MODELING AND ANALYSIS OF AN EPIDEMIC MODEL WITH CLASSICAL KERMACK-MCKENDRICK INCIDENCE RATE UNDER TREATMENT

T. K. KAR1†, ASHIM BATABYAL2, AND R. P. AGARWAL3

1DEPARTMENT OF MATHEMATICS, BENGALENG. AND SCI. UNIV., SHIBPUR, HOWRAH-711103, INDIA
E-mail address: tkar1117@gmail.com, t_k_kar@yahoo.com

2DEPARTMENT OF MATHEMATICS, BALLY NISCHINDA CHITTARANJAN VIDYALAYA, BALLY GHOSHPARA, HOWRAH, INDIA
E-mail address: ashim.bat@gmail.com

3DEPARTMENT OF MATHEMATICS, FLORIDA INSTITUTE OF TECHNOLOGY, MELBOURN, FL 32901, USA
E-mail address: agarwal@fit.edu

ABSTRACT. An epidemic model with Classical Kermack-McKendrick incidence rate under a limited resource for treatment is proposed to understand the effect of the capacity for treatment. We have assumed that treatment function is strictly increasing function of infective individuals and becomes constant when the number of infective is very large. Existence and stability of the disease free and endemic equilibrium are investigated, boundedness of the solutions are shown. Even in this simple version of the model, backward bifurcation and multiple epidemic steady states can be observed with some sets of parameter values. Hopf-bifurcation analyses are given and numerical examples are provided to help understanding.

1. INTRODUCTION

Mathematical models have become important tools in analyzing the spread and control of infectious diseases. By setting up a good epidemic model and understanding well we can have many advantages of preventing invasion of infection to population. In recent times, epidemiological models have received much attention from researchers. Several epidemics models and reviews on theoretical developments are described in Anderson and May [1], Brauer and Castillo-Chavez [2], Britton [3], Diekmann and Heesterbeck [4], Hethcote [5], Lizana and Rivero [6], Ruan and Wang [7], Takeuchi et al. [8], Wang and Ma [9], Wang [10], Stanek [11], Stepan and Hlubinka [12], Brauer [13], Huang and Villasana [14] and Brauer [15]. Also we want to mention here a survey paper (Hethcote [16]) and some books (see Bailey [17], Heesterbeck [18]). The incidence rate in an epidemic model is the rate at which susceptible become infectious. The form of incidence rate that is used in our model is the classical

Received by the editors November 16 2009; Accepted December 7 2009.
2000 Mathematics Subject Classification. 92D30, 34K20.
Key words and phrases. Epidemic, endemic equilibrium, treatment, basic reproductive number, limit cycle.
† Corresponding author.
Kermack-McKendrick model [19], i.e. the simple mass action \( \alpha SI \), \( \alpha \) is called the transmission coefficient. Here susceptible become infectious by contact with infectious individuals. Later they may recover and join the group of immune (or dead) individuals. Treatment plays an important role to control or decrease the spread of diseases such as flu, tuberculosis, and measles (see Feng and Thieme [20], Wu and Feng [21], Hyman and Li [22]). In classical epidemic models, the treatment rate is assumed to be proportional to the number of the infectious, which is almost impossible in real perspective because in that case the resources for treatment should be quite large. In fact, every country or society should have a suitable capacity for treatment. If it is too large, the country or society pays for unnecessary cost. If it is too small, the country or society has the risk of the outbreak of a disease. As the member of infective individuals increases Government or Society of a country extend the existing capacity of treatment accordingly. But each country or society has certain maximum capacity of treatment for infective. Traditionally, the treatment function as taken as \( T(I) = \alpha I \). It embodies the unrealistic feature that there is unbounded linear increase in \( T \) with respect to \( I \). This unrealistic feature can largely be removed and the above feature can be included by adapting the alternative treatment function \( T(I) = bI/(1 + aI) \). Here we see that when \( I \to \infty \) i.e. when the numbers of infectious individuals become very large treatment become constant \( (b/a) \).

The paper is organized as follows. The model and its basic properties are given in section 2. Detailed qualitative analysis is given in section 3 (local stability), section 4 (global stability) and section 5 (bifurcation investigation). Some numerical evidence and discussions are given in section 6.

2. MODEL, BASIC PROPERTIES AND BACKWARD BIFURCATION:

To construct our model we make the following assumptions:

**\( H_1 \):** In the absence of disease the susceptible population grows according to a logistic growth, with carrying capacity \( K \) (> 0), and with an intrinsic birth rate \( r \) (> 0).

**\( H_2 \):** We assume that only susceptible population \( S \) is capable of reproducing with logistic growth rate, the infectious population \( I \) is removed by death with a constant death rate \( d \) (> 0) a portion of susceptible populations are recovered from disease by natural processes and some are recovered by treatment with a treatment rate \( T(I) = bI/(1 + aI) \).

Obviously the treatment function that we consider is a strictly increasing function and become constant when number of infectious individuals becomes very large.

**\( H_3 \):** We assume that the disease spreads with a simple mass action \( \alpha SI \) among the susceptible population only and is not genetically inherited. The infected populations never become susceptible.

