are looking for an inverse polynomial, we can not know anything about the degree of the inverse polynomial, and therefore the right side of the dmod expression.

4. A DESCRIPTION OF THE POLYNOMIAL GCD ALGORITHM

In this section we describe a new algebraic approach to a polynomial GCD computation. Although our point of view is sequential, ideas presented here apply to parallel versions of the algorithm as well. We begin by presenting solution of system of linear equations. Then we describe a new polynomial GCD algorithm and new minimisation criteria.

4.1. Solution of System of Linear Equations

The main problem of system (3) is selection of vector UV. We propose method, which imply usage solving of linear equations.

To solve the system of $AUV = d$ we can take first k rows from A and equate them to zero, see algorithm 1 in Section 4.2. It's true, because first rows of d is zero, see (2). A_k is a matrix, which consists of the first k rows of matrix A. Due to Corollary 1 we take $k = rank(Syl_{n,m}(f,g)) - 1$ and solve system $A_k X = \theta$. Number of equations is lesser than number of variables. We can find result vector X by solving next of linear equations $A_k X = \theta$, x_0 is a non-zero solution of the system of linear equations.

Rank of Sylvester matrix is calculated using the Gaussian algorithm. During reduction of the matrix to a stepped form we there is an accumulation of rounding errors. In the main error increases during the forward stroke, when the leading s-th line is multiplied by the coefficients $a_{i,s}/a_{s,s}$, $i = s+1, ..., n+m$. If the coefficients are greater than 1, then errors obtained in the previous steps are accumulated. To avoid this, a modification of the Gauss method with selection of the main element is applied. At each

step, selection of the maximum element by column is added to the usual scheme. During the elimination of variables system of equations is obtained:

$$x_i + \sum_{j=i+1}^{n+m} a_{ij}^i x_j = b_i^i,\ i = 1, ..., s-1; \quad \sum_{j=s}^{n+m} a_{ij}^{s-1} x_j = b_j^{s-1},\ i = s, ..., n+m.$$

Let find l: $|a_{l,s}^{s-1}| = \max |a_{j,s}^{s-1}|,\ j = s, ..., n+m$. After that we swap s-th and l-th rows. In many cases such transformation significantly reduces sensitivity of solution to rounding errors in calculations.

Another approach to compute rank of the matrix was proposed in [7]. This algorithm is based on the fast Cholesky factorization of $Syl^T Syl$ or $HSyl^T HSyl$ and relies on the stabilized version of the generalized Schur algorithm for matrices with displacement structure. Syl is the Sylvester matrix, $HSyl$ is the Hankel variation of the Sylvester matrix Syl. They introduced $HSyl$ in order to complete the fast Cholesky factorization without pivoting. The generalized Schur algorithm applied to compute the fast Cholesky factorization of $Syl^T Syl$ may break down during the first r steps due to the loss of positive definiteness of the $r * r$ leading principal submatrix of $Syl^T Syl$,

$$HSyl(f,g) = \begin{pmatrix}
0 & \cdots & 0 & 0 & f_0 & 0 & 0 & \cdots & 0 & 0 & g_0 \\
0 & \cdots & 0 & f_0 & f_1 & 0 & 0 & \cdots & 0 & g_0 & g_1 \\
0 & \cdots & f_0 & f_1 & f_2 & 0 & 0 & \cdots & g_0 & g_1 & g_2 \\
\vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\
f_0 & \cdots & f_{n-3} & f_{n-2} & f_{n-1} & g_0 & g_1 & \cdots & g_m & \cdots & 0 \\
f_1 & \cdots & f_{n-2} & f_{n-1} & f_n & g_1 & g_2 & \cdots & 0 & \cdots & 0 \\
f_2 & \cdots & f_{n-1} & f_n & 0 & g_2 & g_3 & \cdots & 0 & \cdots & 0 \\
\vdots & & \vdots & \vdots & & \vdots & & & \vdots & & \vdots \\
f_n & 0 & \cdots & 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0
\end{pmatrix}.$$

The generalized Schur algorithm uses three matrices (G, J, Z). G is a generator, which construct from $Syl^T Syl$ (or $HSyl^T HSyl$) and Euclidean 2-norm of columns of $Syl(f,g)$ (or $HSyl(f,g)$), see $Theorem$ 1 or $Theorem$ 3 in [7]. Z is a lower shift triangular matrix. Z can be constructed from 2 lower shift triangular matrices, $Z = diag(Z_m, Z_n)$. J is a signature matrix of the form $J = \begin{pmatrix} I_{p'} & 0 \\ 0 & -I_{q'} \end{pmatrix}$, $p' + q' = \alpha$. We pass from rank of the Sylvester matrix to apply the generalized Schur factorization of the system $\nabla T = T - ZTZ^T = GJG^T$, where $T = Syl^T Syl$ or $T = HSyl^T HSyl$, α is a diplacement rank of T. In [7] authors introduced the new algorithm for rank constatation (called "HSylRRA"), which is a modification of the generalized Schur algorithm. At each step authors compute Schur complement $\hat{S}_{i+1}$ for $i \geq 1$. If $||\hat{S}_{r+1}|| < \gamma$, then terminate the Schur algorithm and return r as a rank, γ is a tolerance for the Cholesky process.

4.2. Performance of the Polynomial GCD Algorithm

Before running algorithm 1 we have Sylvester matrix with coefficient of two polynomials f, g. If we don't have a matrix, we construct it. At second step we calculate rank of Sylvester matrix. For Gauss method it takes $O((n+m)^3)$, for modified generalized Schur algorithm it takes $O((n+m)^2)$. After that we compute part of X vector. If $d(x) = g(x)$ and $\rho(f,g) = 1$, so it takes $O(n+m)$. After that we compute vector B. If $rank(Syl_{n,m}(f,g)) = (n+m)/2$, so $size(X) = (n+m)/2 - 1$, so computation of vector B takes $(\frac{n+m}{2} - 1)^2 \sim O((n+m)^2)$. Depending on the choice of free variables in the vector X, we must start accurate GCD like Euclidean GCD between input number and output to get clean result

Algorithm 1. GCD of two polynomials

1. $B = [0...0]$

2. $z = n + m - rank(Syl_{n,m}(f, g)) + 1$

for $i = n + m; i > z; --i$ do

$X[i] = random()$

for $i = n + m;\ i > z;\ --i$ do

for $j = n + m;\ j > z;\ --i$ do

$B[i] = B[i] - Syl_{n,m}(f, g)[i][j]X[j]$

3. Solve the new system $AX = B$, where matrix A is obtained by randomly chosen $z - 1$ variables and construct B array by transferring last $z - 1$ members of the equation to the B vector.

return $Syl_{n,m}(f, g)X$

at the end. General assessment of the method: $O((n + m)^2) + O(nm) + O(n + m) \equiv O((n + m)^2)$. Without loss of generality we take asymptotic estimate of Euclidean GCD as $O(nm)$, although it should be less, because we use polynomials of smaller length. If we choose Gauss method to compute numerical rank, general assessment will be $O((n + m)^3)$.

If we have 2 *monic* polynomials we can equate $w = rank(Syl_{n,m}(f, g)) + 1$ row to 1, because GCD polynomial is monic too. In this case we obtain an additional condition regarding the choice of variable and clean GCD without using Euclidean GCD algotithm. This idea can be generalized to non-monic polynomials. Indeed if we know that GCD of two polynomials is not monic, so the first coefficient of GCD (d_0) is contained in the GCD of the first coefficients (f_0, g_0) of the original polynomials. So we also can equate w-th row to the GCD(f_0, g_0). In this case, we also get a clean GCD without calling the Euclidean algorithm.

Example. $f(x) = (x - 1)^2,\ g(x) = (x - 1)(x - 3)$. $Syl_{2,2}(f, g)$:

$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ -2 & 1 & -4 & 1 \\ 1 & -2 & 3 & -4 \\ 0 & 1 & 0 & 3 \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ d_0 \\ d_1 \end{pmatrix}.$$

Rank of matrix is equals to 3, $k = rank - 1 = 2$. We have next system: $x_1 + x_3 = 0, -2x_1 + x_2 - 4x_3 + x_4 = 0$. Let $x_3 = -1, x_4 = 2$, so $X = [1, -4, -1, 2]$ and $Syl_{2,2}(f, g)X = d = [0, 0, -2, 2]^T$. So, result is $-2x + 2$. But we have 2 monic polynomials. So, we can equate $(k + 1)$-th row to 1. So, we will be able to choose the exact solution: $X = [-0.5, 2, 0.5, -1]$. □

Is it possible to reduce the number of free variables? Yes. We must see to the k-th row with the first coefficient of $d(x)$ and we choose $deg(d)$ variables. We want to nullify component of free variables in the k-th row. In some cases it will reduce the value of the final result. $\sum_{i=t+1}^{n+m} a_{t,i}x_i = 0$, where $Syl_{n,m}(f, g) = (a_{i,j}), t = [k, ..., n + m - 1], k = rank(Syl_{n,m}(f, g))$; parameter t doesn't fixed. So, we have $n + m - 1 - k = deg(d) - 1$ equations and $deg(d)$ variables. All equations depend of only one variable. It's called minimization criteria.

In the example above we have new equation $3x_3 - 4x_4 = 0$. For $x_3 = 1$ we have $X = [-1, 1.25, 1, 0.75]$ and $d'(x) = -3.5x + 3.5$. For $x_3 = 0.5$ we have $X = [-0.5, 5/8, 0.5, 3/8]$ and $d'(x) = -1.75x + 1.75$.

5. CONCLUSION

In this article we presented the new polynomial GCD algorithm based on Bezout's identity. Asymptotics of the algorithm is $O((n+m)^2)$. Also we introduced the dmod operation, which can reduce degree of the largest polynomial and save GCD. This operation is available for polynomials over finite fields. Also we need to analyze the GCD of many polynomials. In the future it is necessary to obtain a condition for the minimization criterion, which can find an exact polynomial GCD without using the Euclidean algorithm to reduce spirious factor.

Next work is related to the algorithm parallelization. Firstly we need good parallel algorithm which solves system of linear equations. Besides it we need to analyze this algorithm over finite fields. Now it works only in $Z[x]$. Also in the future we need to modify this algorithm to test big polynomials for co-primality and construct extended version of algorithm.

REFERENCES

1. B. L. van der Waerden, *Einführung in die Algebraische Geometrie* (Springer, New York, 1973).
2. A. G. Kurosh, *Higher Algebra* (Mir, Moscow, 1980).
3. S. Janson, Resultant and Discriminant of Polynomials, Lecture Notes (2007). www2.math.uu.se/∼svante/papers/sjN5.pdf.
4. K. Weber, "The accelerated integer GCD algorithm," ACM Trans. Math. Software **21**, 111–122 (1995).
5. D. Dolgov, "GCD calculation in the search task of pseudoprime and strong pseudoprime numbers," Lobachevskii J. Math. **37**, 734–739 (2016).
6. E. V. Vinberg, *A Course in Algebra* (Am. Math. Soc., Providence, 2003).
7. B. Li, J. Liu, and L. Zhi, "A structured rank-revealing method for Sylvester matrix," J. Comput. Appl. Math. **213**, 212–223 (2008).

J Syst Sci Complex (2018) 31: 508–526

Computing the Determinant of a Matrix with Polynomial Entries by Approximation*

QIN Xiaolin · SUN Zhi · LENG Tuo · FENG Yong

DOI: 10.1007/s11424-017-6033-8
Received: 17 February 2016 / Revised: 19 July 2016

Abstract Computing the determinant of a matrix with the univariate and multivariate polynomial entries arises frequently in the scientific computing and engineering fields. This paper proposes an effective algorithm to compute the determinant of a matrix with polynomial entries using hybrid symbolic and numerical computation. The algorithm relies on the Newton's interpolation method with error control for solving Vandermonde systems. The authors also present the degree matrix to estimate the degree of variables in a matrix with polynomial entries, and the degree homomorphism method for dimension reduction. Furthermore, the parallelization of the method arises naturally.

Keywords Approximate interpolation, dimension reduction, error controllable algorithm, symbolic determinant, Vandermonde systems.

1 Introduction

In the scientific computing and engineering fields, such as computing multipolynomial resultants[1], computing the implicit equation of a rational plane algebraic curve given by its parametric equations[2], computing Jacobian determinant in multi-domain unified modeling[3], computing the determinant of a matrix with polynomial entries (also called symbolic determinant) is inevitable. Therefore, computing symbolic determinants is an active area of research[4−12]. There are several techniques for calculating the determinant of a matrix with

QIN Xiaolin
Department of Mathematics, Sichuan University, Chengdu 610064, *China; Chengdu Institute of Computer Applications, Chinese Academy of Sciences, Chengdu* 610041, *China.* Email: qinxl@casit.ac.cn.
SUN Zhi
Chengdu Institute of Computer Applications, Chinese Academy of Sciences, Chengdu 610041, *China.*
LENG Tuo
School of Computer Engineering and Science, Shanghai University, Shanghai 200444, *China.*
FENG Yong
Chengdu Institute of Computer Applications, Chinese Academy of Sciences, Chengdu 610041, *China.*
*This research was supported by China 973 Project under Grant No. 2011CB302402, the National Natural Science Foundation of China under Grant Nos. 61402537, 11671377, 91118001, and China Postdoctoral Science Foundation funded project under Grant No. 2012M521692.
◇ *This paper was recommended for publication by Editor-in-Chief GAO Xiao-Shan.*

polynomial entries, such as expansion by minors[8], Gaussian elimination over the integers[9, 10], computing the characteristic polynomial of a matrix[11], and the evaluation and interpolation method[5−7]. The first three algorithms belong to symbolic computations. As is well known, symbolic computations are principally exact and stable. However, they have the disadvantage of intermediate expression swell. The last one is the interpolation method, which as an efficient numerical method has been widely used to compute resultants and determinants. In fact, it is not approximate numerical computations but big number computations, which are also exact computations and only improve intermediate expression swell problem. Furthermore, Chen, et al.[12] presented new conditioners to reduce matrix problems to the computation of minimum polynomials on linear algebra problems over finite fields. Therefore, it is particularly suited to the handling of large sparse or structured matrices over finite fields. In this paper, we propose an efficient approximate interpolation approach to remedy these drawbacks.

Hybrid symbolic-numerical computation is a novel method for solving large scale problems, which applies both numerical and symbolic methods in its algorithms and provides a new perspective of them. The approximate interpolation methods are still used to get the approximate results[13]. In order to obtain exact results, one usually applies exact interpolation methods to improve the intermediate expression swell problem arising from symbolic computations[5–7, 13]. Although the underlying floating-point methods in principle allow for numerical approximations of arbitrary precision, the computed results will never be exact. Recently, the exact computation by intermediate of floating-point arithmetic in symbolic computations has been an active area of research[14−18]. The nice feature of the work is as follows: The initial status and final results are accurate, whereas the intermediate of computation is approximate. The aim of this paper is to provide a rigorous and efficient algorithm to compute symbolic determinants by approximate interpolation. In this paper, we restrict our study to a non-singular square matrix with polynomial entries and the coefficients of polynomial over the integers.

Our main contributions in this paper are the following: Based on the Chio's expansion technique, we construct the degree matrix for estimating the degree of variables in a matrix with the univariate and multivariate polynomial entries, and propose the degree homomorphism method for dimension reduction. We also give the Newton's interpolation method with error control for solving Vandermonde systems. The parallelization of the method arises naturally. Moreover, a real application example is presented.

The rest of this paper is organized as follows. Section 2 first constructs the degree matrix of symbolic determinant on variables and gives Theorem 2.5 to estimate the upper bounds degree of variables, and then analyzes the error controlling for solving Vandermonde systems of equations by Newton's interpolation method, finally proposes the reducing dimension method based on degree homomorphism. Section 3 proposes a novel approach for estimating the upper bound on degree of variables in symbolic determinant, and then presents algorithms of dimension reduction and lifting variables and gives a detailed example. Section 4 gives a practical application and some experimental results. Section 5 makes conclusions.

2 Preliminary Results

Throughout this paper, $\mathbb{Z}$ and $\mathbb{R}$ denote the set of the integers and reals, respectively. There are v variables named x_i, for $i = 1$ to v. Denote the highest degree of each x_i by d_i. Denoted by $\Phi_{m,n}(\mathbb{F})$ the set of all m by n matrices over field $\mathbb{F} = \mathbb{R}$, and abbreviate $\Phi_{n,n}(\mathbb{F})$ to $\Phi_n(\mathbb{F})$.

2.1 Estimating Degree of Variables

In this subsection, a brief description to Chio's expansion is proposed. We also give Theorem 2.5 for estimating the upper bound on degree of variables in symbolic determinant.

Lemma 2.1 (see [19]) *Let $A = [a_{ij}]$ be an $n \times n$ matrix and suppose $a_{11} \neq 0$. Let K denote the matrix obtained by replacing each element a_{ij} in A by*

$$\begin{vmatrix} a_{11} & a_{1j} \\ a_{i1} & a_{ij} \end{vmatrix}.$$

Then $|A| = |K|/a_{11}^{n-2}$. That is,

$$|A| = \frac{1}{a_{11}^{n-2}} \begin{vmatrix} \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} & \begin{vmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{vmatrix} & \cdots & \begin{vmatrix} a_{11} & a_{1n} \\ a_{21} & a_{2n} \end{vmatrix} \\ \begin{vmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{vmatrix} & \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} & \cdots & \begin{vmatrix} a_{11} & a_{1n} \\ a_{31} & a_{3n} \end{vmatrix} \\ \vdots & \vdots & \ddots & \vdots \\ \begin{vmatrix} a_{11} & a_{12} \\ a_{n1} & a_{n2} \end{vmatrix} & \begin{vmatrix} a_{11} & a_{13} \\ a_{n1} & a_{n3} \end{vmatrix} & \cdots & \begin{vmatrix} a_{11} & a_{1n} \\ a_{n1} & a_{nn} \end{vmatrix} \end{vmatrix}.$$

Remark 2.2 The proof of Lemma 2.1 is clear. Multiply each row of A by a_{11} except the first, and then perform the elementary row operations, denote $Op(2 - a_{21} \cdot 1)$, $Op(3 - a_{31} \cdot 1)$, $\cdots$, $Op(n - a_{n1} \cdot 1)$, where '1', '2', $\cdots$, 'n' represent for the row index. We can get

$$\begin{aligned} a_{11}^{n-1}|A| &= \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{11}a_{21} & a_{11}a_{22} & \cdots & a_{11}a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{11}a_{n1} & a_{11}a_{n2} & \cdots & a_{11}a_{nn} \end{vmatrix} \\ &= \begin{vmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} & \begin{vmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{vmatrix} & \cdots & \begin{vmatrix} a_{11} & a_{1n} \\ a_{21} & a_{2n} \end{vmatrix} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \begin{vmatrix} a_{11} & a_{12} \\ a_{n1} & a_{n2} \end{vmatrix} & \begin{vmatrix} a_{11} & a_{13} \\ a_{n1} & a_{n3} \end{vmatrix} & \cdots & \begin{vmatrix} a_{11} & a_{1n} \\ a_{n1} & a_{nn} \end{vmatrix} \end{vmatrix} = a_{11}|K|. \end{aligned}$$

We observe that K is $(n-1)\times(n-1)$ matrix, then the above procedure can be repeated until the K is 2×2 matrix. It is a simple and straightforward method for calculating the determinant of a numerical matrix.

Lemma 2.3 *Given two polynomials $f(x_1)$ and $g(x_1)$, the degree of the product of two polynomials is the sum of their degrees, i.e.,*

$$\deg(f(x_1)\cdot g(x_1), x_1) = \deg(f(x_1), x_1) + \deg(g(x_1), x_1).$$

The degree of the sum (or difference) of two polynomials is equal to or less than the greater of their degrees, i.e.,

$$\deg(f(x_1) \pm g(x_1), x_1) \le \max\{\deg(f(x_1), x_1), \deg(g(x_1), x_1)\},$$

where $f(x_1)$ and $g(x_1)$ are the univariate polynomials over field $\mathbb{F}$, and $\deg(f(x_1), x_1)$ represents the highest degree of x_1 in $f(x_1)$.

Definition 2.4 Let $M = [M_{ij}]$ be an $n\times n$ matrix and suppose M_{ij} is a polynomial with integer coefficients consisting of variables $[x_1, x_2, \cdots, x_v]$, where the order of M is $n \ge 2$. Without loss of generality, we call it the degree matrix $\Omega_1 = (\sigma_{ij})$ † for x_1 defined as:

$$\sigma_{ij} = \begin{cases} \text{highest degree of } x_1 \text{ appears in the element } M_{ij}, \text{i.e.}, \deg(M_{ij}, x_1), \\ 0, \quad \text{if } x_1 \text{ does not occur in } M_{ij}. \end{cases}$$

Then we can construct the degree matrices Ω_i from M for each x_i, $2 \le i \le v$, respectively.

Theorem 2.5 *M is the same as Definition* 2.4. *Suppose that the 2×2 degree matrix can be obtained from M for each x_i $(1 \le i \le v)$, denotes*

$$\Omega_i = \begin{bmatrix} \sigma^{(n-2)}_{(n-1)(n-1)} & \sigma^{(n-2)}_{(n-1)n} \\ \sigma^{(n-2)}_{n(n-1)} & \sigma^{(n-2)}_{nn} \end{bmatrix},$$

then

$$\max\deg = \max\left\{\sigma^{(n-2)}_{(n-1)(n-1)} + \sigma^{(n-2)}_{nn}, \sigma^{(n-2)}_{(n-1)n} + \sigma^{(n-2)}_{n(n-1)}\right\}.$$

That is, the highest degree of variable is no more than

$$\max\deg - \sum_{i=3}^{n}(i-2)\sigma^{(n-i)}_{(n-i+1)(n-i+1)},$$

where $\sigma^{(n-2)}_{(n-1)(n-1)} = \deg(M^{(n-2)}_{(n-1)(n-1)}, x_i)$‡, $M^{(n-2)}_{(n-1)(n-1)}$ refers to the following proof.

Proof Considering the order n of symbolic determinant

$$|M| = \begin{vmatrix} M_{11} & M_{12} & \cdots & M_{1n} \\ M_{21} & M_{22} & \cdots & M_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ M_{n1} & M_{n2} & \cdots & M_{nn} \end{vmatrix},$$

† $\Omega_1, \Omega_2, \cdots, \Omega_v$ denote the degree matrix of $[x_1, x_2, \cdots, x_v]$, respectively.

‡ $\sigma^{(\cdot)}_{ij}$ is defined by the same way for the rest of this paper.

by Chio's expansion from Remark 2.2, then

$$\begin{aligned}|M| &= \frac{1}{M_{11}^{n-2}}\begin{vmatrix} M_{22}^{(1)} & M_{23}^{(1)} & \cdots & M_{2n}^{(1)} \\ M_{32}^{(1)} & M_{33}^{(1)} & \cdots & M_{3n}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ M_{n2}^{(1)} & M_{n3}^{(1)} & \cdots & M_{nn}^{(1)} \end{vmatrix} \\ &= \cdots \\ &= \frac{1}{M_{11}^{n-2}}\frac{1}{{M_{22}^{(1)}}^{n-3}}\cdots\frac{1}{M_{(n-2)(n-2)}^{(n-3)}}\begin{vmatrix} M_{(n-1)(n-1)}^{(n-2)} & M_{(n-1)n}^{(n-2)} \\ M_{n(n-1)}^{(n-2)} & M_{nn}^{(n-2)} \end{vmatrix},\end{aligned}$$

where

$$M_{22}^{(1)} = M_{11}M_{22}-M_{12}M_{21}, \quad M_{32}^{(1)} = M_{11}M_{32}-M_{12}M_{31}, \quad \cdots, \quad M_{nn}^{(1)} = M_{11}M_{nn}-M_{1n}M_{n1}.$$

By Lemma 2.3, for each x_i, we can get

$$\begin{aligned}\deg(|M|, x_i) &\le \max\left\{\sigma_{(n-1)(n-1)}^{(n-2)} + \sigma_{nn}^{(n-2)}, \sigma_{(n-1)n}^{(n-2)} + \sigma_{n(n-1)}^{(n-2)}\right\} \\ &\quad -(n-2)\sigma_{11} - (n-3)\sigma_{22}^{(1)} - \cdots - \sigma_{(n-2)(n-2)}^{(n-3)} \\ &= \max\deg - \sum_{i=3}^{n}(i-2)\sigma_{(n-i+1)(n-i+1)}^{(n-i)},\end{aligned}$$

where

$$\max\deg = \max\left\{\sigma_{(n-1)(n-1)}^{(n-2)} + \sigma_{nn}^{(n-2)}, \sigma_{(n-1)n}^{(n-2)} + \sigma_{n(n-1)}^{(n-2)}\right\}.$$

The proof of Theorem 2.5 is completed. ∎

Remark 2.6 We present a direct method to estimate the upper bound on degrees of variables by computation of the degree matrices. Our method only needs the simple recursive arithmetic operations of addition and subtraction. In fact, in many cases, we can obtain the exact degrees of all variables in symbolic determinant.

2.2 Newton's Interpolation with Error Control

Let M be defined as above. Without loss of generality, we consider the determinant of a matrix with bivariate polynomial entries, and then generalize the results to the univariate or multivariate polynomial. A good introduction to the theory of interpolation can be seen in [20].

Definition 2.7 The Kronecker product of $A = [a_{i,j}] \in \Phi_{m,n}(\mathbb{F})$ and $B = [b_{ij}] \in \Phi_{p,q}(\mathbb{F})$ is denoted by $A \otimes B$ and is defined to the block matrix

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix} \in M_{mp,nq}(\mathbb{F}). \tag{1}$$

Notice that $A \otimes B \neq B \otimes A$ in general.

Definition 2.8 With each matrix $A = [a_{ij}] \in \Phi_{m,n}(\mathbb{F})$, we associate the vector $\mathrm{vec}(A) \in \mathbb{F}^{mn}$ defined by

$$\mathrm{vec}(A) \equiv [a_{11}, a_{21}, \cdots a_{m1}, a_{12}, a_{22}, \cdots, a_{m2}, \cdots, a_{1n}, a_{2n}, \cdots, a_{mn}]^{\mathrm{T}},$$

where $^{\mathrm{T}}$ denotes the transpose of matrix or vector.

