ABSTRACT. Ultrasound imaging is a widely used tool for visualizing human body’s internal organs and quantifying clinical parameters. Due to its advantages such as safety, non-invasiveness, portability, low cost and real-time 2D/3D imaging, diagnostic ultrasound industry has steadily grown. Since the technology advancements such as digital beam-forming, Doppler ultrasound, real-time 3D imaging and automated diagnosis techniques, there are still a lot of demands for image quality improvement, faster and accurate imaging, 3D color Doppler imaging and advanced functional imaging modes. In order to satisfy those demands, mathematics should be used properly and effectively in ultrasound imaging. Mathematics has been used commonly as mathematical modelling, numerical solutions and visualization, combined with science and engineering. In this article, we describe a brief history of ultrasound imaging, its basic principle, its applications in obstetrics/gynecology, cardiology and radiology, domestic-industrial products, contributions of mathematics and challenging issues in ultrasound imaging.

1. INTRODUCTION

Ultrasound imaging system visualizes organs inside human body using sound waves with the frequency range of $1\sim15$ MHz higher than human audible frequency. Due to the various advantages of non-invasiveness, safety, portability (see Fig. 1), relatively low cost and real-time imaging over other imaging modalities such as CT and MRI, it is widely used in various diagnostic fields of obstetrics and gynecology, cardiology, radiology, and so on.

Recently, it was reported that ultrasound market size was growing and the ultrasound market including diagnostic ultrasound would be worth about 6.23 billion USD in 2020 [1, 2].
order to have bigger market share, each global company tries to develop competitive imaging technologies and to expand ultrasound use in other applications. For example, one of the most promising technologies affected by increased computational power is real-time 3D ultrasound imaging. It is capable of providing good 3D visualization of organs at the faster rate of about 30 volumes per seconds compared to volumetric imaging modalities such as CT and MRI. However, its resolution is not good enough so that it often fails to discriminate features smaller than a few millimeters. Because of this resolution limitation, real-time 3D ultrasound imaging is not frequently adopted for clinical examinations in radiology. Whereas, it is meaningfully used for observing cardiac wall motion in cardiology or for distinguishing between tissues and surrounding fluids in obstetrics and gynecology. In order to circumvent the resolution problem, each company concentrates on fusion imaging technique that incorporates two images acquired from two different imaging modalities: real-time ultrasound images and CT or MRI images acquired earlier. It is able to provide more anatomical information indistinguishable with the resolution of real-time 3D ultrasound imaging and expand ultrasound use in radiology. As other examples of the promising technologies, we can consider image quality enhancement, faster imaging, 3D color Doppler imaging, computer-aided detection (CADe), computer-aided diagnosis (CADx) and advanced functional imaging modes such as elasticity imaging and vortex flow imaging. They are also very interesting and challenging issues to be improved in ultrasound imaging.

Then how can we advance and innovate ultrasound imaging technologies? Like in other industries, *Mathematics* has played an important role in ultrasound imaging and medical diagnosis already. For acquiring gray-level echo images and color Doppler images, *Mathematics* is used in various parts of time-delay computation for beam-forming, Fourier transform of ultrasound signal for signal processing or filter design, auto-correlation computation for mean frequency estimation, linear interpolation for scan conversion, and so on [4, 5]. *Mathematics*
is also used to extract functional information from ultrasound images. In order to avoid labor-intensive and time-consuming manual measurements on gray-scale echo images, automated segmentation and motion tracking algorithms are demanded. The automated measurement methods are based on partial differential equation, numerical analysis, statistics as well as the interpretation of acoustic fields and speckles as inherent appearance in ultrasound imaging. Therefore, we can see that various levels of mathematical modelling and methods are applied to acquire ultrasound images and to generate functional information for clinical evaluations.

Moreover, Mathematics is essential in cases of elasticity imaging and vortex flow imaging mentioned above. Elasticity imaging is an imaging mode to map the stiffness of soft tissues such as liver or breast, and it can be implemented by reconstructing elastic modulus from measured displacements governed by the elastic equation:

\[
\nabla \cdot (\mu (\nabla u + \nabla u^t)) + \nabla (\lambda \nabla \cdot u) = \rho \frac{\partial^2}{\partial t^2} u, \tag{1.1}
\]

where \( u \) denotes the displacement vector, \( \nabla u^t \) the transpose of the matrix \( \nabla u \), \( \rho \) the density of the soft tissue as elastic material, \( \mu \) the shear modulus with \( \mu = \frac{E}{2(1+\nu)} \) and \( \lambda \) the Lamé coefficient with \( \lambda = \frac{\nu E}{(1+\nu)(1-2\nu)} \). Here, \( E \) and \( \nu \) are Young’s modulus and Poisson’s modulus, respectively.

Vortex flow imaging is an imaging mode to quantify the vorticity of intra-ventricular blood flows and to offer possible clinical indices. Since the vorticity fields are obtained by taking the curl operator to velocity fields of flows, the velocity fields should be exactly computed. The velocity fields can be computed by solving the Navier-Stokes equation:

\[
\begin{align*}
\rho \left( \frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v} \right) &= -\nabla p + \mu \nabla^2 \mathbf{v} \quad \text{in } \Omega_t, \\
\nabla \cdot \mathbf{v} &= 0 \quad \text{in } \Omega_t,
\end{align*}
\tag{1.2}
\]

where \( \Omega_t \) denotes the intra-ventricular domain varying within the heartbeat cycle \( T \), \( \mathbf{v} \) the velocity fields, \( \rho \) the blood density and \( \mu \) the blood viscosity.

The displacements of soft tissues and the motion of blood flows are understood through physics-based mathematical modelling using elastic and fluid equations. Under those mathematical models, we deal with measurements obtained through ultrasound imaging systems, construct the overall phenomenon from the partially measured data and quantify clinically meaningful information. The point is that ultrasound imaging technologies will be advanced and innovated through Mathematics.

The aim of this paper is to describe the importance and value of Mathematics in ultrasound imaging and to share them with other scientists and engineers. In this paper, we describe a brief history of ultrasound imaging, basic imaging principle, its applications in major diagnostic fields, domestic-industrial products and contributions of mathematics, especially for beginners to understand ultrasound imaging and pay much attention to it. Furthermore, we introduce hot topics and challenging issues related to ultrasound imaging. They have to be resolved for the advancements of ultrasound imaging technology. We hope many mathematicians recognize how mathematics is used and contributes much to ultrasound technology innovation.
2. THE HISTORY OF DIAGNOSTIC ULTRASOUND IMAGING

The history of ultrasound imaging starts with the history of sonar, the abbreviation of SOund Navigation And Ranging, measuring the depth of water using sound wave. In the late 1800s, the theoretic and practical foundation for sonar and ultrasound imaging was established by a mathematical equation describing sound wave and the discovery of piezo-electricity, electricity resulting from pressure. L. Rayleigh described sound wave as a mathematical equation in his paper “the Theory of Sound” published in 1877 and P. Curie and his brother J. Curie discovered the piezo-electric effect in certain crystals in 1880 [6–8].

Since the sinking of the Titanic in 1912 and World War I, echo-ranging systems were demanded to detect icebergs or submarines. In 1917, P. Langevin, one of the Curie brothers’ first students developed practical underwater echo-ranging systems using the piezo-electricity for transducer, a device converting a voltage difference to a mechanical stress and reversely converting a mechanical pressure to an electric potential. During World War II, echo-ranging technique was applied to electromagnetic waves and became radar, the abbreviation of RAdio Detection And Ranging.

Echo-ranging system as a diagnostic device to probe the human body was not developed until the 1940’s. K. Dussik used ultrasound to diagnose brain tumor by transmitting an ultrasound beam through the human skull in 1942 [6, 7, 9]. He is regarded as the first-time user of ultrasound for medical diagnosis. In 1948, G. Ludwig used A-mode ultrasound system to detect gallstones [7, 10]. The A-mode shows a signal profile representing the instantaneous echo signal amplitude over time after transmission of the acoustic pulse.

![Transducer and Echo](image)

**Figure 2.** Illustration of the A-mode and B-mode displays. A 2D B-mode image consists of multiple and sequential B-mode lines.

After that time, innovative advances of ultrasound imaging technologies have followed. In the 1950s, 2D B-mode ultrasound imaging system was developed and applied to detect breast tumors and to diagnose in the obstetrics and gynecology fields [6, 11]. The B-mode represents the brightness converted from the amplitude of the echo signals. Fig. 2 illustrates the A-mode and B-mode displays.
Since the first implementation of ultrasonic Doppler techniques in Japan in 1955 [12], various Doppler techniques including the continuous wave Doppler, spectral wave Doppler, and color Doppler ultrasound were developed through the 1960s and the 1970s [7, 13–17]. The emergency of 3D ultrasound technology in the 1980s was capable of capturing 3D images of a fetus [18, 19]. Thanks to the continuous improvements of image quality and 3D imaging capabilities, ultrasound technology became more sophisticated and capable of real-time 3D imaging in the 1990s [20].

From the 2000s to present, 2D arrays of transducers were used for real-time 3D imaging, ultrasound imaging system became miniaturized gradually and ultrasound imaging technology is developing steadily for improving image quality, reducing costs, reducing exam times and being applied to wider diagnostic areas. Recently, hand-carried systems became available. It is becoming common with significant growth and can be used for clinical assessments even in developing countries [21, 22] due to the portability and low cost in addition to higher image quality obtained by using special ultrasound probes for transesophageal echocardiography and transvaginal echo [23, 24].

More detailed information about the history of ultrasound imaging can be found at the references [4, 6].

3. Basic Principle of Ultrasound Imaging

In this section, we describe the basic principle of ultrasound imaging, especially focusing on gray-scale echo images, called the B-mode images. Fig. 3 shows intuitively that mathematics, combined with science and engineering, is used in the process of ultrasound B-mode imaging. Furthermore, we deal with the mathematical models on acoustic fields, fundamental in ultrasound imaging, and describe their applications. We can see that mathematics is applied usually as mathematical modelling, numerical solutions and visualization.
3.1. **Ultrasound B-mode imaging.** An ultrasound image consists of several sequential scanlines (around \(128 \sim 512\) scanlines), each of which is usually acquired through transmission and reception beam-forming using transducer arrays. The transmission beam-forming is to control electronically time delays for ultrasound beam emitted from the transducer arrays to be focused at a point on each scanline. The reception beam-forming is to control electronically time delays to accumulate back-scattered echo signals from a specific position, called the received focal point, on scanlines. We note that the received focal point can be changed dynamically on each scanline.

As sound travels through a material, the amplitude or intensity is decreased due to three factors: absorption, scattering and beam divergence. We start describing mathematically the ultrasound imaging principle, together with beam-forming, considering the time gain compensation of the received and beam-formed echo signals [25, 26].

3.1.1. **Beam-forming and gain compensation.** Let \(N\) be the number of transducers, \(r_j\) the position of the \(j\)-th transducer and \(p_j\) the received signal at \(r_j\) for \(j = 1, \cdots, N\). We assume that transducers are very small, sample positions on each scanline are far away from the transducers and the impulse response is given in the form of Dirac delta function. Then, for sample position \(r = (x, y, z)\) on a given scanline, the amplitude of the received signal at time \(t\) is expressed as

\[
p(t) = \sum_{j=1}^{N} p_j(t - \tau_j) = K \sum_{j=1}^{N} \delta \left( t - \frac{|r_j - r|}{c} \right), \tag{3.1}
\]

where \(\tau_j\) is the delayed time and \(K\) a constant depending on depth.

3.1.2. **Quadrature demodulation and envelope detection.** Let \(f_0\) be the center frequency of transmission signals. Then the amplitude (3.2) of the beam-formed signal along the given scanline can be expressed as

\[
p(t) = a(t) \cos(2\pi f_0 t + \phi(t)) = \frac{a(t)}{2} \left( e^{i(2\pi f_0 t + \phi(t))} + e^{-i(2\pi f_0 t + \phi(t))} \right), \tag{3.2}
\]

where \(a(t)\) and \(\phi(t)\) are the amplitude and the phase of the received signal, respectively. In order to remove signals with high-frequency and obtain meaningful signal with low-frequency, we perform the quadrature demodulation. By setting \(\hat{p}(t) = p(t)e^{-2\pi f_0 t}\) and applying a low-pass filter to \(\hat{p}(t)\), we can obtain a base-band complex signal \(\frac{a(t)}{2}e^{i\phi(t)}\). Its real and imaginary parts are called In-phase and Quadrature signals, denoted by \(I(t)\) and \(Q(t)\), respectively.

\[
I(t) = \Re \left( \frac{a(t)}{2}e^{i\phi(t)} \right), \tag{3.3}
\]

\[
Q(t) = \Im \left( \frac{a(t)}{2}e^{i\phi(t)} \right).
\]

After the envelope detection by taking \(\sqrt{I(t)^2 + Q(t)^2}\) from the quadrature demodulated signals \(I(t)\) and \(Q(t)\), logarithmic compression follows.
3.1.3. Logarithmic compression. Most of clinically meaningful signals are appeared with the small value in the wide range of amplitude. The logarithmic compression is required to display these meaningful signals in the 256 gray-scale values. The image intensity \( I(t) \) is expressed as

\[
I(t) = \frac{I_{\text{max}} - I_{\text{min}}}{\ln A_{\text{max}} - \ln A_{\text{min}}} \ln \left( \frac{A(t)}{A_{\text{min}}} \right) + I_{\text{min}},
\]

where \( I_{\text{max}} \) and \( I_{\text{min}} \) are the maximum and minimum intensity values, to be displayed finally in the gray-scale, respectively, and \( A_{\text{max}} \) and \( A_{\text{min}} \) the maximum and minimum amplitude values with the range of interest, respectively.

3.1.4. Digital scan conversion and display. These intensity values at the sample positions along each scanline are displayed on a video screen through scan conversion process for the spatially arrangement. Scan conversion is performed by using bi-linear interpolation.

In order to improve the B-mode image, some signal/image processing methods may be used additionally.

3.2. Acoustic fields. Jensen [27] derived an inhomogeneous wave equation describing the propagation and scattering of ultrasound in an inhomogeneous medium, derived a mathematical model for the received pulse-echo pressure fields by considering the surface shape of a single transducer element used in diagnostic ultrasound and showed simulated pressure field compared to a measured field. In this subsection, we describe the mathematical models and their applications.

3.2.1. Mathematical modelling. Let \( p(r, t) \) be the pressure variation caused by ultrasound propagation at the position \( r \) and time \( t \), \( c(r) \) the propagation speed and \( \rho(r) \) the material density. Then the inhomogeneous wave equation is derived as

\[
\nabla^2 p - \frac{1}{c_0^2} \frac{\partial^2 p}{\partial t^2} = -\frac{2(\delta c)}{c_0^3} \frac{\partial^2 p}{\partial t^2} + \frac{1}{\rho_0} \nabla(\delta \rho) \cdot \nabla p,
\]

where \( c_0 \) is the mean propagation speed, \( \rho_0 \) the mean density, \( \delta c \) the speed variation from \( c_0 \) and \( \delta \rho \) the density variation from \( \rho_0 \).

Let \( p_r(t) \) the received signal at the transducer surface and \( E_m(t) \) at time \( t \). Then \( p_r(t) \) is written as

\[
p_r(t) = \frac{\rho_0}{2c_0^2} E_m(t) * \frac{\partial^3 v(t)}{\partial t^3} * \left[ \left( \frac{\delta \rho(r)}{\rho_0} - \frac{2(\delta c(r))}{c_0} \right) \ast H(r, t) \right],
\]

where * and \( \ast \) mean the time convolution and spatial convolution, respectively. Note that \( \frac{\rho_0}{2c_0^2} E_m(t) * \frac{\partial^3 v(t)}{\partial t^3} \) is the pulse-echo including the transducer excitation and the electromagnetic impulse response during the emission/reception of the pulse, \( \frac{\delta \rho(r)}{\rho_0} - \frac{2(\delta c(r))}{c_0} \) is the inhomogeneities in tissue and \( H(r, t) \) is the pulse echo spatial impulse response given by \( H(r, t) = h(r, t) \ast h(r, t) \) with \( h(r, t) = \int_S \frac{\delta(t-|r-r_0|/c_0)}{2\pi |r-r_0|} \, dr_0 \) for the transducer surface \( S \).

For validating the derived mathematical model (3.6), a simulation was performed under the setup that a plane reflector was placed at focal point. According to an excited Dirac impulse,
impulse responses were computed at the distances of three different positions from transducer surface and compared to measured responses.

3.2.2. **Field II simulation program.** Afterward Jensen [28, 29] improved the mathematical model (3.6) for the propagation and scattering of acoustic fields by considering acoustic fields from arbitrarily shaped, apodized and excited transducers and dealing with the imaging principle, and developed a simulation program based on the derived mathematical model and running under MATLAB software. It is called Field II. Using Field II simulation program, we can simulate ultrasound transducer fields and generate some gray-level echo and color Doppler images similar to real ultrasound images acquired in ultrasound imaging system as depicted in Fig. 4. Field II is widely used in several companies, universities and departments because it is very helpful for scientists and engineers to understand ultrasound imaging and study image processing methods such as speckle reduction on the synthetic images. It is available at: http://field-ii.dk/.

**4. APPLICATIONS OF DIAGNOSTIC ULTRASOUND IMAGING**

Thanks to the advanced imaging technologies such as image quality enhancement, real-time 3D imaging and the miniaturization of equipment, ultrasound is widely applied to various diagnostic fields of obstetrics and gynecology (OB/GYN), cardiology, radiology, urology, and so on. We note that applications in OB/GYN, cardiology and radiology account for about 70% of the entire ultrasound examinations. In this section, we provide brief explanations of ultrasound imaging applications in major three diagnostic areas.
4.1. OB/GYN ultrasound. Since the 1950’s, ultrasound has been used for the diagnosis of gestation and women’s health. The ultrasound market for OB/GYN examinations has continued to increase. The main issues of ultrasound imaging in OB/GYN is to improve the image quality for detecting endometriosis and gynecological cancer, diagnosing early miscarriage and ectopic pregnancy, and evaluating fetal growth. Especially the performance enhancement of static and real-time 3D imaging is still a hot issue for easier detection of fetal abnormalities in obstetrics.

4.1.1. Ultrasound examinations. Fetal ultrasound examinations are classified into four parts: first-trimester examination, standard second- or third-trimester examination, limited examination and specialized examinations [30]. Most of the examinations performed by observing or measuring target objects: the presence, size, location and number of gestational sac(s), fetal presentation, amniotic fluid volume, cardiac activity, placental position, fetal biometry, fetal heart activity, and so on.

On the other hand, gynecological examinations such as the detection of endometriosis and gynecological cancer are performed commonly by transvaginal scans, with the probe designed to be placed in the vagina of patient. Since the scanhead of the probe is closely adjacent to uterus, the transvaginal scans provide better image qualities and are useful for the early diagnosis of miscarriage and ectopic pregnancies.

Recent advances of static and real-time 3D ultrasound imaging through various image processing and volume rendering techniques have enabled performing more accurate diagnosis for ectopic pregnancy [31] and providing more accurate diagnostic information: abnormalities and anomalies by visualizing fetal face, brain, spine, and skeleton [32], fetal weight obtained by measuring abdominal, thigh, arm volumes and head circumference [33–35], and the volumes of lung and heart [36, 37]. High quality images and accurate boundary extraction are required for fetal examinations.

In addition, Doppler ultrasound is indispensable for detecting fetal heartbeat as well as blood flow characteristics in the various fetal blood vessels: umbilical artery, aorta, middle cerebral arteries, uterine arcuate arteries and inferior vena cava. Power or color Doppler images are used for examinations of fetal middle cerebral artery and cerebral venous [38, 39]. Blood flow to fetus through umbilical cord is also observed by color Doppler images [40, 41]. Note that color Doppler images reflect one-directional velocity components of blood flow along scanlines and power Doppler shows the volume of blood, rather than its velocity.

4.1.2. Innovative technologies for OB/GYN. Samsung Medison, a leading domestic company, has released recently a premium ultrasound system ‘WS80A with Elite’ with superior ultrasound imaging techniques for OB/GYN examinations: S-Harmonic, ClearVision™, Realistic Vue™, Crystal Vue, etc. Among them, we pay attention to Crystal Vue an advanced rendering technique to enhance the image contrast of 3D volume data. It is capable of imaging the interface of bone and soft tissue with high accuracy and providing intuitive information on the contour of fetal skeleton, face and brain. Lately, Crystal Vue has been featured on the front cover of Ultrasound in Obstetrics & Gynecology (UOG), a renowned journal in the women’s
The images of fetal spine and ribs obtained by Crystal Vue were selected as the Picture of the Month by UOG. They provide remarkably clear and realistic information of fetal spine and ribs.

Together with the advanced rendering technique, ElastoScan™ for gynecology is also being promoted. It is used to diagnose benign gynecological disorders and to differentiate uterine fibroids and adenomyosis. We can appreciate the remarkable images acquired by Samsung Medison’s innovative imaging technologies at: http://www.samsungmedison.com/ultrasound/ob-gyn/ws80a-with-elite/.

4.2. **Cardiological ultrasound.** Due to the high temporal resolution of ultrasound imaging, cardiac ultrasound (echocardiography) has been very successful in providing a quick assessment of the overall health of heart, a very fast moving complex organ positioned deep within the body. Ultrasound examinations of heart functions related to cardiac wall motion, valve motion and blood flow are performed by various ultrasound imaging modes: gray-scale imaging, tissue Doppler imaging, color Doppler imaging and real-time 3D imaging. Note that tissue Doppler images represent the velocity of cardiac wall motion.

4.2.1. **Ultrasound examinations.** Echocardiographic examinations include numerous diagnosing items: valvular heart disease, hypertensive heart disease, ischemic heart disease, cardiac tumors, evaluation of the left and right atrium, evaluation of left ventricle(LV) systolic/diastolic function, congenital heart disease, diseases of the aorta, and so on. For more detailed description on echocardiographic examinations, we refer to the literature ‘The Echocardiographer’s Pocket Reference’ by T. Reynolds(2013) [43].

Among echocardiographic examinations, we consider only evaluation of LV function including dimensions, volumes, wall thickness, LV mass, LV mass index, longitudinal motion, radial motion, circumferential, segmental wall motion, wall motion score index, etc. For its quantitative assessment, wall motion tracking and LV volume quantification at each time is needed. To avoid labor-intensive and time-consuming manual LV tracing process, demand for automated LV tracking and analysis methods has been rapidly growing and there have been plenty of studies on LV tracking methods. In Section 6, we represent a problem on LV motion tracking and solutions by mathematical modelling.

4.2.2. **A software for quantifying cardiac function.** Automated LV tracking methods can be applied to analyze and quantify LV function. TomTec is a leading company in medical imaging software solutions and offers industry partners or clinicians various solutions for quantifying cardiac function based on 2D and 3D echocardiography. As an example of software solutions, TomTec developed AutoStrain® by applying automated LV tracking method to visualize LV function quantitatively. AutoStrain® provides automated LV function analysis based on cardiac apical long axis views. It visualizes longitudinal strain results, obtained by 2D speckle tracking, to be color-coded in the individual clips and combined in a Bull’s-eye plot.

4.2.3. **A software for visualizing and quantifying blood flows.** On the other hand, another imaging software company AMID has recently developed ‘hi-def flow tracking’ a software
for visualizing and quantifying the motion of blood flow (http://www.amid.net). It is also called Omega Flow (Siemens, Mountain View, CA). It is not based on color Doppler images, but echo-PIV technique applied to contrast-enhanced echocardiographic images. Along with studies on its clinical possibility and feasibility, this software continues to develop to be used to evaluate cardiac function clinically.

4.3. Radiological ultrasound. Ultrasound examinations in radiology consist of abdominal, vascular, musculoskeletal, breast, thyroid ultrasonographies, and so on. Most of ultrasound examinations are performed by observing main features or measuring the shape and size of target objects using gray-scale B-mode images. Doppler ultrasound is used typically for evaluating blood flow through blood vessels.

4.3.1. Ultrasound examinations. Abdominal ultrasonography is used for diagnosing the diseases of liver, gallbladder, spleen, pancreas, kidneys and bladder. As an example, liver is scanned to examine focal liver disease, diffuse liver disease and hepatic vasculature.

Vascular ultrasound is adopted typically for observing common/internal carotid artery and diagnosing carotid stenosis and chronic mesenteric ischaemia. According to carotid artery stenosis, gray-scale B-mode images is used for characterizing plaques into four types and Doppler ultrasound is used for evaluating both the macroscopic appearance of plaques and flow characteristics [44].

Musculoskeletal ultrasound is a fast and dynamic way to examine shoulder impingement, shoulder instability and rotator cuff disorders. It also allows evaluation of the joint, tendons and ligaments by imaging elbow anatomy and wrist anatomy. Moreover, it is used for diagnosing hand, hip, knee, ankle/foot, etc.

Breast ultrasound enables evaluating a symptomatic young or pregnant patient, to evaluate a palpable lump with negative or equivocal mammographic findings, and to distinguish between benign and malignant characteristics. Benign characteristics of breast lesions include well-circumscribed and hyperechoic tissue, wider than deep and gently curving smooth lobulation. Meanwhile, malignant characteristics contain sonographic spiculation, deeper (taller) than wide, micro-lobulation, thick hyperechoic halo, angular margins and markedly hyperechoic nodule [45].

Ultrasonography of thyroid is used to diagnose multi-nodular goitre, colloid cyst and thyroglossal duct cyst. Especially for diagnosing intra-nodular vascularity, Doppler ultrasound is required. Large cystic component, comet tail artifact and halo are regarded as benign sonographic features of multi-nodular goitre, while intra-nodular blood flow, large size and presence of micro-calcification are classified as malignant features.

In addition, ultrasound elastography is used for evaluating liver disease or detecting tumors of breast and thyroid by providing additional stiffness information of tissues [46, 47]. Different from the conventional ultrasound imaging modes, it is a relatively new imaging mode for visualizing qualitatively or quantitatively the stiffness information of tissues.

4.3.2. High technologies for radiology ultrasound. For radiology examinations, Samsung Medison has released a premium ultrasound system ‘RS80A with Prestige’ with high technologies:


5. Domestic-Industrial Products and Imaging Technologies

Samsung Medison and Alpinion are leading companies in domestic ultrasound imaging industry. They have released many ultrasound imaging systems for OB/GYN, cardiology and radiology examinations. In this section, we deal with major imaging features contained in the ultrasound systems of Samsung Medison and Alpinion, and see how mathematics is related to the imaging features.

Based on the product information provided through the website or brochure, we compare the imaging technologies of WS80A with Elite and E-CUBE 15EX (Women’s Health), which are high-performance imaging systems for OB/GYN application, of Samsung Medison and Alpinion, respectively. Table 1 shows most of imaging technologies are implemented by beamforming methods, signal processing, image processing methods such as speckle reduction, edge enhancement, contrast enhancement, edge detection and image segmentation, digital scan conversion and elasticity imaging methods. 3D volume rendering and visualization techniques are used very much, especially for OB/GYN applications. As we know already, these ultrasound imaging technologies are basically based on mathematical modelling, numerical solution and visualization.

Table 1. Comparison of imaging technologies of two domestic companies: Samsung Medison and Alpinion for OB/GYN ultrasound. (SCI: spatial compound imaging, FCI: frequency compound imaging, DSC: digital scan conversion)

<table>
<thead>
<tr>
<th>Samsung Medison (WS80A with Elite)</th>
<th>Alpinion (E-CUBE 15EX)</th>
<th>Technologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-Harmonic</td>
<td>Harmonic Imaging</td>
<td>signal processing (filter design)</td>
</tr>
<tr>
<td>ClearVision™</td>
<td>Optimal Imaging Suite™</td>
<td>speckle reduction, edge enhancement, contrast enhancement, SCI, FCI</td>
</tr>
<tr>
<td>5D Heart Color (Fetal heart examination)</td>
<td>-</td>
<td>3D volume rendering, visualization</td>
</tr>
<tr>
<td>5D Limb Vol.</td>
<td>-</td>
<td>edge detection, automated measurement</td>
</tr>
<tr>
<td>5D CNS+ (Fetal brain measurement)</td>
<td>-</td>
<td>visualization, segmentation</td>
</tr>
<tr>
<td>5D NT (Nuchal translucency measurement)</td>
<td>Auto NT (Nuchal Translucency)</td>
<td>3D volume rendering, segmentation</td>
</tr>
<tr>
<td>5D Follicle (Follicle measurement)</td>
<td>-</td>
<td>3D visualization, segmentation</td>
</tr>
<tr>
<td>Realistic Vue™</td>
<td>-</td>
<td>elasticity imaging</td>
</tr>
<tr>
<td>Crystal Vue</td>
<td>-</td>
<td>segmentation, edge detection</td>
</tr>
<tr>
<td>ElastoScan™</td>
<td>-</td>
<td>transducer manufacture</td>
</tr>
<tr>
<td>E-Breast™ (ElastoScan™ for Breast)</td>
<td>-</td>
<td>beam-forming, DSC</td>
</tr>
<tr>
<td>E-Thyroid™ (ElastoScan™ for Thyroid)</td>
<td>-</td>
<td>mathematical modelling, optimization</td>
</tr>
<tr>
<td>S-Detect</td>
<td>-</td>
<td>edge detection</td>
</tr>
<tr>
<td>S-Vue Transducer</td>
<td>-</td>
<td>mathematical modelling, optimization</td>
</tr>
<tr>
<td>Wide angle endocavity transducer</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(E3-12A, Max 210°)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
On the other hand, we compare the imaging features of two systems HS70A and E-CUBE 15EX (Cardiology) for cardiology application. Like in OB/GYN ultrasound, most of the imaging technologies shown in Table 2 are implemented by beam-forming methods, signal processing, image processing methods including cardiac wall motion tracking, digital scan conversion and elasticity imaging methods. Compared to OB/GYN ultrasound, two systems do not equip 3D volume rendering methods, but wall motion tracking methods. In order to develop accurate wall motion tracking methods, understanding and mathematical modelling of cardiac wall motion are indispensable for the technology advancements of cardiology ultrasound.