Based on these assumptions the model is described by the following system of nonlinear differential equation

\[
\frac{dS}{dt} = rS \left( 1 - \frac{S + I}{K} \right) - \alpha SI, \tag{2.1}
\]

\[
\frac{dI}{dt} = \alpha SI - (d + \gamma)I - \frac{bI}{1 + aI}, \tag{2.2}
\]
\[
\frac{dR}{dt} = \gamma I + \frac{bI}{1 + aI} - dR, \quad (2.3)
\]

where \( S(t), I(t) \) and \( R(t) \) denotes the numbers of susceptible, infective, and recovered individuals at time \( t \), respectively. \( r(> 0) \) is the intrinsic growth rate of susceptible, \( K \) is the carrying capacity of the susceptible in the absence of infective, \( \alpha \) is the proportionality constant, \( d \) is the natural death rate of the population, \( \gamma \) is the natural recovery rate of infective individuals. Here we consider the treatment function as \( T(I) = bI/(1 + aI) \). In our model recovered individuals never become susceptible. Since the first two equations of (2.1)-(2.3) are independent of the variable \( R \), it suffices to consider the following reduced model

\[
\frac{dS}{dt} = rS \left( 1 - \frac{S + I}{K} \right) - \alpha SI, \quad \frac{dI}{dt} = \alpha SI - (d + \gamma)I - \frac{bI}{1 + aI}.
\]

The system of equation always has the trivial equilibrium \( E_0(0, 0) \) and the boundary equilibrium \( E_1(K, 0) \). To find the positive equilibria, set

\[
\begin{align*}
S &= K - \frac{\alpha K + r}{r} I, \quad (2.6) \\
I &= \frac{d + \gamma}{\alpha} + \frac{b}{\alpha(aI + 1)}. \quad (2.7)
\end{align*}
\]

Hence from (2.6) and (2.7) we get

\[
\begin{align*}
aa(K\alpha + r)I^2 + I \{ \alpha(K\alpha + r) + ar(d + \gamma - K\alpha) \} + r(d + \gamma + b - K\alpha) &= 0. \quad (2.8)
\end{align*}
\]

Thus if \( E^*(S^*, I^*) \) denotes the interior equilibrium,

\[
S^* = K - \left( \frac{K\alpha + r}{r} \right) I^* = \frac{d + \gamma}{\alpha} + \frac{b}{\alpha(1 + aI^*)},
\]

where \( I^* \) is positive root of equation (2.8).

There is a threshold quantity which determines whether an epidemic occurs or the disease simply dies out. This quantity is called basic reproductive number, denoted by \( R_0 \), which can be defined as the number of secondary infections caused by a single infected introduced into a population made up entirely of susceptible individuals over the course of infection of this single infective.

We define the basic reproduction number \( R_0 \) as follows:

\[
R_0 = \frac{K\alpha}{d + \gamma + b}. \quad (2.9)
\]
If \( R_0 > 1 \), then (2.8) admits a unique positive solution \( I^* \), and \( I^* \) is given by

\[
I^* = \frac{-B + \sqrt{B^2 - 4AC}}{2A},
\]

(2.10)

where

\[
A = \frac{a\alpha(K\alpha + r)}{K\alpha}, \quad B = \frac{\alpha(K\alpha + r) + ar(d + \gamma - K\alpha)}{K\alpha}, \quad C = \frac{r(d + \gamma + b - K\alpha)}{K\alpha}.
\]

So, for \( R_0 > 1 \), there exists a unique endemic equilibrium \((S^*, I^*)\) of the system (2.1)-(2.2), where

\[
I^* = \frac{-B + \sqrt{B^2 - 4AC}}{2A}, \quad S^* = \frac{d + \gamma + b}{\alpha} + \frac{b}{\alpha(1 + aI^*)}.
\]

Now if \( R_0 < 1 \), then (2.8) has no positive root if \( B > 0 \), but if \( R_0 < 1 \) and \( B < 0 \), then (2.8) has either two positive roots or no positive root. So for \( R_0 < 1 \) and \( B < 0 \), (2.8) has a pair of positive roots if \( B^2 - 4AC > 0 \).

Equation (2.8) can be written as

\[
AI^2 + BI + r\left(\frac{1}{R_0} - 1\right) = 0.
\]

Therefore, when \( R_0^* < R_0 < 1 \) and \( B < 0 \), (2.8) has a pair of positive roots, where \( R_0^* = 4Ar/(B^2 + 4Ar) \). Again \( B < 0 \), if \( K\alpha > \frac{r[a + a(d + \gamma)]}{ar - \alpha} \).

The above analysis can be stated through a proposition as follows

**Proposition 2.1.**

1. The trivial equilibrium \( E_0(0, 0) \) and the disease free equilibrium \( E_1(K, 0) \) always exist.

2. For \( R_0 > 1 \), a unique positive endemic equilibrium exists in addition to the disease free equilibrium.

3. For \( R_0^* < R_0 < 1 \) and \( \alpha K > r\{\alpha + a(d + \gamma)\}/(ar - \alpha) \), two positive equilibriums exist in addition to the disease free equilibrium.

4. For \( 0 < R_0 < R_0^* \) or \( R_0^* < R_0 < 1 \) and \( \alpha K < r\{\alpha + a(d + \gamma)\}/(ar - \alpha) \) there is only the disease free equilibrium.

5. For \( R_0 = 1 \) and \( \alpha K > r\{\alpha + a(d + \gamma)\}/(ar - \alpha) \), a unique positive endemic equilibrium exists in addition to the disease free equilibrium.

For \( \alpha K > r\{\alpha + a(d + \gamma)\}/(ar - \alpha) \) the model (2.1)-(2.2) exhibit a backward bifurcation. Apart from the disease-free equilibrium \( E_1(K, 0) \) there can exist a single unique endemic steady state or two positive steady states depending on the solutions to a quadratic equation. The endemic steady state value of \( S^* \) is obtained from \( S^* = \frac{d + \gamma}{\alpha} + \frac{b}{\alpha(1 + aI^*)} \), where \( I^* \) is the solution of the quadratic equation (2.8). Equation (2.8) has one or two feasible (i.e., positive, real) solutions, depending on the values of the parameters. As seen in the previous discussion, there is a unique positive endemic equilibrium for the case when \( R_0 > 1 \), For \( R_0 < 1 \), the situation is more complicated. For \( \alpha K > r\{\alpha + a(d + \gamma)\}/(ar - \alpha) \), there is a region of values for \( R_0 < 1 \), where there are two feasible solutions.
To test whether the system (2.1)-(2.3) is positively invariant or not let us suppose that 
\((S(t), I(t), R(t))\) is a solution of (2.1)-(2.3). If \(S(t) > 0, I(t) > 0, R(t) > 0\) for \(0 \leq t < t_0\) and \(I(t_0) = 0\), it is natural to assume that \((S(t), I(t), R(t))\) satisfies

\[
\frac{dS}{dt} = rS \left(1 - \frac{S}{K}\right), \quad I(t) = 0, \quad \frac{dR}{dt} = -dR, \quad \text{for } t \geq t_0
\]

Consequently, \(R_3^+\) is positively invariant for the system (2.1)-(2.3). Now we consider the boundedness of solutions of the system (2.1)-(2.3).

