Summer 2013

Optimal Control Modeling and Simulation, with Application to Cholera Dynamics

Chairat Modnak
Old Dominion University

OPTIMAL CONTROL MODELING AND SIMULATION, WITH APPLICATION TO CHOLERA DYNAMICS

by

Chairat Modnak
B.S. March 1998, Naresuan University, Phitsanulok, THAILAND
M.S. March 2001, King Mongkut's Tech. LAD, Bangkok, THAILAND
M.S. July 2007, Ohio University, Ohio, USA

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

COMPUTATIONAL AND APPLIED MATHEMATICS

OLD DOMINION UNIVERSITY
August 2013

Approved by:

Jin Wang (Director)

Ruhai Zhou (Member)

Yan Peng (Member)

Chung-Hao Chen (Member)

ABSTRACT

OPTIMAL CONTROL MODELING AND SIMULATION, WITH APPLICATION TO CHOLERA DYNAMICS

Chairat Modnak
Old Dominion University, 2013
Director: Dr. Jin Wang

The theory of optimal control, a modern extension of the calculus of variations, has found many applications in a wide range of scientific fields, particularly in epidemiology with respect to disease prevention and intervention. In this dissertation, we conduct optimal control modeling, simulation and analysis to cholera dynamics. Cholera is a severe intestinal infectious disease that remains a serious public health threat in developing countries. Transmission of cholera involves complex interactions between the human host, the pathogen, and the environment. The worldwide cholera outbreaks and their increasing severity, frequency and duration in recent years underscore the gap between the complex mechanism of cholera transmission and our current quantitative understanding and control strategies for this disease.

We incorporate multiple time-dependent intervention strategies, including vaccination, antibiotic treatment, and water sanitation, into cholera epidemiological models and seek solutions that best balance the costs and gains of the controls. Pontryagin's Maximum/Minimum principle allows us to construct the optimal control system that involves the state equations, the adjoint equations, and the optimality condition that characterizes the controls. The system is then numerically solved using an iterative procedure based on the Forward-Backward Sweep Method. We discuss in detail the mathematical models and numerical results for various scenarios and their implications to public health administration on disease control.

In the last part of this dissertation, we investigate new iterative algorithms with improved convergence properties compared to the original Forward-Backward Sweep Method. We discuss the applications of such numerical algorithms to optimal control problems as well as other types of constrained dynamical systems. We conduct careful error analysis and present several numerical examples to validate the analytic results.

ACKNOWLEDGMENTS

I would like to first thank my advisor Dr. Jin Wang who has given me an opportunity to work with him and shared some great ideas in many interesting areas of applied mathematics. I appreciate all of his contributions, suggestions, ideas, funds (from the National Science Foundation), and help that allowed me to accomplish my goal here at Old Dominion University.

I would like to thank my committee members, Dr. Ruhai Zhou, Dr. Yan Peng and Dr. Chung-Hao Chen for their valuable time and advice.

I would also like to thank those faculty members at ODU and other universities, especially Drs. Holly Gaff, Elsa Schaefer, Renee Fister and Suzanne Lenhart, for their valuable advice and help on cholera modeling.

I greatly appreciate Dr. Hideaki Kaneko, Heather Kunkel, Amanda Working, Natalie Hutchinson, Dr. John Adam, Dr. Dorrepaal and all other faculty members at the Department of Mathematics and Statistics for giving me advice, help, time and effort.

I would like to thank the Department of Mathematics and Statistics for providing me with a Graduate Teaching Assistantship so that I could focus on my study without financial worries. Also, my great appreciations go to the secretaries of the department, Barbara, Miriam and Sheila, for helping me during my time as a graduate student with all the paperwork and teaching. They made my life easier.

I also appreciate the supports from my friends, co-workers, and other fellow Monarchs who have helped me to get through some difficult times and given hopes with all the cheers and wishes for me to succeed in my study.

I am greatly thankful to my understanding and encouraging parents whose support helped me to succeed in my studies. Specifically, I would love to thank my mother who has been giving me advice and love that has helped me to succeed in my study, career, and my future. Finally, I would like to dedicate this work to my late father who was my idol.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In recent historic events, we have seen that giving an early warning has saved a lot of lives and has given information to cities to reduce cost from infrastructure destruction, although higher accuracy is still needed for better predictions. Fortunately there are many scientists and mathematicians trying to utilize mathematical models to provide better understanding of the dynamics of nature. There are numerous models that have been proposed and currently been used in society both commercially and publicly such as weather forecast models, hurricane path models, flash flood watch models, stock market forecast models, economy prediction models, and disease outbreak models. As an example, we have witnessed the deadly hurricane Sandy that came to the Atlantic coast in 2012; however, even though it was the most destructive hurricane in 2012, not many lives were lost. The reason is several new models have been used by both national and local television channels to send out early warning for evacuations. As a result, many lives have been saved.

In addition, we need better models to forecast complex systems such as, wildfires, disease outbreaks, floods, horrific tsunamis, earthquakes, meteors, landslides, and droughts. These destructive forces of nature and diseases could possibly be predicted by using mathematical models. In particular, mathematical models have become increasingly important to understand the complex dynamics of diseases.

In this dissertation, we are concerned with cholera, a severe intestinal infectious disease caused by the bacterium *Vibrio cholerae* that remains a serious public health threat in developing countries. We have in recent years, witnessed an increasing number of cholera outbreaks worldwide including one of the largest cholera epidemics in modern history that took place in Haiti from 2010-2012 with more that 530,000 reported cases and over 7,000 deaths. Major cholera outbreaks also include those in Sierra Leone (2012), Ghana (2011), Nigeria (2010), Vietnam (2009), Zimbabwe (2008), and India (2007), among others. Current estimates by World Health Organization show 3-5 million cholera cases every year in the world.

The transmission of cholera involves both direct (i.e., human-to-human) and indirect (i.e., environment-to-human) routes, due to the multiple interactions between

the human host, the pathogen, and the environment. In order to better understand the complex transmission dynamics of cholera, a number of mathematical models have been proposed and analyzed; we will briefly review some of these works in Section 2.2. No doubt these studies have improved our knowledge of cholera dynamics. However, the worldwide cholera outbreaks and their increasing severity, frequency and duration in recent years underscore the gap between the complex mechanism of cholera transmission and our current quantitative understanding and control strategies for this disease.

A focus of this dissertation is to formulate optimal control models to investigate cholera dynamics and explore control strategies that best balance the costs and gains in fighting cholera. In doing so we will combine mathematical modeling, analysis, and numerical simulation to seek optimal control solutions. Our results will improve the understanding of the complex mechanism of cholera transmission and can provide useful guidelines for public health administration for the prevention and intervention of cholera outbreaks. We start our discussion in Chapter 2 with a background of optimal control, followed by a review of some representative mathematical models of epidemic and endemic cholera. In Chapter 3, we discuss our models with optimal controls and numerical simulations. Mathematical equations, model parameters and diagrams are also carefully presented. In addition, we have expanded our study to several interesting cases to better understand the disease outbreak. In Chapter 4, we study new iterative algorithms for solving optimal control and other types of constrained dynamical problems. Algorithms, examples, mathematical formulations and error analysis are presented. In the last Chapter, we conclude our study and discuss future work.

CHAPTER 2

BACKGROUND

2.1 OPTIMAL CONTROL

Optimal control theory is a modern extension of the calculus of variations to find an optimal path or value that gives either maximum or minimum points of functions. An optimal control problem contains state variables, control(s) and an objective function(s). It can be a system of differential equations, partial differential equations, discrete equations, integro-difference equations, and stochastic differential equations. In this chapter, we will show some dynamic problems in order to have a better understanding about the optimal control theory.

First consider the optimal control of growth model formulated by Cohen [1]. The model is a system of two differential equations

$$\begin{aligned} \frac{dx_1}{dt} &= u(t)x_1(t), \\ \frac{dx_2}{dt} &= (1-u(t))x_2(t), \\ 0 \le u(t) &\le 1, \\ x_1(0) > 0,\ x_2(0) &\ge 0, \end{aligned}$$

where $u(t)$ is the fraction of the photosynthate partitioned to vegetative growth, $x_1(t)$ is the weight of the vegetative part at time t and $x_2(t)$ is the weight of the reproductive part. In order to keep plants growing, this problem is maximizing the growth of the reproductive part. Thus, the goal is to find the optimal value of $u(t)$ to maximize the functional

$$\int_0^T ln(x_2(t))dx.$$

This optimal control problem has $\int_0^T$ state variables, $x_1(t)$ and $x_2(t)$, and one control $u(t)$. The objective functional is a natural logarithmic function here because we believe that plants have linear growth.

King and Roughgarden used optimal control to solve this problem. They set up the maximum value of control to be 1 and the minimum value to be 0 for a five day

observation. The results showed that from the starting day to about 2 days, the value of control was constantly 1 and had the constant weight of the reproductive growth. Meanwhile, the weight of the vegetative part was increasing. For the next half day, the weight of the vegetative part was slowing down and the weight of the reproductive growth started increasing. At the time, the values of control were decreasing and their values were between 0.4 and 0.6. After this day, the value of control became constantly zero, the weight of reproductive part kept increasing and the weight of the vegetative part stayed constant. This study shows that at one time, the plant uses all of its photosynthate for vegetative growth and later will use it into some vegetative and some reproductive growth. This is a simple optimal control example that has only one control in the system to affect the state variables in order to maximize or minimize the objective functional in an optimal way. Next we will consider a bigger system that has more state variables and a control.

The following model is developed by Thalya Burden, Jon Ernstherger and K. Renee Fister [5] from the original model discussed in Panetta and Kirschner [6] to investigate using cytokines to treat cancer done in conjunction with adoptive cellular immunotherapy(ACI)

$$\begin{aligned}
\frac{dx}{dt} &= cy - \mu_2 x + \frac{p_1 xz}{g_1 + z} + u(t)s_1, \\
\frac{dy}{dt} &= r_2 y(1 - by) - \frac{axy}{g_2 + y}, \\
\frac{dz}{dt} &= \frac{p_2 xy}{g_3 + y} - \mu_3 z, \\
x(0) &= 1, y(0) = 1, z(0) = 1.
\end{aligned}$$

There are three state variables in this model; $x(t)$, the activated immune system cells; $y(t)$, the tumor cells; $z(t)$, the concentration of IL-2 in the single tumor-site compartment. The function $u(t)$ is the control that represents the percentage of adoptive celluar immunotherapy given. All parameters are assumed to be positive constants. The first equation is for the rate of change for the effector cell population where c is the antigenicity rate of the tumor and s_1 is a critical parameter [6]. The parameter μ_2 in the second term is the rate of death of the effector cells, therefore, the term represents the natural death of the cells. The last term indicates the saturated effects of the immune response. The second equation represents the rate of change of the tumor cells, and the last equation gives the rate of change for the concentration of IL-2.

The control, $u(t)$, is chosen to be a piecewise continuous function with a minimum of 0 and a maximum of 1. When $u(t) = 1$, it shows that the maximal immunotherapy should be applied. In this study, they want to minimize the cost of the control while maximizing the effects of the immunotherapy. Therefore, they define the objective functional as

$$J(u) = \int_0^T \left[x(t) - y(t) + z(t) - \frac{1}{2}(u(t))^2\right] dt.$$

The goal of their study is to characterize the optimal control, u^*, satisfying

$$\max_{0 \le u \le 1} J(u) = J(u^*).$$

The study shows reasonable and promising results that at first the control started at a maximum amount and then reduces sharply to zero. As there is no activity of the control, the cancer cells start to rise in number and that causes the control to come back and reduce the number of the cancer cells. The study seems to give logical and reasonable output. Optimal control theory can be applied to many areas. In the next example, still a simple model, we will look at an economic model of production planning.

This example is an application of optimal control in economics[7]. The model contains one state variable $x(t)$, one known continuous function $r(t)$ and one control $u(t)$

$$\frac{dx}{dt} = -r(t) + u(t), \quad x(0) = x_0, \quad x(t) \ge 0, \quad 0 \le u(t) \le A, \quad 0 \le t \le T,$$

where the state variable $x(t)$ denotes the stock of a commodity at time t, $r(t)$ denotes the rate of demand for the commodity at time t, and $u(t)$ denote the rate of production at time t. The control is a piecewise continuous function that is controlled by the production planner, and x_0 is the initial stock level.

Let h be a function of the rate of production and b be the cost per unit time of storing a unit of commodity. Therefore, the cost per unit time, at time t, of operating the system is

$$f(t, x(t), u(t)) = h(u(t)) + bx(t).$$

The total cost is given by

$$C(u) = \int_0^T f(t, x(t), u(t))dt.$$

The goal is that when the demand and initial stock are given, the production planner tries to determine the control at time t while the total cost is minimized.

From the previous three examples, we can see that the formulation of optimal control problem mainly involves three parts; state variables, controls and objective functional. In general, an optimal control problem can be formed by a system of equations where state variables are described by [7]

$$x(t) = x(x^1(t), x^2(t), \cdots, x^n(t)),$$

in n-dimensional euclidean space with initial conditions at time $t = 0$

$$x(t_0) = x_0 = x(x_0^1, x_0^2, \cdots, x_0^n).$$

The state of the system varies with time according to the system of differential equations or the system of partial differential equations

$$\frac{dx^i}{dt} = f^i(t, x(t), u(t)),$$

where $u(t)$ is the control. The objective functional is in the form

$$J(\phi, u) = \phi(t) + \int_0^T f(t, x(t), u(t))dt,$$

with a function ϕ defined to be a real valued function. Generally, an optimal control problem aims to find the optimal control, $u(t)$, so that the functional $J(\phi, u)$ is minimized or maximized. To achieve this goal, we need to use Pontryagin's Maximum/Minimum principle and some numerical methods [1]. The Pontryagin's Maximum principle is described below; the minimum principle is similar and is not included here.

Theorem 1. *If $u^*(t)$ and $x^*(t)$ are optimal for our problem as described above, then there exists a piecewise differentiable adjoint variable, $\lambda(t)$, such that $H(t, x^*(t), u(t), \lambda(t)) \leq H(t, x^*(t), u^*(t), \lambda(t))$ for all controls u at each time t, where the Hamiltonian H is*

$$H = f(t, x(t), u(t)) + \lambda(t)g(t, x(t), u(t)),$$

and

$$\lambda'(t) = -\frac{\partial H(t, x^*(t), u^*(t), \lambda(t))}{\partial x},$$
$$\lambda(T) = 0,$$

where $t_0 \leq t \leq T$.

After formulating the Hamiltonian and applying the theorem above, our optimal control problem now includes two systems of differential equations that need to be solved. The first system is from the original state equations and the second one is the system of adjoint equations. One necessary condition for the optimality is that at u^*:

$$\frac{\partial H}{\partial u} = 0.$$

Due to the presence of both initial conditions (for the state equations) and final time conditions (for the adjoint equations), and the fact that most models of our interest are nonlinear, the optimal control system has to be solved numerically. We will use the Forward-Backward Sweep Method [1] to conduct the numerical simulation. The steps are described as follows:

Assume that $u = u(t, x, \lambda)$ can be found explicitly from the optimality condition.

Step 1. Make an initial guess for u (usually 0) on the entire domain.

Step 2. Using the initial condition $x(0) = a$ and the values for u, solve x forward in time over the domain.

Step 3. Using the transversality condition $\lambda(T) = b$ (usually 0) and the values for u and x, solve λ backward in time.

Step 4. Update u by the new x and λ values. We use the optimality condition to update control u at this step.

Step 5. Check convergence. If values in this iteration and the last one are negligibly close, output the current values as solutions; otherwise, return to Step 2.

In the next chapter, we will present optimal control in cholera modeling and use this technique to conduct the numerical simulation.

2.2 MATHEMATICAL MODELS ON CHOLERA DYNAMICS

Optimal control theory can be applied to many epidemiological models. In this chapter, we apply optimal control to various cholera models. Cholera is an acute intestinal infectious disease caused by the bacterium Vibrio *cholerae*. The Vibrio *cholerae* could survive for a long time in the water and that an environmental reservoir of Vibrio *cholerae* could be responsible for endemic cholera. Recent cholera outbreaks in Haiti (2010-2011), Nigeria (2010), Kenya (2010), Vietnam (2009), Zimbabwe (2008-2009), etc., continue leading to a large number of infections and receiving worldwide attention [8].

The dynamics of cholera involve multiple interactions between the human host, the pathogen, and the environment [9], which contribute to both direct human-to-human and indirect environment-to-human transmission pathways. In an effort to gain deeper understanding of the complex dynamics of cholera, several mathematical models have been published.

2.2.1 THE KERMACK AND MCKENDRICK MODEL

The very first compartmental model was presented in 1927 by Kermack and McKendrick [2], and it has played a major role in mathematical epidemiology. The model includes three state variables S, I and R, where S represents the susceptible state or the population not yet infected with the disease, I is the state of the population that has been infected and also can spread the disease to the S state, and R is the recovered from the disease state and it is a safe state that cannot give the disease to any other states. The model can be shown in figure 1. The model is a system of differential equations with three state variables

$$\begin{aligned} \frac{dS}{dt} &= -\kappa SI, \\ \frac{dI}{dt} &= \kappa SI - lI, \\ \frac{dR}{dt} &= lI, \end{aligned}$$

FIGURE 1: The SIR model.

where κ is the rate of infection or the contact rate, l is the rate of recovery and $S+I+R=N$, where N is the total population. In their study, the total population was considered as a fixed population and κ and l were constants. The success of the model in predicting disease outbreaks was the beginning of the large body of studies in mathematical epidemiology, particularly, in mathematical modeling of cholera.

Below we summarize several representative cholera models.

2.2.2 THE ROLE OF THE AQUATIC RESERVOIR BY CODEÇO'S WORK

Codeço in 2001 proposed a cholera model [16] that explicitly accounted for the environmental component, i.e., the Vibro *cholerae* concentration in the water supply, into a basic Susceptible-Infective-Recovered model. The incidence (or,the infection force) was modeled by a logistic function to represent the saturation effect. The model was an extension of Capasso's model [3], which was introduced in 1973 to describe the cholera epidemics in Italy and it is shown as follows:

$$\begin{aligned}
\frac{dS}{dt} &= n(H-S) - a\lambda(B)S, \\
\frac{dI}{dt} &= a\lambda(B)S - rI, \\
\frac{dB}{dt} &= B(nb-mb) + eI, \\
S(0) = H, I(0) > 0, B(0) &= 0,
\end{aligned}$$

where $\lambda(B)$ is the probability of a person to catch cholera and it depends on the concentration of Vibro *cholerae* in the consumed water. It can be calculated from the following equation

$$\lambda(B) = \frac{B}{K+B}.$$

The symbols in the equations are listed below:

- S is a state variable that represents the number of susceptibles.
- I is a state variable that denotes the number of infected.
- B is a state variable that represents the concentration of toxigenic Vibro *cholerae* in water.
- H is the total human population.
- n is birth and death rates in the human population.
- a is the rate of exposure to contaminated water.

- K is the concentration of Vibro *cholerae* in water that yields 50% chance of catching cholera.
- r is the rate at which people recovers from *cholerae*.
- nb is the growth rate of Vibro *cholerae* in the aquatic environment.
- mb is the loss rate of Vibro *cholerae* in the aquatic environment.
- e is the contribution of each infected person to the population of Vibro *cholerae* in the aquatic environment.

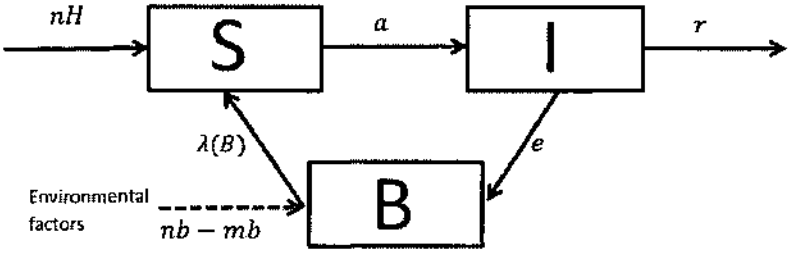

FIGURE 2: The diagrammatic representation of the Codeço model.

2.2.3 THE EXTENSION OF CODÇO'S MODEL BY HARTLEY, MORRIS, AND SMITH

Hartley, Morris and Smith [10] in 2006 extended Codeço's work to include a hyperinfectious state of the pathogen, representing the "explosive" infectivity of freshly shed Vibro *cholerae*, based on the laboratory observations [11, 12]. This model was rigorously analyzed in [51].

The diagram shows that their susceptible state becomes the infected state after consuming concentrations by ingesting water contaminated with rate β_H from HI vibrios or β_L from non-HI vibrios. The modified model now has five state variables

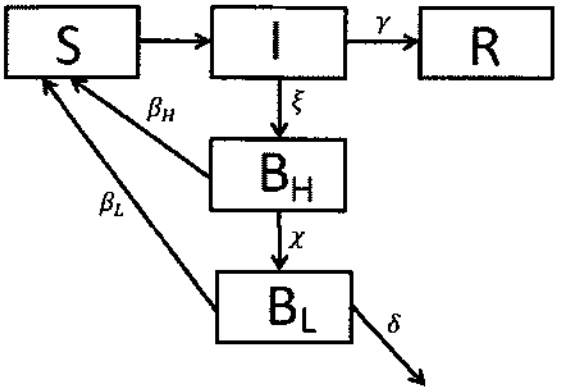

FIGURE 3: The diagrammatic representation of the Hartley, Morris and Smith model.

that give the following system of differential equations

$$\begin{aligned}
\frac{dS}{dt} &= bN - \beta_L S\frac{B_L}{\kappa_L + B_L} - \beta_H S\frac{B_H}{\kappa_H + B_H} - bS,\\
\frac{dI}{dt} &= \beta_L S\frac{B_L}{\kappa_L + B_L} + \beta_H S\frac{B_H}{\kappa_H + B_H} - (\gamma + b)I,\\
\frac{dR}{dt} &= \gamma I - bR,\\
\frac{dB_H}{dt} &= \xi I - \chi B_H,\\
\frac{dB_L}{dt} &= \chi B_H - \delta_L B_L.
\end{aligned}$$

The parameters are described as follows:

- β_L is the rate of drinking non-HI Vibro *cholerae*.
- β_H is the rate of drinking HI Vibro *cholerae*.
- κ_L is the non- HI Vibro *cholerae* concentration.
- b is the natural human birth and death rate.
- χ is the rate of decay from hyper- to reduced infectiousness.
- ξ is the rate of contribution to HI Vibro *cholerae*.
- δ_L is the net of non-HI vibrios in the environment.

- γ is the rate of recovery from cholera.

2.2.4 A THRESHOLD PATHOGEN DENSITY MODEL

Joh, Wang, Weiss et al, [4] in 2009 Modified Codeço's model by a threshold pathogen density for infection, with a careful discussion on human-environment contact and in-reservoir pathogen dynamics. They called their model an iSIR model where "i" represents the indirect transmission dynamics from contact with reservoirs containing human pathogens but not from human to human directly as shown below as state B. They assumed that there is an explicit incorporation of a minimum

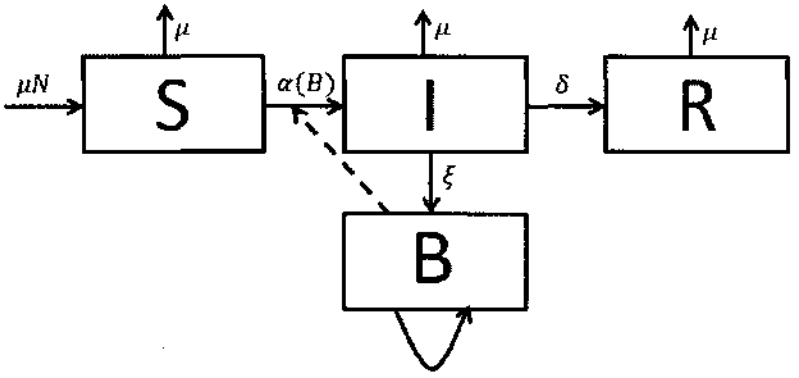

FIGURE 4: The diagram shows the state B which denotes the pathogen density along with other main state variables in the iSIR model.

infectious dose of pathogen to cause infection, and it is an increasing function $\alpha(B)$ in the system of differential equations:

$$\begin{aligned}
\frac{dS}{dt} &= -\alpha(B)S - \mu S + \mu N, \\
\frac{dI}{dt} &= \alpha(B)S - \mu I - \delta I, \\
\frac{dR}{dt} &= \delta I - \mu R, \\
\frac{dB}{dt} &= \pi(B) + \xi I,
\end{aligned}$$

where $\alpha(B)$ is the transmittability of the disease function:

$$\alpha(B) = \begin{cases} 0 & (\mathrm{B} < \mathrm{c}) \\ \frac{a(B-c)^n}{(B-c)^n + H^n} & (\mathrm{B} \geq \mathrm{c}), \end{cases}$$

where n is a positive integer. The additional parameters are described below.

- π is the pathogen growth rate.
- μ is the per capita human birth or death rate.
- δ is the recovery rate. .
- ξ is the pathogen shed rate.
- a is the maximum rate of infection.
- c is the threshold pathogen density of infection.
- r is the maximum per capita pathogen growth efficiency.
- H is the half-saturation pathogen density.

2.2.5 ESTIMATING THE REPRODUCTIVE NUMBERS FOR THE ZIMBABWEAN CHOLERA OUTBREAK

More recently, Mukandavire *et al.* [13] proposed a model to study the 2008-2009 cholera outbreak in Zimbabwe. The model explicitly considered both human-to-human and environment-to-human transmission pathways. The results in this work demonstrated the importance of the human-to-human transmission in cholera epidemics, especially in such places as Zimbabwe, a land-locked country in the middle of Africa.

In this model susceptible individuals become infected by ingesting environmental vibrios with the rates of ingestion of hyperinfectious vibrios given as,

$$\lambda_e = \frac{\beta_e B}{\kappa + B} \qquad \text{and} \qquad \lambda_h = \beta_h I,$$

where e represents environment-to-human transmission and h denotes human-to-human transmission. The parameters β_e and β_h are rates of ingesting vibrios from the contaminated environment and through human-to-human , respectively. The system of differential equations of this model is

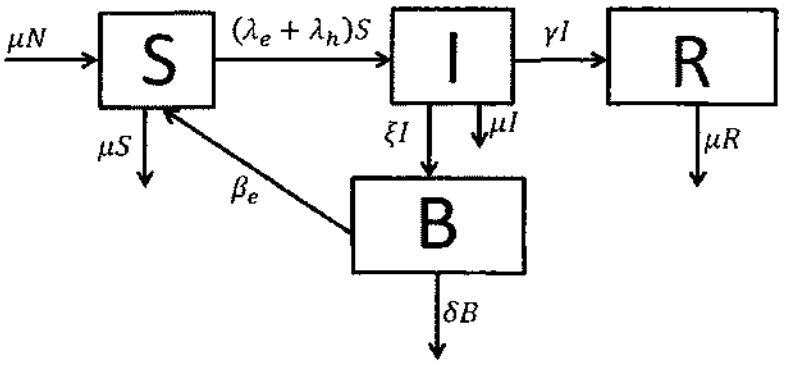

FIGURE 5: The model flow diagram.

$$\begin{aligned}
\frac{dS}{dt} &= \mu N - \beta_e S\frac{B}{\kappa + B} - \beta_h SI - \mu S,\\
\frac{dI}{dt} &= \beta_e S\frac{B}{\kappa + B} + \beta_h SI - (\gamma + \mu)I,\\
\frac{dR}{dt} &= \gamma I - \mu R,\\
\frac{dB}{dt} &= \xi I - \delta B.
\end{aligned}$$

2.2.6 A WATERBORNE PATHOGEN MODEL

Moreover, Tien and Earn [28] in 2010 published a water-borne disease model which also included the dual transmission pathways, with bilinear incidence rates employed for both the environment-to-human and human-to-human infection routes. No saturation effect was considered in Tien and Earn's work. The model was the extension of the SIR model by adding a water state into the system and called the SIWR model. The model contains four state equations, where "W" stands for pathogen concentration in water reservoir:

$$\begin{aligned}
\frac{dS}{dt} &= \mu N - b_W S - b_I SI - \mu S, \\
\frac{dI}{dt} &= b_W WS + b_I SI - \gamma I - \mu I, \\
\frac{dW}{dt} &= \alpha I - \xi W, \\
\frac{dR}{dt} &= \gamma I - \mu R,
\end{aligned}$$

where b_W here is the transmission rate for water-to-person. The model can be described as a diagram below.

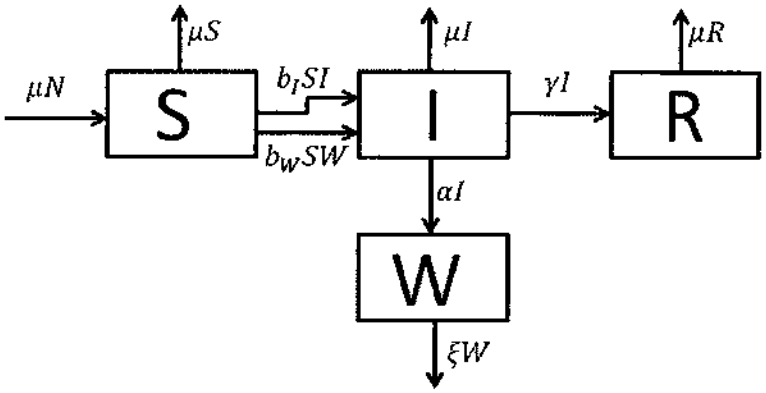

FIGURE 6: The model flow diagram for the SIWR.

2.2.7 A MODEL OF OPTIMAL INTERVENTION STRATEGIES FOR CHOLERA

A rigorous global stability analysis was conducted [19] for many of the aforementioned models. In addition, Neilan *et al.* [20] in 2010 modified the cholera model proposed by Hartley, Morris and Smith [10] and added three control measures into the model. They consequently analyzed the optimal intervention strategies and conducted numerical simulation based on their model. No human-to-human infection route is considered in this work.

The model contains six differential equations and three controls, $u(t)$, $v(t)$ and

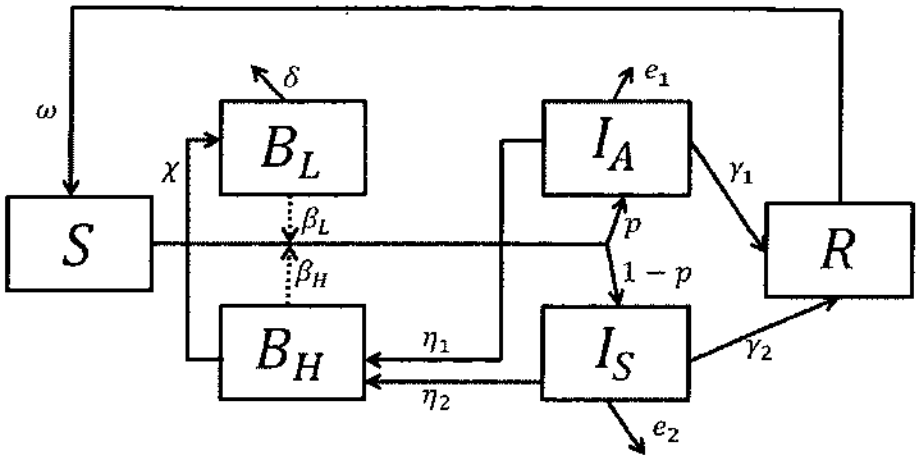

FIGURE 7: The diagram of the model developed by Neilan *et al.* shows two states of B and of I.

$m(t)$, are introduced in the system:

$$\begin{aligned}
\frac{dS}{dt} &= -(1-m(t))\left[\beta_L \frac{B_L(t)}{\kappa_L + B_L(t)} + \beta_H \frac{B_H(t)}{\kappa_H + B_H(t)}\right] S(t) + \omega R(t) - v(t)S(t), \\
\frac{dI_A}{dt} &= p(1-m(t))\left[\beta_L \frac{B_L(t)}{\kappa_L + B_L(t)} + \beta_H \frac{B_H(t)}{\kappa_H + B_H(t)}\right] S(t) - (e_1+\gamma_1)I_A(t), \\
\frac{dI_S}{dt} &= (1-p)(1-m(t))\left[\beta_L \frac{B_L(t)}{\kappa_L + B_L(t)} + \beta_H \frac{B_H(t)}{\kappa_H + B_H(t)}\right] S(t) \\
&\quad - (1-u(t))(e_2+\gamma_2)I_S(t) - u(t)(e_3+\gamma_3)I_S(t) \\
\frac{dR}{dt} &= v(t)S(t) + \gamma_1 I_A(t) + \gamma_2(1-u(t))I_S(t) + \gamma_3 u(t) I_S(t) - \omega R(t), \\
\frac{dB_H}{dt} &= \eta_1 I_A(t) + \eta_2 I_S(t) - \xi B_H(t), \\
\frac{dB_L}{dt} &= \xi B_H(t) - \delta B_L(t),
\end{aligned}$$

where I_S refers to individuals with symptomatic infections, and I_A refers to those with asymptomatic infections. In addition, the control $u(t)$ represents the proportion of people with symptomatic infections who receive combined rehydration and antibiotic treatment, $v(t)$ denotes the vaccination rate that moves susceptible individuals to the immune class, and $m(t)$ is the sanitation rate. All other parameters are described below.

- p is the proportion of infections being asymptomatic.
- e_1 is the cholera-related death rate(asymptomatic).
- e_2 is the cholera-related death rate(symptomatic).
- e_3 is the cholera-related death rate (symptomatic with treatment).
- γ_1 is the recovery rate(asymptomatic).
- γ_2 is the recovery rate(symptomatic).
- γ_3 is the recovery rate(symptomatic with treatment).
- η_1 is the shedding rate(asymptomatic).
- η_2 is the shedding rate(symptomatic).
- κ_H is the half saturation constant(hyperinf.).
- κ_L is the half saturation constant(less-inf.).
- β_H is the ingestion rate(hyperinf.).
- β_L is the ingestion rate(less-inf.).
- ω is the immunity wanning rate.
- δ is the bacteria death rate.
- ξ is the bacteria transition rate.

CHAPTER 3

OPTIMAL CONTROL IN CHOLERA MODELING

3.1 MODELING CHOLERA DYNAMICS WITH CONTROLS

We aim to better understand the effects of different control measures coupled with multiple transmission pathways of cholera, so as to gain useful guidelines for the effective prevention and intervention strategies against cholera epidemics. To that end, we study cholera dynamics with control measures incorporated into the model of Mukandavire *et al.* [13] which involve both the environment-to-human and human-to-human transmission modes. We modify the original model by adding three types of controls: vaccination, therapeutic treatment (including hydration therapy, antibiotics, etc.), and water sanitation. In general, these control measures are functions of time. We will examine how the effects and costs of control measures can be best balanced. Specifically, we will formulate a state-adjoint system and derive the necessary conditions for the optimal control strategies. We will then use numerical simulation to explore various optimal control solutions involving single and multiple controls.

3.1.1 MATHEMATICAL MODEL

Let $S(t)$, $I(t)$ and $R(t)$ denote the susceptible, the infected, and the recovered human population sets, respectively. The total population $N = S + I + R$ is assumed to be a constant, which is a reasonable assumption for a relatively short period of time and for low-mortality disease such as cholera. Let also B denote the concentration of the vibrios in the environment(e.g., contaminated water). The cholera model developed in [13] is a combined system of human populations and the environmental component(SIR-B), with the environment-to-human transmission represented by a logistic(or, Michaelis-Menten type) function and the human-to-human transmission by the standard mass action law.

We now extend this model by adding vaccination, treatment and water sanitation. We assume these controls are implemented continuously; specifically, we make the following assumptions:

- Vaccination is introduced to the susceptible population at a rate of $v(t)$, so that $v(t)S(t)$ individuals per time are removed from the susceptible class and added to the recovered class.

- Therapeutic treatment is applied to the infected people at a rate of $a(t)$, so that $a(t)I(t)$ individuals per time are removed from the infected class and added to the recovered class.

- Water sanitation leads to the death of vibrios at a rate of $w(t)$.

As a result, we obtain the following dynamical system:

$$\frac{dS}{dt} = \mu N - \beta_e S\frac{B}{\kappa + B} - \beta_h SI - \mu S - v(t)S, \tag{1}$$

$$\frac{dI}{dt} = \beta_e S\frac{B}{\kappa + B} + \beta_h SI - (\gamma + \mu)I - a(t)I, \tag{2}$$

$$\frac{dB}{dt} = \xi I - \delta B - w(t)B. \tag{3}$$

In addition, we have the equation for R:

$$\frac{dR}{dt} = \gamma I - \mu R + a(t)I + v(t)S, \tag{4}$$

though this equation is not needed in the model analysis since $R = N - S - I$. The diagram of this model is the same as Figure 5. In this system, the parameters $\mu, \xi, \delta, \gamma, \kappa, \beta_e$, and β_h are all positive constants; μ denotes the natural human birth/death rate, ξ is the rate of human contribution (e.g., shedding) to Vibro *cholerae*, δ is the natural death rate of Vibro *cholerae*, γ is the rate of recovery from cholera, κ is the pathogen concentration that yields 50% chance of catching cholera, and β_e and β_h represent rates of ingesting virbrios from the contaminated water and through human-to-human interaction, respectively. A typical set of numerical values for these parameters are listed in Table 1. In particular, when all controls are set to zero, i.e., $v = a = w = 0$, the above system is reduced to the original model developed in [13].