We note that Samsung Medison and Alpinion lack real-time 3D echocardiography in their released cardiac ultrasound imaging systems, as shown in Table 2. Real-time 3D echocardiography is related to beam-forming algorithms for 2D arrays, 3D rendering and visualization, 3D wall motion tracking methods, spatial/temporal resolution enhancement, and so on.

**Table 2. Comparison of imaging technologies of two domestic companies: Samsung Medison and Alpinion for cardiology ultrasound.**

<table>
<thead>
<tr>
<th>Samsung Medison (HS70A)</th>
<th>Alpinion (E-CUBE 15EX)</th>
<th>Technologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-Harmonic</td>
<td>Advanced Harmonic Imaging</td>
<td>signal processing (filter design)</td>
</tr>
<tr>
<td>Arterial Analysis</td>
<td>Anatomical M-mode</td>
<td>edge detection</td>
</tr>
<tr>
<td>(detection of functional changes of vessels)</td>
<td>Auto IMT</td>
<td>edge detection, strain analysis</td>
</tr>
<tr>
<td>Stress Echo</td>
<td>CUBE Strain™</td>
<td>wall motion analysis, motion tracking</td>
</tr>
<tr>
<td></td>
<td>Stress Echo</td>
<td>wall motion analysis, motion tracking</td>
</tr>
<tr>
<td></td>
<td>Echo Master™</td>
<td>measurement</td>
</tr>
<tr>
<td></td>
<td>Auto Trace PW, CW</td>
<td>edge detection</td>
</tr>
<tr>
<td>S-Detect™</td>
<td></td>
<td>segmentation, edge detection</td>
</tr>
<tr>
<td>E-Breast™</td>
<td></td>
<td>elasticity imaging</td>
</tr>
<tr>
<td>E-Thyroid™</td>
<td></td>
<td>elasticity imaging</td>
</tr>
<tr>
<td>S-Vue Transducer™</td>
<td>high density single crystal phased array transducer</td>
<td>transducer manufacture</td>
</tr>
<tr>
<td>Advanced QuickScan™</td>
<td>Xpeed™</td>
<td>mathematical modelling, optimization</td>
</tr>
</tbody>
</table>

According to radiology applications, the imaging features of RS80A with Prestige and E-CUBE 15EX are listed in Table 3. Among the imaging features, S-Shearwave, S-Fusion and Respiration Auto are added imaging features for radiology ultrasound imaging. S-Shearwave computes the propagate velocity of the shearwave through the targeted lesion using elastic wave equation, while S-Fusion and Respiration Auto are based on image registration methods. S-Fusion enables simultaneous localization of a lesion with a real-time ultrasound in conjunction with other 3D volumetric imaging modalities. Unlike the conventional image fusion technology, Samsung offers a quicker and more precise registration process. Respiration Auto feature minimizes registration differences between the inhaled CT and exhaled ultrasound scan images by generating compensated exhaled CT image.

### 6. Application Cases of Mathematics

As shown in the previous sections, mathematics is widely used in ultrasound imaging. From the understanding of wave propagation and its scattering, beam-forming algorithms, digital
Table 3. Comparison of imaging technologies of two domestic companies: Samsung Medison and Alpinion for radiology ultrasound.

<table>
<thead>
<tr>
<th>Features</th>
<th>Technologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-Harmonic</td>
<td>Harmonic Imaging</td>
</tr>
<tr>
<td>S-Vision beamformer</td>
<td>Harmonic imaging signal processing (filter design)</td>
</tr>
<tr>
<td>HQ Vision</td>
<td>SRI / FullSRI™</td>
</tr>
<tr>
<td>NeedleMate™</td>
<td>Needle Vision™</td>
</tr>
<tr>
<td>CEUS+</td>
<td>edge enhancement, contrast enhancement</td>
</tr>
<tr>
<td>S-Shearwave</td>
<td>contrast enhancement, elasticity imaging, elastic wave equation</td>
</tr>
<tr>
<td>S-Fusion</td>
<td>image registration</td>
</tr>
<tr>
<td>Respiration Auto</td>
<td>edge detection</td>
</tr>
<tr>
<td>S-Tracking</td>
<td>edge detection, strain analysis</td>
</tr>
<tr>
<td>Arterial Analysis</td>
<td>edge detection, strain analysis</td>
</tr>
<tr>
<td>(detection of functional changes of vessels)</td>
<td></td>
</tr>
<tr>
<td>S-3D Arterial Analysis</td>
<td>segmentation, 3D visualization</td>
</tr>
<tr>
<td>(volume measurement of artery plaque)</td>
<td></td>
</tr>
<tr>
<td>Auto IMT+™</td>
<td>edge detection</td>
</tr>
<tr>
<td>Strain+</td>
<td>wall motion analysis, motion tracking</td>
</tr>
<tr>
<td>Stress Echo</td>
<td>wall motion analysis, motion tracking</td>
</tr>
<tr>
<td>S-Detect™ for Breast, S-Detect™ for Thyroid</td>
<td>segmentation, edge detection</td>
</tr>
<tr>
<td>ElastoScan™</td>
<td>elasticity imaging</td>
</tr>
<tr>
<td>E-Strain</td>
<td>elasticity imaging</td>
</tr>
<tr>
<td>E-Breast™</td>
<td></td>
</tr>
<tr>
<td>E-Thyroid™</td>
<td></td>
</tr>
<tr>
<td>S-Vue Transducer</td>
<td>high density single crystal convex transducer, renowned ALPINION’s linear transducer set</td>
</tr>
<tr>
<td>-</td>
<td>transducer manufacture</td>
</tr>
<tr>
<td>Advanced QuickScan™</td>
<td>Panoramic Imaging</td>
</tr>
<tr>
<td>Xpeed™</td>
<td>image registration</td>
</tr>
</tbody>
</table>

scan conversion, signal processing theory, various image processing methods, and so on. In this section, we introduce some practical cases of mathematical applications for ultrasound imaging: LV motion tracking problem, LV vortex flow imaging problem and elasticity imaging study. They were performed for an industry-university cooperation research, by the demands in healthcare or by the demands in ultrasound imaging industry.

6.1. **LV motion tracking problem.** We consider a problem on LV motion tracking and its solutions through mathematical models. The LV motion should be observed and analyzed for evaluations of ventricular systolic/diastolic function, subclinical disease including amyloidosis, LV hypertrophy and hypertrophic cardiomyopathy, coronary artery disease, and so on. Typically, LV motion tracking is performed by observing speckle pattern. Speckle is inherent appearance in ultrasound imaging and its local brightness reflects the local echogeneity of the underlying scatterers. We introduce a case of overcoming limitation in LV motion tracking problem.

6.1.1. **Commonly used LV tracking methods and their limitations.** Among various LV motion tracking methods such as deformable models, statistical methods, classification methods and so on, we can use optical flow methods. Optical flow methods are based on the assumption that
that the intensity of a moving object is constant over time. Let $I(r,t)$ represent the intensity of echocardiography at the location $r=(x,y)$ and the time $t$. Then the noisy time-varying images $I(r,t)$ approximately satisfy

$$u(r,t) \cdot \nabla I(r,t) + \frac{\partial}{\partial t} I(r,t) \approx 0,$$

(6.1)

where $u(r,t)$ is the velocity or displacement vector to be estimated. Especially, Lucas and Kanade [48] used the locally constant motion to compute the velocity $u(r_0,t)$ at a target location $r_0 = (x_0,y_0)$ and time $t$ by forcing constant velocity in a local neighborhood of $r_0 = (x_0,y_0)$ denoted by $N(r_0)$. Following them, Barron et al. [49] proposed an improved model to estimate the velocity $u(r_0,t)$ by minimizing the weighted least square criterion in the neighborhood $N(r_0)$:

$$u(r_0,t) := \arg\min_u \int_{N(r_0)} \left[ w(r-r_0) \left( u(r,t) \cdot \nabla I(r,t) + \frac{\partial}{\partial t} I(r,t) \right)^2 \right] dr,$$

(6.2)

where $w$ is a weight function.

On the other hand, region-based method (also known as the block matching or pattern matching method) can be also used as the LV motion tracking. Duan et al. [50] used the region-based method with cross-correlation coefficient as similarity measure as follows. For given two consecutive images $I(\cdot,t)$ and $I(\cdot,t+\Delta t)$, the displacement $u(r,t)$ at each position $r$ and time $t$ is estimated by maximizing the cross-correlation coefficient:

$$u(r_0,t) := \arg\max_u \left\{ \frac{\int_{N(r_0)}[I(r,t)I(r+u,t+\Delta t)]dr}{\sqrt{\int_{N(r_0)}[I(r,t)]^2dr} \sqrt{\int_{N(r_0)}[I(r+u,t+\Delta t)]^2dr}} \right\}.$$

(6.3)

However, there often exist some incorrectly tracked points in practical environment due to ultrasound artifacts, dropouts, or shadowing phenomena of cardiac wall [51]. It is problematic to track the LV border in ultrasound images with unclear speckle pattern or weak signals. In order to overcome this problems, Ahn [52] proposed a mathematical model for robust myocardial border tracking by considering an affine transformation to describe a global motion that is synthesized by integrating local deformations.

### 6.1.2. A mathematical modelling for robust LV tracking

We denote the LV border traced at initially selected frame by a parametric contour $C^* = \{r(s) = (x^*(s),y^*(s)) \mid 0 \leq s \leq 1\}$ that can be identified as its $n$ tracking points $r^*_1 = r^*(s_1), \ldots, r^*_n = r^*(s_n)$. Here, $0 = s_1 < s_2 < \cdots < s_n = 1$. Let $C(t) = \{r(s,t) = (x(s,t),y(s,t)) \mid 0 \leq s \leq 1\}$ be the contour deformed from $C(0) = C^*$ at time $t$. The motion of the contour $C(t)$ will be determined by an appropriately chosen velocity $U(t)$ indicating a time change of tracking points $(r_1(t),\ldots,r_n(t))$:

$$U(t) := \begin{bmatrix} r_1(t) \\ \vdots \\ r_n(t) \end{bmatrix} = \frac{d}{dt} \begin{bmatrix} r_1(t) \\ \vdots \\ r_n(t) \end{bmatrix} \quad \text{with} \quad \begin{bmatrix} r_1(0) \\ \vdots \\ r_n(0) \end{bmatrix} = \begin{bmatrix} r^*_1 \\ \vdots \\ r^*_n \end{bmatrix}$$
Here, we identify the contour $C(t)$ with tracking points $(r_1(t), \ldots, r_n(t))$. We compute $U(t)$ for each time $t$ by minimizing the following energy functional reflecting local-to-global deformation:

$$
E_t(U) := \frac{1}{2} \sum_{i=1}^{n} \left[ \int_{N(r_i(t))} w(r' - r_i(t)) \left\{ u_i \cdot \nabla I(r', t) + \frac{\partial}{\partial t} I(r', t) \right\}^2 dr' + \lambda \left| r_i(t) + u_i - \begin{bmatrix} a_1(U) \\ a_2(U) \\ a_3(U) \\ a_4(U) \\ a_5(U) \\ a_6(U) \end{bmatrix} \right| \right] (6.4)
$$

where $\lambda$ is a nonnegative parameter and the affine coefficients $a_1(U), \ldots, a_6(U)$ at time $t$ are given by

$$
\begin{bmatrix}
a_1(U) \\
a_2(U) \\
a_3(U) \\
a_4(U) \\
a_5(U) \\
a_6(U)
\end{bmatrix} = \begin{pmatrix} \Phi(C^*)^T \Phi(C^*) \end{pmatrix}^{-1} \Phi(C^*)^T \begin{bmatrix}
(r_1(t) + u_1)^T \\
\vdots \\
(r_n(t) + u_n)^T
\end{bmatrix}, \quad \Phi(C^*) := \begin{bmatrix}
1 \\
\vdots \\
1
\end{bmatrix}.
$$

This study has been performed as an industry-university cooperation research. Based on the proposed model, LV tracking methods of improved performance are being developed.

6.2. A new imaging mode: vortex flow imaging. Vortex flow imaging has recently attracted much attention as a new application for evaluating blood flow [53–55], because it visualizes and quantifies time-varying blood flow inside LV using available ultrasound data. It has shown the potential possibility and availability for evaluating blood flow. In order to compute the velocity fields of blood flow, it is required to model blood flow based on fluid equation.

6.2.1. A reconstruction problem of blood flow. With echo-PIV (particle image velocimetry) [56, 57] being representative, there are several methods to compute and visualize the velocity fields of blood flow inside LV. Echo-PIV is based on optical flow methods tracking the speckle patterns of blood flow to estimate blood motion. However, Echo-PIV is not completely noninvasive because it requires the intravenous injection of a contrast agent to obtain images suitable for the speckle tracking algorithm. To develop less invasive techniques, methods to reconstruct blood flows from color Doppler images have been proposed [58–63]. We represent mathematical models to reconstruct the velocity of blood flow, especially using color Doppler data reflecting one-directional velocity components of blood flow along scanlines.

6.2.2. A mathematical modelling of 3D blood flow. Let $D$ be a 3D imaging domain, $\Omega(t)$ a time-varying LV region satisfying $\Omega(t) \subseteq D \subseteq \mathbb{R}^3$, and $T$ a beat cycle. For the beat cycle $T$, we consider a spatial-temporal domain $\Omega_T$ defined by $\Omega_T := \bigcup_{0 \leq t < T} \Omega(t) \times \{t\} \subseteq D \times (0, T)$.

Let $v(r, t)$ be a velocity field of the blood flow in $\Omega_t \in \Omega_T$ at time $t$. Then the fluid equations governing the blood flow $v$ is given by the 3D incompressible Navier-Stokes equations:

$$
\begin{align*}
\rho \left( \frac{\partial v}{\partial t} + v \cdot \nabla v \right) &= -\nabla p + \mu \nabla^2 v \quad \text{in } \Omega_t, \\
\nabla \cdot v &= 0 \quad \text{in } \Omega_t.
\end{align*}
$$

(6.5)
In order to solve (6.5), we need proper boundary conditions or some conditions based on ultrasound measurement data as follows:

Let \( a(r) \) be scanline directional unit vector at the position \( r \in D \) and \( D(r,t) \) color Doppler image in \( \Omega_t \). Given \( D(r,t) \), we then consider an inverse problem to find a 3D vector field \( v = (u,v,w) \) satisfying the following condition:

\[
a(r) \cdot v(r,t) = D(r,t). \tag{6.6}
\]

However, in clinical practice, it is currently very difficult to acquire 3D color Doppler images given by (6.6). A mathematical model appropriate to the measurements on 2D ultrasound imaging should be considered.

6.2.3. A reconstruction model with mass-source term. Jang et al. [64] suggested a 2D reconstruction model using incompressible Navier-Stokes equations with mass-source term \( s \) to reflect 3D motion of blood flow as the following: Let \( D \) be a 2D imaging domain and \( \Omega(t) \) the cross-section of the LV region in the A3CH view so that they satisfy \( \Omega(t) \subseteq D \subseteq \mathbb{R}^2 \) and \( v(x,t) = (u(x,t), v(x,t)) \) the velocity fields of flow at the position \( x \in D \) and time \( t \). Then the fluid equations governing the blood flow on the imaging plane \( D \) are modeled as

\[
\begin{align*}
\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} &= -\frac{1}{\rho} \frac{\partial p}{\partial x} + \frac{\mu}{\rho} \nabla^2 u + \frac{\mu}{3\rho^2} \frac{s}{\partial x}, \\
\frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} &= -\frac{1}{\rho} \frac{\partial p}{\partial y} + \frac{\mu}{\rho} \nabla^2 v + \frac{\mu}{3\rho^2} \frac{s}{\partial y}, \\
\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} &= \frac{s}{\rho},
\end{align*}
\]

(6.7)

We note that this model (6.7) is equivalent to a 2D incompressible flow having a source-sink distribution \( s(x,t) \) [65]. In order to solve (6.7), we consider color Doppler images measured practically in the imaging plane \( D \). We set the 3D coordinate system for the \( xy \)-plane to contain \( \Omega(t) \) and the \( z \)-axis to be normal to this plane. Let \( D(x,t) \) be the measured color Doppler data, expressed as the inner product of the scanline vector and the velocity vector \( D(x,t) = (a_1(x), a_2(x)) \cdot (u(x,t), v(x,t)) \).

(6.8)

Therefore, we obtain a mathematical problem to find \( v \) satisfying (6.7) and (6.8) simultaneously. Some numerical simulations are performed for solving the given problem and validating the proposed mathematical model.

6.3. Acoustic radiation force-based elastography. Acoustic radiation force(ARF) is caused by absorbed ultrasound wave energy while ultrasound wave propagates through soft tissues. The absorbed energy accumulates in tissue and it generates volume force in the propagation direction of ultrasound beam. This ARF technique enables inducing stress at desired position and measuring response to the stress with a single ultrasound system, whereas other elastography techniques require manual push or additional devices to induce stress or shear waves [66]. We consider ARF-based shear wave elasticity imaging (SWEI) that maps tissue elasticity by
measuring the propagation speed of shear waves generated by ARF. The propagation speed $c$ is expressed as

$$c = \sqrt{\frac{\mu}{\rho}},$$

where $\mu$ and $\rho$ are the shear modulus and density of tissue, respectively. We note that tissue elasticity can be quantified by estimating $c$.

**Figure 5.** A description of elasticity imaging using shear wave. The propagation of shear wave is observed through ultrasound images, acquired from a research ultrasound system E-CUBE 12R (Alpinion, Korea).

6.3.1. **Mathematical model.** An inverse problem for ARF-based ultrasound elastography is to reconstruct the shear modulus $\mu$ from the measured axial movement $w$ which satisfies

$$\rho \frac{\partial^2 w}{\partial t^2} = \mu \nabla^2 w + f_z \text{ in } \mathbb{R}^3 \times (0, T),$$

(6.9)

where $f_z$ is induced ARF in the axial direction, that is, the vertical direction in imaging planes shown in Fig. 5. Note that the excitation $f_z$ and the induced displacement $w$ are not time-harmonic. Let $\text{supp}(f_z)$ be the support of $f_z$ in the spatio-temporal domain. Then, we assume that $\text{supp}(f_z)$ is given because it is possible to change a focal point for inducing ARF. Moreover, we assume that $f_z = 0$ out of $\text{supp}(f_z)$ to avoid the difficulty in quantifying $f_z$ due to the uncertainty of medium [67]. Then, the inverse problem is formulated as a problem to reconstruct the shear modulus $\mu$ from the measured axial movement $w$ satisfying

$$\rho \frac{\partial^2 w}{\partial t^2} = \mu \nabla^2 w \text{ in } (\mathbb{R}^3 \times (0, T)) \setminus \text{supp}(f_z).$$

(6.10)
Nightingale et al. [68] suggested the algebraic inversion of the equation in (6.10) to estimate \( \frac{\mu}{\rho} \):

\[
\frac{\mu}{\rho} = \frac{\partial^2 w}{\partial t^2} \frac{\nabla^2 w}{\nabla^2 w}
\]

(6.11)

However, it could be unstable since numerical approximation of second-order derivatives in (6.10) are sensitive to noise. In many approaches, wave propagation speed is estimated by computing \( \frac{\mu}{\rho} \) instead of the direct inversion [46, 47, 69].

---

**Figure 6.** Illustration of time delays \( T_x \) and \( T_z \) for estimating the lateral speed \( c_x \) and axial speed \( c_z \) of shear wave propagation.

The shear wave propagation speed \( c_x \) in the lateral direction is given by

\[
c_x = \frac{|r_A - r_B|}{T_x}
\]

(6.12)

where \( r_A \) and \( r_B \) are the coordinates of \( A \) and \( B \), respectively. Therefore, \( T_x \) is computed by

\[
T_x = \arg \max_{\tau \in (0, T)} \int_0^T w(x_1, t)w(x_2, t - \tau) \, dt.
\]

(6.13)

Since the wave propagation direction may not be parallel to the lateral direction (Fig. 6) in many cases of ARF-based excitation, we deal with the shear wave propagation speed \( c_z \) in the axial direction. Likewise, let \( T_z \) be the time delay, that is, the difference between arrival times
of the displacement signal by shear wave propagation at two positions $A$ and $C$ placed on an axial line. Then $c_z$ is expressed as

$$c_z = \frac{|r_A - r_C|}{T_z}, \quad (6.14)$$

where $r_C$ is the coordinate of $C$. From $c_x$ and $c_z$, we can compute the propagation speed $c$ [70]:

$$c = \frac{c_x c_y}{\sqrt{c_x^2 + c_z^2}}. \quad (6.15)$$

![Wave propagation](image)

**Figure 7.** A simple description of TOF approach. The disturbance of blue curve corresponds to the displacement at offset from the source. Arrival time of the wave-front is tracked by the red line.

### 6.3.2. Eikonal equation-based approaches.

As illustrated in Fig. 7, wave propagation speed can be estimated from the slope between offset position and arrival time of wave-front. The propagation speed $c$ can be estimated by the gradient of arrival time $T$ using the Eikonal equation [71]:

$$|\nabla T| = \frac{1}{c}, \quad (6.16)$$

where $T(r)$ is defined by

$$T(r) = \inf\{t > 0 : |w(r, t)| > 0\}. \quad (6.17)$$

In a realistic case, $T(r)$ is estimated by using a fixed threshold $\delta$ above a noise level of the estimated $w(r, t)$ [72]:

$$T(r) \approx \inf\{t > 0 : |w(r, t)| > \delta\}. \quad (6.18)$$

As an alternative way, the biased cross-correlation $C$ is suggested by McLaughlin et al. [73]

$$T(r) \approx \arg \max_{\tau \in (0, T)} \frac{1}{T} \int_0^T w(r_{ref}, t) \tilde{w}(r, t - \tau) dt \text{ for } r_{ref} \text{ a reference point}, \quad (6.19)$$
where

\[ \tilde{w}(r, t) = \begin{cases} 
  w(r, t) & \text{if } 0 \leq t \leq T, \\
  w(r, t - T) & \text{if } t \geq T, \\
  w(r, t + T) & \text{if } t \leq 0.
\]  

(6.20)

7. CURRENT HOT TOPICS AND CHALLENGING ISSUES

In this section, we introduce current hot topics in ultrasound imaging, especially in the domestic ultrasound industry. Samsung Medison, a global medical equipment company and a leading domestic company, has recently published some white papers [74–80] on: S-Detect™, 5D CNS, E-Thyroid™, E-Breast™ (breast ElastoScan™), S-Shearwave, Aterial Analysis, and so on. The white papers report the advantages and shortcomings of each technology.

7.1. Hot topics in ultrasound imaging.

- S-Detect™ is a software analyzing the features of lesions and assessing the possibility of malignancy. It uses a machine learning-based algorithm [81]. By showing excellent agreement of 91.2% with the assessment of breast dedicated radiologist in interpreting the breast mass, it is suggested as a good decision-making support for the beginners or non-breast radiologists. However, with the sensitivity of 84.6% at the same time, it missed two breast cancers of relatively circumscribed isoechoic and hypoechoic masses with suspicious clinical findings. Therefore, it is reported that circumscribed malignant masses may be remained as the limitation of S-Detect™ and S-Detect™ is not available to find subtle suspicious features [74].

- A semi-automatic method is proposed for biometric measurements of fetal central nervous system (CNS) from 3D ultrasound volume data of brain. It reduce the number of operations and examination time. Its high success rate >90% is reported under clinical evaluation [75].

- E-Thyroid™ uses carotid artery pulsation, not external compression by free hands, as a compression source. While E-Thyroid™ effectively differentiates malignant from benign in most thyroid nodules including calcified nodule, the extreme location of a nodule in the thyroid can affect the results [76].

- E-Breast™ is helpful for characterizing different regions as a complementary diagnostic imaging technology, based on strain imaging technique. However, it would be important not to evaluate a lesion from an isolated manner with strain ratio generated from E-Breast™ [77].

- S-Shearwave is a technology enabling quantitative analysis of tissue stiffness for assessing liver fibrosis. Unlike the color map used in the conventional elasticity imaging, S-Shearwave displays the stiffness value and Reliable Measurement Index (RMI) for the region of interest [78].

Among them, S-Detect™, 5D CNS and Aterial Analysis are CADe/CADx softwares. There are numerous demands of CADe and CADx for many medical imaging methodologies [82–85] as well as ultrasound [86–88]. According to the applications, CADe is implemented by various
image processing methods: histogram-based thresholding and active contours for segmenting anomaly [83], neural network-based approaches for robust segmentation [89–91]. In the literatures [86, 87], it is reported that neural network and support vector machine techniques are used for CADx.

On the other hand, currently commercialized elasticity imaging modes:E-Thyroid™, E-Breast™, S-Shearwave [76–78] are commonly used as a complementary diagnostic tool. It is required to quantify more accurately the stiffness of tissues.

Fusion imaging techniques have been used for volume navigation of 2D ultrasound images within other 3D CT or MR images. In order to register different images acquired from medical imaging modalities, the fusion imaging techniques have been developed with the help of landmarks, position sensors, or electromagnetic needle tracking [92–94]. As real-time 3D ultrasound imaging becomes feasible, fusion imaging technique for 3D volume data is considered as a promising tool for computer-aided surgery [94–96]. In 3D fusion imaging, external devices such as landmarks, position sensors or electromagnetic needle tracking can be removed [97,98]. Samsung Medison developed the fusion imaging techniques called S-Fusion.

Additionally, we introduce some issues in ultrasound imaging, related to mathematical modelling, numerical solutions, image processing and visualization.

7.2. Issues on LV tracking methods.

• Real-time 3D ultrasound imaging is capable of providing good 3D visualization of organs. However, its resolution is not good enough to discriminate features smaller than a few millimeters and LV tracking methods based on speckles are not appropriate to be applied to 3D ultrasound images. We can consider 3D LV tracking by constructing the motion of 3D LV shape from 2D LV borders, which are extracted by applying LV tracking methods to multiple 2D echocardiography data. How can we model the relationship between the change of LV borders observed on 2D echocardiography and the 3D LV shape?

• LV motion tracking methods include inevitable limitation on 2D echocardiography because of the helical structure of cardiac ventricular anatomy. For a heartbeat cycle, it is difficult to track a portion of LV border designated at the initial stage on a 2D ultrasound image. In fact, it gets out of the imaging plane. How can we model the relationship between the change of LV borders observed on 2D echocardiography and the 3D helical behavior of cardiac motion?

7.3. Issues on vortex flow imaging.

• Vortex flow imaging consists of defining LV geometry from echocardiographic images and reconstructing the velocity field of blood flow in the moving LV region. Let us assume that the acquisition of 3D color Doppler imaging at high frame rate is available. Can we model an inverse problem for reconstructing the 3D velocity field of blood flow from the partial velocity information obtained through the 3D color Doppler images? [99].
7.4. Issues on elasticity imaging.

- Inherent low SNR of the estimated $w$ makes the direct inversion (6.11) difficult so that many models reconstructing $c$ are based on the time-of-flight (TOF)-based approaches. However, they do not fully guarantee stability in practical situations since the inverse of time-delay or $|∇T|$ are still necessary for the speed estimation. In order to overcome the problem, Wang et al. suggest a fitting model for the estimated $T(r)$ [100]

$$T(r) = \beta \cdot r + t_0,$$

where $t_0$ is a fitting parameter and $\beta$ a fitting parameter vector satisfying

$$|\beta| = c.$$

However, in inhomogeneous medium, the fitting model (7.1) could not describe $T(r)$ properly because refraction on the boundary between different tissues makes non-linear distortion of arrival-time. We need a modified fitting model for $T(r)$ which assumes the presence of anomaly.

8. Conclusion

Ultrasound imaging industry is very promising in terms of technology advancements as well as the market size and growth. Like in other industries, the ultrasound imaging technologies have been advanced and innovated through Mathematics. In this article, we described the history of ultrasound imaging, its basic principle, its diagnostic applications, domestic-industrial products, practical use of mathematics, hot topics and challenging issues related to ultrasound imaging. Through them, we confirmed that Mathematics has been used commonly as mathematical modelling, numerical solutions and visualization, combined with science, engineering and medicine in ultrasound imaging.

The practical use of Mathematics in ultrasound imaging requires the understanding of human body and imaging system overall. Based on that understanding, we should perform physics-based mathematical modelling, deal with the measurement data acquired through ultrasound imaging systems and solve various problems and challenging issues in ultrasound imaging.

Currently, there are still numerous demands of technology advancements in healthcare and ultrasound imaging industry, especially in domestic ultrasound industry. As we have seen before, Mathematics can contribute to the advances and innovation of ultrasound imaging technology. We hope many mathematicians contribute much to ultrasound technology innovation.

Acknowledgements

This work was supported by the National Institute for Mathematical Sciences (NIMS) and the National Research Foundation of Korea (NRF) grants funded by the Korea government (No. NRF-20151009350). We would like to thank reviewers for their detailed comments and suggestions to improve the paper. Also, we thank Haeeun Han for her help with rendering the TOC art.
REFERENCES


J. McLaughlin and D. Renzi, Using level set based inversion of arrival times to recover shear wave speed in transient elastography and supersonic imaging, Inverse Problems, 22(2) (2006), 707.


D. J. Lim and M. H. Kim, Experiences of Intrinsic Compression Ultrasound Elastography (E-Thyroid™) in Differentiating Benign From Malignant Thyroid Nodule, White Paper, WP201504-E-Thyroid™, Samsung Medison, 2015.