Let the determinant of M be $f(x_1, x_2) = \sum_{i,j} a_{ij} x_1^i x_2^j$ that is a polynomial with integer coefficients, and d_1, d_2[§] be the bounds on the highest degree of $f(x_1, x_2)$ with x_1, x_2, respectively. We choose the distinct scalars (x_{1i}, x_{2j}) $(i = 0, 1, \cdots, d_1; j = 0, 1, \cdots, d_2)$, and obtain the values of $f(x_1, x_2)$, denoted by $f_{ij} \in \mathbb{R}$ $(i = 0, 1, \cdots, d_1; j = 0, 1, \cdots, d_2)$. The set of monomials is ordered as follows:

$$[1, x_1, x_1^2, \cdots, x_1^{d_1}] \times [1, x_2, x_2^2, \cdots, x_2^{d_2}]$$

and the distinct scalars in the corresponding order is as follows:

$$[x_{10}, x_{11}, \cdots, x_{1d_1}] \times [x_{20}, x_{21}, \cdots, x_{2d_2}].$$

Based on the bivariate interpolate polynomial technique, which is essential to solve the following linear system:

$$(V_{x_1} \otimes V_{x_2})\mathrm{vec}(a) = \mathrm{vec}(F), \tag{2}$$

where the coefficients V_{x_1} and V_{x_2} are Vandermonde matrices:

$$V_{x_1} = \begin{pmatrix} 1 & x_{10} & x_{10}^2 & \cdots & x_{10}^{d_1} \\ 1 & x_{11} & x_{11}^2 & \cdots & x_{11}^{d_1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1d_1} & x_{1d_1}^2 & \cdots & x_{d_1}^{1d_1} \end{pmatrix}, \quad V_{x_2} = \begin{pmatrix} 1 & x_{20} & x_{20}^2 & \cdots & x_{20}^{d_2} \\ 1 & x_{21} & x_{21}^2 & \cdots & x_{21}^{d_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{2d_2} & x_{2d_2}^2 & \cdots & x_{2d_2}^{d_2} \end{pmatrix},$$

and

$$a = \begin{pmatrix} a_{00} & a_{01} & \cdots & a_{0d_2} \\ a_{10} & a_{11} & \cdots & a_{1d_2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{d_10} & a_{d_11} & \cdots & a_{d_1d_2} \end{pmatrix}, \quad F = \begin{pmatrix} f_{00} & f_{01} & \cdots & f_{0d_2} \\ f_{10} & f_{11} & \cdots & f_{1d_2} \\ \vdots & \vdots & \ddots & \vdots \\ f_{d_10} & f_{d_11} & \cdots & f_{d_1d_2} \end{pmatrix}.$$

Marco and Martínez[5] have proved in this way that the interpolation problem has a unique solution. This means that V_{x_1} and V_{x_2} are nonsingular and $V = V_{x_1} \otimes V_{x_2}$, then the coefficient matrix of the linear system (2) is nonsingular. The following lemma shows us how to solve the system (2).

Lemma 2.9 (see [21]) *Let $\mathbb{F}$ denote a field. Matrices $A \in \Phi_{m,n}(\mathbb{F})$, $B \in \Phi_{q,p}(\mathbb{F})$, and $C \in \Phi_{m,q}(\mathbb{F})$ are given and assume $X \in \Phi_{n,p}(\mathbb{F})$ to be unknown. Then, the following equation:*

$$(B \otimes A)\mathrm{vec}(X) = \mathrm{vec}(C) \tag{3}$$

[§] d_1, d_2 are defined by the same way for the rest of this paper.

is equivalent to matrix equation:

$$AXB^{\mathrm{T}} = C. \tag{4}$$

Obviously, Equation (4) *is equivalent to the system of equations*

$$\begin{cases} AY = C, \\ BX^{\mathrm{T}} = Y^{\mathrm{T}}. \end{cases} \tag{5}$$

We notice that the coefficients of System (2) are Vandermonde matrices. Here we give a progressive algorithm, which is significantly more efficient than previous available methods in $O(d_1^2)$ arithmetic operations by the Newton's interpolation method[22] in Algorithm 1.

Algorithm 1 (Björck and Pereyra algorithm)

Input: A set of distinct scalars $(x_i, f_i), 0 \le i \le d_1$;

Output: The solution of coefficients $a_0, a_1, \cdots, a_{d_1}$.

Step 1: $c_i^{(0)} := f_i, i = 0, 1, \cdots, d_1$

for $k = 0$ to $d_1 - 1$ **do**

$c_i^{(k+1)} := \frac{c_i^{(k)} - c_{i-1}^{(k)}}{x_i - x_{i-k-1}}, i = d_1, d_1 - 1, \cdots, k + 1$

end for

Step 2: $a_i^{(d_1)} := c_i^{(d_1)}, i = 0, 1, \cdots, d_1$

for $k = d_1 - 1$ to 0 by -1 **do**

$a_i^{(k)} := a_i^{(k+1)} - x_k a_{i+1}^{(k+1)}, i = k, k + 1, \cdots, d_1 - 1$

end for

Step 3: Return $a_i := a_i^{(0)}, i = 0, 1, \cdots, d_1$.

In general, we can compute the equation (2) after choosing d_1+1 distinct scalars $[x_{10}, x_{11}, \cdots, x_{1d_1}]$ and $d_2 + 1$ distinct scalars $[x_{20}, x_{21}, \cdots, x_{2d_2}]$, then obtain their corresponding exact values $[f_{00}, f_{01}, \cdots, f_{0d_2}, \cdots, f_{10}, f_{11}, \cdots, f_{1d_2}, \cdots, f_{d_1 0}, f_{d_1 1}, \cdots, f_{d_1 d_2}]$. However, in order to improve intermediate expression swell problem arising from symbolic computations and avoid big integer computation, we can get the approximate values of $f(x_1, x_2)$, denoted by $[\widetilde{f}_{00}, \widetilde{f}_{01}, \cdots, \widetilde{f}_{0d_2}, \widetilde{f}_{10}, \widetilde{f}_{11}, \cdots, \widetilde{f}_{1d_2}, \widetilde{f}_{d_1 0}, \widetilde{f}_{d_1 1}, \cdots, \widetilde{f}_{d_1 d_2}]$.

Based on Algorithm 1, together with Lemma 2.9 we can obtain the approximate solution $\widetilde{a} = [\widetilde{a}_{ij}]$ $(i = 0, 1, \cdots, d_1; j = 0, 1, \cdots, d_2)$. Therefore, an approximate bivariate polynomial $\widetilde{f}(x_1, x_2) = \sum_{i,j} \widetilde{a}_{ij} x_1^i x_2^j$ is only produced. However, we usually need the exact results in practice. Next, our main task is to bound the error between approximate coefficients and exact values, and discuss the controlling error ε in Algorithm 1. Feng, et al.[16] gave a preliminary result of this problem. Here, we present a necessary condition on error controlling ε in floating-point arithmetic. In Step 1 of Algorithm 1, it is the standard method for evaluating divided differences $(c_k^{(k)} = f[x_0, x_1, \cdots, x_k])$. We consider the relation on the $f_{ij} - \widetilde{f}_{ij}$ with $a_{ij} - \widetilde{a}_{ij}$ and the propagation of rounding errors in divided difference schemes. We have the following theorem to answer the above question.

Lemma 2.10 *c_i and f_i are defined as in Algorithm 1, and $\widetilde{c}_i$, $\widetilde{f}_i$ are their approximate values by approximate interpolation, $\lambda = \min\{|x_{1i} - x_{1j}| : i \neq j\}$ $(0 < \lambda < 1)$. Then*

$$|c_i - \widetilde{c}_i| \leq \Big(\frac{2}{\lambda}\Big)^{d_1} \max\{|f_i - \widetilde{f}_i|\}.$$

Proof From Algorithm 1, we observe that Step 1 is recurrences for $c_i^{(k+1)}$ $(k = 0, 1, \cdots, d_1 - 1, i = d_1, d_1 - 1, \cdots, k + 1)$, whose form is as follows:

$$c_i^{(d_1)} = \frac{1}{\lambda}(c_i^{(d_1-1)} - c_{i-1}^{(d_1-1)}).$$

However, when we operate the floating-point arithmetic in Algorithm 1, which is recurrences for $\widetilde{c}_i^{(k+1)}$, which form is as follows:

$$\widetilde{c}_i^{(d_1)} = \frac{1}{\lambda}(\widetilde{c}_i^{(d_1-1)} - \widetilde{c}_{i-1}^{(d_1-1)}).$$

Therefore,

$$\begin{aligned} |c_i^{(d_1)} - \widetilde{c}_i^{(d_1)}| &= \frac{1}{\lambda}|c_i^{(d_1-1)} - \widetilde{c}_i^{(d_1-1)} + \widetilde{c}_{i-1}^{(d_1-1)} - c_{i-1}^{(d_1-1)}| \\ &\leq \frac{1}{\lambda}(|c_i^{(d_1-1)} - \widetilde{c}_i^{(d_1-1)}| + |c_{i-1}^{(d_1-1)} - \widetilde{c}_{i-1}^{(d_1-1)}|). \end{aligned}$$

The bounds are defined by the following recurrences,

$$|c_i^{(d_1)} - \widetilde{c}_i^{(d_1)}| \leq \frac{2}{\lambda}|c_{i-1}^{(d_1-1)} - \widetilde{c}_{i-1}^{(d_1-1)}| \leq \cdots \leq \Big(\frac{2}{\lambda}\Big)^{d_1} \max\{|f_i - \widetilde{f}_i|\}.$$

This completes the proof of the lemma. ∎

Theorem 2.11 *Let $\varepsilon = \max\{|f_{ij} - \widetilde{f}_{ij}|\}$, $\lambda = \min\{|x_{1i} - x_{1j}|, |x_{2i} - x_{2j}| : i \neq j\}$ $(0 < \lambda < 1)$. Then*

$$\max\{|a_{ij} - \widetilde{a}_{ij}|\} \leq \Big(\frac{2}{\lambda}\Big)^{d_1}\Big(\frac{2}{\lambda}\Big)^{d_2}\varepsilon.$$

Proof From Equation (2), it holds that

$$V \cdot \mathrm{vec}(\widetilde{a} - a) = \mathrm{vec}(\widetilde{F} - F),$$

where $V = V_{x_1} \otimes V_{x_2}$. By Lemma 2.9, the above equation is equivalent to the following equation:

$$V_{x_2} \cdot (\widetilde{a} - a) \cdot V_{x_1}^{\mathrm{T}} = \widetilde{F} - F.$$

Thus, it is equivalent to

$$V_{x_2} \cdot z = \widetilde{F} - F, \tag{6a}$$

$$V_{x_1} \cdot (\widetilde{a} - a)^{\mathrm{T}} = z^{\mathrm{T}}, \tag{6b}$$

where $z = [z_{ij}]$. Matrix equation (6a) is equivalent to

$$V_{x_2} \cdot z_{\cdot i} = \widetilde{F}_{i\cdot} - F_{i\cdot}, \quad i = 1, 2, \cdots d_2 + 1, \tag{7}$$

where $z_{\cdot i}$ stands for the i-th column of z and $F_{i\cdot}$ stands for the i-th row of matrix F.

From Lemma 2.10 and Algorithm 1, it holds that

$$\max_{j=0}^{d_2}|z_{ji}| < \Big(\frac{2}{\lambda}\Big)^{d_2}|f_{i\cdot} - \widetilde{f}_{i\cdot}|, \text{ for each } i.$$

Hence, we conclude that

$$\max_{i,j}|z_{ji}| < \Big(\frac{2}{\lambda}\Big)^{d_2}|f_{i\cdot} - \widetilde{f}_{i\cdot}|.$$

Let $\delta = (\frac{2}{\lambda})^{d_2}|f_{i\cdot} - \widetilde{f}_{i\cdot}|$, argue Equation (6b) in the same technique as do above, we deduce that

$$\max_{i,j}|a_{ij} - \widetilde{a}_{ij}| \le \Big(\frac{2}{\lambda}\Big)^{d_1}\Big(\frac{2}{\lambda}\Big)^{d_2}\varepsilon.$$

The proof is finished. ∎

In order to avoid the difficulty of computations, we restrict our study to the coefficients of polynomial over $\mathbb{Z}$. Therefore, we only require to solving the Vandermonde system and taking the nearest integer to each component of the solution. The less degree of bounds on variables we obtain, the less the amount of computation is for obtaining approximate multivariate polynomial. Once an upper bound d_1 and d_2 are gotten, we choose $(d_1+1)\cdot(d_2+1)$ interpolate nodes and calculate

$$\varepsilon = 0.5\Big(\frac{\lambda}{2}\Big)^{d_1+d_2}. \tag{8}$$

Then, compute the values $\widetilde{f}_{ij} \approx f(x_{1i}, x_{2j})$ for $i = 0, 1, \cdots, d_1$, $j = 0, 1, \cdots, d_2$ with an error less than ε. By interpolation method, we compute the approximate interpolation polynomial $\widetilde{f}(x_1, x_2)$ with coefficient error less than 0.5.

As for the generalization of the algorithm to the case $v > 2$, we can say that the situation is completely analogous to the bivariate case. It comes down to solving the following system:

$$\underbrace{(V_{x_1} \otimes V_{x_2} \cdots \otimes V_{x_v})}_{v}\,\mathrm{vec}(a) = \mathrm{vec}(F). \tag{9}$$

Of course, we can reduce the multivariate polynomial entries to bivariate ones on symbolic determinant. For more details refer to Subsection 2.3.

We can analyze the computational complexity of the derivation of above algorithm. For the analysis of floating-point arithmetic operations, the result is similar with the exact interpolation situation[5]. However, our method can enable the practical processing of symbolic computations in applications.

Remark 2.12 Our result is superior to the literature[16]. Here we make full use of advantage of arbitrary precision of floating-point arithmetic operations on modern computer and symbolic computation platform, such as Maple. In general, it seems as if at least some problems connected with Vandermonde systems, which traditionally have been considered too ill-conditioned to be attached, actually can be solved with good precision.

2.3 Reducing Dimension Method

As the variables increased, the storage of computations expands severely when calculated high order on symbolic determinant. Moenck[23] proposed a practical method to map the multivariate problem into a univariate one. For the general case, the validity of the method is established by the following lemma.

Lemma 2.13 (see [23]) *In the polynomial ring $R[x_1, x_2, \cdots, x_v], v > 2$. The mapping:*

$$\begin{aligned} &\phi : R[x_1, x_2, \cdots, x_v] \to R[x_1], \\ &\phi : x_i \mapsto x_1^{n_i}, \quad 1 \le i \le v, \end{aligned}$$

where $n_v > n_{v-1} > \cdots > n_1 = 1$ is a homomorphism of rings.

Let $d_i(f(x_1, x_2, \cdots, x_v))$ be the highest degree of the polynomial $f(x_1, x_2, \cdots, x_v)$ with x_i. The following lemma relates the n_i of the mapping to d_i and establishes the validity of the inverse mapping.

Lemma 2.14 (see [23]) *Let ψ be the homomorphism of free R-modules defined by:*

$$\begin{aligned} &\psi : R[x_1] \to R[x_1, x_2, \cdots, x_v], \\ &\psi : x_1^k \mapsto \begin{cases} 1, & \text{if } k = 0, \\ \psi(x_1^r) \cdot x_i^q, & \text{otherwise}, \end{cases} \end{aligned}$$

where $n_{i+1} > k \ge n_i, k = q \cdot n_i + r, 0 \le r < n_i$ and $n_v > n_{v-1} > \cdots > n_1 = 1$. Then for all $f(x_1, x_2, \cdots, x_v) \in R[x_1, x_2, \cdots, x_v], \psi(\phi(f)) = f$, and for all i if and only if

$$\sum_{j=1}^{i} d_j(f) n_j < n_{i+1}, \quad 1 \le i < v. \tag{10}$$

Remark 2.15 We apply the degree homomorphism method to reduce dimension for computing the determinant of a matrix with multivariate polynomial entries, which is distinguished from the practical fast polynomial multiplication[23]. We note that relation (10) satisfying is isomorphic to their univariate images. Therefore, any polynomial ring operation on entries of symbolic determinant, giving results in the determinant, will be preserved by the isomorphism. In this sense ϕ behaves like a ring isomorphism on the symbolic determinant of polynomials. Another way to view the mapping given in the theorems is

$$\phi : x_i \mapsto x_{i-1}^{n_i}, \quad 2 \le i \le v.$$

3 Derivation of the Algorithm

Following our preliminary results in Section 2, the aim of this section is to describe a novel algorithm for estimating the degree of variables on symbolic determinant, and the degree homomorphism method for dimension reduction.

3.1 Description of Algorithm

Algorithm 2 is to estimate the degree of variables on symbolic determinant by computation of the degree matrix, and Algorithms 3 and 4 are applied to reduce dimension and lift variables.

Algorithm 2 (Estimating degree of variables algorithm)

Input: Given the order n of symbolic determinant M, list of variables var$s = [x_1, x_2, \cdots, x_v]$;
Output: The exact degrees or upper bounds on degree of variables.

Step 1: Select variable from var$s$ and repeat the following steps

loop
  Obtain the degree matrix $\Omega = (\sigma_{ij})$ from M by Definition 2.4, $1 \le i, j \le n$;
  if order(Ω)=2 **then**
    $maxdeg := \max\{\sigma_{11} + \sigma_{22}, \sigma_{12} + \sigma_{21}\}$
  else
    for $i = 1$ to $n - 1$ **do**
      for $j = 1$ to $n - 1$ **do**
        $temp := \sigma_{i1} + \sigma_{1j}$
        $\sigma_{ij} := \max\{\sigma_{ij} + \sigma_{11}, temp\}$
      end for
    end for
  end if
  for $i = 1$ to $n - 2$ **do**
    $maxdeg := maxdeg - \sigma_{11}$
  end for
  Return $maxdeg$
end loop

Theorem 3.1 *Algorithm* 2 *works correctly as specified and its complexity is* $O(n^2)$*, where* n *is the order of symbolic determinant.*

Proof Correctness of the algorithm follows from Theorem 2.5. The number of arithmetic operations needs to execute $(n - 1) \times (n - 1)$ additions and simultaneous comparisons, and remains $n - 2$ substructions and one comparison by using degree matrix. Therefore, the total arithmetic operations are $n^2 - n$, that is, $O(n^2)$. ∎

Algorithm 3 (Reducing dimension algorithm)

Input: Given the order n of symbolic determinant M, list of variables vars $= [x_1, x_2, \cdots, x_v]$;
Output: The order n of symbolic determinant M' with bivariate polynomial entries.

Step 1: Call Algorithm 2 to evaluate the bounds on degree of the variables in M, denoted by d_i, $1 \le i \le v$.

Step 2: Reducing dimension

1: Divide the vars into the partitions: $[x_1, x_2, \cdots, x_t], [x_{t+1}, x_{t+2}, \cdots, x_v]$;
2: **for** $i = t-1$ to 1 by -1 **do**
3: $\quad D_i := \prod_{j=i+1}^{t}(d_j + 1)$, $x_i := x_t^{D_i}$
4: **end for**
5: **for** $i = v-1$ to $t+1$ by -1 **do**
6: $\quad D_i := \prod_{j=i+1}^{v}(d_j + 1)$, $x_i := x_v^{D_i}$
7: **end for**

Step 3: Obtain the symbolic determinant M' with $[x_t, x_v]$ of polynomial entries.

Step 4: Return M'.

Remark 3.2 The beauty of reducing dimension algorithm is a substitution trick. In Algorithm 3, $t = \text{ceil}(\frac{n}{2})$, where ceil$(\cdot)$ is a function that returns the smallest following integer. We note that the lexicographic order $x_v \succ x_{v-1} \succ \cdots \succ x_1$ and divide the $[x_1, x_2, \cdots, x_v]$ into two parts. Then the symbolic determinant can be translated into the entries with bivariate polynomial. The reducing dimension algorithm can be used for highly parallel computation when the number of variables is more than three.

Remark 3.3 Based on Algorithm 2, we can estimate the bounds on degree of variables. Then we can reduce dimension for multivariate case to bivariate one by using Algorithm 3. We can solve the Vandermonde coefficient matrix of linear equations with error controlling by using Algorithm 1, and finally lift variables to recover the multivariate polynomial by using Algorithm 4.

In this paper, we consider the general symbolic determinant, which is not sparse. Applying the substitutions to the matrix entries as described above and assuming the monomial exists in the determinant then the bivariate form of unknown polynomial is a highest degree of

$$D = \sum_{i=1}^{\text{ceil}(\frac{n}{2})} \left(d_i \cdot \prod_{k=i+1}^{\text{ceil}(\frac{n}{2})} (d_k + 1) \right). \tag{11}$$

While this upper bound on degree of variable is often much larger than needed, which is the worst case and thus is suitable to all cases.

Algorithm 4 (Lifting variables algorithm)

Input: Given the set of monomial with $[x_t, x_v]$ in L;

Output: The determinant of a matrix with $[x_1, x_2, \cdots, x_v]$ of polynomial entries.

Step 1: Obtain the corresponding power set with $[x_t, x_v]$, respectively.

Step 2: Lifting variables

1: Call Algorithm 3, extract the power D_i, $1 \le i \le t-1, t+1 \le i \le v-1$;
2: **while** nops(L)$\neq$ NULL **do**
3: $temp := deg(x_t)$
4: **for** $i = 1$ to $t-1$ by 1 **do**
5: $d_i := iquo(temp, D_i), temp := irem(temp, D_i)$
6: **end for**
7: $d_i := temp, temp := deg(x_v)$
8: **for** $i = t+1$ to $v-1$ by 1 **do**
9: $d_i := iquo(temp, D_i), temp := irem(temp, D_i)$
10: **end for**
11: $d_i := temp$
12: **end while**

Step 3: Obtain the new set of monomial L' with $[x_1, x_2, \cdots, x_v]c$;

Step 4: Return L'.

3.2 A Small Example in Detail

Example 3.4 For convenience and space-saving purposes, we choose the symbolic determinant is three variables and order 2 as follows:

$$|M| = \begin{vmatrix} 5x_1^2 - 3x_1x_2 + 2x_3^2 & -9x_1 - 3x_2^2 - x_3^2 \\ -x_1 + x_2 + 3x_2x_3 & x_3 - 4x_2^2 \end{vmatrix},$$

At first, based on Algorithm 2 we estimate the degree on x_1, x_2, x_3. For the variable x_1, we get

$$\Omega_1 = \begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix}.$$

Then

$$\max\{2+0, 1+1\} = 2.$$

Therefore, the maximum degree of the variable x_1 is 2. As the same technique for x_2, x_3, we can get 3 and 3.

Call Algorithm 3, by substituting $x_1 = x_2^4$, we get

$$|M'| = \begin{vmatrix} 5x_2^8 - 3x_2^5 + 2x_3^2 & -9x_2^4 - 3x_2^2 - x_3^2 \\ -x_2^4 + x_2 + 3x_2x_3 & x_3 - 4x_2^2 \end{vmatrix}.$$

Then, based on Algorithm 2 we again estimate the degree on x_2, x_3 for $[10, 3]$.

Based on the derivation of algorithm in Subsection 3.1 and Algorithm 1, computing exact polynomial $f(x_2, x_3)$ as follows: Choose the different floating-point interpolation nodes by using the distance between two points 0.5; $\lambda = 0.5$, compute $\varepsilon = 0.745 \times 10^{-8}$ from Theorem 2.11. Compute the approximate interpolate datum $\widetilde{f_{ij}}$ such that $|f_{ij} - \widetilde{f_{ij}}| < \varepsilon$. We get the following approximate bivariate polynomial:

$$\begin{aligned}
&4.99995826234x_2^8x_3 - 20.0000018736x_2^{10} + 24.0010598569x_2^5x_3 + 12.0025760656x_2^7\\
&+2.00000000000x_3^3 - 8.00094828634x_2^2x_3^2 - 9.00045331720x_2^8 + 9.01977448800x_2^5\\
&-3.00897542075x_2^6 + 3.02270681750x_2^3 + 9.00076124850x_2^3x_3 - 1.00207248277x_2^4x_3^2\\
&+1.00018098282x_2x_3^2 + 2.99986559933x_2x_3^3.
\end{aligned}$$

Next, based on Algorithm 4 we lift the variables to obtain the following multivariate polynomial:

$$\begin{aligned}
&4.99995826234x_1^2x_3 - 20.0000018736x_2^2x_1^2 + 24.0010598569x_1x_2x_3 + 12.0025760656x_2^3x_1\\
&+2.00000000000x_3^3 - 8.00094828634x_2^2x_3^2 - 9.00045331720x_1^2 + 9.01977448800x_1x_2\\
&-3.00897542075x_2^2x_1 + 3.02270681750x_2^3 + 9.00076124850x_2^3x_3 - 1.00207248277x_1x_3^2\\
&+1.00018098282x_2x_3^2 + 2.99986559933x_2x_3^3.
\end{aligned}$$

Finally, we easily recover the integer coefficients of above approximate polynomial to the nearest values as follows:

$$\begin{aligned}
&5x_1^2x_3 - 20x_1^2x_2^2 + 24x_1x_2x_3 + 12x_1x_2^3 + 2x_3^3 - 8x_3^2x_2^2\\
&-9x_1^2 + 9x_1x_2 - 3x_2^2x_1 + 3x_2^3 + 9x_2^3x_3 - x_3^2x_1 + x_3^2x_2 + 3x_3^3x_2.
\end{aligned}$$

4 Experimental Results

Our algorithms are implemented in Maple 15. Example 4.1 is a practical application to selective harmonic elimination in power electronics. In Figures 2 and 3, we present the running time (Time) and memory usage (RAM) of computing for symbolic determinants to compare our method with symbolic method (det, see Maple's help), and exact interpolation method[5–7]. The following examples run in the same platform of Maple 15 under Windows and AMD Athlon (tm) 2.70 Ghz, 2.00 GB of main memory.

Example 4.1 Consider a practical switching angles in a multilevel converter example from [24], see Figure 1.