Parameter	Symbol	Value	Source
Total population	N	10,0000	
Natural human birth and death rate	μ	$(43.5yr)^{-1}$	[14]
Concentration of Vibro*cholera* in environment	κ	10^6 cells/ml	[16]
Rate of recovery from cholera	γ	$(5day)^{-1}$	[10]
Rate of human contribution to Vibro *cholerae*	ξ	10 cells/ml-day	[10]
Death rate of vibrios in the environment	δ	$(30day)^{-1}$	[10]
Ingestion rate from the environment	β_e	0.075/day	[13]
Ingestion rate through human-human interaction	β_h	0.00011/day	[13]

TABLE. 1: Cholera model parameters and values.

3.1.2 OPTIMAL CONTROL STUDY

Now we turn to optimal control study on the model (1 - 3) with time-dependent controls $v(t), a(t)$ and $w(t)$. We consider the system on a time interval [0,T]. The functions $v(t), a(t)$ and $w(t)$ are assumed to be at least Lebesgue measurable on [0,T]. The control set is defined as

$$\Omega = \{(v(t), a(t), w(t)) | 0 \leq v(t) \leq v_{\max}, 0 \leq a(t) \leq a_{\max}, 0 \leq w(t) \leq w_{\max}\},$$

where $v_{\max}, a_{\max}$ and $w_{\max}$ denote the upper bounds for the effort of vaccination, treatment, and sanitation, respectively. These bounds reflect practical limitations on the maximum rates of controls in a given time period.

The presence of time-dependent controls makes the analysis of the system (1 - 2) difficult. In fact, the disease dynamics now depend on the evolution of each control profile. In what follows, we perform an optimal control study on this problem. We aim to minimize the total number of infections and the costs of controls over the time interval [0,T]; i.e.,

$$\min_{(v,a,w)\in\Omega} \int_0^T \left[I(t) + c_{21}v(t)S(t) + c_{22}v(t)^2 + c_{31}a(t)I(t) + c_{32}a(t)^2 + c_{41}w(t) + c_{42}w(t)^2 \right] dt. \tag{5}$$

Here the parameters c_{ij} $(i = 2, 3, 4; j = 1, 2)$, with appropriate units, define the appropriate costs associated with these controls. Quadratic terms are introduced to indicate nonlinear costs potentially arising at high intervention levels [17, 18, 20]. Particularly, the cost terms associated with the sanitation, $c_{41}w(t) + c_{42}w(t)^2$, are taken from [20]. The minimization process is subject to the differential equations (1 - 3), which are now referred to as the state equations. Correspondingly, the unknowns S, I and B are now called the state variables, in contrast to the *control variables* v, a and w. Our goal is to determine the optimal control, $v^*(t), a^*(t)$ and $w^*(t)$, so as to minimize the objective functional in (5).

Here we first establish the following theorem on the existence of optimal control:

Theorem 2. *There exists $(v^*(t), a^*(t), w^*(t)) \in \Omega$ such that the objective functional in (5) is minimized.*

Note that the control set Ω is closed and convex, and the integrand of the objective functional in (5) is convex. Therefore, based on the standard optimal control theorem

in [22], the conditions for the existence of optimal control are satisfied, as our model is linear in the control variables.

We will follow the method described in [1, 17], also briefly reviewed in Chapter 2 of this thesis, to seek the optimal control solution. This method is based on Pontryagin's Maximum Principle [21] which introduces the adjoint functions and represents an optimal control in terms of the state and adjoint functions. Essentially, this approach transfers the problem of minimizing the objective functional (under the constraint of the state equations) into minimizing the Hamiltonian with respect to the controls.

Let us first define the adjoint functions λ_S, λ_I and λ_B associated with the state equations for S, I and B, respectively. We then form the Hamiltonian, H, by multiplying each adjoint function with the right-hand side of its corresponding state equation, and adding each of these products to the integrand of the objective functional. As a result, we obtain

$$\begin{aligned} H &= I(t) + c_{21}v(t)S(t) + c_{22}v(t)^2 + c_{31}a(t)I(t) + c_{32}a(t)^2 + c_{41}w(t) + c_{42}w(t)^2 \\ &+ \lambda_S\Big[\mu N - \beta_e S\frac{B}{\kappa+B} - \beta_h SI - \mu S - v(t)S\Big] \\ &+ \lambda_I\Big[\beta_e S\frac{B}{\kappa+B} + \beta_h SI - (\gamma+\mu)I - a(t)I\Big] \\ &+ \lambda_B\Big[\xi I - \delta B - w(t)B\Big]. \end{aligned}$$

To achieve the optimal control, the adjoint functions must satisfy

$$\begin{aligned} \frac{d\lambda_S}{dt} &= -\frac{\partial H}{\partial S} = -c_{21}v(t) + \lambda_S\Big[\beta_e\frac{B}{\kappa+B} + \beta_h I + \mu + v(t)\Big] \\ &- \lambda_I\Big[\beta_e\frac{B}{\kappa+B} + \beta_h I\Big], \end{aligned} \tag{6}$$

$$\frac{d\lambda_I}{dt} = -\frac{\partial H}{\partial I} = -1 - c_{31}a(t) + \lambda_S\beta_h S - \lambda_I\Big[\beta_h S - (\gamma+\mu) - a(t)\Big] - \lambda_B\xi, \tag{7}$$

$$\frac{d\lambda_B}{dt} = -\frac{\partial H}{\partial B} = \lambda_S\beta_e\frac{\kappa S}{(\kappa+B)^2} - \lambda_I\beta_e\frac{\kappa S}{(\kappa+B)^2} + \lambda_B\Big[\delta + w(t)\Big], \tag{8}$$

with transversality conditions (or final time conditions):

$$\lambda_S(T) = 0, \qquad \lambda_I(T) = 0, \qquad \lambda_B(T) = 0. \tag{9}$$

The characterizations of the optimal controls, $v^*(t), a^*(t)$ and $w^*(t)$, are based on the conditions

$$\frac{\partial H}{\partial v} = 0, \qquad \frac{\partial H}{\partial a} = 0, \qquad \frac{\partial H}{\partial w} = 0, \tag{10}$$

respectively, subject to the constraints $0 \leq v \leq v_{\max}, 0 \leq a \leq a_{\max}$, and $0 \leq w \leq w_{\max}$. Specially, we have

$$v^*(t) = \max\Big[0, \min(\tilde{v}(t), v_{\max})\Big], \tag{11}$$

$$a^*(t) = \max\Big[0, \min(\tilde{a}(t), a_{\max})\Big], \tag{12}$$

$$w^*(t) = \max\Big[0, \min(\tilde{w}(t), w_{\max})\Big], \tag{13}$$

where

$$\tilde{v}(t) = \frac{(\lambda_S - c_{21})S(t)}{2c_{22}}, \tag{14}$$

$$\tilde{a}(t) = \frac{(\lambda_I - c_{31})I(t)}{2c_{32}}, \tag{15}$$

$$\tilde{w}(t) = \frac{\lambda_B B(t) - c_{41}}{2c_{42}}, \tag{16}$$

We summarize the above results by the theorem below:

Theorem 3. *Given an optimal control $(v^*(t), a^*(t), w^*(t))$ and corresponding solutions to the state equations (1 - 3), there exist adjoint variables satisfying the system (6 - 8). Furthermore, the optimal control of the problem (5) is represented by (11 - 13).*

The overall optimal system, which consists of the state equations with the initial conditions, the adjoint equations with the transversality conditions, and the optimal control characterization, has to be solved numerically. We apply the Forward-Backward Sweep Method [1] to solve the optimality system in an iterative manner. First, the state equations (1 - 3) are solved forward in time by a fourth-order Runge-Kutta method, with an initial guess for the control variables. Next, the adjoint equations (6 - 8) are solved backward in time using the solutions of the state equations. The control is then updated with the new values of the state and adjoint solutions, and the process is repeated until the solutions converge. See chapter 2 for a more detailed description.

To carry out the numerical simulation, we list the values for the various transmission rates in the state equations (1 - 3) in Table 1. Particularly, we take the values of μ, β_e and β_h from Zimbabwean cholera data [13]; their values are thus specific to Zimbabwe and may be different for other cholera endemic places. Meanwhile, the cost parameters in (5) are assigned with appropriate values [20]. We also set the initial infection number $I(0) = 1000$ and the entire period of time $T = 100$ days.

We first consider the following set of values for the cost parameters

(17) $c_{21} = 2, \quad c_{22} = 10, \quad c_{31} = 10, \quad c_{32} = 10, \quad c_{41} = 10, \quad c_{42} = 20.$

The per capita cost for vaccination, c_{21}, takes a lower value than other costs, based on the fact that vaccination is usually the most commonly used intervention strategy for various infectious diseases. In particular, the World Health Organization [14] has recently strengthened its recommendation for using oral cholera vaccines to control epidemic and endemic cholera.

Figure 8 shows the infection curves for the model without controls, i.e., $v = a = w = 0$, and that with the optimal controls. It is clearly seen that the infection level has been significantly reduced due to the incorporation of the three types of controls. For comparison, let us also consider the case with vaccination being the only control measure. The optimal control problem can be reformulated to determine the optimal strategy for vaccination, by simply setting the other two controls to zero (i.e., $a = w = 0$) and using the same cost parameters for vaccination. The infection curve with this vaccination only strategy is also shown in Figure 8. As can be expected, the infection level with vaccination only is slightly higher than that with multiple controls, yet it still shows significant improvement compared to the no-control infection curve.

Figures 10 and 11 show the profiles of the optimal vaccination rates in these two cases, i.e., with three controls combined and with vaccination only. We observe a common pattern that the optimal vaccination rates are at the maximum ($u_{\max} = 0.7$) initially and remain at that level for several days (about 7 days for the first case, and 9 days for the second case), before decreasing to almost zero. The shorter period that the maximum vaccination rate stays in the first case is due to the combination of the other two types of controls. Additionally, we sketch the profiles of the optimal treatment rate and sanitation rate in Figure 13. We observe that the therapeutic treatment starts with the maximum rate ($v_{max} = 0.5$) but rapidly decays to a level close to zero, whereas the sanitation rate remains at a relatively low level for a much longer period of time.

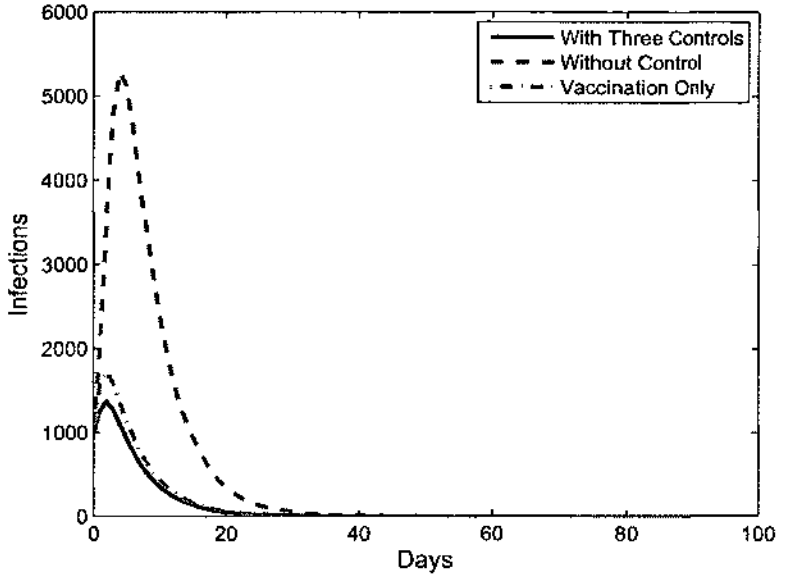

FIGURE 8: Infection curves for the cholera without control ($v = a = w = 0$), with three controls in optimal balance, and with vaccination only ($a = w = 0$) in optimal setting, based on the cost parameters in (17).

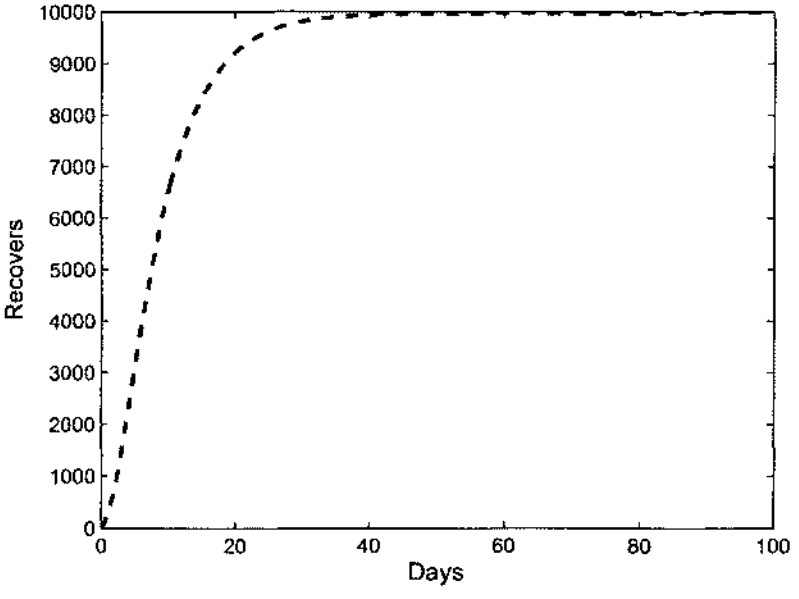

FIGURE 9: The figure shows a regular shape of recover state for the model without control.

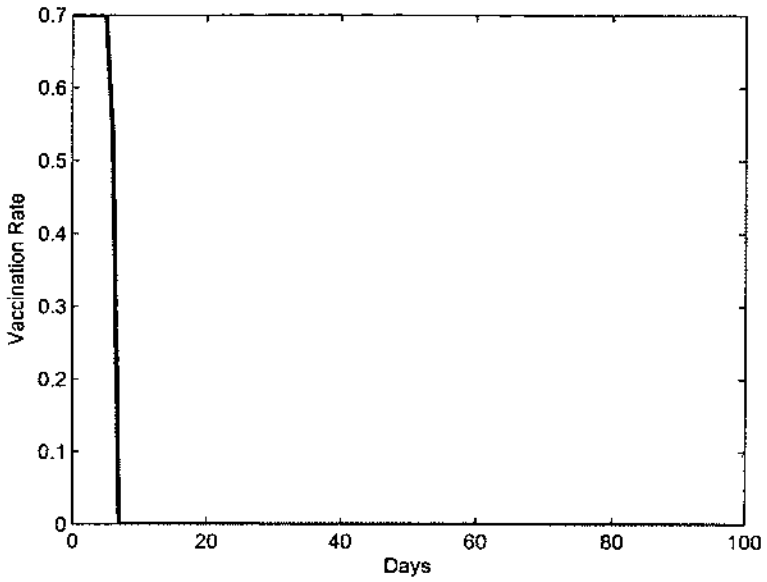

FIGURE 10: Optimal vaccination rate with three controls based on parameters in (17).

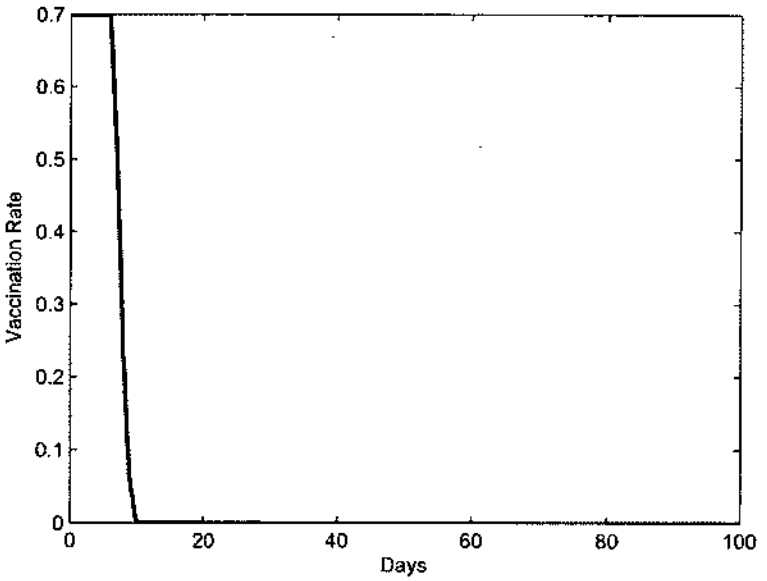

FIGURE 11: Optimal vaccination rate with vaccination only based on parameters in (17).

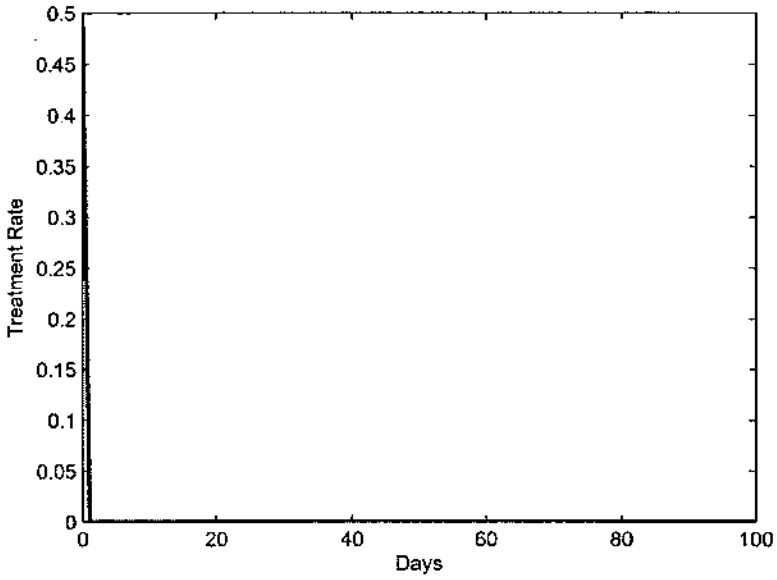

FIGURE 12: Optimal balance of the treatment rate based on parameters in (17).

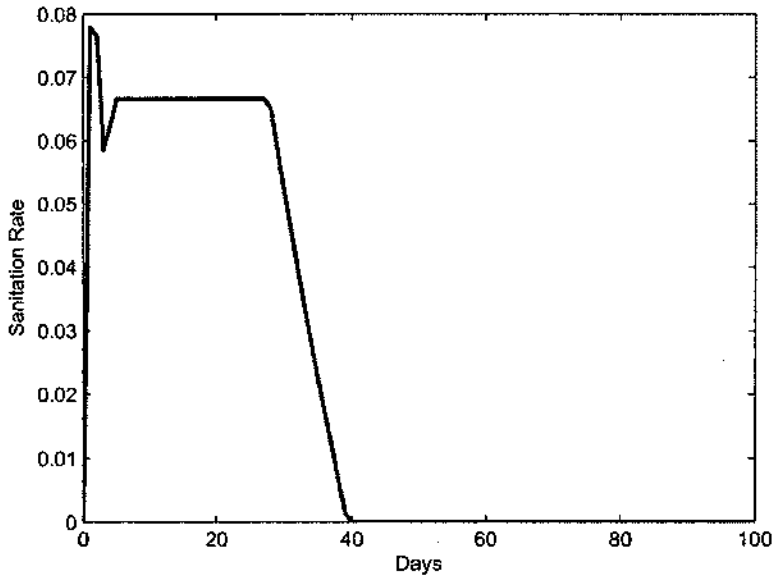

FIGURE 13: Optimal balance of the sanitation rate based on parameters in (17).

Next we consider another set of values for the cost parameters, by decreasing the per capita cost for the therapeutic treatment and increasing the cost for sanitation:

(18) $c_{21} = 2, \quad c_{22} = 10, \quad c_{31} = 2, \quad c_{32} = 10, \quad c_{41} = 100, \quad c_{42} = 20.$

The vaccination cost is kept the same as before. We again conduct simulations for the optimal strategy of the three controls combined and that for vaccination only. The results are presented in Figures 14 - 18.

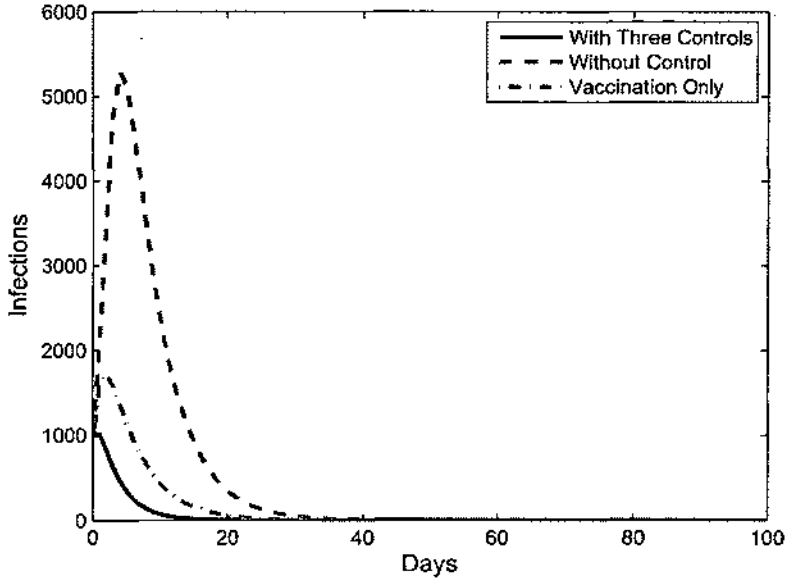

FIGURE 14: Infection curves for the cholera model without control ($v = a = w = 0$), with three controls in optimal balance, and with vaccination only ($a = w = 0$) in optimal setting, based on parameters in (18).

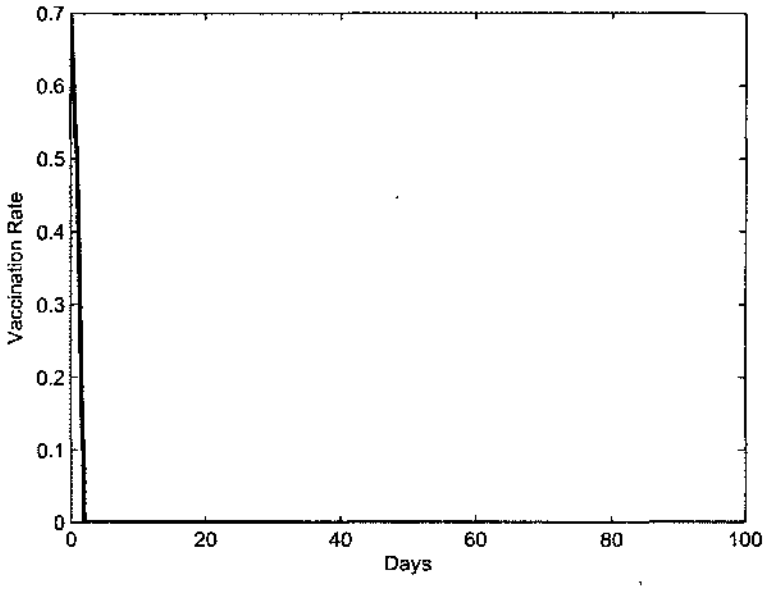

FIGURE 15: Optimal vaccination rate with three controls based on parameters in (18).

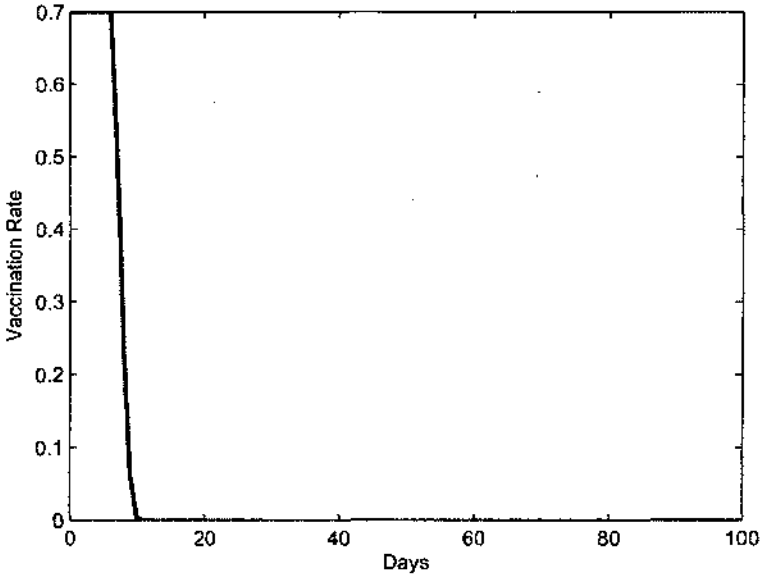

FIGURE 16: Optimal vaccination rate with vaccination only based on parameters in (18).

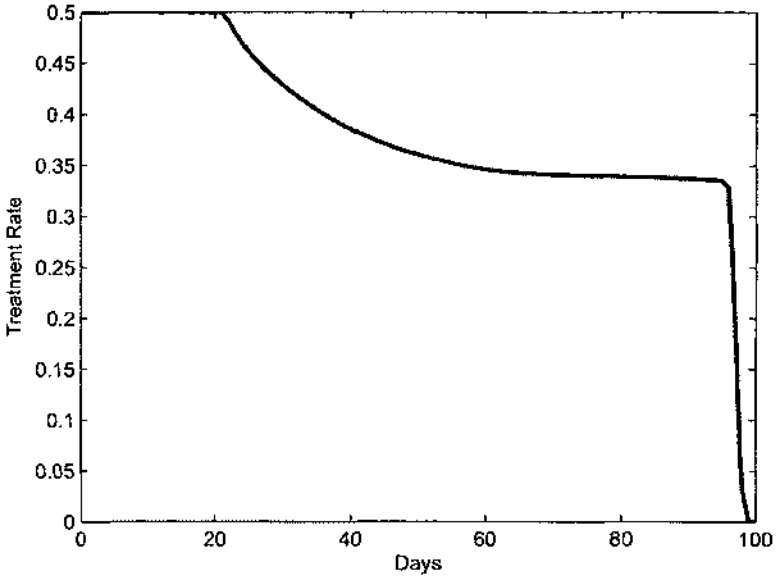

FIGURE 17: Optimal balance of the treatment rate based on parameters in (18).

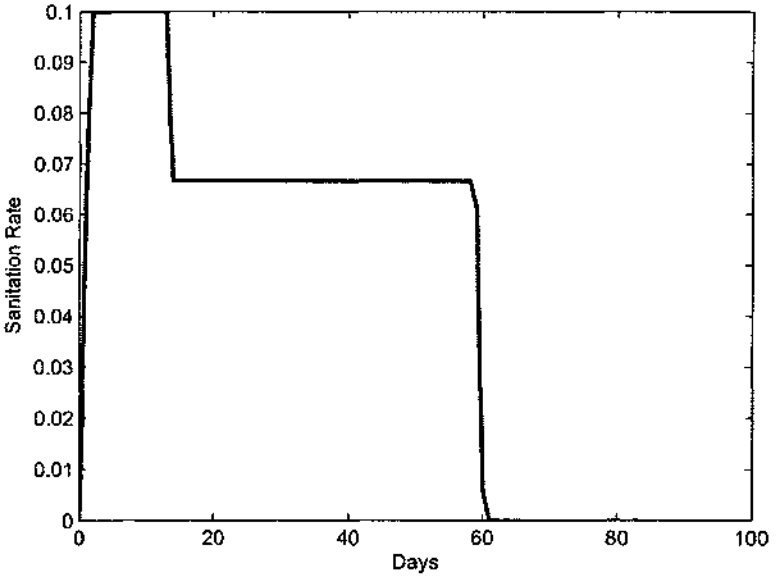

FIGURE 18: Optimal balance of the sanitation rate based on parameters in (18).

With the reduced costs for therapeutic treatment, we observe that both the strength and effective period of the optimal vaccination rate are decreased (See Figure 15) to achieve the optimal balance between controls. The treatment rate starts with the maximum ($v_{\max} = 0.5$) and stays there for more than 20 days, then gradually decays but remains at significant level throughout almost the entire period of 100 days. The increased level of treatment also accounts for the rapid decay of the infection curve starting from the very beginning (see Figure 14). This observation indicates that there is an interaction between the vaccination and treatment in achieving the optimal balance; their relative costs play an important role in determining the length and strength of each control.

In addition, we see there is no significant change to the level of the optimal sanitation rates based on the two different sets of cost parameters (see Figure 13 and 18), which implies that the role of water sanitation in containing a cholera outbreak seems to be minor in the optimal balance of controls, under our model and population settings. Particularly, we note that our model parameters are specific to Zimbabwe, a land-locked country in middle Africa where the level of contact between infected people and the estuarine environment is relatively low.

Finally, we mention that similar patterns are observed for different initial infection sizes and different values of cost parameters, and other results are not shown here.

3.2 A REFINED CHOLERA MODEL WITH OPTIMAL CONTROL

In the previous section, we have presented a basic model to study cholera dynamics with optimal control. We note that the model relies on (unrealistic) assumptions that each control has 100 % efficacy. In what follows, we improve this model by considering more realistic efficacy of the controls, and by introducing a new class of vaccinated people. Thus, our model classifies the human population, denoted by N, into the susceptibles (S), the vaccinated (V), the infected or infectives (I), and the recovered (R).

We assume that individuals are born and die at an average rate μ. The concentration of vibrios in contaminated water is denoted by B. Susceptible individuals acquire cholera infection either by ingesting environmental virbrios from contaminated aquatic reservoirs or through human-to-human transmission, at rates $\lambda_e = (1-\rho)\frac{\beta_e B}{\kappa+B}$ and $\lambda_h = (1-\rho)\beta_h I$ respectively, with the subscripts e and h denoting environment-to-human and human-to-human transmissions, respectively. Here, κ is the pathogen

concentration that yields 50% chance of infection cholera, and $\rho = \epsilon p$ is the sanitation induced preventability to cholera infection, which is a product of the sanitation efficacy ϵ and compliance p. We further assume that susceptible individuals are vaccinated at rate $\phi(t)$, where t is the time variable, with a vaccine that has a degree of protection $\sigma = (1 - \epsilon)$, where ϵ is the vaccine efficacy. Infected individuals are treated at rate $\tau(t)$, and some recover naturally at a rate γ into the recovered class. Infected individuals contribute to Vibro *cholerae* in the aquatic environment at a rate ξ, and vibrios have a net death rate δ in the environment. In addition, water sanitation leads to the death of vibrios at a rate $\nu(t)$.

We thus have the following system of differential equations describing the cholera dynamics with controls:

$$\frac{dS}{dt} = \mu N - (1-\rho)\Big[\beta_e \frac{B}{\kappa + B} + \beta_h I\Big] S - (\phi(t) + \mu) S, \tag{19}$$

$$\frac{dV}{dt} = \phi(t) S - \sigma(1-\rho)\Big[\beta_e \frac{B}{\kappa + B} + \beta_h I\Big] V - \mu V, \tag{20}$$

$$\frac{dI}{dt} = (1-\rho)\Big[\beta_e \frac{B}{\kappa + B} + \beta_h I\Big](S + \sigma V) - (\tau(t) + \gamma + \mu) I, \tag{21}$$

$$\frac{dB}{dt} = \xi I - (\delta + \nu(t)) B, \tag{22}$$

together with

$$\frac{dR}{dt} = (\tau(t) + \gamma) I - \mu R. \tag{23}$$

The diagram of this model is shown in Figure 19. Since N, the total population, is fixed such that $N = S + V + I + R$, we will not need equation (23) in our model analysis.

In what follows, we perform an optimal control study on this problem to explore how the intervention effects, costs can be best balanced, and the control strategy can be best designed to account for the complex and multiple transmission pathways of cholera.

We consider the system on a time interval $[0, T]$ for some $T > 0$. The control is defined as

$$\Gamma = \{(\phi(t), \tau(t), \nu(t)) \quad | \quad 0 \le \phi(t) \le \phi_{\max}, 0 \le \tau(t) \le \tau_{\max}, 0 \le \nu(t) \le \nu_{\max}\}, \tag{24}$$

where $\phi_{\max}, \tau_{\max}$ and $\nu_{\max}$ denote the upper bounds for the effort of vaccination, treatment, and sanitation, respectively. These bounds reflect practical limitations on the maximum rates of controls that can be implemented in a given time period.

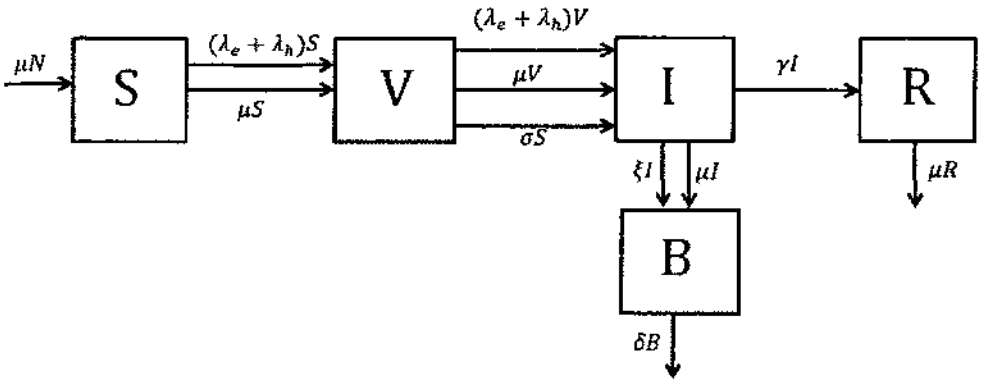

FIGURE 19: The diagram of a refined cholera model with optimal control.

We aim to minimize the total number of infections and the costs of controls over the time interval $[0, T]$;i.e.,

$$\min_{(\phi,\tau,\nu)\in\Gamma} \int_0^T \Big[I(t) + c_{21}\phi(t)S(t) + c_{22}\phi(t)^2 + c_{31}\tau(t)I(t) + c_{32}\tau(t)^2 + c_{41}\nu(t)B(t) + c_{42}\nu(t)^2 \Big] dt. \tag{25}$$

Here again the parameters $c_{ij}, i = 2, 3, 4; j = 1, 2$, define the appropriate costs associated with these controls. Quadratic terms are introduced to account for nonlinear costs potentially arising at high intervention levels [17, 18, 20]. The minimization process is subject to the differential equations in (19 - 22), which we refer to as the state equations. Our goal is then to determine the optimal control, $\phi^*(t), \tau^*(t)$ and $\nu^*(t)$, so as to minimize the objective functional in (25).

We first note that the control set Γ is closed and convex, and the integrand of the objective functional in (25) is also convex. Meanwhile, our model is linear in the control variables. Hence, based on the standard optimal control theorems in [22], we obtain the following theorem.