**Theorem 2.2.** All the solutions of the system (2.1)-(2.3) which initiate in \(R_3^+\) are uniformly bounded.

**Proof.** Let us consider a function

\[
U = S + I + R. \quad (2.11)
\]

Now, the time derivative of (2.11) along the solutions of (2.1)-(2.3) is

\[
\frac{dU}{dt} = rS \left(1 - \frac{S}{K}\right) - \frac{r}{K}SI - \alpha SI + \alpha SI - (d + \gamma)I - \frac{bI}{1 + aI} + \gamma I + \frac{bI}{1 + aI} - dR
\]

\[
= rS \left(1 - \frac{S}{K}\right) - \frac{r}{K}SI - dI - dR
\]

Or,

\[
\frac{dU}{dt} + cU \leq \left\{ r \left(1 - \frac{S}{K}\right) + c \right\} S + (c - d) I + (c - d) R,
\]

where \(c\) is some positive constant. Now choosing \(c = d\),

\[
\frac{dU}{dt} + dU < \left\{ r \left(1 - \frac{S}{K}\right) + d \right\} S,
\]

\[
\Rightarrow \frac{dU}{dt} + dU < \mu,
\]

where \(\mu = K(r + d)^2/4r\). Applying the theory of differential inequality (Birkoff and Rota [23]) we obtain \(0 < U(S, I, R) < (1 - e^{-dt}) \mu/d + U(S(0), I(0), R(0))e^{-dt}\), which, upon letting \(t \to \infty\), yield \(0 < U < \mu/d\). So, we have that all the solutions of the system (2.1)-(2.3) that start in \(R_3^+\) are confined to the region \(G\), where \(G=\{(S, I, R) \in R_3^+ : U = (\mu/d) + \varepsilon,\) for any \(\varepsilon > 0\}\). Hence the theorem. □
3. Dynamic Behavior

In this section we study the stability properties of different equilibriums. Stability analysis is critical in this study since we would be able to know whether endemic equilibrium would be stable so that the disease would persist or not. The Jacobian matrix of the system (2.1)-(2.2) at the trivial equilibrium $E_0(0, 0)$ is given by

$$M_0 = \begin{pmatrix} r & 0 \\ 0 & -(d + \gamma + b) \end{pmatrix}.$$ 

Hence, the eigen values of this system are the roots of the equation

$$(r - \lambda)(-d - \gamma - b - \lambda) = 0.$$ 

Therefore, $E_0(0, 0)$ is always an unstable saddle point.

Jacobian matrix for $E_1(K, 0)$ is given by

$$M_1 = \begin{pmatrix} -r & -(K\alpha + r) \\ 0 & K\alpha - (d + \gamma + b) \end{pmatrix}.$$ 

The eigen values of the matrix are $-r$ and $K\alpha - (d + \gamma + b)$. Hence $E_1(K, 0)$ is locally asymptotically stable when $K\alpha < (d + \gamma + b)$ i.e. when $R_0 < 1$, and unstable when $K\alpha > (d + \gamma + b)$ i.e. when $R_0 > 1$. It is observed that when unique endemic equilibrium $E^*(S^*, I^*)$ exists, $E_1(K, 0)$ is unstable. Also we can observe that, the disease-free equilibrium $E_1(K, 0)$ can exist simultaneously with two positive equilibriums of the system. The Jacobian matrix of the system at $E^*(S^*, I^*)$, the interior equilibrium is given by

$$M = \begin{pmatrix} W & X \\ Y & Z \end{pmatrix},$$

where $W = -(r/K)S^*$, $X = -(K\alpha + r)/K$, $Y = \alpha I^*$, $Z = abI^*/(aI^* + 1)^2$.

The characteristic equation is

$$\lambda^2 + \lambda \text{Trace}(M) + \text{Det}(M) = 0.$$ 

Here

$$\text{Trace}(M) = -\left(\frac{r}{K}\right)S^* + \frac{abI^*}{(aI^* + 1)^2}, \quad \text{Det}(M) = \left[\frac{\alpha(K\alpha + r)}{K} - \frac{r}{K} \frac{ab}{(aI^* + 1)^2}\right]S^*I^*.$$ 

The stability and classification of the equilibrium $E^*(S^*, I^*)$ dependents on the sign of the Trace($M$) and Det($M$).

1. If Trace($M$) < 0, and Det($M$) > 0, then $E^*(S^*, I^*)$ is a stable node or a stable spiral.
2. If Trace($M$) > 0, and Det($M$) > 0, then $E^*(S^*, I^*)$ is an unstable node or unstable spiral.
3. If Det($M$) < 0, then $E^*(S^*, I^*)$ is a saddle point.
4. If Trace($M$) = 0, then there are some limit cycles.
Now
\[ \text{Trace}(M) = -\frac{r}{K} S^* + \frac{abI^*}{(aI^* + 1)^2} = -\frac{r(d + \gamma)(aI^* + 1)^2 - rb(aI^* + 1) + abK\alpha I^*}{K\alpha (aI^* + 1)^2} \]
\[ = -\left[ \frac{r(d + \gamma)(aI^* + 1)^2 + rb + ab(r - K\alpha)I^*}{K\alpha (aI^* + 1)^2} \right]. \]

So, the sufficient condition for negative Trace is \( \alpha < \frac{r}{K} \) i.e. if the proportionality constant less than the biotic potential of the species.