SORET, HALL CURRENT, ROTATION, CHEMICAL REACTION AND THERMAL RADIATION EFFECTS ON UNSTEADY MHD HEAT AND MASS TRANSFER NATURAL CONVECTION FLOW PAST AN ACCELERATED VERTICAL PLATE

M. VENKATESWARLU¹†, D. VENKATA LAKSHMI², AND K. NAGA MALLESWARA RAO³

¹DEPARTMENT OF MATHEMATICS, V. R. SIDDHARTHA ENGINEERING COLLEGE, KANUR, KRISHNA (DIST), ANDHRA PRADESH, INDIA-520007
E-mail address: mvsr2010@gmail.com, mvenkateswarlu@vrsiddhartha.ac.in

²DEPARTMENT OF MATHEMATICS, BAPATLA WOMEN’S ENGINEERING COLLEGE, BAPATLA, GUNTUR (DIST), ANDHRA PRADESH, INDIA-522102

³DEPARTMENT OF MECHANICAL ENGINEERING, V. R. SIDDHARTHA ENGINEERING COLLEGE, KANUR, KRISHNA (DIST), ANDHRA PRADESH, INDIA-520007

ABSTRACT. The heat and mass transfer characteristics of the unsteady hydromagnetic natural convection flow with Hall current and Soret effect of an incompressible, viscous, electrically conducting, heat absorbing and optically thin radiating fluid flow past a suddenly started vertical infinite plate through fluid saturated porous medium in a rotating environment are taken into account in this paper. Derivations of exact analytical solutions are aimed under different physical properties. The velocity, concentration and temperature profiles, Sherwood number and Nusselt number are easily examined and discussed via the closed forms obtained. Soret effect and permeability parameter tends to accelerate primary and secondary fluid velocities whereas hall current, radiation and heat absorption have reverse effect on it. Radiation and heat absorption have tendency to enhance rate of heat transfer at the plate. The results obtained here may be further used to verify the validity of obtained numerical solutions for more complicated transient free convection fluid flow problems.

NOMENCLATURE

\begin{align*}
u^* &= \text{fluid velocity in } x^* \text{ direction} \\
w^* &= \text{fluid velocity in } z^* \text{ direction} \\
g &= \text{acceleration due to gravity} \\
D_T &= \text{thermal diffusivity} \\
t^* &= \text{time} \\
q_r^* &= \text{radiating flux vector} \\
T^* &= \text{fluid temperature} \\
m &= \text{hall current parameter}
\end{align*}

Received by the editors March 21 2016; Revised June 23 2016; Accepted in revised form July 8 2016; Published online September 8 2016.

2000 Mathematics Subject Classification. 76R10, 76W05, 80A20, 80A32.

Key words and phrases. MHD, Hall current, soret effect, Thermal radiation and chemical reaction.

† Corresponding author.
1. INTRODUCTION

In recent years, the problems of hydromagnetic free convective and heat transfer flow in a porous medium plays an important role in several scientific and industrial processes such as problems of boundary layer flow control, plasma studies, thermo nuclear fusion, furnace design, geothermal energy extraction, metallurgy, chemical, mineral and petroleum engineering, solar power technology etc. and on the performance of so many engineering devices using electrically conducting fluids, namely, MHD accelerators, MHD generators, nuclear reactors, MHD pumps, MHD flow-meters, plasma jet engines, etc. Raptis and Kafousias [1] investigated Magnetohydrodynamic free convection flow and mass transfer through porous medium bounded by an infinite vertical porous plate with constant heat flux. Raptis [2] discussed free convection and mass transfer effects on the oscillatory flow past an infinite moving vertical isothermal plate with constant suction and heat sources. Unsteady hydromagnetic free convection flow with Hall current mass transfer and variable suction through a porous medium near an infinite vertical porous plate with constant heat flux was studied by Sattar [3]. Lai and Kulacki [4] analyzed Non-Darcy mixed convection along a vertical wall in a saturated porous medium. Comprehensive reviews of natural convection boundary layer flow over various geometrical bodies with heat and mass transfer in porous and non-porous media are well documented by
Eckert and Drake [5], Gebhart et al. [6], Nield and Bejan [7], Pop and Ingham [8] and Incropera et al. [9]. Seth et al. [10] investigated on effects of hall current on hydromagnetic free convection flow with heat and mass transfer of a heat absorbing fluid past an impulsively moving vertical plate with ramped temperature.

Moreover, the chemical reaction, heat and mass transfer on MHD flow over a vertical stretching surface with heat source and thermal stratification have been presented by Kandasamy et al. [11]. Dursunkaya and Worek [12] analyzed diffusion-thermo and thermal diffusion effects in transient and steady natural convection from vertical surface. Postelnicu [13] studied the influence of a magnetic field on heat and mass transfer by natural convection from vertical surfaces in porous media considering soret and dufour effects. Alam and Rahman [14] considered dufour and soret effects on MHD free convective heat and mass transfer flow past a vertical flat plate embedded in a porous medium.

It is observed that when the density of an electrically conducting fluid is low and/or applied magnetic field is strong, Hall current is produced in the flow-field which plays an important role in determining flow features of the problems because it induces secondary flow in the flow-field. Keeping in view this fact, significant investigations on hydromagnetic free convection flow past a flat plate with Hall effects under different thermal conditions are carried out by several researchers in the past. Mention may be made of the research studies of Pop and Watanabe [15], Abo-Eldahab and Elbarbary [16], Takhar et al. [17] and Saha et al. [18]. It is worthy to note that Hall current induces secondary flow in the flow-field which is also the characteristics of Coriolis force. Therefore, it becomes very important to compare and contrast the effects of these two agencies and also to study their combined effects on such fluid flow problems. Satya Narayana et al. [19] studied the effects of Hall current and radiation–absorption on MHD natural convection heat and mass transfer flow of a micropolar fluid in a rotating frame of reference. Seth et al. [20] investigated effects of Hall current and rotation on unsteady hydromagnetic natural convection flow of a viscous, incompressible, electrically conducting and heat absorbing fluid past an impulsively moving vertical plate with ramped temperature in a porous medium taking into account the effects of thermal diffusion.

Al-Odat and Al-Azab [21] studied the influence of magnetic field on unsteady free convective heat and mass transfer flow along an impulsively started semi-infinite vertical plate taking into account a homogeneous chemical reaction of first order. The effect of radiation on the heat and fluid flow over an unsteady stretching surface has been analyzed by El-Aziz [22]. Singh et al. [23] studied the heat transfer over stretching surface in porous media with transverse magnetic field. Singh et al. [24] and [25] also investigated MHD oblique stagnation-point flow towards a stretching sheet with heat transfer for steady and unsteady cases. Elbashbeshy et al. [26] investigated the effects of thermal radiation and magnetic field on unsteady boundary layer mixed convection flow and heat transfer problem from a vertical porous stretching surface. The opposing buoyancy effects on simultaneous heat and mass transfer by natural convection in a fluid saturated porous medium investigated by Angirasa et al. [27]. Ahmed [28] investigates the effects of unsteady free convective MHD flow through a porous medium bounded by an infinite vertical porous plate. Ahmed Sahin [29] studied the Magneto hydrodynamic and chemical reaction effects on unsteady flow, heat and mass transfer characteristics in
a viscous, incompressible and electrically conduction fluid over a semi-infinite vertical porous plate in a slip-flow regime. Recently, Venkateswarlu et al. [30-35] investigated heat and mass transfer effects on convective MHD flows through porous medium in presence of thermal radiation and chemical reaction.

The objective of the present investigation is to study an unsteady natural convection flow of a viscous, incompressible and electrically conducting fluid with radiative heat and mass transfer past an impulsively moving vertical plate embedded in a fluid saturated porous medium taking into account the effects of the thermal diffusion, hall current, rotation, chemical reaction and thermal radiation.

The rest of the present paper is arranged in the following fashion. The problem is formulated in section 2. Analysis of the existence of unique exponential type solutions are given in section 3. Results and discussions in section 4 are followed by the concluding remarks in section 5.

2. FORMATION OF THE PROBLEM

![Physical model and coordinate system](image)

**FIGURE 1.** Physical model and coordinate system
We consider unsteady MHD natural convection flow of an incompressible, viscous, electrically conducting and temperature dependent heat absorbing and optically thin radiating fluid past a moving infinite vertical plate embedded in a porous medium taking Soret and Hall effects into account. Select the coordinate system in such way that $x^*$-axis is along the length of the plate in the upward direction and $y^*$-axis normal to the plane of the plate in the fluid. A uniform transverse magnetic field $B_0$ is applied parallel to $y^*$-axis. Both the fluid and plate are in rigid body rotation with uniform angular velocity $\Omega$ about $y^*$-axis. Initially i.e. at time $t^* \leq 0$, both the fluid and plate are at rest and at uniform temperature $T_\infty^*$. Also species concentration within the fluid is maintained at uniform concentration $C_\infty^*$. At time, $t^* > 0$ plate starts moving with time dependent velocity $U(t^*)$ in $x^*$ direction and temperature of the plate is raised or lowered to $T_\infty^* + (T^* - T_\infty^*)t^*/t_0^*$ when $t^* \leq t_0^*$, and thereafter, i.e. at $t^* > t_0^*$ plate is maintained at uniform temperature $T_w^*$. Also species concentration at the surface of the plate is raised to uniform species concentration $C_w^*$ and is maintained thereafter. Geometry of the flow problem is presented in Fig. 1. Since plate is of infinite extent along $x^*$ and $z^*$ directions and is electrically non-conducting, all physical quantities depend on $y^*$ and $t^*$ only. It is assumed that the induced magnetic field produced by fluid motion is negligible in comparison to the applied one. This assumption is valid because magnetic Reynolds number is very small for metallic liquids and partially ionized fluids [36]. Also no external electric field is applied so the effect of polarization of fluid is negligible.

Keeping in view of the above assumptions, the governing equations for unsteady MHD natural convection flow of a viscous, incompressible, electrically conducting and temperature dependent heat absorbing and optically thin heat radiating fluid in a uniform porous medium taking Hall current, Soret, Rotation and Chemical reaction effects into account are given by

Continuity equation:
$$\frac{\partial v^*}{\partial y^*} = 0$$

Momentum conservation equations:
$$\frac{\partial u^*}{\partial t^*} + 2\Omega w^* = \nu \frac{\partial^2 u^*}{\partial y^*^2} - \frac{\sigma B_0^2}{\rho} \left[ \frac{u^* + mw^*}{1 + m^2} \right] + g\beta_T (T^* - T_\infty^*)$$
$$+ g\beta_C (C^* - C_\infty^*) - \frac{u^*}{K_1^r}$$

$$\frac{\partial w^*}{\partial t^*} - 2\Omega u^* = \nu \frac{\partial^2 w^*}{\partial y^*^2} + \frac{\sigma B_0^2}{\rho} \left[ \frac{mu^* - w^*}{1 + m^2} \right] - \frac{w^*}{K_1^r}$$

Energy conservation equation:
$$\frac{\partial T^*}{\partial t^*} = \frac{k_T}{\rho c_p} \frac{\partial^2 T^*}{\partial y^*^2} - \frac{Q_0}{\rho c_p} (T^* - T_\infty^*) - \frac{1}{\rho c_p} \frac{\partial q_v^*}{\partial y^*}$$

Mass diffusion equation:
$$\frac{\partial C^*}{\partial t^*} = D_M \frac{\partial^2 C^*}{\partial y^*^2} + D_T \frac{\partial^2 T^*}{\partial y^*^2} - K_r (C^* - C_\infty^*)$$
where \( m = \omega_e \tau_e \) is the Hall current parameter, \( u^* \) — fluid velocity in \( x^* \) direction, \( w^* \) — fluid velocity along \( z^* \) direction, \( g^* \) — acceleration due to gravity, \( \rho^* \) — fluid density, \( \beta_T \) — coefficient of thermal expansion, \( \beta_C \) — coefficient of concentration expansion, \( t^* \) — time, \( \Omega \) — angular velocity, \( K_1^* \) — permeability of porous medium, \( B_0^* \) — magnetic induction, \( T^* \) — fluid temperature, \( k_T \) — thermal conductivity, \( C^* \) — species concentration, \( \sigma^* \) — electrical conductivity, \( \epsilon_p^* \) — specific heat at constant pressure, \( D_M^* \) — chemical molecular diffusivity, \( D_T^* \) — thermal diffusivity, \( q_{r^*}^* \) — radiating flux vector, \( \nu^* \) — kinematic coefficient of viscosity, \( K_r^* \) — chemical reaction parameter, \( \omega_e \) — cyclotron frequency and \( \tau_e \) — electron collision time respectively.

The initial and boundary conditions for the fluid flow problem are given below

\[
\begin{align*}
  u^* &= w^* = 0, \quad T^* = T_w^*, \quad C^* = C_w^* \quad \text{at} \quad y^* \geq 0 \quad \text{and} \quad t^* \leq 0 \\
  u^* &= U(t^*), \quad w^* = 0, \quad C^* = C_w^* \quad \text{at} \quad y^* = 0 \quad \text{and} \quad t^* > 0 \\
  T^* &= \left( \frac{T_{w^*} + (T_w^* - T_{w}^*)t^*}{t_0} \right) \quad \text{at} \quad y^* = 0 \quad \text{and} \quad 0 < t^* \leq t_0 \\
  T^* &= T_w^* \quad \text{at} \quad y^* = 0 \quad \text{and} \quad t^* > t_0 \\
  u^* &\rightarrow 0, \quad w^* \rightarrow 0, \quad T^* \rightarrow T_w^*, \quad C^* \rightarrow C_w^* \quad \text{as} \quad y^* \rightarrow \infty \quad \text{and} \quad t^* > 0
\end{align*}
\]

(2.6)

where \( T_{w^*} \) — temperature of the wall, \( C_{w^*} \) — concentration of the wall, \( T_{w^*}^* \) — fluid temperature in the free stream, \( C_{w^*}^* \) — species concentration in the free stream and \( U(t^*) \) — time dependent velocity respectively.

In the case of an optically thin gray fluid the local radiant [37] absorption is expressed as

\[
\frac{\partial q_{r^*}^*}{\partial y^*} = -4a^* \sigma^* \left( T_{\infty}^* - T^* \right)
\]

(2.7)

where \( a^* \) is mean absorption coefficient and \( \sigma^* \) is Stefan–Boltzmann constant.

It is assumed that the temperature difference within the fluid flow is sufficiently small such that the fluid temperature \( T_{\infty}^* \) may be expressed as a linear function of the temperature. This is accomplished by expanding \( T_{\infty}^* \) in a Taylor’s series about free stream temperature \( T_{\infty}^* \). By neglecting second and higher order terms, \( T_{\infty}^* \) is expressed as

\[
T_{\infty}^* \approx 4T_{\infty}^* T^* - 3T_{\infty}^*^3
\]

(2.8)

Using the equations (2.7) and (2.8) in the last term of equation (2.4) we obtain

\[
\frac{\partial T^*}{\partial t^*} = \frac{k_T}{\rho \epsilon_p} \frac{\partial^2 T^*}{\partial y^*} - \frac{Q_0}{\rho \epsilon_p} \left( T^* - T_{\infty}^* \right) - \frac{16a^* \sigma^* T_{\infty}^3}{\rho \epsilon_p} \left( T^* - T_{\infty}^* \right)
\]

(2.9)

In order to write the governing equations and the boundary conditions in dimensional form, the following non-dimensional quantities are introduced.

\[
y = \frac{y^*}{U_0(t_0)}, \quad u = \frac{u^*}{U_0}, \quad w = \frac{w^*}{U_0}, \quad t = \frac{t^*}{t_0}, \quad T = \frac{T^* - T_{w}^*}{T_{w}^* - T_{\infty}^*}, \quad C = \frac{C^* - C_w^*}{C_{w^*} - C_{w^*}^*}
\]

(2.10)

Equations (2.2), (2.3), (2.5), and (2.9) reduce to the following dimensional form.

\[
\frac{\partial u}{\partial t} + 2K^2 w = \frac{\partial^2 u}{\partial y^2} - \frac{M(u + mw)}{1 + m^2} + G_T + G_m C - \frac{u}{K_1}
\]

(2.11)
\[
\frac{\partial w}{\partial t} - 2K^2 u = \frac{\partial^2 w}{\partial y^2} + \frac{M(mu - w)}{1 + m^2} - \frac{w}{K_1} \tag{2.12}
\]
\[
\frac{\partial T}{\partial t} = \frac{1}{Pr} \frac{\partial^2 T}{\partial y^2} - (R + \phi)T \tag{2.13}
\]
\[
\frac{\partial C}{\partial t} = \frac{1}{Sc} \frac{\partial^2 C}{\partial y^2} + S_0 \frac{\partial^2 T}{\partial y^2} - K_r C \tag{2.14}
\]

where \(K^2 = \frac{\Omega_k}{U_0^2}\) is the rotation parameter, \(M = \frac{\sigma B_0^2 \mu}{\rho \nu^2}\) is the magnetic parameter, \(K_1 = \frac{K^2 U_0^2}{\nu^2}\) is the permeability parameter, \(G_r = \frac{g \beta_r V(T^*_0 - T^*_\infty)}{U_0^3}\) is the thermal Grashof number, \(G_m = \frac{g \beta \nu (C^*_w - C^*_\infty)}{U_0^3}\) is the Solutal Grashof number, \(Pr = \frac{\nu \rho c p \sigma B_0}{k_f}\) is the Prandtl number, \(R = \frac{16a^* \sigma^* \nu T^*_\infty}{\rho c p U_0^3}\) is the radiation parameter, \(\phi = \frac{\nu Q_0}{\rho c p U_0^3}\) is the heat absorption parameter, \(Sc = \frac{\nu}{D_M}\) is the Schmidt number, \(S_0 = \frac{D_r (T^*_0 - T^*_\infty)}{\nu (C^*_w - C^*_\infty)}\) is the Soret number, \(K_r = \frac{K^2 \nu}{U_0^3}\) is the chemical reaction number respectively.

It is noticed that characteristic time \(t_0\) may be defined according to the non-dimensional method mentioned above as
\[
t_0 = \frac{\nu}{U_0^2} \tag{2.15}
\]
where \(U_0\) is characteristic velocity.

The initial and boundary conditions, presented by equation (2.6) reduced to the following non-dimensional form
\[
\begin{aligned}
&u = 0, \ w = 0, \ T = 0, \ C = 0 \text{ at } y \geq 0 \text{ and } t \leq 0 \\
&u = f(t), \ w = 0, \ C = 1 \text{ at } y = 0 \text{ and } t > 0 \\
&T = t \text{ at } y = 0 \text{ and } 0 < t \leq 1 \\
&T = 1 \text{ at } y = 0 \text{ and } t > 1 \\
&u \to 0, \ w \to 0, \ T \to 0, \ C \to 0 \text{ as } y \to \infty \text{ and } t > 0
\end{aligned} \tag{2.16}
\]

By combining the equations (2.11) and (2.12), we obtain
\[
\frac{\partial F}{\partial t} - 2iK^2 F = \frac{\partial^2 F}{\partial y^2} + \frac{F}{K_1} - NF + G_r T + G_m C \tag{2.17}
\]
where \(F = u + iw\) and \(N = \frac{M}{1 + m^2}\).

The initial and boundary conditions, presented by equation (2.16), in compact form, are given by
\[
\begin{aligned}
&F = 0, \ T = 0, \ C = 0 \text{ at } y \geq 0 \text{ and } t \leq 0 \\
&F = f(t), \ C = 1 \text{ at } y = 0 \text{ and } t > 0 \\
&T = t \text{ at } y = 0 \text{ and } 0 < t \leq 1 \\
&T = 1 \text{ at } y = 0 \text{ and } t > 1 \\
&F \to 0, \ T \to 0, \ C \to 0 \text{ as } y \to \infty \text{ and } t > 0
\end{aligned} \tag{2.18}
\]
In order to investigate the flow features of the fluid generated due to uniformly accelerated movement of the plate, we consider $f(t) = R_1 t$, where $R_1$ a non-dimensional constant.

It is now important to calculate physical quantities of primary interest, which are the local wall shear stress or skin friction coefficient, the local surface heat flux and the local surface mass flux. Given the velocity, temperature and concentration fields in the boundary layer, the shear stress $\tau_w$, the heat flux $q_w$ and mass flux $j_w$ are obtained by

\begin{align*}
\tau_w &= \mu \frac{\partial F}{\partial y} \bigg|_{y=0} \quad (2.19) \\
q_w &= -\alpha \frac{\partial T}{\partial y} \bigg|_{y=0} \quad (2.20) \\
j_w &= -D_m \frac{\partial C}{\partial y} \bigg|_{y=0} \quad (2.21)
\end{align*}

In non-dimensional form the skin-friction coefficient $C_f$, heat transfer coefficient $Nu$ and mass transfer coefficient $Sh$ are defined as

\begin{align*}
C_f &= \frac{\tau_w}{\rho \nu^2} \quad (2.22) \\
Nu &= \frac{\nu}{U_o k_T} \frac{q_w}{(T_w^* - T_\infty^*)} \quad (2.23) \\
Sh &= \frac{\nu}{U_o D_m} \frac{j_w}{(C_w^* - C_\infty^*)} \quad (2.24)
\end{align*}

Using non-dimensional variables in equation (2.10) and equations (2.19) to (2.21) into equations (2.22) to (2.24), we obtain the physical parameters

\begin{align*}
C_f &= \left[ \frac{\partial F}{\partial y} \right]_{y=0} \quad (2.25) \\
Nu &= -\left[ \frac{\partial T}{\partial y} \right]_{y=0} \quad (2.26) \\
Sh &= -\left[ \frac{\partial C}{\partial y} \right]_{y=0} \quad (2.27)
\end{align*}

3. Solution of the Problem

Equations (2.13), (2.14) and (2.17) are coupled, nonlinear partial differential equations and these cannot be solved in closed form. So, we reduced these nonlinear partial differential equations into a set of ordinary differential equations, which can be solved analytically. This can be done by assuming the trial solutions for the velocity, temperature and concentration of the fluid as

\begin{align*}
F(y, t) &= F_0(y) + \varepsilon \exp (i \omega t) F_1(y) + o(\varepsilon^2) + \ldots \\
T(y, t) &= T_0(y) + \varepsilon \exp (i \omega t) T_1(y) + o(\varepsilon^2) + \ldots 
\end{align*}

\begin{align*}
C_w(y, t) &= C_0(y) + \varepsilon \exp (i \omega t) C_1(y) + o(\varepsilon^2) + \ldots \quad (3.1)
\end{align*}
\[ C(y, t) = C_0(y) + \varepsilon \exp(i\omega t) C_1(y) + o(\varepsilon^2) + \ldots \ldots \] (3.3)

where \( \omega \) is frequency of oscillation and \( \varepsilon \ll 1 \).

Substituting (3.1), (3.2), and (3.3) in equations (2.13), (2.14), and (2.17), then equating the harmonic and non-harmonic terms and neglecting the higher order terms of \( o(\varepsilon^2) \), we obtain

\[
F_0'' = \left[ N + \frac{1}{K_1} - 2iK^2 \right] F_0 = -(G_r T_0 + G_m C_0)
\] (3.4)

\[
F_1'' = \left[ N + \frac{1}{K_1} - 2iK^2 + i\omega \right] F_1 = -(G_r T_1 + G_m C_1)
\] (3.5)

\[
T_0'' - P_r (R + \phi) T_0 = 0
\] (3.6)

\[
T_1'' - P_r (R + \phi + i\omega) T_1 = 0
\] (3.7)

\[
C_0'' - S_c K_r C_0 = -S_c S_0 T_0''
\] (3.8)

\[
C_1'' - S_c (K_r + i\omega) C_1 = -S_c S_0 T_1''
\] (3.9)

where prime denotes the ordinary differentiation with respect to \( y \).

The corresponding boundary conditions can be written as

\[
\begin{align*}
F_0 &= 0, \quad F_1 = 0, \quad T_0 = 0, \quad T_1 = 0, \quad C_0 = 0, \quad C_1 = 0 \quad \text{at} \quad y \geq 0 \quad \text{and} \quad t \leq 0 \\
F_0 &= f(t), \quad F_1 = 0, \quad C_0 = 1, \quad C_1 = 0 \quad \text{at} \quad y = 0 \quad \text{and} \quad t > 0 \\
T_0 &= t, \quad T_1 = 0 \quad \text{at} \quad y = 0 \quad \text{and} \quad 0 < t \leq 1 \\
T_0 &= 1, \quad T_1 = 0 \quad \text{at} \quad y = 0 \quad \text{and} \quad t > 1 \\
F_0 &\to 0, \quad F_1 &\to 0, \quad T_0 &\to 0, \quad T_1 &\to 0, \quad C_0 &\to 0, \quad C_1 &\to 0 \quad \text{as} \quad y \to \infty \quad \text{and} \quad t > 0
\end{align*}
\] (3.10)

Solving equations (3.4)-(3.9) under the boundary condition (3.10) we obtain the velocity, temperature and concentration distribution in the boundary layer as

\[
F(y, t) = A_{11} \exp(-A_3 y) + A_{13} \exp(-A_{11} y) + A_{14} \exp(-A_{12} y)
\] (3.11)

\[
T(y, t) = t \exp(-A_{11} y)
\] (3.12)

\[
C(y, t) = A_6 \exp(-A_{11} y) + A_7 \exp(-A_{12} y)
\] (3.13)

The Skin friction, Nusselt number and Sherwood number are important physical parameters for this type of boundary layer flow.

**Skin friction coefficient:**

Knowing the velocity field, the skin friction coefficient can be obtained, which is in non-dimensional form is given by

\[
C_f = \left( \frac{\partial F}{\partial y} \right)_{y=0} = -[A_3 A_{11} + A_1 A_{13} + A_9 A_{14}]
\] (3.14)
Nusselt number:
Knowing the temperature field, the rate of heat transfer coefficient can be obtained, which is in non-dimensional form is given, in terms of the Nusselt number, is given by

\[
N_u = -\left(\frac{\partial T}{\partial y}\right)_{y=0} = tA_1
\]  

(3.15)

Sherwood number:
Knowing the concentration field, the rate of mass transfer coefficient can be obtained, which is in non-dimensional form, in terms of the Sherwood number, is given by

\[
S_h = -\left(\frac{\partial C}{\partial y}\right)_{y=0} = A_1A_6 + A_3A_7
\]  

(3.16)

Here the constants are not given due to shake of brevity.

4. RESULTS AND DISCUSSIONS

In order to investigate the influence of various physical parameters such as hall current parameter \(m\), Soret effect \(S_0\), rotation parameter \(K^2\), magnetic parameter \(M\), radiation parameter \(R\), permeability parameter \(K_1\), heat absorption parameter \(\phi\), Prandtl number \(P_r\), thermal buoyancy force \(G_r\), solutal buoyancy force \(G_m\), chemical reaction parameter \(K_r\), mass diffusion parameter \(S_c\) and time \(t\) on the flow-field, primary velocity \(u\), secondary velocity \(w\), temperature \(T\) and concentration \(C\) have been studied analytically and computed results of the analytical solutions, presented by equations (3.11)-(3.13) are displayed graphically from Figs. 2 to 22. In the present study following default parameter values are adopted for computations: \(G_r = 4\), \(G_m = 3\), \(P_r = 0.71\), \(K_r = 0.5\), \(R = 2.0\), \(m = 0.5\), \(K^2 = 2.0\), \(\omega = 0.5\), \(M = 1.0\), \(S_c = 0.6\), \(R_1 = 0.5\), \(t = 0.5\), \(K_1 = 0.5\), \(\phi = 3.0\) and \(S_0 = 0.5\). It is observed from Figures. 2 to 14 fluid primary velocity \(u\) and fluid secondary velocity \(w\) attain a distinctive maximum value near the surface of the plate and then decrease properly on increasing boundary layer coordinate \(y\) to approach free stream value. Fig. 2 depicts the influence of hall current parameter \(m\) on the primary velocity \(u\) and secondary velocity \(w\).

It is evident from Fig. 2, \(u\) increases on increasing \(m\) in a region near to the plate and it decreases on increasing \(m\) in the region away from the plate whereas \(w\) increases on increasing \(m\) throughout the boundary layer region. This implies that, hall current tends to accelerate secondary fluid velocity throughout the boundary layer region which is consistent with the fact that hall current induces secondary flow in the flow field. Hall current tends to accelerate primary fluid velocity in a region close to the plate whereas it has a reverse effect on fluid primary velocity in the region away from the plate. It is perceived from Fig. 3 that, primary velocity \(u\) decreases on increasing rotation parameter \(K^2\) whereas secondary velocity \(w\) increases on increasing \(K^2\) in the region near to the plate and it decreases on increasing \(K^2\) in the region away from the plate. This implies that, rotation tends to retard primary fluid velocity throughout the boundary layer region. Although rotation is known to induce secondary fluid velocity
in the flow-field by suppressing the primary fluid velocity, its accelerating effect is prevalent only in the region near to the plate whereas it has a reverse effect on secondary fluid velocity in the region away from the plate. This is due to the reason that Coriolis force is dominant in the region near to the axis of rotation. Fig. 4 depicts the effect of radiation parameter \( R \) on the primary velocity \( u \) and secondary velocity \( w \) of the flow field.

The radiation parameter \( R \) is found to decelerate both the primary velocity \( u \) and secondary velocity \( w \) of the flow field at all points. This is because radiation parameter have tendency to reduce fluid temperature which is clearly evident from Fig. 16. Higher the radiation parameter, the more sharper is the reduction in velocity.
It is observed that from Fig. 5 both $u$ and $w$ decreases on increasing heat absorption parameter $\phi$. This implies that heat absorption tend to retard the primary and secondary fluid velocities.

![Figure 5. Influence of $\phi$ on velocity profiles.](image)

This is because heat absorption have tendency to reduce fluid temperature which is clearly evident from Fig. 15. The effect of Grashof numbers for heat transfer $G_r$ and mass transfer $G_m$ on the primary velocity $u$ and secondary velocity $w$ of the flow field are presented in Figs. 6 and 7.