Theorem 4. *There exists* $(\phi^*(t), \tau^*(t), \nu^*(t)) \in \Omega$ *such that the objective functional in (25) is minimized.*

To proceed, we will again use the Pontryagin's Maximum/Minimum principle [21] to seek the optimal control solution. We define the adjoint functions $\lambda_S, \lambda_V, \lambda_I$ and λ_B associated with the state equations for S, V, I and B, respectively. We then form

the Hamiltonian, H, by multiplying each adjoint function with the right-hand side of its corresponding state equation, and adding each of these products to the integrand of the objective functional. As a result, we obtain

$$\begin{aligned}
H &= I(t) + c_{21}\phi(t)S(t) + c_{22}\phi(t)^2 + c_{31}\tau(t)I(t) + c_{32}\tau(t)^2 + c_{41}\nu(t)B(t) + c_{42}\nu(t)^2 \\
&+ \lambda_S\Big[\mu N - (1-\rho)\big[\beta_e\frac{B}{\kappa+B} + \beta_h I\big]S - (\phi(t)+\mu)S\Big] \\
&+ \lambda_V\Big[\phi(t)S - \sigma(1-\rho)\big[\beta_e\frac{B}{\kappa+B} + \beta_h I\big]V - \mu V\Big] \\
&+ \lambda_I\Big[(1-\rho)\big[\beta_e\frac{B}{\kappa+B} + \beta_h I\big](S+\sigma V) - (\tau(t)+\gamma+\mu)I\Big] \\
&+ \lambda_B\Big[\xi I - (\delta+\nu(t))B\Big].
\end{aligned}$$

To achieve the optimal control, the adjoint functions must satisfy

$$\begin{aligned}
\frac{d\lambda_S}{dt} &= -\frac{\partial H}{\partial S}, \\
\frac{d\lambda_V}{dt} &= -\frac{\partial H}{\partial V}, \\
\frac{d\lambda_I}{dt} &= -\frac{\partial H}{\partial I}, \\
\frac{d\lambda_B}{dt} &= -\frac{\partial H}{\partial B}.
\end{aligned}$$

These yield

$$\begin{aligned}
\frac{d\lambda_S}{dt} &= -c_{21}\phi + \lambda_S\Big[(1-\rho)\big[\beta_e\frac{B}{\kappa+B} + \beta_h I\big] + (\phi+\mu)\Big] \\
&\quad - \lambda_V\phi - \lambda_I\Big[(1-\rho)\big[\beta_e\frac{B}{\kappa+B} + \beta_h I\big]\Big], \qquad (26) \\
\frac{d\lambda_V}{dt} &= \lambda_V\Big[\sigma(1-\rho)\big[\beta_e\frac{B}{\kappa+B} + \beta_h I\big] + \mu\Big] \\
&\quad - \lambda_I\Big[\sigma(1-\rho)\big[\beta_e\frac{B}{\kappa+B} + \beta_h I\big]\Big], \qquad (27) \\
\frac{d\lambda_I}{dt} &= -1 - c_{31}\tau + \lambda_S(1-\rho)\beta_h S + \lambda_V\sigma(1-\rho)\beta_h V \\
&\quad - \lambda_I\big[(1-\rho)\beta_h(S+\sigma V) - (\tau+\gamma+\mu)\big] - \lambda_B\xi, \qquad (28) \\
\frac{d\lambda_B}{dt} &= -c_{41}\nu + \lambda_S(1-\rho)\beta_e\frac{\kappa S}{(\kappa+B)^2} + \lambda_V\sigma(1-\rho)\beta_e\frac{\kappa V}{(\kappa+B)^2} \\
&\quad - \lambda_I(1-\rho)\beta_e\frac{\kappa(S+\sigma V)}{(\kappa+B)^2} + \lambda_B(\delta+\nu), \qquad (29)
\end{aligned}$$

with transversality conditions (i.e., final time conditions):

$$\lambda_S(T) = 0, \quad \lambda_V(T) = 0, \quad \lambda_I(T) = 0, \quad \lambda_B(T) = 0. \tag{30}$$

The characterizations of the optimal controls, $\phi^*(t), \tau^*(t)$ and $\nu^*(t)$, are based on the conditions

$$\frac{\partial H}{\partial \phi} = 0, \quad \frac{\partial H}{\partial \tau} = 0, \quad \frac{\partial H}{\partial \nu} = 0, \tag{31}$$

respectively, subject to the constraints given in (24). Thus we have

$$\phi^*(t) = \max[0, \min(\tilde{\phi}(t), \phi_{\max})], \tag{32}$$

$$\tau^*(t) = \max[0, \min(\tilde{\tau}(t), \tau_{\max})], \tag{33}$$

$$\nu^*(t) = \max[0, \min(\tilde{\nu}(t), \nu_{\max})], \tag{34}$$

where

$$\tilde{\phi}(t) = \frac{(\lambda_S - \lambda_V - c_{21})S(t)}{2c_{22}},$$
$$\tilde{\tau}(t) = \frac{(\lambda_I - c_{31})I(t)}{2c_{32}},$$
$$\tilde{\nu}(t) = \frac{(\lambda_B - c_{41})B(t)}{2c_{42}}.$$

Again to summarize, our optimal control problem consists of the state system (19 - 22) with initial conditions, the adjoint equations (26) - (29) with the transversality conditions, and equations (32) - (34) to characterize the optimal controls. Such a problem has to be solved numerically. Similar to the previous model, we apply the Forward-Backward Sweep Method [1] to solve it in an iterative manner.

The various transmission rates are listed in Table 1. We assume that the unit costs for vaccination and treatment are about the same, whereas the cost per percent reduction of vibrio concentration through sanitation is doubled. We further assume that for each control, the per capita cost at the quadratic level is about 20% of that at the linear level. Thus, we assign the following set of values (with appropriate units) to the cost parameters in (25):

$$c_{21} = 50, \quad c_{22} = 10, \quad c_{31} = 50, \quad c_{32} = 10, \quad c_{41} = 100, \quad c_{42} = 20. \tag{35}$$

Based on practical observations and empirical values [8, 14], we set the upper bounds of the rates for the three controls as $\phi_{\max} = 0.7$, $\tau_{\max} = 0.5$, and $\nu_{\max} = 0.1$.

We also set the initial infection number $I(0) = 1000$ and the entire period of time $T = 100$ days.

Figure 20 shows the infection curves for the model without controls, i.e., $\phi(t) = \tau(t) = \nu(t) = 0$, and that with the optimal controls implemented. We clearly see that the number of infections has been significantly reduced due to the incorporation of the three types of controls. Here we also present the case with vaccination being the only control measure; in this case the optimal control problem is reformulated to determine the optimal strategy for vaccination, by setting $\tau(t) = \nu(t) = 0$ and using the same cost parameters for vaccination. The infection curve with this vaccination-only strategy is shown in Figure 20 with the dash-dot-line. We observe that with vaccination only, the infections still reach a high level, though the peak value is lower than that without controls. This is evidence that multiple intervention methods, which target both the direct and indirect transmission routes of cholera, would achieve better results than a single control such as vaccination only.

Figure 21 - 23 further show the profile of each of the three controls in their optimal balance. We observe that both the vaccination rate and the treatment rate start with their maximum values and remain at that level for a number of days (about 7 days for vaccination, and 35 days treatment), before decreasing to lower levels of strength. The pattern is similar for sanitation, except that the start of maximum sanitation rate lags for 1 - 2 days.

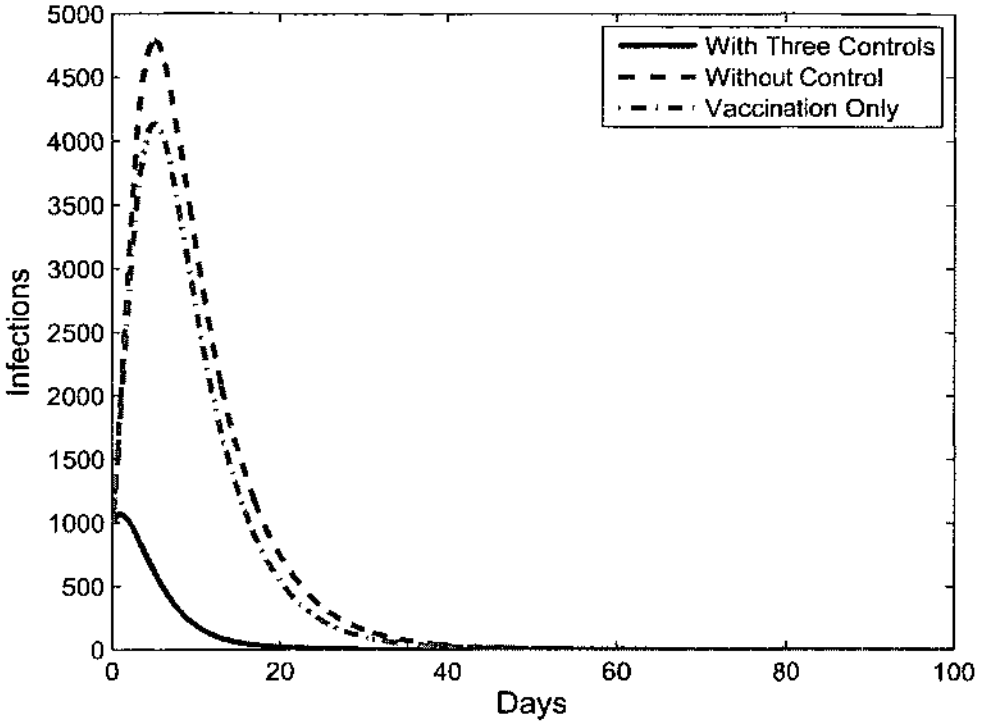

FIGURE 20: The infection curves for the model without controls and with the optimal controls implemented.

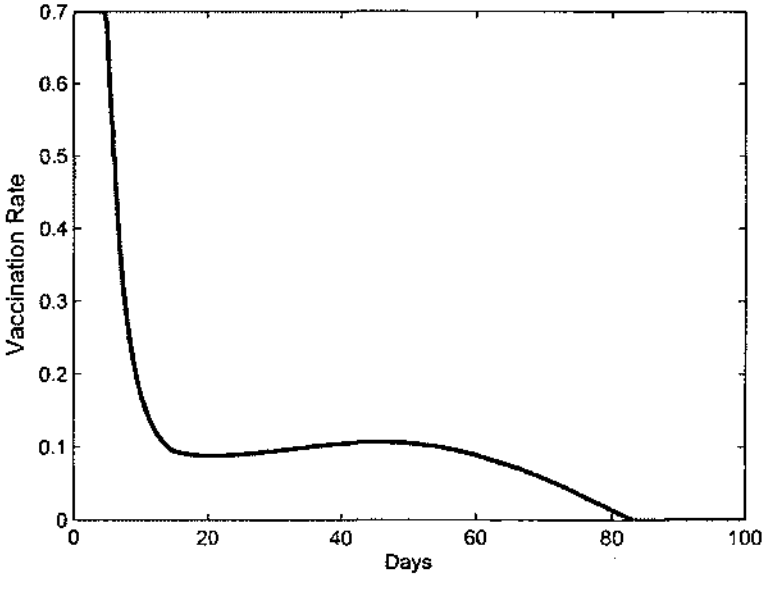

FIGURE 21: The optimal vaccination rate.

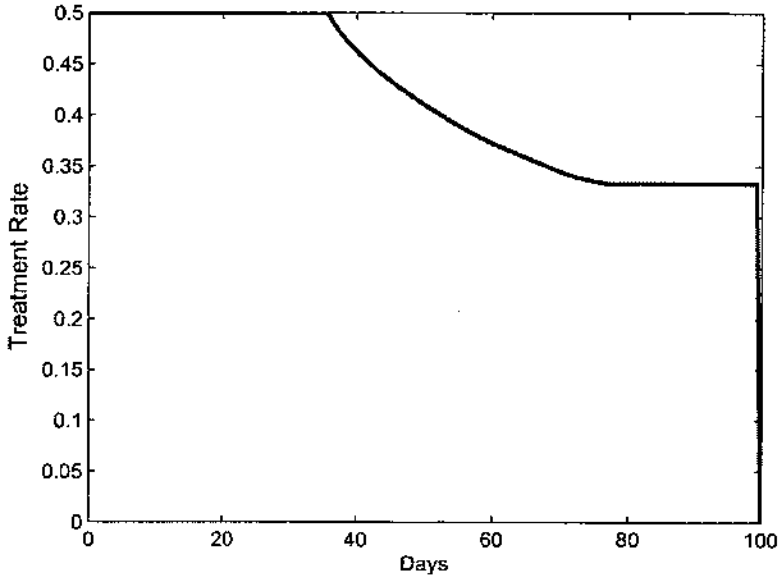

FIGURE 22: The optimal treatment rate.

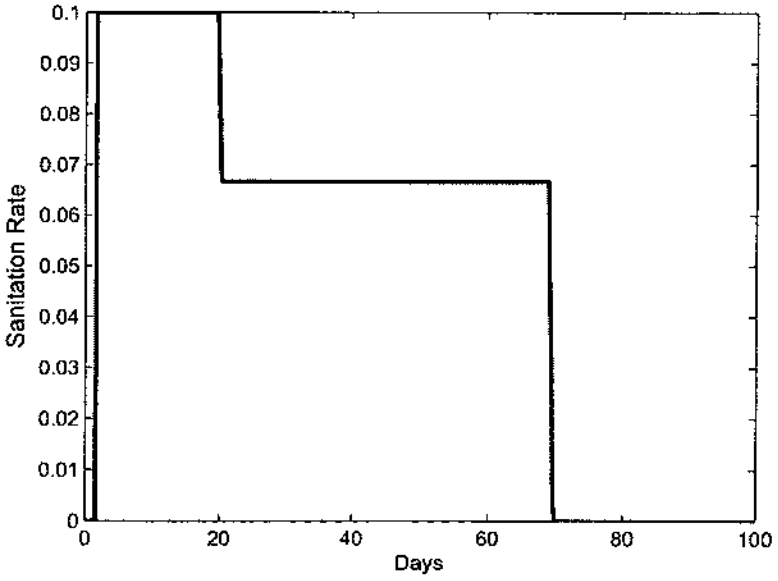

FIGURE 23: The optimal sanitation rate.

As can be expected, the costs of these controls directly affect the strength and duration of the controls in their optimal balance. For demonstration, we now assume that the linear per capita costs for vaccination and treatment (i.e., c_{21} and c_{31}) can be significantly reduced compared to those in (35), and consider the following cost parameters:

$$c_{21} = 2, \quad c_{22} = 10, \quad c_{31} = 2, \quad c_{32} = 10, \quad c_{41} = 100, \quad c_{42} = 20. \tag{36}$$

This set of cost parameters lead to improved control results. The reason is that with lower costs for vaccination and treatment, these two controls can be implemented with higher average strength and longer duration, thus further reducing the infections. Indeed, Figure 24 and 25 show the profiles for the optimal vaccination and treatment rates in this case, where we see that the vaccination and treatment remain at their maximum levels for about 12 days and 60 days, respectively; both are significantly longer than those in Figure 20 - 23. In addition, the vaccination rate now stays at the level of 0.3 from day 20 until near the end of the time interval.

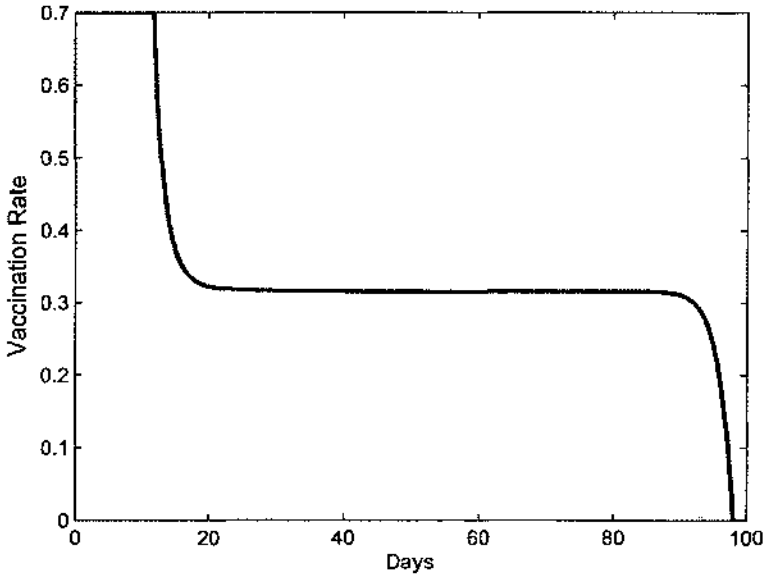

FIGURE 24: The optimal vaccination rate.

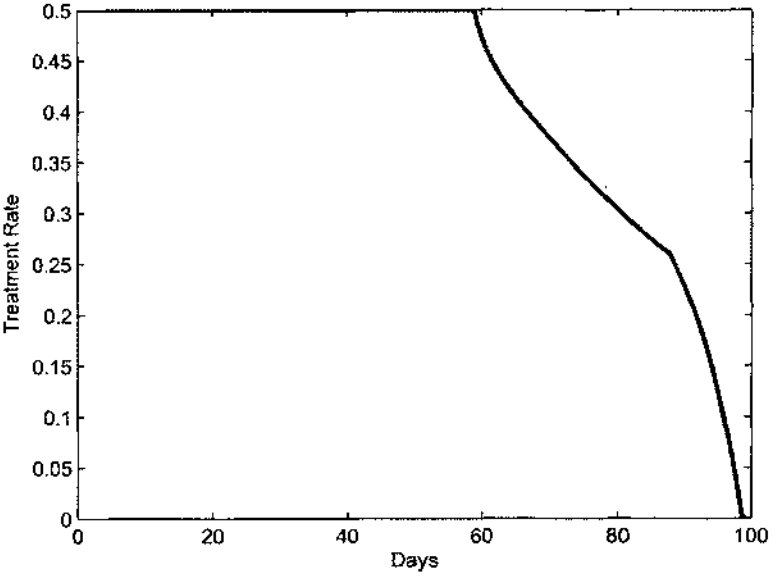

FIGURE 25: The optimal treatment rate.

3.3 SIMULATING OPTIMAL VACCINATION TIMES DURING CHOLERA OUTBREAKS

Although the use of vaccines has been increasingly recognized as an effective control measure in cholera endemic regions, current strategies and experiences for reactive use of cholera vaccines after an outbreak has started to remain limited, and guidelines for cholera vaccination in complex epidemic and emergency settings (e.g., refugee camps) are urgently needed [52]. In particular, question on the optimal time frame to pursue mass vaccination during a cholera outbreak remains to be answered.

In this section, we use mathematical modeling and simulation techniques to shed light on the value of vaccination in controlling ongoing cholera outbreaks. Particularly, we will formulate a new optimal control model and explore optimal times during epidemics for deploying cholera vaccines that best balance the gains and costs of vaccination.

3.3.1 METHODS

We modified our model in Section 3.2 with vaccination as the only control:

$$
\begin{aligned}
\frac{dS}{dt} &= \mu N - (\beta_e \frac{B}{\kappa + B} + \beta_h I)S - (\phi(t) + \mu)S, \\
\frac{dV}{dt} &= \phi(t)S - \sigma(\beta_e \frac{B}{\kappa + B} + \beta_h I)V - \mu V, \\
\frac{dI}{dt} &= (\beta_e \frac{B}{\kappa + B} + \beta_h I)(S + \sigma V) - (\gamma + \mu)I, \\
\frac{dB}{dt} &= \xi I - \delta B.
\end{aligned} \tag{37}
$$

In addition, the equation for the recovered individuals is given by

$$
\frac{dR}{dt} = \gamma I - \mu R. \tag{38}
$$

The diagram of this model is the same as Figure 19. We will construct an optimal control model [17, 18, 20] to seek an answer on the "best time", in terms of the total costs (or efforts), to deploy the vaccine after the onset of an outbreak. We aim to minimize the following objective functional:

$$
\int_0^T \Big[c_0 I(t) + c_1 \phi(t) S(t) \Big] \, dt \, ; \tag{39}
$$

$$
\phi(t) = 0 \ \text{ when } \ t < d \, , \tag{40}
$$

where the parameter c_0 denotes the average cost incurred by each infected individual and c_1 denotes the per capita cost of the vaccines, and where d is the time *when*

vaccination starts. For simplicity, we consider only the linear control in the objective functional in (39). We emphasize that the constraint (40) distinguishes our optimal control model from those common ones [1]. The parameter d can be varied, which allows us to explore the optimal vaccination strategies with different starting times.

Since the realistic value of the cost parameter c_0 is difficult to estimate, we will consider the following objective functional, instead, to make the discussion slightly simpler:

$$(41)\quad J(\phi,\, d) = \int_0^T \Big[I(t) + c_{21}\phi(t)S(t)\Big]\, dt\,; \qquad \phi(t) = 0 \ \text{ when } \ t < d\,.$$

That is, we have normalized the cost of each infection to 1 in the above equation; correspondingly, c_{21} is now the normalized cost parameter associated with vaccination. This objective functional can also be interpreted as a balance between the gain of the control (i.e., reduction of infections) and the cost of the control.

For a given value of d $(0 \le d < T)$, a control set is defined as

$$(42)\qquad \Gamma(d) = \big\{\, \phi(t) \big|\ 0 \le \phi(t) \le \phi_{\max}\,,\ \ \phi(t) = 0 \ \text{ if } \ t < d \,\big\},$$

over which the optimal control problem (41) can be solved. Here $\phi_{\max}$ denotes the upper bound of the vaccination rates which reflects practical limitations on the effort of vaccination that can be pursued in a given cholera epidemic setting.

We first note that for each d, the control set $\Gamma(d)$ is closed and convex, and the integrand of the objective functional in (41) is also convex. Meanwhile, our model is linear in the control variable ϕ. Hence, based on the standard optimal control theorems in [22], there exists a $\phi^*(t)$ for any $0 \le d < T$ such that the objective functional in (41) is minimized. Indeed, the optimal control is also unique for small T due to the Lipschitz property of the state equations and the boundedness of the state variables [22, 24]. In view of this, let us denote

$$(43)\qquad f(d) = J(\phi^*,\, d) = \min_{\phi} J(\phi,\, d).$$

In what follows, we will again apply the Pontryagin's Maximum/Minimum Principle [21] to seek the optimal control solution. We first define the adjoint functions λ_S, λ_V, λ_I, and λ_B associated with the state equations for S, V, I and B, respectively. We then form the Hamiltonian H, using the objective functional in (41) and the state equations in (37):

$$(44)\qquad H = I(t) + c_{21}\phi(t)S(t) + \lambda_S\frac{dS}{dt} + \lambda_V\frac{dV}{dt} + \lambda_I\frac{dI}{dt} + \lambda_B\frac{dB}{dt}\,.$$

To achieve the optimal control, the adjoint functions must satisfy

$$(45)\quad \frac{d\lambda_S}{dt} = -\frac{\partial H}{\partial S}, \quad \frac{d\lambda_V}{dt} = -\frac{\partial H}{\partial V}, \quad \frac{d\lambda_I}{dt} = -\frac{\partial H}{\partial I}, \quad \frac{d\lambda_B}{dt} = -\frac{\partial H}{\partial B}.$$

These yield

$$\begin{aligned}
\frac{d\lambda_S}{dt} &= -c_{21}\phi(t) + \lambda_S(\beta_e\frac{B}{\kappa+B} + \beta_h I) \\
&\quad + \lambda_S(\phi(t)+\mu) - \lambda_V\phi(t) - \lambda_I(\beta_e\frac{B}{\kappa+B} + \beta_h I), \\
\frac{d\lambda_V}{dt} &= \lambda_V\sigma(\beta_e\frac{B}{\kappa+B} + \beta_h I) + \lambda_V\mu - \lambda_I(\sigma(\beta_e\frac{B}{\kappa+B} + \beta_h I)), \\
\frac{d\lambda_I}{dt} &= -1 + \lambda_S\beta_h S + \lambda_V\sigma\beta_h V - \lambda_I(\beta_h(S+\sigma V)) + \lambda_I(\gamma+\mu) - \lambda_B\xi, \\
\frac{d\lambda_B}{dt} &= \lambda_S\beta_e S\frac{\kappa}{(\kappa+B)^2} + \lambda_V\sigma\beta_e V\frac{\kappa}{(\kappa+B)^2} - \lambda_I\beta_e(S+\sigma V)\frac{\kappa}{(\kappa+B)^2} + \lambda_B\delta,
\end{aligned}$$

with transversality conditions (i.e., final time conditions):

$$(46)\qquad \lambda_S(T) = 0, \qquad \lambda_V(T) = 0, \qquad \lambda_I(T) = 0, \qquad \lambda_B(T) = 0,$$

and

$$\frac{\partial H}{\partial \phi} = (c_{21} + \lambda_V - \lambda_S)S(t).$$

Meanwhile, the characterization of the optimal control, $\phi^*(t)$, is based on the switching condition [1, 23]:

$$(47)\qquad \phi^* = \phi_{\max} \text{ if } \frac{\partial H}{\partial \phi} < 0\,; \quad \phi^* = 0 \text{ if } \frac{\partial H}{\partial \phi} > 0\,,$$

subject to the constraints given in (42). Numerically, we have verified that the value of the switching function $\frac{\partial H}{\partial \phi}$ is never zero on a non-empty time interval; that is, the case of singular control does not occur in our optimal control study.

In summary, given the optimal vaccination $\phi^*(t)$ and corresponding solution to the state system (37), there exist adjoint variables satisfying the system (45). Furthermore, the optimal control is characterized by equations (42) and (47). Using the forward-backward sweeping method, we have conducted numerical simulation to our optimal control model, and the simulation results are presented in next section.

3.3.2 RESULTS

We list in Table 2 the model parameters and their values used in our numerical simulation. Figures 26 and 27 show typical scenarios for the infected number versus

time without vaccination (the solid line) and that with vaccination (the dashed line). In Figure 26, vaccination starts right at the beginning of the cholera epidemic, i.e., $d = 0$ in the objective functional in equation (41), whereas in Figure 27, vaccination starts 2 days after the onset of the epidemic, i.e., $d = 2$. For each case, the model is numerically solved to obtain the optimal control solution. The cost parameter is set as $c_{21} = 1.0$ in the optimal control simulation. It is clear that in both cases vaccination significantly reduces the number of infections throughout the cholera outbreak. In particular, the peak value of the infection is about $5,200$ if no vaccination is implemented; this number is reduced to 800 (or, 85% reduction) with optimal vaccination starting from the very beginning, and to $1,500$ (or, 71% reduction) with optimal vaccination starting from day 2.

Figures 28 and 29 show the profiles of the optimal control, $\phi^*(t)$, corresponding to Figures 26 and 27, respectively. In each case, we see that vaccination starts at the maximum rate 70% (note that vaccination starts at $d = 2$ in case b) and remains at this level for most of the 100-day period, then decreases to zero toward the end of the time. A bang-bang type of control is displayed for each case. The vaccination rate in case (a) stays at the maximum strength longer than that in case (b), as the infection curve corresponding to case (a) decays to 0 slightly slower than that corresponding to case (b); see the dashed lines in Figures 26 and 27.

Our focus, however, is to investigate the settings where vaccination is pursued after the onset of the cholera outbreak and to seek optimal times to deploy the vaccines. To that end, we have carefully examined different choices of the cost parameter c_{21} and the vaccination starting time d. For a given value of c_{21}, we vary the value of d in the range $0 < d < T$, where T is set as 100 days; for each value of d, we conduct the optimal control simulation for the objective functional (41) on the domain (42), and evaluate the minimized cost $f(d)$. We then plot the curve $f(d)$ vs. d to display the variation of the minimal cost with respect to the time for starting vaccination after the onset of the outbreak.

We note that when d is large; i.e., when vaccination is applied near or after the end of the cholera outbreak, the effects of vaccination are tiny and $f(d)$ is approximated by $\int_0^T I(t)\,dt \approx 5 \times 10^4$. In fact $f(d) \to 5 \times 10^4$ as $d \to T$. This is demonstrated in Figures 30 and 31. For case (a), $c_{21} = 0.1$; for case (b), $c_{21} = 0.5$. Though the two curves look similar, the difference in c_{21}, the normalized per capita cost for vaccination, is represented by the slight difference of the initial positions, i.e., the

values $f(0)$, between the two curves. We observe that $f(0) \approx 3.4 \times 10^4$ in case (a) and $f(0) \approx 3.8 \times 10^4$ in case (b). We also observe that $f(d)$ is increasing with d for both curves.

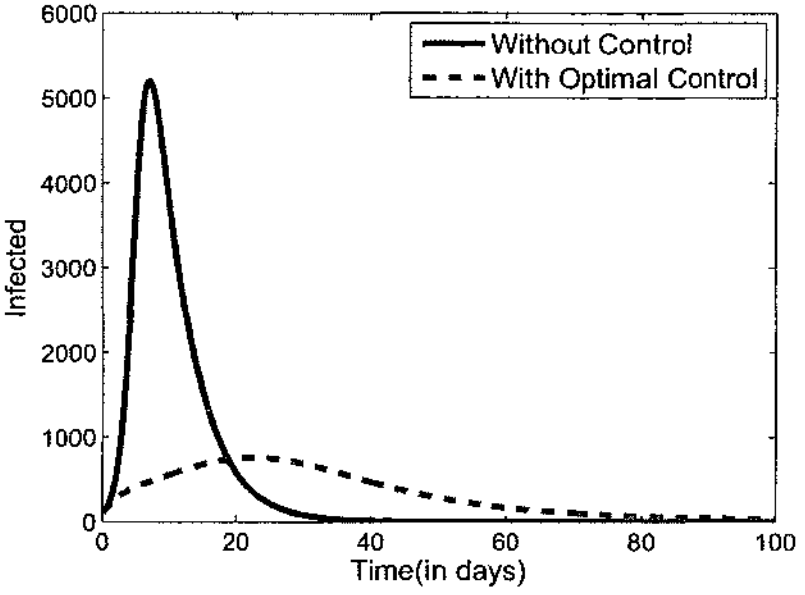

FIGURE 26: The number of infected individuals versus time for the case without vaccination and that with optimal vaccination when optimal vaccination starts from t=0.

Next, we have conducted the simulation using various values of the cost parameter c_{21}. For each value of c_{21}, we run the optimal control simulation for d in the range of $0 < d < T$ and plot the curve for $f(d)$ vs. d. Some typical results are presented in Figure 32 - 35. We clearly observe that as the value of c_{21} increases, the initial value of $f(d)$, i.e., $f(0)$, also increases, a natural consequence of the increased per capita cost of vaccination. Meanwhile, for each value of c_{21}, the curve of $f(d)$ approaches 5×10^4 as $d \to T$. The more interesting pattern, however, is that different sizes of c_{21} lead to two different behaviors of the curve $f(d)$. We found that when $c_{21} < 1.4$ (*e.g.*, Figure 30 - 33), $f(d)$ always increases with d and approaches 5×10^4 eventually. In contrast, when $c_{21} \geq 1.4$ (see Figure 34 and 35 for two typical examples), then $f(d)$ remains approximately a constant, 5×10^4, for all values of d. This latter scenario illustrates an extreme in the control strategy, $\phi^* = 0$; i.e., no vaccination. The implication is that if the unit cost of the vaccines is too high (i.e., higher than a certain value), then we should give up vaccination to achieve the optimal control,

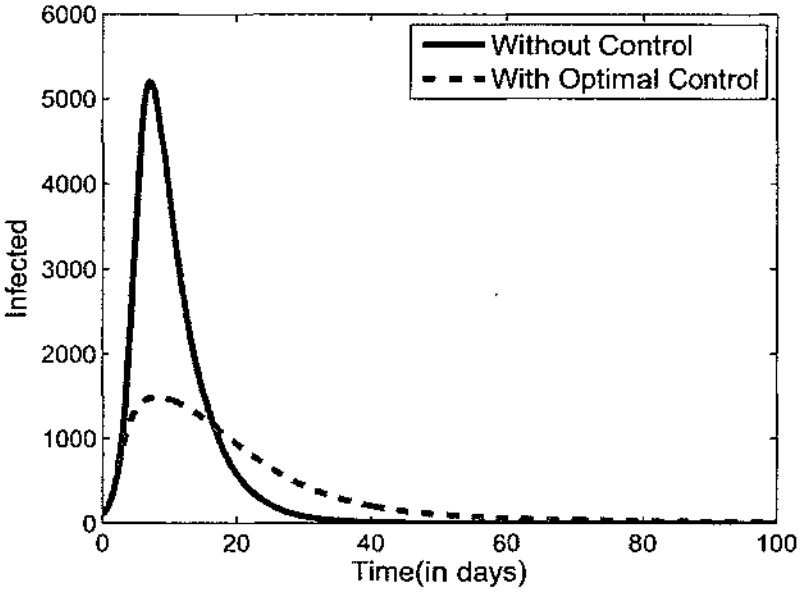

FIGURE 27: The number of infected individuals versus time for the case without vaccination and that with optimal vaccination when optimal vaccination starts from t=2.

in the sense to minimize the objective functional in (41).

Based on our model, the main numerical observation is that if the normalized per capita cost of the cholera vaccine is lower than a critical value, $c_{21}^* \approx 1.4$, then the earlier the vaccine is deployed, the better; in other words, the optimal time would be $d = 0$. In contrast, if c_{21} is higher than the critical value c_{21}^*, then vaccination would not contribute to the optimal control; consequently, there is no optimal time for vaccination deployment.

3.3.3 DISCUSSION

In this section we have formulated an optimal control model to analyze the effects of cholera vaccines in epidemic settings in an attempt to understand optimal times to pursue vaccines that best balance the gains and costs of vaccination for cholera.

The optimal strategy depends on what we aim to optimize. Obviously, if the sole purpose of the disease control is to minimize the total number of infected people, i.e.,

$$\min \int_0^T I(t)\, dt, \tag{48}$$

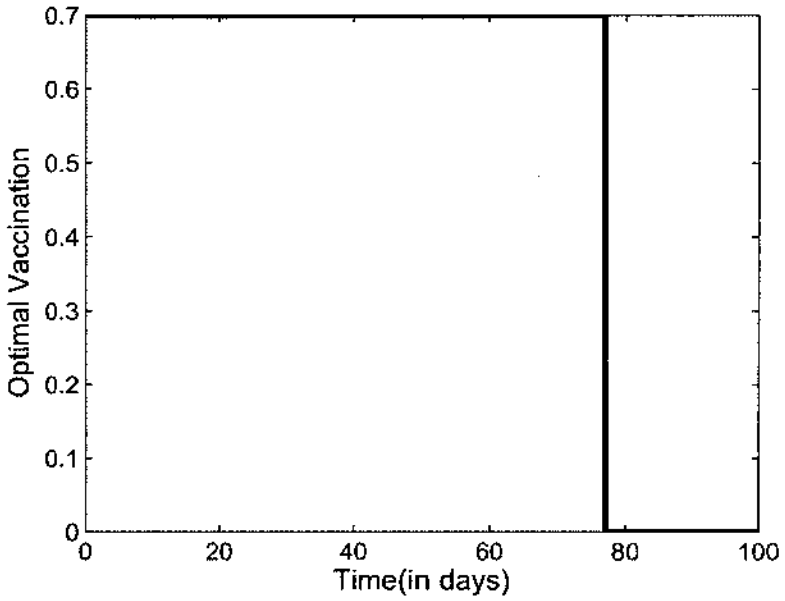

FIGURE 28: Case(a): Profile of the optimal control, $\phi^*(t)$ when optimal vaccination starts from t=0.

then vaccination should always be applied regardless of the price and should always be applied in the very beginning ($t = 0$) to achieve this minimum. However, medical resources are generally limited and cost factors have to be taken into account. Meanwhile, immediate response with mass vaccination at the onset of a cholera outbreak may not always be feasible. In this study, we have incorporated the costs of vaccination into the objective functional so as to seek an optimal balance between the reduction of infection and the costs of the control measure. Our results imply that as long as the cholera vaccine prices are sufficiently low, vaccination should be deployed and should always start from, or immediately after, the onset of a cholera outbreak. If, however, the vaccine prices are higher than a certain value, then other types of control measures should be sought to replace vaccination so as to achieve the best outcome in balancing the gains and costs.

The findings in this work contribute to our knowledge base on the value of vaccination in epidemic cholera settings and could provide useful guidelines for public health administration to better control ongoing cholera outbreaks. There are, however, several limitations in this study. We have assumed that the cholera vaccines are of

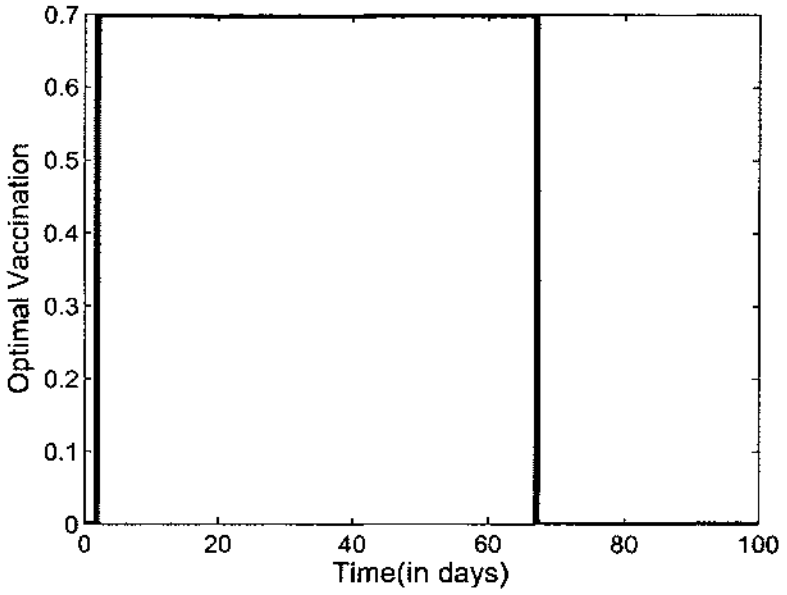

FIGURE 29: Case(b): Profile of the optimal control, $\phi^*(t)$ when optimal vaccination starts from t=2.

sufficient quantity for mass administration. In practical cholera epidemic and emergency settings, the availability of such vaccines could be limited and, consequently, the vaccination strategy would likely change. Meanwhile, we did not distinguish the human population with ages in our model, whereas in reality different age groups might have different transmission dynamics of the disease and exhibit different degrees of protection from vaccination. A model incorporating age structure will be constructed and analyzed in Section 3.4 of this dissertation. In addition, though our study here has focused on vaccination as the control measure against cholera outbreaks, as already demonstrated in previous sections, a combination with water sanitation, hygiene, rehydration, and other appropriate medical treatments would most likely yield the best results in fighting cholera.