In Figure 1, the Fourier series expansion of the output voltage waveform is as follows:

$$V(\omega t) = \frac{4V_{dc}}{\pi} \sum_{n=1,3,5,\cdots}^{\infty} \frac{1}{n} \times (\cos(n\theta_1) + \cos(n\theta_2) + \cdots + \cos(n\theta_s)) \sin(n\omega t), \tag{12}$$

where s is the number of dc sources. The goal is to choose the switching angles $0 \le \theta_1 < \theta_2 < \cdots < \theta_s \le \pi/2$ so as to make the first harmonic equal to the given desired fundamental voltage

V_1 and the specific higher harmonics of $V(\omega t)$ equal to zero. In particular, in the case of $s = 5$ dc sources, the mathematical statement of Equation (12) for a three-phase system is

$$\left.\begin{aligned} \frac{4V_{dc}}{\pi}(\cos(\theta_1) + \cos(\theta_2) + \cdots + \cos(\theta_5)) &= V_1, \\ \cos(5\theta_1) + \cos(5\theta_2) + \cdots + \cos(5\theta_5) &= 0, \\ \cos(7\theta_1) + \cos(7\theta_2) + \cdots + \cos(7\theta_5) &= 0, \\ \cos(11\theta_1) + \cos(11\theta_2) + \cdots + \cos(11\theta_5) &= 0, \\ \cos(13\theta_1) + \cos(13\theta_2) + \cdots + \cos(13\theta_5) &= 0. \end{aligned}\right\} \tag{13}$$

Here, it is a system of five transcendental equations in the five unknowns $[\theta_1, \theta_2, \cdots, \theta_5]$. The goal is to determine when the equations of (13) have a solution. Since a Chebyshev expansion is related to a Fourier cosine series, define $x_i = \cos\theta_i\,(i = 1, 2, \cdots, 5)$, we can transform the equations (13) into the nonlinear polynomial systems as follows:

$$\left.\begin{aligned} f_1 &= \sum_{i=1}^{5}(x_i) - V_1/(4V_{dc}/\pi), \\ f_2 &= \sum_{i=1}^{5}(16x_i^5 - 20x_i^3 + 5x_i), \\ f_3 &= \sum_{i=1}^{5}(64x_i^7 - 112x_i^5 + 56x_i^3 - 7x_i), \\ f_4 &= \sum_{i=1}^{5}(1024x_i^{11} - 2816x_i^9 + 2816x_i^7 - 1232x_i^5 + 220x_i^3 - 11x_i), \\ f_5 &= \sum_{i=1}^{5}(-364x_i^3 + 2912x_i^5 + 13x_i - 9984x_i^7 - 13312x_i^{11} + 16640x_i^9 + 4096x_i^{13}), \end{aligned}\right\} \tag{14}$$

where $m = V_1/(4V_{dc}/\pi)$. Following the approach described in [24], the system of Equations (14) is symmetric polynomials. For five variables, define the elementary symmetric polynomials as follows:

$$s_1 = \sum_{1\le i\le 5} x_i, \quad s_2 = \sum_{1\le i<j\le 5} x_i x_j, \quad s_3 = \sum_{1\le i<j<k\le 5} x_i x_j x_k, \quad \cdots, \quad s_5 = \prod_{i=1}^{5} x_i. \tag{15}$$

Therefore, we can rewrite the system of Equation (14) as new expressions with $s_1, s_2, \cdots, s_5$, the information of which is in Table 1.

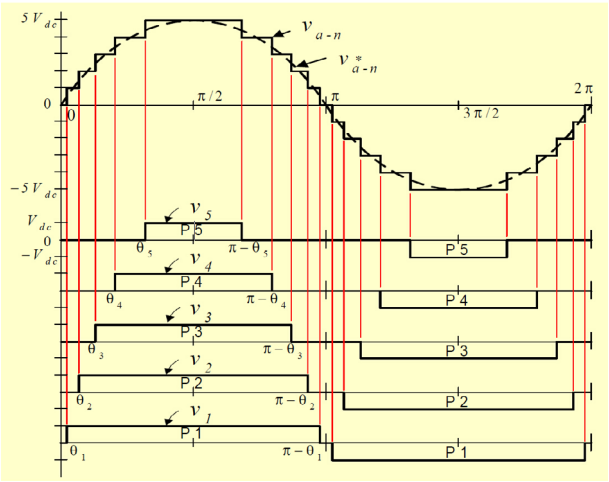


Figure 1 Output waveform of an 11-level cascade multilevel inverter

Table 1 The basic information of the system of equations

equation	variable	term	degree				
			deg s_1	deg s_2	deg s_3	deg s_4	deg s_5
f_1	s_1, m	2	1	0	0	0	0
f_2	$s_1, s_2, \cdots, s_5$	11	5	2	1	1	1
f_3	$s_1, s_2, \cdots, s_5$	24	7	3	2	1	1
f_4	$s_1, s_2, \cdots, s_5$	84	11	5	3	2	2
f_5	$s_1, s_2, \cdots, s_5$	141	13	6	4	3	2

Since the system of Equation (14) has the specific structure $f_1 = s_1 - m$, we can replace s_1 with m. From Table 2, we can get much less than the highest degrees of original system of Equation (14). Here we apply a novel Dixon resultant elimination method to compute the new system of polynomials with $s_1, s_2, \cdots, s_5$. First, we can get five Dixon resultant matrices from [25] in Table 2. Following our Algorithms 2, 3 and 4, we obtain the results of Dixon resultant matrices in Table 2.

Table 2 The results of matrices with polynomial entries

eliminated variable	univariate polynomial	Dixon resultant matrix	results	
			degree	term
s_3, s_4, s_5	s_2	18×18	17	512
s_2, s_4, s_5	s_3	25×25	11	398
s_2, s_3, s_5	s_4	35×35	9	507
s_2, s_3, s_4	s_5	42×42	9	464

Remark 4.2 Our result is consistent with the existing method, which of result is using the Sylvester resultant[24]. However, we only need to compute the elimination procedure once.

The existing approach needs 6 times to obtain the same result. In general, we can know that the time of elimination procedure is C_n^2 from the system of multivariate polynomial equations to a univariate polynomial based on Sylvester resultant, where n is the number of equations.

In Figure 2, we compare the running time of our method with that of two other algorithms. In Figure 3, we compare the memory consumption of our method with that of two other algorithms, where the order of x-coordinate represents for the order of symbolic determinants.

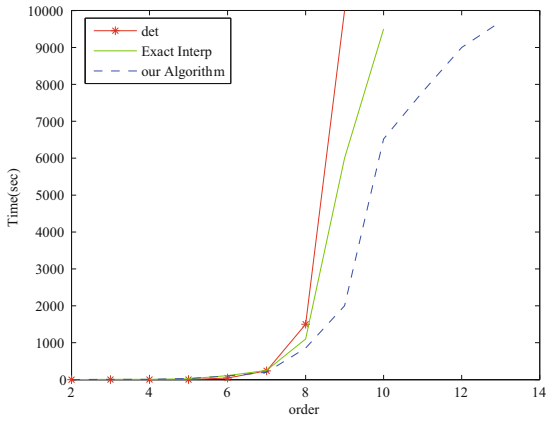


Figure 2 Computing time for symbolic determinant with different algorithms

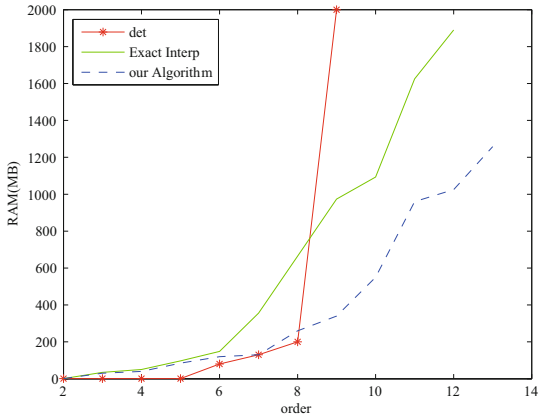


Figure 3 Computing memory for symbolic determinant with different algorithms

From Figures 2 and 3, we have the observations as follows:

1) In general, the Time and RAM of algorithm det are reasonable when the order is less than nine, and two indicators increase very rapidly when the order is to nine. However, two indicators of interpolation algorithm is steady growth.

2) Compared with the exact interpolation method, the approximate interpolation algorithm has the obvious advantages on the Time and RAM when the order is more than eight.

Remark 4.3 All examples are randomly generated using the command of Maple. The symbolic method has the advantage of the low order or sparse symbolic determinants, such as expansion by minors, Gaussian elimination over the integers. However, a purely symbolic algorithm is powerless for many scientific computing problems, such as resultants computing, Jacobian determinants and some practical engineering always involving high-order symbolic determinants. Therefore, it is necessary to introduce numerical methods to improve intermediate expression swell problem arising from symbolic computations.

5 Conclusions

In this paper, we propose a hybrid symbolic-numerical method to compute the symbolic determinants. Meanwhile, we also present a novel approach for estimating the bounds on degree of variables by the extended numerical determinant technique, and introduce the reducing dimension algorithm. Combined with these methods, our algorithm is more efficient than exact interpolation algorithm for computing the high order symbolic determinants. It can be applied in scientific computing and engineering fields, such as computing Jacobian determinants in particular. Thus we can take fully advantage of approximate methods to solve large scale symbolic computation problems.

References

[1] Cox D A, Little J, and O'Shea D, *Using Algebraic Geometry*, 2nd Edition, Springer-Verlag, Berlin Heidelberg, 2005.

[2] Delvaux S, Marco A, Martínez J J, et al., Fast computation of determinants of Bézout matrices and application to curve implicitization, *Linear. Algebra. Appl.*, 2009, **430**(1): 27–33.

[3] Qin X L, Wu W Y, Feng Y, et al., Structural analysis of high-index DAE for process simulation, *Int. J. Model. Simul. Sci. Comput.*, 2013, **4**(4): 1342008.

[4] Horowitz E and Sahni S, On computing the exact determinant of matrices with polynomial entries, *J. ACM*, 1975, **22**(1): 38–50.

[5] Marco A and Martínez J J, Parallel computation of determinants of matrices with polynomial entries, *J. Symb. Comput.*, 2004, **37**(6): 749–760.

[6] Li Y, An effective hybrid algorithm for computing symbolic determinants, *Appl. Math. Comput.*, 2009, **215**(7): 2495–2501.

[7] Chen L Y and Zeng Z B, Parallel computation of determinants of matrices with multivariate polynomial entries, *Sci. China Inform. Sci.*, 2013, **56**(11): 1–16.

[8] Gentleman W M and Johnson S C, Analysis of algorithms, a case study: Determinants of matrices with polynomial entries, *ACM T. Math. Software*, 1976, **2**: 232–241.

[9] Sasaki T and Murao H, Efficient Gaussian elimination method for symbolic determinants and linear systems, *ACM T. Math. Software*, 1982, **8**(3): 277–289.

[10] Kaltofen E, On computing determinants of matrices without divisions, *Proc. ISSAC* 1992, ACM Press, New York, 1992, 342–349.

[11] Lipson J D, Symbolic methods for the computer solution of linear equations with applications to flowgraphs, *Proc.* 1968 *Summer Institute on Symbolic Mathematical Computation*, 1969, 233–303.

[12] Chen L, Eberly W, Kaltofen E, et al., Efficient matrix preconditioners for black box linear algebra, *Linear Algebra & Its Applications*, 2002, (343–344): 119–146.

[13] Kaltofen E and Yang Z F, On exact and approximate interpolation of sparse rational functions, *Proc. ISSAC* 2007, ACM Press, New York, 2007, 203–210.

[14] Chèze G and Galligo A, From an approximate to an exact absolute polynomial factorization, *J. Symb. Comput.*, 2006, **41**: 682–696.

[15] Zhang J Z and Feng Y, Obtaining exact value by approximate computations, *Sci. China Math.*, 2007, **50**(9): 1361–1368.

[16] Feng Y, Qin X L, Zhang J Z, et al., Obtaining exact interpolation multivariate polynomial by approximation, *Journal of Systems Science and Complexity*, 2011, **24**(4): 803–815.

[17] Kaltofen E, Li B, Yang Z F, et al., Exact certification in global polynomial optimization via sums-of-squares of rational functions with rational coefficients, *J. Symb. Comput.*, 2012, **47**(1): 1–15.

[18] Qin X L, Feng Y, Chen J W, et al., A complete algorithm to find exact minimal polynomial by approximations, *Int. J. Comput. Math.*, 2012, **89**(17): 2333–2344.

[19] Howard E, *Elementary Matrix Theory*, Dover Publications, New York, 1966.

[20] Boor C d, Polynomial interpolation in several variables, *In Studies in Computer Science* (in Honor of Samuel D. Conte), Eds. by DeMillo R and Rice J R, Plenum Press, New York, 1994, 87–119.

[21] Horn R A and Johnson C R, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, 1991.

[22] Björck A and Pereyra V, Solution of vandermonde systems of equations, *Math. Comput.*, 1970, **24**(112): 893–903.

[23] Moenck R T, Practical fast polynomial multiplication, *Proc. ACM Symposium on Symbolic and Algebraic Computation*, 1976, 136–148.

[24] Chiasson J N, Tolbert L M, McKenzie K J, et al., Elimination of harmonics in a multilevel converter using the theory of symmetric polynomials and resultants, *IEEE Trans. Control Systems Technology*, 2005, **13**(2): 216–223.

[25] Kapur D, Saxena T, and Yang L, Algebraic and geometric reasoning using Dixon resultants, *Proc. ISSAC* 1994, ACM Press, New York, 1994, 99–107.

AAECC (2018) 29:245–259
https://doi.org/10.1007/s00200-017-0340-0

ORIGINAL PAPER

Factoring RSA moduli with primes sharing bits in the middle

Omar Akchiche[1] · **Omar Khadir**[1]

Received: 19 October 2016 / Revised: 11 May 2017 / Accepted: 3 August 2017 /
Published online: 21 August 2017

Abstract We address the problem of factoring a large RSA modulus $N = pq$ with p and q sharing a portion of bits in the middle. New polynomial time algorithms for computing the prime decomposition of N under certain conditions are presented. As an application, several attacks against RSA system using this class of moduli with low public exponent are described. Our results suggest that such integers are not appropriate for cryptographic purposes.

Keywords Integer factorization problem · RSA system · Coppersmith's method · Public key cryptography

Mathematics Subject Classification 11Y05 · 94A60

1 Introduction

Factoring integers is a major issue in number theory and cryptography. The security of many cryptosystems such as RSA [16] is based on its presumed intractability. Fermat factoring method, see e.g. [7], is efficient when the integer is a product of two primes that are close to one another. In 1931, the continued fraction method [9] was discovered. Some decades later, J. Pollard published the $p-1$ [12] and ρ [13] algorithms. The success of the former hinges on $p-1$ having small factors for some prime divisor

Omar Khadir
khadir@hotmail.com

Omar Akchiche
omar.akchiche@hotmail.com

[1] Laboratory of Mathematics, Cryptography, Mechanics and Numerical Analysis, Fstm, University Hassan II of Casablanca, Casablanca, Morocco

p of N. The latter applies a cycle finding process. With the advent of public key cryptography, integer factorization problem receives much research interest. This has leaded to more developed techniques such as the quadratic sieve [14] and the elliptic curve factoring algorithm [11]. The most sophisticated method up to date is the general number field sieve [10, p. 103]. It was conceived by J. Pollard and allows to factor integers with more than 110 digits.

There are several results about integer factorization when given extra information about the prime factors. In 1985, Rivest and Shamir [15] proved that for an RSA modulus $N = pq$, the knowledge of $\frac{1}{3}\log N$ of the bits of p or q enables us to factor N. Later, Coppermsith ameliorates the result to $\frac{1}{4}\log N$ in 1997 [6]. To achieve this purpose, he applied a lattice based method to find small roots of bivariate integer polynomials. Boneh, Durfee and Frankel [5] showed, that for a short public exponent RSA, only a quarter of the bits of the secret key d suffices to reconstruct the whole value of d, and thus to factor the public modulus. They obtained similar results for larger values of e as well, under some conditions.

It is well known that one can factor $N = pq$ in an almost instantaneous time by using Fermat's method, see e.g. [7], provided that the primes p and q share a sufficient amount of most significant bits, namely $\frac{1}{4}\log N$. Furthermore, in [7], the author proved that a modulus with a small difference of its factors is unsafe in an RSA system using short private keys. The security of an RSA system with primes sharing low-order bits was investigated in [17] and [18]. In [18], the authors proposed an efficient method to recover the prime decomposition of N when p and q have in common more that $\frac{1}{4}\log N$ least significant bits. The paper extended the partial key exposure attacks initially presented in [5]. Some improvements of the attacks depicted in [17] and [18] were given in [20]. In [22], it was shown that such RSA protocol with small private exponent is more vulnerable to the lattice based attack of Boneh–Durfee [2,3], than the original RSA scheme. The Boneh–Durfee's work [2,3], has ameliorated the Wiener's continued fraction technique [21] published in 1990. The results of [22] were further improved in [19].

Our work is devoted to the factorization of RSA moduli $N = pq$ when the primes p and q share bits in the middle. To the best of our knowledge, this class of integers has never been studied before. We present new algorithms for factoring N under certain conditions. In particular, our results improve the Coppersmith's "factoring with a hint" [6] and Fermat method, see e.g. [7], for this kind of integers. Furthermore, new attacks against RSA system using such moduli with low public exponent are described.

The paper is organized as follows. In Sect. 2, we give our main results. In Sect. 3, we study the security of an RSA system with the new class of moduli. Finally, we conclude in Sect. 4.

Throughout the sequel, $\mathbb{N}$ and $\mathbb{Z}$ are the sets of natural numbers and integers respectively, and $\mathbb{N}^* = \mathbb{N} - \{0\}$. For integers a, b and c, we write $a \equiv b \pmod{c}$ if c divides the difference $a - b$, and $a = b \mod c$ if a is the remainder in the division of b by c. We denote by $\phi(.)$ the Euler phi function. Let n be a natural number. Then, $\mathbb{Z}_n = \{0, 1, 2, \ldots, n-1\}$. Moreover, $\mathbb{Z}_n^*$ is the set of all the elements of $\mathbb{Z}_n$ that are

invertible modulo n. If x is a real number, then $\lfloor x \rfloor$ and $\lceil x \rceil$ are respectively the floor and the ceiling of x. All the logarithms should be interpreted as logarithms to the base 2.

We start by showing how to factor $N = pq$ when p and q share some bits in the middle.

2 Factoring the RSA modulus

In this section, we establish sufficient conditions for factoring a large RSA modulus $N = pq$ when the primes p and q share bits in the middle. But first, we need some necessary preliminaries. In particular, we provide tools for solving modular equations of the form $x^2 \equiv a \pmod{2^t}$ where a and t are given integers and x is the unknown variable. When a is odd, the next lemma holds:

Lemma 1 ([1, p. 192]) *Let $t \geq 3$ be a natural number.*

(a) *Element a is a square modulo 2^t if and only if $a \equiv 1 \pmod 8$.*
(b) *If $x^2 \equiv a \pmod{2^t}$, then the square roots of a modulo 2^t are $\pm x, \pm x + 2^{t-1}$.*
(c) *A square root of a modulo 2^t can be computed in $O(t^2)$ bit operations.*

Proof See [1, Ex. 7.9.38, p. 192]. □

When a is even, the equation $x^2 \equiv a \pmod{2^t}$ was studied in [18], and the authors proved the following assertions:

Lemma 2 ([18]) *The set of solutions in $\mathbb{Z}_{2^t}$ to the modular equation $x^2 \equiv a \pmod{2^t}$ is summarised as follows. Let $a = 2^\alpha b$ where b is an odd integer and $\alpha \in \mathbb{N}^*$.*

(a) *If $t \leq \alpha$, there are $2^{\lfloor \frac{t}{2} \rfloor}$ solutions $x \equiv 0 \pmod{2^{\lceil \frac{t}{2} \rceil}}$.*
(b) *If $t > \alpha$, there are no solutions if α is odd, and three subcases if α is even.*
(b.1) *If $t = \alpha + 1$, there are $2^{\frac{\alpha}{2}}$ solutions $x \equiv 2^{\frac{\alpha}{2}} \pmod{2^{\frac{\alpha}{2}+1}}$.*
(b.2) *If $t = \alpha + 2$, there are $2^{\frac{\alpha}{2}+1}$ solutions $x \equiv \pm 2^{\frac{\alpha}{2}} \pmod{2^{\frac{\alpha}{2}+2}}$ if $b \equiv 1 \pmod 4$ and none if not.*
(b.3) *If $t \geq \alpha + 3$, there are $2^{\frac{\alpha}{2}+2}$ solutions of the form $x \equiv 2^{\frac{\alpha}{2}}(\pm s + \delta . 2^{t-\alpha-1}) \pmod{2^{t-\frac{\alpha}{2}}}$ with $\delta \in \{0, 1\}$ if $b \equiv 1 \pmod 8$, and no solutions if $b \not\equiv 1 \pmod 8$. Here s is any solution to $x^2 \equiv b \pmod{2^{t-\alpha}}$.*

Proof See [18]. □

Coppersmith [6], using a lattice based method, proved that it is possible to factor an RSA modulus $N = pq$ when $\frac{1}{4} \log N$ least or most significant bits of one of its prime divisors are revealed.

Theorem 1 ([6]) *In polynomial time, we can find the factorization of $N = pq$, where p and q are primes of the same bit-size, if we know the low or high-order $\frac{1}{4} \log N$ bits of p.*

Proof See [6]. □

From now on, we denote by $T_1(N)$ the polynomial running time indicated in Theorem 1. Assume that p and q have the same bits in the middle. In our following lemma, we show that it is possible to compute the least significant bits of the prime factors of N when given access to an oracle that outputs a part of $|p-q|$.

Lemma 3 *Let $N = pq$ be an RSA modulus where p and q share bits from position t_1 to t_2, for given t_1 and t_2. We denote by α the multiplicity of 2 in $p+q$. If there exists an oracle that outputs $|p-q| \mod 2^{t_1}$, then there exist $O(2^{\min(\alpha-1,\frac{t_2}{2})})$ candidates for the t_2 least significant bits of p, and each one of them can be computed in time $O((\log N)^2)$.*

Proof As the primes p and q share bits from position t_1 to t_2, they are expressed as $p = p_2 2^{t_2} + r2^{t_1} + p_0$ and $q = q_2 2^{t_2} + r2^{t_1} + q_0$. Reducing $N = pq$ modulo 2^{t_2}, it follows that $N \equiv (r2^{t_1} + p_0)(r2^{t_1} + q_0) \pmod{2^{t_2}}$. This entails $N \equiv (r2^{t_1})^2 + (p_0 + q_0)r2^{t_1} + p_0q_0 \pmod{2^{t_2}}$. By completing the square, we get $\left(r2^{t_1} + \frac{p_0+q_0}{2}\right)^2 \equiv N + \left(\frac{p_0-q_0}{2}\right)^2 \pmod{2^{t_2}}$. Both p_0 and q_0 are odd integers, so the quantities $\frac{p_0 \pm q_0}{2}$ are well defined. Without loss of generality, assume that $q_0 < p_0$. As $p \mod 2^{t_2} = r2^{t_1} + p_0 < 2^{t_2}$, we obtain $r2^{t_1} + \frac{p_0+q_0}{2} < 2^{t_2}$. We are given $|p-q| \mod 2^{t_1}$. So, we can determine $|p_0 - q_0|$. Set $X = r2^{t_1} + \frac{p_0+q_0}{2}$. Therefore, X is a square root of $a = N + \left(\frac{p_0-q_0}{2}\right)^2$ modulo 2^{t_2}.

We first handle the case when $N \equiv 1 \pmod 4$. Clearly, $p \equiv q \pmod 4$. Using the fact that $\left(\frac{p+q}{2}\right)^2 \equiv a \pmod{2^{t_2}}$, we deduce that a is odd. By b) and c) of Lemma 1, there are four square roots of a modulo 2^{t_2}, and they can be computed in time $O((\log N)^2)$. One of them equals X modulo 2^{t_2}. Inequality $p_0, q_0 < 2^{t_1}$ leads to $p_0 + q_0 = 2(X \mod 2^{t_1})$. Moreover, we know the difference $p_0 - q_0$, so we are able to determine p_0 and q_0. On another hand, $X = r2^{t_1} + \frac{p_0+q_0}{2}$, so it is possible to recover the value of r. We compute $r2^{t_1} + p_0$ which reveals the t_2 least significant bits of p.

Suppose now that $N \equiv 3 \pmod 4$. It is not difficult to see that $a = N + \left(\frac{p_0-q_0}{2}\right)^2$ is even. Write $p+q = s2^{\alpha}$ where s is an odd integer and $\alpha \in \mathbb{N}^*$. Since $\left(\frac{p+q}{2}\right)^2 \equiv a \pmod{2^{t_2}}$, if $2(\alpha-1) \geq t_2$, then $N + \left(\frac{p_0-q_0}{2}\right)^2 \equiv 0 \pmod{2^{t_2}}$. By a) of Lemma 2, $X \equiv 0 \pmod{2^{\lceil \frac{t_2}{2} \rceil}}$. There are $\lfloor \frac{t_2}{2} \rfloor$ missing bits of X. Hence, we have $2^{\lfloor \frac{t_2}{2} \rfloor}$ candidates for X.

When $2(\alpha-1) < t_2$, we use part b) of Lemma 2 to compute the square roots of $N + \left(\frac{p_0 - q_0}{2}\right)^2$ modulo 2^{t_2}. If $t_2 = 2\alpha - 1$ or $t_2 = 2\alpha$, then there are $2^{\alpha-1}$ or 2^{α} possible values of X respectively. Otherwise, there are at most four candidates for λ, that can be computed in time $O((\log N)^2)$, such that $r2^{t_1} + \frac{p_0+q_0}{2} \equiv \lambda \pmod{2^{t_2-(\alpha-1)}}$. It remains to find $\alpha - 1$ bits. So, we have $2^{\alpha+1}$ candidates for X. The knowledge of X enables us to compute p_0 and r. Thus, the t_2 least significant bits of p are obtained from $r2^{t_1} + p_0$. □

From Lemma 3, we prove the following theorem:

Theorem 2 *Let $N = pq$ be an RSA modulus where p and q are primes of the same bit-size. Assume that p and q share bits from position t_1 to t_2 such that $t_2 \geq \frac{1}{4}\log N$, for given t_1 and t_2. We denote by α the multiplicity of 2 in $p+q$. If there exists an oracle that outputs $|p-q| \mod 2^{t_1}$, then we can factor N in time $T_2(N) = O(2^{\min(\alpha-1, \frac{t_2}{2})}(T_1(N) + (\log N)^2))$.*

Proof We set $l = p \mod 2^{t_2}$. By Lemma 3, given $|p-q| \mod 2^{t_1}$, it is possible to find $O(2^{\min(\alpha-1, \frac{t_2}{2})})$ candidates for l. For each one, we apply Theorem 1 of Coppersmith. By hypothesis, $t_2 \geq \frac{1}{4}\log N$. So, the true guess of l yields the prime decomposition of N in time $T_1(N)$. □

Example 1 We illustrate Theorem 2 through two examples. The numerical computations were done using Maple software.