![Figure 6. Influence of $G_r$ on velocity profiles.](image)  
![Figure 7. Influence of $G_m$ on velocity profiles.](image)

A study of the curves shows that the Grashof numbers for heat transfer $G_r$ and mass transfer $G_m$ accelerate the primary velocity $u$ and secondary velocity $w$ of the flow field at all points. Comparing the curves of Figs. 6 and 7, it is further observed that the increase in velocity of the
flow fields more significant in presence of mass transfer. Thus, mass transfer has a dominant effect on the flow field.

The nature of primary velocity $u$ and secondary velocity $w$ in presence of foreign species such as Hydrogen ($S_c = 0.20$), Helium ($S_c = 0.30$), Water vapour ($S_c = 0.60$), Ammonia ($S_c = 0.78$) is shown in Fig. 8. The flow field suffers a decrease in primary velocity $u$ and secondary velocity $w$ at all points in presence of heavier diffusing species. It is observed from Fig. 9 primary velocity $u$ and secondary velocity $w$ increase on increasing time $t$. This implies that primary and secondary fluid velocities are getting accelerated with the progress of time.

Fig. 10 displays the effect of the chemical reaction parameter $K_r$ on primary velocity $u$ and secondary velocity $w$ of the flow field at all points. It is seen, that primary velocity $u$ and secondary velocity $w$ decreases with increasing the chemical reaction parameter $K_r$. Fig. 11
displays the effect of the Soret number $S_0$ on primary velocity $u$ and secondary velocity $w$ of the flow field at all points. It is seen, that primary velocity $u$ and secondary velocity $w$ increases with increasing the Soret number $S_0$.

Fig. 12 depicts the effect of Magnetic parameter $M$ on primary velocity $u$ and secondary velocity $w$ of the flow field. The Magnetic parameter $M$ is found to decrease the primary velocity $u$ and secondary velocity $w$ of the flow field at all points. Fig. 13 depicts the influence of porosity parameter $K_1$ on primary velocity $u$ and secondary velocity $w$ of the flow field. It is evident from Fig. 13, $u$ increases on increasing $K_1$ in a region near to the plate and it decreases on increasing $K_1$ in the region away from the plate whereas $w$ increases on increasing $K_1$ throughout the boundary layer region.

Fig. 14 shows the plot of primary velocity $u$ and secondary velocity $w$ of the flow field against different values of prandtl number $Pr$ taking other parameters are constant. The values
of the prandtl number are chosen for air \((P_r = 0.71)\), electrolytic solution \((P_r = 1.00)\), water \((P_r = 7.00)\) and water at \(4^{0}C\) \((P_r = 11.40)\). It is observed that the primary velocity \(u\) and secondary velocity \(w\) of the flow field decreases in magnitude as prandtl number \(P_r\) increases. Thus higher prandtl number leads to faster cooling of the plate.

Figs. 15 and 16 depict the effect of heat absorption parameter \(\phi\) and radiation parameter \(R\) on the temperature \(T\) of the flow field. It is observed that heat absorption parameter \(\phi\) and radiation parameter \(R\) are found to decelerate the temperature \(T\) of the flow field at all point.

Figs. 17 and 18 depict the effect of time \(t\) and prandtl number \(P_r\) on the temperature \(T\) of the flow field. The prandtl number defines the ratio of momentum diffusivity to thermal diffusivity. The values of the prandtl number are chosen for air \((P_r = 0.71)\), electrolytic
solution \((P_r = 1.00)\), water \((P_r = 7.00)\) and water at \(40^\circ C\) \((P_r = 11.40)\). It is noticed that the fluid temperature \(T\) increases in magnitude on increasing time \(t\) and it decreases on increasing the prandtl number \(P_r\) at all points of the flow field.

It is observed from Figs. 19 and 20 that species concentration \(C\) decreases on increasing Schmidt number \(S_c\) whereas it increases on increasing time \(t\). This implies that there is an enhancement in species concentration with the progress of time throughout the boundary layer region.

Figs. 21 and 22 depict effect of Soret number \(S_o\) and chemical reaction parameter \(K_r\) on concentration distribution of the flow field. The species concentration \(C\) is found to increases on increasing Soret number \(S_o\) whereas it is decreases on increasing the chemical reaction parameter \(K_r\).
parameter $K_r$. Table 1. Represents the numerical values of heat transfer coefficient $N_u$ for different values of radiation parameter $R$, Prandtl number $P_r$ and heat absorption parameter $\phi$. Nusselt number $N_u$ increases on increasing $R$, $P_r$, and $\phi$. Also the value of $N_u$ is least for Mercury and highest for Water at $4^0C$.

### Table 1. Nusselt number $N_u$ when $t = 0.5$.

<table>
<thead>
<tr>
<th>$R$</th>
<th>$P_r$</th>
<th>$\phi$</th>
<th>$N_u$</th>
<th>$R$</th>
<th>$P_r$</th>
<th>$\phi$</th>
<th>$N_u$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>0.71</td>
<td>3.0</td>
<td>0.9421</td>
<td>2.0</td>
<td>0.71</td>
<td>3.0</td>
<td>2.9580</td>
</tr>
<tr>
<td>4.0</td>
<td>0.71</td>
<td>3.0</td>
<td>1.1147</td>
<td>2.0</td>
<td>11.40</td>
<td>3.0</td>
<td>3.7749</td>
</tr>
<tr>
<td>6.0</td>
<td>0.71</td>
<td>3.0</td>
<td>1.2639</td>
<td>2.0</td>
<td>0.71</td>
<td>1.0</td>
<td>0.7297</td>
</tr>
<tr>
<td>8.0</td>
<td>0.71</td>
<td>3.0</td>
<td>1.3973</td>
<td>2.0</td>
<td>0.71</td>
<td>3.0</td>
<td>0.9421</td>
</tr>
<tr>
<td>2.0</td>
<td>0.71</td>
<td>3.0</td>
<td>0.9421</td>
<td>2.0</td>
<td>0.71</td>
<td>5.0</td>
<td>1.1147</td>
</tr>
<tr>
<td>2.0</td>
<td>1.00</td>
<td>3.0</td>
<td>1.1180</td>
<td>2.0</td>
<td>0.71</td>
<td>7.0</td>
<td>1.2639</td>
</tr>
</tbody>
</table>

Table 2. Represents the numerical values of skin friction coefficient $C_f$ for different values of Hall current parameter $m$, rotation parameter $K^2$, radiation parameter $R$ and heat absorption parameter $\phi$. Skin friction coefficient $C_f$ increases on increasing $m$ whereas $C_f$ decreases on increasing $R$ and $\phi$ for both primary and secondary velocities. It is observed that skin

### Table 2. Skin friction coefficient when $K_r = 0.5, S_o = 0.5, M = 1, K_1 = 0.5, G_r = 4, G_m = 3, R_1 = 0.5, \omega = 0.5, S_c = 0.6, t = 0.5, P_r = 0.71$.

<table>
<thead>
<tr>
<th>$m$</th>
<th>$K^2$</th>
<th>$R$</th>
<th>$\phi$</th>
<th>Skin friction coefficient $C_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Primary</td>
</tr>
<tr>
<td>0.5</td>
<td>2.0</td>
<td>2.0</td>
<td>3.0</td>
<td>1.0138</td>
</tr>
<tr>
<td>1.0</td>
<td>2.0</td>
<td>2.0</td>
<td>3.0</td>
<td>1.0232</td>
</tr>
<tr>
<td>1.5</td>
<td>2.0</td>
<td>2.0</td>
<td>3.0</td>
<td>1.0429</td>
</tr>
<tr>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>3.0</td>
<td>1.0600</td>
</tr>
<tr>
<td>0.5</td>
<td>2.0</td>
<td>2.0</td>
<td>3.0</td>
<td>1.0138</td>
</tr>
<tr>
<td>0.5</td>
<td>4.0</td>
<td>2.0</td>
<td>3.0</td>
<td>0.5745</td>
</tr>
<tr>
<td>0.5</td>
<td>6.0</td>
<td>2.0</td>
<td>3.0</td>
<td>0.2914</td>
</tr>
<tr>
<td>0.5</td>
<td>8.0</td>
<td>2.0</td>
<td>3.0</td>
<td>0.0857</td>
</tr>
<tr>
<td>0.5</td>
<td>2.0</td>
<td>2.0</td>
<td>3.0</td>
<td>1.0138</td>
</tr>
<tr>
<td>0.5</td>
<td>2.0</td>
<td>4.0</td>
<td>3.0</td>
<td>0.9870</td>
</tr>
<tr>
<td>0.5</td>
<td>2.0</td>
<td>6.0</td>
<td>3.0</td>
<td>0.9665</td>
</tr>
<tr>
<td>0.5</td>
<td>2.0</td>
<td>8.0</td>
<td>3.0</td>
<td>0.9501</td>
</tr>
<tr>
<td>0.5</td>
<td>2.0</td>
<td>1.0</td>
<td>3.0</td>
<td>1.0525</td>
</tr>
<tr>
<td>0.5</td>
<td>2.0</td>
<td>2.0</td>
<td>3.0</td>
<td>1.0138</td>
</tr>
<tr>
<td>0.5</td>
<td>2.0</td>
<td>5.0</td>
<td>3.0</td>
<td>0.9870</td>
</tr>
<tr>
<td>0.5</td>
<td>2.0</td>
<td>7.0</td>
<td>3.0</td>
<td>0.9665</td>
</tr>
</tbody>
</table>
friction coefficient $C_f$ decreases for primary velocity and increases for secondary velocity on increasing rotation parameter $K^2$.

Table 3. Represents the numerical values of skin friction coefficient $C_f$ for different values of Grashof number $G_r$, modified Grashof number $G_m$, Schmidt number $S_c$ and chemical reaction parameter $K_r$. Skin friction coefficient $C_f$ increases on increasing $G_r$ and $G_m$ whereas $C_f$ decreases on increasing $S_c$ and $K_r$ for both primary and secondary velocities.

<table>
<thead>
<tr>
<th>$G_r$</th>
<th>$G_m$</th>
<th>$S_c$</th>
<th>$K_r$</th>
<th>Skin friction coefficient $C_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Primary</td>
</tr>
<tr>
<td>1.0</td>
<td>3.0</td>
<td>0.60</td>
<td>0.5</td>
<td>0.6564</td>
</tr>
<tr>
<td>2.0</td>
<td>3.0</td>
<td>0.60</td>
<td>0.5</td>
<td>0.7755</td>
</tr>
<tr>
<td>3.0</td>
<td>3.0</td>
<td>0.60</td>
<td>0.5</td>
<td>0.8947</td>
</tr>
<tr>
<td>4.0</td>
<td>3.0</td>
<td>0.60</td>
<td>0.5</td>
<td>1.0138</td>
</tr>
<tr>
<td>4.0</td>
<td>2.0</td>
<td>0.60</td>
<td>0.5</td>
<td>0.6679</td>
</tr>
<tr>
<td>4.0</td>
<td>4.0</td>
<td>0.60</td>
<td>0.5</td>
<td>1.3597</td>
</tr>
<tr>
<td>4.0</td>
<td>6.0</td>
<td>0.60</td>
<td>0.5</td>
<td>2.0516</td>
</tr>
<tr>
<td>4.0</td>
<td>8.0</td>
<td>0.60</td>
<td>0.5</td>
<td>2.7434</td>
</tr>
<tr>
<td>4.0</td>
<td>3.0</td>
<td>0.20</td>
<td>0.5</td>
<td>1.0503</td>
</tr>
<tr>
<td>4.0</td>
<td>3.0</td>
<td>0.30</td>
<td>0.5</td>
<td>1.0376</td>
</tr>
<tr>
<td>4.0</td>
<td>3.0</td>
<td>0.60</td>
<td>0.5</td>
<td>1.0138</td>
</tr>
<tr>
<td>4.0</td>
<td>3.0</td>
<td>0.78</td>
<td>0.5</td>
<td>1.0045</td>
</tr>
<tr>
<td>4.0</td>
<td>3.0</td>
<td>0.60</td>
<td>0.5</td>
<td>1.0138</td>
</tr>
<tr>
<td>4.0</td>
<td>3.0</td>
<td>0.60</td>
<td>1.0</td>
<td>0.9498</td>
</tr>
<tr>
<td>4.0</td>
<td>3.0</td>
<td>0.60</td>
<td>1.5</td>
<td>0.9047</td>
</tr>
<tr>
<td>4.0</td>
<td>3.0</td>
<td>0.60</td>
<td>2.0</td>
<td>0.8693</td>
</tr>
</tbody>
</table>

Table 4. Represents the numerical values of mass transfer coefficient $S_h$ for different values of Schmidt number $S_c$, chemical reaction parameter $K_r$, Soret number $S_o$, radiation parameter $R$, Prandtl number $P_r$ and heat absorption parameter $\phi$. Sherwood number $S_h$ increases on increasing $S_c$ and $K_r$. Sherwood number $S_h$ decreases on increasing $S_o$, $R$, $P_r$ and $\phi$. Also the value of $S_h$ is least for Hydrogen and highest for Propyl benzene.

Table 5. Represents the numerical values of skin friction coefficient $C_f$ for different values of Soret number $S_o$, Magnetic parameter $M$, permeability parameter $K_1$ and prandtl number $P_r$. Skin friction coefficient $C_f$ increases on increasing $S_o$ and $K_1$ whereas $C_f$ decreases on increasing $M$ and $P_r$ for both primary and secondary velocities.
### Table 4. Sherwood number $S_h$ when $t = 0.5$.  

<table>
<thead>
<tr>
<th>$S_c$</th>
<th>$K_r$</th>
<th>$S_0$</th>
<th>$R$</th>
<th>$P_r$</th>
<th>$\phi$</th>
<th>$S_h$</th>
<th>$S_c$</th>
<th>$K_r$</th>
<th>$S_0$</th>
<th>$R$</th>
<th>$P_r$</th>
<th>$\phi$</th>
<th>$S_h$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.20</td>
<td>0.5</td>
<td>0.5</td>
<td>2.0</td>
<td>0.71</td>
<td>3.0</td>
<td>0.0611</td>
<td>0.60</td>
<td>0.5</td>
<td>0.5</td>
<td>2.0</td>
<td>0.71</td>
<td>3.0</td>
<td>0.3288</td>
</tr>
<tr>
<td>0.30</td>
<td>0.5</td>
<td>0.5</td>
<td>2.0</td>
<td>0.71</td>
<td>3.0</td>
<td>0.0846</td>
<td>0.60</td>
<td>0.5</td>
<td>0.5</td>
<td>4.0</td>
<td>0.71</td>
<td>3.0</td>
<td>0.2793</td>
</tr>
<tr>
<td>0.60</td>
<td>0.5</td>
<td>0.5</td>
<td>2.0</td>
<td>0.71</td>
<td>3.0</td>
<td>0.1479</td>
<td>0.60</td>
<td>0.5</td>
<td>0.5</td>
<td>6.0</td>
<td>0.71</td>
<td>3.0</td>
<td>0.2361</td>
</tr>
<tr>
<td>0.78</td>
<td>0.5</td>
<td>0.5</td>
<td>2.0</td>
<td>0.71</td>
<td>3.0</td>
<td>0.1826</td>
<td>0.60</td>
<td>0.5</td>
<td>0.5</td>
<td>8.0</td>
<td>0.71</td>
<td>3.0</td>
<td>0.1972</td>
</tr>
<tr>
<td>0.60</td>
<td>0.5</td>
<td>0.5</td>
<td>2.0</td>
<td>0.71</td>
<td>3.0</td>
<td>0.1479</td>
<td>0.60</td>
<td>0.5</td>
<td>0.5</td>
<td>2.0</td>
<td>0.71</td>
<td>3.0</td>
<td>0.3288</td>
</tr>
<tr>
<td>0.60</td>
<td>1.0</td>
<td>0.5</td>
<td>2.0</td>
<td>0.71</td>
<td>3.0</td>
<td>0.4089</td>
<td>0.60</td>
<td>0.5</td>
<td>0.5</td>
<td>2.0</td>
<td>1.00</td>
<td>3.0</td>
<td>0.2783</td>
</tr>
<tr>
<td>0.60</td>
<td>1.5</td>
<td>0.5</td>
<td>2.0</td>
<td>0.71</td>
<td>3.0</td>
<td>0.6055</td>
<td>0.60</td>
<td>0.5</td>
<td>0.5</td>
<td>2.0</td>
<td>7.00</td>
<td>3.0</td>
<td>-0.0264</td>
</tr>
<tr>
<td>0.60</td>
<td>2.0</td>
<td>0.5</td>
<td>2.0</td>
<td>0.71</td>
<td>3.0</td>
<td>0.7692</td>
<td>0.60</td>
<td>0.5</td>
<td>0.5</td>
<td>2.0</td>
<td>11.40</td>
<td>3.0</td>
<td>-0.5082</td>
</tr>
<tr>
<td>0.60</td>
<td>0.5</td>
<td>0.2</td>
<td>2.0</td>
<td>0.71</td>
<td>3.0</td>
<td>0.4601</td>
<td>0.60</td>
<td>0.5</td>
<td>0.5</td>
<td>2.0</td>
<td>0.71</td>
<td>1.0</td>
<td>0.3885</td>
</tr>
<tr>
<td>0.60</td>
<td>0.5</td>
<td>0.4</td>
<td>2.0</td>
<td>0.71</td>
<td>3.0</td>
<td>0.3725</td>
<td>0.60</td>
<td>0.5</td>
<td>0.5</td>
<td>2.0</td>
<td>0.71</td>
<td>3.0</td>
<td>0.3288</td>
</tr>
<tr>
<td>0.60</td>
<td>0.5</td>
<td>0.6</td>
<td>2.0</td>
<td>0.71</td>
<td>3.0</td>
<td>0.2850</td>
<td>0.60</td>
<td>0.5</td>
<td>0.5</td>
<td>2.0</td>
<td>0.71</td>
<td>5.0</td>
<td>0.2793</td>
</tr>
<tr>
<td>0.60</td>
<td>0.5</td>
<td>0.8</td>
<td>2.0</td>
<td>0.71</td>
<td>3.0</td>
<td>0.1974</td>
<td>0.60</td>
<td>0.5</td>
<td>0.5</td>
<td>2.0</td>
<td>0.71</td>
<td>7.0</td>
<td>0.2361</td>
</tr>
</tbody>
</table>

### Table 5. Skin friction coefficient when $G_r = 4, G_m = 3, K_r = 0.5, \phi = 3, R = 2, m = 0.5, K^2 = 2, \omega = 0.5, R_1 = 0.5, S_c = 0.6, t = 0.5$.  

<table>
<thead>
<tr>
<th>$S_o$</th>
<th>$M$</th>
<th>$K_1$</th>
<th>$P_r$</th>
<th>Skin friction coefficient $C_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Primary</td>
</tr>
<tr>
<td>0.2</td>
<td>1.0</td>
<td>0.5</td>
<td>0.71</td>
<td>0.9865</td>
</tr>
<tr>
<td>0.4</td>
<td>1.0</td>
<td>0.5</td>
<td>0.71</td>
<td>1.0047</td>
</tr>
<tr>
<td>0.6</td>
<td>1.0</td>
<td>0.5</td>
<td>0.71</td>
<td>1.0229</td>
</tr>
<tr>
<td>0.8</td>
<td>1.0</td>
<td>0.5</td>
<td>0.71</td>
<td>1.0411</td>
</tr>
<tr>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>0.71</td>
<td>1.0138</td>
</tr>
<tr>
<td>0.5</td>
<td>2.0</td>
<td>0.5</td>
<td>0.71</td>
<td>0.8966</td>
</tr>
<tr>
<td>0.5</td>
<td>3.0</td>
<td>0.5</td>
<td>0.71</td>
<td>0.7895</td>
</tr>
<tr>
<td>0.5</td>
<td>4.0</td>
<td>0.5</td>
<td>0.71</td>
<td>0.6918</td>
</tr>
<tr>
<td>0.5</td>
<td>1.0</td>
<td>0.1</td>
<td>0.71</td>
<td>0.3167</td>
</tr>
<tr>
<td>0.5</td>
<td>1.0</td>
<td>0.2</td>
<td>0.71</td>
<td>0.7323</td>
</tr>
<tr>
<td>0.5</td>
<td>1.0</td>
<td>0.3</td>
<td>0.71</td>
<td>0.8926</td>
</tr>
<tr>
<td>0.5</td>
<td>1.0</td>
<td>0.4</td>
<td>0.71</td>
<td>0.9705</td>
</tr>
<tr>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>0.71</td>
<td>1.0138</td>
</tr>
<tr>
<td>0.5</td>
<td>1.0</td>
<td>1.00</td>
<td>0.9865</td>
<td>0.8579</td>
</tr>
<tr>
<td>0.5</td>
<td>1.0</td>
<td>7.00</td>
<td>0.8335</td>
<td>0.7937</td>
</tr>
<tr>
<td>0.5</td>
<td>1.0</td>
<td>11.40</td>
<td>0.8013</td>
<td>0.7853</td>
</tr>
</tbody>
</table>
5. Conclusion

An investigation of the effects of hall current, Soret number, chemical reaction parameter and rotation parameter on unsteady hydromagnetic natural convection flow with heat and mass transfer of a viscous, incompressible, electrically conducting and optically thick radiating fluid past an impulsively moving vertical plate embedded in a fluid saturated porous medium is carried out. Important findings are as follows:

- Hall current tends to accelerate secondary fluid velocity throughout the boundary layer region. Hall current tends to accelerate primary fluid velocity in a region close to the plate whereas it has a reverse effect on the primary fluid velocity in the region away from the plate.
- Rotation tends to retard primary fluid velocity throughout the boundary layer region and accelerate secondary fluid velocity only in the region near to the plate whereas it has a reverse effect on secondary fluid velocity in the region away from the plate.
- Thermal and concentration buoyancy forces tend to accelerate both the primary and secondary fluid velocities throughout the boundary layer region.
- Primary and secondary fluid velocities are getting accelerated with the progress of time throughout the boundary layer region.
- Heat absorption parameter, thermal radiation and Prandtl number tend to retard fluid temperature and there is an enhancement in fluid temperature with the progress of time throughout the boundary layer region.
- Mass diffusion and chemical reaction parameter tends to retard species concentration and there is an enhancement in species concentration with the progress of time and Soret number throughout the boundary layer region.
- Radiation parameter, heat absorption parameter, Schmidt number, chemical reaction parameter, magnetic parameter and Prandtl number tends to retard skin friction coefficient and there is an enhancement in skin friction coefficient on increasing Hall current parameter, Grashof number, modified Grashof number, Soret number and permeability parameter for both primary and secondary velocities. But skin friction coefficient decreases on increasing rotation parameter for primary velocity whereas it has a reverse effect for secondary velocity.
- Radiation parameter, Prandtl number and heat absorption parameter have tendency to increase the heat transfer coefficient.
- Schmidt number and chemical reaction parameter have tendency to accelerate the mass transfer coefficient. Soret number, radiation parameter, Prandtl number and heat absorption parameter retard the fluid concentration.

Acknowledgements

The first author is thankful to V. R. Siddhartha Engineering College, Kanuru, Vijayawada, Andhra Pradesh, India for providing necessary research facilities to publish this paper. The authors are thankful for the suggestions and comments of the referees, which have led to improvement of the paper.
REFERENCES


AN OPTIMIZATION APPROACH FOR COMPUTING A SPARSE MONO-CYCLIC POSITIVE REPRESENTATION

KYUNGSUP KIM

DEPARTMENT OF COMPUTER ENGINEERING, CHUNGNAK NATIONAL UNIVERSITY, DAEJEON, 305-764, KOREA
E-mail address: kskim@cse.cnu.ac.kr

ABSTRACT. The phase-type representation is strongly connected with the positive realization in positive system. We attempt to transform phase-type representation into sparse mono-cyclic positive representation with as low order as possible. Because equivalent positive representations of a given phase-type distribution are non-unique, it is important to find a simple sparse positive representation with lower order that leads to more effective use in applications. A Hypo-Feedback-Coxian Block (HFCB) representation is a good candidate for a simple sparse representation. Our objective is to find an HFCB representation with possibly lower order, including all the eigenvalues of the original generator. We introduce an efficient nonlinear optimization method for computing an HFCB representation from a given phase-type representation. We discuss numerical problems encountered when finding efficiently a stable solution of the nonlinear constrained optimization problem. Numerical simulations are performed to show the effectiveness of the proposed algorithm.

1. INTRODUCTION

The positive realization problem has been independently studied in the system theory and probability theory communities [1, 2, 3, 4]. The phase-type representation (so called, positive representation) is strongly connected with the positive realization in system theory [5]. The relation between the Laplace-Stieltjes transform (LST) of a probability distribution and a corresponding phase-type representation is similar to that between a transfer function and a corresponding state space realization.

Because equivalent positive representations of a given phase-type(PH) distribution are non-unique [6], finding smaller and simpler matrix representations for PH-distributions has become an important theoretical and practical issue. In modeling and simulation applications, the importance of simple sparse representations comes from the fact that the computational complexity of PH-distributed random-variate generation depends on the representation [7, 8]. Finding
smaller, simpler and structured representations with lower order can lead to more effective use of PH-distribution random number generators in stochastic modeling and simulation.

There are some important approaches for minimal and specially structured positive realization such as the triangular, bi-diagonal, mono-cyclic, and uni-cyclic representations [5, 9]. Every phase-type distribution has a mono-cyclic representation with larger order [9]. The mono-cyclic representation is a natural extension of a generator with a bi-diagonal matrix. The mono-cyclic representations are referred to as feed-back Erlang representations (or feed-back Coxian representations) due to their structures [9]. We introduce a Hypo-Feedback-Coxian Block (HFCB) representation covering Erlang, Coxian, Feedback Erlang (FE) and Feedback Coxian (FC) representations. These approach can be applied to the sparse positive realization of positive system since it has the same structure [3, 5]. Similar constructive methods of simple compact positive realizations for general transfer functions have been extensively discussed in the positive system [2, 4, 10, 11]. The main objective of this paper is to compute an HFCB positive representation with a lower order using nonlinear optimization method.

Finding the roots of polynomials is an important task for finding a proper HFCB representation. A coefficient relation function between the roots and coefficients of a polynomial can be used to find all of the roots of a given polynomial simultaneously and in parallel, instead of using a deflation method finding the roots one by one such as Laguerre’s and Newton-Raphson’s [12]. A constrained nonlinear optimization problem computing a sparse mono-cyclic representation is formulated by using coefficient relation function. We discuss the numerical method to find a optimal solution of the constrained nonlinear optimization problem depending on how to choose an initial value.

The rest of this paper is organized as follows. Section 2 provides some relevant background material, including definitions and preliminary results. In Section 3, various structured phase-type representations classified into acyclic and cyclic forms and their properties are considered. Numerical optimization methods for finding a sparse mono-cyclic representation with possibly small order are discussed in Section 4. The numerical experiments are presented in Section 5. Finally, Section 6 concludes the paper.

2. PHASE-TYPE DISTRIBUTIONS

Before proceeding, we introduce some basic notations. An \( n \times n \) non-negative matrix \( A \) is denoted by \( A \geq 0 \) if its entries are non-negative and at least one entry is positive. A strict positive matrix \( A = [a_{ij}] \) is denoted \( A > 0 \) if all entries \( a_{ij} > 0 \). We discuss the phase-type distribution for a non-negative random variable \( X \) in terms of a continuous-time Markov process. A continuous-time Markov process is defined on an \( n + 1 \) finite state space. The row vector \( \alpha \) gives the initial probability vector. A phase-type (PH) distribution is defined as the distribution of the time needed for absorption in a Continuous-Time Markov Chain (CTMC) with one absorbing state. If the \( n + 1 \) state is an absorbing state and all other states are transient, the infinitesimal generator matrix of the Markov chain in the form of the augmented matrix
tuple \((\bar{T}, \bar{b}, \bar{\alpha})\), called the augmented phase-type representation, can be defined by
\[
\begin{align*}
\bar{T} &= \begin{bmatrix} T & -T \mathbf{1} \\ 0 & 0 \end{bmatrix}, \\
\bar{b} &= \mathbf{e}_{n+1}, \\
\bar{\alpha} &= \begin{bmatrix} \alpha & \alpha_{n+1} \end{bmatrix},
\end{align*}
\] (2.1)
where \(\mathbf{0}\) refers to the column vector, row vector or matrix with all entries equal to zero in the case without ambiguity, \(\mathbf{e}_k\) is a \(k\)-th standard basis vector with zeros, except the \(k\)-th one, which is 1, and \(\mathbf{1}\) is the column vector with all entries being one. We can see that \(\alpha_{n+1} = 0\) if \(\alpha \mathbf{1} = 1\), and \(\alpha_{n+1} = 1 - \alpha \mathbf{1}\) otherwise. PH distributions are commonly represented by a vector-matrix tuple \((\alpha, T)\) that describes the transient part of the CTMC. The vector-matrix tuple \((\alpha, T)\) denotes a phase-type representation (or a phase-type generator) of a phase-type distribution if \(\alpha\) and \(T\) have following properties:
\[
T_{ii} < 0, \quad T_{ij} \geq 0 \quad \text{for} \quad i \neq j, \quad \alpha \mathbf{1} \leq 1, \quad \alpha \geq 0, \quad \text{and} \quad T \mathbf{1} \leq 0.
\]
Equivalently, a phase-type representation is regarded as the positive realization (representation) of positive system [5].

The probability density function (PDF), cumulative distribution function (CDF), and Laplace-Stieltjes Transform (LST) of the PDF, respectively, are defined by
\[
\begin{align*}
f(x) &= \alpha \exp(Tx)(-T \mathbf{1}) \\
F(x) &= 1 - \alpha \exp(Tx) \mathbf{1} \\
F^*(s) &= E(\exp(sX)) = \alpha(sI - T)^{-1}(-T \mathbf{1})
\end{align*}
\]
where \(E(X)\) means an expectation of a random variable \(X\).