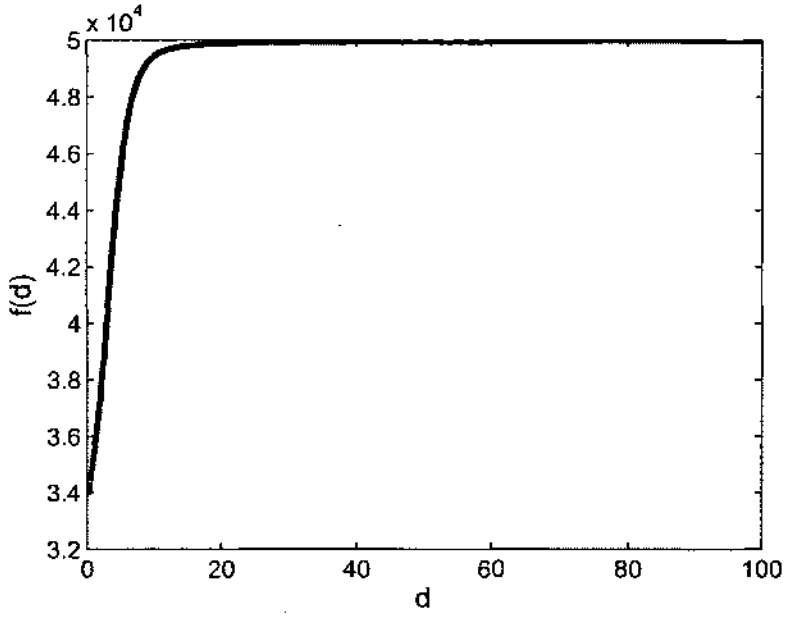

FIGURE 30: The curves of $f(d)$ vs. d when $c_{21} = 0.1$.

Parameter	Symbol	Value	Source
Total population	N	10,000	-
Natural human birth and death rate	μ	$(43.5\text{yr})^{-1}$	[13]
Concentration of Vibro *cholerae* in environment	κ	10^6 cells/ml	[16]
Rate of recovery from cholera	γ	$(5\,\text{day})^{-1}$	[10]
Rate of human contribution to Vibro *cholerae*	ξ	10 cells/ml-day	[10]
Death rate of vibrios in the environment	δ	$(30\,\text{day})^{-1}$	[10]
Ingestion rate from the environment	β_e	0.075/day	[13]
Ingestion rate through human-human interaction	β_h	0.00011/day	[13]
Maximum vaccination rate	$\phi_{\max}$	70%	[14]
Efficacy of cholera vaccines	ϵ	75%	[15]

TABLE. 2: Parameter values for the cholera model with vaccination.

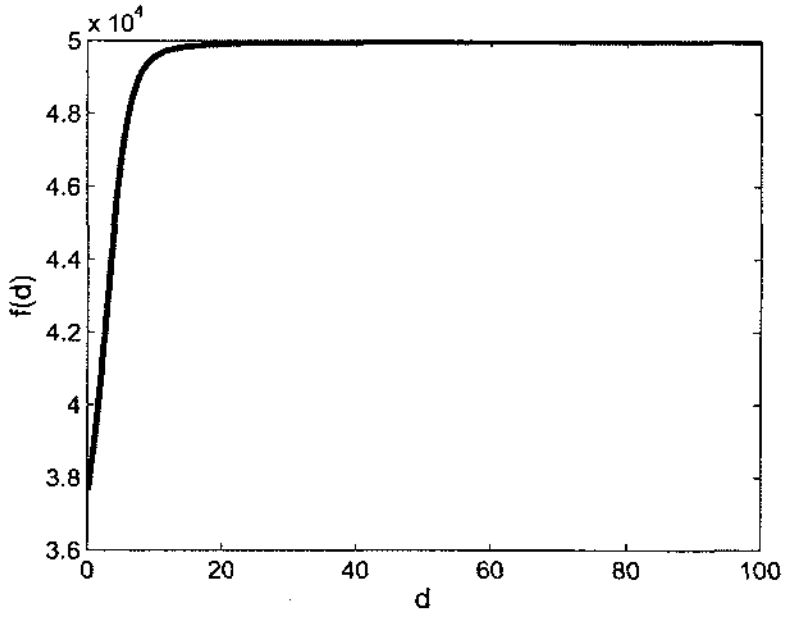

FIGURE 31: The curves of $f(d)$ vs. d when $c_{21} = 0.5$.

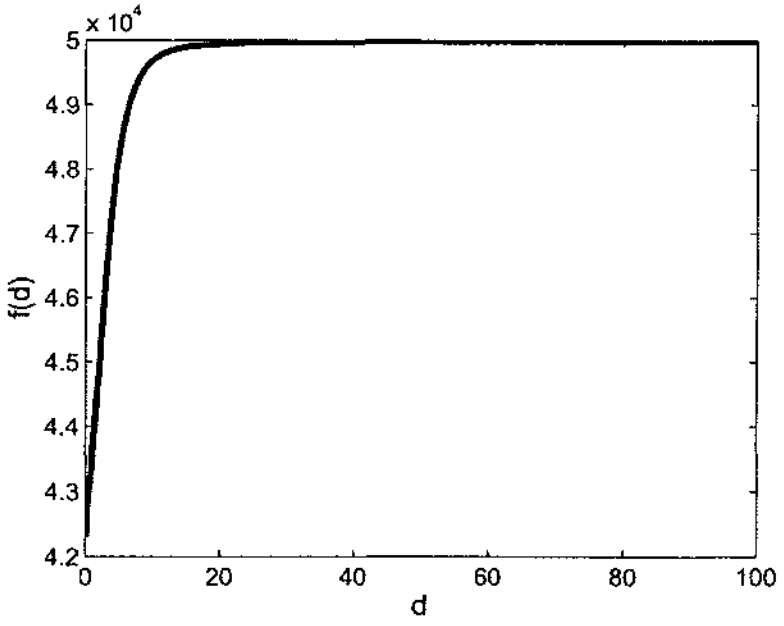

FIGURE 32: The curves of $f(d)$ vs. d when $c_{21} = 1.0$.

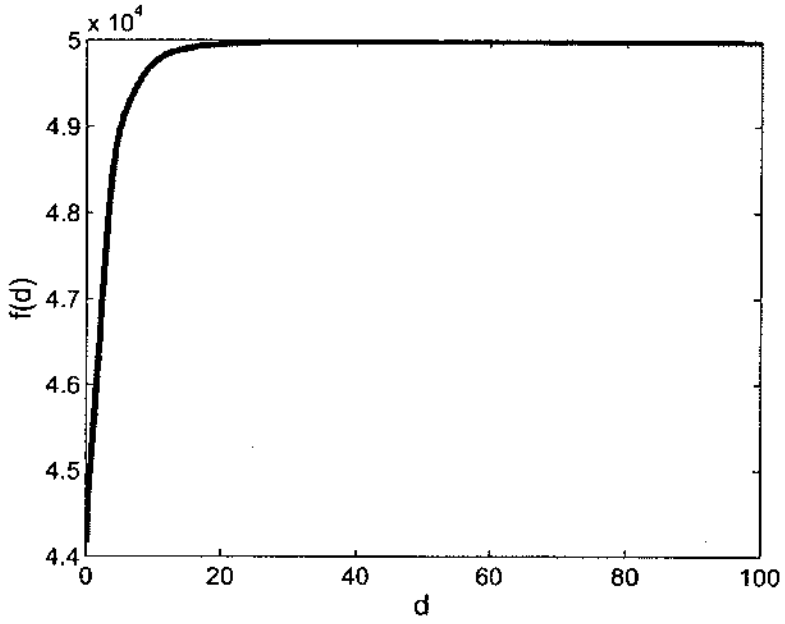

FIGURE 33: The curves of $f(d)$ vs. d when $c_{21} = 1.2$.

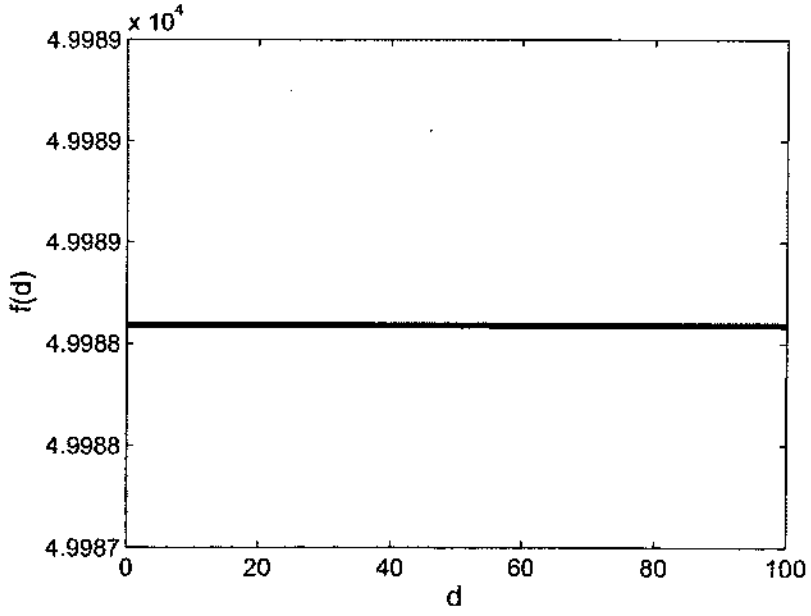

FIGURE 34: The curves of $f(d)$ vs. d when $c_{21} = 1.4$.

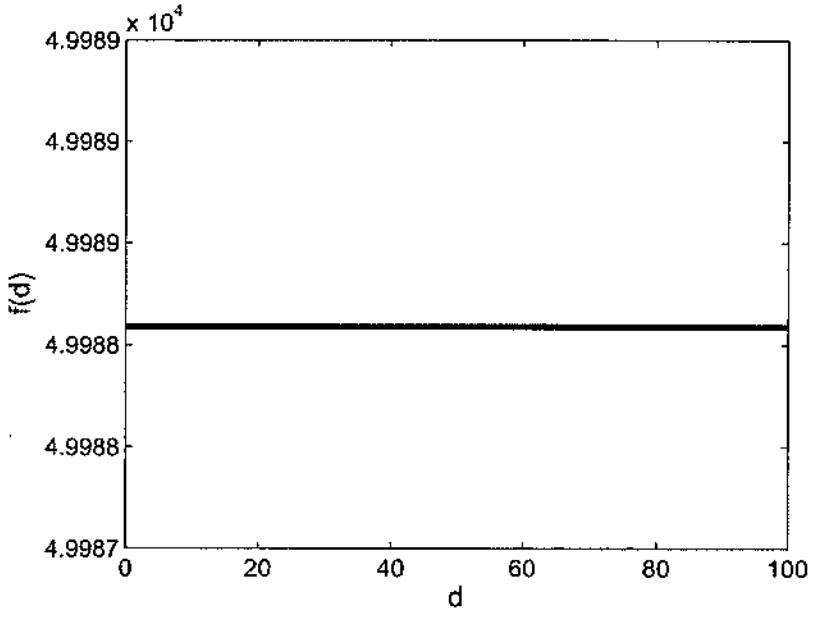

FIGURE 35: The curves of $f(d)$ vs. d when $c_{21} = 2.0$.

3.4 OPTIMAL CONTROL APPLIED TO CHOLERA MODEL WITH AGE STRUCTURE

In this section, we extend our study to an age-structure model to investigate the impact of different ages on cholera dynamics and the corresponding control strategy. Furthermore, inspired by the model in [20] (see Sec.2.2.7), we divide the human population into the symptomatic classes (denoted as S, I_S and R_S), the asymptomatic classes (denoted as $\hat{S}, I_A$, and R_A), and the vaccinated class (V). Now the system contains seven partial differential equations and two ordinary differential equations:

$$\begin{aligned}\frac{\partial S}{\partial t}+\frac{\partial S}{\partial a} &= -\Big[\beta_L(a)\frac{B_L(t)}{\kappa_L(a)+B_L(t)}+\beta_H(a)\frac{B_H(t)}{\kappa_H(a)+B_H(t)}\Big]S(a,t)\\ &\quad +b(a)\Big[S(a,t)+\hat{S}(a,t)+I_S(a,t)+I_A(a,t)+R_S(a,t)+R_A(a,t)\\ &\quad +V(a,t)\Big]-d(a)S(a,t)+\omega_3\hat{S}(a,t)+\omega_4V(a,t)-u(a,t)S(a,t),\end{aligned} \tag{49}$$

$$\begin{aligned}\frac{\partial I_S}{dt}+\frac{\partial I_S}{\partial a} &= p\Big[\beta_L(a)\frac{B_L(t)}{\kappa_L(a)+B_L(t)}+\beta_H(a)\frac{B_H(t)}{\kappa_H(a)+B_H(t)}\Big]S(a,t)\\ &\quad -\Big[d(a)+\gamma_2(a)+e_2(a)\Big]I_S(a,t),\end{aligned} \tag{50}$$

$$\frac{\partial R_S}{\partial t}+\frac{\partial R_S}{\partial a} = -\Big[d(a)+\omega_2(a)\Big]R_S(a,t)+\gamma_2(a)I_S(a,t), \tag{51}$$

$$\begin{aligned}\frac{\partial \hat{S}}{\partial t}+\frac{\partial \hat{S}}{\partial a} &= -\Big[\beta_L(a)\frac{B_L(t)}{\kappa_L(a)+B_L(t)}+\beta_H(a)\frac{B_H(t)}{\kappa_H(a)+B_H(t)}\Big]\hat{S}(a,t)\\ &\quad -\Big[d(a)+\omega_3(a)+u(a,t)\Big]\hat{S}(a,t)+\omega_1(a)R_A(a,t)+\omega_2(a)R_S(a,t),\end{aligned} \tag{52}$$

$$\begin{aligned}\frac{\partial I_A}{\partial t}+\frac{\partial I_A}{\partial a} &= \Big[\beta_L(a)\frac{B_L(t)}{\kappa_L(a)+B_L(t)}+\beta_H(a)\frac{B_H(t)}{\kappa_H(a)+B_H(t)}\Big]\hat{S}(a,t)\\ &\quad -\Big[d(a)+e_1(a)+\gamma_1(a)\Big]I_A(a,t)\\ &\quad +(1-p)\Big[\beta_L(a)\frac{B_L(t)}{\kappa_L(a)+B_L(t)}+\beta_H(a)\frac{B_H(t)}{\kappa_H(a)+B_H(t)}\Big]S(a,t),\end{aligned} \tag{53}$$

$$\frac{\partial R_A}{\partial t}+\frac{\partial R_A}{\partial a} = -\Big[d(a)+\omega_1(a)\Big]R_A(a,t)+\gamma_1(a)I_A(a,t), \tag{54}$$

$$\frac{\partial V}{\partial t}+\frac{\partial V}{\partial a} = u(a,t)\Big[S(a,t)+\hat{S}(a,t)\Big]-\Big[\omega_4(a)+d(a)\Big]V(a,t), \tag{55}$$

$$\frac{dB_H}{dt} = \int_0^A \eta_1(a)I_A(a,t)da+\int_0^A \eta_2(a)I_S(a,t)da-\chi(t)B_H(t), \tag{56}$$

$$\frac{dB_L}{dt} = \chi(t)B_H(t)-\delta(t)B_L(t), \tag{57}$$

The natural domain for this system is

$$\{(t,a) \quad | \quad 0 \leq t \leq T, \quad 0 \leq a \leq A\}$$

where $T > 0$ is the final time and $A > 0$ is the maximum age under consideration.

The initial conditions are

$$\begin{aligned} & S(a,0) = S_0(a), \quad I_S(a,0) = I_{S_0}(a), \quad R_S(a,0) = R_{S_0}(a), \\ & \hat{S}(a,0) = \hat{S}_0(a), \quad I_A(a,0) = \hat{I}_{A_0}(a), \quad R_A(a,0) = R_{A_0}(a), \\ & V(a,0) = V_0(a), \quad B_H(0) = B_{H_0}, \quad B_L(0) = B_{L_0}, \end{aligned} \tag{58}$$

and the boundary conditions are

$$\begin{aligned} & S(0,t) = 0, \\ & R_S(0,t) = \int_0^A \Big[I_S(a,t) + I_A(a,t) + R_S(a,t) + R_A(a,t)\Big] f(a)da, \\ & R_A(0,t) = 0, \quad I_A(0,t) = 0 = I_S(0,t) = \hat{S}(0,t) = V(0,t). \end{aligned} \tag{59}$$

We aim to minimize the following objective functional

$$\begin{aligned} J(u) = \int_0^T \int_0^A \Big[& A_1(a) I_S(a,t) + A_2(a) u(a,t) \Big(S(a,t) + \hat{S}(a,t) \\ & + I_A(a,t) + R_A(a,t) \Big) + \frac{1}{2} A_3(a) u^2(a,t) \Big] dadt, \end{aligned} \tag{60}$$

where A_1, A_2 and A_3 are appropriate cost parameters, generally depending on the age a.

The adjoint equations are as follows:

$$-((\lambda_1)_t + (\lambda_1)_a) = -\zeta\lambda_1 + b\lambda_1 - (d+\mu)\lambda_1 + p\zeta\lambda_2$$
$$+ (1-p)\zeta\lambda_5 + \mu\lambda_7 + A_2(a)u, \tag{61}$$

$$-((\lambda_2)_t + (\lambda_2)_a) = -(d+\gamma_2+e_2)\lambda_2 + b\lambda_1 + \gamma_2\lambda_3$$
$$+ \eta_2\lambda_3 + \lambda_3(0,t)f(a) + A_1(a), \tag{62}$$

$$-((\lambda_3)_t + (\lambda_3)_a) = -(d+\omega_3)\lambda_3 + b\lambda_1 + \omega_2\lambda_4 + \lambda_3(0,t)f(a), \tag{63}$$

$$-((\lambda_4)_t + (\lambda_4)_a) = -\zeta\lambda_4 - (d+\omega_3)\lambda_4 - \lambda_4 u(a,t) + b\lambda_1$$
$$+ \lambda_1\omega_3 + \zeta\lambda_5 + u(a,t)\lambda_7 + A_2(a)u(a,t), \tag{64}$$

$$-((\lambda_5)_t + (\lambda_5)_a) = -(d+e_1+\gamma_1)\lambda_5 + b\lambda_1 + \omega_1\lambda_6 + \eta_1\lambda_8$$
$$+ \lambda_3(0,t)f(a) + A_2(a)u(a,t), \tag{65}$$

$$-((\lambda_6)_t + (\lambda_6)_a) = -(d+\omega_1)\lambda_6 + b\lambda_1 + \omega_1\lambda_4$$
$$+ \lambda_3(0,t)f(a) + A_2(a)u(a,t), \tag{66}$$

$$-((\lambda_7)_t + (\lambda_7)_a) = -(d+w_1)\lambda_7 + b\lambda_1 + \omega_4\lambda_1, \tag{67}$$

$$-\frac{d\lambda_8}{dt} = \chi(\lambda_9 - \lambda_8) + H(t)\int_0^A \Big[(-\lambda_1 + p\lambda_2 + (1-p)\lambda_5)S(a,t)$$
$$+ (\lambda_5 - \lambda_4)\hat{S}(a,t)\Big]da, \tag{68}$$

$$-\frac{d\lambda_9}{dt} = -\delta\lambda_9 + K(t)\int_0^A \Big[(-\lambda_1 + p\lambda_2 + (1-p)\lambda_5)S(a,t)$$
$$+ (\lambda_5 - \lambda_4)\hat{S}(a,t)\Big]da, \tag{69}$$

with

$$\lambda_i(a,T) = 0 \quad \text{for} \quad a \in (0,A), \quad 1 \le i \le 7, \tag{70}$$

$$\lambda_i(A,t) = 0 \quad \text{for} \quad a \in (0,T), \quad 1 \le i \le 7, \tag{71}$$

$$\lambda_8(T) = \lambda_9(T) = 0. \tag{72}$$

Note that we have used the notations

$$\zeta(a,t) = \beta_L(a)\frac{B_L}{\kappa_L + \beta_L} + \beta_H(a)\frac{B_H}{\kappa_H + B_H},$$
$$H(t) = \beta_H\frac{K_H}{(K_H + B_H)^2} \quad \text{and} \quad K(t) = \beta_L\frac{K_L}{(K_L + B_L(t))^2}.$$

The characterization of the optimal control is

$$u^*(a,t) = \Im\left(\frac{\lambda_1 S + \lambda_4\hat{S} - \lambda_7(S+\hat{S}) - A_2(S+\hat{S}+I_A+R_A)}{A_3}\right), \tag{73}$$

where

$$\Im(x) = \begin{cases} 0 & \text{if} \quad x < 0 \\ x & \text{if} \quad 0 \leq x \leq \text{Max} \\ \text{Max} & \text{if} \quad x > \text{Max}. \end{cases}$$

We use Max = 1 in this study and

$$f(a) = \begin{cases} \frac{1}{5}\sin^2\left[\left(\frac{a-15}{30}\right)\pi\right] & \text{for} \quad 15 < a < 45 \\ 0 & \text{otherwise.} \end{cases}$$

The parameters are presented in Table 3. We have conducted some preliminary numerical simulation for this optimal control age-structure model, and the results are presented in Figures (36)-(43).

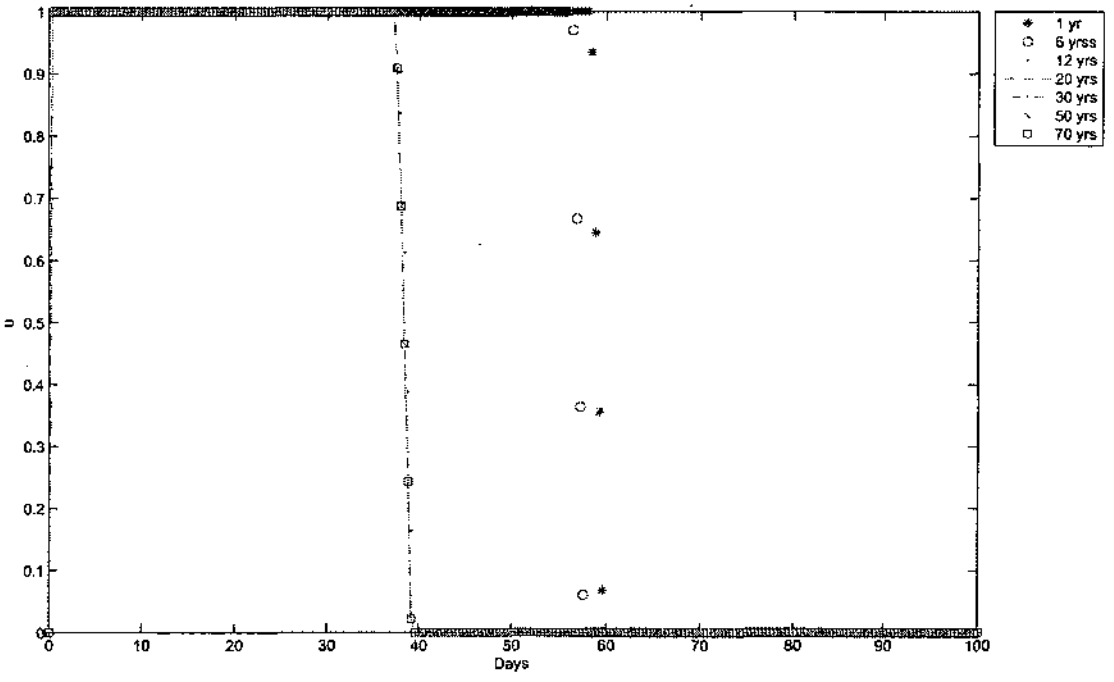

FIGURE 36: Profile of the optimal control, u.

Parameter	Value
Ingestion rate of non-hyperinfective cholera bacteria, β_L	0.02 day^{-1}
Ingestion rate of hyperinfective cholera bacteria, β_H	0.002 day^{-1}
Saturation constant of non-hyperinfective cholera bacteria, κ_L	10^3 cells/ml
Saturation constant of hyperinfective cholera bacteria, κ_H	$\frac{10}{7}$ cells/ml
Natural mortality rate of humans, d	$\frac{0.01619}{365}$ day^{-1}
Natural birth rate of humans, b	$\frac{0.03149}{365}$ day^{-1}
Proportion of infections being symptomatic, p	0.1
Waning immunity rate of asymptomatic recovered humans, ω_1	0.01 day^{-1}
Waning immunity rate of symptomatic recovered humans, ω_2	0.0022 day^{-1}
Rate of transfer from asymptomatic susceptibles to symptomatic susceptibles, ω_3	$\frac{1}{(10\times365)}$ day^{-1}
Waning immunity rate of vaccinated people, ω_4	$\frac{1}{(10\times365)}$ day^{-1}
Recovery rate of asymptomatic infections, γ_1	0.5 day^{-1}
Recovery rate of symptomatic infections, γ_2	0.2 day^{-1}
Cholera induced death rate of asymptomatic infections, e_1	0.000205 day^{-1}
Cholera induced death rate of symptomatic infections, e_2	0.0041 day^{-1}
Shedding rate of asymptomatic infected individuals, η_1	0.008 $\frac{cells}{ml-day-human}$
Shedding rate of symptomatic infected individuals, η_1	0.8 $\frac{cells}{ml-day-human}$
Rate of vibrio moving from HI to non-Hi state, χ	5 day^{-1}
Death rate of vibrio in the environment, δ	$\frac{1}{30}$ day^{-1}
Initial symptomatic susceptible population, S_0	9000
Initial asymptomatic susceptible population, $\hat{S}_0$	1000
Initial asymptomatic infections, $I_A(0)$	0
Initial symptomatic infections, $I_S(0)$	0
Initial recovered population from asymptomatic state, $R_A(0)$	0
Initial recovered population from symptomatic state, $R_S(0)$	0
Initial vaccinated population, V_0	0
Initial non-hyperinfective state, B_{L_0}	10
Initial hyperinfective state, B_{H_0}	0
Final time, T	100 days
Maximum age, A	72 years

TABLE. 3: Parameter values for the age structure model [26].

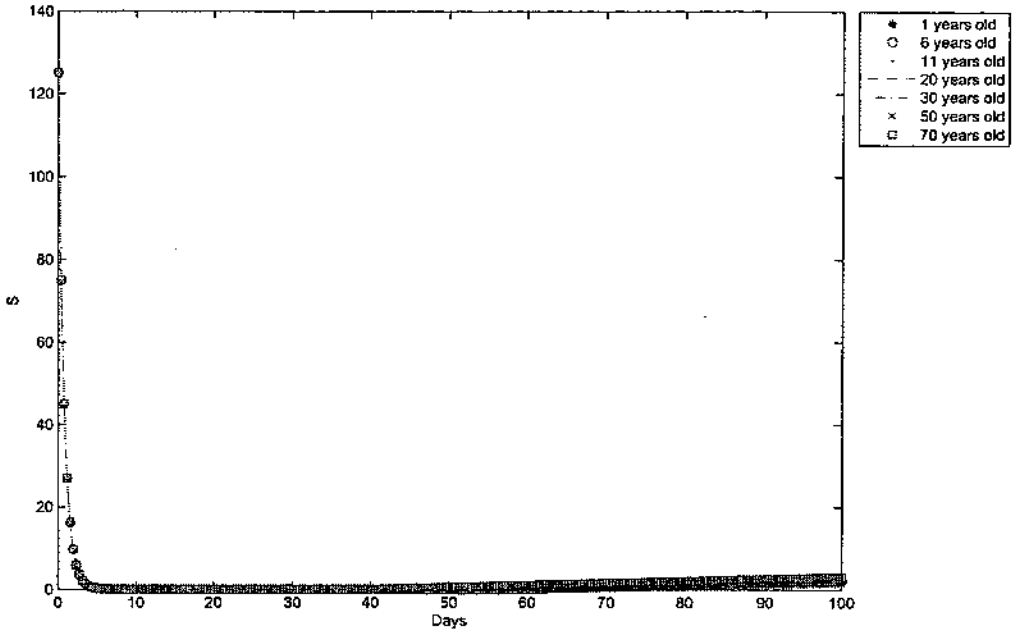

FIGURE 37: Profile of the susceptible state, S.

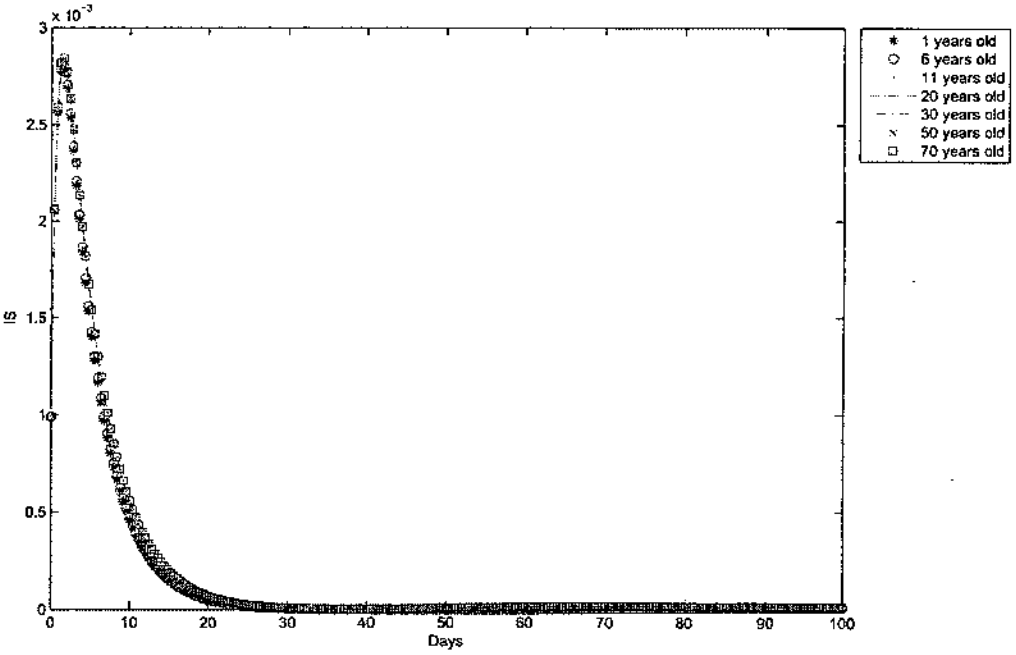

FIGURE 38: Profile of I_S.

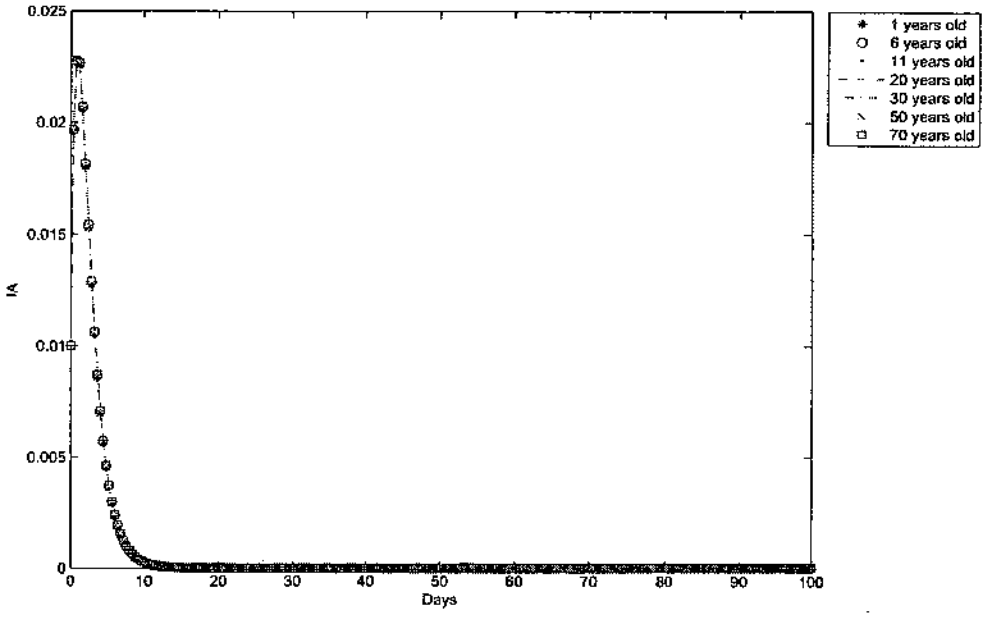

FIGURE 39: Profile of I_A.

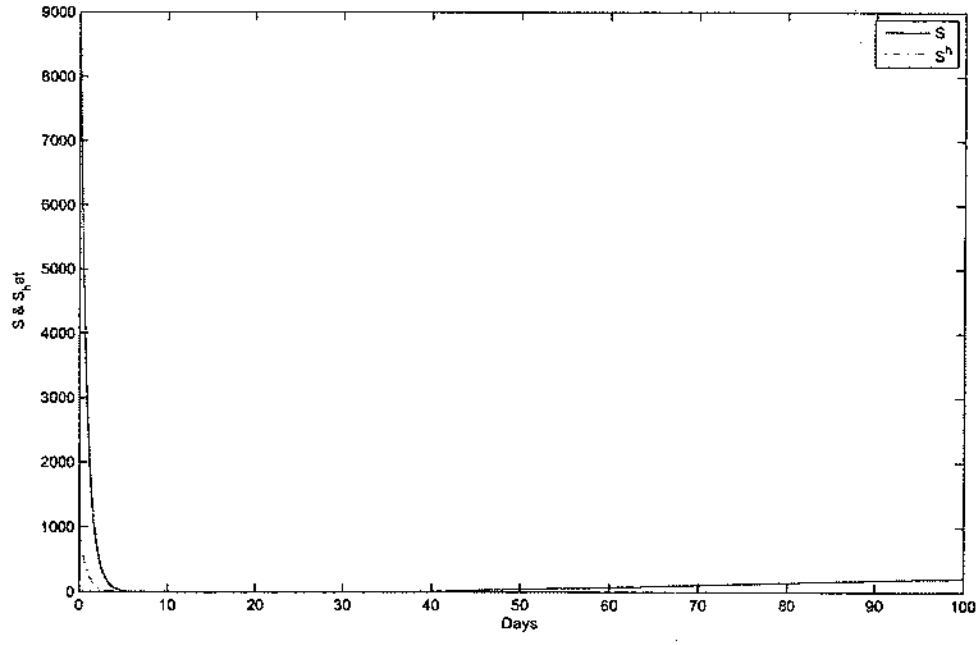

FIGURE 40: Profile of S and $\hat{S}$.

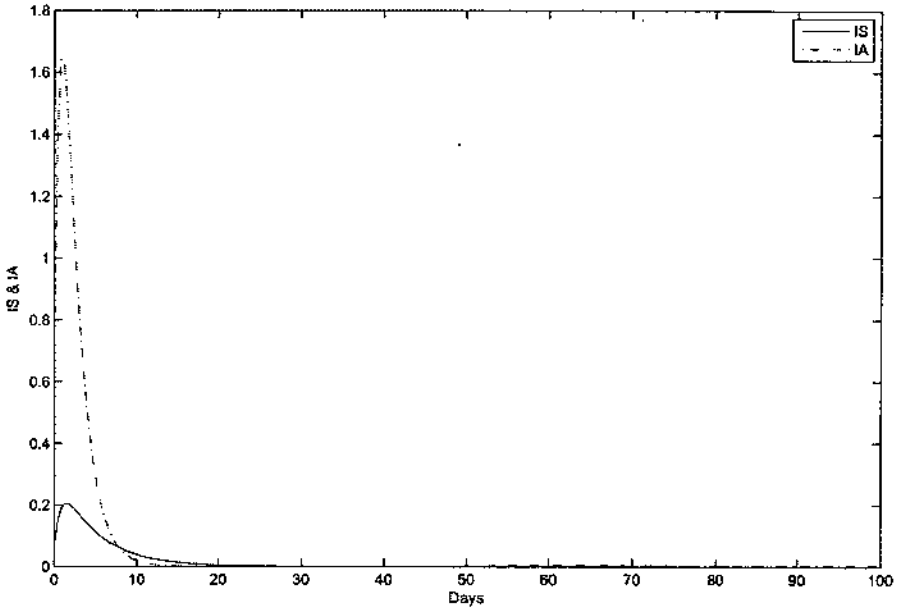

FIGURE 41: Profile of I_S and I_A.