Consider the following 128-bit RSA modulus:

$$N = 202914989268620230739444582780476305109.$$

By an oracle, we know that the prime factors p and q share the bits from position $t_1 = 11$ to $t_2 = 37$, and $|p-q| \mod 2^{t_1} = 908$.

We have $N \mod 4 = 1$, so we are in the first case of Lemma 3. The absolute value of $p_0 - q_0$ is either $|p-q| \mod 2^{t_1}$ or $-(|p-q| \mod 2^{t_1}) \mod 2^{t_1}$. Suppose that $|p_0 - q_0| = |p-q| \mod 2^{t_1}$ and set $a = N + \left(\frac{|p-q| \mod 2^{t_1}}{2}\right)^2 \mod 2^{t_2} = 13773755385$. Next, we solve $x^2 \equiv a \pmod{2^{t_2}}$ where x is the unknown. We get four solutions:

$$\{48786752181, 19932724555, 88652201291, 117506228917\}.$$

Among the roots of the previous modular equation, there is one that equals $r2^{t_1} + \frac{p_0+q_0}{2}$. Assuming that $r2^{t_1} + \frac{p_0+q_0}{2} = x$ with $x = 19932724555$, it follows that $p_0 + q_0 = 2(x \mod 2^{t_1}) = 2710$. Since $|p_0 - q_0| = |p-q| \mod 2^{t_1}$, we obtain $p_0 = 1809$ and $q_0 = 901$. Moreover, $r = 9732775$ as $r2^{t_1} + \frac{p_0+q_0}{2} = x$.

Finally, we get the t_2 least significant bits of p from $r2^{t_1} + p_0 = 19932725009$. This information is sufficient to recover the factorization of N using Coppersmith's method since $t_2 \geq \frac{\log N}{4}$. So, $p = 15764252793534496529$ and $q = 12871843145768582021$.

Now, consider the following 128-bit RSA modulus:

$$N = 281158159401182057324315156925939510931.$$

Furthermore, we are given the following hints. The primes p and q have in common the bloc of bits from $t_1 = 11$ to $t_2 = 37$, and $|p-q| \mod 2^{t_1} = 1650$. We are in the second situation of Lemma 3 as $N \mod 4 = 3$. The absolute value of $p_0 - q_0$ is either $|p-q| \mod 2^{t_1}$ or $-(|p-q| \mod 2^{t_1}) \mod 2^{t_1}$. Assume that $|p_0-q_0| = |p-q| \mod 2^{t_1}$. It follows that the solutions of $x^2 \equiv a \pmod{2^{t_2}}$, where $a = N + \left(\frac{|p-q| \mod 2^{t_1}}{2}\right)^2 \mod 2^{t_2} = 59950977348$, are :

$$\{100485641234, 71313050606, 2593573870, 134845379602, 31766164498, 105672788974, 66125902866, 36953312238\}$$

One of these $2^{\frac{\alpha}{2}+2}$ elements, with $\alpha = 2$, equals $r2^{t_1} + \frac{p_0+q_0}{2}$. Suppose that $r2^{t_1} + \frac{p_0+q_0}{2} = x$ with $x = 105672788974$. Then $p_0 + q_0 = 2(x \mod 2^{t_1}) = 2012$. Using the fact that $|p_0 - q_0| = 1650$, we deduce that $p_0 = 1831$ and $q_0 = 181$. From $r2^{t_1} + \frac{p_0+q_0}{2} = x$, we find that $r = 51598041$. So, the t_2 least significant bits of p are obtained from $r2^{t_1} + p_0 = 105672789799$. By Coppersmith's method, as $t_2 \geq \frac{\log N}{2}$, the factors of N are $p = 16884512291466694439$ and $q = 16651837763965339829$.

When the hypothesis $t_2 \geq \frac{\log N}{4}$ is not satisfied, we obtain a weaker result.

Proposition 1 *Let $N = pq$ be an RSA modulus where p and q are primes of the same bit-size. Assume that p and q share bits from position t_1 to t_2 such that $t_2 < \frac{\log N}{4}$, for given t_1 and t_2. We denote by α the multiplicity of 2 in $p+q$. If there exists an oracle that outputs $|p-q| \mod 2^{t_1}$, then we can factor N in time $T_3(N) = O(2^{\frac{\log N}{4} - t_2 + \min(\alpha - 1, \frac{t_2}{2})}(T_1(N) + (\log N)^2))$.*

Proof We put $l = p \mod 2^{\lceil \frac{\log N}{4} \rceil}$. Using Lemma 3, there are $O(2^{\frac{\log N}{4} - t_2 + \min(\alpha-1, \frac{t_2}{2})})$ possible values for l. By Theorem 1 of Coppersmith, once we have the true value of l, we can factor N which ends the proof. □

From our Theorem 2 and Proposition 1 , we derive an improvement of the "factoring with a hint" [6] result when applied to RSA moduli $N = pq$ such that p and q share a block of bits in the middle. In this case, the number of low-order bits required to be known is reduced. More precisely:

Corollary 1 *Let $N = pq$ be an RSA modulus where p and q are primes of the same bit-size. Assume that p and q share bits from position t_1 to t_2, for given t_1 and t_2. If we know the t_1 least significant bits of p, then N can be factored in time $T_2(N)$ if $t_2 \geq \frac{1}{4}\log N$ and $T_3(N)$ otherwise.*

Proof Let N be an RSA modulus satisfying the hypothesis of the above statement. As we have access to the t_1 least significant bits of p, we easily compute the corresponding low-order bits of q. Thus, it is possible to compute $|p-q| \mod 2^{t_1}$. The result follows immediately from Theorem 2 and Proposition 1. □

Let $N = pq$ be an RSA modulus where p and q are unknowns. Fermat method, see e.g. [8, p. 144] or [7], is efficient for factoring a product of two integers that are close one to another. In particular, in [7] was proved that if $|p - q| < cN^{\frac{1}{4}}$ for a certain positive constant c, then one can find the prime decomposition of N in at most $\frac{1}{4}c^2$ iterations. In the next statement, we show that the Fermat technique can be improved, when p and q have in common bits in the middle.

Proposition 2 *Let $N = pq$ be an RSA modulus such that p and q share bits from position t_1 to t_2, for given t_1 and t_2. We denote by α the multiplicity of 2 in $p + q$. Suppose that $|p - q| < 2^{\frac{c+\log N}{4}}$ where c is a positive constant such that $c \leq 2t_2$. If there exists an oracle that outputs $|p - q| \mod 2^{t_1}$, then we can factor N in time $T_4(N) = O(2^{\min(\alpha-1,\frac{t_2}{2})}(\log N)^2)$.*

Proof Let $N = pq$ be the RSA modulus to be factored. In the Fermat method, see e.g. [7], we look for two integers a and b such that $a^2 - b^2 = 4N$. We try $a = \lceil 2\sqrt{N}\rceil$, $\lceil 2\sqrt{N}\rceil + 1$, $\lceil 2\sqrt{N}\rceil + 2, \ldots$, until $a^2 - 4N$ is a perfect square. Having computed a and b, the factors of N are $p = \frac{1}{2}(a + b)$ and $q = \frac{1}{2}(a - b)$. The proof of our proposition is based on a slight modification of this technique as depicted in [8, p. 144]. We have $\left(\frac{p+q}{2}\right)^2 - \left(\frac{p-q}{2}\right)^2 = N$. So, we search $x = \frac{p+q}{2}$ and $y = \frac{p-q}{2}$ respectively instead of $a = p + q$ and $b = p - q$. In such case, we must try $x = \lfloor\sqrt{N}\rfloor + 1$, $\lfloor\sqrt{N}\rfloor + 2, \ldots$, until $x^2 - N$ is a square. In other words, there exists a positive integer j that verifies $(\lfloor\sqrt{N}\rfloor + 1 + j)^2 - \left(\frac{p-q}{2}\right)^2 = N$. We started with $\lfloor\sqrt{N}\rfloor + 1$ since $\sqrt{N} < \frac{p+q}{2}$, [7]. As p and q share bits from t_1 to t_2, $p = p_2 2^{t_2} + r2^{t_1} + p_0$ and $q = q_2 2^{t_2} + r2^{t_1} + q_0$. Thus $(\lfloor\sqrt{N}\rfloor + 1 + j)^2 \equiv N + \left(\frac{p_0 - q_0}{2}\right)^2 \pmod{2^{t_2}}$ for some $j \in \mathbb{N}$. Set $X = \lfloor\sqrt{N}\rfloor + 1 + j$. Since $|p - q| \mod 2^{t_1}$ is given, we are left with the task of solving a quadratic modular equation. We proceed as in the proof of Lemma 3.

If $N \equiv 1 \pmod 4$, then $N + \left(\frac{p_0 - q_0}{2}\right)^2$ is odd. So, using Lemma 1, we efficiently compute the four square roots of $N + \left(\frac{p_0 - q_0}{2}\right)^2$ modulo 2^{t_2}. Hence, we find $(j \mod 2^{t_2})$ in time $O((\log N)^2)$.

If $N \equiv 3 \pmod 4$, then the quantity $N + \left(\frac{p_0 - q_0}{2}\right)^2$ is even. We have $\left(\frac{p+q}{2}\right)^2 \equiv N + \left(\frac{p_0 - q_0}{2}\right)^2 \pmod{2^{t_2}}$. Writing $p + q = s2^{\alpha}$ for some odd number s and $\alpha \in \mathbb{N}^*$, we see that $N + \left(\frac{p_0 - q_0}{2}\right)^2 \equiv 0 \pmod{2^{t_2}}$ whenever $2(\alpha - 1) \geq t_2$. By part a) of Lemma 2, $X \equiv 0 \pmod{2^{\lceil \frac{t_2}{2} \rceil}}$. We find the $\lfloor \frac{t_2}{2} \rfloor$ missing bits in at most $2^{\lfloor \frac{t_2}{2} \rfloor}$ trials. If $2(\alpha - 1) < t_2$, then we make use of part b) of Lemma 2 to solve the quadratic equation. If $t_2 = 2\alpha - 1$ or $t_2 = 2\alpha$, we have $2^{\alpha - 1}$ or 2^{α} possible choices for $X \mod 2^{t_2}$ respectively. Otherwise, there are at most four candidates for λ, that can be computed in time $O((\log N)^2)$, such that $X \equiv \lambda \pmod{2^{t_2 - (\alpha - 1)}}$. The remaining $\alpha - 1$ bits are exhaustively searched in $2^{\alpha - 1}$ steps. So, we recover $(j \mod 2^{t_2})$ in time $O(2^{\alpha + 1}(\log N)^2)$.

The number of iterations is j such that $j + \lfloor \sqrt{N} \rfloor + 1 = \frac{p+q}{2}$, and so $j = \frac{p+q}{2} - \lfloor \sqrt{N} \rfloor - 1$. It was established in the lemma of Sect. 2 of [7] that $p + q - 2\sqrt{N} < \frac{(p-q)^2}{4\sqrt{N}}$. We deduce that $j < \frac{(p-q)^2}{8\sqrt{N}}$. In particular, this means that if $\frac{(p-q)^2}{8\sqrt{N}} \leq 2^{t_2}$, then it is possible to recover the entire value of j. By hypothesis, $|p - q| < 2^{\frac{c + \log N}{4}}$ where c is a positive constant. Thus, it suffices that $c \leq 2t_2 + 6$, so the condition $\frac{(p-q)^2}{8\sqrt{N}} \leq 2^{t_2}$ holds. The knowledge of j reveals the factorization of N which ends the proof. □

The next corollary ensues from Proposition 2:

Corollary 2 *Let $N = pq$ be an RSA modulus such that p and q share bits from position t_1 to t_2, for given t_1 and t_2. Suppose that $|p - q| < 2^{\frac{c + \log N}{4}}$ where c is a positive constant such that $c \leq 2t_2$. If there exists an oracle that outputs the t_1 least significant bits of p, then we can factor N in time $T_4(N)$.*

Proof By division, given the t_1 least significant bits of p, one can determine the t_1 low-order bits of q. We then have $|p - q| \mod 2^{t_1}$. By Proposition 2, we compute the factors of N in time $T_4(N)$. □

Let $N = pq$ be an RSA modulus where p and q have the same block of bits from t_1 to t_2. It is worth noting that, given the t_1 least significant bits of p, computing the

prime decomposition of N by using Theorem 1 of Coppersmith requires an exhaustive search of all the bits from t_1 to $\frac{\log N}{2}$. We denote by α the multiplicity of 2 in $p+q$.

Our Theorem 2 and Corollary 2 yield polynomial time algorithms for factoring N provided that $N \equiv 1 \pmod 4$ and $t_2 \geq \frac{\log N}{4}$. When $N \equiv 3 \pmod 4$ and $t_2 > \max(2(\alpha-1), \frac{\log N}{4})$, we must have $\alpha = O(\log\log N)$ in order to get similar efficient results.

Moreover, Proposition 2 and Corollary 2 improve the Fermat technique for this class of integers N such that $N \equiv 1 \pmod 4$. If $N \equiv 3 \pmod 4$ and $t_2 > 2(\alpha-1)$, the methods derived from both statements enable to factor N once that $\alpha = O(\log\log N)$.

On another hand, for the sake of simplicity, we have assumed so far that t_1 and t_2 are known parameters. Actually, this condition can be omitted. Since $t_1, t_2 < \log N$, one can apply our statements with all possible candidates for t_1 and t_2 in a polynomial running time.

In the next section, we investigate the security of an RSA system with the new class of moduli N.

3 Cryptanalysis of the RSA system

The aim of this section is to analyse the security of an RSA system with moduli $N = pq$ such that p and q share a block of bits in the middle. In [5], Boneh, Durfee and Frankel presented the partial key exposure attacks. They showed that only a fraction of bits in the secret exponent d suffices to reconstruct all of d. Their paper was later revised in [4]. In 2004, in [18], the authors extended the results presented in [5]. Our following theorem examines the aforementioned attacks when applied to a particular class of RSA moduli.

Theorem 3 *Let $N = pq$ be an RSA modulus where p and q are primes of the same bit-size. Assume that p and q share bits from position t_1 to t_2 such that $t_2 \geq \frac{1}{4}\log N$, for given t_1 and t_2. We denote by α and β the multiplicities of 2 in $p+q$ and $p-q$ respectively. Let $(e,d) \in \mathbb{Z} \times \mathbb{Z}^*_{\phi(N)}$ be the public-key/secret-key pair satisfying $ed \equiv 1 \pmod{\phi(N)}$. Suppose that $e \leq 2^{t_1-3}$. If there exists an oracle that outputs the t_1 least significant bits of d, then we can factor N in time $T(N) = O(2^{\min(\alpha-1, \frac{t_2}{2})+\beta}\, e \log e\, (T_1(N) + (\log N)^2))$.*

Proof We first look for the value of $|p-q| \mod 2^{t_1}$, then we apply Theorem 2 to get the factorization of the public modulus. Our proof is inspired by the method used in [5] and [18]. The RSA key equation is $ed \equiv 1 \pmod{\phi(N)}$. Hence, there exists an integer k such that $ed = 1 + k\phi(N)$ and so $ed - 1 = k(N-(p+q)+1)$. By hypothesis, the t_1 least significant bits of d are provided. That is, $d_0 = d \mod 2^{t_1}$ is known. Working modulo 2^{t_1}, it results that $\frac{ed_0-1}{2} \equiv k(\frac{N+1}{2} - \frac{p+q}{2}) \pmod{2^{t_1-1}}$. Since $k = \frac{ed-1}{\phi(N)}$ and $d < \phi(N)$, $k < e$. For each candidate for k, we apply the

following process. Write $k = w2^{\mu_k}$ where w is an odd integer and $\mu_k \in \mathbb{N}$. As $ed \equiv 1 \pmod{\phi(N)}$, $ed \equiv 1 \pmod{2^{\mu_k+1}}$ and so $\frac{p+q}{2} \equiv \frac{N+1}{2} - w^{-1}\frac{ed_0-1}{2^{\mu_k+1}} \pmod{2^{t_1-\mu_k-1}}$. Substituting $\left(\frac{p+q}{2}\right)^2 = \left(\frac{p-q}{2}\right)^2 + N$ into the last formula, it ensues that $\left(\frac{p-q}{2}\right)^2 \equiv \left(w^{-1}\frac{ed_0-1}{2^{\mu_k+1}} - \frac{N+1}{2}\right)^2 - N \pmod{2^{t_1-\mu_k-1}}$. For the sake of simplicity, put $a(k) = \left[\left(w^{-1}\frac{ed_0-1}{2^{\mu_k+1}} - \frac{N+1}{2}\right)^2 - N\right] \mod 2^{t_1-\mu_k-1}$. Obviously, $\frac{|p-q|}{2}$ is a square root of $a(k)$ modulo $2^{t_1-\mu_k-1}$. Without loss of generality, suppose that $q < p$. Let $p - q = u2^\beta$ for some odd integer u and $\beta \in \mathbb{N}^*$. Assume that $N \equiv 1 \pmod 4$. Clearly, $\beta > 1$. If $2(\beta-1) \geq t_1-\mu_k-1$, then $a(k) \equiv 0 \pmod{2^{t_1-\mu_k-1}}$ and any candidate k that does not satisfy the last criterion must be rejected. Using part a) of Lemma 2, $\frac{p-q}{2} \equiv 0 \pmod{2^{\lceil \frac{t_1-\mu_k-1}{2} \rceil}}$. So, we know the $\lceil \frac{t_1-\mu_k+1}{2} \rceil$ least significant bits of $p-q$. The remaining $\lfloor \frac{t_1+\mu_k-1}{2} \rfloor$ bits will be found exhaustively and this requires at most $2^{\lfloor \frac{t_1+\mu_k-1}{2} \rfloor}$ trials.

If $2(\beta-1) < t_1-\mu_k-1$, then the multiplicity of 2 in $a(k)$ is $2(\beta-1)$. So, we must eliminate any k that does not fulfill this condition. If $t_1-\mu_k-1 = 2\beta-1$ or $t_1-\mu_k-1 = 2\beta$, then from parts b.1) and b.2) of Lemma 2, we have $2^{\mu_k+\beta-1}$ or $2^{\mu_k+\beta}$ possible values for the t_1 bits of $p-q$ respectively. Otherwise, by part b) of Lemma 2, we compute, in time $O((\log N)^2)$, at most four candidates for λ such that $\frac{p-q}{2} \equiv \lambda \pmod{2^{t_1-\mu_k-1-(\beta-1)}}$. So, the $t_1-\mu_k-\beta+1$ least significant bits of $p-q$ are revealed. We have to try at most $2^{\mu_k+\beta-1}$ possible choices in order to find the $\mu_k+\beta-1$ missing bits. For all candidates for k, the number of tries needed is upper bounded by $2^{\mu_k+\beta+1}$. In [18], the authors observed that $\sum_{i=1}^{e-1} 2^{\mu_i} \leq \lceil \log e \rceil 2^{\lceil \log e \rceil}$. Then, we apply Theorem 2 to recover the factors of N. Therefore, the whole complexity is $O(2^\beta\, e \log e\, (T_1(N) + (\log N)^2))$.

Let us now turn to the case when $N \equiv 3 \pmod 4$. In this case, $\beta = 1$. Since by hypothesis $e < 2^{t_1-3}$, it is straightforward that $3 \leq t_1-\mu_k-1$. We use parts b) and c) of Lemma 1 to find the solutions to the equation $x^2 \equiv a(k) \pmod{2^{t_1-\mu_k-1}}$. There are at most four candidates for λ, that can be computed in time $O((\log N)^2)$, such that $\frac{p-q}{2} \equiv \lambda \pmod{2^{t_1-\mu_k-1}}$. The remaining μ_k bits of $p-q$ are exhaustively searched in at most 2^{μ_k} trials. By Theorem 2, we are able to factor N. Hence, the complexity is $O(2^{\alpha-1}\, e \log e\, (T_1(N) + (\log N)^2))$ if $t_2 > 2(\alpha-1)$ and $O(2^{\frac{t_2}{2}}\, e \log e\, (T_1(N) + (\log N)^2))$ if not, which ends the proof. □

Let the public exponent e be upper bounded by a polynomial in $\log N$. When $N \equiv 1 \pmod 4$, the attack depicted in the previous theorem is feasible if $\beta = O(\log \log N)$. If

$N \equiv 3 \pmod 4$ and $t_2 > 2(\alpha - 1)$, we get an efficient algorithm if $\alpha = O(\log \log N)$. In our next result, the CRT-RSA system is treated.

Theorem 4 *Let $N = pq$ be an RSA modulus where p and q are primes of the same bit-size. Assume that p and q share bits from position t_1 to t_2 such that $t_2 \geq \frac{1}{4}\log N$, for given t_1 and t_2. We denote by α the multiplicity of 2 in $p+q$. Let e be the public-key and d_p satisfy $ed_p \equiv 1 \pmod{p-1}$. Suppose that $e \leq 2^{t_1-3}$. If there exists an oracle that outputs the t_1 least significant bits of d, then we can factor N in time $T(N) = O((2^{\min(\alpha-1, \frac{t_2}{2})}\, e \log e\, (T_1(N) + (\log N)^2))$.*

Proof From the RSA key equation, we have $ed_p \equiv 1 \pmod{p-1}$. Hence, there exists an integer k such that $ed_p = 1 + k(p-1)$. Taking the last formula modulo 2^{t_1}, we get $ed_p \equiv 1 + k(p-1) \pmod{2^{t_1}}$. For each candidate for k, we perform the following steps. By hypothesis, $d_0 = d_p \mod 2^{t_1}$ is given. Writing $k = w2^{\mu_k}$ where w is odd and $\mu_k \in \mathbb{N}$, we obtain $w^{-1}\dfrac{ed_0 - 1}{2^{\mu_k}} + 1 \equiv p \pmod{2^{t_1-\mu_k}}$. Since $k < e\dfrac{d_p}{p-1}$ and $e \leq 2^{t_1-3}$, $\mu_k < t_1$. We have so far computed the $t_1 - \mu_k$ least significant bits of p. The μ_k missing bits are recovered in at most 2^{μ_k} tries. Once the t_1 low-order bits of p are known, we factor N by applying Corollary 1. The complexity is $O(T_2(N)\sum_{i=1}^{e-1} 2^{\mu_i})$ and using the fact that $\sum_{i=1}^{e-1} 2^{\mu_i} \leq \lceil \log e \rceil 2^{\lceil \log e \rceil}$, the result follows. □

It is noteworthy that unlike Theorem 3, when $N \equiv 1 \pmod 4$, the method described in Theorem 4 is always efficient for low public exponent.

Let $N = pq$ where p and q have in common some bits from position t_1 to t_2 ($t_1 < t_2$). In the next result, we show that if there exists a partial key exposure attack using the t_2 least significant bits of the secret exponent d, then it is possible to build a factoring algorithm with only the t_1 low-order bits of p.

Proposition 3 *Let $N = pq$ be an RSA modulus where p and q are primes of the same bit-size. Assume that p and q share bits from position t_1 to t_2, for given t_1 and t_2. We denote by α the multiplicity of 2 in $p+q$. Let $(e,d) \in \mathbb{Z} \times \mathbb{Z}^*_{\phi(N)}$ be the public-key/secret-key pair satisfying $ed \equiv 1 \pmod{\phi(N)}$. Suppose that there exists an algorithm $\mathcal{A}$ that, given (N, e, d_0) where d_0 consists of the t_2 least significant bits of d, factors N in time $T_{\mathcal{A}}(N)$. Then there exists an algorithm $\mathcal{B}$ that, given (N, e, p_0) where p_0 consists of the t_1 least significant bits of p, factors N in time $T_{\mathcal{B}}(N) = O(2^{\min(\alpha-1, \frac{t_2}{2})}\, e(T_{\mathcal{A}}(N) + (\log N)^2))$.*

Proof Let $N = pq$ such that $p = p_2 2^{t_2} + r2^{t_1} + p_0$ and $q = q_2 2^{t_2} + r2^{t_1} + q_0$. As p_0 is given, we can find q_0 by division. In the proof of Lemma 3, it was shown that $O(2^{\min(\alpha-1, \frac{t_2}{2})})$ candidates for r can be found by studying the modular equation $\left(r2^{t_1} + \dfrac{p_0+q_0}{2}\right)^2 \equiv N + \left(\dfrac{p_0-q_0}{2}\right)^2 \pmod{2^{t_2}}$. Using the RSA key equation, there exists an integer k such that $ed = 1 + k\phi(N)$. Therefore $ed \equiv 1 + k(N - 2r2^{t_1} - (p_0+q_0) + 1) \pmod{2^{t_2}}$. Since $d < \phi(N)$ and $k = \dfrac{ed-1}{\phi(N)}$, $k < e$. For

each candidate for k, we compute $d_0 = e^{-1}[1 + k(N - 2r2^{t_1} - (p_0 + q_0) + 1)] \mod 2^{t_2}$. The public exponent e is odd, so the inverse of e modulo 2^{t_2} is well defined. Hence, one can apply algorithm $\mathcal{A}$ with inputs (N, e, d_0) to get the factors of N. The whole running time is $T_{\mathcal{B}}(N) = O(e(T_{\mathcal{A}}(N) + (\log N)^2))$ if $N \equiv 1 \pmod 4$. When $N \equiv 3 \pmod 4$, $T_{\mathcal{B}} = O(2^{\alpha-1}\, e(T_{\mathcal{A}}(N) + (\log N)^2))$ if $t_2 > 2(\alpha - 1)$ and $T_{\mathcal{B}} = O(2^{\frac{t_2}{2}}\, e(T_{\mathcal{A}}(N) + (\log N)^2))$ otherwise. Our statement is then proved. □

Consider an RSA modulus $N = pq$ such that p and q share bits from t_1 to t_2. The following theorem establishes that we can factor N if the cryptosystem leaks both the t_1 bits of p and the bits in the middle of the secret exponent d from position t_2 to $\frac{\log N}{4}$.