The particular structured representation problem has been studied as an important work in the theory of phase-type representations [5]. Depending on their structures, we can typically divide a phase-type representation into two classes: an acyclic subclass and a cyclic subclass. In the concept of graph theory, the acyclic class consists of a vector-matrix pair \((\alpha, T)\), which the generator \(T\) represents an acyclic transition graph and \(\alpha\) is a positive vector.

The concepts of PH-simplicity and PH-majorization are useful tools in the study of PH-distributions [6, 13]. For a given \(T\), \(PH(T)\) denotes the set of all distributions with a phase-type representation \((\alpha, T)\). For two generators \(T\) and \(S\), \(S\) is said to PH-majorize \(T\) if \(PH(T) \subset PH(S)\). \(T\) is called by PH-simple if for any different \(\alpha\) and \(\beta\), \((\alpha, T)\) and \((\beta, T)\) represent different distributions. \(S\) PH-majorizes \(T\) if and only if there exists a non-negative matrix \(P\) such that
\[
TP = PS, \quad P \mathbf{1} = \mathbf{1}.
\] (2.2)

This means that \((\alpha, T)\) and \((\alpha P, S)\) represent the same phase-type distribution. We note that a non-negative condition of \(P\) is not necessary for the existence of a new phase-type representation \((\alpha P, S)\). The non-negative condition of \(P\) is replaced by \(\alpha P \geq 0\).

A multiset is defined by a collection of elements in which certain elements may occur more than once. Let \(|A|\) be the cardinality of \(A\). A multiset \(A \uplus B\) denotes an additive union (counting multiplicities) in multiset such that \(|A \uplus B| = |A| + |B|\).
3. Specially structured positive representations

3.1. Coxian representation. Any PH-distribution with a triangular representation \((\alpha, T)\) has a bi-diagonal representation \((\beta, S(\tilde{\lambda}))\) with the same order. Every phase-type representation whose LST has only real poles has an ordered Coxian representation with larger order than the original [13]. By using an optimization method, several algorithms have been introduced for finding a Coxian generator, \(S(\lambda)\) PH-majorizing a PH-generator [14].

We consider the computation problem for PH-majorizing \(S(\lambda)\) for a PH-generator \((\alpha, T)\) such that for a given vector \(\tilde{\lambda} = (\lambda_1, \cdots, \lambda_m)\), \(TP = PS(\tilde{\lambda})\) and a bi-diagonal matrix \(S(\tilde{\lambda})\) is defined by

\[
S(\tilde{\lambda}) = \begin{bmatrix}
-\lambda_1 & \lambda_1 & \cdots & 0 & 0 \\
0 & -\lambda_2 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & -\lambda_{m-1} & \lambda_{m-1} \\
0 & 0 & \cdots & 0 & -\lambda_m
\end{bmatrix}
\]

A spectral polynomial algorithm for computing a canonical bi-diagonal representation of a state space representation has been proposed [14]. We assume that all eigenvalues of \(T\) are nonzero and real-valued. The spectral polynomial algorithm is given as follows:

1. Find the descending ordered multiset \(\{-\lambda_1, -\lambda_2, \cdots, -\lambda_m\}\) including all the eigenvalues (counting multiplicities) of \(T\).
2. Let \(p_m = -T \mathbb{1} / \lambda_m\). For \(2 \leq k \leq m\), recursively, set \(k = k - 1\) and compute \(p_k\) such that

\[
p_k = (\lambda_{k+1} I + T)p_{k+1}/\lambda_k.
\]
3. Construct the bi-diagonal matrix \(S(\tilde{\lambda})\) for \(\tilde{\lambda} = (\lambda_1, \cdots, \lambda_m)\) and \(n \times m\) matrix \(P = [p_1 \ p_2 \ \cdots \ p_m]\).
4. Check that \(TP - PS(\tilde{\lambda}) = 0\) and \(\beta = \alpha P \geq 0\).

PH-generators with only real eigenvalues can be transformed into Coxian generators by the spectral algorithm [14]. By using the spectral polynomial algorithm, a Coxian representation is induced based on the next theorem.

**Theorem 3.1.** We assume that a vector-matrix pair \((\alpha, T)\) is a PH-generator such that all the eigenvalues of \(T\) are real and \(\alpha \geq 0\). Then, a Coxian representation \((\beta, S(\tilde{\lambda}))\) is an equivalent bi-diagonal representation of a vector-matrix pair \((\alpha, T)\) such that \(TP = PS(\tilde{\lambda})\), \(\mathbb{1} = P \mathbb{1}\) and \(\beta = \alpha P\).

3.2. Monocyclic positive representation. We attempt to find sparse representations representing phase-type distributions with complex and real poles. The presence of complex poles in LST implies the presence of backward transitions in the associated Markov chain. Several candidates for canonical sparse positive realization, such as feedback Erlang representation [13], mono-cyclic Coxian representation [9] and unicyclic representation [15, 16], have been introduced.
An $n \times n$ matrix $A$ is said to be reducible if there exists a permutation matrix $P$ such that $P^t A P = \begin{bmatrix} B & C \\ 0 & D \end{bmatrix}$ where $B$ and $D$ are square matrices and $P^t$ is the transpose of $P$. Otherwise, $A$ is said to be irreducible. Define $E_{ij}$ be the $n \times n$ matrix that has only a one in the position $(i, j)$ and zeros otherwise. If $A$ is irreducible and for any $a_{ij} \neq 0$, $A - a_{ij} E_{ij}$ is reducible, then $A$ is said to be nearly reducible. Suppose $A$ is an $n \times n$ nearly reducible matrix [17]. Then, there exist permutation matrices $P$ and $Q$ and an integer $s > 1$ such that

$$PAQ = \begin{bmatrix} A_1 & B_1 & \cdots & 0 & 0 \\ 0 & A_2 & \cdots & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & \cdots & A_{s-1} & B_{s-1} \\ B_s & 0 & \cdots & 0 & A_s \end{bmatrix}$$

(3.1)

where each $B_i$ has exactly one entry equal to one and each $A_i$ is nearly reducible. When all $B_i$’s are $1 \times 1$, this is a simplified form for nearly reducible matrices. We note that this form is a good candidate as a simple sparse representation of a phase-type distribution with complex poles.

The simplest cyclic representations are obtained by adding a feedback on an Erlang representation, which is called a Feedback Erlang (FE) representation or mono-cyclic representation [16, 9]. A new type of representation can be obtained by adding a feedback to the Coxian representation. It is specially denoted as a Feedback Coxian (FC) representation if the transition matrix of a vector matrix pair can be obtained by adding feedback to the Coxian representation. Thus, for a given vector $\vec{\lambda} = (\lambda_1, \cdots, \lambda_m)$ and $0 \leq z < 1$, an FC representation $(\alpha, M(\vec{\lambda}, z))$ is defined by

$$M(\vec{\lambda}, z) = \begin{bmatrix} -\lambda_1 & \lambda_1 & \cdots & 0 & 0 \\ 0 & -\lambda_2 & \cdots & 0 & 0 \\ & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -\lambda_{m-1} & \lambda_{m-1} \\ z\lambda_m & 0 & \cdots & 0 & -\lambda_m \end{bmatrix}$$

(3.2)

We can see that the matrix $M(\vec{\lambda}, z)$ is a simple nearly reducible matrix.

**Lemma 3.2.** Let $P$ be a permutation matrix and $\vec{\lambda}' = \vec{\lambda}P$. Then an FC representation $(\alpha, M(\vec{\lambda}', z))$ has the same eigenvalue multiset.

**Proof.** The characteristic polynomial $p(s) = \det(sI - M(\vec{\lambda}, z))$ is given by $p(s) = (s + \lambda_1)(s + \lambda_2) \cdots (s + \lambda_m) - \theta$ where $\theta = z\lambda_1\lambda_2 \cdots \lambda_m$. Since the characteristic polynomial is independent of the order of the entries of $\vec{\lambda}$, the FC representation $(\alpha, M(\vec{\lambda}', z))$ has the same eigenvalue multiset. □
We can resort \( \tilde{\lambda} \) in ascending order representing the given characteristic polynomial. FC representations have more free parameters than FE representations. We define a Hypo-Feedback-Coxian Block (HFCB) representation as a generalized form, covering Erlang, Coxian, FE and FC block representation.

**Definition 3.1.** A Hypo-Feedback-Coxian Block (HFCB) representation is defined as a transient generator having the following structure

\[
M = \begin{bmatrix}
M_1 & M_1^* & \cdots & 0 & 0 \\
0 & M_2 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & M_{J-1} & M_{J-1}^* \\
0 & 0 & \cdots & 0 & M_J \\
\end{bmatrix}
\]  

(3.3)

where \( M_i \) is a feedback Coxian \( M(\tilde{\lambda}, z) \) or Coxian \( S(\lambda) \) for each \( i \) and \( M_j^* \) is defined by

\[
M_j^* = (-M_j)\text{e}_1, \text{ where } \text{e}_1 \text{ is the first standard basis vector.}
\]

In the above definition, if cyclic \( M_i \) is an FE, it is denoted by a Hypo-Feedback-Erlang Block (HFEB) as a special case of HFCB [8]. It has been proven that every phase-type representation can be transformed into a mixture of HFEB representations [5, 9].

**Algorithm 1** A cyclic spectral polynomial algorithm

**Input:** \((\alpha, T)\)  
**Output:** \( P \) and \( \beta = \alpha P \)

1. Choose an appropriate HFCB generator \( M \) such that the eigenvalue multiset of \( M \) includes all eigenvalues of \( T \) counting multiplicities.
2. Give an initial column vector \( p_m = -T1 \).
3. \( \zeta_k \) is defined by the number of nonzero elements of the \( k \)-th column of \( M \) and \( c_{j,k} \) is the \((j,k)\)-th element of \( M \). For a given \( p_k \) and \( 1 < k \leq m \), recursively, set \( k = k - 1 \) and compute \( p_k \) as follows
   (a) if \( \zeta_k = 2 \),
   \[
p_k = (\lambda_{k+1}I + T)p_{k+1}/\lambda_k
   \]  
   (3.4)
   (b) if \( \zeta_k = 3 \) and \( c_{k+n',k} \neq 0 \),
   \[
p_k = ((\lambda_{k+1}I + T)p_{k+1} - c_{k+n',k+1}p_{k+n'})/c_{k,k+1}
   \]  
   (3.5)
4. Construct \( P = [p_1 \ p_2 \ \cdots \ p_m] \)
5. Finally, check that \( TP = PM \) and \( \alpha P \geq 0 \), and return \( \beta = \alpha P \) and \( P \).

We propose a cyclic spectral polynomial algorithm to compute a transform matrix \( P \) generating \( M \) from a representation \((\alpha, T)\) in Algorithm 1. Finally, check two facts: \( TP = PM \) and \( \alpha P \geq 0 \). In the next section, we will discuss how to find an appropriate HFCB generator \( M \) with lower order in the first step in Algorithm 1.
Theorem 3.3. We assume that a vector-matrix pair \((\alpha, T)\) is a PH-generator such that all the eigenvalues of \(T\) are real and complex numbers and \(\alpha \geq 0\). Then the matrix \(P\) satisfies that \(TP = PM\) and the eigenvalue multiset (counting multiplicities) of \(T\) is included all eigenvalues of the HFCB matrix \(M\) in Eq. (3.3).

Proof. First, we assume that an eigenvalue multiset (counting multiplicities) of \(T\) is included in the multiset of eigenvalues of the HFCB matrix \(M\). However, it is postponed how to find \(M\) satisfying the assumption to the next section. A matrix \(P\) is computed by Algorithm 1. We will show that we can obtain \(P\) such that \(TP = PM\). We have

\[
TP = \begin{bmatrix} Tp_1 & Tp_2 & \cdots & Tp_{m-1} & Tp_m \end{bmatrix}
\]

and set \(p_m = -T\mathbb{1}\). The last column of \(M\) has only two nonzero values. From Algorithm 1, we obtain

\[
Tp_m = \lambda_{m-1}p_{m-1} - \lambda_mp_m.
\]

For \(1 < k < m\), since \(k\)-th column of \(M\) has two or three nonzero elements, we have two cases

\[
Tp_k + 1 = c_{k,k+1}p_k - \lambda_{k+1}p_{k+1} + c_{k+n',k+1}p_{k+n'}\quad \text{or}\quad Tp_k + 1 = \lambda_kp_k - \lambda_{k+1}p_{k+1}
\]

in each step. For the first column of \(M\), we have two cases

\[
Tp_1 = -\lambda_1p_1 + c_{1+n',1}p_{1+n'}\quad \text{or}\quad Tp_1 = -\lambda_1p_1.
\]

Therefore, we can see that \(TP = PM\). \(\square\)

4. Optimization method for finding a sparse HFCB representation

We consider the properties of FE blocks and FC blocks. For a given FE block \(M(\bar{\lambda}, z)\) with \(\lambda_i = \lambda > 0\) and a constant \(\lambda\), the eigenvalue multiset of \(M(\bar{\lambda}, z)\) can be computed explicitly in a closed form [9].

Lemma 4.1. [9] Assume that we have a real root \(\mu_0 = -\lambda + \lambda z^{1/m}\) and a complex root \(\mu_1 = a + ib\). Then each \(n\)-th root \(\mu_k\) of its characteristic polynomial equation of an FE representation \(M(\bar{\lambda}, z)\) is simply obtained by

\[
\mu_k = -\lambda(1 - z^{1/m} \cos \frac{2k\pi}{m}) + iz^{1/m} \lambda \sin \frac{2k\pi}{m} \quad (4.1)
\]

for \(0 \leq k < m\), where \(\bar{\lambda} = (\lambda, \cdots, \lambda)\), and the parameters of the representation can uniquely be determined by

\[
\lambda = \frac{1}{2} \left( 2a - b \tan \frac{\pi}{m} + b \cot \frac{\pi}{m} \right) \quad (4.2)
\]

\[
z = \frac{b(\tan \frac{\pi}{m} + \cot \frac{\pi}{m})}{2\lambda} \quad (4.3)
\]

for a given \(\mu_1\) and an integer \(m > 2\).

For a complex root \(\mu_1 = a + ib\), we can obtain an FE block \(M(\bar{\lambda}, z)\) such that \(\bar{\lambda} = (\lambda, \cdots, \lambda)\) and \(z\) are defined by Eqs (4.2) and (4.3). We consider a lower bound problem of the order of a phase-type generator \(T\) including a given complex eigenvalue. It is related to the continuous version of famous Kolmogorov’s remark: “for a fixed integer \(m\), what is the multiset of all complex numbers that are eigenvalues of an order-\(n\) stochastic matrix?” [13].
Lemma 4.2. [13] Let \((\alpha, T)\) be a generator of phase-type distribution with a maximal real eigenvalue \(\mu_0\) and with any complex eigenvalue \(\mu_1 = -a + bi\) \((a > 0)\). Then the order \(m\) of \(T\) satisfies

\[
\frac{|b|}{a - |\mu_0|} \leq \cot \frac{\pi}{m}
\]

(4.4)

It has been shown that a wedge as in (4.4) contains all eigenvalues of \(T\) using an invariant polytope argument [13]. By using the result of Lemma 4.2, we can determine the lower bound of the order of an FC block including \(\mu_1\) because an FC block is regarded as a special case of a phase-type representation. Since the eigenvalue multiset of FE block computed by using Lemma 4.1 covers at most two or three eigenvalues of the original matrix \(T\), the total order of \(M\) gets larger. On other hand, because FC blocks have more free parameters, we can find an FC block including more common eigenvalues of \(T\) than that of FE block. So we can reduce the total order of \(M\).

The relationship of the roots of a polynomial equation and its coefficients has been studied in the complex analysis field [18]. It is closely related to a typical theorem of [18], which shows a connection between the roots of the polynomial and the generator \((\bar{\lambda}, z)\) of an FC matrix \(M(\bar{\lambda}, z)\).

Lemma 4.3. Assume that a diagonal vector \(\vec{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_m)\) of the mono-cyclic matrix \(M(\bar{\lambda}, z)\) in (3.2) lies in a circular region whose center is \(\bar{\lambda}\) and a given radius \(r > 0\).

1. Then, all of the roots \(\{\mu_1, \cdots, \mu_m\}\) of the polynomial equation \(p(s) = (s + \lambda_1)(s + \lambda_2) \cdots (s + \lambda_m) - \theta = 0\) lie on or within one of the \(n\) circles, which have the common radius \(r\) and whose centers, \(\bar{\mu}_k\)'s, are given by \(\bar{\mu}_k = -\bar{\lambda} + \exp\left(\frac{2ik\pi}{m}\right)|\theta|^{1/m}\), where \(\theta = z\lambda_1\lambda_2 \cdots \lambda_m\).

2. If these \(m\) circles are mutually external, each circle contains precisely one of the roots \(\{\mu_1, \cdots, \mu_m\}\) of the polynomial equation \(p(s)\).

3. If \(\bar{\lambda} = E(\lambda_k)\), then we have \(|\theta|^{1/m} \leq |\bar{\lambda}|^{1/2}\), and equality holds if and only if \(\lambda_k = \bar{\lambda}\) for all \(k\).

Proof. Proofs for the first and the second can be shown by using the results of [18]. The arithmetic mean is larger than or equal to the geometric mean and the equality only holds if \(\lambda_k = \bar{\lambda}\) for all \(k\). The third is trivial.

Lemma 4.3 shows the connection between \(\{\lambda_k\}\) in \(M(\bar{\lambda}, z)\) and the root multiset \(\{\mu_k\}\) of \(p(s)\). With the help of a numerical procedure, we can depict the feasible regions of the eigenvalues of \(M(\bar{\lambda}, z)\) for a given \(r\).

Example 4.1. By using a random number generator, for given \(0 < z < 1\) and \(r > 0\), we generate a random number vector \(\bar{\lambda}\) such that \(\bar{\lambda} - r \leq \lambda_k \leq \bar{\lambda} + r\) and \(\bar{\lambda} = E(\lambda_k)\). We compute all the roots \(\{\mu_k\}\) of the characteristic polynomial \(p(s)\) depending on a random number vector \(\bar{\lambda}\) with respect to \(z = 0.5\) and \(\bar{\lambda} = 2\). Figure 1 shows a relationship between the eigenvalues \(\{\mu_k\}\) of an FC block \(M(\bar{\lambda}, z)\) and a random number vector \(\bar{\lambda}\) for \(n = 5\) or \(n = 6\). In Fig. 1,
the green dots, the red dots and the blue dots denote the eigenvalues points corresponding to $0.8 \leq r \leq 1$, $1.12 \leq r \leq 1.4$, and $1.44 \leq r \leq 1.8$, respectively. The diamond points present the eigenvalues of an FE block computed by Eq. (4.1).

From Example 4.1, we can predict the moving pattern of the eigenvalues corresponding to the change of $(\vec{\lambda}, z)$. We can verify that there will be just one root in each circle with a given radius $r$, remarked in Lemma 4.3. From the simulation, we can see that the eigenvalue points $\{\mu_k\}$ are distributed to the inside direction from the vertices of the polygons.

An eigenvalue multiset $\mathcal{A} = \{\nu_1, \cdots, \nu_n\}$ (counting multiplicities) is a multiset of eigenvalues of $T$. An eigenvalue multiset $\mathcal{B} = \{\mu_1, \cdots, \mu_m\}$ (counting multiplicities) is the multiset of eigenvalues of the HFCB matrix $M$. Let $\vec{\nu} = (\nu_1, \cdots, \nu_n)$ be an eigenvalue vector such that $\nu_i$ for each $i$ is an eigenvalue of $T$. The multiset $\mathcal{B}$ must include the eigenvalue multiset $\mathcal{A}$. An eigenvalue multiset $\mathcal{B}^{(i)} = \{\mu_1^{(i)}, \cdots, \mu_m^{(i)}\}$ (counting multiplicities) denotes the eigenvalue multiset of an FC block $M(\vec{\lambda}^{(i)}, z^{(i)})$ for each $1 \leq i \leq r$. A vector pair $(\vec{\lambda}^{(i)}, z^{(i)})$ denotes a generator matrix of $M(\vec{\lambda}^{(i)}, z^{(i)})$ for each $1 \leq i \leq r$. Let a generator $\Upsilon = \{(\vec{\lambda}^{(1)}, z^{(1)}), \cdots, (\vec{\lambda}^{(r)}, z^{(r)})\}$ generate the HFCB matrix $M$. The problem can be reformulated as follows.

**Problem 4.1.** Find a generator $\Upsilon = \{(\vec{\lambda}^{(1)}, z^{(1)}), \cdots, (\vec{\lambda}^{(r)}, z^{(r)})\}$ to minimize the order of the HFCB matrix $M$ such that

$$\mathcal{A} \subset \mathcal{B} = \bigoplus_{i=1}^{r} \mathcal{B}^{(i)}$$

where $\mathcal{B}^{(i)}$ is the eigenvalue multiset of $M(\vec{\lambda}^{(i)}, z^{(i)})$. 

**Figure 1.** The location of the eigenvalues $\{\mu_k\}$'s of FE, and FC for orders $n = 5$ and $n = 6$. 

- (a) $n = 5$
- (b) $n = 6$
We note that the problem is closely related to the simultaneous root finding problem of a polynomial \([12]\). Finding the roots of polynomials is an important task for various areas of signal processing, such as filter and wavelet design, spectral estimation, phase unwrapping, and communication. The constrained relation between the roots and coefficients of a polynomial can be used to find all of the roots of a given polynomial simultaneously, instead of using sequential methods, such as Laguerre’s and Newton-Raphson’s \([12]\). Since \(p_i(s)\) the characteristic is polynomial of each FC block matrix \(M(\vec{X}^{(i)}, \vec{z}^{(i)})\), the characteristic function \(p(s)\) of the HFCB matrix \(M\) is \(p(s) = p_1(s)p_2(s) \cdots p_r(s)\).

**Lemma 4.4.** For any generator \((\vec{X}^{(i)}, \vec{z}^{(i)}) \in \Upsilon\), assume that \(B^{(i)}\) denotes the eigenvalue multiset (counting multiplicities) of the sub-block \(M(\vec{X}^{(i)}, \vec{z}^{(i)})\) of the HFCB matrix \(M\). Then it satisfies the coefficient relation equations such that each \(\phi_k\) is defined by

\[
\phi_1 \triangleq \sum_{k=1}^{m_i} \lambda_k^{(i)} - \sum_{k=1}^{m_i} (-\mu_k^{(i)}) = 0
\]

\[
\phi_2 \triangleq \sum_{k<j} \lambda_k^{(i)} \lambda_j^{(i)} - \sum_{i<j} (-1)^2 \mu_k^{(i)} \mu_j^{(i)} = 0
\]

\[
\vdots
\]

\[
\phi_{m_i} \triangleq (1 - z^{(i)})\lambda_1^{(i)} \cdots \lambda_{m_i}^{(i)} - (-1)^{m_i} \mu_1^{(i)} \cdots \mu_{m_i}^{(i)} = 0
\]

for each \(\mu_k^{(i)} \in B^{(i)}\).

**Proof.** We will show that for given generator \(\Upsilon\) and \(B^{(i)}\), the coefficient relation equations between roots and coefficients of a polynomial as \((4.6)\) can be defined. The characteristic polynomial \(p_i(s) = \det(sI - M(\vec{X}^{(i)}, \vec{z}^{(i)})))\) for each \(i\) is given by

\[
p_i(s) = (s + \lambda_1^{(i)})(s + \lambda_2^{(i)}) \cdots (s + \lambda_{m_i}^{(i)}) - \theta
\]

where \(\theta = z^{(i)}\lambda_1^{(i)} \lambda_2^{(i)} \cdots \lambda_{m_i}^{(i)}\) and \(m_i\) is an order of \(M(\vec{X}^{(i)}, \vec{z}^{(i)})\). The eigenvalue multiset of \(M(\vec{X}^{(i)}, \vec{z}^{(i)})\) is equal to the root multiset of the polynomial \(p_i(s)\). The polynomial \(p_i(s)\) can be rewritten in the form

\[
p_i(s) = \prod_{k=1}^{m_i} (s - \mu_k^{(i)}).
\]

All coefficient relation functions \(\phi_i\)’s can be defined by comparing with coefficients of \((4.7)\) and \((4.8)\) with respect to \((\vec{X}^{(i)}, \vec{z}^{(i)}) \triangleq (\lambda_1^{(i)}, \cdots, \lambda_{m_i}^{(i)}, \vec{z}^{(i)})\) and \(\vec{\mu}^{(i)} = (\mu_1^{(i)}, \mu_2^{(i)} \cdots, \mu_{m_i}^{(i)})\), respectively. Therefore we can obtain the relation equations, \(\phi_k\)’s, between roots and coefficients of a polynomial, comparing the coefficients induced from the eigenvalue multiset \(B^{(i)}\) with the coefficients of the characteristic polynomial \(p_i(s)\) of each FC block matrix \(M(\vec{X}^{(i)}, \vec{z}^{(i)})\).

\(B^{(i)}\) is partitioned into a fixed multiset \(F^{(i)}\) and a flexible multiset \(U^{(i)} = B^{(i)} - F^{(i)}\). Let us set \(\vec{s} = (\vec{X}^{(i)}, \vec{z}^{(i)})\). For each \(U^{(i)}\), we define a vector \(\vec{u}\) by \(\vec{u} = (u_1, \cdots, u_d)\) for \(u_k \in U^{(i)}\)
AN OPTIMIZATION APPROACH FOR COMPUTING A SPARSE MONO-CYCLIC POSITIVE REPRESENTATION 235

and \( d = |\mathcal{U}(i)| \) where \(|A|\) is defined by the cardinality of \( A \). A free parameter vector \( \vec{x} \) is defined by \( \vec{x} = (Re(\vec{u}), Im(\vec{u})) \). We define a fixed vector \( \vec{x}_c = (Re(\vec{v}), Im(\vec{v})) \), where \( \vec{v} = (v_1, \cdots, v_d) \) for each \( v_k \in \mathcal{F}(i) \).

For all integers with \( 1 \leq k \leq m_i \), the coefficient relation functions \( \phi_k \) can be defined in the form, \( \phi_k(\vec{x}, \vec{x}_c, \vec{s}) \). Set \( \vec{\phi} = (\phi_1, \cdots, \phi_{m_i}) \) as in (4.6). The coefficient relation vector function \( \vec{\phi} : \mathbb{R}^\tilde{n} \to \mathbb{R}^m \) represents smooth nonlinear functions defined by Eq. (4.6), where \( \tilde{n} \) is the dimension of the vector \( (\vec{x}, \vec{x}_c, \vec{s}) \). We formulate a constrained nonlinear optimization problem to compute an FC block \( M(\vec{\lambda}(i); \vec{s}(i)) \) whose eigenvalue multiset includes \( \mathcal{F}(i) \).

**Problem 4.2.** For each multiset \( \mathcal{F}(i) \), find a solution \((\vec{x}, \vec{s})\) of a nonlinear objective function \( f(\vec{x}, \vec{s}) \) subject to equation or inequality constraints such that

\[
\min_{\vec{x}, \vec{s}} \quad f(\vec{x}, \vec{s}) \\
\text{s.t.} \quad b_l \leq \vec{x} \leq b_u \\
\quad \quad 0 < \vec{s} \leq c_u \\
\quad \vec{\phi}(\vec{x}, \vec{x}_c, \vec{s}) = 0.
\]

for a fixed parameter \( \vec{x}_c \), where \( b_l \) is a lower bound vector, \( b_u \) is an upper bound vector, and the objective function \( f(\vec{x}, \vec{s}) \) in (4.9) is given by

\[
f(\vec{x}, \vec{s}) = -\sum_{k=1}^{m_i} \ln \lambda_k.
\]

Now we consider what reasonable constraints are given in the constrained optimization problem in (4.9). Because \( M(\vec{\lambda}(i); \vec{s}(i)) \) and \( T \) are phase-type, all the real values of eigenvalues are negative, i.e., \( Re(\mu_k(i)) < 0 \) for any \( \mu_k(i) \in \mathcal{B}(i) \). We can set the upper bounds zero \( b_u = 0 \) for the real numbers, because \( Re(\mu_k(i)) < 0 \). The other bounds, \( b_l \) and \( b_u \), are given in a proper bound from initial values for the algorithm to be convergence. We can see that the entries \( \vec{\lambda}(i) \) are positive and \( 0 \leq \vec{z}(i) < 1 \). The upper bound \( c_u \) for \( \vec{\lambda}(i) \) are given in an appropriate bound from initial values.

**Remark 4.5.**

- We can assume that first and second derivatives of the objective functions and constraints are available because \( \vec{\phi} \) consists of polynomials.
- We note that roots of the polynomial in (4.8) depend continuously on its coefficients. That is well-known as the use of Rouche’s Theorem [19]. A small change in the coefficients of \( p(s) \) causes only a small change in the roots of the polynomial.
- Using the result of Lemma 4.4, Problem 4.1 can be reformulated in the form of a constrained nonlinear optimization problem for the coefficient relation functions with respect to \((\vec{x}, \vec{s})\) in Problem 4.2. By solving a constrained nonlinear optimization in Problem 4.2, we find a generator \((\vec{\lambda}(i), \vec{z}(i))\) and \( \mathcal{U}(i) \) for a fixed \( \mathcal{F}(i) \).