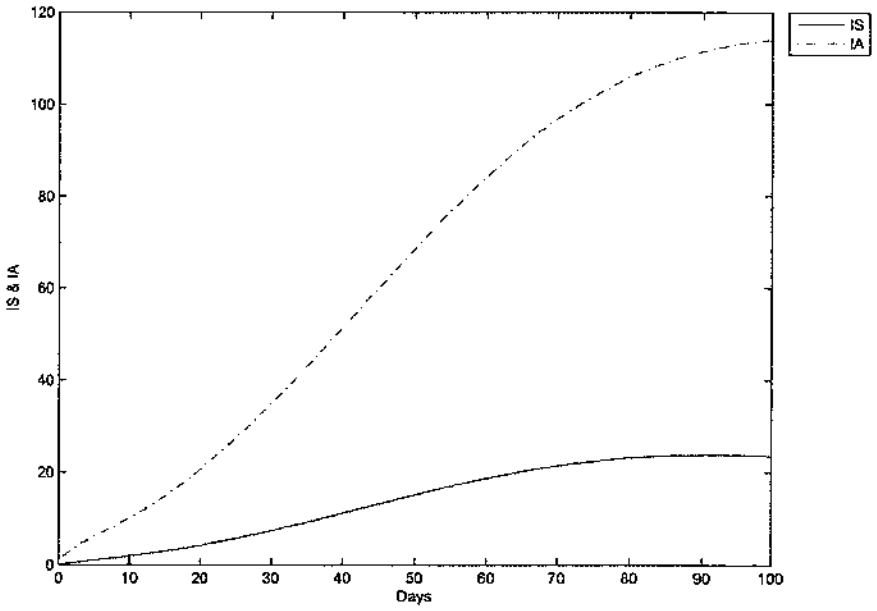

FIGURE 42: Profile of I_S and I_A.

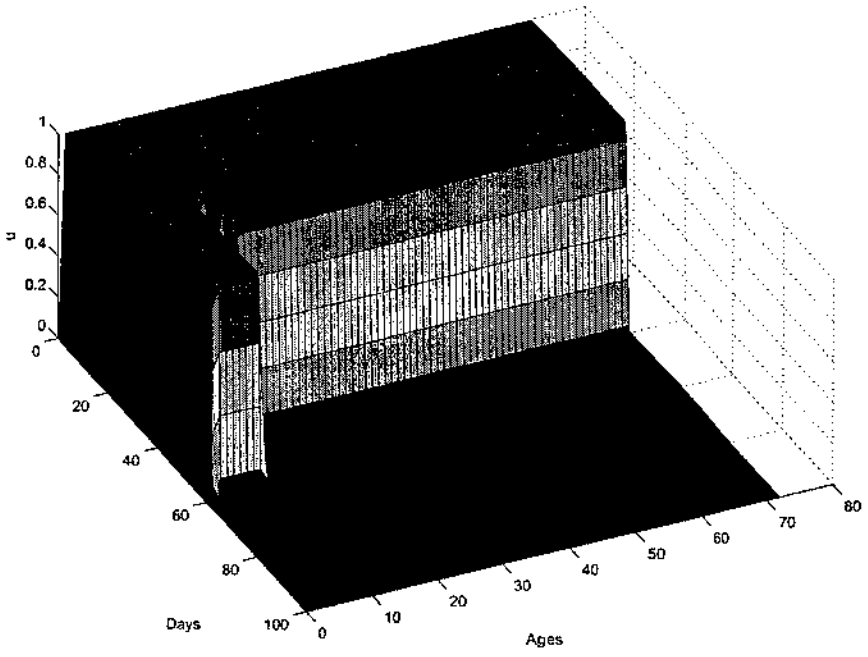

FIGURE 43: Three dimensional plot of u with time and ages.

3.5 OPTIMAL CONTROL FOR MULTIGROUP CHOLERA MODELING

So far in our discussion we have focused on the time evolution of cholera dynamics and have not considered the spatial spread of the infection. It is well known that the transmission and spread of infectious diseases are complicated by spatial heterogeneity that involves distinctions in ecological and geographical environments, demographic structures, human activity levels, contact and mixing patterns, and many other factors. Consequently, such spatial heterogeneity will likely lead to distinct control strategies in different regions.

One of the most successful approaches to investigate spatial heterogeneity in mathematical epidemiology is the multigroup modeling, where the entire population is divided into $n \quad (n \geq 2)$ distinct groups, and disease transmission occurs both within the same group and between different groups (reflecting the movement of human hosts and/or pathogen from one region to another). In this section, we will take a first step to conduct an optimal control study for a two-group cholera model. For simplicity, we will consider bilinear incidence for both the direct and indirect transmission pathways (see Sec. 2.2.6 for such work in Tien and Earn's model), and we will use vaccination as the only control measure.

Our two-group cholera model thus takes the form below:

$$
\begin{aligned}
\frac{dS_1}{dt} &= \mu_1 N_1 - (\lambda_{11} S_1 B_1 + \lambda_{12} S_1 B_2) - (\beta_{11} S_1 I_1 + \beta_{12} S_1 I_2) - \mu_1 S_1 - u_1(t) S_1, \\
\frac{dI_1}{dt} &= (\lambda_{11} S_1 B_1 + \lambda_{12} S_1 B_2) + (\beta_{11} S_1 I_1 + \beta_{12} S_1 I_2) - (\mu_1 + \gamma_1) I_1, \\
\frac{dB_1}{dt} &= \xi_1 I_1 - \delta_1 B_1, \\
\frac{dS_2}{dt} &= \mu_2 N_2 - (\lambda_{21} S_2 B_1 + \lambda_{22} S_2 B_2) - (\beta_{21} S_2 I_1 + \beta_{22} S_2 I_2) - \mu_2 S_2 - u_2(t) S_2, \\
\frac{dI_2}{dt} &= (\lambda_{21} S_2 B_1 + \lambda_{22} S_2 B_2) + (\beta_{21} S_2 I_1 + \beta_{22} S_2 I_2) - (\mu_2 + \gamma_2) I_2, \\
\frac{dB_2}{dt} &= \xi_2 I_2 - \delta_2 B_2.
\end{aligned} \tag{74}
$$

The parameters $\lambda_{ij} \quad (i, j = 1, 2)$ represents the cross transmission rate from vibrios in group j to susceptibles in group i, and $\beta_{ij} (i, j = 1, 2)$ represents the cross transmission rate from infectives in group j to susceptibles in group i. Other parameters have similar meanings as before, with subscript $i (i = 1, 2)$ referring to the ith group.

In our optimal control study, we aim to minimize the total number of infections

and the (linear) costs of vaccination for both groups over the time domain $[0, T]$; i.e.,

$$\min_{(u_1(t),u_2(t))} \int_0^T (I_1(t) + cu_1(t)S_1(t) + I_2(t) + cu_2(t)S_2(t))dt. \tag{75}$$

The Hamiltonian H in this case is given by

$$\begin{aligned}
H &= I_1(t) + cu_1(t)S_1(t) + I_2(t) + cu_2(t)S_2(t) + \lambda_{S_1}(\frac{dS_1}{dt}) + \lambda_{S_2}(\frac{dS_2}{dt}) \\
&+ \lambda_{I_1}(\frac{dI_1}{dt}) + \lambda_{I_2}(\frac{dI_2}{dt}) + \lambda_{B_1}(\frac{dB_1}{dt}) + \lambda_{B_2}(\frac{dB_2}{dt}), \\
&= I_1(t) + cu_1(t)S_1(t) + I_2(t) + cu_2(t)S_2(t) \\
&+ \lambda_{S_1}\Big[\mu_1 N_1 - (\lambda_{11}S_1B_1 + \lambda_{12}S_1B_2) - (\beta_{11}S_1I_1 + \beta_{12}S_1I_2) - \mu_1 S_1 - u_1(t)S_1\Big], \\
&+ \lambda_{S_2}\Big[\mu_2 N_2 - (\lambda_{21}S_2B_1 + \lambda_{22}S_2B_2) - (\beta_{21}S_2I_1 + \beta_{22}S_2I_2) - \mu_2 S_2 - u_2(t)S_2\Big], \\
&+ \lambda_{I_1}\Big[(\lambda_{11}S_1B_1 + \lambda_{12}S_1B_2) + (\beta_{11}S_1I_1 + \beta_{12}S_1I_2) - (\mu_1 + \gamma_1)I_1\Big], \\
&+ \lambda_{I_2}\Big[(\lambda_{21}S_2B_1 + \lambda_{22}S_2B_2) + (\beta_{21}S_2I_1 + \beta_{22}S_2I_2) - (\mu_2 + \gamma_2)I_2\Big], \\
&+ \lambda_{B_1}\Big[\xi_1 I_1 - \delta_1 B_1\Big] + \lambda_{B_2}\Big[\xi_2 I_2 - \delta_2 B_2\Big].
\end{aligned}$$

Here the adjoint functions must satisfy

$$\begin{aligned}
\frac{d\lambda_{S_1}}{dt} &= -\frac{\partial H}{\partial S_1} \\
&= -cu_1 + \lambda_{S_1}(\lambda_{11}B_1 + \lambda_{12}B_2 + \beta_{11}I_1 + \beta_{12}I_2 + \mu_1 + u_1) \\
&- \lambda_{I_1}(\lambda_{11}B_1 + \lambda_{12}B_2 + \beta_{11}I_1 + \beta_{12}I_2), \qquad (76) \\
\frac{d\lambda_{S_2}}{dt} &= -\frac{\partial H}{\partial S_2} \\
&= -cu_2 + \lambda_{S_2}(\lambda_{21}B_1 + \lambda_{22}B_2 + \beta_{21}I_1 + \beta_{22}I_2 + \mu_2 + u_2) \\
&- \lambda_{I_2}(\lambda_{21}B_1 + \lambda_{22}B_2 + \beta_{21}I_1 + \beta_{22}I_2), \qquad (77) \\
\frac{d\lambda_{I_1}}{dt} &= -\frac{\partial H}{\partial I_1} \\
&= -1 + \lambda_{S_1}\beta_{11}S_1 + \lambda_{S_2}\beta_{21}S_2 - \lambda_{I_1}(\beta_{11}S_1 - \mu_1 - \gamma_1) \\
&- \lambda_{I_2}(\beta_{21}S_2 - \lambda_{B_1}\xi_1), \qquad (78)
\end{aligned}$$

$$\frac{d\lambda_{I_2}}{dt} = -\frac{\partial H}{\partial I_2}$$
$$= -1 + \lambda_{S_1}\beta_{12}S_1 + \lambda_{S_2}\beta_{22}S_2$$
$$- \lambda_{I_1}(\beta_{12}S_1) - \lambda_{I_2}(\beta_{22}S_2 - \mu_2 - \gamma_2) - \lambda_{B_1}\xi_2, \quad (79)$$
$$\frac{d\lambda_{B_1}}{dt} = -\frac{\partial H}{\partial B_1}$$
$$= \lambda_{S_1}\lambda_{11}S_1 + \lambda_{S_2}\lambda_{21}S_2 - \lambda_{I_1}\lambda_{11}S_1 - \lambda_{I_2}\lambda_{21}S_2 + \lambda_{B_1}\delta_1, \quad (80)$$
$$\frac{d\lambda_{B_2}}{dt} = -\frac{\partial H}{\partial B_2}$$
$$= \lambda_{S_1}\lambda_{12}S_1 + \lambda_{S_2}\lambda_{22}S_2 - \lambda_{I_1}\lambda_{12}S_1 - \lambda_{I_2}\lambda_{22}S_2 + \lambda_{B_2}\delta_2, \quad (81)$$

with transversality conditions:

$$\lambda_{S_1}(T) = 0, \lambda_{S_2}(T) = 0, \lambda_{I_1}(T) = 0, \lambda_{I_2}(T) = 0, \lambda_{B_1}(T) = 0, \lambda_{B_2}(T) = 0. \quad (82)$$

Meanwhile, we note that

$$\frac{\partial H}{\partial u_1} = c_1 S_1 - \lambda_{S_1} S_1 \qquad \text{and} \qquad \frac{\partial H}{\partial u_2} = c_2 S_2 - \lambda_{S_2} S_2. \quad (83)$$

Therefore, again, the characterization of the optimal controls are based on the switching conditions [1, 23]:

$$u_1^* = u_{1_{\max}} \text{ if } \frac{\partial H}{\partial u_1} < 0\,; \quad u_1^* = 0 \text{ if } \frac{\partial H}{\partial u_1} > 0\,, \quad (84)$$

and

$$u_2^* = u_{2_{\max}} \text{ if } \frac{\partial H}{\partial u_2} < 0\,; \quad u_2^* = 0 \text{ if } \frac{\partial H}{\partial u_2} > 0\,. \quad (85)$$

For our numerical simulation we list all the parameters in Table 4 and Table 5. We have performed some preliminary runs in our optimal control simulation to this two-group model. In particular, we have chosen cross transmission rates in such a way that $\beta_{12} > \beta_{21}$ and $\lambda_{12} > \lambda_{21}$, while keeping other transmission parameters the same between the two groups: $\xi_1 = \xi_2 = \xi$, $\lambda_{11} = \lambda_{22} = \lambda$, $\mu_1 = \mu_2 = \mu$, $\gamma_1 = \gamma_2 = \gamma$, $\delta_1 = \delta_2 = \delta$, $\beta_{11} = \beta_{22} = \beta$. This simple setting allows us to investigate (and focus on) the impact of distinct cross transmission rates on the optimal control strategy for each group. The results for this scenario are presented in Figures (44)-(47). As can be naturally expected, the higher disease transmission from group 2 to group 1 (than that in converse route) results in higher levels of infections and pathogen concentration in group 1, which necessitates longer duration of vaccination in group 1.

Parameter	Symbol	Value	Source
Total population	N	10,000	-
Natural human birth and death rate of group 1	μ_1	$(43.5\text{yr})^{-1}$	[13]
Natural human birth and death rate of group 2	μ_2	$(43.5\text{yr})^{-1}$	[13]
Rate of recovery from cholera of group 1	γ_1	$(5\,\text{day})^{-1}$	[10]
Rate of recovery from cholera of group 2	γ_2	$(5\,\text{day})^{-1}$	[10]
Rate of human contribution to Vibro *cholerae* of group 1	ξ_1	10 cells/ml-day	[10]
Rate of human contribution to Vibro *cholerae* of group 2	ξ_2	10 cells/ml-day	[10]
Death rate of vibrios in the environment of group 1	δ_1	$(30\,\text{day})^{-1}$	[10]
Death rate of vibrios in the environment of group 2	δ_2	$(30\,\text{day})^{-1}$	[10]
Ingestion rate from the environment to humans in group 1	λ_{11}	0.0001/day	-
Ingestion rate from the environment to humans in group 2	λ_{22}	0.0001/day	-
Ingestion rate through human-human interaction of group 1	β_{11}	0.00011/day	[13]
Ingestion rate through human-human interaction of group 2	β_{22}	0.00011/day	[13]
Maximum vaccination rate of group 1	u_1	0.7	-
Maximum vaccination rate of group 2	u_2	0.7	-

TABLE. 4: Parameter values for the two-group cholera model.

Parameter	Value
Cross transmission rate from vibrios in group 2 to susceptibles in group 1, β_{12}	0.0005
Cross transmission rate from vibrios in group 1 to susceptibles in group 2, β_{21}	0.0001
Cross transmission rate from infectives in group 2 to susceptibles in group 1, λ_{12}	0.00008
Cross transmission rate from infectives in group 1 to susceptibles in group 2, λ_{21}	0.00001
The initial population, $N_1(0)$	5000
The initial population, $N_2(0)$	5000
The initial infectious population, $I_1(0)$	100
The initial infectious population, $I_2(0)$	10
The initial susceptible human populations, $S_1(0)$	4900
The initial susceptible human populations, $S_2(0)$	4990
The initial population of the concentration of the vibrios, $B_1(0)$	0
The initial population of the concentration of the vibrios, $B_2(0)$	0
The cost parameter, c	3
Time	100 days

TABLE. 5: The additional parameter values for the numerical implementation.

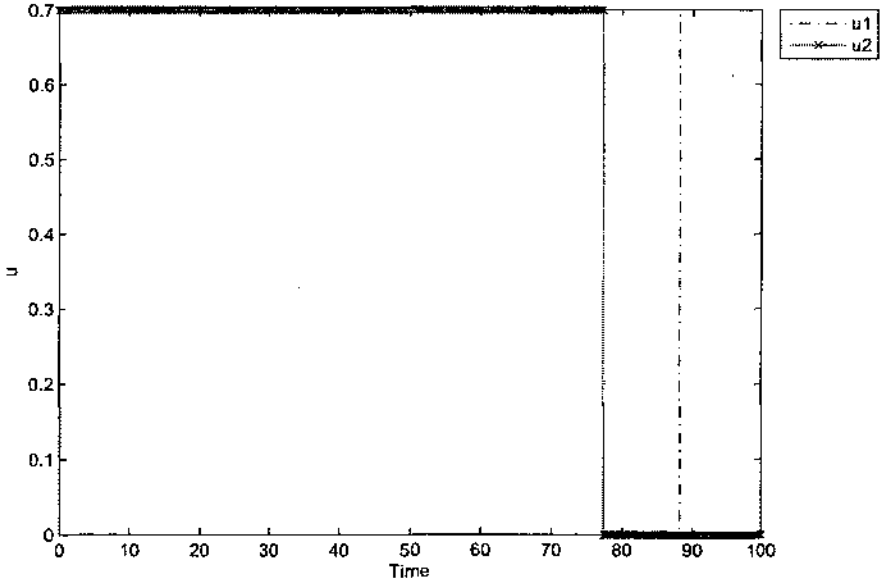

FIGURE 44: Profile of the optimal controls u_1 and u_2 from two groups.

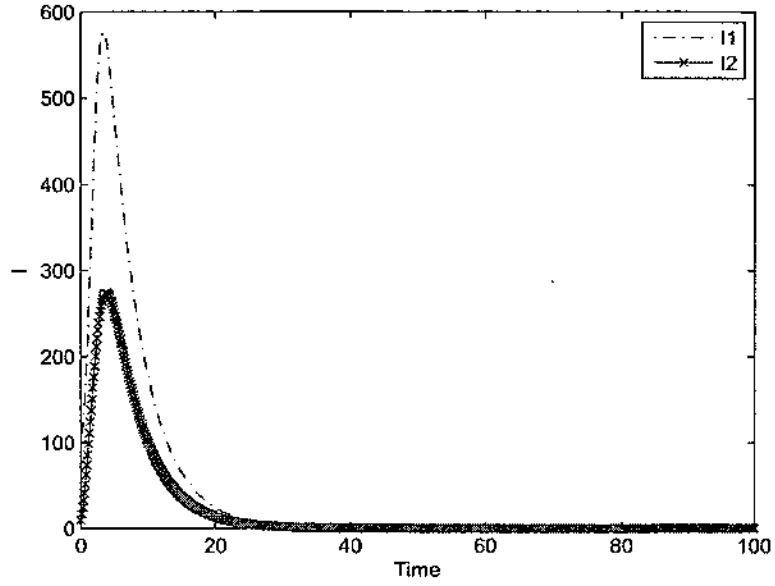

FIGURE 45: Profile of the infected populations I_1 and I_2 from two groups.

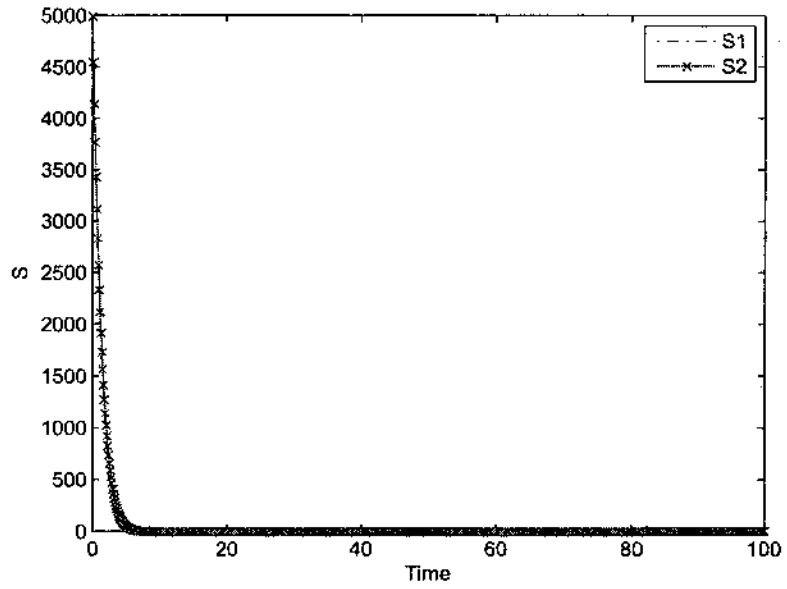

FIGURE 46: Profile of the susceptible populations S_1 and S_2 from two groups.

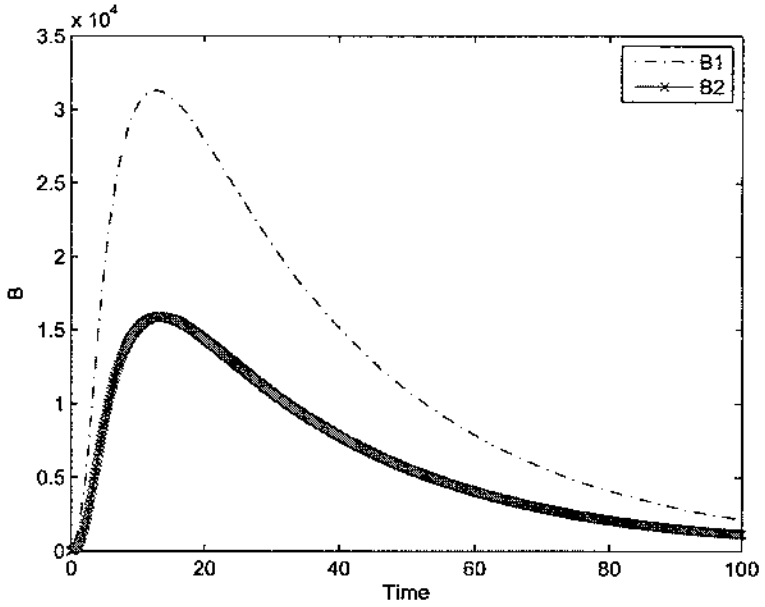

FIGURE 47: Profile of the concentration of the vibrios in the environment B_1 and B_2 from two groups.

CHAPTER 4

ITERATIVE ALGORITHM

As we have seen from our previous discussion, for most optimal control problems, numerical simulation is the only feasible means to gain the (approximate) solution. The numerical method we have been using is the Forward-Backward Sweeping Method (FBSM), which is essentially an iterative procedure. The advantages of the FBSM are that it is straightforward to implement, easy to code up, and it can be applied to a wide range of problems. The disadvantage, however, is that the FBSM often encounters problem in convergence: generally the method does not have fast convergence, and in some situations it does not converge at all. Using different ways of control update, e.g., replacing the control u by a convex combination between the previous control values and the current value, could help speed the convergence in some cases but such improvement is limited by the type of the problems, and often at the cost of losing accuracy [1].

In this chapter, we will explore several possible ways to improve the convergence of the original FBSM, by introducing the idea of locally refined iterative procedures. We will conduct rigorous error analysis for both the FBSM and our proposed methods for comparison. Next, we attempt to extend the concept of local iterative algorithms to more general problems of the kind. To that end, we will illustrate the application of local iterative algorithms to a class of constrained dynamical problems involving second-order differential-algebraic equations.

4.1 LOCALLY REFINED FBSM FOR OPTIMAL CONTROL PROBLEMS

4.1.1 ERROR ANALYSIS FOR FBSM

Despite its popularity, there is no rigorous error analysis on the FBSM that has been published so far. In what follows, we will first conduct a careful analysis on the convergence and accuracy for the FBSM, based on which we will explore improvements.

Let's consider the governing equations

$$\dot{x}(t) = g(t, x(t), u(t)), \tag{86}$$

$$\dot{\lambda}(t) = -[f_x(t, x, u) + \lambda(t) g_x(t, x, u)], \tag{87}$$

$$u(t) = h(t, x, u), \tag{88}$$

with $0 \leq t \leq T$. Meanwhile, we have an initial condition for x and a final time condition for λ :

$$x(0) = a,$$
$$\lambda(T) = b.$$

Equation (88) is the optimality condition, and it is assumed that $u(t)$ can be represented in terms of t, x and λ. The numerical formulation of this system is

$$\dot{x}^{k+1} = g(t, x^{k+1}, u^k), \quad 0 \leq t \leq T, \tag{89}$$

$$\dot{\lambda}^{k+1} = -\Big[f_x(t, x^{k+1}, u^k) + \lambda^{k+1} g_x(t, x^{k+1}, u^k)\Big], \quad 0 \leq t \leq T, \tag{90}$$

$$u^{k+1} = h(t, x^{k+1}, \lambda^{k+1}), \quad 0 \leq t \leq T, \tag{91}$$

for the $(k+1)^{th}$ iteration (k = 0,1,2, $\cdots$). In particular, $x^{k+1}(0) = a$ and $\lambda^{k+1}(T) = b$. To start, i.e., for the 0^{th} iteration, an initial guess $u^0 = 0 \quad (0 \leq t \leq T)$ is usually made. Let x, λ, and u denote the exact solutions of equations (86 - 88). Let also x^{k+1}, λ^{k+1}, and u^{k+1} denote the exact solutions of equations (89 - 91) at the $(k+1)^{th}$ iteration, $k = 0, 1, 2, \cdots$.

Next, we introduce the error functions

$$e_x^k(t) = x^k(t) - x(t), \quad 0 \leq t \leq t, \tag{92}$$

$$e_\lambda^k(t) = \lambda^k(t) - \lambda(t), \quad 0 \leq t \leq T, \tag{93}$$

$$e_u^k(t) = u^k(t) - u(t), \quad 0 \leq t \leq T, \tag{94}$$

for $k = 1, 2, 3, \cdots$. Note that

$$e_x^k(0) = 0, \tag{95}$$

and

$$e_\lambda^k(T) = 0, \tag{96}$$

for each k. Note also that we have not considered yet the truncation errors due to the ODE solvers employed. We will use the maximum norm in this error analysis: for any continuous function $y(t)$ defined on $[0,T]$,

$$||y|| = \max_{0\le t\le T} |y(t)|. \tag{97}$$

Definition 5. *A function F: $\Re^m \to \Re$ is called Lipschitz continuous with a Lipschitz constant L if, for all $X = (x_1, x_2, \cdots, x_m)$ and $Y = (y_1, y_2, \cdots, y_m)$ in $\Re^m$,*

$$|F(X) - F(Y)| \le Ld(X,Y), \tag{98}$$

where $d(X,Y) = \Big[\sum_{i=1}^{m}(x_i - y_i)^2\Big]^{\frac{1}{2}}$.

We will establish the following result.

Theorem 6. *Let $h(t,x,\lambda), g(t,x,u), f_x(t,x,u)$ and $g_x(t,x,u)$ be all Lipschitz continuous with Lipschitz constants $L_h, L_g, L_{f_x}, L_{g_x}$, respectively. Denote*

$$M_\lambda = \max_{0\le t\le T} |\lambda(t)|, \qquad M_{g_x} = \max_{0\le t\le T} |g_x(t)|. \tag{99}$$

Also assume $TL_g < 1, TM_{g_x} < 1$. Then

$$||e_u^{k+1}|| \le \gamma ||e_u^k||, \tag{100}$$

where γ is defined in (109), and if $\gamma < 1$, the iterative scheme (89 - 91) converges on $[0,T]$ with any start-up error.

Remark 7. *If a function is Lipschitz continuous, then it must be continuous.*

Proof of Theorem. Consider that

$$\begin{aligned}
|e_u^{k+1}(t)| = |u^{k+1}(t) - u(t)| &= |h(t,x^{k+1},\lambda^{k+1}) - h(t,x,\lambda)| \\
&\le L_h\sqrt{(x^{k+1}-x)^2 + (\lambda^{k+1}-\lambda)^2} \\
&\le L_h\Big(|x^{k+1}(t) - x(t)| + |\lambda^{k+1}(t) - \lambda(t)|\Big) \\
&\le L_h||x^{k+1} - x|| + L_h||\lambda^{k+1} - \lambda||,
\end{aligned}$$

for all $0 \le t \le T$. Hence,

$$||e_u^{k+1}|| \le L_h||e_x^{k+1}|| + L_h||e_\lambda^{k+1}||. \tag{101}$$

Next,

$$\begin{aligned} e_x^{k+1}(t) &= e_x^{k+1}(0) + \int_0^t \dot{e}_x^{k+1}(\tau)d\tau \\ &= 0 + \int_0^t \Big[g(\tau, x^{k+1}, u^k) - g(\tau, x, u)\Big]d\tau. \end{aligned}$$

Thus

$$\begin{aligned} |e_x^{k+1}(t)| &\leq \int_0^t \Big|g(\tau, x^{k+1}, u^k) - g(\tau, x, u)\Big|d\tau \\ &\leq \int_0^t \Big[L_g|x^{k+1} - x| + L_g|u^k - u|\Big]d\tau \\ &\leq TL_g||x^{k+1} - x|| + TL_g||u^k - u||, \end{aligned}$$

which yields

(102)

$$||e_x^{k+1}|| \leq TL_g||e_x^{k+1}|| + TL_g||e_u^k||.$$

Hence,

(103)

$$||e_x^{k+1}|| \leq T\beta||e_u^k||,$$

where

(104)

$$\beta = \frac{L_g}{1 - TL_g}.$$

Similarly,

$$\begin{aligned} e_\lambda^{k+1}(t) &= e_\lambda^{k+1}(T) - \int_t^T \dot{e}_\lambda^{k+1}(\tau)d\tau \\ &= 0 - \int_t^T -\Big[f_x(\tau, x^{k+1}, u^k) - f_x(\tau, x, u) \\ &\qquad + \lambda^{k+1}g_x(\tau, x^{k+1}, u^k) - \lambda g_x(\tau, x, u)\Big]d\tau. \end{aligned}$$

Thus

$$
\begin{aligned}
|e_\lambda^{k+1}(t)| \le & \int_t^T \Big| f_x(\tau, x^{k+1}, u^k) - f_x(\tau, x, u) \Big| d\tau + \\
& \int_t^T \Big| \lambda^{k+1} g_x(\tau, x^{k+1}, u^k) - \lambda g_x(\tau, x^{k+1}, u^k) \Big| d\tau + \\
& \int_t^T \Big| \lambda g_x(\tau, x^{k+1}, u^k) - \lambda g_x(\tau, x, u) \Big| d\tau \\
\le & \int_t^T \Big[L_{f_x} ||x^{k+1} - x|| + L_{f_x} ||u^k - u|| \Big] d\tau + \int_t^T M_{g_x} ||\lambda^{k+1} - \lambda|| d\tau \\
& + \int_t^T M_\lambda L_{g_x} \Big[||x^{k+1} - x|| + ||u^k - u|| \Big] d\tau,
\end{aligned}
$$

which yields

$$
\begin{aligned}
||e_\lambda^{k+1}|| \le & T L_{f_x} \Big(||e_x^{k+1}|| + ||e_u^k|| \Big) + T M_{g_x} ||e_\lambda^{k+1}|| \\
& + T M_\lambda L_{g_x} \Big(||e_x^{k+1}|| + ||e_u^k|| \Big).
\end{aligned} \tag{105}
$$

Hence,

$$
||e_\lambda^{k+1}|| \le \omega T ||e_x^{k+1}|| + \omega T ||e_u^k||, \tag{106}
$$

where

$$
\omega = \frac{L_{f_x} + M_\lambda L_{g_x}}{1 - T M_{g_x}}. \tag{107}
$$

Substituting (103) and (106) into (101), we obtain

$$
||e_u^{k+1}|| \le \gamma ||e_u^k||, \tag{108}
$$

with

$$
\gamma = T L_h [(1 + \omega T)\beta + \omega]. \tag{109}
$$

Since (108) holds for any k, we obtain, by iterating back on k:

$$
||e_u^{k+1}|| \le \gamma^2 ||e_u^{k-1}|| \le \cdots \le \gamma^{k+1} ||e_u^0||, \tag{110}
$$

where $||e_u^0||$ measures the start-up error. Particularly, if $u^0 = 0$ is set, then $||e_u^0|| = ||u||$. Based on (103), (106) and (110), it is clear that the iterations converge for all $t \in [0, T]$ if $\gamma < 1$. □

Remark 8. *Base on (109), it is always possible to pick T small enough so that $\gamma < 1$ is satisfied; and the smaller T, the faster convergence. However, in most practical applications, we may not have the choice of changing the domain. This is one motivation for our locally refined FBSM.*

Remark 9. *When numerically implementing the FBSM, since the exact solutions are unknown, the convergence check is usually based on the backward errors $u^{k+1} - u^k$, etc. If we redefine the error functions*

(111)
$$e_u^{k+1}(t) = u^{k+1}(t) - u^k(t),$$

(112)
$$e_x^{k+1}(t) = x^{k+1}(t) - x^k(t),$$

(113)
$$e_\lambda^{k+1}(t) = \lambda^{k+1}(t) - \lambda^k(t),$$

then it is easy to see that the error estimates (103), (106) and (108) still hold.

4.1.2 LOCALLY REFINED FBSM AND ERROR ANALYSIS

The result in Theorem 6 from the previous section, particularly the criterion $\gamma < 1$, does provide some insight into the convergence property of the FBSM. Note that γ depends on T, the length of the time domain, as well as the Lipschitz constants associated with the functions involved in the system. Hence, for larger time domains (necessary for long-term study of disease dynamics) and/or more sophisticated epidemic models where those Lipschitz constants may not exist or may not be small enough, the condition $\gamma < 1$ can be easily violated which could cause trouble in convergence. Numerically, this issue stems from the way the iterative procedure is performed in the FBSM: for each iteration, the state and adjoint equations are solved in the entire domain. Thus, errors due to initial guess can easily accumulate through each global iteration, leading to divergent results.

Based on this observation, better iterative methods can be naturally designed with focus on reducing the errors from each global iteration. A starting point is to introduce an inner cycle of *local* iterations within each global iteration. That is, we divide the global domain $[0, T]$ into a set of intervals $[t_j, t_{j+1}]$ $(j = 0, 1, \cdots, N)$, then we solve the optimal control system on each interval $[t_j, t_{j+1}]$ through an iterative manner (referred to as the *local* iteration) until convergence achieved, and then move to next interval. A shooting method, either going forward or backward in time, can be combined with the local iterative procedure to deal with the separation of the initial condition for the state variables and the final condition for the adjoint variables. Below we will discuss in detail the locally refined iterative algorithms applied to optimal control problems, with a focus on error analysis.

We first consider the locally refined FBSM in a special case:

(114)
$$\dot{x}(t) = g(t, x, u),$$

(115)
$$\dot{\lambda}(t) = s(t, u, \lambda) \quad \text{independent of } x,$$

(116)
$$u(t) = h(t, \lambda) \quad \text{independent of } x.$$

The iterative procedure is only needed for λ and u. Specifically, we solve λ and u backward in time from $t = T$ to $t = 0$, with local iterative procedure performed on

each interval $[t_{n-1}, t_n]$. At the $(k+1)^{th}$ iteration, we have

(117)
$$u^{k+1} = h(t, \lambda^k), \qquad t_{n-1} \le t \le t_n,$$

(118)
$$\dot{\lambda}^{k+1} = s(t, u^{k+1}, \lambda^{k+1}), \quad t_{n-1} \le t \le t_n.$$

The error analysis can be easily conducted, assuming that s is also Lipschitz continuous with Lipschitz constant L_s. First,

(119)
$$||u^{k+1} - u|| \le L_h ||\lambda^k - \lambda||.$$

Next,

$$\begin{aligned}\lambda^{k+1}(t) &= \lambda^{k+1}(t_n) - \int_t^{t_n} \dot{\lambda}^{k+1}(\tau) d\tau \\ &= \lambda^{k+1}(t_n) - \int_t^{t_n} s(\tau, u^{k+1}, \lambda^{k+1}) d\tau.\end{aligned}$$

Note that $\lambda^{k+1}(t_n)$ is a constant independent of k, it is determined from the calculation on the previous interval $[t_n, t_{n+1}]$. At the current interval $[t_{n-1}, t_n]$, $\lambda^0(t_n) = \lambda^1(t_n) = \cdots = \lambda^k(t_n)$ for any k. Thus,

$$\begin{aligned}|\lambda^{k+1}(t) - \lambda(t)| \le\ & |\lambda^{k+1}(t_n) - \lambda(t_n)| \\ &+ \int_t^{t_n} |s(\tau, u^{k+1}, \lambda^{k+1}) - s(\tau, u, \lambda)| d\tau.\end{aligned}$$

Denote the constant $|\lambda^{k+1}(t_n) - \lambda(t_n)| = c_n \ge 0$. Then

$$||\lambda^{k+1} - \lambda|| \le c_n + L_s \triangle t ||u^{k+1} - u|| + L_s \triangle t ||\lambda^{k+1} - \lambda||.$$

Let $\triangle t$ be small enough such that $\triangle t L_s < 1$. Then

(120)
$$||\lambda^{k+1} - \lambda|| \le \frac{c_n}{1 - \triangle t L_s} + \frac{\triangle t L_s}{1 - \triangle t L_s} ||u^{k+1} - u||.$$

Substitution of (119) into (120) yields

(121)
$$||\lambda^{k+1} - \lambda|| \le d_n + \gamma ||\lambda^k - \lambda||,$$

where

(122)

$$d_n = \frac{c_n}{1 - \Delta t L_s}, \qquad \gamma = \frac{\Delta t L_s L_h}{1 - \Delta t L_s}.$$

Iterating back on k, (121) yields

$$\begin{aligned} ||\lambda^{k+1} - \lambda||_{[t_{n-1},t_n]} &\leq d_n + \gamma[d_n + \gamma||\lambda^{k-1} - \lambda||] \leq \cdots \\ &\leq d_n(1 + \gamma + \cdots + \gamma^k) + \gamma^{k+1}||\lambda^0 - \lambda|| \\ &= d_n \frac{1 - \gamma^{k+1}}{1 - \gamma} + \gamma^{k+1}||\lambda^0 - \lambda||. \end{aligned}$$

Clearly, the iterations converge if $\gamma < 1$, which can always be satisfied with sufficiently small Δt (see 122).