Again,
\[ \text{Det}(M) = \left[ \frac{\alpha(K\alpha + r)}{K} - \frac{r}{K} \right] S^* I^* \]
\[ = \frac{S^* I^*}{K(aI^* + 1)^2} \left[ \frac{\alpha(K\alpha + r)a^2 I^* + 2aa(K\alpha + r)I^* + \alpha(K\alpha + r) - rab}{\alpha(K\alpha + r)} \right]. \]

Obviously, \( \text{Det}(M) > 0 \) for \( K\alpha > \frac{r}{\alpha} (ab - \alpha) \).

Therefore the unique positive equilibrium of the system is asymptotically stable if
\[ \min \left\{ d + \gamma + b, \frac{r}{\alpha} (ab - \alpha) \right\} < K\alpha < r. \] (3.1)

But when there exist two positive equilibriums, the positive equilibrium which satisfies (3.1) must be asymptotically stable.

Now there exists a limit cycle near the unique endemic equilibrium point if Trace\( (M) = 0 \) at that point. Trace\( (M) = 0 \) implies \( J_1 I^* + J_2 I^* + J_3 = 0 \), where
\[ J_1 = ra^2(d + \gamma), \quad J_2 = 2ar(d + \gamma) + ab(r - K\alpha), \quad J_3 = r(b + \gamma + d). \]

Again from the equation of equilibrium we have \( J_1 I^* + J_2 I^* + J_3 = 0 \), where
\[ J_1' = a\alpha(K\alpha + r), \quad J_2' = \alpha(K\alpha + r) + ar(d + \gamma - K\alpha), \quad J_3' = d + \gamma + b - K\alpha. \]

Therefore, if \( R_0 > 1 \), and \( I^* = \frac{J_3' - J_2'}{J_2' - J_1'} \), there exists a limit cycle near the unique endemic equilibrium point.

4. Global stability

In this section we investigate the global stability of both the disease free equilibrium and endemic equilibrium.

4.1. The global stability of the disease free equilibrium.

**Theorem 4.1.** The disease-free equilibrium \( E_1(K, 0) \) always exists and is globally stable if \( K\alpha < (d + \gamma - r) \).
Proof. To investigate the global stability of the disease-free equilibrium $E_1(K, 0)$, we consider the Lyapunov function as follows

$$V_1 = S - K - K \ln \left( \frac{S}{K} \right) + I.$$ 

Calculating the derivative of $V_1$ along the solutions of the system (2.4)-(2.5)

$$\frac{dV_1}{dt} = (S - K) \frac{1}{S} \frac{dS}{dt} + \frac{dI}{dt}$$

$$= (S - K) \left\{ r \left( 1 - \frac{S}{K} \right) - \left( \alpha + \frac{r}{K} \right) I \right\} + \alpha SI - (d + \gamma)I - \frac{bI}{1 + aI}$$

$$\leq - \frac{r}{K}(S - K)^2 - (d + \gamma - K\alpha - r)I < 0$$

in the interior of $R^2_+$ for $K\alpha \leq (d + \gamma - r)$.

Therefore, by Lyapunov-La Salle(Hale [24]), it follows that $E_1(K, 0)$ is locally asymptotically stable and all trajectories starting in $\text{Int} R^2_+$ approach $E_1(K, 0)$ as $t$ goes to infinity if the inequality $K\alpha \leq (d + \gamma - r)$ holds. This establishes the global stability of disease-free equilibrium $E_1(K, 0)$.

So, $E_1(K, 0)$ always exists and is globally stable for $K\alpha \leq (d + \gamma - r)$.

\[\Box\]

4.2. The global stability of the endemic equilibrium. To investigate the global dynamic behavior of the endemic equilibrium, the following discussion consists of two cases: $R_0 > 1$ and $R_0^* < R_0 < 1$.

Case 1. $R_0 > 1$.

Theorem 4.2. For $R_0 > 1$, the unique endemic equilibrium exists and is globally stable in the region $G_1 = \{(S, I) \in R^2_+: S > \frac{Kb}{4r}\}$ if $\frac{r}{\alpha}(ab - \alpha) < K\alpha < r$.

Proof. We know that the existence and stability of limit cycle is related to the existence and stability of a positive equilibrium. Also we know that, if there is no limit cycle and the endemic equilibrium is unique, the unique endemic equilibrium is globally stable. So, the limit cycles of (2.1)-(2.2) plays crucial roles on structure of dynamical behavior of the model. For this reason, we take Dulac function as $D(S, I) = 1/SI$.

Let $f_1(S, I) = rS \left( 1 - \frac{S}{K} \right) - \left( \alpha + \frac{r}{K} \right) SI$, and $f_2(S, I) = \alpha SI - (d + \gamma)I - \frac{bI}{1 + aI}$. Then we have,

$$\frac{\partial}{\partial S}(Df_1) + \frac{\partial}{\partial I}(Df_2) = - \frac{r}{KI} + \frac{ab}{S(1 + aI)^2} = - \frac{r}{KSI} \left[ S - \frac{KabI}{r(1 + aI)^2} \right] \leq 0,$$

if $S > \max \left\{ \frac{KabI}{r(1 + aI)^2} \right\}$ i.e. $S > \frac{Kb}{4r}$.