Our objective is to find a HFCB representation with possible lower order, whose eigenvalue multiset includes all eigenvalues of \( T \). We will discuss numerical problems for finding a HFCB
representation with lower order by using a nonlinear optimization method. Our strategy is to obtain the optimal solutions by extending some partially known solutions recursively as follows:

1. Initially, compute an FE block $M(\vec{\lambda}^{(i)}, z^{(i)})$ such that $\mathcal{F}^{(i)} = \mathcal{A} \cap \mathcal{B}^{(i)}$ is nonempty, using Lemma 4.1.
2. We choose a target multiset $\mathcal{T}$ such that $\mathcal{A} \supset \mathcal{T} \supset \mathcal{F}^{(i)}$. Set $\mathcal{F}_k = \mathcal{F}^{(i)}$
3. We choose a new multiset $\mathcal{F}_k(\subset \mathcal{T})$ corresponding to $\vec{x}_c$ after $\vec{x}_c$ moves a little and solve Problem 4.2 for a new multiset $\mathcal{F}_k$, recursively. If Problem 4.2 is solved finally, then we can obtain a generator $(\vec{\lambda}^{(i)}, z^{(i)})$ such that $\mathcal{F}^{(i)} = \mathcal{T} = \mathcal{A} \cap \mathcal{B}^{(i)}$.
4. If there is more additional target multiset $\mathcal{T}$, then go to Step 2. Otherwise, go to the next step.
5. Finally, we obtain a $M(\vec{\lambda}^{(i)}, z^{(i)})$ whose eigenvalue multiset includes $\mathcal{F}^{(i)} = \mathcal{A} \cap \mathcal{B}^{(i)}$

We need to discuss the efficient numerical algorithm at the third step in the above method. While a nonlinear optimization method is a powerful tool to find an optimal solution, there are several potential difficulties to solve some problems such as divergence and instability. Most of the numerical methods for solving the nonlinearly constrained optimization problems meet a danger converging to a local minimal point or diverging, depending on initial values.

We propose a linear search algorithm in Algorithm 2. We can generate a sequence of iterates that converges to an optimal solution provided that the initial value is sufficiently close to that optimal solution. A basic idea is to make shift by following a straight line to a final point from an initial point knowing a solution, and find a solution of the next stage using the previous value as an initial value.

We choose a target multiset $\mathcal{T} \subset \mathcal{A}$, to which the multiset $\mathcal{F}_k$ expands in each iteration. Initially, assume that an optimal solution $\vec{x}_{c,0}$ of Problem 4.2 is already computed and can be used as an initial vector, because the $\mu_k$'s in (4.8) are explicitly computed by using Lemma 4.1. $n_d$ denotes the number of segments. It is related with the length of change. If the procedure fails, then in order to avoid diverging we can try to execute the algorithm again after increasing $n_d$ recursively until some threshold. For $0 \leq k \leq n_d - 1$, compute $\vec{x}_{c,k+1}$ such that

$$\vec{x}_{c,k+1} = \vec{x}_{c,k} + \vec{d}$$

where $\vec{d} = \frac{\vec{x}_{c,k} - \vec{x}_{c,0}}{n_d}$. Let $\mathcal{F}_k$ correspond to $\vec{x}_{c,k}$. We find $(\vec{x}, \vec{s})$ corresponding to $\mathcal{F}_k$ by solving Problem 4.2 recursively, until arriving at the final destination. If the multiset $\mathcal{F}_k$ arrives at the final destination multiset $\mathcal{T}$ successfully, then the adaptive process succeeds. We can obtain a vector $(\vec{x}, \vec{s})$ for a target multiset $\mathcal{T}$. Additionally, if the procedure fails, then, in order to avoid diverging, we can try to execute the algorithm again after increasing $n_d$ recursively until a given threshold.

Finally, we implement a complete algorithm transforming from a given phase-type representation to an HFCB representation $(\beta, \mathcal{M})$, which has as minimal order as possible in Algorithm 3. First, we compute an eigenvalue multiset $\mathcal{A}$ of $\mathcal{T}$. Recursively, we find an FC block $M(\vec{\lambda}^{(i)}, z^{(i)})$ including a chosen eigenvalue $\mu \in \mathcal{A}$. We typically have two cases: an FC block for a complex eigenvalue and a Coxian block for a real eigenvalue. For a complex eigenvalue
Algorithm 2 LineSearch($B, F, T, s$)

Input: $B, F, T, s$

Output: $\bar{s}, \text{Flag}, B$

1: Compute a vector $\bar{x}$ with respect to $U = B - F$
2: Compute an initial vector $\bar{x}_{c,0}$ with respect to $F$
3: Compute $\bar{x}_{c}$ with respect to a target multiset $T$
4: Set a segment number $n_d$
5: $\tilde{d} = \bar{x}_{c} - \bar{x}_{c,0}$
6: for $0 \leq k < n_d$ do
7: Set $\bar{x}_{c,k+1} = \bar{x}_{c,k} + \tilde{d}$
8: Compute a new solution $(\bar{x}, \bar{s})$ of Problem 4.2 for an old value $(\bar{x}, \bar{s})$ and $\bar{x}_{c,k+1}$
9: Compute $F$ corresponding to $\bar{x}_{c,k+1}$
10: if No solution then
11: return $\text{Flag} = \text{False}$
12: else
13: $\text{Flag} = \text{True}$
14: end if
15: Resort $\bar{\lambda}^{(i)}$ in ascending order
16: update $s$ using $\bar{\lambda}^{(i)}$
17: end for
18: $B$ is updated by using $(\bar{x}, \bar{x}_{c})$
19: return $\bar{s}, \text{Flag}, B$

$\mu$, an FC induced from $\mu$ possibly includes other elements of $A$. Choose a subset collection $\mathfrak{P} = \{S_1, \cdots, S_N\}$ such that $S_k \subset \Omega$, $S_k \cap F = \emptyset$, and $|S_k| \leq |S_j|$ for $k > j$. If $w \in S_k$, then $\bar{w} \in S_k$. Using $\mathfrak{P}$, we can choose a new target multiset $T \cup F$ as large as possible. Compute $(\bar{\lambda}^{(i)}, z^{(i)})$ corresponding to $T$ by applying the linear search algorithm. We can reduce the total order of the HFCB matrix $M$ depending on how to choose a target multiset $T$. The process continues until all complex numbers from $\Omega$ are removed. Next, for a selected real eigenvalue $\mu \in \Omega$, we compute $\bar{\lambda}^{(i)} = \mu, m_i = 1$ and $z = 0$, which generate the Coxian block. We then resort a generator $T$ in the ascending order of the minimal number of $\bar{\lambda}^{(i)}$. Finally, we compute an HFCB representation $M$ and $P$ using a generator $T$ such that $TP = PM$. We can obtain a transformation matrix $P$ and $M$ satisfying $TP = PM$.

Finally, we note the initial vector $\beta$ corresponding to the obtained mono-cyclic representation can be negative. Even though $T$ has a Coxian generator $S(\bar{\lambda})$ as Theorem 3.1, the positivity of $\beta$ is not guaranteed. The order of the resulting Coxian generator can be larger than that of the original one. The process for finding such $P$ and $S$ is called the Coxianization of PH-generator $T$. The Coxianization of a PH-generator with only real eigenvalues has been proven to be feasible; three numerical methods for Coxianization have been introduced and their performances were compared [14].
**Algorithm 3** Compute an HFCB representation.

**Input:** \((\alpha, T)\)

**Output:** \((\beta, M)\)

1. Compute the eigenvalue multiset \(\mathcal{A}\) of \(T\)
2. Index \(\mathcal{A}\) in descending order of their real values
3. \(\Omega = \mathcal{A}, i = 0\)
4. **while** \(\Omega\) is nonempty **do**
   5. \(i = i + 1\)
   6. **if** \(\Omega\) has any complex number **then**
      7. Select a complex number \(\mu \in \Omega\) with a maximal real value
      8. Choose a proper number \(m_i\) satisfying the inequality in (4.4)
      9. Compute \((\vec{\lambda}(i), z(i))\) and an eigenvalue multiset \(\mathcal{B} = \{\mu_0, \cdots, \mu_{m_i - 1}\}\) of \(M(\vec{\lambda}(i), z(i))\) by using Lemma 4.1.
   10. \(\mathcal{F} = \{\mu, \bar{\mu}\}\)
   11. Choose a subset collection \(\mathcal{P} = \{S_1, \cdots, S_N\}\) such that \(S_k \subset \Omega\), \(S_k \cap \mathcal{F} = \emptyset\), \(|S_k| \leq |S_j|\) for \(k > j\), and if \(w \in S_k\), then \(\bar{w} \in S_k\).
   12. **for** \(t = 1 : N\) **do**
      13. \(\mathcal{T} = \mathcal{F} \cup S_t\)
      14. \([\bar{s}, Flag, \mathcal{B}] = \text{LineSearch}(\mathcal{B}, \mathcal{F}, \mathcal{T}, \bar{s})\)
      15. **if** \(Flag == True\) **then**
        16. Compute \((\vec{\lambda}(i), z(i))\) using \(\bar{s}\), and insert it into \(\Upsilon\).
        17. **Break**
      18. **end if**
   19. **end for**
   20. **else**
      21. Select a maximal real value \(\mu \in \Omega\) with
      22. \(\mathcal{B} = \{\mu\}\)
      23. \(\vec{\lambda}(i) = \mu, m_i = 1\) and \(z_i = 0\).
   24. **end if**
25. \(\Omega = \Omega - \mathcal{B} \cap \mathcal{A}\)
26. **end while**
27. Resort a generator \(\Upsilon\) in the ascending order of the minimal number of \(\vec{\lambda}(i)\)
28. By using Algorithm 1, compute a HFCB representation \(M\) and \(P\) using a generator \(\Upsilon\) such that \(TP = MP\)
29. Compute \(\beta = \alpha P\)
30. **return** \((\beta, M)\)

To make the initial vector non-negative, additional states have to be added to the distribution similarly. If \(\beta\) contains at least one negative element then a further transformation is required to obtain a PH representation. We can extend the PH generator by adding an Erlang tail as
follows. For sufficiently large $\lambda$, we define a vector $\vec{\lambda} = (\lambda, \ldots, \lambda)$ with repeat times $N$. Set a new generator $M_1 = \begin{bmatrix} M & -M_1 \\ 0 & S(\vec{\lambda}) \end{bmatrix}$. We have $TP = PM_1$, $\mathbf{1} = P_1$ and $\beta = \alpha P$. For sufficiently large $N$, the matrix $\alpha P$ is non-negative.

5. NUMERICAL RESULTS AND DISCUSSION

We have implemented the proposed optimization methods to compute an HFCB representation $\beta, M$ by using Matlab. In order to solve constrained nonlinear problems, we use an interior point option in the function ‘fmincon’, consisting of sequential quadratic programming and trust region techniques [20]. Extensive numerical experiments have been carried out. The orders of the PH-generators are between 4 and 6 in our numerical simulation. This section presents typical numerical examples to illustrate the effects of the proposed method.

Example 5.1. Consider a phase-type distribution with the PH-representation $((1,0,0,0), T)$, where

$$
T = \begin{bmatrix}
-4 & 0 & 0 & 0 \\
1 & -4 & 0 & 2.7 \\
0 & 1.5 & -2 & 0 \\
0.1 & 0 & 0.9 & -1
\end{bmatrix}
$$

as introduced in [21]. By using Algorithm 3, we can obtain a $4 \times 4$ transformation matrix $P$ and $M$ satisfying $TP = PM$, where $M$ is constructed in the form (3.3) by using $\vec{\lambda}^{(1)} = [1.435, 1.435, 3.716]$ and $z^{(1)} = 0.611$, $\vec{\lambda}^{(2)} = 4.412$, $z^{(2)} = 0$ and

$$
P = \begin{bmatrix}
1.325 & 0.514 & 0.571 & 2.000 \\
1.685 & 0.954 & 41.472 & 0.300 \\
-0.146 & 2.910 & 1.148 & 0.500 \\
2.349 & 1.611 & 0.450 & 0
\end{bmatrix}.
$$

Because the computed matrix $P$ is not non-negative, the other constraint as $\alpha P \geq 0$ should be needed instead of $P \geq 0$. The dimension is less than that of [9]. Therefore, we obtain an FC representation $\alpha P, M(\vec{\lambda}, z)$ with the same order of 4, while Mocanu and Commault’s method [9] can transform an HFEB representation with an order of 5.

We note that the positivity of $P$ is not necessary for the existence of $\beta = \alpha P > 0$ from Example 5.1. The PH-majorization of $T$ is a too strong condition. Even for Coxianization algorithm, making $P$ a positive matrix enforces the order to increase [14]. If $\beta$ contains at least one negative element then a further transformation is required to obtain a monocyclic representation. We can resort the diagonal elements of the HFCB be in an ascending order. The ordering of blocks affects whether the vector $\beta$ will be non-negative. We can extend the monocyclic generator by adding an Erlang tail for sufficient large $\lambda$ similar to that of [8]. We give a counterexample condition for the nonexistence of $\beta = \alpha P > 0$. We present a counterexample condition, $\max(1^t M) > 0$ for $\beta > 0$ not to exist. In Algorithm 1, let us observe the recursion $p_k = ((\lambda_{k+1}I + T)p_{k+1} - c_{k+n', k+1}p_{k+n'})/c_{k,k+1}$ for $\max(1^t M) > 0$. 
When $p_{k+1}$ and $p_{k+n'}$ are positive, and $c_{k+n',k} > \lambda_k$ (or $1^T M > 0$), negative entries in $p_{k-1}$ can appear. Therefore, we cannot obtain $\alpha P \geq 0$ for any $\alpha \geq 0$. In the next example, we can verify different properties of two cases, $1^T M \leq 0$ and $1^T M > 0$ by using numerical simulations.

**Example 5.2.** Consider a phase-type distribution with the PH-representation $(\alpha, T)$ such that

$$T = \begin{bmatrix}
-6.0 & 0.6 & 0.6 & 0.6 & 3 \\
2 & -4.5 & 0 & 0 & 0.1 \\
0 & 2 & -3.5 & 0 & 0 \\
0 & 0 & 2 & -3 & 0 \\
0 & 0 & 0 & 2 & -2
\end{bmatrix}$$

While Mocanu and Commault’s method [9] can transform $T$ into an HFEB representation with order of 8, we can obtain a $5 \times 6$ matrix $P$ and an HFCB generator $M$ with an order of 6 by using Algorithm 3. We can see that the matrix $P$ is not non-negative. The proposed method can derive multiple solutions depending on a given initial value. For $\max(1^T M) \leq 0$, we could obtain $\alpha \geq 0$ such that $\alpha P \geq 0$ after attempting several times. For example, there are two typical cases after applying the proposed algorithm.

1. The first case has $\lambda^{(1)} = 1.120$, $z^{(1)} = 0$, $\lambda^{(2)} = (2.673, 2.678, 3.116, 4.49201, 6.366)$ and $z^{(2)} = 0.0591$. For $\max(1^T M) \leq 0$, we obtained $\alpha \geq 0$ such that $\alpha P \geq 0$ after several trials.

2. In the second case, we have $\lambda^{(1)} = 1.12084$, $z^{(1)} = 0$, $\lambda^{(2)} = (1.517, 3.660, 3.660, 3.660, 6.420)$ and $z^{(2)} = 0.096$. For $\max(1^T M) > 0$, we compute $P$ for several times. Typically, we can obtain $P$ such as

$$P = \begin{bmatrix}
-0.202 & 2.084 & 0.920 & 0.858 & 0.940 & 1.2 \\
-0.107 & 0.644 & 0.099 & 1.156 & 2.507 & 1.5 \\
-0.113 & 0.539 & 0.936 & 1.686 & 1.753 & 1.0 \\
-0.259 & 2.840 & 1.468 & 1.206 & 0.5463 & 0
\end{bmatrix}.$$ 

Because all entries of the first column are negative, we can see that $\min(\alpha P) < 0$ for any $\alpha \geq 0$.

In order to show that the proposed method has good performance, a large set of random PH generators samples were generated in the next example. The random matrices $T$ are dense, but the canonical representations obtained by the proposed method are sparse.

**Example 5.3.** A 31 samples of $T$ with order 5 and 4 complex eigenvalues are generated randomly. Then for each sample, the proposed method and Mocanu and Commault’s method [9] are applied to obtain HFCB and HFEB, respectively. For each algorithm, we determine $M$ and $\beta = \alpha P$ and check non-negativity of $\beta$. We compared the orders of the HFCB generators and HFEB generators. Let $d_{FC}$ and $d_{FE}$ be the order of HFCB and HFEB, respectively.
In Fig. 2, we plot $d_{FE} - d_{FC}$. We could see that the order of the solution obtained by using the proposed method is smaller than that of the solution obtained by using Mocanu and Commault’s method [9].

We note that there are several similar approaches to find compact PH representations with a unicyclic structure [15, 21]. A unicyclic block has more flexibility than FE block or FC block, since the representation has more free parameter and is denser. Since a unicyclic block has too many free parameters $(2m - 1)$ comparing to the number of equations $(m)$, it is impossible to obtain a stable optimal solution using numerical method. So a modified unicyclic block was introduced [15]. A modified unicyclic block approach reduces the free parameter number to $m$. Thus, the values of $m$ free parameters can be obtained by solving $m$ equations. When the given eigenvalue set frequently has no feasible solution, it cannot be applied. On the other hand, for the case, the proposed method can obtain a feasible solution by solving Problem 4.2.

Furthermore, the importance of simple sparse representations comes from the fact that the computational complexity of PH-distributed random-variate generation depends on the representation in simulation applications [8]. The FC block approach can lead to more effective random number generation of PH-distributions in stochastic modeling and analysis because the number of its feedback loop can be reduced.

6. Conclusion

We tried to transform a phase-type representation into a sparse HFCB representation with an order as low as possible. We introduced a numerical method for computing an HFCB representation by solving a nonlinear optimization problem. To find a sparse HFCB representation with lower order, including the multiset of eigenvalues of the original generator, we developed a numerical algorithm solving a constrained nonlinear optimization program. We discussed some numerical issues in efficiently finding a stable solution of the nonlinear constrained optimization problem. The proposed method can be applied to the sparse positive realization of positive system.
ACKNOWLEDGEMENTS

The author thanks the editor and the referees for their careful reading of the manuscript. This work was supported by research fund of Chungnam National University.

REFERENCES

A MASS LUMPING AND DISTRIBUTING FINITE ELEMENT ALGORITHM FOR MODELING FLOW IN VARIOUS SATURATED POROUS MEDIA

M. S. ISLAM

DEPARTMENT OF MATHEMATICS, SHAHJALAL UNIVERSITY OF SCIENCE & TECHNOLOGY, SYLHET-3114, BANGLADESH
E-mail address: sislam_25@yahoo.com

ABSTRACT. The Richards equation for water movement in unsaturated soil is highly nonlinear partial differential equations which are not solvable analytically unless unrealistic and oversimplifying assumptions are made regarding the attributes, dynamics, and properties of the physical systems. Therefore, conventionally, numerical solutions are the only feasible procedures to model flow in partially saturated porous media. The standard Finite element numerical technique is usually coupled with an Euler time discretizations scheme. Except for the fully explicit forward method, any other Euler time-marching algorithm generates nonlinear algebraic equations which should be solved using iterative procedures such as Newton and Picard iterations. In this study, lumped mass and distributed mass in the frame of Picard and Newton iterative techniques were evaluated to determine the most efficient method to solve the Richards equation with finite element model. The accuracy and computational efficiency of the scheme and of the Picard and Newton models are assessed for three test problems simulating one-dimensional flow processes in unsaturated porous media. Results demonstrated that, the conventional mass distributed finite element method suffers from numerical oscillations at the wetting front, especially for very dry initial conditions. Even though small mesh sizes are applied for all the test problems, it is shown that the traditional mass-distributed scheme can still generate an incorrect response due to the highly nonlinear properties of water flow in unsaturated soil and cause numerical oscillation. On the other hand, non oscillatory solutions are obtained and non-physics solutions for these problems are evaded by using the mass-lumped finite element method.

1. INTRODUCTION

The Richards equation is the governing equation for movement of water flow in partially saturated porous media, contains nonlinearity arising from pressure head dependencies in soil moisture and hydraulic conductivity. It is practically impossible to solve the equation analytically in unsaturated soil profiles with complex initial boundary conditions due to its highly nonlinear nature. Therefore numerical approximations are typically used to solve the unsaturated flow equations. The standard approximations that are applied to the spatial domain
are the finite difference and finite element methods are two popular schemes \cite{1, 2}. Variations of the standard finite difference and finite element methods, such as the sub domain finite element method have also been successfully used to solve these problems \cite{3}. The choice of using either finite difference or finite element method for the solution of variably saturated flow problems is largely personal and promoters of a particular method can easily support the strengths of their preferred solution approach. This can cause problems when a choice has to be made between these approaches, as it is difficult to independently evaluate the methods in a substantial technique.

Several studies have explored the finite element solution \cite{3, 4, 5, 6, 7}. Discussion about efficiency of mass types, element orders, and matrix solution methods, as well as influence of methods of numerical integration for the Richards equation are found few of these studies. The main reason of this is the consistent (distributed) mass scheme with high order elements has been typically accepted to be superior for solving complex nonlinear physical problems. Previous studies established that the numerical solution with consistent mass formulation shows oscillation \cite{3, 5, 4, 6} but they are not explained about the reason of oscillations and methods how to avoid the oscillations. Numerical oscillations is a significant factor to diverge the solution of Richards equation when simulate a sharp wetting front entering a dry soil profile. To overcome this difficulty, it is necessary to reduce the size of the element and time step size. But the reduction in element size could drastically increase the simulation time, as a result the numerical method becomes less attractive.

Numerical solution using pressure based formulation coupled with backward Euler time discretization is shown to produce unacceptably large mass balance errors for many example calculations. It is true for both finite difference and finite element approximation in space, although finite elements are generally inferior to finite differences. Because use of pressure head formulation with a simple time-stepping method is common, these findings appear to have significant practical implications. Almost, all unsaturated flow simulations use either the head-based or the moisture content-based formulation of Richards equation. A variety of finite difference and finite element solution techniques have been used with each of these equation forms \cite{3, 8, 9, 10, 11}. Numerical algorithm based on mixed form of Richards equation have been proposed and mass balance errors occur in standard head-based numerical solution \cite{12}.

Fully implicit backward Euler time approximation applied to the mixed form of Richards equation is mass conserving solution procedure for the unsaturated flow equation. Proper expansion of the time derivative produces a simple computational algorithm that is perfectly mass conservative for numerical approximation that preserves spatial symmetry. Thus the finite difference and finite element approximations using this mixed formulation are perfectly mass conservative. This approach is show to be superior to the standard head-based approximations while requiring no more computational effort. However conservation of mass is shown to be inadequate to guarantee good numerical solutions. For infiltration into dry soils, finite element approximations produce oscillatory solutions even while conserving mass. It is shown that diagonalization of the time matrix, which occurs naturally in finite difference approximations, is necessary and sufficient to guarantee oscillations free solutions. Mass conserving solution procedure is presented by the use of modified definition of capacity term to force global mass
balance and illustrated the importance of mass lumping in finite element solutions to unsaturated flow problems [5]. The significance of mass lumping is also shown and report good finite element solutions without mass lumping [3, 4, 10, 11]. Most of the studies are used a one-step Euler time-marching algorithm with the head-based version of Richards equation.

For stability reasons an implicit time discretization requiring evaluation of the nonlinear coefficients at the current time level, is generally used to solve the equation numerically. Newton and Picard iterative schemes are commonly used to linearize the resulting discrete system of equations, with the Picard scheme being the more popular of the two [3, 11, 13, 14, 15]. Picard method is the most intuitive linearization of Richards equation, computationally inexpensive on a per-iteration basis, and preserves symmetry of the discrete system of equations. However, some studies shown experimentally, the method may diverge under certain conditions [7, 10] and verified theoretically [16].

On the other hand, the Newton scheme, which is quadratically convergent [7, 17, 18] yields complex non symmetric system matrices and expensive than Picard linearization, and can be more robust than Picard for certain flow problems. The drawback of Newton’s method is that it is only locally convergent and involves the computation of derivatives. Although the use of the solution of the last time step to start the Newton iterations improves considerably the robustness of Newton’s method, in the degenerate case (saturated/unsaturated flow) the convergence is ensured only when a regularization step is applied and under additional constraints on the discretization parameters. While the Newton and the Picard schemes are generally robust, these iterative methods entail computational costs associated with having to evaluate and solve the system of equations repeatedly for each time step.

The objective of this study is to investigate the numerical behavior of finite element solution of Richards equation including the suitability of lumped mass and distributed mass. The work is also focused on determining the method which would offer a stable solution without requiring the resizing of the finite element mesh structure. Realistic initial and Dirichlet boundary conditions are imposed in the numerical simulator to the head-based form of Richards equation. To reduce the CPU time and maintain small truncation error, an adaptive time-stepping strategy is implemented. The nonlinear matrix equations are solved using the Picard and Newton iteration schemes. The performance of the algorithm is shown to be superior to the conventional pressure head-based form and can be easily used in layered porous media without any extraordinary treatment.

2. Governing Equation and Discretization

The partial differential equation describing fluid flow in partially saturated porous media, Richards equation, is obtained by combining Darcy’s law with the continuity equation. For one-dimensional vertical flow in unsaturated soils, pressure head-based Richards equation is written as

\[ C(\psi) \frac{\partial \psi}{\partial t} = \frac{\partial}{\partial z} \left( K(\psi) \left( \frac{\partial \psi}{\partial z} + 1 \right) \right) \]  

(2.1)
where, $\psi$ is the pressure head [L], $t$ is time [T], $z$ denotes the vertical distance from the soil surface assumed positive downward [L], $K(\psi)$ is the hydraulic conductivity [LT$^{-1}$], $C(\psi) = \frac{d\theta}{d\psi}$ is the specific fluid capacity [L$^{-1}$], $\theta$ is the volumetric water content.

The pressure head-based form can be successfully used in both saturated and unsaturated zones as well as in layered and composite porous materials. However, several studies have shown that it suffers from convergence difficulties and poor mass balance in modeling infiltration into very dry media unless very fine discretizations are used [5, 7, 14, 19, 20, 21, 22, 23] which, in turn, makes the computation very expensive. The time steps required for convergence are several orders of magnitude smaller than is required for reasonable temporal discretization [19]. The reason for these problems is highly nonlinear nature of the saturation-pressure function under dry initial conditions, causing very high fluid pressure gradient near the wetting front and huge computational cost.

Solution of Richards equation requires knowledge of hydraulic conductivity and fluid content versus fluid pressure head. These relationships are known as the hydraulic properties of the porous media. In this study, the most commonly used relationships are the van Genuchten [24] model. This model illustrated in detail as follows

$$\theta(\psi) = \theta_r + \frac{\theta_s - \theta_r}{[1 + |\alpha\psi|^n]^m} \text{ if } \psi \leq 0 \quad (2.2)$$
$$\theta(\psi) = \theta_s \text{ if } \psi > 0 \quad (2.3)$$
$$K(\psi) = K_s \left[ \frac{\theta - \theta_r}{\theta_s - \theta_r} \right]^{\frac{1}{2}} \left\{ 1 - \left[ 1 - \left( \frac{\theta - \theta_r}{\theta_s - \theta_r} \right)^{\frac{1}{m}} \right]^m \right\}^{2} \text{ if } \psi \leq 0 \quad (2.4)$$
$$K(\psi) = K_s \text{ if } \psi > 0 \quad (2.5)$$
$$c(\psi) = \alpha mn \frac{\theta_s - \theta_r}{[1 + |\alpha\psi|^n]^{m+1}} |\alpha\psi|^{n-1} \text{ if } \psi \leq 0 \quad (2.6)$$
$$c(\psi) = 0 \text{ if } \psi > 0 \quad (2.7)$$

Due to the nonlinear nature of the Richards equation (2.1), it should be solve numerically, finite element Galarkin discretization in space and a finite difference discretization of the time derivative term are used. To develop the finite element approximation of the pressure head-based Richards equation, the weak formulation of the dependent variable and the constitutive relations were approximated using interpolating polynomials [6, 25]. It was assumed that the hydraulic conductivity as well as capacitance varies linearly within each element [26]:

$$\psi(z,t) \approx \hat{\psi}(z,t) = \sum_{j=1}^{M} \psi_j(t) N_j(z) \quad (2.8)$$
$$K \approx \hat{K} = \sum_{j=1}^{M} K_j N_j(z) \quad (2.9)$$
\[ C \approx \hat{C} = \sum_{j=1}^{M} C_j N_j(z) \] (2.10)

where \( M \) is the number of elements, \( N_j(z) \) is the selected basis function, and \( \psi_j(t) \) is the associated and time-dependent unknown coefficients representing the solution of flow equation at nodes within the domain. The goal of the finite element approximation is to minimize the error and this can be accomplished by introducing the weight function, \( N_j(z) \), and setting operator \( L(\hat{\psi}) \) to be orthogonal to \( N_j(z) \):

\[
\int_{\Omega} L(\hat{\psi}) N_j(z) \, d\Omega = \int_{0}^{\Delta z} \left( \hat{C}(\psi) \frac{\partial \hat{\psi}(z,t)}{\partial z} - \frac{\partial}{\partial z} \left[ \hat{K}(\psi) \frac{\partial \hat{\psi}(z,t)}{\partial z} \right] + \frac{\partial \hat{K}(\hat{\psi})}{\partial z} \right) N_j(z) \, dz = 0
\] (2.11)

Performing integration by parts to reduce the second derivative and using fully implicit backward Euler time-marching algorithm with the solution is assumed to be known at time level \( n \) and unknown at time level \( n+1 \), one can discretize the time derivative in (2.11) to yield [27]:

\[
[A] \{\psi\}_{j}^{n+1} + [F] \left\{ \frac{\psi_{j}^{n+1} - \psi_{j}^{n}}{\Delta t} \right\} = \{B\} \quad (2.12)
\]

where:

\[
[A] = \int_{0}^{\Delta z} \hat{K} \frac{\partial N_i(z)}{\partial z} \frac{\partial N_j(z)}{\partial z} \, dz \quad (2.13)
\]

\[
[F] = \int_{0}^{\Delta z} \hat{C} N_i(z) N_j(z) \, dz \quad (2.14)
\]

\[
\{B\} = - \int_{0}^{\Delta z} \frac{\partial \hat{K}}{\partial z} N_j(z) \, dz \quad (2.15)
\]

One can found the detail evaluations of the above integrals and assembly of the global matrices in the literature [28]. The final results are highlighted herein, first for the stiffness mass matrix \([F]\):

\[
[F] = \frac{\Delta z}{12} \begin{pmatrix}
3C_1 + C_2 & C_1 + C_2 & 0 & \cdots & 0 \\
C_1 + C_2 & C_1 + 6C_2 + C_3 & C_2 + C_3 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & C_{N-2} + C_{N-1} & C_{N-2} + 6C_{N-1} + C_N \\
0 & 0 & \cdots & C_{N-1} + C_N & C_{N-1} + 3C_N
\end{pmatrix}
\]
Similarly matrix \([A]\) is assemble to form:

\[
[A] = \frac{1}{2\Delta z} \begin{pmatrix}
  K_1 + K_2 & -K_1 - K_2 & 0 & \cdots & 0 \\
-1K_1 - K_2 & K_1 + 2K_2 + K_3 & -K_2 - K_3 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
\vdots & \vdots & \cdots & -K_{N-2} - K_{N-1} & K_{N-2} + 2K_{N-1} + K_N \\
0 & 0 & \cdots & -K_{N-1} - K_N & K_{N-1} + K_N
\end{pmatrix}
\]

Finally the driving force vector \([B]\) can be assembled to produce:

\[
[B] = \frac{1}{2} \begin{pmatrix}
  K_1 + K_2 \\
  K_3 - K_1 \\
\vdots \\
  K_{i+1} - K_{i-1} \\
  K_N - K_{N-1}
\end{pmatrix} + \begin{pmatrix}
  q_0(t) \\
  0 \\
\vdots \\
  0 \\
  -q_L(t)
\end{pmatrix}
\]

where \(q_0(t)\) is the net flux at the upper end and \(q_L(t)\) is the imposed flux at the lower end of the spatial domain.