Remark 10. *By the assumption that convergence is already achieved in the previous interval* $[t_n, t_{n+1}]$, *we have* $c_n = |\lambda^{k+1}(t_n) - \lambda(t_n)| < \epsilon$ *for some (arbitrarily) small* ϵ.

Next we consider another special case

(123)

$$\dot{x}(t) = g(t, x, u),$$

(124)

$$u(t) = h(t, x) \quad \text{independent of } \lambda,$$

(125)

$$\dot{\lambda}(t) = -[f_x(t, x, u) + \lambda g_x(t, x, u).$$

Iterations are only performed for x and u. We solve x and u forward in time from $t = 0$ to $t = T$. On each interval $[t_{n-1}, t_n]$, we conduct the local iterations

(126)

$$u^{k+1} = h(t, x^k), \quad t_{n-1} \leq t \leq t_n,$$

(127)

$$\dot{x}^{k+1} = g(t, x^{k+1}, u^{k+1}), t_{n-1} \leq t \leq t_n,$$

for $k = 0, 1, 2, \cdots$.

Similar analysis can be conducted to obtain that if

(128)

$$\gamma = \frac{\Delta t L_g L_h}{1 - \Delta t L_g} < 1,$$

then the iterative procedure will converge. Again this can be satisfied by setting Δt small enough.

Now for a general case, we can perform the following steps.

Step 1. Make initial guesses for x and u on $[0,T]$; denote these values by $x^0(t)$ and $u^0(t), 0 \le t \le T$.

Step 2. Solve $\lambda(t)$ backward from $t = T$ to $t = 0$, using

(129)
$$\dot{\lambda}(t) = -[f_x(t, x^0, u^0) + \lambda g_x(t, x^0, u^0)],$$

(130)
$$\lambda(T) = b.$$

Denote the solution by $\lambda^0(t), 0 \le t \le T$.

Step 3. Solve x, u and λ forward in time with local iterative procedure . On each interval $[t_{n-1}, t_n]$, conduct the iterations

(131)
$$u^{k+1} = h(t, x^k, \lambda^k),$$

(132)
$$\dot{x}^{k+1} = g(f, x^{k+1}, u^{k+1}),$$

(133)
$$\dot{\lambda}^{k+1} = -[f_x(t, x^{k+1}, u^{k+1}) + \lambda^{k+1} g_x(t, x^{k+1}, u^{k+1})],$$

for $k = 0, 1, 2, \cdots$.

Step 4. After convergent solutions are achieved on each interval throughout $[0,T]$, re-set these solutions to $x^0(t), u^0(t)$ and $\lambda^0(t), 0 \le t \le T$. Then check the value of $|\lambda^0(T) - b|$: If $|\lambda^0(T) - b| < \epsilon$, task completed; otherwise, go to *Step 2* for next round of iterations.

Using similar analysis as before, we obtain, based on the formulation (131 - 133),

(134)
$$||e_x^k|| \le \Delta t \beta ||e_u^k||,$$

(135)
$$||e_u^{k+1}|| \le L_h ||e_x^k|| + L_h ||e_\lambda^k||,$$

(136)
$$||e_\lambda^k|| \le \omega \Delta t ||e_x^k|| + \omega \Delta t ||e_u^k||.$$

Substitution of (134) and (136) into (135) yields

(137)

$$||e_u^{k+1}|| \leq ||e_u^k||,$$

where

(138)

$$\gamma = \triangle t L_h[\omega + \beta(1 + \triangle t\omega)],$$

and where β and ω are defined in (104), (107), respectively, but with T replaced by $\triangle t$. That is,

(139)

$$\beta = \frac{L_g}{1 - \triangle t L_g}, \qquad \omega = \frac{L_{f_x} + M_\lambda L_{g_x}}{1 - \triangle t M_{g_x}}.$$

The condition $\gamma < 1$ would ensure the convergence. Thus, when $\triangle t$ is small enough, the convergence is guaranteed. Compare (138) to (109) and we can see the condition is much weaker for our current method. If, instead, (131 - 133) are replaced by

(140)

$$\dot{x}^{k+1} = g(t, x^{k+1}, u^k),$$

(141)

$$u^{k+1} = h(t, x^{k+1}, \lambda^k),$$

(142)

$$\dot{\lambda}^{k+1} = -\Big[f_x(t, x^{k+1}, u^{k+1}) + \lambda^{k+1} g_x(t, x^{k+1}, u^{k+1})\Big].$$

Then we have

(143)

$$||e_x^{k+1}|| \leq \triangle t\beta ||e_u^k||,$$

(144)

$$||e_u^{k+1}|| \leq L_h||e_x^{k+1}|| + L_h||e_\lambda^k||,$$

(145)

$$||e_\lambda^k|| \leq \omega\triangle t||e_x^k|| + \omega\triangle t||e_u^k||.$$

Substituting (143) and (145) into (144), we obtain

$$||e_u^{k+1}|| \leq \triangle t L_h(\beta + \omega)||e_u^k|| + \triangle t^2 L_h\omega\beta||e_u^{k-1}||.$$

This inequality can be rewritten as

(146)
$$||e_u^{k+1}|| + \alpha||e_u^k|| \leq \gamma\Big[||e_u^k|| + \alpha||e_u^{k-1}||\Big],$$

where

$$\alpha = \triangle t\frac{-L_h(\beta+\omega) + \sqrt{L_h^2(\beta+\omega)^2 + 4L_h\omega\beta}}{2},$$

and

(147)
$$\gamma = \triangle t\frac{L_h(\beta+\omega) + \sqrt{L_h^2(\beta+\omega)^2 + 4L_h\omega\beta}}{2}.$$

We clearly observe from (146) that as long as $\gamma < 1$ (which is guaranteed for small $\triangle t$), the iterative method will converge.

A simple example described below is to illustrate the methods mentioned in this section:

$$\min \quad x(t) + \int_0^T x(t)u^2(t)dt,$$
$$\text{subject to } \dot{x}(t) = \frac{1}{2}x(t) - x(t)u(t), \quad x(0) = a,$$

and it is very easy to find the exact solution as

$$u^*(t) = \frac{1}{1+e^{0.5(t-T)}},$$
$$\lambda^*(t) = \frac{2}{1+e^{0.5(t-T)}}.$$

The analytical solution will be used to compare with our numerical solution.

We use $T = 10$ and $a = 1$ in the numerical simulations shown below.

We also check the number of iterations before the convergence of each method and it is shown in Table 6.

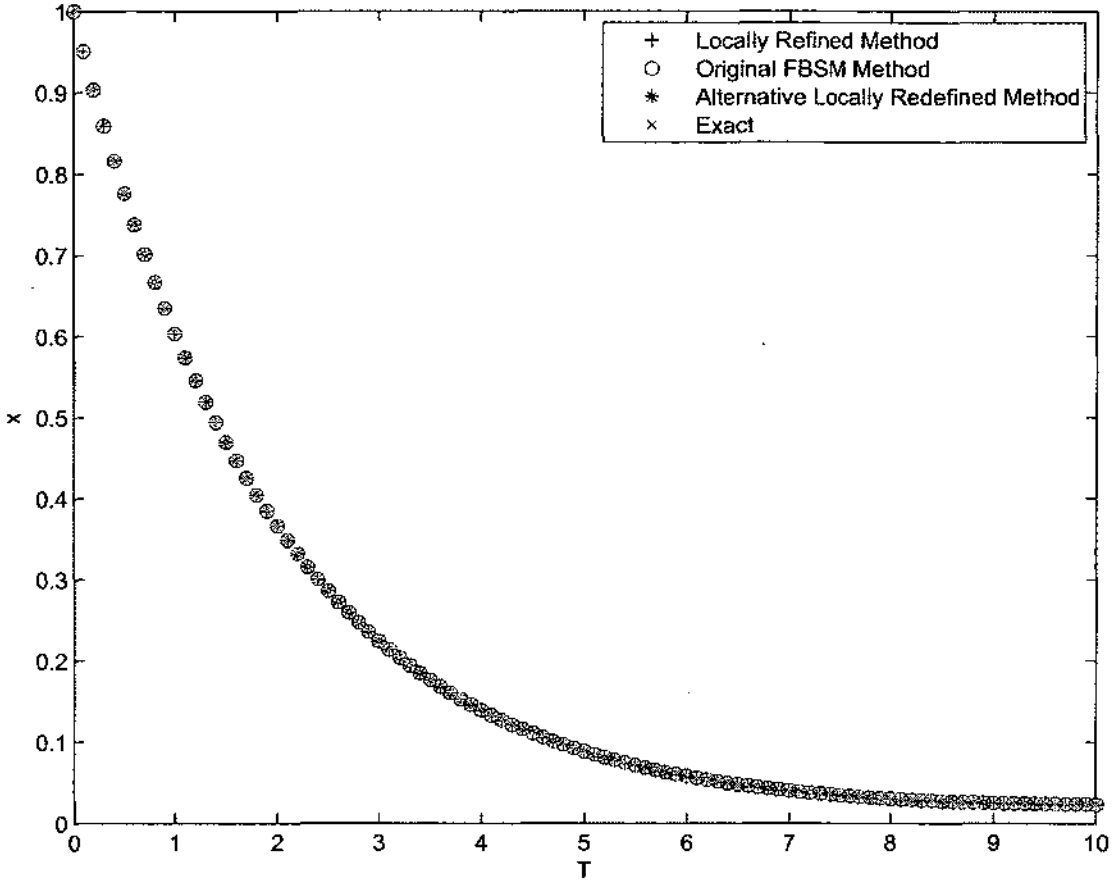

FIGURE 48: The graph shows that all methods can solve x very well.

Method	Number of Iterations
Regular Forward-Backward Sweep Method	32
Locally Refined FBSM	20
Alternative Locally Redefined Method	20

TABLE. 6: Number of iterations for convergence.

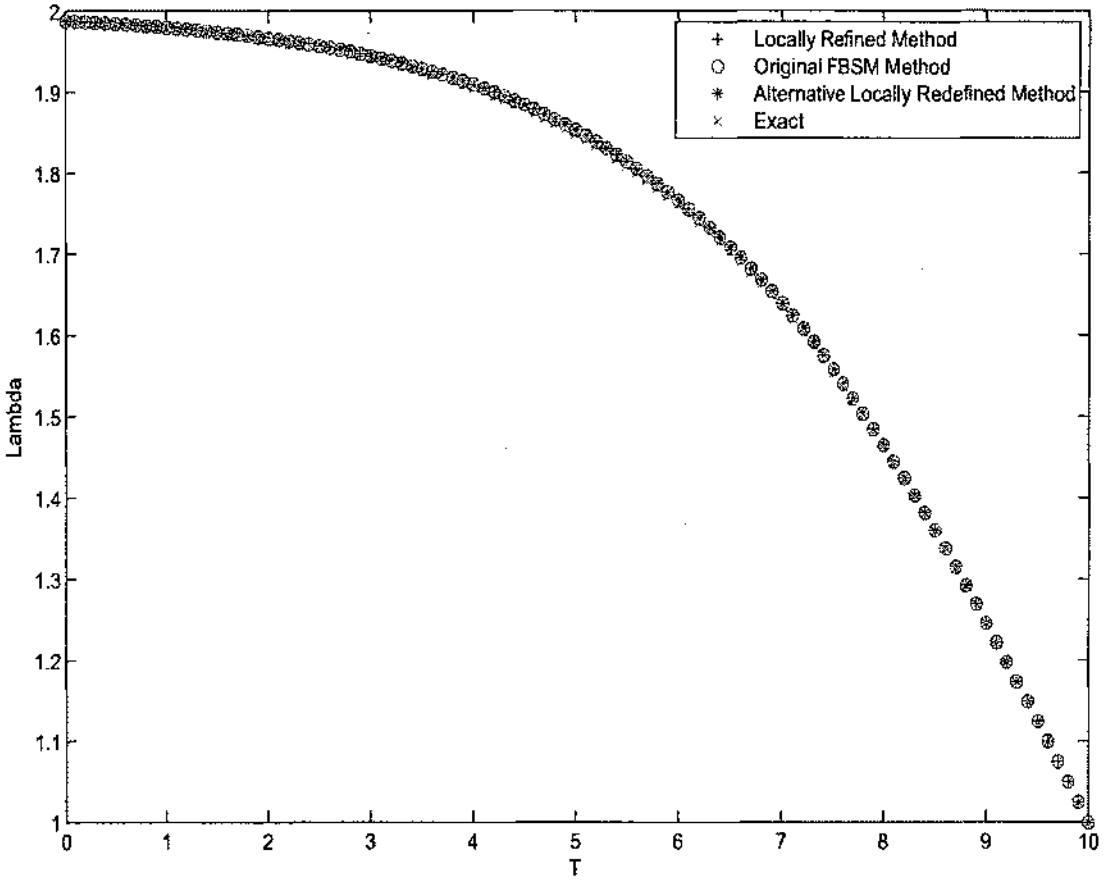

FIGURE 49: The graph shows the numerical results for the adjoint variable, λ.

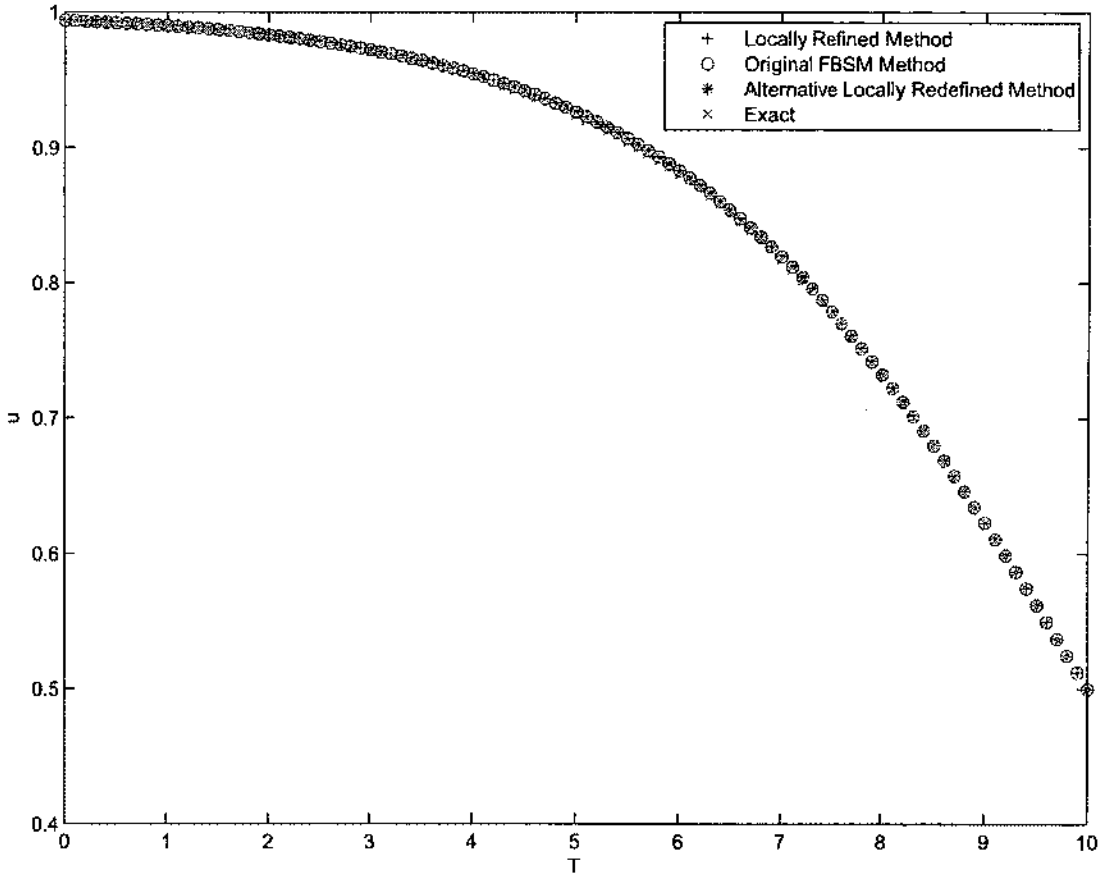

FIGURE 50: The graph shows the control variable, u, calculated numerically by all methods.

4.2 LOCAL ITERATIVE ALGORITHMS FOR A CLASS OF CONSTRAINED DYNAMICAL PROBLEMS

In this section, we extend the idea of locally refined iterative procedures to the study for a class of constrained dynamical problems involving second-order linear differential-algebraic equations:

(148)
$$A_1(t)\ddot{x} + E_1(t)\dot{x} + C_1(t)x = r_1(t) - B_1(t)\lambda,$$

(149)
$$A_2(t)\ddot{y} + E_2(t)\dot{y} + C_2(t)y = r_2(t) - B_2(t)\lambda,$$

(150)
$$D_1(t)\ddot{x} + D_2(t)\ddot{y} = d(t),$$

where x and y are unknown vectors with the same or different dimensions, $\dot{x}$ and $\dot{y}$ denote the first derivatives, and $\ddot{x}$ and $\ddot{y}$ the second derivatives, in time for x and y, respectively, and λ is an *algebraic* unknown (in the sense that its derivative does not appear) which couples x and y. Here the unknowns x, y and λ have dimensions $n_1 \times 1, n_2 \times 1$ and $n_3 \times 1$, respectively. Correspondingly, A_1, E_1 and C_1 are square matrices of dimensions $n_1 \times n_1, n_j \times 1$ for $j = 1, 2$. Meanwhile, the matrices B_j and $D_j (j = 1, 2)$ have dimensions $n_j \times n_3$ and $n_3 \times n_j$, respectively, and d is a vector of dimension $n_3 \times 1$. Generally B_j andD_j may not be square matrices; we typically have $n_3 < n_j (j = 1, 2)$ in most practical applications, meaning that equation (150) merely supply partial information, as additional constraint, for x and y when making connection between the two solutions. Finally, we note that the entries of these matrices or vectors are in general functions of t, where $t \in [0, T]$ for some constant $T > 0$. We assume all these functions are continuous on $[0, T]$ in this work.

Equations (148) - (150) represent a class of dynamical systems where the entire dynamics are determined by those from two sub-systems, x and y, subject to certain constraints. Such problems could have many applications in science and engineering, especially in physical/mechanical discipline. Two-body rigid motion [27], two-phase flow [46, 47] and fluid-structure interaction [33, 31, 48] are just a few typical examples that can be modeled, at the linear level, by equations (148) - (150). In many of these practical applications, the two sub-systems for x and y often have different computational requirements in terms of accuracy and efficiency. The proposed iterative approach will not only improve the accuracy of the numerical solution but

also allow us to employ different ODE solvers and use different local meshes for x and y, since the solution procedures for x and y will be uncoupled through the local iterations.

To ensure the well-posedness of the dynamical problem and the validity of the numerical algorithm, we will further make the following assumptions for the system (148 - 150):

(**A1**) The matrices $A_1(t)$ and $A_2(t)$ are invertible for all $t \in [0, T]$.

(**A2-1**) The $n_3 \times n_3$ matrix $D_1(t)A_1^{-1}(t)B_1(t)$ is invertible for all $t \in [0, T]$.

(**A2-2**) The $n_3 \times n_3$ matrix $D_2(t)A_2^{-1}(t)B_2(t)$ is invertible for all $t \in [0, T]$.

We will assume that (A1) and at least one of (A2-1) and (A2-2) hold in this study.

4.2.1 ALGORITHMS AND ERROR ANALYSIS

The dynamic problem (148) - (150) can be assembled into a block-matrix system

(151)
$$\begin{bmatrix} A_1 & 0 & B_1 \\ 0 & A_2 & B_2 \\ D_1 & D_2 & 0 \end{bmatrix} \begin{bmatrix} \ddot{x} \\ \ddot{y} \\ \lambda \end{bmatrix} + \begin{bmatrix} E_1 & 0 & 0 \\ 0 & E_2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x} \\ \dot{y} \\ \lambda \end{bmatrix} + \begin{bmatrix} C_1 & 0 & 0 \\ 0 & C_2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ \lambda \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ d \end{bmatrix}.$$

Regardless the dimensions of B_j and $D_j (j = 1, 2)$, the two coefficient block-matrices (each consisting of 9 blocks) are always square matrices of dimensions $(n_1 + n_2 + n_3) \times (n_1 + n_2 + n_3)$.

The iterative method presented below decoupled the computation for x and y while strictly maintaining the constraint. It is based on splitting the leading coefficient block matrix in equation (151). The first approach is to impose the constraint (150) to equation (148). Consequently, the leading terms (i. e., those with the second derivatives of the unknowns) in equation (151) are decomposed as

(152)
$$\begin{bmatrix} A_1 & 0 & B_1 \\ 0 & A_2 & 0 \\ D_1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \ddot{x} \\ \ddot{y} \\ \lambda \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & B_2 \\ 0 & D_2 & 0 \end{bmatrix} \begin{bmatrix} \ddot{x} \\ \ddot{y} \\ \lambda \end{bmatrix}.$$

The first part in (152),as well as all other terms in equation (151) associated with the unknowns and their first derivatives, will be computed at the current iteration, labeled as $i + 1$, whereas the second part in equation (152) will be evaluated using

the values from the previous iteration, i. This leads to an iterative scheme with two separated equations,

(153)
$$\begin{bmatrix} A_1 & B_1 \\ D_1 & 0 \end{bmatrix} \begin{bmatrix} \ddot{x}^{i+1} \\ \lambda^{i+1} \end{bmatrix} + \begin{bmatrix} E_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}^{i+1} \\ \lambda^{i+1} \end{bmatrix} + \begin{bmatrix} C_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x^{i+1} \\ \lambda^{i+1} \end{bmatrix} = \begin{bmatrix} r_1 \\ d - D_2\ddot{y}^i \end{bmatrix},$$

and

(154)
$$A_2\ddot{y}^{i+1} + E_2\dot{y}^{i+1} + C_2y^{i+1} = r_2 - B_2\lambda^{i+1},$$

where the superscript i denotes the solution at the $i$tth iteration, $i = 0, 1, 2, \cdots$. At each iteration, equations (153) and (154) will be solved separately by employing some ODE solvers.

The second approach is to impose the constraint (150) to equation (149), so that the leading terms in equation (151) are decomposed as

(155)
$$\begin{bmatrix} A_1 & 0 & 0 \\ 0 & A_2 & B_2 \\ 0 & D_2 & 0 \end{bmatrix} \begin{bmatrix} \ddot{x} \\ \ddot{y} \\ \lambda \end{bmatrix} + \begin{bmatrix} 0 & 0 & B_1 \\ 0 & 0 & 0 \\ D_1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \ddot{x} \\ \ddot{y} \\ \lambda \end{bmatrix}.$$

Correspondingly, an iterative scheme based on (155) can be constructed by

(156)
$$\begin{bmatrix} A_2 & B_2 \\ D_2 & 0 \end{bmatrix} \begin{bmatrix} \ddot{y}^{i+1} \\ \lambda^{i+1} \end{bmatrix} + \begin{bmatrix} E_2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{y}^{i+1} \\ \lambda^{i+1} \end{bmatrix} + \begin{bmatrix} C_2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y^{i+1} \\ \lambda^{i+1} \end{bmatrix} = \begin{bmatrix} r_2 \\ d - D_1\ddot{x}^i \end{bmatrix},$$

and

(157)
$$A_1\ddot{x}^{i+1} + E_1\dot{x}^{i+1} + C_1x^{i+1} = r_1 - B_1\lambda^{i+1}.$$

Equations (156) and (157) will also be solved separately by some ODE solvers at each iteration.

Lemma 11. *Let* $F = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$ *be a square matrix, where* A, B, C *and* D *are matrix sub-blocks of arbitrary size, with* A *and* D *being square. Then the matrix* F

is invertible if and only if A and $D - CA^{-1}B$ are invertible, and the inverse of F is given by

(158)
$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix}.$$

See [53], for example, for a proof of Lemma 11. Based on this result, it is straightforward to obtain

Lemma 12. *The leading coefficient matrix, $\begin{bmatrix} A_1 & B_1 \\ D_1 & 0 \end{bmatrix}$, of equation (153) is invertible under the assumptions (A1) and (A2-1). Similarly, the leading coefficient matrix, $\begin{bmatrix} A_2 & B_2 \\ D_2 & 0 \end{bmatrix}$, of equation (156) is invertible under the assumptions (A1) and (A2-2).*

The following theorem is a direct consequence of Lemma 12 and classical differential equation theory.

Theorem 13. *Let assumptions (A1) and (A2-1) hold. Also assume $\ddot{y}^i$ is continuous. Then the system of (153) and (154) has a unique solution $\{x^{i+1}, y^{i+1}, \lambda^{i+1}\}$ on $[0, T]$ with a given initial condition. Similarly, if the assumptions (A1) and (A2-2) hold and $\ddot{x}^i$ is continuous, then the system of (156) and (157) has a unique solution $\{x^{i+1}, y^{i+1}, \lambda^{i+1}\}$ on $[0, T]$ with a given initial condition.*

Proof. We show the second half of the theorem; the proof for the first half is similar. Assume (A1) and (A2-2) hold. Then the matrix $\begin{bmatrix} A_2 & B_2 \\ D_2 & 0 \end{bmatrix}$ is invertible based on Lemma 12. Let $\begin{bmatrix} G_1 & G_2 \\ G_3 & G_4 \end{bmatrix}$ denote the inverse and multiply it to both sides of equation (156) to obtain

(159) $$\ddot{y}^{i+1} + G_1E_2\,\dot{y}^{i+1} + G_1C_2\,y^{i+1} = G_1r_2 + G_2d - G_2D_1\ddot{x}^i,$$

and

(160) $$\lambda^{i+1} + G_3E_2\,\dot{y}^{i+1} + G_3C_2\,y^{i+1} = G_3r_2 + G_4d - G_4D_1\ddot{x}^i.$$

Since all coefficient matrices are assumed to be continuous, clearly equation (159) has a unique solution for y^{i+1} on $[0, T]$. Consequently, λ^{i+1} is uniquely solvable based on equation (160). Finally, since A_1 is invertible, equation (157) has a unique solution for x^{i+1}. □

Theorem 13 ensures the solvability of the iterative formulation based on equations (156) and (157), or (153) and (154). Consequently, they build the ground for a local iterative procedure to advance the solution from $t = t_n$ to $t = t_{n+1}$. The procedure is presented below, and equations (156) and (157) are employed here for illustration.

Assume that $x(t_n), y(t_n), \lambda(t_n)$, and $\ddot{x}(t_n)$ are known. We construct a local mesh by dividing the interval $[t_n, t_{n+1}]$ into a set of subintervals. For example, if it is uniformly divided into N subintervals, then the local mesh points are given by $t_{n,j} = t_n + j \cdot \Delta t/N$ for $0 \leq j \leq N$. We also assume that $x^i(t), y^i(t), \lambda^i(t)$, and $\ddot{x}^i(t)$ are known for $t_n \leq t \leq t_{n+1}$. Particularly, for the first iteration with $i = 0$, we make an initial guess by setting

$$x^0(t) = x(t_n), \quad y^0(t) = y(t_n), \quad \lambda^0(t) = \lambda(t_n), \quad \ddot{x}^0(t) = \ddot{x}(t_n), \quad \text{for} \quad t_n \leq t \leq t_{n+1}.$$

For $i > 0$, we carry out the following steps to obtain the solution at the (i+1)th iteration.

Step 1. Compute y^{i+1} and λ^{i+1} on $[t_n, t_{n+1}]$ by solving equation (156) with the initial conditions

(161)
$$y^{i+1}(t_n) = y(t_n), \quad \dot{y}^{i+1}(t_n) = \dot{y}(t_n), \quad \lambda^{i+1}(t_n) = \lambda(t_n).$$

The numerical solution of (156) and (161) is then reported at $t = t_{n+1}$. We calculate the norms of the backward errors

$$\varepsilon_y = ||y^{i+1}(t_{n+1}) - y^i(t_{n+1})||, \quad \text{and} \quad \varepsilon_\lambda = ||\lambda^{i+1}(t_{n+1}) - \lambda^i(t_{n+1})||.$$

Step 2. Compute x^{i+1} on $[t_n, t_{n+1}]$ by solving equation (157) with the initial conditions

(162)
$$x^{i+1}(t_n) = x(t_n), \qquad \text{and} \quad \dot{x}^{i+1}(t_n) = \dot{x}(t_n).$$

We calculate the norm of the backward error

$$\varepsilon_x = ||x^{i+1}(t_{n+1}) - x^i(t_{n+1})||.$$

Step 3. Check the convergence. If

$$\max(\varepsilon_x, \varepsilon_y, \varepsilon_\lambda) \leq \varepsilon_0,$$

where ε_0 is the given error tolerance, then the convergence has been achieved. Start the iterative procedure for the next time step, $n = n + 1$. Otherwise, set $i = i + 1$ and return to *Step 1.*

To facilitate the error analysis, we will use the L_∞ norm, denoted by $||\cdot||$, for constant matrices and vectors. For any constant vector $u = [u_1, u_2, \cdots, u_n]^T$ and matrix $A = [a_{ij}]_{m\times n}$, we have

$$||u|| = \max_{1\leq i\leq n} |u_i|, \quad ||A|| = \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}|,$$

and

$$||A\,u|| \leq ||A||\,||u||. \tag{163}$$

Meanwhile, we introduce the following M-norm, denoted by $||\cdot||_M$, for both vector-valued and matrix-valued continuous functions. Let $u = [u_1(t), u_2(t), \cdots, u_n(t)]^T$ be a vector whose entries are continuous functions of t on some closed interval $[a, b]$. We define

$$||u(t)||_M = \max_{1\leq i\leq n} \max_{a\leq t\leq b} |u_i(t)| = \max_{a\leq t\leq b} \max_{1\leq i\leq n} |u_i(t)|. \tag{164}$$

Similarly, let $A(t) = [a_{ij}(t)]_{m\times n}$ denote a matrix with each entry being a continuous function of t on $[a, b]$. We define

$$||A(t)||_M = \max_{a\leq t\leq b} \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}(t)| = \max_{1\leq i\leq m} \max_{a\leq t\leq b} \sum_{j=1}^{n} |a_{ij}(t)|. \tag{165}$$

Next, we show that the matrix M-norm (for $A(t)$) and vector M-norm (for $u(t)$) are consistent, and this result will be frequently used in our error analysis.

Lemma 14.

$$||A(t)\,u(t)||_M \leq ||A(t)||_M\,||u(t)||_M \tag{166}$$

Proof. We have

$$A\,u(t) = \Big[\sum_{j=1}^{n} a_{1j}\,u_j(t), \sum_{j=1}^{n} a_{2j}\,u_j(t), \cdots, \sum_{j=1}^{n} a_{mj}\,u_j(t)\Big]^T$$

Hence,

$$||A\,u(t)||_M \;=\; \max_{1\le i\le m}\max_{a\le t\le b}\Big|\sum_{j=1}^{n} a_{ij}(t)\,u_j(t)\Big| \;\le\; \max_{1\le i\le m}\max_{a\le t\le b}\sum_{j=1}^{n}\big|a_{ij}(t)\big|\,\big|u_j(t)\big|$$

$$\le \max_{1\le i\le m}\max_{a\le t\le b}\sum_{j=1}^{n}\big|a_{ij}(t)\big|\,\Big(||u(t)||_M\Big) \;=\; ||A(t)||_M\,||u(t)||_M$$

The proof is then complete. □

Remark 15. *It is easy to observe that if I_1 and I_2 are two intervals such that $u(t)$ and $A(t)$ are defined on both intervals and that $I_1 \subset I_2$, then*

$$||u(t)||_{M,I_1} \;\le\; ||u(t)||_{M,I_2}\,, \quad \textit{and} \quad ||A(t)||_{M,I_1} \;\le\; ||A(t)||_{M,I_2}$$

where $||\cdot||_{M,I_j}$ refers to the M-norm imposed on the interval I_j, $j = 1, 2$.

Lemma 16. *Let $U(t)$ be the solution vector of the "one-step" initial value problem*

$$(167)\quad \begin{cases} \ddot{U}(t) \;=\; S(t)\,\dot{U}(t) + R(t)\,U(t) + H(t)\,, & t_n \le t \le t_{n+1} \\ U(t_n) \;=\; 0\,, \quad \dot{U}(t_n) \;=\; 0 \end{cases}$$

where $S(t)$ is a continuously differentiable matrix-valued function, and $R(t)$ and $H(t)$ are continuous matrix- and vector-valued functions, respectively. Then for sufficiently small Δt, $U(t)$ satisfies

$$(168)\qquad ||U(t)||_M \;\le\; \frac{1}{2}\Delta t^2\,\big(1 + O(\Delta t)\big)\,||H(t)||_M,$$

where the M-norm is imposed on $[t_n\,,\ t_{n+1}]$.

Proof. For $t \in [t_n\,,\ t_{n+1}]$, we have

$$\begin{aligned} U(t) \;&=\; U(t_n) + \int_{t_n}^{t} \dot{U}(\tau)\,d\tau \\ &=\; U(t_n) + (t - t_n)\,\dot{U}(t_n) + \int_{t_n}^{t}\int_{t_n}^{\tau} \ddot{U}(r)\,dr\,d\tau \\ &=\; 0 + 0 + \int_{t_n}^{t}\int_{t_n}^{\tau} \Big[S(r)\,\dot{U}(r) + R(r)\,U(r) + H(r)\Big]\,dr\,d\tau. \end{aligned}$$

Using integration by parts, we obtain

$$U(t) \;=\; \int_{t_n}^{t} S(\tau)U(\tau)\,d\tau + \int_{t_n}^{t}\int_{t_n}^{\tau} \Big[\,\big(R(r) - \dot{S}(r)\big)U(r) + H(r)\,\Big]\,dr\,d\tau.$$

Taking the M-norm on both sides, we have

$$||U||_M \;\le\; \Delta t\,||S||_M||U||_M + \frac{1}{2}\Delta t^2\big(\,||R-\dot{S}||_M||U||_M + ||H||_M\,\big),$$

which yields

$$\Big(1-\Delta t\,||S||_M - \frac{1}{2}\Delta t^2||R-\dot{S}||_M\Big)||U||_M \;\le\; \frac{1}{2}\Delta t^2||H||_M.$$

It is thus clear that (168) holds for sufficiently small Δt. □

Lemma 17. *(Gronwall Inequality [44, 49])* *Let $u(t) \ge 0$ and $g(t) \ge 0$ be continuous real-valued functions on the interval $[a\,,\,b]$. If there are constants $K \ge 0$ and $L \ge 0$ such that*

$$u(t) \;\le\; L + K\int_a^t g(s)\,u(s)\,ds$$

for all $t \in [a\,,\,b]$, then the inequality

$$u(t) \;\le\; L\,\exp\big[\,K\int_a^t g(s)\,ds\,\big]$$

holds for $a \le t \le b$.

In what follows, we let $x(t)$, $y(t)$ and $\lambda(t)$ be the *exact solution* of the original problem (151) at time t. It is necessary to introduce several other notations before conducting the error analysis. Below we will concentrate on the iterative scheme (156) and (157), and similar analysis can be conducted for the other method, (153) and (154).