For $R_0 > 1$, the unique positive equilibrium exists and is globally stable in the region $G_1 = \{(S, I) \in R^2_+: S > \frac{Kb}{4r}\}$ for $\frac{r}{\alpha}(ab - \alpha) < K\alpha < r$. Hence proved.

\[\Box\]

Case 2. $R_0^* < R_0 < 1$. 

Theorem 4.3. For $R^*_0 < R_0 < 1$, the positive equilibrium will be globally stable if

$$\frac{r}{\alpha} (ab - \alpha) < K\alpha < \min \{r, d + \gamma\}.$$ 

Proof. To test the global stability of endemic equilibrium, we construct the Lyapunov function as $V = S - K \ln(S) + I$.

Now taking time derivative of $V$ along the solutions of equations (2.1)-(2.2), we get

$$\frac{dV}{dt} = \frac{dS}{dt} - K \frac{dS}{S} + \frac{dI}{dt} = (1 - K\alpha) \left\{ rS \left( 1 - \frac{S + I}{K} \right) - \alpha SI \right\} + \alpha SI - (d + \gamma)I - \frac{bI}{1 + aI}$$

$$= -(K - S) \left\{ r \left( 1 - \frac{S + I}{K} \right) \right\} + \alpha KI - \alpha SI + \alpha SI - (d + \gamma)I - \frac{bI}{1 + aI}$$

$$= -(K - S) \left\{ r \left( 1 - \frac{S + I}{K} \right) \right\} - (d + \gamma - K\alpha)I - \frac{bI}{1 + aI}.$$ 

Therefore, $\frac{dV}{dt} \leq 0$, if $K\alpha < d + \gamma$. Hence by La Salle’s theorem(Khalil [25]) the trajectories starting in $\text{Int}R^2_+$ approach $E^*(S^*, I^*)$ as $t$ goes to infinity if $K\alpha < d + \gamma$. This establishes the global stability of $E^*(S^*, I^*)$. Hence for $R^*_0 < R_0 < 1$, the positive equilibrium will be globally stable if $\frac{r}{\alpha} (ab - \alpha) < K\alpha < \min \{r, d + \gamma\}$. □

5. Hopf bifurcation and periodic solutions

Epidemiological models with constant parameters are often found to approach a steady state in which the species coexist in equilibrium. But if parameters used in the model are changed, other types of dynamical behavior may occur and the critical parameter values at which such transitions happen are called bifurcation points. When a stable steady state goes through a bifurcation will in general either lose its stability or disappear entirely. Even if the system ends up in another steady state the transition to that state will often involve the extinction of one or more species of the system. On the other hand the entire system may survive in a non-stationary state, but further bifurcation may lead to local extinction of species. In order to preserve the system under consideration in its natural state, crossing bifurcation should be avoided and in doing so it is of great importance to determine the critical parameter values at which bifurcation occur. However in order to understand the general mechanisms leading to bifurcations in ecosystems much simpler models are needed.

From the equation of equilibrium we have

$$P_1 I^{*2} + P_2 I^* + r(P_3 + b) = 0,$$ 

where $P_1 = a\alpha(K\alpha + r)$, $P_2 = \alpha(K\alpha + r) + a\alpha(d + \gamma - K\alpha)$, $P_3 = d + \gamma - K\alpha$.

$$I^* = -\frac{P_2 + \sqrt{P_2^2 - 4rP_3P_3 - 4rP_3b}}{2P_1} = \frac{\sqrt{P_4 - P_5b}}{2P_1} - \frac{P_2}{2P_1} = P_6\sqrt{P_4 - P_5b} - P_7,$$
where $P_4 = P_2^2 - 4rP_1P_3$, $P_5 = 4rP_1$, $P_6 = 1/2P_1$, $P_7 = P_2/2P_1$. Now \[ \text{Trace}(M) = 0 \implies P_8 I^2 + (P_9 + P_{10}b)I^* + r(P_{11} + b) = 0, \] (5.2)

where $P_8 = ra^2(d + \gamma)$, $P_9 = 2ar(d + \gamma)$, $P_{10} = a(r - K\alpha)$, $P_{11} = (d + \gamma)$.

Now putting the value of $I^*$ in (5.2), we get

\[
P_8 \left\{ P_6 \sqrt{P_4 - bP_5 - P_7} \right\}^2 + \left( P_6 \sqrt{P_4 - bP_5 - P_7} \right) (P_9 + bP_{10}) + r(P_{11} + b) = 0.
\]

Or,

\[
P_{12} + bP_{13} = \left( \sqrt{P_4 - bP_5} \right) (P_{14} - bP_{15}),
\]

where

\[
P_{12} = P_8P_4P_6^2 + P_8P_7^2 - P_7P_5 + rP_{11},
\]

\[
P_{13} = -P_8P_5P_6^2 - P_7P_{10} + r, P_{14} = 2P_6P_7P_8 - P_6P_9,
\]

\[
P_{15} = P_6P_{10}.
\]

Or,

\[
(P_{12} + bP_{13})^2 = (P_4 - bP_5)(P_{14} - bP_{15})^2.
\]

Or,

\[
P_{12}^2 + 2P_{12}P_{13}b + P_{13}^2b^2 = (P_4 - bP_5)(P_{14} - 2P_{14}P_{15}b + P_{15}^2b^2).
\]

Or,

\[
a_0b^3 + 3a_1b^2 + 3a_2b + a_3 = 0,
\]

where

\[
a_0 = P_{15}^2P_5
\]

\[
a_1 = (P_{13}^2 - P_4P_{15}^2 - 2P_5P_{14}P_{15}) / 3
\]

\[
a_2 = (2P_{12}P_{13} + 2P_4P_{14}P_{15} + P_5P_{14}^2) / 3
\]

\[
a_3 = (P_{12}^2 - P_4P_{14}^2) / 3.
\]

Now applying Cardan’s method we get $b = \{(p - a_1) - (H/p)\}/a_0$, where $H = a_0a_2 - a_1^2$.