Theorem 5 *Let $N = pq$ be an RSA modulus where p and q are primes of the same bit-size. Assume that p and q share bits from position t_1 to t_2, for given t_1 and t_2. We denote by α and β the multiplicities of 2 in $p + q$ and $p - q$ respectively. Let $(e, d) \in \mathbb{Z} \times \mathbb{Z}^*_{\phi(N)}$ be the public-key/secret-key pair satisfying $ed \equiv 1 \pmod{\phi(N)}$. Suppose that $e \leq \frac{N^{\frac{1}{4}}}{8}$. If there exists an oracle that outputs the t_1 least significant bits of p and the bits of d from position t_2 to $\frac{\log N}{4}$, then we can factor N in time $T(N) = O(2^{\min(\alpha-1, \frac{t_2}{2})+\beta}\, e \log e\, (T_1(N) + (\log N)^2))$.*

Proof By hypothesis, $p = p_2 2^{t_2} + r2^{t_1} + p_0$ and $q = q_2 2^{t_2} + r2^{t_1} + q_0$. As p_0 is given, it is possible to find q_0 by division. We proceed as in the proof of Lemma 3 in order to retrieve the value of r by solving $\left(r2^{t_1} + \frac{p_0 + q_0}{2}\right)^2 \equiv N + \left(\frac{p_0 - q_0}{2}\right)^2 \pmod{2^{t_2}}$. From the RSA key equation, $ed = 1 + k(N - (p+q) + 1)$ where k is an integer. The number k ranges over the set $\{1, 2, \ldots, e-1\}$. Reducing modulo 2^{t_2}, we get $ed \equiv 1 + k(N - 2r2^{t_1} - (p_0 + q_0) + 1) \pmod{2^{t_2}}$. The public exponent e is odd. So, for each candidate for k, $d \mod 2^{t_2} = e^{-1}[1 + k(N - 2r2^{t_1} - (p_0 + q_0) + 1)] \mod 2^{t_2}$. The oracle outputs the bits of d from t_2 to $\frac{\log N}{4}$. Hence, $d_0 = d \mod 2^{\lceil \frac{\log N}{4} \rceil}$ is determined. We are left with a classical partial key exposure attack as in [5] and [18]. To estimate the complexity of the algorithm, the details are provided. We compute the $\frac{\log N}{4}$ least significant bits of p in order to get the factorization of N by applying Theorem 1 of Coppersmith. Set $k = w2^{\mu_k}$ where w is an odd integer and $\mu_k \in \mathbb{N}$. The prime p is a root of the modular congruence $\left(x - \frac{p+q}{2}\right)^2 \equiv \left(\frac{p-q}{2}\right)^2 \pmod{2^{\lceil \frac{\log N}{4} \rceil - \mu_k - 1}}$. However, $\frac{p+q}{2} \equiv \frac{N+1}{2} - w^{-1}\frac{ed_0 - 1}{2^{\mu_k+1}} \pmod{2^{\lceil \frac{\log N}{4} \rceil - \mu_k - 1}}$ and $\left(\frac{p-q}{2}\right)^2 = \left(\frac{p+q}{2}\right)^2 - N$. For convenience, set $a(k) = \left(w^{-1}\frac{ed_0 - 1}{2^{\mu_k+1}} - \frac{N+1}{2}\right)^2 - N \pmod{2^{\lceil \frac{\log N}{4} \rceil - \mu_k - 1}}$.

Then, we have $\left(x + w^{-1}\frac{ed_0 - 1}{2^{\mu_k+1}} - \frac{N+1}{2}\right)^2 \equiv a(k) \pmod{2^{\lceil \frac{\log N}{4} \rceil - \mu_k - 1}}$. Let $|p - q| = v2^{\beta}$ where v is an odd integer and $\beta \in \mathbb{N}^*$.

If $N \equiv 1 \pmod 4$, then $\beta > 1$. The multiplicity of 2 in $\left(\frac{p-q}{2}\right)^2$ is $2(\beta - 1)$. When $2(\beta - 1) \geq \lceil \frac{\log N}{4} \rceil - \mu_k - 1$, $a(k) \equiv 0 \pmod{2^{\lceil \frac{\log N}{4} \rceil - \mu_k - 1}}$. By a) of Lemma 2, $\left(x + w^{-1}\frac{ed_0 - 1}{2^{\mu_k+1}} - \frac{N+1}{2}\right) \equiv 0 \pmod{2^{\lceil \frac{\log N}{8} - \frac{\mu_k+1}{2} \rceil}}$. Thus, at most $2^{\lceil \frac{\log N}{8} + \frac{\mu_k+3}{2} \rceil}$ tries suffice to obtain the $\frac{\log N}{4}$ low-order bits of p.

If $2(\beta - 1) < \lceil \frac{\log N}{4} \rceil - \mu_k - 1$, then according to the part b) of Lemma 2, there are three sub-cases. If $\lceil \frac{\log N}{4} \rceil - \mu_k - 1 = 2\beta - 1$ or $\lceil \frac{\log N}{4} \rceil - \mu_k - 1 = 2\beta$, then we have $2^{\mu_k+\beta}$ or $2^{\mu_k+\beta+1}$ candidates for the $\lceil \frac{\log N}{4} \rceil$ bits of p respectively. Otherwise, there are at most four candidate for λ, that can be computed in time $O((\log N)^2)$, such that $\left(x + w^{-1}\frac{ed_0 - 1}{2^{\mu_k+1}} - \frac{N+1}{2}\right) \equiv \lambda \pmod{2^{\lceil \frac{\log N}{4} \rceil - 1 - \mu_k - (\beta-1)}}$. It remains to perform an exhaustive search for the $\mu_k + \beta$ missing bits. Each time that $2(\beta - 1) \geq \lceil \frac{\log N}{4} \rceil - \mu_k - 1$, the inequality $2^{\lceil \frac{\log N}{8} + \frac{\mu_k+3}{2} \rceil} \leq 2^{\mu_k+\beta+2}$ holds. It was seen that $\sum_{i=1}^{e-1} 2^{\mu_i} \leq \lceil \log e \rceil 2^{\lceil \log e \rceil}$. Hence, the whole complexity is $O(2^{\beta}\, e \log e\, (T_1(N) + (\log N)^2))$.

Suppose that $N \equiv 3 \pmod 4$. The quadratic modular equation $\left(x + w^{-1}\frac{ed_0 - 1}{2^{\mu_k+1}} - \frac{N+1}{2}\right)^2 \equiv a(k) \pmod{2^{\lceil \frac{\log N}{4} \rceil - \mu_k - 1}}$ can be solved in time $O((\log N)^2)$ by using Lemma 1. However, we still have to find the $\mu_k + 1$ remaining bits. The whole running time follows.

It should be noted that if $t_2 \geq \frac{\log N}{4}$, we come back to the scenario of Corollary 1. □

Now we prove another result for the CRT-RSA system. More precisely:

Theorem 6 *Let $N = pq$ be an RSA modulus where p and q are primes of the same bit-size. Assume that p and q share bits from position t_1 to t_2, for given t_1 and t_2. We denote by α the multiplicity of 2 in $p + q$. Let e be the public-key and let d_p satisfy $ed_p \equiv 1 \pmod{p - 1}$. Suppose that $e \leq \frac{N^{\frac{1}{4}}}{8}$. If there exists an oracle that outputs the t_1 least significant bits of p and the bits of d_p from position t_2 to $\frac{\log N}{4}$, then we can factor N in time $T(N) = O(2^{\min(\alpha-1, \frac{t_2}{2})}\, e \log e\, (T_1(N) + (\log N)^2))$.*

Proof By hypothesis, $p = p_2 2^{t_2} + r2^{t_1} + p_0$ and $q = q_2 2^{t_2} + r2^{t_1} + q_0$. The value of q_0 is easily found by division as p_0 is known. We recover the value of r by following the

proof of Lemma 3. The RSA key equation gives $ed_p = 1+k(p-1)$ where k is an integer with $k \in \{1, 2, \ldots, e-1\}$. Working modulo 2^{t_2}, we have $ed_p \equiv 1 + k(r2^{t_1} + p_0 - 1) \pmod{2^{t_2}}$. So, for each candidate for k, $d_p \mod 2^{t_2} = e^{-1}(1 + k(r2^{t_1} + p_0 - 1)) \mod 2^{t_2}$ as e is odd. Given the bits in middle of d_p from position t_2 to $\frac{\log N}{4}$, it is easy to determine $d_0 = d_p \mod 2^{\lceil \frac{\log N}{4} \rceil}$. Let $k = w2^{\mu_k}$ where w is an odd integer and $\mu_k \in \mathbb{N}$. Then $p \equiv w^{-1}\frac{ed_0 - 1}{2^{\mu_k}} + 1 \pmod{2^{\lceil \frac{\log N}{4} \rceil - \mu_k}}$. At most 2^{μ_k} tries are sufficient to find $p \mod 2^{\lceil \frac{\log N}{4} \rceil}$. By Coppersmith's result in Theorem 1, we compute the prime divisors of N in polynomial time $T_1(N)$ since $\frac{\log N}{4}$ least significant bits of p are revealed. We have seen that $\sum_{i=1}^{e-1} 2^{\mu_i} \leq \lceil \log e \rceil 2^{\lceil \log e \rceil}$. For each possible value for k, the described steps are performed. The running time follows.

Note that the case where $t_2 \geq \frac{\log N}{4}$ was previously treated in Corollary 1. □

4 Conclusion

In this paper, we studied the factorization of large RSA moduli $N = pq$ with p and q sharing bits in the middle. In particular, we presented polynomial time algorithms for computing the prime divisors of N under certain conditions. As a consequence, new partial key exposure attacks with such class of integers were depicted. Our results suggest that these numbers N should be used with care.

Acknowledgements This work is supported by the project PHC Maghreb 14MAG14.

References

1. Bach, E., Shallit, J.: Algorithmic Number Theory: Efficient Algorithms. MIT press, Cambridge (1996)
2. Boneh, D., Durfee, G.: Cryptanalysis of RSA with private key d less than $N^{0.292}$. In: Stern, J. (ed.) Advances in Cryptology, EUROCRYPT'99, pp. 1–11. Springer, Berlin (1999)
3. Boneh, D., Durfee, G.: Cryptanalysis of RSA with private key d less than $N^{0.292}$. IEEE Trans. Inf. Theory **46**(4), 1339–1349 (2000)
4. Boneh, D., Durfee, G., Frankel, Y.: Exposing an RSA private key given a small fraction of its bits. Available at Boneh's web page at: http://crypto.stanford.edu/~dabo/abstracts/bits_of_d.html. Revised version of Asiacrypt'98 paper
5. Boneh, D., Durfee, G., Frankel, Y.: An attack on RSA given a small fraction of the private key bits. In: Ohta, K., Pei, D. (eds.) Advances in Cryptology, ASIACRYPT'98, pp. 25–34. Springer, Berlin (1998)
6. Coppersmith, D.: Small solutions to polynomial equations, and low exponent RSA vulnerabilities. J. Cryptol. **10**(4), 233–260 (1997)
7. De Weger, B.: Cryptanalysis of RSA with small prime difference. Appl. Algebra Eng. Commun. Comput. **13**(1), 17–28 (2002)
8. Koblitz, N.: A Course in Number Theory and Cryptography. Springer, Berlin (1994)
9. Lehmer, D.H., Powers, R.E.: On factoring large numbers. Bull. Am. Math. Soc. **37**(10), 770–776 (1931)
10. Lenstra, A.K., Lenstra Jr., H.W.: The Development of the Number Field Sieve, vol 1554. Lecture Notes in Mathematics. Springer (1993)

11. Lenstra Jr., H.W. : Factoring integers with elliptic curves. Ann. Math. 649–673 (1987)
12. Pollard, J.M. :Theorems on factorization and primality testing. In: *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 76. Cambridge University Press, pp. 521–528 (1974)
13. Pollard, J.M.: A Monte Carlo method for factorization. BIT Numer. Math. **15**(3), 331–334 (1975)
14. Pomerance, C.: The quadratic sieve factoring algorithm. In: Beth, T., Cot, N., Ingemarsson, I., (eds.), *Advances in Cryptology, EUROCRYPT'84* . pp. 169–182 (1985)
15. Rivest, R.L., Shamir, A.: Efficient factoring based on partial information. In: Pichler, F. (ed.) Advances in Cryptology, EUROCRYPT'85, pp. 31–34. Springer, Berlin (1985)
16. Rivest, R.L., Shamir, A., Adleman, L.: A method for obtaining digital signatures and public-key cryptosystems. Commun. ACM **21**(2), 120–126 (1978)
17. Steinfeld, R., Zheng, Y.: An advantage of low-exponent RSA with modulus primes sharing least significant bits. In: Naccache, D. (ed.) Topics in Cryptology, CT-RSA 2001, pp. 52–62. Springer, Berlin (2001)
18. Steinfeld, R., Zheng, Y.: On the security of RSA with primes sharing least-significant bits. Appl. Algebra Eng. Commun. Comput. **15**(3–4), 179–200 (2004)
19. Sun, H.-M., Wu, M.-E., Steinfeld, R., Guo, J., Wang, H.: Cryptanalysis of short exponent RSA with primes sharing least significant bits. In: Franklin, M.K., Hui, L.C.K., Wong, D.S. (eds.) Cryptology and Network Security, CANS 2008, pp. 49–63. Springer, Berlin (2008)
20. Sun, H.-M., Wu, M.-E., Wang, H., Guo, J.: On the improvement of the BDF attack on LSBS-RSA. In: Mu, Y., Susilo, W., Seberry, J. (eds.) Information Security and Privacy, ACISP 2008, pp. 84–97. Springer, Berlin (2008)
21. Wiener, M.J.: Cryptanalysis of short RSA secret exponents. IEEE Trans. Inf. Theory **36**(3), 553–558 (1990)
22. Zhao, Y.-D., Qi, W.-F.: Small private-exponent attack on RSA with primes sharing bits. In: Garay, J.A., Lenstra, A.K., Mambo, M., Peralta, R. (eds.) Information Security, ISC 2007, pp. 221–229. Springer, Berlin (2007)

Neural Comput & Applic (2017) 28:585–595
DOI 10.1007/s00521-015-2080-z

ORIGINAL ARTICLE

Extension of Laplace transform–homotopy perturbation method to solve nonlinear differential equations with variable coefficients defined with Robin boundary conditions

U. Filobello-Nino[1] · H. Vazquez-Leal[1] · Yasir Khan[6] · M. Sandoval-Hernandez[2] · A. Perez-Sesma[1] · A. Sarmiento-Reyes[3] · Brahim Benhammouda[4] · V. M. Jimenez-Fernandez[1] · J. Huerta-Chua[5] · S. F. Hernandez-Machuca[1] · J. M. Mendez-Perez[1] · L. J. Morales-Mendoza[5] · M. Gonzalez-Lee[5]

Received: 22 May 2015 / Accepted: 11 October 2015 / Published online: 18 November 2015

Abstract This article proposes the application of Laplace transform–homotopy perturbation method with variable coefficients, in order to find analytical approximate solutions for nonlinear differential equations with variable coefficients. As case study, we present the oxygen diffusion problem in a spherical cell including nonlinear Michaelis–Menten uptake kinetics. It is noteworthy that this important problem introduces the Robin boundary conditions as an additional difficulty. In fact, after comparing figures between approximate and exact solutions, we will see that the proposed solutions are highly accurate. What is more, we will see that the square residual error of our solutions is $1.808511632 \times 10^{-7}$ and $2.560574954 \times 10^{-10}$ which confirms the accuracy of the proposed method, taking into account that we will just keep the first-order approximation.

Keywords Homotopy perturbation method · Nonlinear differential equation · Approximate solutions · Laplace transform · Laplace transform–homotopy perturbation method · Oxygen cell · Diffusion

List of symbols

x	Radial distance
y	Oxygen concentration
α	Maximum reaction rate
K	Michaelis constant
H	Permeability of the cell membrane

1 Introduction

Laplace transform (L.T.) has played an important role in mathematics, because its application allows solving, in a simple fashion, many problems in science and engineering [1]. As it is well known that the L.T. is useful for solving linear ordinary differential equations (ODEs) with constant coefficients and initial conditions, also it is useful in some cases of differential equations with variable coefficients and partial differential equations [1]. The applications of L.T. for nonlinear ODEs mainly focus on finding approximate solutions; thus, in [2] was reported a combination of homotopy perturbation method (HPM) and L.T. methods (LT–HPM), in order to solve approximately nonlinear problems with initial conditions [2, 3]. On the other hand, LT–HPM was adopted in order to apply it to the case of nonlinear problems with boundary conditions defined on finite intervals [4–6]. This work proposes the Laplace transform–homotopy perturbation method with variable coefficients (VCLT–HPM), as an extension of

✉ H. Vazquez-Leal
hvazquez@uv.mx

1 Electronic Instrumentation and Atmospheric Sciences School, Universidad Veracruzana, Circuito Gonzalo Aguirre Beltrán S/N, 91000 Xalapa, Veracruz, Mexico

2 Universidad de Xalapa, Km 2 Carretera Xalapa-Veracruz, 91190 Xalapa, Veracruz, Mexico

3 National Institute for Astrophysics, Optics and Electronics, Luis Enrique Erro #1, Sta. María Tonantzintla, 72840 Puebla, Mexico

4 Higher Colleges of Technology, Abu Dhabi Men's College, P.O. Box 25035, Abu Dhabi, United Arab Emirates

5 Department of Electronics Engineering, Universidad Veracruzana, Venustiano Carranza S/N, Col. Revolución, 93390 Poza Rica, Veracruz, Mexico

6 Department of Mathematics, Zhejiang University, Hangzhou, China

Laplace transform–homotopy perturbation method seen, from several points of view. First, the proposed method will follow a strategy which introduces an initial trial function and proposes to cancel the residual error in several points of the interest interval. We will see that the above approach will accelerate the convergence of the proposed solution. Later on, we will apply the method for the case of nonlinear problems with variable coefficients. As case study, we present the problem of finding an analytical approximate solution for the steady-state reaction–diffusion nonlinear differential equation representing the oxygen diffusion in a spherical cell, including nonlinear Michaelis–Menten uptake kinetics [7, 8]. The importance of this model is that it was originally proposed in order to represent the distribution of oxygen inside a cell [9].

We will extend the proposed methodology to problems with Robin boundary conditions, which are indeed difficult to model [7]. Our results will show the potential of VCLT–HPM in the search for analytical approximate solutions for ODEs under the mentioned conditions. A relevant point of our proposal is its contribution to the search for solutions to nonlinear problems. As it is well known that the importance of research on nonlinear differential equations relies on the fact that many phenomena, practical or theoretical, are of nonlinear nature. For the same reason, several methods focused on finding approximate solutions to nonlinear differential equations have been reported, such as those based on variational approaches [10–13], tanh method [14], exp-function [15, 16], Adomian's decomposition method (ADM) [17–20], parameter expansion [21], homotopy perturbation method (HPM) [2–6, 22–41], homotopy analysis method (HAM) [42–46], series method [47–50], group method of data handling (GMDH) [51], differential transform method and the Padé approximation (DTM–Padé technique) [52, 53], and perturbation method (P.M.) [54, 55] among many others.

The rest of this work is organized as follows. In Sect. 2, we introduce the basic idea of standard HPM. Section 3 presents a review of LT–HPM, although above all explain the VCLT–HPM, emphasizing in the modifications mentioned above. Additionally, Sect. 4 presents the interesting case study of the oxygen diffusion in a spherical cell. In addition, a discussion on the results is presented in Sect. 5. Finally, a brief conclusion is given in Sect. 6.

2 Standard HPM

The standard homotopy perturbation method (HPM) was proposed by Ji Huan He, and it was introduced to approach various kinds of nonlinear problems. The HPM is considered as a combination of the classical perturbation technique and the homotopy (whose origin is in the topology), but it is not restricted to small parameters as occurred with traditional perturbation methods [23, 24].

To figure out how HPM works, consider a general nonlinear differential equation in the form

$$A(u) - f(r) = 0, \quad r \in \Omega, \tag{1}$$

with the following boundary conditions

$$B(u, \partial u/\partial n) = 0, \quad r \in \Gamma, \tag{2}$$

where A is a general differential operator, B is a boundary operator, $f(r)$ is a known analytical function, and Γ is the domain boundary for Ω. A can be divided into two operators, L and N, where L is linear and N nonlinear, so that (1) can be rewritten as

$$L(u) + N(u) - f(r) = 0. \tag{3}$$

Generally, a homotopy can be constructed as [23, 24]

$$\begin{aligned} H(U,p) &= (1-p)[L(U) - L(u_0)] + p[L(U) + N(U) - f(r)] \\ &= 0, \quad p \in [0,1], \quad r \in \Omega. \end{aligned} \tag{4}$$

or

$$\begin{aligned} H(U,p) &= L(U) - L(u_0) + p[L(u_0) + N(U) - f(r)] \\ &= 0, \quad p \in [0,1], \quad r \in \Omega, \end{aligned} \tag{5}$$

where p is a homotopy parameter, whose values are within the range of 0 and 1 and u_0 is the first approximation for the solution of (3) that satisfies the boundary conditions.

Assume that solution for (4) or (5) can be written as a power series of p as

$$U = v_0 + v_1 p + v_2 p^2 + \cdots \tag{6}$$

Substituting (6) into (5) and equating identical powers of p terms, values for the sequence v_0, v_1, v_2, … can be found.

When $p \to 1$, it yields the approximate solution for (1) in the form

$$U = v_0 + v_1 + v_2 + v_3 \cdots \tag{7}$$

3 Description of VCLT–HPM

The objective of this section is to show how VCLT–HPM can be employed to find analytical approximate solutions for ODEs such as (3), but with variable coefficients and Robin boundary conditions. We start introducing the basic idea of LT–HPM [4].

3.1 Basic idea of the method LT–HPM

LT–HPM follows the same steps of standard HPM until (5); next we apply L.T. on both sides of homotopy Eq. (5), to obtain [1–6, 22]

$$\Im\{L(U) - L(u_0) + p[L(u_0) + N(U) - f(r)]\} = 0, \tag{8}$$

and using the differential property (49) of L.T., we have [1]:

$$s^n\Im\{U\} - s^{n-1}U(0) - s^{n-2}U'(0) - \cdots - U^{(n-1)}(0) = \Im\{L(u_0) - pL(u_0) + p[-N(U) + f(r)]\} \quad (9)$$

or

$$\Im(U) = \left(\frac{1}{s^n}\right)\left\{s^{n-1}U(0) + s^{n-2}U'(0) + \cdots + U^{(n-1)}(0)\right\} + \left(\frac{1}{s^n}\right)\Im\{L(u_0) - pL(u_0) + p[-N(U) + f(r)]\} \quad (10)$$

Applying inverse L.T. to both sides of (10), we obtain

$$U = \Im^{-1}\left\{\left(\frac{1}{s^n}\right)\left\{s^{n-1}U(0) + s^{n-2}U'(0) + \cdots + U^{(n-1)}(0)\right\} + \left(\frac{1}{s^n}\right)\Im\{L(u_0) - pL(u_0) + p[-N(U) + f(r)]\}\right\} \quad (11)$$

Assume that the solutions of (3) can be expressed as a power series of p

$$U = \sum_{n=0}^{\infty} p^n v_n. \quad (12)$$

Then substituting (12) into (11), we get

$$\sum_{n=0}^{\infty} p^n v_n = \Im^{-1}\left\{\begin{array}{l}\left(\frac{1}{s^n}\right)\left\{s^{n-1}U(0) + s^{n-2}U'(0) + \cdots + U^{(n-1)}(0)\right\} \\ \quad + \left(\frac{1}{s^n}\right)\Im\left\{L(u_0) - pL(u_0) + p\left[-N\left(\sum_{n=0}^{\infty} p^n v_n\right) + f(r)\right]\right\}\end{array}\right\}, \quad (13)$$

and comparing coefficients of p with the same power leads to

$$p^0 : v_0 = \Im^{-1}\left\{\left(\frac{1}{s^n}\right)\left(s^{n-1}U(0) + s^{n-2}U'(0) + \cdots + U^{(n-1)}(0)\right) + \Im\{L(u_0)\}\right\},$$

$$p^1 : v_1 = \Im^{-1}\left\{\left(\frac{1}{s^n}\right)(\Im\{-N(v_0) - L(u_0) + f(r)\})\right\},$$

$$p^2 : v_2 = \Im^{-1}\left\{\left(\frac{1}{s^n}\right)\Im\{-N(v_0, v_1)\}\right\},$$

$$p^3 : v_3 = \Im^{-1}\left\{\left(\frac{1}{s^n}\right)\Im\{-N(v_0, v_1, v_2)\}\right\}, \quad (14)$$

…

$$p^j : v_j = \Im^{-1}\left\{\left(\frac{1}{s^n}\right)\Im\{-N(v_0, v_1, v_2, \ldots, v_{j-1})\}\right\},$$

…

Assume that $U(0) = u_0 = \alpha_0$, $U'(0) = \alpha_1, \ldots, U^{n-1}(0) = \alpha_{n-1}$; the exact solution may be obtained as follows

$$u = \lim_{p\to 1} U = v_0 + v_1 + v_2 + \cdots \quad (15)$$

3.2 VCLT–HPM

To obtain (9), we assumed that the coefficient of $L(U)$ is one. In this work, we will consider the case where the mentioned coefficient is a positive whole power of r; thus, we rewrite explicitly (8) for this case as

$$\Im\{(r^m L(U) - r^m w(r)) + p[r^m w(r) + N(U) - f(r)\} = 0, \quad (16)$$

where m is a positive integer, and employing the versatility and freedom of homotopy formulation, we have substituted $L(u_0)$ for an arbitrary function $w(r)$. An adequate guide for the kind of problems proposed in this work would be to choose $w(r)$ as a polynomial trial function provided with some unknown parameters $A, B, C, ..$ to be determined (see Sect. 5).