Based on equation (148), it is clear that $x(t_{n+1})$ is the value of the *exact solution* to the following "one-step" initial value problem at $t = t_{n+1}$:

$$(169)\quad \begin{cases} A_1\,\ddot{z} + E_1\,\dot{z} + C_1\,z \;=\; r_1 - B_1\,\lambda(t)\,, & t_n \le t \le t_{n+1}, \\ z(t_n) \;=\; x(t_n)\,, \quad \dot{z}(t_n) \;=\; \dot{x}(t_n). \end{cases}$$

Meanwhile, based on equation (157), we denote the *exact solution* of the following one-step initial value problem at $t = t_{n+1}$ by x_{n+1}^{i+1}:

$$(170)\quad \begin{cases} A_1\,\ddot{z} + E_1\,\dot{z} + C_1\,z \;=\; r_1 - B_1\,\lambda^{i+1}(t)\,, & t_n \le t \le t_{n+1}, \\ z(t_n) \;=\; x(t_n)\,, \quad \dot{z}(t_n) \;=\; \dot{x}(t_n), \end{cases}$$

where $\lambda^{i+1}(t)$ represents the solution of the $(i+1)$th iteration for $\lambda(t)$ on $[t_n\,,\,t_{n+1}]$.

Furthermore, we introduce the hat notation to denote the *numerical values* of the solutions and their derivatives. For example, $\widehat{x_{n+1}}$ and $\widehat{\dot{x}_{n+1}}$ refer to the numerical

values of $x(t_{n+1})$ and $\dot{x}(t_{n+1})$, respectively. Also, $\widehat{x_{n+1}^{i+1}}$ $(i = 0, 1, \cdots)$ denotes the numerical solution at t_{n+1} for the $(i+1)$th iteration. Similar hat notations are used for y and λ.

In addition, we let $\overline{x_{n+1}^{i+1}}$ denote the *exact solution* of the initial value problem given below at $t = t_{n+1}$:

$$(171)\quad \begin{cases} A_1\,\ddot{z} + E_1\,\dot{z} + C_1\,z \;=\; r_1 - B_1\,\lambda^{i+1}(t)\,, & t_n \le t \le t_{n+1}, \\ z(t_n) = \widehat{x_n}\,, \quad \dot{z}(t_n) = \widehat{\dot{x}_n}. \end{cases}$$

Here in the initial conditions, $\widehat{\dot{x}_n}$ is the numerical approximation of $\dot{x}(t_n)$ and can be calculated, for example, using a backward difference formula. We may assume that $\widehat{\dot{x}_n}$ is computed at the same order of accuracy as that of $\widehat{x_n}$ with an appropriate numerical formula as long as $x(t)$ is reasonably smooth. This means, there exists a constant $c_1 > 0$ such that

$$(172)\qquad ||\widehat{\dot{x}_n} - \dot{x}(t_n)|| \;\le\; c_1\,||\widehat{x_n} - x(t_n)||.$$

We may define y_{n+1}^{i+1}, $\overline{y_{n+1}^{i+1}}$, and etc., in a similar way. Below we will focus our attention on the error analysis for x, since the error estimates of y and λ will immediately follow the results for x.

Theorem 18. *Let the assumptions (A1) and (A2-2) hold. Let also the matrix J be defined as*

$$(173)\qquad J = -(D_2 A_2^{-1} B_2)^{-1}(D_1 A_1^{-1} B_1).$$

If $||J||_M < 1$, then for sufficiently small $\triangle t$, the iteration (156)(157) converges for all $t \in [t_n, t_{n+1}]$ with any start-up error and any tolerance.

Proof. Let $e_x^{i+1}(t)$ denote the difference between the exact solutions of (169) and (170) on the interval $[t_n\,,\, t_{n+1}]$. Also define $e_y^{i+1}(t)$ in a similar way. In particular, we have

$$(174)\quad e_x^{i+1}(t_{n+1}) \;=\; x_{n+1}^{i+1} - x(t_{n+1})\,, \qquad e_y^{i+1}(t_{n+1}) \;=\; y_{n+1}^{i+1} - y(t_{n+1}),$$

and

$$(175)\quad e_x^{i+1}(t_n) \;=\; e_y^{i+1}(t_n) \;=\; 0\,, \qquad \dot{e}_x^{\,i+1}(t_n) \;=\; \dot{e}_y^{\,i+1}(t_n) \;=\; 0.$$

In addition, we let $e_\lambda^{i+1}(t) = \lambda(t) - \lambda^{i+1}(t)$. When $t_n \le t \le t_{n+1}$, these error functions satisfy

$$A_1 \ddot{e}_x^{i+1} + E_1 \dot{e}_x^{i+1} + C_1 e_x^{i+1} = -B_1 e_\lambda^{i+1}, \tag{176}$$

$$A_2 \ddot{e}_y^{i+1} + E_2 \dot{e}_y^{i+1} + C_2 e_y^{i+1} = -B_2 e_\lambda^{i+1}, \tag{177}$$

$$\text{and} \quad D_2 \ddot{e}_y^{i+1} + D_1 \ddot{e}_x^{i} = 0. \tag{178}$$

In what follows, we proceed in several steps to complete the proof of this theorem.

Step 1: We establish the relationship between the norms of those error functions introduced above.

Equation (176) yields

$$\ddot{e}_x^{i+1} = -A_1^{-1}C_1 e_x^{i+1} - A_1^{-1}E_1 \dot{e}_x^{i+1} - A_1^{-1}B_1 e_\lambda^{i+1}, \tag{179}$$

from which we can solve e_x^{i+1} in terms of $e_\lambda^{i+1}(t)$. Based on Lemmas 14 and 16, we obtain the estimate

$$\|e_x^{i+1}\|_M \le \frac{1}{2}\Delta t^2 \left(1 + O(\Delta t)\right) \|A_1^{-1}B_1\|_M \|e_\lambda^{i+1}\|_M, \tag{180}$$

where the norm $\|\cdot\|_M$, defined in equation (164), is evaluated on the interval $[t_n, t_{n+1}]$. Similarly, from equation (177) we can derive

$$\|e_y^{i+1}\|_M \le \frac{1}{2}\Delta t^2 \left(1 + O(\Delta t)\right) \|A_2^{-1}B_2\|_M \|e_\lambda^{i+1}\|_M. \tag{181}$$

Meanwhile, if we take the M-norm on both sides of equation (179), we have

$$\|\ddot{e}_x^{i+1}\|_M \le \|A_1^{-1}B_1\|_M\|e_\lambda^{i+1}\|_M + \|A_1^{-1}C_1\|_M\|e_x^{i+1}\|_M + \|A_1^{-1}E_1\|_M\|\dot{e}_x^{i+1}\|_M. \tag{182}$$

Using Taylor series expansions at $t = t_n$ for $e_x^{i+1}(t)$ and $e_y^{i+1}(t)$, and applying the initial conditions (175), we can easily obtain

$$\|\dot{e}_x^{i+1}\|_M \le \Delta t \|\ddot{e}_x^{i+1}\|_M \quad \text{and} \quad \|e_x^{i+1}\|_M \le \frac{1}{2}\Delta t^2 \|\ddot{e}_x^{i+1}\|_M, \tag{183}$$

as well as

$$\|\dot{e}_y^{i+1}\|_M \le \Delta t \|\ddot{e}_y^{i+1}\|_M \quad \text{and} \quad \|e_y^{i+1}\|_M \le \frac{1}{2}\Delta t^2 \|\ddot{e}_y^{i+1}\|_M. \tag{184}$$

Substituting (183) into (182), we obtain

$$\left(1 - \Delta t\|A_1^{-1}E_1\|_M - \frac{1}{2}\Delta t^2\|A_1^{-1}C_1\|_M\right)\|\ddot{e}_x^{i+1}\|_M \le \|A_1^{-1}B_1\|_M\|e_\lambda^{i+1}\|_M. \tag{185}$$

Let us denote

$$\beta_1 = 1 - \Delta t||A_1^{-1}E_1||_M - \frac{1}{2}\Delta t^2||A_1^{-1}C_1||_M. \tag{186}$$

Then $\beta_1 > 0$ for sufficiently small Δt. Thus we have

$$||\ddot{e_x}^{i+1}||_M \leq \beta_1^{-1}||A_1^{-1}B_1||_M||e_\lambda^{i+1}||_M. \tag{187}$$

In a similar way, we obtain

$$||\ddot{e_y}^{i+1}||_M \leq \beta_2^{-1}||A_2^{-1}B_2||_M||e_\lambda^{i+1}||_M, \tag{188}$$

with

$$\beta_2 = 1 - \Delta t||A_2^{-1}E_2||_M - \frac{1}{2}\Delta t^2||A_2^{-1}C_2||_M. \tag{189}$$

Step 2: We establish a recurrence relation for $||e_\lambda^{i+1}||_M$.

Let us put equations (177) and (178) into a block-matrix form

(190)
$$\begin{bmatrix} A_2 & B_2 \\ D_2 & 0 \end{bmatrix}\begin{bmatrix} \ddot{e_y}^{i+1} \\ e_\lambda^{i+1} \end{bmatrix}+\begin{bmatrix} E_2 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \dot{e_y}^{i+1} \\ e_\lambda^{i+1} \end{bmatrix}+\begin{bmatrix} C_2 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} e_y^{i+1} \\ e_\lambda^{i+1} \end{bmatrix} = \begin{bmatrix} 0 \\ -D_1\ddot{x}^i \end{bmatrix}.$$

Based on Lemma 12, the matrix $\begin{bmatrix} A_2 & B_2 \\ D_2 & 0 \end{bmatrix}$ is invertible; let $\begin{bmatrix} G_1 & G_2 \\ G_3 & G_4 \end{bmatrix}$ denote its inverse. Using the formula (158), we have

(191)
$$\begin{bmatrix} G_1 & G_2 \\ G_3 & G_4 \end{bmatrix} = \begin{bmatrix} A_2^{-1} - A_2^{-1}B_2\big(D_2A_2^{-1}B_2\big)^{-1}D_2A_2^{-1} & A_2^{-1}B_2\big(D_2A_2^{-1}B_2\big)^{-1} \\ \big(D_2A_2^{-1}B_2\big)^{-1}D_2A_2^{-1} & -\big(D_2A_2^{-1}B_2\big)^{-1} \end{bmatrix}.$$

From equation (190), it is easy to obtain

$$e_\lambda^{i+1} = -G_4D_1\ddot{e_x}^i - G_3E_2\dot{e_y}^{i+1} - G_3C_2e_y^{i+1}. \tag{192}$$

Substituting equation (179) for $\ddot{e_x}^i$, we obtain

$$e_\lambda^{i+1} = G_4D_1A_1^{-1}B_1\,e_\lambda^i + G_4D_1\Big(A_1^{-1}C_1\,e_x^i + A_1^{-1}E_1\,\dot{e_x}^i\Big) - G_3E_2\dot{e_y}^{i+1} - G_3C_2e_y^{i+1}. \tag{193}$$

Then using the results in (183) and (184), we have

$$\begin{aligned} ||e_\lambda^{i+1}||_M \leq\ & ||G_4D_1A_1^{-1}B_1||_M||e_\lambda^i||_M + \Big(\Delta t||G_4D_1A_1^{-1}E_1||_M \\ & + \frac{1}{2}\Delta t^2||G_4D_1A_1^{-1}C_1||_M\Big)||\ddot{e_x}^i||_M + \Big(\Delta t||G_3E_2||_M \\ & + \frac{1}{2}\Delta t^2||G_3C_2||_M\Big)||\ddot{e_y}^{i+1}||_M. \end{aligned} \tag{194}$$

Finally, substitution of (187) and (188) into (194) yields

$$(195)\quad ||e_\lambda^{i+1}||_M \;\le\; ||G_4D_1A_1^{-1}B_1||_M\,||e_\lambda^i||_M + \alpha_1\,\Delta t\,||e_\lambda^i||_M + \alpha_2\,\Delta t\,||e_\lambda^{i+1}||_M,$$

where

$$\begin{aligned}\alpha_1 &= \Big(||G_4D_1A_1^{-1}E_1||_M + \frac{1}{2}\Delta t||G_4D_1A_1^{-1}C_1||_M\Big)\beta_1^{-1}||A_1^{-1}B_1||_M,\\ \alpha_2 &= \Big(||G_3E_2||_M + \frac{1}{2}\Delta t||G_3C_2||_M\Big)\beta_2^{-1}||A_2^{-1}B_2||_M.\end{aligned}$$

Note that $G_4 = -(D_2A_2^{-1}B_2)^{-1}$ from equation (191). Let us denote

$$(196)\qquad J \;=\; G_4D_1A_1^{-1}B_1 \;=\; -(D_2A_2^{-1}B_2)^{-1}(D_1A_1^{-1}B_1).$$

Let us also denote

$$(197)\qquad \gamma \;=\; (1-\alpha_2\,\Delta t)^{-1}\big(\,||J||_M + \alpha_1\,\Delta t\big).$$

Then $\gamma > 0$ for sufficiently small Δt. Thus the inequality in (195) becomes

$$(198)\qquad ||e_\lambda^{i+1}||_M \;\le\; \gamma\,||e_\lambda^i||_M,$$

which holds for any i. Iterating back on i, we obtain

$$(199)\qquad ||e_\lambda^{i+1}||_M \;\le\; \gamma^i\;||e_\lambda^1||_M.$$

Step 3: We establish similar error estimates for $||e_x^{i+1}||_M$ and $||e_y^{i+1}||_M$.

By evaluating equation (192) at $i = 0$ and using the results in (184) and (188), we obtain

$$(200)\qquad ||e_\lambda^1||_M \;\le\; ||G_4D_1||_M\,||\ddot{e}_x^{\,0}||_M + \alpha_2\,\Delta t\,||e_\lambda^1||_M.$$

That is,

$$(201)\qquad ||e_\lambda^1||_M \;\le\; (1-\alpha_2\,\Delta t)^{-1}||G_4D_1||_M\,||\ddot{e}_x^{\,0}||_M,$$

where $\ddot{e}_x^{\,0}$ measures the start-up error of our iterative procedure.

Combining the results in (180), (199) and (201), we obtain

$$(202)\quad ||e_x^{i+1}||_M \;\le\; \gamma^i\cdot\frac{1}{2}\Delta t^2\,\big(1+O(\Delta t)\big)\,||A_1^{-1}B_1||_M\,||G_4D_1||_M\,||\ddot{e}_x^{\,0}||_M.$$

Meanwhile, if we combine the results in (181), (199) and (201), we obtain

$$(203)\quad ||e_y^{i+1}||_M \;\le\; \gamma^i\cdot\frac{1}{2}\Delta t^2\,\big(1+O(\Delta t)\big)\,||A_2^{-1}B_2||_M\,||G_4D_1||_M\,||\ddot{e}_x^{\,0}||_M.$$

Based on the results in (199), (202) and (203), it is clear that the iterative scheme is convergent if and only if $\gamma < 1$. Hence, Theorem 18 holds for sufficiently small Δt. □

If, instead, the iterative method (153) and (154) is used, the analysis can be conducted in the same way as above and we can easily obtain the following Corollary.

Corollary 19. *Let the assumptions (A1) and (A2-1) hold. Let also*

(204)
$$\tilde{J} = -(D_1 A_1^{-1} B_1)^{-1}(D_2 A_2^{-1} B_2).$$

If $||\tilde{J}||_M < 1$, *then for sufficiently small* $\triangle t$, *the iteration (153) and (154) converges for all* $t \in [t_n, t_{n+1}]$ *with any start-up error and any tolerance.*

4.2.2 EXAMPLES

Example 1. We first consider a one-dimensional example where x and y are both scalars:

(205)
$$\ddot{x} = 8x + 9\lambda,$$

(206)
$$\ddot{y} = y + \lambda,$$

subjected to the constraint

(207)
$$x = 2y.$$

The initial conditions are given as

(208)
$$x(0) = 0, \qquad \dot{x}(0) = 2.$$

The exact solution of this problem can be easily found as

(209)
$$x = 2\sin(t), \quad y = \sin(t), \quad \lambda = -2\sin(t).$$

Based on the formulas (173) and (204), it is straightforward to obtain

(210)
$$||J||_M = 9 > 1 \quad \text{and} \quad ||\tilde{J}||_M = \frac{1}{9} < 1.$$

Hence, the iterative method (153) and (154) should be used to ensure convergence; that is, the constraint should be placed on x. This yields the following iterative procedure at the $(i+1)$th iteration:

(211)
$$x^{i+1} = 2y^i,$$

(212)
$$\lambda^{i+1} = \frac{1}{9}(\ddot{x}^{i+1} - 8x^{i+1}),$$

and

(213)
$$\ddot{y}^{i+1} = y^{i+1} + \lambda^{i+1}.$$

We choose the computational domain as $[0, 10]$, which is equally divided into a set of intervals, each with length $\triangle t$. We consider different values of $\triangle t$ in the numerical tests. Meanwhile, we divide each interval $[t_n, t_{n+1}]$ into N uniform subintervals when implementing the iterative procedure, so that the local mesh has a grid spacing of $\triangle t/N$. The convergence tolerance is set as $\varepsilon_0 = 10^{-7}$. Using the analytical solution (209), we can easily verify the overall accuracy of the numerical approach by checking the errors at the end of the computation, $t = 10$. Table 7 shows a typical set of results when combining the iterative procedure (211)-(213) with two common ODE solvers, the forward Euler method (explicit, first-order) and the trapezoidal rule (implicit, second-order), respectively. The quantity R_x is defined as

(214)
$$R_x = \left|\left| \frac{x_{\triangle t}(10) - x(10)}{x_{\triangle t/2}(10) - x(10)} \right|\right|,$$

where $x(10)$ stands for the exact solution at $t = 10$ and $x_{\triangle t}(10)$ denotes the numerical

	R_x	R_y	R_λ
Euler	1.99	1.99	1.90
Trapezoidal rule	4.02	4.02	4.02

TABLE. 7: Example 1 - Order of accuracy for the iterative algorithm using the forward Euler method and the trapezoidal rule as the ODE solvers, respectively.

approximation to $x(10)$ with the stepsize Δt. In a similar way we define R_y and R_λ. We clearly see first-order accuracy (with Euler) and second-order accuracy (with trapezoidal rule) achieved in the end of the computation, which are consistent with the accuracy of the ODE solvers.

Figure 51 shows a set of simulation results using the iterative approach with the Euler method, for $\Delta t = 0.25$. We observe excellent agreement between the numerical solution and the exact solution. The results with the trapezoidal rule show similar pattern and are not presented here.

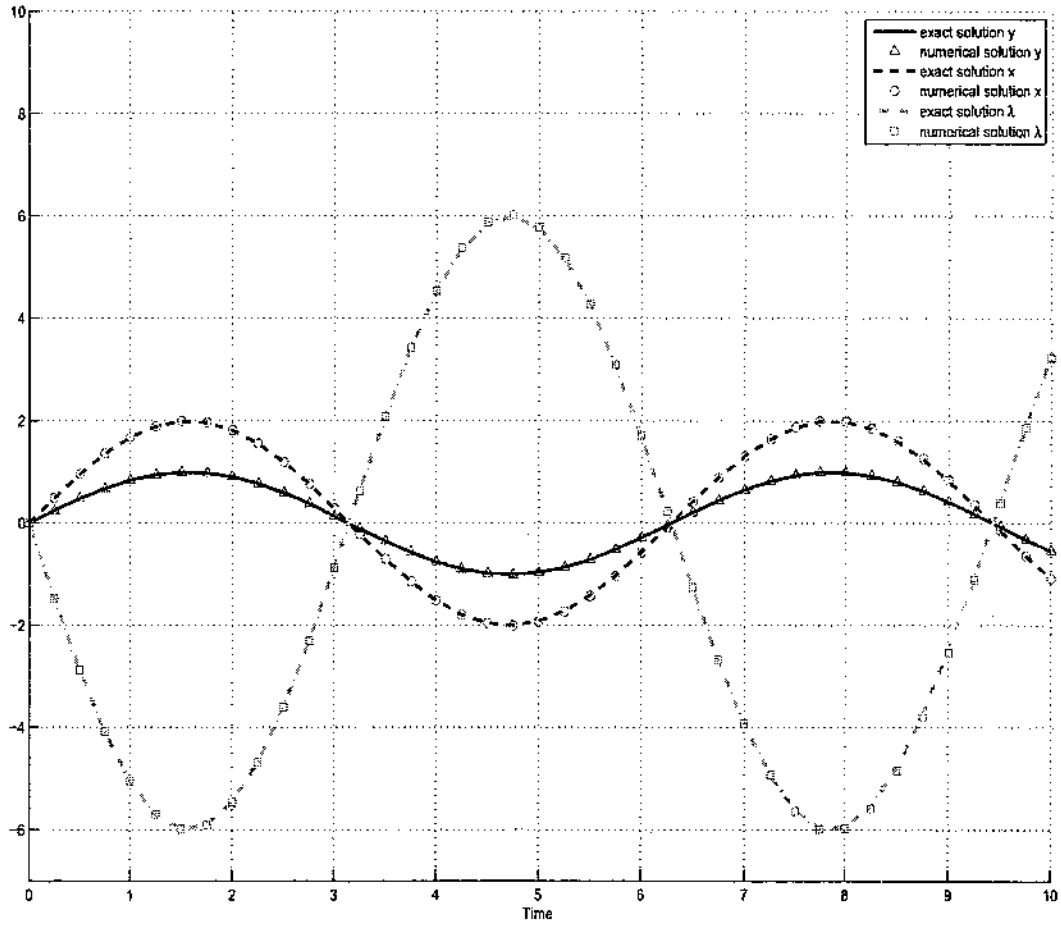

FIGURE 51: Comparison between the numerical approximation and the exact solution for Example 1.

Example 2. Next, we consider the following dynamical equations where x and y are both vectors of two components:

(215)
$$\begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} \ddot{x}_1 \\ \ddot{x}_2 \end{bmatrix} = \begin{bmatrix} -2 & -1 \\ -7 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 0 \\ 2\lambda \end{bmatrix},$$

(216)
$$\begin{bmatrix} 3 & 4 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} \ddot{y}_1 \\ \ddot{y}_2 \end{bmatrix} = \begin{bmatrix} -5 & -3 \\ -1 & -5 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} \lambda \\ 0 \end{bmatrix},$$

subjected to the constraint

(217)
$$x_2 = 2y_1, \quad \text{or} \quad \ddot{x}_2 = 2\ddot{y}_1.$$

The initial conditions are given as

(218)
$$x_1(0) = 1, \quad \dot{x}_1 = 0, \quad y_1(0) = 0, \quad \dot{y}_1(0) = 1, \quad y_2(0) = 2, \quad \text{and} \quad \dot{y}_2(0) = 0.$$

The exact solution of this problem can be found as

$$x_1 = \cos(t), \quad x_2 = 2\sin(t), \quad y_1 = \sin(t),$$

(219)
$$y_2 = 2\cos(t), \quad \lambda = 2\sin(t) - 2\cos(t).$$

Equations (215) - (217) can be assembled into a block-matrix system in the form of (151):

(220)
$$\begin{bmatrix} 2 & 1 & 0 & 0 & 0 \\ 3 & 4 & 0 & 0 & 2 \\ 0 & 0 & 3 & 4 & -1 \\ 0 & 0 & 1 & 5 & 0 \\ 0 & 1 & -2 & 0 & 0 \end{bmatrix} \begin{bmatrix} \ddot{x}_1 \\ \ddot{x}_2 \\ \ddot{y}_1 \\ \ddot{y}_2 \\ \lambda \end{bmatrix} - \begin{bmatrix} -2 & -1 & 0 & 0 & 0 \\ -7 & -2 & 0 & 0 & 0 \\ 0 & 0 & -5 & -3 & 0 \\ 0 & 0 & -1 & -5 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ y_1 \\ y_2 \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

If the local iterative scheme (156) and (157) is applied, we obtain, at the $(i+1)$th iteration,

(221)
$$\begin{bmatrix} 3 & 4 & -1 \\ 1 & 5 & 0 \\ -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \ddot{y}_1^{i+1} \\ \ddot{y}_2^{i+1} \\ \lambda^{i+1} \end{bmatrix} + \begin{bmatrix} 5 & 3 & 0 \\ 1 & 5 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} y_1^{i+1} \\ y_2^{i+1} \\ \lambda^{i+1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -\ddot{x}_2^i \end{bmatrix},$$

and

(222)
$$\begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} \ddot{x}_1^{i+1} \\ \ddot{x}_2^{i+1} \end{bmatrix} = \begin{bmatrix} -2 & -1 \\ -7 & -2 \end{bmatrix} \begin{bmatrix} x_1^{i+1} \\ x_2^{i+1} \end{bmatrix} - \begin{bmatrix} 0 \\ 2\lambda^{i+1} \end{bmatrix}.$$

With some algebra, equation (221) may be further manipulated to decouple the computation for y^{i+1} and λ^{i+1} so as to simplify the solution procedure. Using equation (173), we obtain

$$J = -(D_2 A_2^{-1} B_2)^{-1}(D_1 A_1^{-1} B_1) = \frac{11}{10} \cdot \frac{4}{5} = 0.88 < 1.$$

Based on Theorem 13, the local iterative procedure (156) and (157) will converge for this problem.

If, instead, the local iterative scheme (153) and (154) is applied, we have

(223)
$$\begin{bmatrix} 2 & 1 & 0 \\ 3 & 4 & 2 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \ddot{x}_1^{i+1} \\ \ddot{x}_2^{i+1} \\ \lambda^{i+1} \end{bmatrix} + \begin{bmatrix} 2 & 1 & 0 \\ 7 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1^{i+1} \\ x_2^{i+1} \\ \lambda^{i+1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 2\ddot{y}_1^i \end{bmatrix},$$

and

(224)
$$\begin{bmatrix} 3 & 4 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} \ddot{y}_1^{i+1} \\ \ddot{y}_2^{i+1} \end{bmatrix} = \begin{bmatrix} -5 & -3 \\ -1 & -5 \end{bmatrix} \begin{bmatrix} y_1^{i+1} \\ y_2^{i+1} \end{bmatrix} + \begin{bmatrix} \lambda^{i+1} \\ 0 \end{bmatrix}.$$

Using equation (204), we obtain

$$\tilde{J} = -(D_1 A_1^{-1} B_1)^{-1}(D_2 A_2^{-1} B_2) = \frac{5}{4} \cdot \frac{10}{11} > 1.$$

Hence, the iterative scheme (153) and (154) will not converge.

	R_x	R_y	R_λ
Euler	2.03	2.04	2.07
Trapezoidal rule	4.04	4.00	4.01

TABLE. 8: Example 2 - Order of accuracy for the iterative algorithm using the forward Euler method and the trapezoidal rule as the ODE solvers, respectively.

Below we present numerical results to confirm these predictions. We again set the computational domain as $[0, 10]$, and divide each interval $[t_n, t_{n+1}]$ into equally spaced subintervals when implementing the iterative procedure. The convergence tolerance is set as $\varepsilon_0 = 10^{-7}$. We focus on the iterative scheme (156) and (157), and employ again the forward Euler method and trapezoidal rule as the ODE solvers.

Figure 52 shows the simulation results using the iterative approach with the trapezoidal rule, for $\triangle t = 0.25$. We again observe that the numerical solution closely matches the exact solution. The results with the Euler method show similar pattern (not presented).

Next, we verify the overall accuracy of the numerical simulation by checking the errors at the end of the computation, $t = 10$. Table 8 shows the values of R_x, R_y and R_λ for the Euler method and trapezoidal rule, respectively. We clearly observe first-order accuracy for the Euler and second-order accuracy for the trapezoidal rule.

In addition, we have applied the iterative scheme (153) and (154) to compute this problem. We find the numerical solution does not converge no matter what ODE solvers are applied and how small $\triangle t$ is. This is consistent with the analytical prediction.

Example 3. Now we consider an example where the coefficient matrices and vectors are functions of t:

$$\begin{bmatrix} 1 & 0 \\ -2t & t+1 \end{bmatrix} \begin{bmatrix} \ddot{x}_1 \\ \ddot{x}_2 \end{bmatrix} + \begin{bmatrix} -2\sin(t) & \cos(t) \\ 2\cos(t) & \sin(t) \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} + \begin{bmatrix} -1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

(225)

$$= \begin{bmatrix} 2 \\ 2(\cos(t) - \sin(t)) \end{bmatrix} - \lambda \begin{bmatrix} -1 \\ t-1 \end{bmatrix},$$

$$\begin{bmatrix} 5 & -3 \\ 10 & 6 \end{bmatrix} \begin{bmatrix} \ddot{y}_1 \\ \ddot{y}_2 \end{bmatrix} + \begin{bmatrix} 2\cos(t) & -\sin(t) \\ 2\sin(t) & \cos(t) \end{bmatrix} \begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \end{bmatrix} + \begin{bmatrix} 3 & -2 \\ 12 & 5 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

(226)

$$= \begin{bmatrix} 2 \\ 0 \end{bmatrix} - \lambda \begin{bmatrix} 1 \\ -1 \end{bmatrix},$$

subjected to the constraint

(227)

$$x_2 = 2y_1, \quad \text{or} \quad \ddot{x}_2 = 2\ddot{y}_1.$$

The initial conditions of this problem are given as

(228)

$$x_1(0) = 1, \quad \dot{x}_1(0) = 0, \quad y_1(0) = 0, \quad \dot{y}_1(0) = 1, \quad y_2(0) = 0, \quad \text{and} \quad \dot{y}_2(0) = 0.$$

The exact solution of this problem is the same as given in (219).

Similar to Example 2, the local iterative schemes (153) and (154) and (156) and (157) can be applied to this problem (note the assumptions (A1), (A2-1) and (A2-2) are all satisfied). From equation (173), we have

$$J = -(D_2 A_2^{-1} B_2)^{-1}(D_1 A_1^{-1} B_1) = -10.$$

Thus, the iterative method (156) and (157) will not converge for this problem. Instead, from equation (204), we obtain

$$\tilde{J} = -(D_1 A_1^{-1} B_1)^{-1}(D_2 A_2^{-1} B_2) = -\frac{1}{10},$$

which indicates that the iterative method (153) and (154) will converge for this problem.

Based on the iterative scheme (153) and (154), the following two major steps are performed for the $(i+1)$th iteration:

Step 1. Solve for x_1^{i+1}, x_2^{i+1} and λ^{i+1} from the system

$$\begin{bmatrix} 1 & 0 & -1 \\ -2t & t+1 & t-1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \ddot{x}_1^{i+1} \\ \ddot{x}_2^{i+1} \\ \lambda^{i+1} \end{bmatrix} + \begin{bmatrix} -2\sin(t) & \cos(t) & 0 \\ 2\cos(t) & \sin(t) & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1^{i+1} \\ \dot{x}_2^{i+1} \\ \lambda^{i+1} \end{bmatrix}$$

(229)

$$+ \begin{bmatrix} -1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1^{i+1} \\ x_2^{i+1} \\ \lambda^{i+1} \end{bmatrix} = \begin{bmatrix} 2 \\ 2(\cos(t) - \sin(t)) \\ 2y_1^i \end{bmatrix}.$$

With some algebra, the system (229) can be reduced to three separate equations:

(230)

$$x_2^{i+1} = 2y_1^i,$$

$$-(t+1)\ddot{x}_1^{i+1} + [2\cos(t) - 2(t-1)\sin(t)]\dot{x}_1^{i+1} - (t-1)x_1^{i+1} = - (t+1)\ddot{x}_2^{i+1} - [(t-1)\cos(t) + \sin(t)]\dot{x}_2^{i+1}$$

(231)

$$- tx_2^{i+1} + 2(t-1) + 2(\cos(t) - \sin(t)),$$

for x_1^{i+1} and x_2^{i+1}, respectively, and

(232)

$$\lambda^{i+1} = \ddot{x}_1^{i+1} = 2\sin(t)\dot{x}_1^{i+1} + \cos(t)\dot{x}_2^{i+1} - x_1^{i+1} + x_2^{i+1} - 2,$$

for λ^{i+1}.

Step 2. Solve for y_1^{i+1} and y_2^{i+1} from the system

$$\begin{bmatrix} 5 & -3 \\ 10 & 6 \end{bmatrix} \begin{bmatrix} \ddot{y}_1^{i+1} \\ \ddot{y}_2^{i+1} \end{bmatrix} + \begin{bmatrix} 2\cos(t) & -\sin(t) \\ 2\sin(t) & \cos(t) \end{bmatrix} \begin{bmatrix} \dot{y}_1^{i+1} \\ \dot{y}_2^{i+1} \end{bmatrix}$$

(233)

$$+ \begin{bmatrix} 3 & -2 \\ 12 & 5 \end{bmatrix} \begin{bmatrix} y_1^{i+1} \\ y_2^{i+1} \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix} - \begin{bmatrix} \lambda^{i+1} \\ -\lambda^{i+1} \end{bmatrix}.$$

Figure 53 shows the simulation result using this iterative approach with the trapezoidal rule, for $\Delta t = 0.05$. The pattern is similar to that in Figure 52; i.e., the numerical solution closely follows the exact solution. Also note that Examples 2 and 3 have the same exact solution. Meanwhile, we observe similar accuracy (i.e., first order with the Euler and second order with the trapezoidal rule) for the numerical solution(see Table 9).

	R_x	R_y	R_λ
Euler	2.01	2.03	1.99
Trapezoidal rule	4.00	3.83	4.00

TABLE. 9: Example 3 - Order of accuracy for the iterative algorithm using the forward Euler method and the trapezoidal rule as the ODE solvers, respectively.

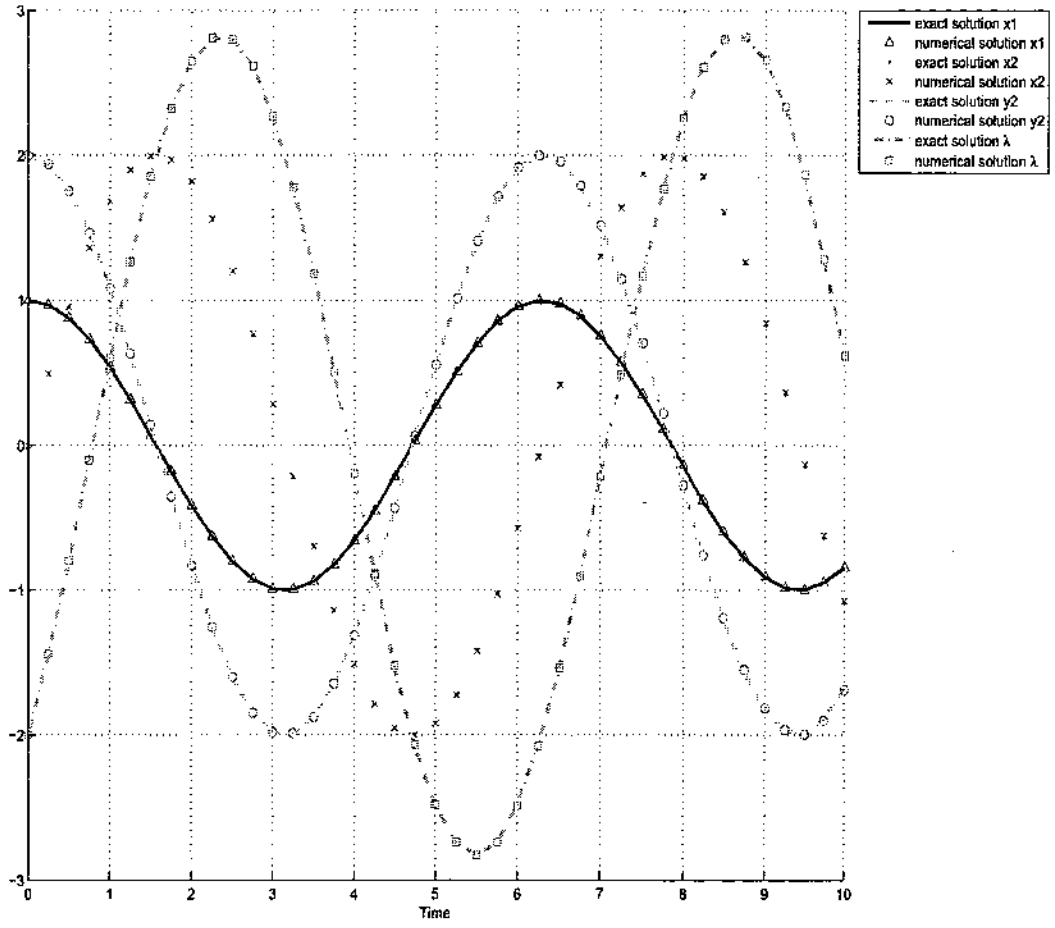

FIGURE 52: Comparison between the numerical approximation and the exact solution for Example 2.