$G = a_0^3a_3 - 3a_0a_1a_2 + 2a_0^3$, $G^2 + 4H^3 > 0$ and $p = \left\{ \left( -G + \sqrt{G^2 + 4H^3} \right) / 2 \right\}^{\frac{1}{3}}$.

But if $G^2 + 4H^3 < 0$, three real roots of $b$ are $2\sqrt{-H} \cos \frac{\theta}{3}$, $2\sqrt{-H} \cos \left( \frac{2\pi + \theta}{3} \right)$, $2\sqrt{-H} \cos \left( \frac{4\pi + \theta}{3} \right)$, where $r_1(\cos \theta) = -G/2$, $r_1(\sin \theta) = \sqrt{-\left( G^2 + 4H^3 \right)} / 2$, and $r_1^2 = -H^3$.

We now present a Hopf–bifurcation analysis of the system (2.1)-(2.2). Suppose that the parameter ‘$b$’ is such that Trace($M$) = 0 at the endemic equilibrium. Now if for any $b = b^*$, $\frac{\partial}{\partial b}$ (Trace($M$)) $\neq 0$, then by using the Hopf-bifurcation theorem (Hassard et al [26]), the system (2.1)-(2.2) enters into Hopf type small amplitude periodic solution at parametric value $b = b^*$ near the positive interior equilibrium.

For our model it is very difficult to test $\frac{\partial}{\partial b}$ (Trace($M$)) $\neq 0$ analytically, but examples are provided to help understanding.
6. Numerical Analysis and Discussion

Our proposed model consists of three nonlinear ordinary differential equations, namely, a susceptible population, an infective population and recovered population. We have shown that all the solutions are bounded in the region \( G = \{(S, I, R) \in \mathbb{R}^3_+ : U = (\mu/d) + \varepsilon, \text{ for any positive } \varepsilon \} \). For the reduced model, both the trivial equilibrium \( E_0(0, 0) \) and a disease free equilibrium \( E_1(K, 0) \) exist, among which \( E_0(0, 0) \) is always unstable. Conditions for the unique endemic equilibrium or two endemic equilibriums are obtained. We have considered the treatment function as \( T(I) = bI/(1 + aI) \), and it is realized that it plays an important role for the existence of different equilibriums. We take the values of different parameters as follows, \( r = 3, \ K = 100, \ \alpha = 0.1, \ d = 0.5, \ \gamma = 0.3, \ a = 0.1, \ 0 < b < 18 \).

![Figure 1. Existence diagram of different equilibrium points in \( (b, R_0) \) plane.](image-url)

In Figure 1, we see that when \( b < 9.2 \) the basic reproductive number, \( R_0 > 1 \) and there exists a unique positive equilibrium. For \( b = 8 \), we get the corresponding unique positive equilibrium \((42.31783, 13.31127)\), where \( R_0 = 1.13636 > 1 \). Now for \( b \in [9.2, 10.566) \) the corresponding values of \( R_0 \in [R_0^*, 1) \), where \( R_0^* = 0.879752 \). We see that, when \( b = 9.2 \) there exists unique positive equilibrium \((51.33, 11.23)\) and when \( b = 9.3 \), there exist two positive endemic equilibriums \((52.24, 11.02)\) and \((99.09, 0.209)\), corresponding value of \( R_0 \) is \( .990099 \). But if we take \( b = 10.567 \), value of \( R_0 \) becomes \( 0.8797396 \) which is less than \( R_0^* \). So in that case there exists no real interior equilibrium point.

To understand the role of treatment to the stability of unique endemic equilibrium, we have drawn Figure 2. Here we take the values of the parameters same as before.

In Figure 2, we see that, \( \det(M) \) is always positive and \( \text{Trace}(M) < 0 \) for \( b < 1.956 \). So, for \( 0 < b < 1.956, \ R_0 > 1 \) and the unique endemic equilibrium is stable. For \( b = 1 \) and
values of all other parameters same as before, we get the unique endemic equilibrium point (11.28, 20.47). Here $R_0=5.555556$, $\text{Trace}(M) = -0.1179772$, $\text{Det}(M) =2.928028$ respectively, and hence the equilibrium is a stable spiral, see Figure 3.

For $b = 2.1$, the unique endemic equilibrium is (15.096, 19.593) and the corresponding values of $\text{Trace}(M)$, $\text{Det}(M)$ and $R_0$ are 0.01694339, 3.632401 and 3.448276 respectively. So, here both of $\text{Trace}(M)$ and $\text{Det}(M)$ are positive and hence the unique endemic equilibrium is unstable (see Figures 4).
System (2.1)-(2.2) has a unique endemic equilibrium $(15.096, 19.593)$, which is an unstable focus. There exists a unique stable periodic solution. Again for $b=1.965$, $\text{Trace}(M) = 0$ and $\text{Det}(M) > 0$ and in that case limit cycle occur, see Figure 5.

Now, if we take the values of parameters as follows: $r = 1.5$, $K = 120$, $\alpha = 0.05$, $d = 1$, $\gamma = 0.1$, $a = 2.7$, $b = 5$, we get the corresponding value of $R_0 = 0.9836066$ and $R_0^* = 0.08771911$ So, here $R_0^* < R_0 < 1$, and we get two interior equilibriums $(23.89, 19.22)$ and...
(119.96, 0.77). For the equilibrium (23.89, 19.22) the corresponding values of \( \text{Trace}(M) \) and \( \text{Det}(M) \) are -0.2058972 and 1.407367 respectively, and for the equilibrium (119.96, 0.77), the corresponding values of \( \text{Trace}(M) \) and \( \text{Det}(M) \) are -1.399669 and -0.1468361 respectively. Therefore the equilibrium point (23.89, 19.22) is stable and (119.96, 0.77) is unstable, see Figures 6.