Upon substituting these matrices in (2.12) and rearranging the finite element approximation of Richards equation can be written in matrix form as [28]:

\[
\left[ (C_{i-1}^{n+1} + C_i^{n+1}) \frac{\Delta z}{12\Delta t} + \frac{(K_{i-1}^{n+1} + K_i^{n+1})}{2\Delta z} \right] \psi_{i-1}^{n+1} \\
+ \left[ (C_i^{n+1} + 6C_i^{n+1} + C_{i+1}^{n+1}) \frac{\Delta z}{12\Delta t} + \frac{(K_{i-1}^{n+1} + 2K_i^{n+1} + K_{i+1}^{n+1})}{2\Delta z} \right] \psi_i^{n+1} \\
+ \left[ (C_i^{n+1} + C_{i+1}^{n+1}) \frac{\Delta z}{12\Delta t} + \frac{(K_i^{n+1} + K_{i+1}^{n+1})}{2\Delta z} \right] \psi_{i+1}^{n+1} \\
= \left[ (C_{i-1}^{n+1} + 6C_i^{n+1} + C_{i+1}^{n+1}) \frac{\Delta z}{12\Delta t} \right] \psi_i^{n+1} - \frac{K_{i+1}^{n+1} - K_{i-1}^{n+1}}{2}
\]

This formulation is called the finite element approximation of the pressure head-based form Richards equation with distributed mass matrix.

The finite element formulation of flow equation with lumped mass matrix is as follows [27]:

\[
\left[ \frac{K_{i-1}^{n+1} + K_i^{n+1}}{2\Delta z} \right] \psi_{i-1}^{n+1} + \left[ \frac{\Delta z}{\Delta t} C_i^{n+1} + \frac{(K_{i-1}^{n+1} + 2K_i^{n+1} + K_{i+1}^{n+1})}{2\Delta z} \right] \psi_i^{n+1} \\
+ \left[ \frac{K_i^{n+1} + K_{i+1}^{n+1}}{2\Delta z} \right] \psi_{i+1}^{n+1} = \frac{\Delta z}{\Delta t} C_i^{n+1} \psi_i^{n} - \frac{K_{i+1}^{n+1} - K_{i-1}^{n+1}}{2}
\]
3. Iterative Methods

The Richards equation (2.12) can be written in the following vector form

\[ A \left( \Psi^{n+1} \right) \Psi^{n+1} + F \left( \Psi^{n+1} \right) \frac{\Psi^{n+1} - \Psi^n}{\Delta t} = q \left( t^{n+1} \right) - b \left( \Psi^{n+1} \right) \]  

(3.1)

3.1. Newton scheme. Let us consider

\[ f \left( \Psi^{n+1} \right) = A \left( \Psi^{n+1} \right) \Psi^{n+1} + F \left( \Psi^{n+1} \right) \frac{\Psi^{n+1} - \Psi^n}{\Delta t} - q \left( t^{n+1} \right) - b \left( \Psi^{n+1} \right) = 0 \]  

(3.2)

Here \( m \) stands for iteration index, so the Newton scheme [28] is

\[ f' \left( \Psi^{n+1}, m \right) h = -f \left( \Psi^{n+1}, m \right) \]  

(3.3)

where

\[ h = \Psi^{n+1,m+1} - \Psi^{n+1,m} \]  

(3.4)

and

\[ f'_{ij} = \lambda A_{ij} + \frac{1}{\Delta t^{n+1}} F_{ij} + \sum_s \frac{\partial A_{is}}{\partial \psi^{n+1}_j} \psi^{n+1}_s \]  

\[ + \frac{1}{\Delta t^{n+1}} \sum_s \frac{\partial F_{is}}{\partial \psi^{n+1}_j} \left( \psi^{n+1}_s - \psi^n_s \right) + \frac{\partial b_i}{\partial \psi^{n+1}_j} \]  

(3.5)

is the \( ij \) th component of the Jacobian matrix \( f' \left( \Psi^{n+1} \right) \).

3.2. Picard Scheme. The simple formulation of Picard scheme [28] can be obtained directly from (2.12) by iterating with all linear occurrences of \( \Psi^{n+1} \) taken at the current iteration level \( m + 1 \) and all nonlinear occurrences at the previous level \( m \). We get,

\[ \left[ A \left( \lambda \Psi^{n+1,m+1} \right) + \frac{1}{\Delta t} F \left( \Psi^{n+1,m+1} \right) \right] h = -f \left( \Psi^{n+1,m+1} \right) \]  

(3.6)

4. Numerical Simulations

All numerical simulations are done by CATHY (CATchment HYdrology) model that features elements of the sequential iterative coupling schemes. CATHY is a physically-based hydrological model where the surface module resolves the one-dimensional (1D) diffusion wave equation and the subsurface module solves the three-dimensional (3D) Richards equation. Coupling between these two equations is based on an extension of the boundary condition switching procedure used in some subsurface models for the handling of atmospheric inputs on the land surface boundary of the catchment. The main objective of this work is to assess, via sensitivity analysis, the accuracy, computational effort and mass balance limitations for the CATHY model over the frame of lumped mass and distributed mass along with the three tests of soil hydraulic parameters which make soil retention functions are highly nonlinear. For the case of convergence criterion, dynamic time stepping control is used to adjust step size of time during simulation according to the convergence behavior of the nonlinear iteration scheme. Nonlinear tolerance (\( tol = 10^{-3} \)) is specified for each time step, along with a maximum number of
iterations, \(maxit(= 10)\). The simulation begins with a time step size of \(\Delta t_0\) and proceeds until time \(T_{max}\). The current time step size is increased by a factor of \(\Delta t_{mag}(= 1.20)\) to a maximum size of \(\Delta t_{max}\) if convergence is achieved in fewer than \(maxit_1(= 5)\) iterations, it is remain unchanged if convergence required between \(maxit_1\) and \(maxit_2 (= 8)\) iterations, and it is decreased by a factor of \(\Delta t_{red}(= 0.5)\) to a minimum of \(\Delta t_{min}\) if convergence required more than \(maxit_2\) iterations. If convergence is not achieved within \(maxit\), the solution at the current time level is recomputed using a reduced time step size to the minimum time step size \(\Delta t_{min}\). For the first time step of simulation, the initial conditions are used as the first solution estimate for the iterative procedure. For subsequent time steps of simulation the pressure head solution from the previous step is used as the first estimate. Thus time step size has a direct effect on convergence behavior, via its influence on the quality of the initial solution estimate. Back-stepping is also triggered if linear solver failed or if the convergence or residual errors become larger than maximum allowable convergence or residual error in the nonlinear solution. In the nonlinear iterative methods, the infinity norm \((l_{\infty})\) of the convergence error is used as the termination criterion; that is, when \(\|\Psi^{n+1,m+1} - \Psi^{n+1,m}\| \leq tol\) is satisfied, convergence is achieved [29]. The residual error \((\|f(\Psi^{n+1,m})\|)\) is computed using \(l_{\infty}\) and \(l_2\) norms.

One measure of a numerical simulator is its ability to conserve global mass over the domain of interest. Satisfying the mass balance is necessary but not completely adequate prerequisite for a correct solution [7]. To measure the ability of the simulator to conserve mass, one of the most widely used criteria for evaluating the accuracy of a numerical scheme [7] is the mass balance error given by

\[
\text{Mass Balance Error} = \left| 1 - \frac{\text{Total additional mass in the domain}}{\text{Total net flux into the domain}} \right|
\]

where the additional mass in the domain is the difference between the mass measured at any instant \(t\) and the initial mass in the domain, and the total net flux into the domain is the flux balance integrated in time up to \(t\). For the finite element approximation, this is calculated by the following formula [7]:

\[
MB(t) = \frac{\sum_{i=1}^{E-1} (\theta_{i+1}^{n+1} - \theta_i^0) (\Delta z) + (\theta_0^{n+1} - \theta_0^0) \left(\frac{\Delta z}{2}\right) + (\theta_{E}^{n+1} - \theta_E^0) \left(\frac{\Delta z}{2}\right)}{\sum_{j=1}^{n+1} \left\{ (q_0 - q_N) (\Delta t) \right\}}
\]

with \(N = E + 1\) nodes \(\{z_0, z_1, z_2, \ldots, z_E\}\), and constant nodal spacing \(\Delta z\) is considered and \(q_0\) and \(q_N\) being boundary fluxes calculated from the finite element equations associated with the boundary modes \(z_0\) and \(z_N\).

5. RESULTS AND DISCUSSIONS

To assure the purposes of this work, sum up three one-dimensional test problems on which these investigations were based, examine methods for quantifying the efficiency, accuracy and applicability of the resultant solutions, and draw methods to evaluate the computational work.
required to achieve the results. A set of numerical experiments along with lumped and consistent mass was performed, to assess the robustness of the approach, to investigate methods for improving the efficiency of solutions to Richards equation and to evaluate the advantage of using the technique.

5.1. Test Problem 1. It is a benchmark test problem that has been previously examined [22, 23, 31, 32, 33]. The domain of this test problem is short and saturated conditions are not developed. To run CATHY model, two sets of nodal spacing are used (that is, 126 and 251 nodes are used). To assess the oscillation of the solution profile between mass lumping and distributed mass, very small time step \( \Delta t_{\text{min}} = 10^{-8} \text{s} \) is used. In this test case, constant pressure head boundary conditions are imposed, at the bottom and top of the soil column are \(-10m \) and \(-0.75m \) respectively. The initial pressure head is \(-10m \). Since the initial and boundary conditions are not consistent, so a steep gradient in the pressure head is setup. A 0.3m column of soil with van Genuchten parameters \( \theta_s = 0.368, \theta_r = 0.102, \alpha = 3.35/m, n = 2.0 \) and \( K_s = 7.970m/\text{day} \).

The pressure head profiles of the finite element results for lumped and distributed mass matrix for 126 and 251 nodes are shown in the FIGURE 1 and FIGURE 2 respectively. Note that Pic=Picard & New=Newton in FIGURES. The agreement is quite perfect if using the lumped mass for both the Picard and Newton iteration. Few numerical oscillations are produces at the bottom of the soil column based on the consistent mass scheme and it does not improve even with larger grid spacing. This dissimilarity occurs only for the handling of time (mass) matrices in the two solution procedures.

![Figure 1](image1.png)  
**Figure 1.** Pressure head profiles at 21600s for the Picard and Newton iterative schemes for the lumped and distributed mass of 125 layers.  

![Figure 2](image2.png)  
**Figure 2.** Pressure head profiles at 21600s for the Picard and Newton iterative schemes for the lumped and distributed mass of 250 layers.
Satisfactory cumulative mass balance errors and nonlinear iterations per time step for the Picard scheme are shown in FIGURE 3 and FIGURE 4 respectively. Similar behaviors are observed for the Newton iteration technique. The evaluated methods clearly outperform the predictable algorithm and these methods seem to handle this test case without any significant problems. It is also noted that the performance of the schemes in all cases is very similar to that of the published reports [23, 31, 32, 33]. The computational statistics of the methods under these conditions is shown in TABLE 1.

TABLE 1. Computational statistics of the Test Problem 1.
Note: Dist=Distributed, ite=iteration, CMBE=Cumulative mass balance error.

<table>
<thead>
<tr>
<th>No. of layers</th>
<th>Technique</th>
<th>Mass</th>
<th>No. of time steps</th>
<th>NL ite/ Time step</th>
<th>CMBE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Picard</td>
<td>Lump</td>
<td>2622</td>
<td>5.02</td>
<td>1.2954e-4</td>
</tr>
<tr>
<td>125</td>
<td></td>
<td>Dist</td>
<td>3783</td>
<td>5.02</td>
<td>5.9864e-5</td>
</tr>
<tr>
<td></td>
<td>Newton</td>
<td>Lump</td>
<td>6130</td>
<td>5.02</td>
<td>1.2940e-4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dist</td>
<td>9620</td>
<td>5.01</td>
<td>5.9181e-5</td>
</tr>
<tr>
<td>250</td>
<td>Picard</td>
<td>Lump</td>
<td>1372</td>
<td>5.02</td>
<td>1.1228e-4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dist</td>
<td>1360</td>
<td>5.02</td>
<td>3.7021e-5</td>
</tr>
<tr>
<td></td>
<td>Newton</td>
<td>Lump</td>
<td>3008</td>
<td>5.14</td>
<td>1.1114e-4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dist</td>
<td>4281</td>
<td>5.01</td>
<td>3.3256e-5</td>
</tr>
</tbody>
</table>
Graphical results and statistics of the simulation clearly indicate that the mass lumping technique is robust than the mass distributing technique for the numerical solution of Richards equation.

5.2. Test Problem 2. It is a very difficult vertical infiltration one-dimensional problem of 10m high soil column for the van Genuchten model with the soil parameters $\theta_s = 0.301$, $\theta_r = 0.093$, $\alpha = 5.47/m$, $n = 4.264$ and $K_s = 5.040m/days$. In order to evaluate the influence of the algorithm, two sets of grid spacing, one of them is very dense (e.g., $\Delta z = 0.05m$ and $0.025m$) and very small time step size (i.e., for $\Delta z = 0.05m$, $\Delta t_{min} = 10^{-5}s$ and for $\Delta z = 0.025 m$, $\Delta t_{min} = 10^{-15} s$) are considered and has been already analyzed in details [23, 30, 33]. It has constant head boundary conditions at both top ($\psi(10, t) = 0.1$) and bottom ($\psi(0, t) = 0.0$) boundaries and a hydrostatic equilibrium initial condition ($\psi(z, 0) = -z$). The set of conditions investigated represents range of medium and auxiliary conditions that are representative of a difficult class of infiltration problem frequently solved using Richards equation.

The pressure head profile from the solution for Test Problem 2 of 200 layers is shown in FIGURE 5. It is shown that mass distributed finite element method suffers from numerical oscillations at the wetting front but the lumped mass scheme offered oscillations free stable solution for dry initial conditions. The results hold for the iterative solution to the discretized version of the pressure head Richards equation including Picard and Newton techniques. To avoid the divergence, mesh and step size are reduced and FIGURE 6 represents pressure head solution of 400 layers with very small time step size $\Delta t_{min} = 10^{-15} s$ for mass lumped and distributed mass. Still, the consistent mass scheme of Picard and Newton methods exhibits
significant numerical errors ahead of the infiltration front but such oscillations are not present in the mass lumping formulation. Because the only difference between two solution procedures is the treatment of the time derivative term, these results imply that diagonalized time (mass) matrices are to be preferred. The cumulative mass balance errors plots are shown in

**FIGURE 6.** Pressure head profiles at 17280s for the Picard and Newton iterative schemes for the lumped and distributed mass of 400 layers.

**FIGURE 7.** Mass balance result for the Picard iterative scheme for the lumped and distributed mass of 400 layers.

**FIGURE 8.** Convergence result for the Picard iterative scheme for the lumped and distributed mass of 400 layers.

FIGURE 7 for lumped and distributed mass of Picard scheme of 400 layers case. It is evident
that excellent mass balance errors are shown for mass lumping case and this implies the numerical results are strictly maintained accuracy. Same error profile is obtained by the Newton method. FIGURE 8 is the graphical representation of the convergence behavior of the lumped and distributed mass of Picard scheme in terms of the number of nonlinear iterations required at each time step. It is found most striking here the very different behavior between the lumped and distributed schemes during the simulation period. Where the consistent mass scheme is forced to take very small step sizes from the beginning to the end of the simulation, that is, constants convergence oscillations are shown in the simulation of distributed mass procedure. A comparison of computational statistics, such as the cumulative mass balance errors, the total number of time steps, and the nonlinear iterations per time steps for the various runs for the two time matrices approaches, are tabulated in TABLE 2.

<table>
<thead>
<tr>
<th>No. of layers</th>
<th>Technique</th>
<th>Mass</th>
<th>No. of time steps</th>
<th>NL ite/Time step</th>
<th>CMBE</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>Picard</td>
<td>Lump</td>
<td>2109</td>
<td>5.49</td>
<td>5.3998e-2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dist</td>
<td>4679</td>
<td>5.06</td>
<td>6.0757e-2</td>
</tr>
<tr>
<td></td>
<td>Newton</td>
<td>Lump</td>
<td>1137</td>
<td>5.73</td>
<td>5.4068e-2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dist</td>
<td>764</td>
<td>5.49</td>
<td>6.1042e-2</td>
</tr>
<tr>
<td>400</td>
<td>Picard</td>
<td>Lump</td>
<td>4028</td>
<td>5.60</td>
<td>5.0243e-2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dist</td>
<td>9760</td>
<td>5.20</td>
<td>5.9236e-2</td>
</tr>
<tr>
<td></td>
<td>Newton</td>
<td>Lump</td>
<td>2539</td>
<td>5.56</td>
<td>5.0267e-2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dist</td>
<td>2072</td>
<td>5.42</td>
<td>5.9502e-2</td>
</tr>
</tbody>
</table>

From these statistics in TABLE 2 it can be concluded that all runs have adequate and comparable accuracy for the case of mass lumping. Thus the flow equation is one that benefits from mass lumping in finite element approximation.

5.3. Test Problem 3. This test study consists of vertical infiltration with redistribution [30, 35]. Current problem considers a one-dimensional soil column of 5m deep discretized for two sets of vertical resolution $\Delta z = 0.0250m$ and 0.0125m. Constant head boundary condition $\psi(0,t) = 0.0$ at the bottom of the domain and a time dependent boundary condition $\psi(10,t) = -10(1.0 - 1.01e^{-t})$ at the top of the domain with hydrostatic equilibrium initial conditions $\psi(z,0) = -z$ are applied. The time varying boundary condition yields a difficult two-front problem. This soil column is parameterized using the van Genuchten relationships with $\theta_s = 0.301$, $\theta_r = 0.093$, $\alpha = 5.47/m$, $n = 4.264.0$ and $K_s = 5.040m/days$.

FIGURE 9 and FIGURE 10 show the comparison of pressure head solution profiles for the cases of lumped and consistent mass of 200 and 400 layers respectively and the solution obtained with lumped mass is very similar to the previous studies [30, 36]. Picard and Newton iterative techniques with lumped and distributed mass schemes are achieved convergence with the minimum time step size $\Delta t_{min} = 10^{-18}days$. Run of Newton iteration of distributed
mass method had to forced stopped due to the time of simulation is significantly increased for both the grid spacing and it is true even for infinitesimal minimum time step $\Delta t_{\text{min}} = 10^{-20} \text{days}$. It is evident from these figures (FIGURE 9 and FIGURE 10) that there is a rapid infiltration of water from the surface, followed by a period of redistribution of the water due to the dynamic boundary condition at the top of the domain based on the mass lumping process. Mass-distributed scheme can still generate numerical oscillations due to the highly nonlinear properties of water flow in unsaturated soil. Similar numerical results hold for the larger grid spacing.

Figure 11 shows the cumulative mass balance errors profile of the Picard iterative scheme for the mass lumping and distributed mass of 400 layers. The magnitude of the errors is remarkably small for the lumped mass scheme throughout the entire domain of integration, attesting to the consistency of the mass lumping approximation. Similar results are obtained with respect to 201 nodes of spatial discretization, again confirming the robustness of the mass lumping mechanism. The number of iterations required for convergence per time step with the mass lumping scheme is compared to the distributed mass scheme in FIGURE 12. Lumped finite element solution converge rapidly, conversely the number of iterations for the distributed mass solution increases at the beginning of the simulation. The cumulative mass balance errors, number of time steps, nonlinear iterations per time step for different runs of the mass lumping method and their comparison with the mass distributed method are presented in TABLE 3 and it is clear that good performances are exhibited by the mass lumping algorithm. These results
are illustrative of numerical comparisons and indicate clearly that diagonal time matrices are superior to the distributed matrices.

### Table 3. Computational statistics of the Test Problem 3.
Note: Dist=Distributed, ite=iteration, CMBE=Cumulative mass balance error, NF=Simulation not finished.

<table>
<thead>
<tr>
<th>No. of layers</th>
<th>Technique</th>
<th>Mass</th>
<th>No. of time steps</th>
<th>NL ITE/ Time step</th>
<th>CMBE</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>Picard</td>
<td>Lump</td>
<td>2010</td>
<td>3.77</td>
<td>1.3358e-2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dist</td>
<td>3402</td>
<td>3.44</td>
<td>1.5059e-2</td>
</tr>
<tr>
<td></td>
<td>Newton</td>
<td>Lump</td>
<td>1736</td>
<td>3.61</td>
<td>1.3354e-2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dist</td>
<td>NF</td>
<td>NF</td>
<td>NF</td>
</tr>
<tr>
<td>400</td>
<td>Picard</td>
<td>Lump</td>
<td>3579</td>
<td>4.37</td>
<td>1.3419e-2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dist</td>
<td>6896</td>
<td>3.49</td>
<td>1.5226e-2</td>
</tr>
<tr>
<td></td>
<td>Newton</td>
<td>Lump</td>
<td>2865</td>
<td>4.13</td>
<td>1.3416e-2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dist</td>
<td>NF</td>
<td>NF</td>
<td>NF</td>
</tr>
</tbody>
</table>

### 6. Conclusion

A simple one-dimensional finite element model for flow equation into homogeneous soils was presented and the algorithm was implemented by applying the traditional iterative formulations, Picard and Newton based on lumped and distributed mass. Lumping scheme ensures
mass conservation and allows the handling of highly nonlinear phenomena. The similarity of number of time steps to complete the simulation and rate of convergence between lumped and distributed cases suggests that lumping formulation is generally superior to the distributing approximation. Presented results indicate that non-oscillatory numerical solution becomes very critical in solving problems of infiltration into dry soils with distributing mass. By the choice of mass lumping numerical method performs very well for unsaturated flow problems involving vertical infiltration as well as redistribution. According to the experiences learned from this study, very small time step size or very dense spatial discretization does not eliminate numerical oscillations when solving the highly nonlinear Richards equation based on distributed mass. The numerical evidence is complemented and explained by the fact that, based on the pressure head form of Richards equation, coupled with a lumped matrix, yields consistently reliable and robust for unsaturated flow problems, even for very dry initial conditions. Numerical evaluations are presented by this study confirmed that the suitable methodology for unsaturated flow problems is one that is based on the pressure head form of Richards equation, and uses a lumped form of the time matrix and can easily be extended to multidimensional problem in saturated-unsaturated regions in heterogeneous and layered porous media.

REFERENCES

NUMERICAL SOLUTION OF AN INTEGRO-DIFFERENTIAL EQUATION ARISING IN OSCILLATING MAGNETIC FIELDS

Kourosh Parand and Mehdi Delkhosh

1 Department of Computer Sciences, Shahid Beheshti University, Tehran, Iran.
E-mail address: k_parand@sbu.ac.ir
E-mail address: mehdidelkhosh@yahoo.com

2 Department of Cognitive Modelling, Institute for Cognitive and Brain Sciences, Shahid Beheshti University, Tehran, Iran.

Abstract. In this paper, an integro-differential equation which arises in oscillating magnetic fields is studied. The generalized fractional order Chebyshev orthogonal functions (GFCF) collocation method used for solving this integral equation. The GFCF collocation method can be used in applied physics, applied mathematics, and engineering applications. The results of applying this procedure to the integro-differential equation with time-periodic coefficients show the high accuracy, simplicity, and efficiency of this method. The present method is converging and the error decreases with increasing collocation points.

1. INTRODUCTION

In this section, Spectral methods and some basic definitions and theorems which are useful for our method have been introduced.

1.1. Spectral methods. Spectral methods have been developed rapidly in the past two decades. They have been successfully applied to numerical simulations in many fields, such as heat conduction, fluid dynamics, quantum mechanics, etc. These methods are powerful tools to solve differential equations. The key components of their formulation are the trial functions and the test functions. The trial functions, which are the linear combinations of suitable trial basis functions, are used to provide an approximate representation of the solution. The test functions are used to ensure that the differential equation and perhaps some boundary conditions are satisfied as closely as possible by the truncated series expansion. This is achieved by minimizing the residual function that is produced by using the truncated expansion instead of the exact solution with respect to a suitable norm [1, 2, 3, 4, 5, 6, 7, 8, 9, 10].

Received by the editors April 26 2016; Revised September 12 2016; Accepted in revised form September 12 2016; Published online September 20 2016.

2000 Mathematics Subject Classification. 34B05, 34K28, 78A35, 47G20.

Key words and phrases. Fractional order of the Chebyshev functions, Integro-differential equations, Oscillating magnetic fields, Collocation method, Integral equations.

† Corresponding author.
1.2. **Basical definitions.** In this section, some basic definitions and theorems which are useful for our method have been introduced [11].

**Definition 1.** For any real function \( f(t), t > 0 \), if there exists a real number \( p > \mu \), such that \( f(t) = t^p f_1(t) \), where \( f_1(t) \in C(0, \infty) \), is said to be in space \( C_\mu, \mu \in \mathbb{R} \), and it is in the space \( C^m_\mu \) if and only if \( f^n \in C_\mu, n \in \mathbb{N} \).

**Definition 2.** The fractional derivative of \( f(t) \) in the Caputo sense by the Riemann-Liouville fractional integral operator of order \( \alpha > 0 \) is defined as [12, 13]:

\[
D_\alpha f(t) = \frac{1}{\Gamma(m - \alpha)} \int_0^t (t - s)^{m-\alpha-1} D^m f(s) \, ds, \quad \alpha > 0,
\]

for \( m - 1 < \alpha \leq m, m \in \mathbb{N}, t > 0, m \) is the smallest integer greater than \( \alpha \), and \( f \in C^m_\alpha \).

Some properties of the operator \( D_\alpha \) are as follows. For \( f \in C_\mu, \mu \geq -1, \alpha, \beta \geq 0, \gamma \geq -1, N_0 = \{0, 1, 2, \ldots \} \) and constant \( C \):

1. \( D_\alpha C = 0 \),
2. \( D_\alpha D_\beta f(t) = D_{\alpha + \beta} f(t) \),
3. \( D_\alpha t^\gamma = \begin{cases} 
0, & \gamma \in N_0 \text{ and } \gamma < [\alpha], \\
\Gamma(\gamma + 1) t^{\gamma - \alpha}, & \gamma \in N_0 \text{ and } \gamma \geq [\alpha] \text{ or } \gamma \notin N \text{ and } \gamma > [\alpha], 
\end{cases} \)
4. \( D_\alpha (\sum_{i=1}^n c_i f_i(t)) = \sum_{i=1}^n c_i D_\alpha f_i(t), \quad \text{where} \ c_i \in \mathbb{R}. \)

**Definition 3.** Suppose that \( f(t), g(t) \in C[0, \eta] \) and \( w(t) \) is a weight function, then

\[
\| f(t) \|^2_w = \int_0^\eta f^2(t) w(t) \, dt,
\]

\[
\langle f(t), g(t) \rangle_w = \int_0^\eta f(t) g(t) w(t) \, dt.
\]

**Theorem 1.** (Generalized Taylor’s formula) Suppose that \( f(t) \in C[0, \eta] \) and \( D^{k+\alpha} f(t) \in C[0, \eta], \) where \( k = 0, 1, \ldots, m, 0 < \alpha \leq 1 \) and \( \eta > 0 \). Then we have

\[
f(t) = \sum_{i=0}^{m-1} \frac{t^i \alpha}{\Gamma(i \alpha + 1)} D^{i \alpha} f(0^+) + \frac{t^m \alpha}{\Gamma(m \alpha + 1)} D^{m \alpha} f(\xi), \quad \text{with } 0 < \xi \leq t, \ \forall t \in [0, \eta].
\]

And thus

\[
|f(t) - \sum_{i=0}^{m-1} \frac{t^i \alpha}{\Gamma(i \alpha + 1)} D^{i \alpha} f(0^+)| \leq M_\alpha \frac{t^m \alpha}{\Gamma(m \alpha + 1)}, \quad \text{where } M_\alpha = |D^{m \alpha} f(\xi)|.
\]

**Proof:** See Ref. [14].

The organization of the paper is expressed as follows: in section 2, the mathematical preliminaries to the problem is expressed. In section 3, the GFCFs and their properties are obtained.
In section 4, the work method is explained. Applications of the proposed method are shown in section 5. Finally, a brief conclusion is given in the last section.