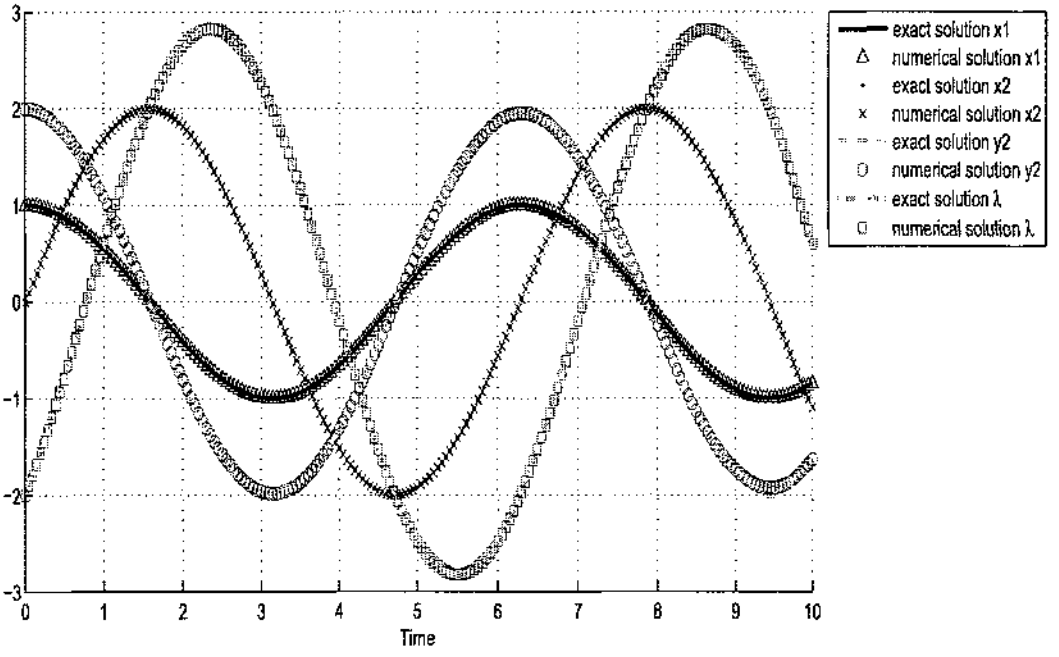

FIGURE 53: Comparison between the numerical approximation and the exact solution for Example 3.

Example 4. Now we consider a linearized plane stress problem. Figure 54 shows a solid arch whose left and right edges are completely constrained on the walls. The arch is discretized into a finite element model with 6 nodes and 3 Constant Strain Triangular (CST) elements. The whole structure can be viewed as two separate substructures joined at Node 1: the left substructure is made of two (triangular) elements and the right substructure has one element.

The equations of motion for the left and right substructures, respectively, are given by

$$M_L\,\ddot{x}_L + J_L(t)\,x_L = 0, \tag{234}$$

and

$$M_R\,\ddot{x}_R + J_R(t)\,x_R = 0, \tag{235}$$

where M_L and M_R are the (constant) mass matrices, and $J_L(t)$ and $J_R(t)$ are the stiffness matrices which are assumed to be time-dependent. The displacement vector x_L of equation (234) consists of the displacements at Nodes 1 and 2 as

$$x_L = \left(u_1^L, v_1^L, u_2^L, v_2^L\right)^T.$$

On the other hand, the displacement vector x_R of equation (235) consists of the displacements at Node 1 of the right substructure,

$$x_R = \left(u_1^R, v_1^R\right)^T.$$

The stiffness matrix $J_L(t)$ and the mass matrix M_L of equation (234) are given by

$$J_L(t) = \begin{bmatrix} 0.3 + 0.1\cos t & -0.04 & 0.04 & 0.2\cos t \\ -0.04 & 0.6 & -0.4 & -0.02 \\ 0.04 & -0.4 & 0.2 & 0.02 \\ 0.2\cos t & -0.02 & 0.02 & 0.15 + 0.44\cos t \end{bmatrix},$$

and

$$M_L = \begin{bmatrix} 0.04 & 0 & 0.02 & 0 \\ 0 & 0.04 & 0 & 0.02 \\ 0.02 & 0 & 0.15 & 0 \\ 0 & 0.02 & 0 & 0.15 \end{bmatrix}.$$

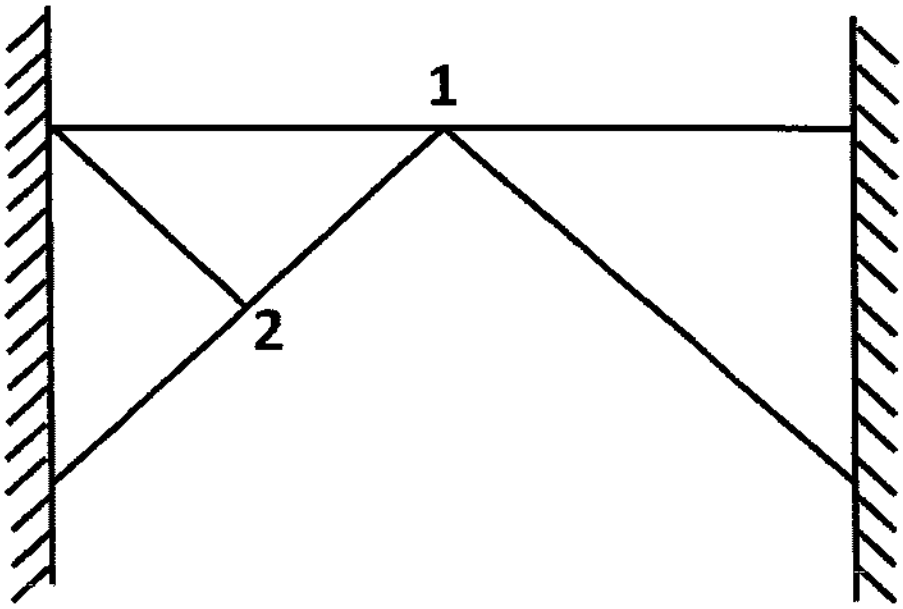

FIGURE 54: Structural configurations in Example 4.

Meanwhile, the matrices $J_R(t)$ and M_R of equation (235) are given as

$$J_R(t) = \begin{bmatrix} 0.1\cos t & 0 \\ 0 & 0.2 \end{bmatrix} \quad \text{and} \quad M_R = \begin{bmatrix} 0.26 & 0 \\ 0 & 0.26 \end{bmatrix}.$$

The joint conditions between the left and the right substructures at Node 1 can be represented by the following constraint which has to be maintained at all times:

$$u_1^L = u_1^R \quad \text{and} \quad v_1^L = v_1^R. \tag{236}$$

The initial displacements are set as all zero:

$$x_L(0) = \big(0\,,\ 0\,,\ 0\,,\ 0\big)^T \quad \text{and} \quad x_R(0) = \big(0\,,\ 0\big)^T, \tag{237}$$

whereas the initial velocities are given by

$$\dot{x}_L(0) = \big(1\,,\ -2\,,\ 2\,,\ -1\big)^T \quad \text{and} \quad \dot{x}_R(0) = \big(1\,,\ -2\big)^T. \tag{238}$$

The exact solution to this problem is

$$x_L(t) = \big(\sin t\,,\ -\sin(2t)\,,\ \sin(2t)\,,\ -\sin t\big)^T \quad \text{and} \quad x_R(t) = \big(\sin t\,,\ -\sin(2t)\big)^T,$$

which can be used to measure the accuracy of our numerical solution.

Based on the Theorem of Lagrange Multipliers [25], we obtain the following equations to describe the constrained dynamics of the entire structural system:

$$M_L\,\ddot{x}_L + J_L(t)\,x_L = -B_L\,\lambda, \tag{239}$$

$$M_R\,\ddot{x}_R + J_R(t)\,x_R = -B_R\,\lambda, \tag{240}$$

$$\text{and} \quad D_L\,x_L + D_R\,x_R = 0. \tag{241}$$

Here λ is the algebraic unknown (or, the Lagrange multiplier in this case) and $\lambda = [\lambda_1,\ \lambda_2]^T$. Meanwhile,

$$D_L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \qquad D_R = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix},$$

and $B_L = D_L^T$ and $B_R = D_R^T$.

Equations (239 - 241) constitute a constrained dynamical problem, and we will then apply our proposed iterative approach to solve this problem. Before applying the numerical algorithm, we can easily evaluate, based on equations (173) and (204), that

$$||J|| = ||(D_R M_R^{-1} B_R)^{-1}(D_L M_L^{-1} B_L)|| \approx 6.96 > 1,$$

and

$$||\tilde{J}|| = ||(D_L M_L^{-1} B_L)^{-1}(D_R M_R^{-1} B_R)|| \approx 0.14 < 1.$$

Hence, the local iterative scheme (153) and (154) should be used to ensure convergence. Specifically, we perform the following steps to obtain the solution at the $(i+1)$th iteration:

Step 1. First evaluate part of x_L^{i+1} (i.e., $(u_1^L)^{i+1}$ and $(v_1^L)^{i+1}$) based on (241):

$$\left(u_1^L\right)^{i+1} = \left(u_1^R\right)^i, \qquad \text{and} \quad \left(v_1^L\right)^{i+1} = \left(v_1^R\right)^i. \tag{242}$$

Then compute the other components of x_L^{i+1} (i.e., $(u_2^L)^{i+1}$ and $(v_2^L)^{i+1}$) and λ^{i+1} by

$$M_L \begin{bmatrix} (\ddot{u}_1^L)^{i+1} \\ (\ddot{v}_1^L)^{i+1} \\ (\ddot{u}_2^L)^{i+1} \\ (\ddot{v}_2^L)^{i+1} \end{bmatrix} + J_L(t) \begin{bmatrix} (u_1^L)^{i+1} \\ (v_1^L)^{i+1} \\ (u_2^L)^{i+1} \\ (v_2^L)^{i+1} \end{bmatrix} = - \begin{bmatrix} \lambda_1^{i+1} \\ \lambda_2^{i+1} \\ 0 \\ 0 \end{bmatrix}. \tag{243}$$

This systems consists of four scalar equations; the last two equations are used to calculate $(u_2^L)^{i+1}$ and $(v_2^L)^{i+1}$, then the first two equations are applied to evaluate λ_1^{i+1} and λ_2^{i+1}.

Calculate the backward errors ε_λ and ε_x.

Step 2. Compute x_R^{i+1} by solving

$$M_R\,\ddot{x}_R^{i+1} + J_R(t)\,x_R^{i+1} = \lambda^{i+1}. \tag{244}$$

Also calculate the backward error ε_y.

Step 3. Check the convergence.

Applying this iterative procedure to the system (239 - 241), we observed similar patterns of convergence and accuracy as those demonstrated in the previous two examples. A typical set of numerical results are presented in Figures 55 - 58 which show excellent agreement between the numerical solution and the exact solution on the interval $[0,\ 10]$.

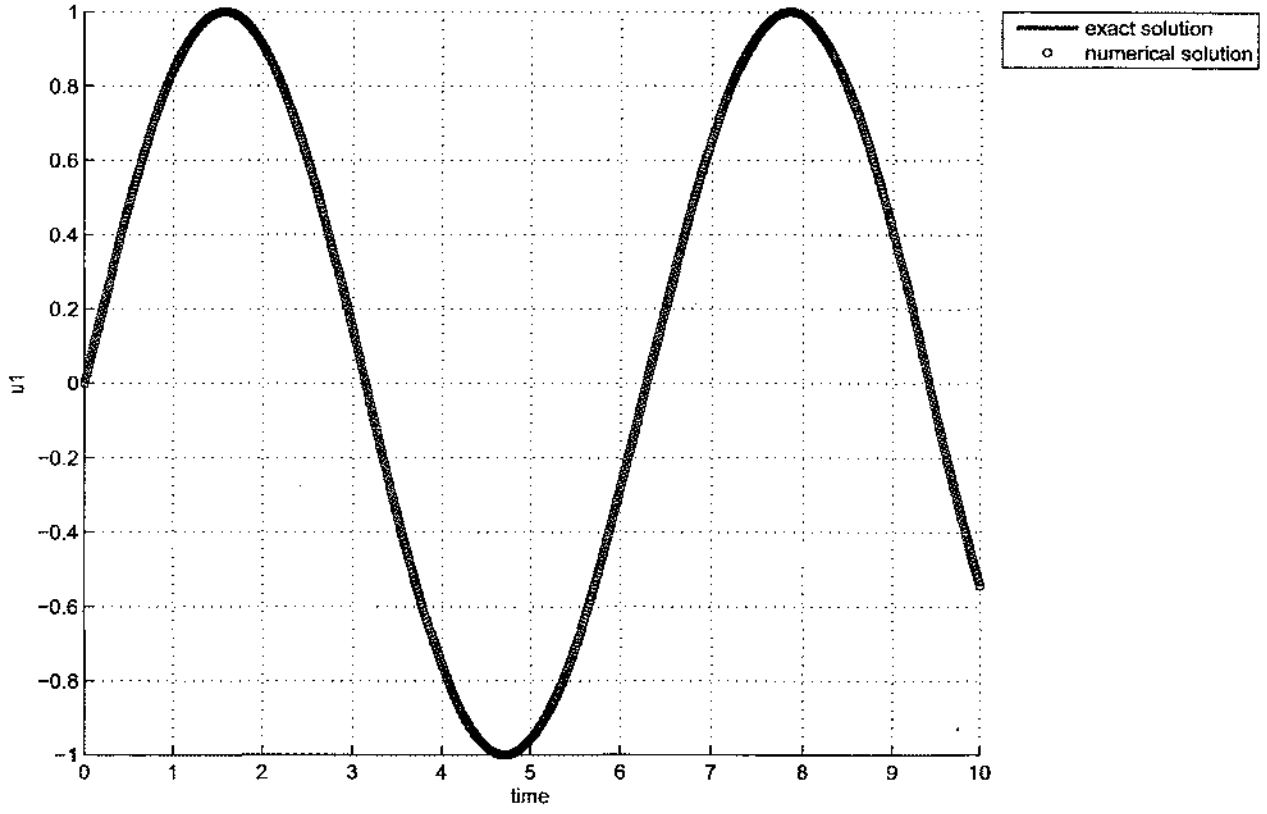

FIGURE 55: Comparison between the numerical and the exact solutions for Example 4 for $u_1^L = u_1^R$.

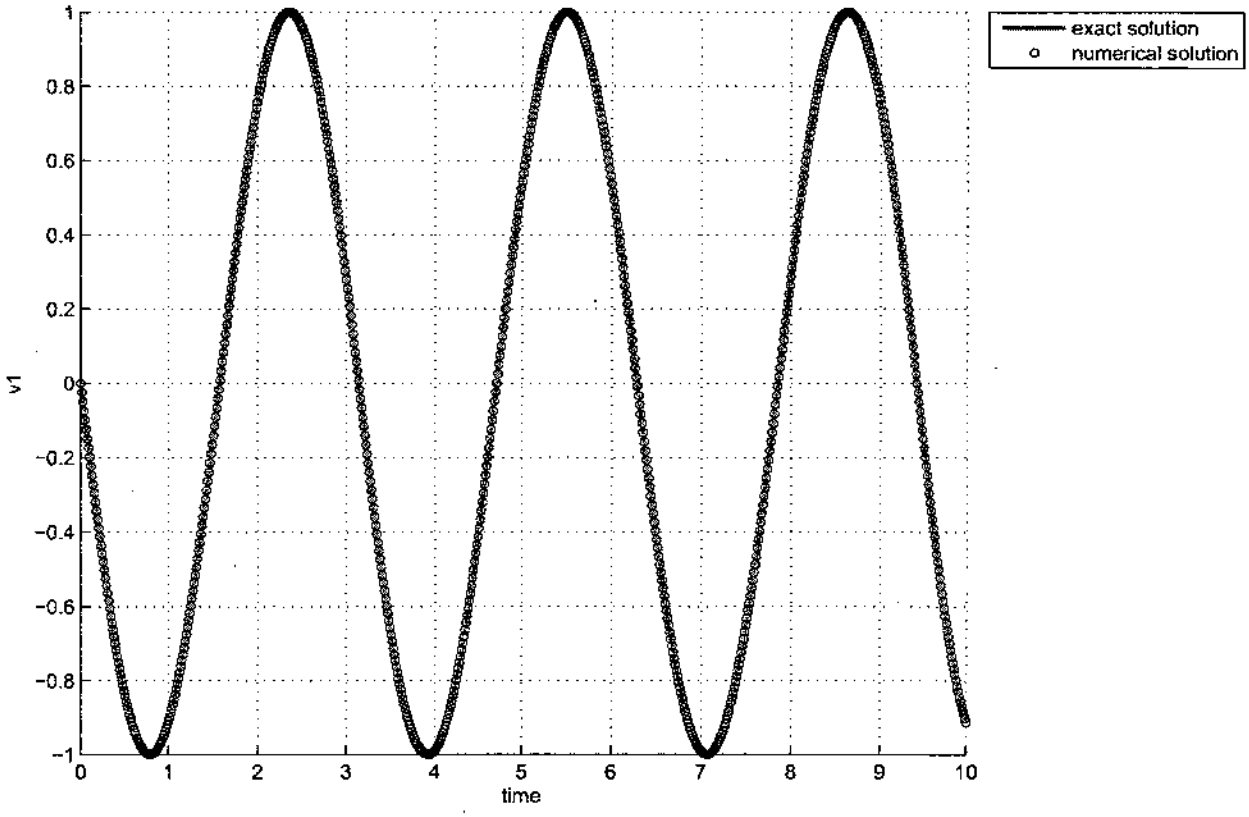

FIGURE 56: Comparison between the numerical and the exact solutions for Example 4 for $v_1^L = v_1^R$.

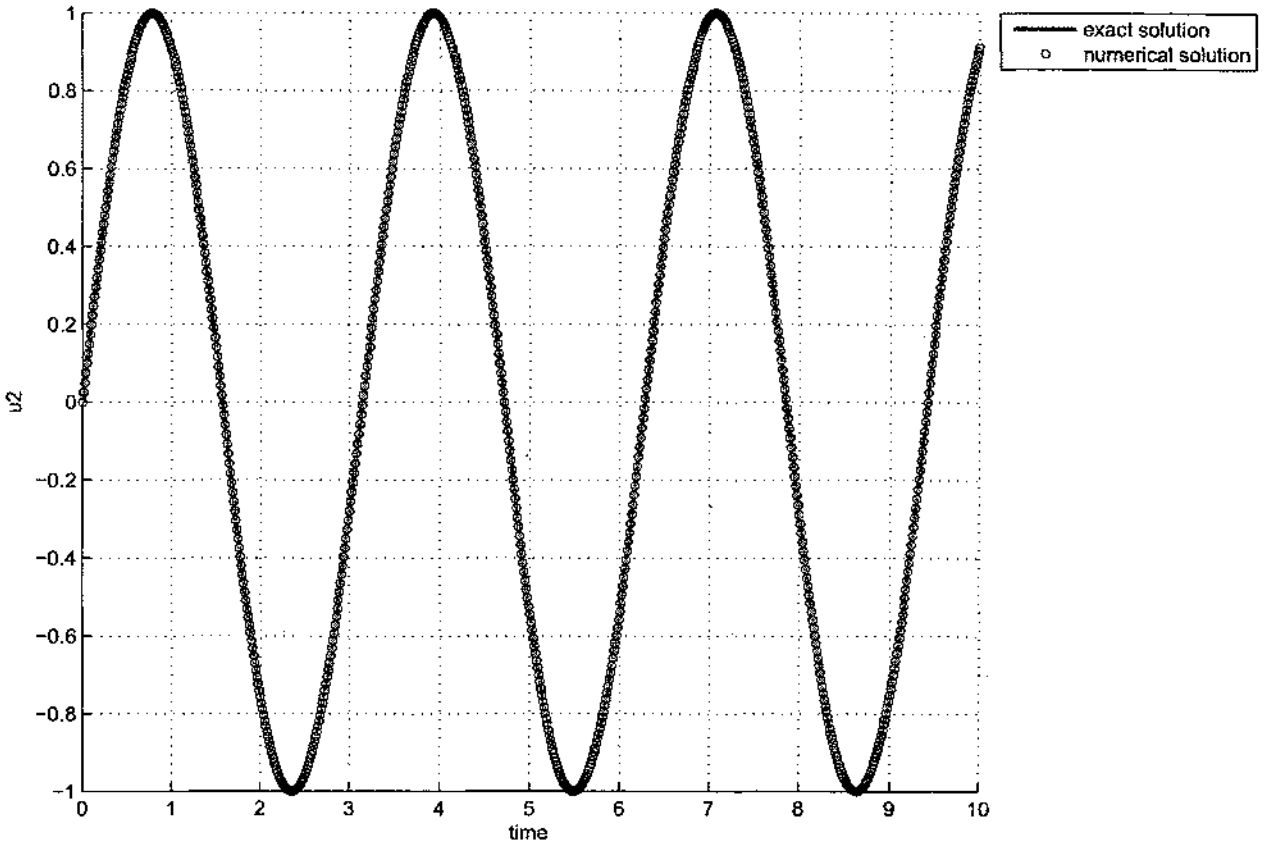

FIGURE 57: Comparison between the numerical and the exact solutions for Example 4 for u_2^R.

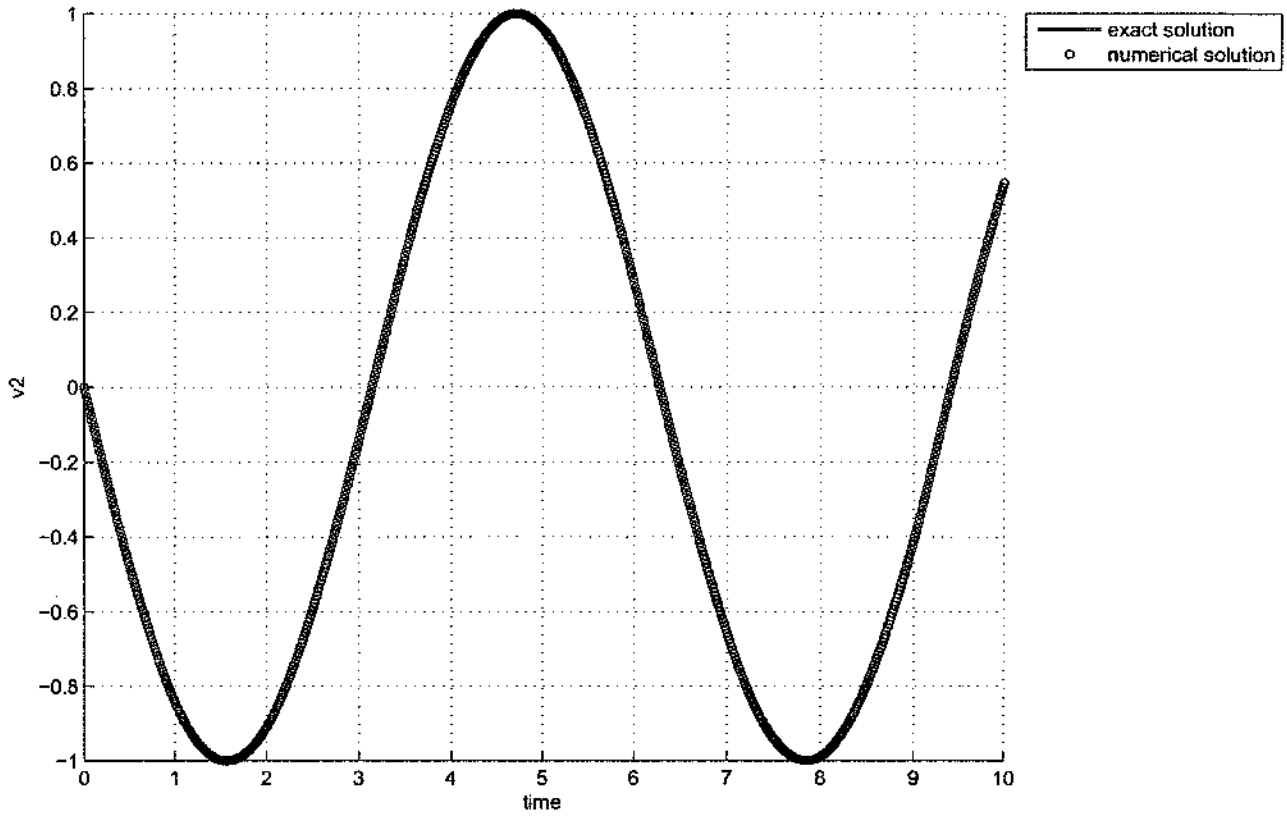

FIGURE 58: Comparison between the numerical and the exact solutions for Example 4 for v_2^L.

CHAPTER 5

CONCLUSIONS

In this dissertation, mathematical modeling of cholera dynamics with optimal controls and new iterative algorithms for solving optimal control problems have been investigated. In the beginning of Chapter 3, we extended the original model proposed by Mukandavire *et al.* to investigate optimal control strategies. We modified the model by adding three types of controls: vaccination, therapeutic treatment, and water sanitation. We then followed the Forward-Backward Sweep Method to find the optimal control solutions. Numerical simulations showed that with the incorporation of any of the three controls, the level of infection was reduced compared to the original model without controls. Specifically, with all three controls, such reduction is significant. An interesting result showed that the improvement has been achieved even with vaccination only.

Results from this simple model motivated our interest for more careful exploration on cholera dynamics with controls. By using more realistic assumptions on control parameters and by introducing an additional class of vaccinated individuals, we formulated a refined optimal control model for cholera. The numerical simulations showed similar patterns, i.e., the number of infections has been significantly reduced when all three controls were applied. This confirms our observation that multiple intervention methods should be used whenever possible.

We have seen from our numerical simulations that vaccination is a very effective control measure. However, with the cost of vaccines and their limited availability, a mass vaccination may not always be possible in reality. This led us to propose a new optimal control model to investigate optimal times during epidemics for deploying cholera vaccines that best balance the gains and costs of vaccination. Our study showed that if one has access to sufficient resources of vaccines, a mass vaccination should be deployed immediately after an outbreak. In reality, however, this may not always be feasible, and it suggests that vaccination should be deployed strategically in combination with other types of control methods.

The complexity of cholera dynamics is yet to be better understood. We continued our study to an age-structure model to investigate the impact of different ages on

cholera dynamics and the corresponding control strategies. The numerical simulations demonstrated that ages could affect the infections. We expect more results on this topic to be reported in near future.

We closed Chapter 3 with our interest in a multigroup cholera modeling. We proposed a first step to a two-group cholera model. Preliminary numerical results showed that the higher disease transmission from group 2 to group 1 results in higher levels of infections and pathogen concentration in group 1. Correspondingly, that leads to longer duration of vaccination in group 1.

In Chapter 4, we turned to the study on iterative algorithms for optimal control problems. We introduced a locally refined Forward-Backward Sweeping Method to improve the convergence of the original FBSM. Error analysis and numerical simulations were also presented. We then extended the idea of locally refined iterative procedure to the study for a class of constrained dynamical problems involving second-order linear differential-algebraic equations. Examples and numerical results were presented to validate our analysis. We expect applications of such iterative algorithms to a wider range of scientific and engineering problems.

All numerical simulations were conducted by Matlab program (license number : 347959) on a personal computer owned by Old Dominion University running Windows 7 Enterprise, Intel(R) Core(TM)2 Duo E7300 2.66 GHz processor, 4.00 GB memory, and 64-bit Operating System.

All parameter values in this dissertation were taken from published research papers and they were cited in this work.

REFERENCES

1. S. Lenhart and J. T. Workman, *Optimal Control Applied to Biological Models*,Chapman and Hall/CRC, London, (2007).

2. W. O. Kermack, A. G. McKendrick, *A Contribution to the Mathematical Theory of Epidemics*, Proceedings of the Royal Society, **115**, No.772, (1927), 700-721.

3. V. Capasso, S. L. Paveri-Fontana, *A mathematical model for the 1973 cholera epidemic in the European mediteranean region*, Revue d épidémoligié et de santé Publiqué, **2**, No.27, (1979), 121-132.

4. R. I. Joh, H. Wang, J. S. Weiss, *Dynamics of indirectly transmitted infectious disease with immunological threshold*, Bulletin of Mathematical Biology, **71**, (2009), 845-862.

5. T.Burden, J. Ernstberger, and K. R. Fister,*Optimal Control Applied to Immunotherapy*, Discrete and Continuous Dynamical Systems-Series B, **4**, No.1, (2004), 135-146.

6. D. Kirschner and J. C. Panetta,*Modeling immunotherapy of the tumor - immune interaction*, J. Math. Biol., **37**, (1998), 235-252. 585 (1998).

7. L. D. Berkovitz, *Optimal Control Theory*, Springer-Verlag, New York, (1974).

8. Center for Disease Control and Prevention web page: www.cdc.gov.

9. E. J. Nelson, J. B. Harris, J. G. Morris, S. B. Calderwood, *Cholera transmission: the host, pathogen and bacteriophage dynamic*, A. Camilli, Nat. Rev. Microboil. **10**, (2009), 693-702.

10. D. M. Hartley, Jr. J. G. Morris, D. L. Smith,*Hyperinfectivity: A Critical Element in the Ability of V. cholerae to Cause Epidemics?*, PLoS Med. **3**, (2005), e7.

11. A. Alam, R. C. Larocque, J. B. Harris, et al.,*Hyperinfectivity of human-passaged Vibrio cholerae can be modeled by growth in the infant mouse*,Infection and Immunity, **73**, No.10, (2005), 6674-6679.

12. D. S. Merrell, S. M. Butler, F. Qadri, et al., *Host-induced epidemic spread of the cholera bacterium*, Nature, **417**, (2002), 642-645.

13. Z. Mukandavire, S. Liao, J. Wang, H. Gaff, D. L. Smith, J. G. Morris,*Estimating the reproductive numbers for the 2008-2009 cholera outbreaks in Zimbabwe*, Proceedings of the National Academy of Sciences, **108**, (2011), 8767-8772.

14. World Health Organization web page: www.who.org.

15. K.A. Date, A. Vicari, T.B. Hyde et al.,*Consideration for oral cholera vaccine use during outbreak after earthquake in Haiti, 2010-2011*, Emerging Infectious Diseases, **17**, No.11, (2011), 2105-2112.

16. C. T. Codeço, *Endemic and epidemic dynamics of cholera: the role of the aquatic reservoir*, BMC Infectious Diseases, **1**, (2001), 1-14.

17. E. Asano, L. J. Gross, S. Lenhart, L. A. Real, *Optimal control of vaccine distribution in a rabies metapopulation model*, Mathematical Biosciences and Engineering, **5**, (2008), 219-238.

18. B. Buonomo, *A simple analysis of vaccination strategies for Rubella*, Mathematical Biosciences and Engineering, **8**, (2011), 677-687.

19. J. Tian, J. Wang, *Global stability for cholera epidemic models*, Mathematical Biosciences, **232**, (2011), 31-41.

20. R. L. M. Neilan, E. Schaefer, H. Gaff, K. R. Fister, S. Lenhart, *Modeling optimal intervention strategies for cholera*, Bulletin of Mathematical Biology, **72**, (2010), 2004-2018.

21. L. S. Pontryagin, V. G. Boltyanski, R. V. Gamkrelize, E. F. Mishchenko, *The mathematical theory of optimal processes*, Wiley, New York, (1967).

22. W. H. Fleming, R. W. Rishel, *Deterministic and stochastic optimal control*, Springer, New York, (1975).

23. H. D. Gaff, E. Schaefer, S. Lenhart, *Use of optimal control model to predict treatment time for managing tick-borne disease*, Journal of Biological Dynamics, **5**, (2011), 517-530.

24. E. Jung, S. Iwami, Y. Takeuchi, T. -C. Jo, *Optimal Control strategy for prevention of avian influenza pandemic*, Journal of Theoretical Biology, **260**, (2009), 220-229.

25. A. D. Belegundu, T. R. Chandrupatla *Optimization Concepts and Applications in Engineering*, Cambridge University Press, New York, (1999).

26. A. Alexanderian, M. K. Gobbert, K. R. Fister, H. Gaff, S. Lenhart, El Schaefer, *An age-structured model for the spread of epidemic cholera: Analysis and simulation*, Nonlinear Analysis: Real World Applications, **12**, (2011), 3483-3498.

27. E. J. Haug *Intermediate Dynamics*, Prentice Hall, (1992).

28. J. H. Tien, D. J. D. Earn, *Multiple transmission pathways and disease dynamics in a water borne pathogen model*, Bulletin of Mathematical Biology, **72**, (2010), 1502-1533.

29. D. A. Knoll, D. E. Keyes, *Jacobian-free Newton-Krylov methods: a survey of approaches and applications*, Journal of Computational Physics, **193**, (1996), 2303-2317.

30. K. C. Park, C. A. Felippa, *A Variational principle for the formulations of partitioned structural systems*, International Journal for Numerical Methods in Engineering, **47**, No.1-3, (2000), 395-418.

31. R. Unger, M. C. Haupt, P. Horst, *Application of Lagrange Multipliers for Coupled Problems in Fluid and Structural Interactions*, Computers and Structures, **85**, (2007), 796-809.

32. Y. Xu, *A matrix free Newton/Krylow method for coupling multi-physics subsystems*, Ph.D. Dissertation, Purdue University, (2004).

33. R. Glowinski, T. W, Pan, T. I. Hesla, D. D. Joseph, J. Periauz, *A distributed Lagrange multiplier/fictitious domain method for the simulation of flow around moving rigid bodies: application to particular flow*, Computer Methods in Applied Mechanics and Engineering, **184**, (2000), 241-267.

34. A. R. Gourlay, *A note on trapezoidal methods for the solution of initial value problems*, Mathematics of Computation, **24**, (1970), 629-633.

35. A. Prothero, A. Robinson, *On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations*, Mathematics of Computation, **28**, (1974), 145-162.

36. P. Kunkel, V. Mehrmann, *Differential-Algebraic Equations: Analysis and Numerical Solution*, European Mathematical Society Publishing House, Switzerland, (2006).

37. U. M. Ascher, L. R. Petzold, *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*, Society for Industrial and Applied Mathematics, Philadelphia, (1998).

38. K. E. Brenan, S. L. Campbell, L. R. Petzold, *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, Elsevier Science Publishing, North-Holland, (1989).

39. J. R. Cash, *Efficient numerical methods for the solution of stiff initial-value problems and differential algebraic equations*, Proceedings of the Royal Society of London A, **459**, (2003), 797-815.

40. E. Hairer, G. Wanner, *Solving Ordinary Differential Equations II: Stiff and Differential Algebraic Problems*, Springer, (1996).

41. G. H. Golub, C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, (1996).

42. T. Sauer, *Numerical Analysis*, Person Addison-Wesley, (2004).

43. S. H. Strogatz, *Nonlinear Dynamics and Chaos: with Applications to Physics, Biology, Chemistry, and Engineering*, Westview Press, (2001).

44. E. A. Coddington, N. Levinson, *Theory of Ordinary Differential Equations*, 6th edition, Tata McGraw-Hill, New Dehli, (1982).

45. C. Moler, C. Van Loan, *Nineteen dubios ways to compute the exponential of a matrix, twenty-five years later*, SIAM Review, **45**, (2003), 1-46.

46. R. Scardovelli, S. Zaleski, *Direct numerical simulation of free surface and interfacial flow*, Annual Review of Fluid Mechanics, **31**, (1999), 567-603.

47. J. Wang, G. Baker, *A numerical algorithm for viscous incompressible interfacial flows*, Journal of Computational Physics, **228**, (2009), 5470-5489.

48. J. Wang, A. Layton, *Numerical simulations of fiber sedimentation in Navier-Stokes flows*, Communications in Computational Physics, **5**, No. 1, (2009), 61-83.

49. T. H. Gronwall, *Note on the derivative with respect to a parameter of the solutions of a system of differential equations*, Annals of Mathematics, **20**, (1919), 292-296.

50. J. Wang, C. Modnak, G. Hou, *Convergence analysis of an iterative algorithm for a class of constrained dynamic problems*,Applied Mathematics and Computation, **219**, (2012), 1200-1221.

51. S. Liao, J. Wang, *Stability analysis and application of a mathematical cholera model*, Mathematical Biosciences and Engineering, No. 8, (2011), 733-752.

52. K. A. Date, A. Vicari, T. B. Hyde et al., *Consideration for oral cholera vaccine use during outbreak after earthquake in Haiti, 2010-2011*, Emerging Infectious Diseases, **17**, No. 11, (2011), 2105-2112.

53. D. Bernstein, *Matrix Mathematics*, Princeton University Press, (2005).

VITA

Chairat Modnak
Department of Computational and Applied Mathematics
Old Dominion University
Norfolk, VA 23529

PREVIOUS DEGREES:
B.S. Mathematics, March 1998, Naresuan University.
M.S. Applied Mathematics, March 2001, King Mongkut's Institute of Technology.
M.S. Mathematics, July 2007, Ohio University.

SCHOLARSHIP:
The Higher Educational Strategic Scholarships for Frontier Research Network (SFR Network), Thailand, 2002-2007.
The National Science Foundation support (One Quarter Project), Ohio University, USA, 2005.
The Graduate Teaching Assistantship, Old Dominion University, USA, 2007-2011.
The National Science Foundation support (Summer Project), Old Dominion University, USA, 2008-2012.

Typeset using LaTeX.