![Figure 6](image-url)

**Figure 6.** Figures correspond the case \( R_0^* < R_0 < 1 \), where one of the positive endemic equilibrium is globally stable.

Again, if we take \( r = 3, K = 100, \alpha = 0.1, d = 0.8, \gamma = 1.4, a = 0.3, b = 8 \), we get \( R_0 = 0.9803922, R_0^* = 0.588543 \) and the corresponding two endemic equilibriums are (36.9, 14.56) and (99.54, 0.105656). For (36.9, 14.56), the values of \( \text{Trace}(M) \) and \( \text{Det}(M) \) are 0.1055624 and 5.642884 respectively. Also for (99.54, 0.105656), those values are -2.748032 and -0.5747029 respectively. Therefore in this case both of two equilibriums are unstable and only the disease-free equilibrium \( E_1(K, 0) \) is stable here (see Figures 7).

In this paper, by combining qualitative and bifurcation and analyses we have studied the global behavior of an SIR epidemic model with treatment. This model can be more significant when it is transformed to adjust to a specific transmittable disease; by using specific parameter values and adding some extra terms into a model would do this.

**References**


Figures correspond the case $R_0^* < R_0 < 1$, where both of the positive endemic equilibriums are unstable and only the disease free equilibrium is globally stable.


SOLUTION OF TENTH AND NINTH-ORDER BOUNDARY VALUE PROBLEMS
BY HOMOTOPY PERTURBATION METHOD

SYED TAUSEEF MOHYUD-DIN\textsuperscript{1} AND AHMET YILDIRIM\textsuperscript{2}

\textsuperscript{1}HITEC UNIVERSITY, TAXILA CANTT, PAKISTAN
\textit{E-mail address: syedtauseefs@hotmail.com}

\textsuperscript{2}EGE UNIVERSITY, SCIENCE FACULTY, DEPT. OF MATHEMATICS, 35100 BORNOVA-IZMIR, TURKEY

ABSTRACT. In this paper, we apply homotopy perturbation method (HPM) for solving ninth and tenth-order boundary value problems. The suggested algorithm is quite efficient and is practically well suited for use in these problems. The proposed iterative scheme finds the solution without any discretization, linearization or restrictive assumptions. Several examples are given to verify the reliability and efficiency of the method. The fact that the proposed homotopy perturbation method solves nonlinear problems without using Adomian’s polynomials can be considered as a clear advantage of this technique over the decomposition method.

1. INTRODUCTION

In the last two decades, with the rapid development of nonlinear science, there has appeared ever-increasing interest of scientists, physicists and engineers in the analytical techniques for nonlinear problems. It is well known, that perturbation methods provide the most versatile tools available in nonlinear analysis of engineering problems, see [7-14, 22-29, 42] and the references therein. The Perturbation methods, like other nonlinear analytical techniques, have their own limitations. At first, almost all perturbation methods are based on an assumption that a small parameter must exist in the equation. This so-called small parameter assumption greatly restricts applications of perturbation techniques. As is well known, an overwhelming majority of nonlinear problems have no small parameters at all. Secondly, the determination of small parameter seems to be a special art requiring special techniques. An appropriate choice of small parameters leads to the ideal results but, an unsuitable choice may create serious problems. Furthermore, the approximate solutions solved by perturbation methods are valid, in most cases, only for the small values of the parameters. It is obvious that all these limitations come from the small parameter assumption. These facts have motivated to suggest alternate techniques such as, variational iteration [1-3, 14-21, 25-35], decomposition [39, 40], variation
of parameters [36] and exp-function [37, 38]. In order to overcome these drawbacks, combining
the standard homotopy method and perturbation, which is called the homotopy perturbation,
modifies the homotopy method.

This paper is devoted to the study of boundary value problems of tenth and ninth-order which
are known to arise in the study of astrophysics, hydrodynamic and hydro magnetic stability, see
[4-6, 25, 34, 35, 39, 40]. A class of characteristic-value problems of higher order (as higher
as twenty four) is known to arise in hydrodynamic and hydro magnetic stability. In addition,
it is well known that when a layer of fluid is heated from below and is subject to the action of
rotation, instability may set in as ordinary convection which may be modeled by a tenth-order
boundary value problem, see [4-6, 34, 25, 35, 39, 40] and the references therein. The boundary
value problems of higher-order have been investigated due to their mathematical importance
and the potential for applications in diversified applied sciences. Several techniques includ-
ing the finite-difference, decomposition, variational iteration and modified variational iteration
have been employed for solving such problems, see [6, 25, 34, 35, 39, 40] and the references
therein. He [9-14] developed the homotopy perturbation method for solving linear, nonlinear,
initial and boundary value problems. Moreover, He realized the physical significance of the
homotopy perturbation method, its compatibility with the physical problems and applied this
promising technique to a wide class of linear and nonlinear, ordinary, partial, deterministic or
stochastic differential equation; see [9-14]. In this method the solution is given in an infinite
series usually converging to an accurate solution [7-14, 22-29, 42]. The basic motivation of this
paper is to apply the homotopy perturbation method for solving boundary value problems of
tenth and ninth-order. It is worth mentioning that the suggested method is applied without any
discretization, restrictive assumption or transformation and is free from round off errors. The
selection of initial value is done very carefully because the approximants are heavily depen-
dant upon the initial value. Unlike the method of separation of variables that require initial and
boundary conditions, the method provides an analytical solution by using the initial conditions
only. The proposed method work efficiently and the results are very encouraging and reli-
able. The fact that the proposed HPM solves nonlinear problems without using the Adomian’s
polynomials can be considered as a clear advantage of this technique over the decomposition
method. Several examples are given to verify the reliability and efficiency of the homotopy
perturbation method.