Using the properties (49) and (50), we have [1]

$$(-1)^m \frac{d^m\left[s^n\Im\{U\} - s^{n-1}U(0) - s^{n-2}U'(0) - \cdots - U^{(n-1)}(0)\right]}{\mathrm{d}s^m} = \Im\{r^m w(r) + p[-w(r)r^m - N(U) + f(r)]\} \quad (17)$$

and after integrating m times, we obtain

$$s^n\Im\{U\} - s^{n-1}U(0) - s^{n-2}U'(0) - \cdots - U^{(n-1)}(0) = (-1)^m \int\int\cdots\int \Im\{r^m w(r) + p[-w(r)r^m - N(U) + f(r)]\}\mathrm{d}s\mathrm{d}s'\cdots\mathrm{d}s'' \ (m\text{-times}) \quad (18)$$

or

$$U = \Im^{-1}\left\{\frac{1}{s^n}\left\{s^{n-1}U(0) + s^{n-2}U'(0) + \ldots + U^{(n-1)}(0).\right.\right.$$
$$+ (-1)^m \int\int\cdots\int \Im\{r^m w(r) + p[-w(r)r^m - N(U)$$
$$\left.\left. + f(r)]\}\mathrm{d}s\mathrm{d}s'\cdots\mathrm{d}s''\right\}\right\} \quad (19)$$

Assume also that the solutions of the ODE to solve can be expressed as a power series of p

$$U = \sum_{n=0}^{\infty} p^n v_n. \quad (20)$$

Then substituting (20) into (19), we get

$$\sum_{n=0}^{\infty} p^n v_n = \Im^{-1}\left\{\begin{array}{l}\left(\frac{1}{s^n}\right)\left\{s^{n-1}U(0) + s^{n-2}U'(0) + \cdots + U^{(n-1)}(0)\right\} \\ + \left(\frac{(-1)^m}{s^n}\right)\int\int\cdots\int \Im\left\{r^m w(r) + p\left[-r^m w(r) - N\left(\sum_{n=0}^{\infty} p^n v_n\right) + f(r)\right]\right\}\mathrm{d}s\mathrm{d}s'\cdots\mathrm{d}s''\end{array}\right\}, \quad (21)$$

Assume that $U(0) = u_0 = \alpha_0, U'(0) = \alpha_1, \ldots, U^{n-1}(0) = \alpha_{n-1}$; then an approximate solution may be obtained as follows

$$u = \lim_{p\to 1} U = v_0 + v_1 + v_2 + \cdots \quad (23)$$

For boundary value problems, it is expected that some of the initial conditions mentioned above are initially unknown; therefore, (23) can be expressed as

$$u = u(r, A, B, C, \ldots, \alpha_i) \quad (24)$$

In order to calculate adequately the values for $A, B, C, \ldots, \alpha_i$, we will deduce an algebraic system for them, in the following way

and comparing coefficients of p, with the same power leads to

$$p^0 : v_0 = \Im^{-1}\left\{\left(\frac{1}{s^n}\right)(s^{n-1}U(0) + s^{n-2}U'(0) + \cdots + U^{(n-1)}(0)) + (-1)^m\int\int\cdots\int \Im\{r^m w(r)\}\mathrm{d}s\mathrm{d}s'\cdots\mathrm{d}s''\right)\right\},$$

$$p^1 : v_1 = \Im^{-1}\left\{\left(\frac{1}{s^n}\right)(-1)^m\int\int\cdots\int (\Im\{-N(v_0) - r^m w(r) + f(r)\})\mathrm{d}s\mathrm{d}s'\cdots\mathrm{d}s''\right\},$$

$$p^2 : v_2 = \Im^{-1}\left\{\left(\frac{1}{s^n}\right)(-1)^m\int\int\cdots \int (\Im\{-N(v_0, v_1)\})\mathrm{d}s\mathrm{d}s'\cdots\mathrm{d}s''\right\},$$

$$p^3 : v_3 = \Im^{-1}\left\{\left(\frac{1}{s^n}\right)(-1)^m\int\int\cdots \int (\Im\{-N(v_0, v_1, v_2)\})\mathrm{d}s\mathrm{d}s'\cdots\mathrm{d}s''\right\}, \quad (22)$$

$$\ldots$$

$$p^j : v_j = \Im^{-1}\left\{\left(\frac{1}{s^n}\right)(-1)^m\int\int\cdots \int \left(\Im\left\{-N(v_0, v_1, v_2, ..v_{j-1})\right\}\right)\mathrm{d}s\mathrm{d}s'\cdots\mathrm{d}s''\right\},$$

$$\ldots$$

1. It is required that (24) satisfies the Robin boundary conditions at the endpoint of the interval.
2. In order to find the values of the total number of parameters, we will require adding more algebraic equations to those found in (1), until obtaining as many equations as parameters to determine. If they are required, let us say j additional equations, then a convenient possibility is to incorporate the following j equations $R(r_1, A, B, C, \ldots, \alpha_i) = R(r_2, A, B, C, \ldots, \alpha_i) = \cdots = R(r_j, A, B, C, \ldots, \alpha_i) = 0$ [25], where the residual value is defined by the substitution of (24) into (3), to obtain $R(r, A, B, C, \ldots, \alpha_i) = r^m L(u(r, A, B, C, \ldots, \alpha_i)) + N(u(r, A, B, C, \ldots, \alpha_i)) - f(r)$ (we have considered the case of ODEs with variable coefficients). The above points $r_1, r_2, r_3, .., r_j$ belong to the interest interval, and it is assumed that $u(r, A, B, C, .., \alpha_i)$ is the approximate solution of (3) given by (24).

For $m = 1$, the above procedure involves only one integration. For the case of second-order ODEs, where $y'(0) = A \neq 0$, it is possible to show that the term containing $y'(x)$ may give rise to an inappropriate term $\Im^{-1}\left\{\frac{\mathrm{Ln}(s)}{s^2}\right\}$. To avoid this problem, we perform an adequate transformation in order to express our differential equation in its normal form [56]. As it is known, this normal form does not contain the offending term $y'(x)$ and

therefore allows applying VCLT–HPM algorithm in accordance with the above procedure.

It is worthwhile to mention that this procedure can be employed in a similar manner for the case of nonlinear differential equations with variable coefficients defined with Dirichlet, mixed and Neumann boundary conditions.

From the above, we see that there are several points to highlight LT–HPM, with respect to HPM.

1. While LT–HPM calculates the solutions of the different orders in a systematic way using basic Laplace transforms, the HPM generates a cumbersome set of coupled differential equations to calculate the solutions of the different orders arising from its iterative process.
2. Unlike the standard HPM, which incorporates from the beginning the boundary conditions of the problem, LT–HPM incorporates also one of such conditions from the start, while the other one, to the end of the process, i.e., to the final approximate solution, stemming from LT–HPM algorithm.
3. Unlike HPM standard, LT–HPM (also VCLT–HPM) may incorporate some adjustment parameters whose values are adequately determined in order to get handy and accurate analytical approximate solutions.
4. In particular, VCLT–HPM is a suitable method for solving ODEs with variable coefficients, while the standard formulation HPM could be inadequate if the differential equations of its iterative process have singular points in the domain of study.

We summarize the key points of this proposal as follows:

(a) First, VCLT–HPM requires that the coefficient of $L(U)$ is a power whole of r. If such function was constant, then LT–HPM would be adequate.

(b) Next, we apply the homotopy formulation given by (16), but substitute $L(u_0)$ for a trial function containing some parameters.

(c) Then, we apply the VCLT–HPM algorithm described above. Since the proposed method is iterative, we assume that the sought solution can be expressed as a power series of p (20).
It is worthwhile to mention that VCLT–HPM, unlike HPM and LT–HPM, calculates the different orders, in terms of elementary integrals of the lower orders, previously calculated.

(d) Once we get a series solution (23), we determine adequately the above-mentioned quantities in (b) through a system of algebraic equations, requiring that proposed solution satisfies some boundary conditions, and by residual error cancelation in some points of the interest interval.

(e) Our final approximate solution is obtained by substituting the calculated coefficients from step (d), into (23).

4 Case study

The objective of this section is to employ VCLT–HPM, to find an analytical approximate solution for the nonlinear singular boundary value problem in the Lane–Emden form, given in dimensionless form as [7].

$$y''(x)+\frac{2y'(x)}{x}=\alpha\frac{y(x)}{K+y(x)}\quad 0<x\leq 1, \tag{25}$$

where radius of the cell corresponds to $x=1$ and the boundary conditions $y'(0)=0$, $y'(1)+Hy(1)=H$ (Robin boundary condition).

The above differential equation describes the oxygen diffusion in a spherical cell with Michaelis–Menten uptake kinetics, given at the right-hand side of (25) [7].

In order to obtain an approximate analytical solution, we rewrite (25) as

$$Kxy''+xyy''+2Ky'+2yy'-\alpha xy=0, \tag{26}$$

where prime denotes differentiation with respect to x.

We identify terms:

$$L(y)=Kxy'', \tag{27}$$

$$N(y)=xyy''+2Ky'+2yy'-\alpha xy. \tag{28}$$

Next we construct a homotopy as follows

$$(1-p)(Kxy''-KxA)+p[Kxy''+xyy''+2Ky'+2yy'-\alpha xy]=0, \tag{29}$$

where we have chosen as a polynomial trial function $w(x)=A$ (where A is a constant; see (16) and following comments).

Next we rewrite homotopy Eq. (29) as

$$Kxy''=KxA+p[-KxA-xyy''-2Ky'-2yy'+\alpha xy]. \tag{30}$$

Applying L.T. algorithm, we get

$$\Im\{Kxy''\}=\Im\{KxA+p[-KxA-xyy''-2Ky'-2yy'+\alpha xy]\}. \tag{31}$$

As it is explained in [1], it is possible to rewrite (31) as (50)

$$-K\frac{\mathrm{d}(s^2Y(s)-sB)}{\mathrm{d}s}=\Im\{KxA+p[-KxA-xyy''-2Ky'-2yy'+\alpha xy]\}. \tag{32}$$

After integrating (32), we obtain

$$Y(s) = \frac{B}{s} + \left(\frac{-1}{Ks^2}\right) \int \Im\{KxA + p[-KxA - xyy'' - 2Ky' - 2yy' + \alpha xy]\}\mathrm{d}s,$$

where we defined $B = y(0)$ and employed the condition $y'(0) = 0$.

After applying $\Im^{-1}$ to the previous integral expression, we get

$$y(x) = \Im^{-1}\left\{\frac{B}{s} + \left(\frac{-1}{Ks^2}\right) \int \Im\{KxA + p[-KxA - xyy'' - 2Ky' - 2yy' + \alpha xy]\}\mathrm{d}s\right\} \quad (33)$$

Next, we assume a series solution for $y(x)$, in the form

$$y(x) = \sum_{n=0}^{\infty} p^n v_n, \quad (34)$$

Substituting (34) into (33), we get

$$\sum_{n=0}^{\infty} p^n v_n = \Im^{-1}\left\{\frac{B}{s} + \left(\frac{-1}{Ks^2}\right) \int \Im\left\{KxA + p\left[\begin{array}{c} -KxA - x\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} p^m p^n v_m v_n'' - 2K\sum_{n=0}^{\infty} p^n v_n' \\ -2\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} p^m p^n v_m v_n' + \alpha x \sum_{n=0}^{\infty} p^n v_n \end{array}\right]\right\}\mathrm{d}s\right\} \quad (35)$$

On comparing the coefficients of identical powers of p, and after performing the indicated operations, we have

$$p^0 : v_0 = \Im^{-1}\left\{\frac{B}{s} + \left(\frac{-1}{Ks^2}\right) \int \Im\{KxA\}\mathrm{d}s\right\}, \quad (36)$$

$$p^1 : v_1 = \Im^{-1}\left\{\left(\frac{-1}{Ks^2}\right) \int \Im\{[-KxA - xv_0v_0'' - 2Kv_0' - 2v_0v_0' + \alpha xv_0]\}\mathrm{d}s\right\}, \quad (37)$$

…

Solving the above operations for $v_0(x)$ and $v_1(x)$, we obtain

$$p^0 : v_0(x) = B + \frac{A}{2}x^2, \quad (38)$$

$$p^1 : v_1(x) = -\frac{A}{2}x^2 + \frac{(-2KA - 3AB + \alpha B)}{2K}x^2 + \frac{\alpha A - 3A^2}{24K}x^4, \quad (39)$$

…

and so on.

By substituting solutions (38) and (39) into (34) and calculating the limit when $p \to 1$, it results in a handy first-order approximation.

$$y(x) = B + \frac{(-2KA - 3AB + \alpha B)}{2K}x^2 + \frac{\alpha A - 3A^2}{24K}x^4. \quad (40)$$

In accordance with VCLT–HPM algorithm (Sect. 3.2), we will calculate the values of A and B, through a system of algebraic equations, as follows:

1. The first equation for A and B is obtained, requiring that (40) satisfies the Robin boundary condition $y'(1) + Hy(1) = H$.
2. On the other hand, a second equation is obtained by substituting (40) into (26) and evaluating the expression obtained for some value, let us say $x = 1/5$, which lies into the interval under study $0 \le x \le 1$ (see discussion below).
3. Finally, we solve the above system of algebraic equations.

Considering the case study: $\alpha = 1$, $K = 10$, $H = 4$, and $\alpha = 0.1$, $K = 5$, $H = 4$, we obtain the following values

$$A = 0.02968850376, \quad B = 0.9775229350, \quad (41)$$

and

$$A = 0.005536482514, \quad B = 0.9958404758, \quad (42)$$

respectively.

After substituting (41) into (40), we obtain a first-order approximation

$$y(x) = 0.9775229350 + 0.01483446399x^2 + 0.0001126845083x^4, \quad (43)$$

and in the same way, the substitution of (42) into (40) leads to the following handy first-order approximation.

$$y(x) = 0.9958404758 + 0.002767886230x^2 + 0.000003847419462x^4. \quad (44)$$

5 Discussion

In order to use a pure numerical solution as reference, we utilized the scheme based on trapezoid combined with Richardson extrapolation from the built-in numerical

routines provided by Maple 15. Moreover, the command was set up with an absolute error (A.E.) tolerance of 1×10^{-12}.

This article generalizes what has been published in [4], since it considers the solution of nonlinear differential equations with singular points; as it was noted, these problems can be rewritten in terms of a differential equation with variable coefficients. Therefore, this paper proposes the VCLT–HPM in order to find approximate solutions to these problems. An important fact is to mention that the methodology is based on the systematic use of basic operations on the L.T. which can be found in many references and textbooks [1] [see also "Appendix"; it is easy to verify from our case study that the proposed procedure relies heavily on the use of the simpler results (47), (48), (49), (50)]. VCLT–HPM expresses the problem to be solved in terms of a differential equation for L.T. $Y(s)$, but unlike the original nonlinear differential equation to solve, this is easily separable [see (17–19)]. Once $Y(s)$ is expressed in terms of $y(x)$ after applying Laplace inverse transform $\Im^{-1}$, we assume a series solution for $y(x)$ in the form (20), and from here on, VCLT–HPM calculates the different approximate orders in a similar fashion as LT–HPM [4], but unlike this, its nth iterative process is expressed in terms of integrals of the lower-order approximations which are previously calculated [compare (14) and (22)]. On the other hand, since some boundary value problems entail serious difficulties, even to get approximate solutions, this proposed article exploits the freedom of VCLT–HPM, explained in Sect. 3.2, in order to obtain a highly accurate solution to the proposed nonlinear problem. LT–HPM has been employed successfully to solve nonlinear problems defined for both initial and finite boundary conditions [2–6, 22, 40, 41]; however, those works followed one assumption from original HPM, which avoids LT–HPM to have a better performance. In accordance with HPM, u_0 is the first approximation for the solution of (3) that satisfies the boundary conditions, which results too restrictive, since HPM admits in principle the validity of homotopy formulation (5) for any choosing of the aforementioned function.

On the other hand, [4–6] showed that the searching for polynomial solutions of nonlinear problems with finite boundary conditions is a possibility with high potential; therefore, VCLT–HPM substitutes $L(u_0)$ as it was originally defined, for polynomial functions $w(r)$, provided with unknown parameters $A, B, C\ldots$, which have to be adequately determined through an algebraic system of equations in order to obtain analytical approximate solutions, following the procedure explained in Sect. 3.2. We highlight that these equations were obtained, in the one hand requiring that the residual becomes zero, for several points distributed along the interest interval, and on the other hand applying the Robin boundary condition. We note that one of the important features of VCLT–HPM is that the high complexity of problems with singular points and Robin boundary conditions was effectively handled by this technique and for the same reason, the proposed method is considered a generalization of [4] where LT–HPM was used to find approximate solutions for nonlinear problems with Dirichlet, mixed, and Neumann boundary conditions. Unlike [4], this work presents the important case study of the oxygen diffusion in a spherical cell, with a Robin right boundary condition. This kind of boundaries (also called third-type boundary condition) are considered as a combination of Dirichlet and Neumann conditions and are often used to solve problems of Sturm–Liouville that appear in science and engineering.

As it is well known from literature, modeling the closer region to the unknown endpoints of an interval turns out to be sometimes indeed, very difficult [54] (such as occurred with the case of Robin boundary conditions). Indeed, the problem studied in this work turns out to be particularly complicated because of the presence of singular points and because the endpoints are not given from start.

Figures 1 and 2 show the comparison between numerical solution of (25) and VCLT–HPM first-order approximations (43) and (44). From these figures, it is clear that (40) provides a good approximation to the solution to (25). In more precise terms, it is possible to verify the accuracy of our results by calculating the square residual error (S.R.E) of (43) and (44) defined as $\int_a^b R^2(u(r))\mathrm{d}r$, where a and b are the end points, the residual $R(u(r))$, which already was defined in Sect. 3, and $u(r)$ is an approximate solution to the equation to be solved, in our case (25) [25]. As it can be seen, the square residual error (S.R.E.) is in general terms a positive number, which is representative of the total error committed, by using the approximate solution $u(r)$. S.R.E will be zero only for the case where $u(r)$ is

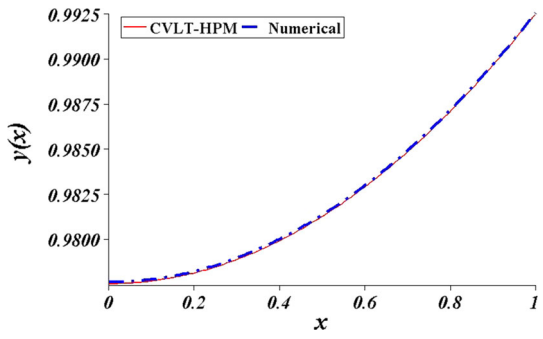


Fig. 1 Comparison between numerical solution of (25) and VCLT–HPM first-order approximation (43)

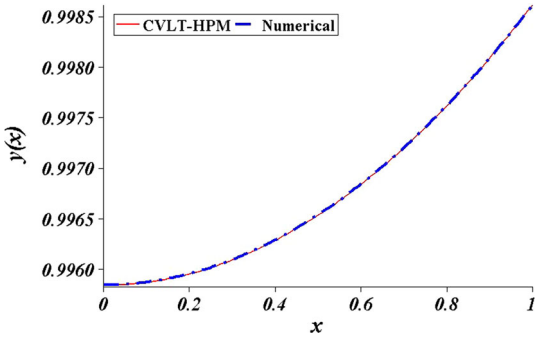


Fig. 2 Comparison between numerical solution of (25) and VCLT–HPM first-order approximation (44)

the exact solution for the differential equation under study. The resulting values were of $1.808511632 \times 10^{-7}$ and $2.560574954 \times 10^{-10}$, respectively, which confirm the accuracy of the proposed method. If more accuracy has to be required, one must consider higher-order approximations [remember that (40) is just a first-order approximation of (25)]. Another possibility would be to use a higher-order polynomial than $w(x) = A$ (A constant), since it would contain more adjustment parameters, whereby it is expected to get a better approximation; what is more could be employed both strategies for better results. We note that residual is other useful manner to show the accuracy of an approximate analytical solution. From Fig. 3, we conclude the high precision of the proposed solutions (43) and (44).

Should be mentioned that this problem has been successfully studied for several authors, although most of them have proposed just numerical approximations [9, 57]. Rach et al. [7] proposed an analytical approximate solution for (25), and its methodology consisted in expressing the resulting equation (25), first in terms of a Volterra integral equation, and then the last one in an equivalent Fredholm–Volterra integral form. Finally the Adomian decomposition method is employed in order to solve the mentioned Fredholm–Volterra integral representation of (25). Unlike this methodology, the method proposed in this paper is not restricted only to this kind of problems, but it can be extended in principle to other problems with singular points [see (16) and Sect. 3.2]. On the other hand, it is necessary to comment that the above procedure is much more complicated than the one introduced by this work for the solution of nonlinear problems like (25), since VCLT–HPM is handy and given that it is based on both elementary Laplace transforms and integrals (see "Appendix"), making it an ideal tool for practical applications (it is known that although Adomian is a powerful technique, the process of obtaining its polynomial solutions is not straightforward for practical applications). In the same way, it is expected that the proposed method will contribute to overcoming some problems that face other semi-analytic techniques. Unlike VCLT–HPM, perturbation method depends on a parameter assumed small, which is considered as a disadvantage of P.M. On the other hand, HAM is accurate and powerful, but sometimes its approximate solutions turn out to be long and cumbersome, and for the same reason, they are not adequate for practical applications, while LT–HPM [it is expected that also VCLT–HPM; see (40)] has already reported handy accurate analytical solutions, by using polynomial with only a few terms [4–6, 40, 41].

The advantages of LT–HPM and VCLT–HPM, with respect to HPM, were already discussed in detail in Sect. 3.2.

Emphasizing the general characteristics of the method VCLT–HPM, it is an iterative method which is based on elementary L.T. properties [1] and simple algebraic steps which make it an ideal technique for practical applications. In particular, for the case of boundary value problems, the proposed method expresses the problem of solving a nonlinear differential equation in terms of the resolution of a system of algebraic equations for some unknown initial conditions and some unknown parameters. The calculation of these quantities is made so that the solutions arising from the proposed method are accurate and practical.

From the aforementioned discussion, it is expected that LT–HPM (VCLT–HPM for nonlinear ODEs with variable coefficients) can be applied to other areas of the knowledge like fluids.

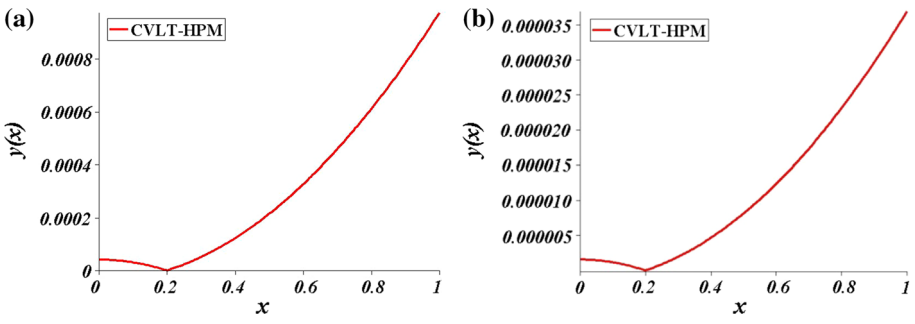


Fig. 3 Residual for VCLT–HPM approximations (43) and (44), **a** and **b,** respectively

In fact, LT–HPM was successfully applied in order to find an approximate solution for the problem of an axisymmetric Newtonian fluid squeezed between two large parallel plates [5].

In the same manner, Hossein [22] employed LT–HPM in order to find an approximate solution to the Blasius nonlinear differential equation that describes the boundary layer of a two-dimensional viscous laminar flow over a semi-infinite flat plate. We noted that applications for boundary layer problems for more general conditions (for instance, for a coupled set of nonlinear ordinary differential equations [45]) may require a generalization of the proposed method, which could be useful for the solution of problems in other areas of knowledge.

6 Conclusions

This work introduced VCLT–HPM as a novel modification of LT–HPM in order to find analytical approximate solutions for nonlinear ordinary differential equations defined on finite intervals with singular points (as it was noted, these problems can be rewritten in terms of a differential equation with variable coefficients) and Robin boundary conditions. Keeping the benefits of LT–HPM [4], the proposed method exploits the flexibility of homotopy formulation, in order to obtain initially unknown parameters, which are adequately determined, in order to obtain highly accurate analytical approximate solutions to nonlinear problems. This is accomplished through the solution of an algebraic system of equations which is derivative, demanding in the one hand that Robin condition is satisfied at the right end of the interval, and on the other hand by canceling the residual error in several points of the interval of interest. Our case study, the important problem of the oxygen diffusion in a spherical cell, showed that VCLT–HPM is a method with potential to accelerate the convergence of the solution for a given nonlinear problem including the region close to unknown endpoints. Despite the fact that we only employed the first-order approximation, we obtained a handy precise solution.

Acknowledgments We gratefully acknowledge the financial support from the National Council for Science and Technology of Mexico (CONACyT) through Grant CB-2010-01 #157024. The authors would like to thank Rogelio Alejandro Callejas-Molina and Roberto Ruiz-Gomez for their contribution to this project.

Compliance with ethical standards

Conflict of interest The authors declare that there is no conflict of interests regarding the publication of this paper.

Appendix

Laplace transform of $F(t)$ is denoted by $\Im\{F(t)\}$ and is defined by the integral [1]

$$\Im\{F(t)\} = f(s) = \int_0^\infty e^{-st} F(t)\mathrm{d}t. \tag{45}$$

Linearity of L.T. is an important property, which is expressed as

$$\Im\{c_1F_1(t) + c_2F_2(T)\} = c_1f_1(s) + c_2f_2(s), \tag{46}$$

where c_1 and c_2 are constants, and we have denoted: $\Im\{F_1(t)\} = f_1(s)$, $\Im\{F_2(t)\} = f_2(s)$.

Some known properties of L.T. widely employed in this study are

1. $$\Im\{1\} = \frac{1}{s}\ (s > 0) \tag{47}$$

2. $$\Im\{t^n\} = \frac{n!}{s^{n+1}}\ (s > 0) \tag{48}$$

3. $$\Im\left\{F^{(n)}(t)\right\} = s^n f(s) - s^{n-1}F(0) - s^{n-2}F'(0) - \cdots - F^{(n-1)}(0), \tag{49}$$

where $F^{(n)}(t)$ denotes the nth derivative of $F(t)$ and $\Im\{F(t)\} = f(s)$.

4. $$\Im\{t^nF(t)\} = (-1)^n \frac{\mathrm{d}^n f(s)}{\mathrm{d}t^n}, \tag{50}$$

n denotes a positive integer.

If L.T. of $F(t)$ is $f(s)$, then $F(t)$ is called the inverse L.T. of $f(s)$ and is expressed by $F(t) = \Im^{-1}\{f(s)\}$, where $\Im^{-1}$ is called the inverse L.T. operator.

From Eqs. (47) and (48), it is clear that

$$1 = \Im^{-1}\left(\frac{1}{s}\right), \tag{51}$$

$$t^n = \Im^{-1}\left(\frac{n!}{s^{n+1}}\right), \tag{52}$$

and so on.