2. Mathematical Preliminaries

The integro-differential equation [15, 16]

\[
\frac{d^2 y}{dt^2} = g(t) - a(t)y(t) + b(t) \int_0^t \cos(w_p s) y(s) ds
\]  

(2.1)

where \(a(t)\), \(b(t)\) and \(g(t)\) are given periodic functions of time may be easily found in the charged particle dynamics for some field configurations. Taking for instance the three mutually orthogonal magnetic field components \(B_x = B_1 \sin(w_p t)\), \(B_y = 0\), \(B_z = B_0\), the nonrelativistic equations of motion for a particle of mass \(m\) and charge \(q\) in this field configuration are

\[
m \frac{d^2 x}{dt^2} = q \left( B_0 \frac{dy}{dt} \right),
\]

(2.2)

\[
m \frac{d^2 y}{dt^2} = q \left( B_1 \sin(w_p t) \frac{dz}{dt} - B_0 \frac{dx}{dt} \right),
\]

(2.3)

\[
m \frac{d^2 z}{dt^2} = q \left(-B_1 \sin(w_p t) \frac{dy}{dt} \right).
\]

(2.4)

By integration of (2.2) and (2.4) and replacement of the time first derivatives of \(z\) and \(x\) in (2.3) one has (2.1) with

\[
a(t) = w_c^2 + w_f^2 \sin^2(w_p t),\quad b(t) = w_f^2 w_p \sin(w_p t),
\]

(2.5)

\[
g(t) = w_f \sin(w_p t) z'(0) + w_c^2 y(0) + w_c x'(0),
\]

(2.6)

where \(w_c = qB_0/m\) and \(w_f = qB_1/m\). Making the additional simplification that \(x'(0) = 0\) and \(y(0) = 0\), equation (2.1) is finally written as [16]

\[
\frac{d^2 y}{dt^2} = w_f \sin(w_p t) z'(0) - \left(w_c^2 + w_f^2 \sin^2(w_p t)\right) y(t)
\]

\[+ \left(w_f^2 w_p \sin(w_p t)\right) \int_0^t \cos(w_p s) y(s) ds
\]

(2.7)

In this study, we consider the equation (2.1) with the following initial conditions

\[
y(0) = \beta_0, \quad y'(0) = \beta_1
\]

(2.8)

There are methods to solve this equation, such as, He’s Homotopy perturbation method [16], Chebyshev wavelet [17], Legendre multi-wavelets [18], Local polynomial regression [19], Shannon wavelets [20], Variational iteration method [21] and Homotopy analysis method [22].

In this paper, we attempt to introduce a new method based on the generalized fractional order Chebyshev orthogonal functions (GFCFs) of the first kind for solving the equation (2.1) with the initial conditions (2.8).
3. THE GENERALIZED FRACTIONAL ORDER CHEBYSHEV FUNCTIONS

In this section, first, the generalized fractional order of the Chebyshev functions (GFCF) have been defined, and then some properties and convergence of them for our method have been introduced.

3.1. The Chebyshev functions. The Chebyshev polynomials have been used in numerical analysis, frequently, including polynomial approximation, Gauss-quadrature integration, integral and differential equations and spectral methods. Chebyshev polynomials have many properties, for example orthogonal, recursive, simple real roots, complete in the space of polynomials. For these reasons, many researchers have employed these polynomials in their research [23, 24, 25, 26, 27, 28].

Using some transformations, the number of researchers extended Chebyshev polynomials to semi-infinite or infinite domains, for example by using \( x = \frac{t - L}{t + L}, L > 0 \) the rational functions introduced [29, 30, 31, 32, 33, 34].

In the proposed work, by transformation \( z = \frac{1}{1 - 2(\frac{t}{\eta})^{\alpha}}, \alpha > 0 \) on the Chebyshev polynomials of the first kind, the fractional order of the Chebyshev orthogonal functions in the interval \([0, \eta]\) have been introduced, that they can use to solve these integro-differential equations.

3.2. The GFCFs definition. The efficient methods have been used by many researchers to solve the differential equations (DE) is based on the series expansion of the form \( \sum_{i=0}^{n} c_i t^i \), such as Adomian decomposition method [35] and Homotopy perturbation method [36]. But the exact solution of many DEs can’t be estimated by polynomial basis. Therefore, we have defined a new basis for Spectral methods to solve them as follows:

\[
\Phi_n(t) = \sum_{i=0}^{n} c_i t^{i\alpha}.
\]

Now by transformation \( z = 1 - 2(\frac{t}{\eta})^{\alpha}, \alpha, \eta > 0 \) on classical Chebyshev polynomials of the first kind, we defined the GFCFs in the interval \([0, \eta]\), that be denoted by \( \eta FT_\alpha^n(t) = T_n(1 - 2(\frac{t}{\eta})^{\alpha}) \).

By this definition, the singular Sturm-Liouville differential equation of classical Chebyshev polynomials become:

\[
\sqrt{\frac{\eta^2 - t^\alpha}{t^{\alpha - 1}}} \frac{d}{dt} \left[ \sqrt{\frac{\eta^2 - t^\alpha}{t^{\alpha - 1}}} \frac{d}{dt} \eta FT_\alpha^n(t) \right] + n^2 \alpha^2 \eta FT_\alpha^n(t) = 0, \quad t \in [0, \eta]. \tag{3.1}
\]

The \( \eta FT_\alpha^n(t) \) can be obtained using the recursive relation as follows \( n = 1, 2, \cdots \):

\[
\eta FT_\alpha_0(t) = 1, \quad \eta FT_\alpha_1(t) = 1 - 2(\frac{t}{\eta})^{\alpha}, \\
\eta FT_\alpha_{n+1}(t) = (2 - 4(\frac{t}{\eta})^{\alpha}) \eta FT_\alpha_n(t) - \eta FT_\alpha_{n-1}(t).
\]

The analytical form of \( \eta FT_\alpha^n(t) \) of degree \( n\alpha \) is given by

\[
\eta FT_\alpha^n(t) = \sum_{k=0}^{n} (-1)^k \frac{n^{2k}(n + k - 1)!}{(n - k)!(2k)!} \left( \frac{t}{\eta} \right)^{\alpha k}.
\]
SOLVING THE INTEGRO-DIFFERENTIAL EQUATIONS

\[ \sum_{k=0}^{n} \beta_{n,k,\eta,\alpha} t^{\alpha k}, \quad t \in [0, \eta], \quad (3.2) \]

where

\[ \beta_{n,k,\eta,\alpha} = (-1)^k \eta^{2k} (n+k-1)! \frac{(n-k)!}{(n-k)!(2k)!} \eta^{\alpha k} \] and \( \beta_{0,k,\eta,\alpha} = 1. \)

Note that \( \eta FT_0^\alpha (0) = 1 \) and \( \eta FT_\eta^\alpha (\eta) = (-1)^n. \)

The GFCFs are orthogonal with respect to the weight function \( w(t) = \frac{t^{\alpha - 1}}{\sqrt{\eta^{\alpha} - t}} \) in the interval \([0, \eta]\):

\[ \int_0^\eta \eta FT_n^\alpha (t) \eta FT_m^\alpha (t) w(t) dt = \frac{\pi}{2\alpha} c_n \delta_{mn}. \quad (3.3) \]

where \( \delta_{mn} \) is Kronecker delta, \( c_0 = 2 \), and \( c_n = 1 \) for \( n \geq 1. \) Eq. (3.3) is provable using properties of orthogonality in the Chebyshev polynomials.

Figs. 1 shown graphs of GFCFs for various values of \( n \) and \( \alpha \) and \( \eta = 5. \)

**Figure 1.** (a) Graph of the GFCFs with \( \alpha = 0.25 \) and various values of \( n. \)
(b) Graph of the GFCFs with \( n = 5 \) and various values of \( \alpha. \)

### 3.3. Approximation of functions.

Any function \( y(t) \in C[0, \eta] \) can be expanded as follows:

\[ y(t) = \sum_{n=0}^{\infty} a_n \eta FT_n^\alpha (t), \]

where the coefficients \( a_n \) obtain by inner product:

\[ \langle y(t), \eta FT_n^\alpha (t) \rangle_w = \langle \sum_{n=0}^{\infty} a_n \eta FT_n^\alpha (t), \eta FT_n^\alpha (t) \rangle_w \]

and using the property of orthogonality in the GFCFs:

\[ a_n = \frac{2\alpha}{\pi c_n} \int_0^\eta \eta FT_n^\alpha (t)y(t)w(t)dt, \quad n = 0, 1, 2, \ldots \]
In practice, we have to use first $m$-terms GFCFs and approximate $y(t)$:

$$y(t) \approx y_m(t) = \sum_{n=0}^{m-1} a_n \eta^{FT_n^\alpha}(t) = A^T \Phi(t), \quad (3.4)$$

with

$$A = [a_0, a_1, \cdots, a_{m-1}]^T, \quad (3.5)$$

$$\Phi(t) = [\eta^{FT_0^\alpha}(t), \eta^{FT_1^\alpha}(t), \cdots, \eta^{FT_{m-1}^\alpha}(t)]^T. \quad (3.6)$$

3.4. Convergence of the method. The following theorem shows that by increasing $m$, the approximation solution $f_m(t)$ is convergent to $f(t)$ exponentially.

**Theorem 2.** Suppose that $D^k f(t) \in C[0, \eta]$ for $k = 0, 1, \cdots, m$, and $\eta^{F_m^\alpha}$ is the subspace generated by $\{\eta^{FT_0^\alpha}(t), \eta^{FT_1^\alpha}(t), \cdots, \eta^{FT_{m-1}^\alpha}(t)\}$. If $f_m = A^T \Phi$ (in Eq. (3.4)) is the best approximation of $f(t)$ from $\eta^{F_m^\alpha}$, then the error bound is presented as follows

$$\| f(t) - f_m(t) \|_w \leq \frac{\eta^m M_\alpha}{2^m \Gamma(m\alpha + 1)} \sqrt{\frac{\pi}{\alpha m!}},$$

where $M_\alpha \geq |D^{m\alpha} f(t)|, \ t \in [0, \eta]$.

**Proof.** By theorem 1, $y = \sum_{i=0}^{m-1} \frac{t^i}{\Gamma(i\alpha + 1)} D^{i\alpha} f(0^+)$ and

$$|f(t) - y(t)| \leq M_\alpha \frac{t^{m\alpha}}{\Gamma(m\alpha + 1)},$$

since $A^T \Phi(t)$ is the best approximation to $f(t)$ in $\eta^{F_m^\alpha}$ and $y \in \eta^{F_m^\alpha}$, one has

$$\| f(t) - f_m(t) \|_w^2 \leq \| f(t) - y(t) \|_w^2 \leq \frac{M_\alpha^2}{\Gamma(m\alpha + 1)^2} \int_0^\eta \frac{t^{\frac{\alpha}{2} + 2m\alpha - 1}}{\sqrt{\eta^\alpha - t^\alpha}} dt = \frac{M_\alpha^2 \eta^{2m\alpha}}{\Gamma(m\alpha + 1)^2 \alpha^{2m} m!}.$$

Now by taking the square roots, the theorem can be proved. $\Box$

**Theorem 3.** The generalized fractional order of the Chebyshev function $\eta^{FT_n^\alpha}(t)$, has precisely $n$ real zeros on interval $(0, \eta)$ in the form

$$t_k = \eta \left( \frac{1 - \cos\left(\frac{2k-1}{2n}\pi\right)}{2} \right)^{\frac{1}{\alpha}}, \ k = 1, 2, \cdots, n.$$

Moreover, $\frac{d}{dt} \eta^{FT_n^\alpha}(t)$ has precisely $n - 1$ real zeros on interval $(0, \eta)$ in the following points:

$$t'_k = \eta \left( \frac{1 - \cos\left(\frac{k\pi}{n}\right)}{2} \right)^{\frac{1}{\alpha}}, \ k = 1, 2, \cdots, n - 1.$$
Proof. The Chebyshev polynomial $T_n(x)$ has $n$ real zeros [37, 38]:

$$x_k = \cos\left(\frac{(2k - 1)\pi}{2n}\right), \ k = 1, 2, \cdots, n,$$

therefore $T_n(x)$ can be written as

$$T_n(x) = (x - x_1)(x - x_2)\cdots(x - x_n).$$

Using transformation $x = 1 - 2\left(\frac{t}{\eta}\right)^\alpha$ yields to

$$FT_n^\alpha(t) = ((1 - 2\left(\frac{t}{\eta}\right)^\alpha) - x_1)((1 - 2\left(\frac{t}{\eta}\right)^\alpha) - x_2)\cdots((1 - 2\left(\frac{t}{\eta}\right)^\alpha) - x_n),$$

so, the real zeros of $\eta FT_n^\alpha(t)$ are $t_k = \eta\left(\frac{1-x_k}{2}\right)^\frac{1}{\alpha}$.

Also, the real zeros of $\frac{d}{dt} T_n(t)$ occurs in the following points [37]:

$$x'_k = \cos\left(\frac{k\pi}{n}\right), \ k = 1, 2, \cdots, n - 1.$$

Same as in previous, the absolute extremes of $\eta FT_n^\alpha(t)$ are $t'_k = \eta\left(\frac{1-x'_k}{2}\right)^\frac{1}{\alpha}$.

4. Application of the GFCF collocation method

In this section, the GFCFs collocation method is applied to solve the integro-differential equation in the Eq. (2.1). For satisfying the boundary conditions, the conditions in the Eq. (2.8) are satisfied as follows:

$$\widehat{y}_m(t) = \beta_0 + \beta_1 t + t^2 y_m(t), \quad (4.1)$$

where $y_m(t)$ is defined in the Eq. (3.4). Now, $\widehat{y}_m(t) = \beta_0$ and $\widehat{y}_m'(t) = \beta_1$ when $t$ tends to zero, so the conditions in the Eq. (2.8) are satisfied.

To apply the collocation method, the residual function is constructed by substituting $\widehat{y}_m(t)$ in the Eq. (4.1) for $y(t)$ in the integro-differential equation (2.1):

$$Res(t) = \frac{d^2 y}{dt^2} - g(t) + a(t)y(t) - b(t)\int_0^t \cos(w_p s)y(s)ds. \quad (4.2)$$

The equations for obtaining the coefficient $\{a_i\}_{i=0}^{m-1}$ arise from equalizing $Res(t)$ to zero on $m$ collocation points:

$$Res(t_i) = 0, \quad i = 0, 1, \ldots, m - 1. \quad (4.3)$$

In this study, the roots of the GFCFs in the interval $[0, \eta]$ (Theorem 3) are used as collocation points. By solving the obtained set of equations, we have the approximating function $\widehat{y}_m(t)$. And also consider that all of the computations have been done by Maple 18 on a laptop with CPU Core i7, Windows 8.1 64bit, and 8GB of RAM.
5. ILLUSTRATIVE EXAMPLES

In this section, by using the present method, some well-known examples are solved to show efficiently and applicability GFCFs method based on Spectral method. The present method is applied to solve the integro-differential equation (2.1) and their outputs are compared with the corresponding analytical solution. These examples studied also by Dehghan [16], Khan [18] and Pathak [21], we will compare our results with their results to show the effectiveness of the present method.

Example 1. Consider the equation (2.1) with [16, 18, 21]

\[ w_p = 2, \quad a(t) = \cos(t), \quad b(t) = \sin(t/2), \quad \beta_0 = 1, \quad \beta_1 = 0 \]
\[ g(t) = \cos(t) - t \sin(t) + \cos(t) (t \sin(t) + \cos(t)) \]
\[ - \sin(t/2) \left( \frac{2}{9} \sin(3t) - \frac{t}{6} \cos(3t) + \frac{t}{2} \cos(t) \right). \]

The exact solution of this equation is \( y(t) = t \sin(t) + \cos(t). \) By applying the technique described in the last section, for satisfying the boundary conditions, the conditions are satisfied as \( \hat{y}_m(t) = 1 + t^2 \). The residual function is constructed as follows:

\[ \text{Res}(t) = \frac{d^2 y}{dt^2} - g(t) + a(t)y(t) - b(t) \int_0^t \cos(w_p s) y(s) ds. \]

Therefore, to obtain the coefficient \( \{a_i\}_{i=0}^{m-1} \); \( \text{Res}(t) \) is equalized to zero at \( m \) collocation point. By solving this set of nonlinear algebraic equations, we can find the approximating function \( \hat{y}_m(t) \). Figure 2 shows the logarithmic graph of the absolute error and the residual error of the approximate solution and the analytic solution for \( m = 20 \) and \( \alpha = 0.50 \).

![Figure 2](image-url)
The resulting graph in comparison to the presented method and the exact solution is shown in Figure 3(a). In Figure 3(b) to show the convergence of the present method, we showed that by increasing the \(m\) the residual function decreases, where \(\alpha = 0.50\).

![Graph comparison](image)

**Figure 3.** (a) Obtained graph in comparison to the exact solution with \(m = 20\) and \(\alpha = 0.50\). (b) Residual functions for \(m = 10, 15, 20\) and \(\alpha = 0.50\), to show the convergence rate of the GFCF method for example 1.

Table 1 compares the error norm \(\|y - y_m\|_2\) by the present method and Dehghan [16] for examples 1-3.

<table>
<thead>
<tr>
<th>Example</th>
<th>Dehghan [16]</th>
<th>Present method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1</td>
<td>6.826965141905116e-13</td>
<td>1.4691408679507079e-20</td>
</tr>
<tr>
<td>Example 2</td>
<td>2.440936203513114e-13</td>
<td>8.8876680000000000e-39</td>
</tr>
<tr>
<td>Example 3</td>
<td>3.279510650660257e-12</td>
<td>2.6061709227703975e-95</td>
</tr>
</tbody>
</table>

Table 2 compares the obtained values of \(y(t)\) by the present method and the values given by Khan [18] (Legendre multi-wavelets) and Pathak [21] (Variational iteration method), it shows that the results obtained in the present method are more accurate.
Table 2. Obtained values of \(y(t)\) for example 1 by Khan, Pathak, and the present method with \(m = 20\).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>1.006711649</td>
<td>1.004987505</td>
<td>1.004987506944</td>
<td>1.004987506942</td>
<td>1.997e-12</td>
<td>3.329e-9</td>
</tr>
<tr>
<td>0.2</td>
<td>1.019729643</td>
<td>1.0198004356</td>
<td>1.019800444000</td>
<td>1.019800444000</td>
<td>5.546e-12</td>
<td>3.391e-9</td>
</tr>
<tr>
<td>0.3</td>
<td>1.043967189</td>
<td>1.04399255130</td>
<td>1.043992551124</td>
<td>1.043992551124</td>
<td>6.216e-12</td>
<td>1.032e-9</td>
</tr>
<tr>
<td>0.4</td>
<td>1.078484444</td>
<td>1.076823090</td>
<td>1.076828330926</td>
<td>1.076828330926</td>
<td>1.072e-11</td>
<td>3.318e-9</td>
</tr>
<tr>
<td>0.5</td>
<td>1.116769872</td>
<td>1.117295331180</td>
<td>1.117295331192</td>
<td>1.117295331192</td>
<td>1.182e-11</td>
<td>1.190e-9</td>
</tr>
<tr>
<td>0.6</td>
<td>1.164117090</td>
<td>1.164067124</td>
<td>1.164121098922</td>
<td>1.164121098922</td>
<td>2.436e-11</td>
<td>3.746e-9</td>
</tr>
<tr>
<td>0.7</td>
<td>1.216007719</td>
<td>1.215666930</td>
<td>1.215794568347</td>
<td>1.215794568347</td>
<td>3.654e-12</td>
<td>1.388e-9</td>
</tr>
<tr>
<td>0.8</td>
<td>1.270604841</td>
<td>1.270327190</td>
<td>1.270591582096</td>
<td>1.270591582096</td>
<td>2.975e-11</td>
<td>2.415e-9</td>
</tr>
<tr>
<td>0.9</td>
<td>1.326616801</td>
<td>1.326109990</td>
<td>1.326604186976</td>
<td>1.326604186976</td>
<td>4.096e-11</td>
<td>3.827e-9</td>
</tr>
</tbody>
</table>

Example 2. Consider the equation (2.1) with [16, 18, 21]

\[
\begin{align*}
wp &= 1, & a(t) &= -\sin(t), & b(t) &= \sin(t), & \beta_0 &= 1, & \beta_1 &= 2/3 \\
g(t) &= \frac{1}{9}e^{-\frac{t}{3}} - \sin(t) \left( e^{-\frac{t}{3}} + t \right) \\
&\quad - \sin(t) \left( -\frac{3}{10} \cos(t) e^{\frac{-t}{3}} + \frac{9}{10} \sin(t) e^{\frac{-t}{3}} + \cos(t) + t \sin(t) - \frac{7}{10} \right).
\end{align*}
\]

The exact solution of this equation is \(y(t) = e^{-\frac{t}{3}} + t\). By applying the technique described in the last section, the conditions are satisfied as: \(\hat{y}_m(t) = 1 + \frac{2}{3}t + t^2 \hat{y}_m(t)\), and construct the residual functions as follows:

\[
Res(t) = \frac{d^2y}{dt^2} - g(t) + a(t)y(t) - b(t) \int_0^t \cos(wp s) y(s) ds.
\]

Figure 4 shows the logarithmic graph of the absolute error and the residual error of the approximate solution and the exact solution for \(m = 20\) and \(\alpha = 0.50\).

Figure 4. The logarithmic graphs of the (a) absolute error and the (b) residual error for example 2 with \(m = 20\) and \(\alpha = 0.50\).
The resulting graph in comparison to the present method and the exact solution is shown in Figure 5 (a). To show the convergence of the present method to solve this example with $\alpha = 0.50$ in Figure 5 (b), we showed that by increasing the $m$ the residual function decreases.

![Figure 5. (a) Obtained graph in comparison with the exact solution with $m = 20$ and $\alpha = 0.50$. (b) Residual functions for $m = 10, 15, 20$ and $\alpha = 0.50$, to show the convergence rate of GFCF method for example 2.](image)

Table 1 compares the error norm $\|y - y_m\|_2$ by the present method and Dehghan [16] for examples 1-3. Table 3 compares the obtained values of $y(t)$ by the present method and the values given by Khan [18] (Legendre multi-wavelets) and Pathak [21] (Variational iteration method), it shows that the results obtained in the present method are more accurate.

**Table 3.** Obtained values of $y(t)$ for example 2 by Khan (Legendre multi-wavelets), Pathak (Variational iteration method), and the present method with $m = 20$

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>1.067409867</td>
<td>5.0900e-13</td>
<td>1.06721610048200590204</td>
<td>1.06721610048200590204</td>
<td>2.5e-11</td>
<td>4.1e-18</td>
</tr>
<tr>
<td>0.2</td>
<td>1.135498873</td>
<td>1.3237e-10</td>
<td>1.13550698503161773772</td>
<td>1.13550698503161773772</td>
<td>6.6e-11</td>
<td>4.0e-18</td>
</tr>
<tr>
<td>0.3</td>
<td>1.204831254</td>
<td>3.4433e-09</td>
<td>1.20483741803959597317</td>
<td>1.20483741803959597317</td>
<td>6.8e-11</td>
<td>1.1e-18</td>
</tr>
<tr>
<td>0.4</td>
<td>1.275363888</td>
<td>3.4759e-08</td>
<td>1.27517331904294754000</td>
<td>1.27517331904294754000</td>
<td>1.2e-20</td>
<td>3.7e-18</td>
</tr>
<tr>
<td>0.5</td>
<td>1.346379637</td>
<td>2.0843e-07</td>
<td>1.34648172489061407403</td>
<td>1.34648172489061407403</td>
<td>1.2e-20</td>
<td>1.2e-18</td>
</tr>
<tr>
<td>0.6</td>
<td>1.418729558</td>
<td>————</td>
<td>1.41873075307798185864</td>
<td>1.41873075307798185864</td>
<td>2.6e-20</td>
<td>3.9e-18</td>
</tr>
<tr>
<td>0.7</td>
<td>1.491932333</td>
<td>————</td>
<td>1.49188956633678166437</td>
<td>1.49188956633678166437</td>
<td>5.1e-21</td>
<td>1.4e-18</td>
</tr>
<tr>
<td>0.8</td>
<td>1.565927886</td>
<td>————</td>
<td>1.5659283386464869274</td>
<td>1.5659283386464869274</td>
<td>2.8e-20</td>
<td>2.4e-18</td>
</tr>
<tr>
<td>0.9</td>
<td>1.640830867</td>
<td>————</td>
<td>1.64081822068171786606</td>
<td>1.64081822068171786606</td>
<td>4.1e-20</td>
<td>3.8e-18</td>
</tr>
</tbody>
</table>
Example 3. Consider equation (2.1) with \([16, 18, 21]\)

\[ w_p = 3, \quad a(t) = 1, \quad b(t) = \sin(t) + \cos(t), \beta_0 = 2, \quad \beta_1 = -5, \]

\[ g(t) = -t^3 + t^2 - 11t + 4 - (\sin(t) + \cos(t)) \]

\[
\left( -\frac{t^3}{3} \sin(3t) - \frac{t^2}{3} \cos(3t) - \frac{13}{27} \cos(3t) - \frac{13}{9} t \sin(3t) \right) \\
+ \frac{t^2}{3} \sin(3t) + \frac{16}{27} \sin(3t) + \frac{2t}{9} \cos(3t) + \frac{13}{27},
\]

The exact solution of this equation is \(y(t) = -t^3 + t^2 - 5t + 2\). By applying the technique described in the last section, the conditions are satisfied as: \(\hat{y}_m(t) = 2 - 5t + t^2 y_m(t)\). The residual function is constructed as follows:

\[ \text{Res}(t) = \frac{d^2 y}{dt^2} - g(t) + a(t)y(t) - b(t) \int_0^t \cos(w_p s)y(s)ds. \]

Figure 6 shows the logarithmic graph of the absolute error and the residual error of the approximate solution and the exact solution for \(m = 5\) and \(\alpha = 0.50\).

**Figure 6.** The logarithmic graphs of the (a) absolute error and the (b) residual error for example 2 with \(m = 5\) and \(\alpha = 0.50\).
The resulting graph in comparison to the present method and the exact solutions is shown in Figure 7.

![Graph comparison](image)

**Figure 7.** Obtained graph in comparison to the exact solution for example 3 with $m = 5$ and $\alpha = 0.50$.

Table 1 compares the error norm $\|y - y_m\|_2$ by the present method and Dehghan [16] for examples 1-3. Table 4 compares the obtained values of $y(t)$ by the present method and the values given by Khan [18] (Legendre multi-wavelets) and Pathak [21] (Variational iteration method), it shows that the results obtained in the present method are more accurate.

**Table 4.** Obtained values of $y(t)$ for example 3 by Khan, Pathak, and the present method with $m = 5$.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>1.510232801</td>
<td>—</td>
<td>1.50900000</td>
<td>1.509</td>
<td>0</td>
<td>4.331e-49</td>
</tr>
<tr>
<td>0.2</td>
<td>1.031768407</td>
<td>0.0000e-00</td>
<td>1.03200000</td>
<td>1.032</td>
<td>0</td>
<td>3.845e-50</td>
</tr>
<tr>
<td>0.3</td>
<td>0.563149603</td>
<td>—</td>
<td>0.56300000</td>
<td>0.563</td>
<td>0</td>
<td>2.869e-49</td>
</tr>
<tr>
<td>0.4</td>
<td>0.096626905</td>
<td>1.5399e-12</td>
<td>0.09600000</td>
<td>0.096</td>
<td>0</td>
<td>5.244e-49</td>
</tr>
<tr>
<td>0.5</td>
<td>-0.373166748</td>
<td>—</td>
<td>-0.37500000</td>
<td>-0.375</td>
<td>0</td>
<td>6.837e-49</td>
</tr>
<tr>
<td>0.6</td>
<td>-0.856012437</td>
<td>4.1554e-11</td>
<td>-0.85600000</td>
<td>-0.856</td>
<td>0</td>
<td>7.799e-49</td>
</tr>
<tr>
<td>0.7</td>
<td>-1.353869376</td>
<td>—</td>
<td>-1.35300000</td>
<td>-1.353</td>
<td>0</td>
<td>8.296e-49</td>
</tr>
<tr>
<td>0.8</td>
<td>-1.871984718</td>
<td>2.1376e-10</td>
<td>-1.87200000</td>
<td>-1.872</td>
<td>0</td>
<td>8.489e-49</td>
</tr>
<tr>
<td>0.9</td>
<td>-2.419488059</td>
<td>—</td>
<td>-2.41900000</td>
<td>-2.419</td>
<td>0</td>
<td>8.517e-49</td>
</tr>
</tbody>
</table>
6. CONCLUSION

The main goal of this paper was to introduce a new orthogonal basis, namely the generalized fractional order of the Chebyshev orthogonal functions (GFCF) to construct an approximation to the solution of the integro-differential equation arising in oscillating magnetic fields. The presented results show that the introduced basis for the collocation spectral method is efficient and applicable. Our results have better accuracy with lesser \( m \), and the absolute error as compared to the exact solution. A comparison was made of the exact solution and the present method. As shown, the method is converging and has an approximate accuracy and stability, and the error decreases with increasing \( m \).

ACKNOWLEDGEMENTS

The authors are very grateful to reviewers and editor for carefully reading the paper and for their comments and suggestions which have improved the paper.

REFERENCES