The following important result is obtained from (46) and denotes the linearity property of $\Im^{-1}$

$$\Im^{-1}\{c_1f_1(s) + c_2f_2(s)\} = c_1F_1(t) + c_2F_2(T). \tag{53}$$

References

1. Spiegel MR (1998) Teoría y Problemas de Transformadas de Laplace, primera edición. Serie de compendios Schaum, McGraw-Hill, México

2. Aminikhan H, Hemmatnezhad M (2012) A novel effective approach for solving nonlinear heat transfer equations. Heat Transf Asian Res 41(6):459–466
3. Aminikhah Hossein (2012) The combined Laplace transform and new homotopy perturbation method for stiff systems of ODEs. Appl Math Model 36:3638–3644
4. Filobello-Nino U, Vazquez-Leal H, Khan Y, Perez-Sesma A, Diaz-Sanchez A, Jimenez-Fernandez VM, Herrera-May A, Pereyra-Diaz D, Mendez-Perez JM, Sanchez-Orea J (2013) Laplace transform–homotopy perturbation method as a powerful tool to solve nonlinear problems with boundary conditions defined on finite intervals. Comput Appl Math. ISSN:0101-8205. doi:10.1007/s40314-013-0073-z
5. Filobello-Nino U, Vazquez-Leal H, Cervantes-Perez J, Benhammouda B, Perez-Sesma A, Hernandez-Martinez L, Jimenez-Fernandez VM, Herrera-May AL, Pereyra-Diaz D, Marin-Hernandez A, Huerta Chua J (2014). A handy approximate solution for a squeezing flow between two infinite plates by using of Laplace transform–homotopy perturbation method. SpringerPlus, 3:421, 10 pp. doi:10.1186/2193-1801-3-421
6. Filobello-Nino U, Vazquez-Leal H, Benhammouda B, Hernandez-Martinez L, Hoyos-Reyes C, Perez-Sesma JAA, Manuel Jimenez-Fernandez V, Pereyra-Diaz D, Marin-Hernandez A, Diaz-Sanchez A, Huerta-Chua J, Cervantes-Perez J (2014) Nonlinearities distribution Laplace transform–homotopy perturbation method. SpringerPlus 3:594. doi:10.1186/2193-1801-3-594
7. Rach R, Wazwaz A-M, Duan J-S (2014) A reliable analysis of oxygen diffusion in a spherical cell with nonlinear oxygen uptake kinetics. Int J Biomath 7(2):1450020. doi:10.1142/S17935245 1450020X
8. Vazquez-Leal H, Sandoval-Hernandez M, Castaneda-Sheissa R, Filobello-Nino U, Sarmiento-Reyes A (2015) Modified Taylor solution of equation of oxygen diffusion in a spherical cell with Michaelis-Menten uptake kinetics. J Appl Math Res 4(2): 253–258. doi:10.14419/ijamr.v4i2.4273
9. Lin SH (1976) Oxygen diffusion in a spherical cell with nonlinear oxygen uptake kinetics. J Theor Biol 60:449–457
10. Assas LMB (2007) Approximate solutions for the generalized K-dV- Burgers' equation by He's variational iteration method. Phys Scr 76:161–164. doi:10.1088/0031-8949/76/2/008
11. He JH (2007) Variational approach for nonlinear oscillators. Chaos Solitons Fractals 34:1430–1439. doi:10.1016/j.chaos.2006.10.026
12. Kazemnia M, Zahedi SA, Vaezi M, Tolou N (2008) Assessment of modified variational iteration method in BVPs high-order differential equations. J Appl Sci 8:4192–4197. doi:10.3923/jas.2008.4192.4197
13. Noorzad R, Tahmasebi Poor A, Omidvar M (2008) Variational iteration method and homotopy-perturbation method for solving Burgers equation in fluid dynamics. J Appl Sci 8:369–373. doi:10.3923/jas.2008.369.373
14. Evans DJ, Raslan KR (2005) The Tanh function method for solving some important nonlinear partial differential. Int J Comput Math 82:897–905. doi:10.1080/00207160412331336026
15. Xu F (2007) A generalized soliton solution of the Konopelchenko–Dubrovsky equation using exp-function method. Zeitschrift Naturforschung Sect A J Phys Sci 62(12):685–688
16. Mahmoudi J, Tolou N, Khatami I, Barari A, Ganji DD (2008) Explicit solution of nonlinear ZK–BBM wave equation using Exp-function method. J Appl Sci 8:358–363. doi:10.3923/jas.2008.358.363
17. Adomian G (1988) A review of decomposition method in applied mathematics. Math Anal Appl 135:501–544
18. Babolian E, Biazar J (2002) On the order of convergence of Adomian method. Appl Math Comput 130(2):383–387. doi:10.1016/S0096-3003(01)00103-5
19. Sheikholeslami M, Ganji DD, Ashorynejad HR, Rokni HB (2012) Analytical investigation of Jeffery–Hamel flow with high magnetic field and nanoparticle by Adomian decomposition method. Appl Math Mech English Edn 33(1):25–36. doi:10.1007/s10483-012-1531-7
20. Sheikholeslami M, Ganji DD, Ashorynejad HR (2013) Investigation of squeezing unsteady nanofluid flow using ADM. Powder Technol 239:259–265
21. Zhang L-N, Xu L (2007) Determination of the limit cycle by He's parameter expansion for oscillators in a potential. Zeitschriftfür Naturforschung Sect A J Phys Sci 62(7–8):396–398
22. Aminikhah H (2011) Analytical approximation to the solution of nonlinear Blasius viscous flow equation by LTNHPM. International Scholarly Research Network ISRN Mathematical Analysis, Volume 2012, Article ID 957473, 10 pp. doi:10.5402/2012/957473
23. He JH (1998) A coupling method of a homotopy technique and a perturbation technique for nonlinear problems. Int J Non-Linear Mech 351:37–43. doi:10.1016/S0020-7462(98)00085-7
24. He JH (1999) Homotopy perturbation technique. Comput Methods Appl Mech Eng 178:257–262. doi:10.1016/S0045-7825(99)00018-3
25. Marinca V, Herisanu N (2011) Nonlinear dynamical systems in engineering, 1st edn. Springer, Berlin
26. Khan M, Gondal MA, Hussain I, Karimi Vanani S (2011) A new study between homotopy analysis method and homotopy perturbation transform method on a semi infinite domain. Math Comput Model 55:1143–1150
27. He JH (2006) Homotopy perturbation method for solving boundary value problems. Phys Lett A 350(1–2):87–88
28. Vazquez-Leal H, Hernandez-Martinez L, Khan Y, Jimenez-Fernandez VM, Filobello-Nino U, Diaz-Sanchez A, Herrera-May AL, Castaneda-Sheissa R, Marin-Hernandez A, Rabago-Bernal F, Huerta-Chua J (2014) Multistage HPM applied to path tracking damped oscillations of a model for HIV infection of CD4 + T cells. Br J Math Computer Sci 4(8):1035–1047
29. Vazquez-Leal H, Sarmiento-Reyes A, Khan Y, Filobello-Nino U, Diaz-Sanchez A (2012) Rational biparameter homotopy perturbation method and laplace-padé coupled version. J Appl Math Article ID 923975, 21 pp. doi:10.1155/2012/923975
30. Ganji DD, Mirgolbabaei H, Me M, Mo M (2008) Application of homotopy perturbation method to solve linear and non-linear systems of ordinary differential equations and differential equation of order three. J Appl Sci 8:1256–1261. doi:10.3923/jas.2008.1256.1261
31. Biazar J, Eslami M (2012) A new homotopy perturbation method for solving systems of partial differential equations. Comput Math Appl 62:225–234
32. Vazquez-Leal H, Filobello-Niño U, Castañeda-Sheissa R, HernandezMartinez L, Sarmiento-Reyes A (2012) Modified HPMs inspired by homotopy continuation methods. Math Problems Eng Article ID 309123:20. doi:10.155/2012/309123
33. Vazquez-Leal H, Castañeda-Sheissa R, Filobello-Niño U, Sarmiento-Reyes A, Sánchez-Orea J (2012) High accurate simple approximation of normal distribution related integrals. Math Problem Eng 2012:Article ID 124029, 22 pp. doi:10.1155/2012/124029
34. Filobello-Niño U, Vazquez-Leal H, Castañeda-Sheissa R, Yildirim A, Hernandez ML, Pereyra Díaz D, Pérez Sesma A, Hoyos Reyes C (2012) An approximate solution of Blasius equation by using HPM method. Asian J Math Stat 10. doi:10.3923/ajms.2012
35. Biazar J, Ghazvini H (2009) Convergence of the homotopy perturbation method for partial differential equations. Nonlinear Anal Real World Appl 10(5):2633–2640
36. Sheikholeslami M, Ganji DD, Rokni HB (2013) Nanofluid flow in a semi-porous channel in the presence of uniform magnetic field. IJE Trans C Aspects 26(6):653–662

37. Sheikholeslami M, Ashorynejad HR, Ganji, DD, Kolahdooz A (2011) Investigation of rotating MHD viscous flow and heat transfer between stretching and porous surfaces using analytical method. Hindawi Publishing Corporation Mathematical Problems in Engineering, vol 2011. Article ID 258734, 17. doi:10.1155/2011/258734
38. Sheikholeslami M, Ganji DD (2013) Heat transfer of Cu-water nanofluid flow between parallel plates. Powder Technol 235:873–879
39. Sheikholeslami M, Ashorynejad HR, Ganji D, Yıldırım A (2012) Homotopy perturbation method for three-dimensional problem of condensation film on inclined rotating disk. Scientia Iranica 19(3):437–442
40. Filobello-Nino U, Vazquez-Leal H, Benhammouda B, Shukla AK, Cervantes-Perez J, Perez-Sesma A et al. The study of heat transfer phenomena by using of Laplace transform–homotopy perturbation method. Appl Math Inf Sci (Article currently under review)
41. Filobello-Nino U, Vazquez-Leal H, Benhammouda B, Perez-Sesma A, Jimenez-Fernandez VM, Cervantes-Perez J et al. An easy computable approximate solution for Troesch's problem by using of Laplace transform–homotopy perturbation method. Appl Math Inf Sci (Article currently under review)
42. Rashidi MM, Rastegari MT, Asadi M, Bg OA (2012) A study of non-newtonian flow and heat transfer over a non-isothermal wedge using the homotopy analysis method. Chem Eng Commun 199:231–256
43. Rashidi M, Pour SM, Hayat T, Obaidat S (2012) Analytic approximate solutions for steady flow over a rotating disk in porous medium with heat transfer by homotopy analysis method. Comput Fluids 54:1–9
44. Rashidi MM, Domairry G, Dinarvand S (2009) The homotopy analysis method for explicit analytical solutions of Jaulent–Miodek equations. Numer Methods Partial Differ Equ 25(2): 430–439. doi:10.1002/num.20358
45. Anwar Bég O, Rashidi MM, Bég TA, Asadi M (2012) Homotopy analysis of transient magneto-bio-fluid dynamics of micropolar squeeze film in porous medium: a model for magneto-bio-rheological lubrication. J Mech Med Biol 12(1):1–21. doi:10.1142/S0219519412004648
46. Sheikholeslami M, Ashorynejad HR, Domairry HR, Hashim I (2012) Flow and heat transfer of Cu–Water Nanofluid between a stretching sheet and a porous surface in a rotating system. Hindawi Publishing Corporation. J Appl Math Article ID 421320:18. doi:10.1155/2012/421320
47. Sheikholeslami M, Ellahi R, Ashorynejad HR, Domairry G, Hayat T (2014) Effects of heat transfer in flow of nanofluids over a permeable stretching wall in a porous medium. J Comput Theor Nanosci 11(2):486–496
48. Ince E (1956) Ordinary differential equations. Dover, New York
49. Forsyth A (1906) Theory of differential equations. Cambridge University Press, New York
50. Filobello-Nino U, Vazquez-Leal H, Benhammouda H, Perez-Sesma A, Jimenez-Fernandez VM et al (2015) Analytical solutions for systems of singular partial differential-algebraic equations. Hindawi Publishing Corporation Discrete Dynamics in Nature and Society, vol 2015, Article ID 752523, 9 pp. doi:10.1155/2015/752523
51. Sheikholeslami M, BaniSheykholeslami F, Khoshhal S, Mola-Abasia H, Ganji DD, Rokni HB (2014) Effect of magnetic field on Cu–water nanofluid heat transfer using GMDH-type neural network. Neural Comput Appl 25:171–178. doi 10.1007/s00521-013-1459-y
52. Rashidi MM, Gangi D (2009) New analytical solution of the three dimensional Navier–Stokes equations. Modern Phys Lett B World Sci 23(26). ISSN:1793-6640
53. Rashidi MM, Erfani E The modified differential transform method for investigating nano boundary-layers over stretching surfaces. Int J Numer Methods Heat Fluid Flow 21(7). ISSN:0961-5539
54. Filobello-Nino U, Vazquez-Leal H, Sarmiento-Reyes A, Perez-Sesma A, Hernandez-Martinez L, Herrera-May A, Jimenez-Fernandez VM, Marin-Hernandez A, Pereyra-Diaz D, Diaz-Sanchez A (2013) The study of heat transfer phenomena using PM for approximate solution with Dirichlet and mixed boundary conditions. Appl Comput Math 2(6):143–148. doi:10.11648/j.acm.20130206.16
55. Filobello-Nino U, Vazquez-Leal H, Khan Y, Yildirim A, Jimenez-Fernandez VM, Herrera-May AL, Castaneda-Sheissa R, Cervantes-Perez J (2013) Perturbation method and Laplace–Padé approximation to solve nonlinear problems. Miskolc Math Notes 14(1):89–101
56. Simmons GF (1993) Ecuaciones Diferenciales: Con Aplicaciones Y Notas Históricas, 2da Edición. McGraw-Hill, España
57. McElwain DLS (1978) A re-examination of oxygen diffusion in a spherical cell with nonlinear oxygen uptake kinetics. J Theor Biol 71:255–263

ISSN 0001-4346, Mathematical Notes, 2018, Vol. 104, No. 2, pp. 167–175. © Pleiades Publishing, Ltd., 2018.
Original Russian Text © P.V. Bibikov, 2018, published in Matematicheskie Zametki, 2018, Vol. 104, No. 2, pp. 163–173.

On Differential Invariants and Classification of Ordinary Differential Equations of the Form $y'' = A(x, y)y' + B(x, y)$

P. V. Bibikov[1]*

[1] *Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, 117997 Russia*

Received September 17, 2017

Abstract—The class of second-order ordinary differential equations $y'' = A(x, y)y' + B(x, y)$ is studied by methods of the geometry of jet spaces and the geometric theory of differential equations. The symmetry group of this class of equations is calculated, and the field of differential invariants of its action on equations is described. These results are used to state and prove a criterion for the local equivalence of two nondegenerate ordinary differential equations of the form $y'' = A(x, y)y' + B(x, y)$, in which the coefficients A and B are rational in x and y.

DOI: 10.1134/S0001434618070180

Keywords: *ordinary differential equation, symmetry group, jet space, differential invariant.*

1. INTRODUCTION

The theory of differential equations dates back to the work of Isaac Newton, whose celebrated anagram

"6accdae13eff7i3l9n4o4qrr4s8t12ux"

can be solved as

"Data aequatione quotcunque fluentes quantitates involvente
fluxiones invenire et vice versa,"

which means

"it is useful to solve differential equations."

However, it had shortly turned out that far from all of differential equations can be "solved," i.e., have a solution which can be written in terms of integrals of elementary functions. In particular, the solution of Liouville's equation $y' = y^2 - x$ cannot be written in this form. Therefore, it was required to find different approaches to the study of differential equations. One of the possible approaches was proposed by Sophus Lie, who became the founder of the geometric theory of differential equations and first applied geometric ideas and constructions to study differential equations. In the 1970s, due to the effort of V. V. Lychagin, I. S. Krasil'shchik, A. M. Vinogradov, and other mathematicians, these ideas transformed into a unified harmonious theory with incredibly deep results and wide possibilities for applications. In particular, the theory of differential invariants, which is a differential-geometric analog of the classical algebraic invariant theory, has been developed. The theory of differential invariants has made it possible to solve previously unapproachable problems related to differential equations. For example, in the framework of this theory, Kruglikov [1] solved the problem of point classification of second-order ordinary differential equations, which was being tackled by many outstanding mathematicians during the whole twentieth century (see, e.g., [2]–[6]), Dubrov constructed a series of relative contact differential invariants (see [7], [8]), which has made it possible to solve the so-called trivialization problem for ordinary differential equations (i.e., the problem of reducing such equations to the form

*E-mail: tsdtp4u@proc.ru

$y^{(n)} = 0$ by a change of variables), and Kushner [9] completed the classification of the Monge–Ampère equations, which include virtually all equations of mathematical physics (see also [10]). A detailed survey of results of this theory can be found in [11] and [12].

In the recent paper [13], the author solved Lie's problem of point classification of ODEs of the form $y'' = F(x, y)$ (see also [14] and [15]). Therefore, it is natural to consider a larger class of second-order ODEs. Note that the largest class of equations not fitting into Kruglikov's classification [1] has the form

$$y'' = a_3(x, y)(y')^3 + a_2(x, y)(y')^2 + a_1(x, y)(y') + a_0(x, y).$$

Such equations are closely related to projective geometry (see [16], [17]) and have been extensively studied. A final classification of generic equations of this form was obtained by Yumaguzhin in [18]. However, equations of the form $y'' = F(x, y)$ do not fit into Yumaguzhin's classification and, therefore, must be studied separately.

Another class of equations, which do not fit Yumaguzhin's classification either, is formed by equations of the form $y'' = A(x, y)y' + B(x, y)$. This paper is devoted to a point classification of equations of this type. We calculate a group of point symmetries preserving the class of such equations and the field of differential invariants of the action of this group on the coefficients A and B of our equations; finally, using these invariants, we obtain a criterion for the local equivalence of two generic equations of the form $y'' = A(x, y)y' + B(x, y)$ with rational coefficients on the right-hand side.

2. THE FIELD OF DIFFERENTIAL INVARIANTS

In this section, we give computational results related to the problem of point classification of ODEs of the form $y'' = A(x, y)y' + B(x, y)$ needed in what follows; namely, we find a point symmetry group G preserving the class of such equations, obtain the explicit form of its action on the coefficients A and B, determine the transcendence degrees of the fields of differential invariants of fixed order for this action, and, finally, calculate all fields of differential invariants. We begin by recalling the necessary definitions and facts of the geometric theory of differential equations and the theory of differential invariants (see [11] and [12] for more details).

2.1. The Necessary Definitions and Facts

Let $J^2(\mathbb{C})$ denote the space of 2-jets of germs of holomorphic functions $f\colon \mathbb{C} \to \mathbb{C}$ with coordinates (x, y, p_1, p_2). Then to each differential equation of the form

$$y'' = A(x, y)y' + B(x, y)$$

we can assign the hypersurface

$$\mathscr{E} = \{p_2 = A(x, y)p_1 + B(x, y)\} \subset J^2(\mathbb{C})$$

in the space of 2-jets. Thus, by a *differential equation* we mean a hypersurface $\mathscr{E}$ in the 2-jet space $J^2(\mathbb{C})$.

Now, consider the point pseudogroup of germs of holomorphic transformations of the plane (x, y). The action of this pseudogroup is naturally lifted to an action on the 2-jet space $J^2(\mathbb{C})$:

$$x \mapsto X = X(x, y), \qquad y \mapsto Y = Y(x, y), \qquad p_1 \mapsto P_1 = \frac{DY}{DX}, \qquad p_2 \mapsto P_2 = \frac{DP_1}{DX}$$

(here $D = d/dx = \partial_x + p_1\,\partial_y + p_2\,\partial_{p_1} + \cdots$ is the total differentiation operator).

Our immediate goal is to find a subgroup of the point pseudogroup that preserves the class of equations of the form $y'' = A(x, y)y' + B(x, y)$, i.e., speaking the language of geometry, a subgroup G such that, for each equation $\mathscr{E} \subset J^2(\mathbb{C})$ of the form specified above and any element $g \in G$, the equation $g \circ \mathscr{E}$ has the same form.

2.2. The Symmetry Group G

The following assertion is valid.

Proposition 1. *The group G of symmetries of equations of the form $y''=A(x,y)y'+B(x,y)$ consists of the germs of holomorphic transformations of the form*

$$x\mapsto X(x),\qquad y\mapsto Y_1(x)y+Y_2(x).$$

The Lie algebra $\mathfrak{g}$ of G consists of the germs of vector fields of the form

$$X=\xi(x)\,\partial_x+(\eta(x)y+\zeta(x))\,\partial_y.$$

Proof. It is convenient to prove this proposition in the infinitesimal language of vector fields. Namely, a vector field X is an element of the Lie algebra $\mathfrak{g}$ of the symmetry group G if and only if

$$L_X(p_2-(A(x,y)p_1+B(x,y)))|_{p_2=A(x,y)p_1+B(x,y)}=\widetilde{A}(x,y)p_1+\widetilde{B}(x,y)$$

for any functions A and B (the functions $\widetilde{A}$ and $\widetilde{B}$ are uniquely determined by A and B). This condition is a system of differential equations for the unknown components $\alpha(x,y)$ and $\beta(x,y)$ of the vector field $X=\alpha(x,y)\,\partial_x+\beta(x,y)\,\partial_y$. Solving this system by using the `Maple` software, which was developed by I. Anderson, we obtain the form of X specified in the statement of the proposition. After that, it is easy to find the transformations which form the group G itself. □

Thus, we have found the point symmetry group G of the class of differential equations under consideration. The group G and the algebra $\mathfrak{g}$ act on these equations and, thereby, induce an action on the coefficients A and B. In coordinates, the action of the group G on the coefficients A and B can be written as

$$\begin{aligned}
&A\mapsto\frac{Y_1^2}{X'^3}\,A+\frac{2Y_1'X'-Y_1X''}{X'^3},\\
&B\mapsto\frac{Y_1}{X'^2}\,B+AY_1\frac{Y_1'y+Y_2'}{X'^3}+\frac{Y_1''X'-Y_1'X'}{X'^3}\,y+\frac{Y_2''X'-Y_2'X''}{X'^3}.
\end{aligned}\tag{2.1}$$

We denote the corresponding transformation group by $\widehat{G}$. Its Lie algebra $\widehat{\mathfrak{g}}$ consists of vector fields of the form

$$\begin{aligned}
\widehat{X}=\xi(x)\,\partial_x+(\eta(x)y+\zeta(x))\,\partial_y+(-\xi'(x)A+2\eta'(x)-\xi''(x))\,\partial_A\\
+\big(-Ay\eta'(x)+B\eta(x)+y\eta''(x)-2\xi'(x)B-A\zeta'(x)+\zeta''(x)\big)\,\partial_B.
\end{aligned}\tag{2.2}$$

Thus, we have reduced the problem to classifying the orbits of the action of $\widehat{G}$ on the space of pairs (A,B) of rational functions. To solve this problem, we apply the technique of differential invariants.

2.3. The Transcendence Degree of Fields of Differential Invariants

We denote the space of k-jets of pairs of germs of functions $(A,B)\colon\mathbb{C}^2\to\mathbb{C}^2$ by $\mathbf{J}^k$. The coordinates in this space are denoted by $(x,y,a,b,a_{10},a_{01},b_{10},b_{01},\dots)$. The actions (2.1) and (2.2) of the pseudogroup $\widehat{G}$ and the Lie algebra $\widehat{\mathfrak{g}}$ on the space $\mathbf{J}^0$ extend canonically to an action on the k-jet space for any k and to an action on the space $\mathbf{J}^\infty:=\varprojlim\mathbf{J}^k$ of infinite jets. We denote the extensions of vector fields $\widehat{X}\in\widehat{\mathfrak{g}}$ by $\widehat{X}^{(k)}$, the extension of the entire Lie algebra $\widehat{\mathfrak{g}}$ by $\widehat{\mathfrak{g}}^{(k)}$, and the extension of the pseudogroup $\widehat{G}$ by $\widehat{G}^{(k)}$.

We recall the following definition.

Definition 1. (1) A *differential invariant of order* $\leq k$ of the action of $\widehat{G}$ on the space $\mathbf{J}^k$ is a rational function I constant along all vector fields $\widehat{X}^{(k)} \in \widehat{\mathfrak{g}}^{(k)}$, i.e., such that

$$\widehat{X}^{(k)}(I) = 0.$$

(2) An *invariant derivation* is a derivation ∇ with rational coefficients on the space of functions on $\mathbf{J}^\infty$ which commutes with the action of the Lie algebra $\widehat{\mathfrak{g}}^{(\infty)}$:

$$\nabla \circ \widehat{X}^{(\infty)} = \widehat{X}^{(\infty)} \circ \nabla \qquad \text{for all fields} \quad \widehat{X}^{(\infty)} \in \widehat{\mathfrak{g}}^{(\infty)}.$$

Remark 1. According to the Lie–Tresse theorem, the algebra of differential invariants is locally generated by a finite set of differential invariants and invariant derivations. In [19], it was proved that, under certain assumptions, this assertion holds not only locally but also globally. It is also valid for the action of the symmetry group $\widehat{G}$ on the space $\mathbf{J}^\infty$ of infinite jets (see Theorem 2).

Let $\mathscr{I}_k$ be the set of all differential invariants of pure order k. Clearly, the entire field $\mathscr{I}$ of differential invariants is the union of the sets $\mathscr{I}_k$:

$$\mathscr{I} = \bigcup_k \mathscr{I}_k.$$

The first main result of this paper gives the transcendence degree $t_k := \operatorname{tr\,deg} \mathscr{I}_k$ of $\mathscr{I}_k$, i.e., the number of independent differential invariants of pure order k, for each k.

Theorem 1. *The transcendence degrees t_k of the sets $\mathscr{I}_k$ of differential invariants are given in the table.*

Table

Order k of invariants	≤ 2	3	4	5	...	k
Transcendence degree t_k	0	5	7	9	...	$2k-1$

Proof. To calculate the number of independent differential invariants, we employ a construction which has already been used to solve similar problems, namely, the point classification problem for various classes of second-order ODEs [1], [13], [14] and the problem of point classification of smooth functions on the 1-jet space [20].

Consider the natural projections $\pi_{i,i-1}\colon \mathbf{J}^i \to \mathbf{J}^{i-1}$, a sequence of generic jets θ_i each of which is projected on the preceding one (i.e., such that $\pi_{i,i-1}(\theta_i) = \theta_{i-1}$), and the fiber $V_{\theta_{k-1}}$ of the projection $\pi_{k,k-1}$ over the jet θ_{k-1}.

Let $\widehat{\mathfrak{g}}_{\theta_{k-1}} \subset \widehat{\mathfrak{g}}^{(k)}$ be the isotropy subalgebra, which consists of all vector fields $\widehat{X}^{(k)} \in \widehat{\mathfrak{g}}^{(k)}$ vanishing at θ_{k-1}:

$$\widehat{\mathfrak{g}}_{\theta_{k-1}} = \{\widehat{X}^{(k)} \in \widehat{\mathfrak{g}}^{(k)} : \widehat{X}^{(k-1)}_{\theta_{k-1}} = 0\}.$$

The isotropy subalgebra $\widehat{\mathfrak{g}}_{\theta_{k-1}}$ acts on the fiber $V_{\theta_{k-1}}$. The transcendence degree t_k of the field $\mathscr{I}_k$ of differential invariants is equal to the codimension of a generic orbit of this action.

The dimension of the fiber $V_{\theta_{k-1}}$ equals $2(k+1)$. Therefore, to find the codimension of a generic orbit, it suffices to explicitly calculate all isotropy subalgebras. It follows from (2.2) that the vector fields in the isotropy subalgebra $\widehat{\mathfrak{g}}_{\theta_{k-1}}$ depend on the quantities

$$\xi_i := \xi^{(i)}(a), \qquad \eta_j := \eta^{(j)}(a), \qquad \zeta_m := \zeta^{(m)}(a),$$

where $i, m \leq k+3$, $j \leq k+2$, and $a = \pi_{k,0}(\theta_k)$, and have the forms

$$\widehat{\mathfrak{g}}_{\theta_0} = \Big\{ [-\xi_2 a_{0,0} + 2\eta_2 - \xi_3 - 2\xi_1 a_{1,0} - a_{0,1}\zeta_1]\, \partial_{a_{1,0}} - [a_{0,1}(\xi_1 + \eta_0)]\, \partial_{a_{0,1}}$$

$$
\begin{aligned}
&+\left[-\frac{3}{2}\xi_2 b_{0,0}-\frac{3}{2}b_{0,0}\xi_1 a_{0,0}-a_{0,0}^2\zeta_1+a_{0,0}b_{0,0}\eta_0+\zeta_3-a_{1,0}\zeta_1\right.\\
&\qquad\left.+b_{1,0}\eta_0-3\xi_1 b_{1,0}-b_{0,1}\zeta_1\right]\partial_{b_{1,0}}\\
&+\left[-\frac{1}{2}\xi_2 a_{0,0}-\frac{1}{2}\xi_1 a_{0,0}^2+\eta_2-a_{0,1}\zeta_1-2b_{0,1}\xi_1\right]\partial_{b_{0,1}}\Big\},\\
\widehat{\mathfrak{g}}_{\theta_1}=&\Big\{\left[-\xi_1 a_{0,0}^3+2a_{0,0}\xi_1 a_{1,0}-2a_{0,0}a_{0,1}\zeta_1-4a_{0,0}b_{0,1}\xi_1+2\eta_3-\xi_4-3\xi_2 a_{1,0}\right.\\
&\qquad\left.-3a_{0,1}\xi_1 b_{0,0}-2a_{1,1}\zeta_1-3\xi_1 a_{2,0}\right]\partial_{a_{2,0}}\\
&+\left[-\frac{3}{2}\xi_2 a_{0,1}-\frac{1}{2}a_{0,1}\xi_1 a_{0,0}-\xi_1 a_{1,1}-a_{0,2}\zeta_1\right]\partial_{a_{1,1}}+\left[a_{0,2}\xi_1\right]\partial_{a_{0,2}}\\
&+\left[-4b_{0,0}\xi_1 a_{0,0}^2-2b_{0,0}\xi_1 a_{1,0}-3a_{0,0}a_{1,0}\zeta_1-b_{0,0}\xi_2 a_{0,0}-3a_{0,0}\xi_1 b_{1,0}\right.\\
&\qquad-b_{0,0}a_{0,1}\zeta_1-9b_{0,0}b_{0,1}\xi_1-2a_{0,0}b_{0,1}\zeta_1-4\xi_2 b_{1,0}-a_{2,0}\zeta_1\\
&\qquad\left.-2b_{1,1}\zeta_1-5\xi_1 b_{2,0}+\zeta_4-a_{0,0}^3\zeta_1\right]\partial_{b_{2,0}}\\
&+\left[-\frac{1}{2}\xi_2 a_{0,0}^2-\frac{1}{2}\xi_1 a_{0,0}^3-2a_{0,0}a_{0,1}\zeta_1-2a_{0,0}b_{0,1}\xi_1+\eta_3-3a_{0,1}\xi_1 b_{0,0}-2b_{0,1}\xi_2\right.\\
&\qquad\left.-\frac{1}{2}\xi_2 a_{1,0}-\frac{1}{2}a_{0,0}\xi_1 a_{1,0}-a_{1,1}\zeta_1-3\xi_1 b_{1,1}-b_{0,2}\zeta_1\right]\partial_{b_{1,1}}\\
&+\left[-\xi_2 a_{0,1}-a_{0,1}\xi_1 a_{0,0}-a_{0,2}\zeta_1-\xi_1 b_{0,2}\right]\partial_{b_{0,2}}\Big\},\\
\widehat{\mathfrak{g}}_{\theta_{k-1}}=&\left\{\left[2\eta_{k+1}-\xi_{k+2}\right]\partial_{a_{k,0}}+\left[\zeta_{k+2}\right]\partial_{b_{k,0}}+\left[\eta_{k+1}\right]\partial_{b_{k-1,1}}\right\}
\end{aligned}
$$

(here $k\geq 3$; for convenience, the coefficients multiplying vectors are enclosed in brackets).

Thus, the codimension of the orbits of the action of the isotropy algebra $\widehat{\mathfrak{g}}_{\theta_{k-1}}$ equals 0 for $k\leq 2$, and $2k-1$ for $k\geq 3$, as required. □

Remark 2. The proof of Theorem 1 also provides a description of singular k-jets (i.e., those k-jets for which the dimension of the orbits of the action of the isotropy subalgebras $\widehat{\mathfrak{g}}_{\theta_{k-1}}$ is not maximal). Namely, singular jets are those projected on one of the sets

$$\{a_{0,1}=0\},\qquad \{a_{0,2}=0\}.$$

The right-hand sides of the ODEs $y''=A(x,y)y'+B(x,y)$ in these singular cases are

$$A(x,y)=C_0(x)\qquad\text{and}\qquad A(x,y)=C_0(x)y+C_1(x).$$

2.4. The Structure of the Field of Differential Invariants

Now we are ready to describe the field $\mathscr{I}=\bigcup_k \mathscr{I}_k$ of differential invariants.

Theorem 2. 1. *The field $\mathscr{I}$ of differential invariants of the action of the group $\widehat{G}$ on the space $\mathbf{J}^\infty$ of infinite jets is generated by the five differential invariants*

$$
\begin{aligned}
J_1&=a_{0,1}a_{0,3}/a_{0,2}^2,\\
J_2&=\big(-2a_{0,0}a_{0,1}^2a_{0,3}+(3a_{0,0}a_{0,2}^2+a_{0,2}a_{1,2}-a_{0,3}(3b_{0,2}-2a_{1,1}))a_{0,1}\\
&\qquad+4a_{0,2}^2(b_{0,2}-a_{1,1})\big)/a_{0,2}a_{0,1}^3,\\
J_3&=\big(-4a_{0,0}a_{0,1}^4a_{0,2}+((4a_{1,1}-6b_{0,2})a_{0,2}+4a_{0,0}^2a_{0,3})a_{0,1}^3\\
&\qquad+((-6a_{0,0}^2+3b_{0,1}-a_{1,0})a_{0,2}^2-4a_{0,0}a_{0,2}a_{1,2}+4a_{0,0}a_{0,3}(3b_{0,2}-2a_{1,1}))a_{0,1}^2\\
&\qquad+\big(a_{0,2}^3b_{0,0}+(a_{2,1}+17a_{1,1}a_{0,0}-18a_{0,0}b_{0,2})a_{0,2}^2-2a_{1,2}(3b_{0,2}-2a_{1,1})a_{0,2}
\end{aligned}
$$

$$
\begin{aligned}
&\qquad + a_{0,3}(3b_{0,2} - 2a_{1,1})^2\big)a_{0,1} - 12a_{0,2}^2(b_{0,2} - a_{1,1})^2\big)/a_{0,1}^6,\\
J_4 = {}&\big(-2a_{0,0}a_{0,1}^2a_{0,3} + (3a_{0,0}a_{0,2}^2 + b_{0,3}a_{0,2} - a_{0,3}(3b_{0,2} - 2a_{1,1}))a_{0,1}\\
&\qquad + 3a_{0,2}^2(b_{0,2} - a_{1,1})\big)/a_{0,1}^3a_{0,2},\\
J_5 = {}&\big(-4a_{0,0}a_{0,1}^4a_{0,2} + ((-6b_{0,2} + 4a_{1,1})a_{0,2} + 4a_{0,0}^2a_{0,3})a_{0,1}^3\\
&\qquad + ((-6a_{0,0}^2 + 2b_{0,1})a_{0,2}^2 - 2a_{0,0}a_{0,2}(b_{0,3} + a_{1,2} + 4a_{0,0}a_{0,3}(3b_{0,2} - 2a_{1,1})))a_{0,1}^2\\
&\qquad + \big(a_{0,2}^3b_{0,0} + (-15a_{0,0}b_{0,2} + 14a_{0,0}a_{1,1} + b_{1,2})a_{0,2}^2\\
&\qquad\qquad - a_{0,2}(b_{0,3} + a_{1,2})(3b_{0,2} - 2a_{1,1}) + a_{0,3}(3b_{0,2} - 2a_{1,1})^2\big)a_{0,1}\\
&\qquad - 8a_{0,2}^2(b_{0,2} - a_{1,1})^2\big)/a_{0,1}^6
\end{aligned}
$$

of order 3 *and the two invariant derivations*

$$
\nabla_1 = \frac{a_{0,1}}{a_{0,2}}\cdot\frac{d}{dy} \qquad \textit{and} \qquad \nabla_2 = \frac{1}{a_{0,1}^2}\biggl(a_{0,2}\cdot\frac{d}{dx} - (2a_{0,0}a_{0,1} - 2a_{1,1} + 3b_{0,2})\cdot\frac{d}{dy}\biggr).
$$

2. *The sets $\mathscr{I}_k$ of differential invariants of pure order k are generated by the differential invariants*

$$
J_1^{(p,q)}, \quad J_4^{(p,q)}, \quad J_2^{(0,k-3)}, \quad J_3^{(0,k-3)}, \quad J_5^{(0,k-3)}, \qquad \textit{where} \quad p+q = k-3
$$

(*here $J_i^{(p,q)} = \nabla_1^p\nabla_2^qJ_i$*) *and separate the $\widehat{G}$-orbits of nonsingular k-jets.*

Proof. First, the invariance of the functions $J_1,\dots,J_5$ and the derivations ∇_1 and ∇_2 is verified by straightforward calculations with `Maple`.

Next, we prove assertion 2 of the theorem, which, in turn, implies assertion 1. To this end, it suffices to calculate the symbols of the invariants specified in assertion 2. Let $\alpha_1^i\alpha_2^j$ denote the symbol of the function $a_{i,j}$, and let $\beta_1^i\beta_2^j$ denote the symbol of $b_{i,j}$. Then, up to constants, the symbols of our invariants are

$$
\begin{aligned}
\sigma(J_1^{(p,q)}) &= \alpha_2^k + \alpha_1\alpha_2^{k-1} + \dots + \alpha_1^q\alpha_2^{p+3},\\
\sigma(J_4^{(p,q)}) &= \alpha_2^k + \alpha_1\alpha_2^{k-1} + \dots + \alpha_1^q\alpha_2^{p+3} + \beta_2^k + \beta_1\beta_2^{k-1} + \dots + \beta_1^q\beta_2^{p+3},\\
\sigma(J_2^{(0,k-3)}) &= \alpha_2^k + \alpha_1\alpha_2^{k-1} + \dots + \alpha_1^{k-2}\alpha_2^2,\\
\sigma(J_3^{(0,k-3)}) &= \alpha_2^k + \alpha_1\alpha_2^{k-1} + \dots + \alpha_1^{k-2}\alpha_2^2 + \alpha_1^{k-1}\alpha_2,\\
\sigma(J_5^{(0,k-3)}) &= \alpha_2^k + \alpha_1\alpha_2^{k-1} + \dots + \alpha_1^{k-2}\alpha_2^2 + \beta_2^k + \beta_1\beta_2^{k-1} + \dots + \beta_1^{k-2}\beta_2^2.
\end{aligned}
$$

It is seen that all these symbols are linearly independent; therefore, the invariants themselves are (functionally) independent. Moreover, all these invariants are affine along the fibers of the projection π. According to Rosenlicht's theorem (see [21], [22]), the invariants

$$
J_1^{(p,q)}, \quad J_4^{(p,q)}, \quad J_2^{(0,i)}, \quad J_3^{(0,i)}, \quad J_5^{(0,i)}, \qquad \text{where} \quad i, p+q \le k-3,
$$

generate the entire field $\bigcup_{j\le k}\mathscr{I}_k$ of rational differential invariants of order $\le k$.

Therefore, the differential invariants $J_1,\dots,J_5$ and the invariant derivations ∇_1 and ∇_2 generate the entire field $\mathscr{I} = \bigcup_k\mathscr{I}_k$ of differential invariants, as required. □

3. CLASSIFICATION THEOREM

We proceed to the classification of differential equations of the form

$$
y'' = A(x,y)y' + B(x,y)
$$

with respect to the action of the symmetry group G, or, equivalently, the classification of pairs of functions (A,B) with respect to the action of the group $\widehat{G}$. We begin with the following definition.

Definition 2. We say that a pair of functions (A, B) is *regular* if both functions are rational in the variables x and y, the restrictions of the invariants $J_1, \dots, J_5$ to this pair are defined, and the restrictions of the invariant derivations ∇_1 and ∇_2 are linearly independent in a Zariski open subset of the plane $\mathbb{C}^2$.

Consider a regular pair of functions (A, B) and the set of basis differential invariants of orders 3 and 4. The restrictions of these invariants to our pair of functions form a set of rational functions of x and y in the plane $\mathbb{C}^2$. This set determines the rational morphism

$$\pi_{(A,B)}\colon \mathbb{C}^2 \to \mathbb{C}^{12}, \qquad \pi_{(A,B)}(a) = (J_1([A,B]^4_a), J_2([A,B])^4_a, \dots).$$

Let $\mathscr{C}_{(A,B)}$ denote the closure of the image of $\pi_{(A,B)}$ in the Zariski topology. Then $\mathscr{C}_{(A,B)}$ is an algebraic variety. Its zero ideal $\mathscr{D}_{(A,B)}$ is the ideal of polynomial dependences between the rational functions

$$J_1([A,B]^4_a), \quad J_2([A,B]^4_a), \quad \dots .$$

The following theorem is valid.

Theorem 3. 1. *Two regular pairs of functions (A, B) and $(\widetilde{A}, \widetilde{B})$ are $\widehat{G}$-equivalent if and only if $\mathscr{C}_{(A,B)} = \mathscr{C}_{(\widetilde{A},\widetilde{B})}$.*

2. *Two regular pairs of functions (A, B) and $(\widetilde{A}, \widetilde{B})$ are $\widehat{G}$-equivalent if and only if $\mathscr{D}_{(A,B)} = \mathscr{D}_{(\widetilde{A},\widetilde{B})}$.*

Remark 3. We emphasize that the equivalence criterion given by this theorem is *effective*, i.e., can be verified in finite time on a computer. Namely, the ideal $\mathscr{D}_{(A,B)}$ can be calculated by using the apparatus of Gröbner bases (see, e.g., [23]).

Proof of Theorem 3. First, note that an affine algebraic variety is uniquely determined by its zero ideal (see, e.g., [21]); therefore, the conditions

$$\mathscr{C}_{(A,B)} = \mathscr{C}_{(\widetilde{A},\widetilde{B})} \qquad \text{and} \qquad \mathscr{D}_{(A,B)} = \mathscr{D}_{(\widetilde{A},\widetilde{B})}$$

are equivalent. We use these conditions interchangeably, depending on the situation.

Obviously, the equivalence of pairs of functions implies the coincidence of the corresponding varieties $\mathscr{C}$ and ideals $\mathscr{D}$. Let us prove the converse. We set

$$\mathscr{C}_{(A,B)} = \mathscr{C}_{(\widetilde{A},\widetilde{B})} =: \mathscr{C} \qquad \text{and} \qquad \mathscr{D}_{(A,B)} = \mathscr{D}_{(\widetilde{A},\widetilde{B})} =: \mathscr{D}.$$

The condition $\mathscr{C}_{(A,B)} = \mathscr{C}_{(\widetilde{A},\widetilde{B})}$ implies that, for any generic point $a_1 = (x_0, y_0) \in \mathbb{C}^2$, there exists a point $a_2 \in \mathbb{C}^2$ for which

$$\pi_{(A,B)}(a_1) = \pi_{(\widetilde{A},\widetilde{B})}(a_2).$$

This means that the values of all basis differential invariants of order at most 4 at the 4-jets $[A,B]^4_{a_1}$ and $[\widetilde{A},\widetilde{B}]^4_{a_2}$ coincide. According to Theorem 2, these jets are $\widehat{G}^{(4)}$-equivalent, i.e., there exists an element $g^4_{(a_1,a_2)} \in \widehat{G}^{(4)}_{(a_1,a_2)}$ for which

$$g^4_{(a_1,a_2)} \circ [A,B]^4_{a_1} = [\widetilde{A},\widetilde{B}]^4_{a_2}$$

(here $\widehat{G}^{(4)}_{(a_1,a_2)} \subset \widehat{G}^{(4)}$ is the subgroup of $\widehat{G}^{(4)}$ consisting of the 4-extensions of those transformations in $\widehat{G}$ which take a_1 to a_2).

Differentiating the relations in the ideal $\mathscr{D}$ by using the invariant derivations ∇_1 and ∇_2, we obtain relations in which the basis differential invariants of order 5 occur linearly. Hence the values of their restrictions to the pair (A, B) are uniquely determined by the restrictions to (A, B) of the basis invariants of order at most 4. Therefore, the values of basis invariants of order 5 at the 5-jets $[A,B]^4_{a_1}$

and $[\widetilde{A},\widetilde{B}]^5_{a_2}$ coincide. According to Theorem 2, these jets are $\widehat{G}^{(5)}$-equivalent, i.e., there exists an element $g^5_{(a_1,a_2)}\in\widehat{G}^{(5)}_{(a_1,a_2)}$ for which

$$g^5_{(a_1,a_2)}\circ[A,B]^5_{a_1}=[\widetilde{A},\widetilde{B}]^5_{a_2}.$$

A completely similar argument proves that, for any $k\geq 5$, the k-jets $[A,B]^k_{a_1}$ and $[\widetilde{A},\widetilde{B}]^k_{a_2}$ are $\widehat{G}^{(k)}$-equivalent, i.e., there exists an element $g^k_{(a_1,a_2)}\in\widehat{G}^{(k)}_{(a_1,a_2)}$ for which

$$g^k_{(a_1,a_2)}\circ[A,B]^k_{a_1}=[\widetilde{A},\widetilde{B}]^k_{a_2}.$$

Now consider the infinite jet

$$g^\infty_{(a_1,a_2)}=\{g^k_{(a_1,a_2)}\}\in G^{(\infty)}_{(a_1,a_2)}.$$

We have $g^\infty_{(a_1,a_2)}\circ[A,B]^\infty_{a_1}=[\widetilde{A},\widetilde{B}]^\infty_{a_2}$, i.e., the pairs (A,B) and $(\widetilde{A},\widetilde{B})$ are equivalent at the formal level. Our objective is to prove their $\widehat{G}$-equivalence, i.e., the existence of an element $g\in\widehat{G}$ for which

$$g\circ(A,B)=(\widetilde{A},\widetilde{B}).$$

We use an idea of Lychagin that appears in [24] and which has already worked for similar problems (see, e.g., [13]–[15]). Namely, consider the pair of functions

$$g\circ(A,B)-(\widetilde{A},\widetilde{B})=(F_1,F_2)=(0,0).$$

We obtain the system of ordinary differential equations

$$F_1|_{y=y_0}=F_2|_{y=y_0}=(F_1)_y|_{y=y_0}=(F_2)_y|_{y=y_0}=0$$

with unknown functions X, Y_1, and Y_2 included in g (see (2.1)). This system is consistent, because the 1-jets of the pairs $[A,B]^1_{a_1}$ and $[\widetilde{A},\widetilde{B}]^1_{a_2}$ are $\widehat{G}^{(1)}$-equivalent. Therefore, the above system of differential equations has a solution, which determines an element $g\in\widehat{G}$ for which

$$[g]^1\circ[A,B]^1_{a_1}=[\widetilde{A},\widetilde{B}]^1_{a_2},$$

i.e.,

$$F_i|_{y=y_0}=(F_i)_y|_{y=y_0}=0.$$

Since the infinite jets $[A,B]^\infty_{a_1}$ and $[\widetilde{A},\widetilde{B}]^\infty_{a_2}$ are $\widehat{G}^{(\infty)}$-equivalent, it follows that

$$[g]^\infty\circ[A,B]^\infty_{a_1}=[\widetilde{A},\widetilde{B}]^\infty_{a_2}.$$

Therefore, all partial derivatives of the functions F_1 and F_2 vanish at $y=y_0$. It remains to note that these functions are analytic in y; hence $F_1=F_2=0$, which means that $g\circ(A,B)=(\widetilde{A},\widetilde{B})$, as required. □

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research under grant 16-31-60018-mol_a_dk.

REFERENCES

1. B. S. Kruglikov, "Point classification of second order ODEs: Tresse classification revisited and beyond," in *Differential Equations*: *Geometry, Symmetries and Integrability* (Springer, Berlin, 2009).
2. A. Tresse, *Détermination des invariants ponctuels de l'équation différentielle ordinaire du second ordre* $y'' = \omega(x, y, y')$ (Leipzig, 1896).
3. R. A. Sharipov, *Effective Procedure of Point Classification for the Equation* $y'' = P + 3Qy' + 3Ry'^2 + Sy'^3$, arXiv: 9802027 (1998).
4. S. Lie, "Klassification und Integration von gewöhnlichen Differentialgleichungen zwischen x, y, die eine Gruppe von Transformationen gestatten. III," Lie Arch. **8**, 371–458 (1883).
5. R. Liouville, "Sur les invariants de certaines eǫuations differentielles et sur leurs applications," J. de l'Ećole Polytech. **59**, 7–76 (1889).
6. S. Yu. Slavyanov, "Polynomial degree reduction of a Fuchsian 2×2 system," Teoret. Mat. Fiz. **182** (2), 223–230 (2015) [Theoret. and Math. Phys. **182** (2), 182–188 (2015)].
7. B. Dubrov, "Contact trivialization of ordinary differential equations," in *Differential Geometry and Its Applications* (Silesian Univ. Opava, Opava, 2001), pp. 73–84.
8. B. Dubrov, "On a class of contact invariants of systems of ordinary differential equations," Izv. Vyssh. Uchebn. Zaved. Mat., No. 1, 76–77 (2006) [Russian Math. (Iz. VUZ) **50** (1), 73–74 (2006)].
9. A. Kushner, *Classifications of Monge–Ampère Equations*, Doctoral Dissertation in Physics and Mathematics (Kazan State Univ., Kazan, 2011).
10. A. Kushner, V. Lychagin, and V. Rubtsov, *Contact Geometry and Non-Linear Differential Equations* (Cambridge Univ. Press, Cambridge, 2007).
11. D. V. Alekseevskii, A. M. Vinogradov, and V. V. Lychagin, *Basic Ideas and Concepts of Differential Geometry*, in *Current Problems in Mathematics*: *Fundamental Directions*, Vol. 28, Itogi Nauki i Tekhniki (VINITI, Moscow, 1988), pp. 5–289 [in Russian].
12. A. M. Vinogradov, I. S. Krasil'shchik, and V. V. Lychagin, *Introduction to the Geometry of Nonlinear Differential Equations* (Nauka, Moscow, 1986) [in Russian].
13. P. V. Bibikov, "On Lie's problem and differential invariants of ODEs $y'' = F(x, y)$," Funktsional. Anal. Prilozhen. **51** (4), 16–25 (2017) [Functional Anal. Appl. **51** (4), 255–262 (2017)].
14. P. Bibikov and A. Malakhov, "On Lie problem and differential invariants for the subgroup of the plane Cremona group," J. Geom. Phys. **121**, 72–82 (2017).
15. P. Bibikov, "Generalized Lie problem and differential invariants for the third order ODEs," Lobachevskii J. Math. **38** (4), 622–629 (2017).
16. V. I. Arnold, *Geometric Methods in the Theory of Ordinary Differential Equations* (Regulyarnaya i Khaoticheskaya Dinamika, Izhevsk, 2000) [in Russian].
17. E. Cartan, "Sur les variétés à connexion projective," Bull. Soc. Math. France **52**, 205–241 (1924).
18. V. A. Yumaguzhin, *Differential Invariants of* 2*-Order ODEs, I*, arXiv: 0804.0674.
19. B. Kriglikov and V. Lychagin, "Global Lie–Tresse theorem," Selecta Math. (N. S.) **22** (3), 1357–1411 (2016).
20. P. V. Bibikov, "Point equivalence of functions on the 1-jet space $J^1\mathbb{R}$," Funktsional. Anal. Prilozhen. **48** (4), 19–25 (2014) [Functional Anal. Appl. **48** (4), 250–255 (2014)].
21. É. Vinberg and V. Popov, *The Theory of Invariants* (VINITI, Moscow, 1989) [in Russian].
22. M. Rosenlicht, "A remark on quotient spaces," An. Acad. Brasil. Ci. **35**, 487–489 (1963).
23. *Computer Algebra: Symbolic and Algebraic Computation*, 2nd ed., Ed. by B. Buchberger, G. E. Collins, and R. Loos in cooperation with R. Albrecht (Springer-Verlag, Vienna, 1983; Mir, Moscow, 1986).
24. V. V. Lychagin, "Feedback differential invariants," Acta Appl. Math. **109** (1), 211–222 (2010).