2. Homotopy Perturbation Method

To explain the homotopy perturbation method, we consider a general equation of the type,
\[ L(u) = 0, \quad (2.1) \]
where \( L \) is any integral or differential operator. We define a convex homotopy \( H(u, p) \) by
\[ H(u, p) = (1 - p)F(u) + pL(u), \quad (2.2) \]
where \( F(u) \) is a functional operator with known solutions \( u_0 \), which can be obtained easily. It
is clear that, for
\[ H(u, p) = 0, \quad (2.3) \]
we have
\[ H(u, 0) = F(u), \quad H(u, 1) = L(u). \]
This shows that \( H(u, p) \) continuously traces an implicitly defined curve from a starting point \( H(v_0, 0) \) to a solution function \( H(f, 1) \). The embedding parameter monotonically increases from zero to unit as the trivial problem \( F(u) = 0 \) is continuously deforms the original problem \( L(u) = 0 \). The embedding parameter \( p \in (0, 1) \) can be considered as an expanding parameter \([7-14, 22-29, 42]\). The homotopy perturbation method uses the homotopy parameter \( p \) as an expanding parameter \([9-14]\) to obtain
\[
    u = \sum_{i=0}^{\infty} p^i u_i = u_0 + p u_1 + p^2 u_2 + p^3 u_3 + \cdots, \quad (2.4)
\]
if \( p \to 1 \), then (2.4) corresponds to (2.2) and becomes the approximate solution of the form,
\[
    f = \lim_{p \to 1} u = \sum_{i=0}^{\infty} u_i. \quad (2.5)
\]
It is well known that series (2.5) is convergent for most of the cases and also the rate of convergence is dependent on \( L(u) \), see \([9-14]\). We assume that (2.5) has a unique solution. The comparisons of like powers of \( p \) give solutions of various orders.

### 3. Numerical Applications

In this section, we apply the homotopy perturbation method (HPM) for solving the boundary value problems of tenth and ninth-order. The selection of initial value is done carefully because the approximants are heavily dependant upon initial value.

**Example 3.1** \([25, 34, 35, 39]\) Consider the following nonlinear boundary value problem of tenth-order
\[
y^{(x)}(x) = e^{-x}y^2(x), \quad 0 < x < 1,
\]
with boundary condition
\[
y(0) = 1, \quad y''(0) = y^{(iv)}(0) = y^{(vi)}(0) = y^{(viii)}(0) = 1,
y(1) = e, \quad y''(1) = y^{(iv)}(1) = y^{(vi)}(1) = y^{(viii)}(1) = e.
\]
The exact solution of the problem is
\[
y(x) = e^x.
\]
Applying the convex homotopy method
\[
y^{(x)}_0(x) + p y^{(x)}_1(x) + p^2 y^{(x)}_2(x) + \cdots = p \left( e^{-x} \left(y_0(x) + p y_1(x) + p^2 y_2(x) + \cdots \right)^2 \right).
\]
Comparing the co-efficient of like powers of $p$

$p^{(0)}: y_0(x) = 1,$

$p^{(1)}: y_1(x) = Ax + \frac{1}{2!} x^2 + \frac{1}{3!} B x^3 + \frac{1}{4!} x^4 + \frac{1}{5!} C x^5 + \frac{1}{6!} x^6 + \frac{1}{7!} D x^7$

\[+ \frac{1}{8!} x^8 + \frac{1}{9!} E x^9 + \frac{1}{10!} x^{10} + \frac{1}{11!} x^{11} + \frac{1}{12!} x^{12} + \cdots , \]

$p^{(2)}: y_2(x) = \frac{2}{11!} A x^{11} + \left( -\frac{4}{12!} A + \frac{1}{239500800} \right) x^{12} + \cdots , \]

The series solution is given as:

\[y(x) = 1 + A x + \frac{1}{2!} x^2 + \frac{1}{3!} B x^3 + \frac{1}{4!} x^4 + \frac{1}{5!} C x^5 + \frac{1}{6!} x^6 + \frac{1}{7!} D x^7 \]

\[+ \frac{1}{8!} x^8 + \frac{1}{9!} E x^9 + \frac{1}{10!} D x^{10} + \left( -\frac{1}{19958400} A + \frac{1}{39916800} \right) x^{11} \]

\[+ \left( -\frac{1}{119750400} A + \frac{1}{159667200} \right) x^{12} + O(x^{13}) , \]

where

\[A = y'(0), \quad B = y^{(3)}(0), \quad C = y^{(5)}(0), \quad D = y^{(7)}(0), \quad E = y^{(9)}(0). \]

Imposing the boundary conditions at $x = 1$, we obtain

\[A = 1.00001436, \quad B = 0.999858964, \quad C = 1.001365775, \]

\[D = 0.987457318, \quad E = 1.0932797434 . \]

The series solution is given as:

\[y(x) = 1 + 1.00001436 x + \frac{1}{2!} x^2 + 0.1666431607 x^3 + \frac{1}{4!} x^4 + 0.008344714791 x^5 \]

\[+ \frac{1}{6!} x^6 + 0.00019524071 x^7 + \frac{1}{8!} x^8 + 3.013 \times 10^{-6} x^9 + \frac{1}{10!} x^{10} \]

\[+ 2.51 \times 10^{-8} x^{11} - 2.087 \times 10^{-9} x^{12} + \cdots . \]

Table 1 exhibits the exact solution and the series solution along with the errors obtained by using the homotopy perturbation method. It is obvious that the errors can be reduced further and higher accuracy can be obtained by evaluating more components of $y(x)$.

Figure 1 clearly indicates the accuracy of the proposed homotopy perturbation method (HPM).

**Example 3.2** [34, 35, 40] Consider the following linear boundary value problem of tenth-order

\[y^{(x)}(x) = -8 e^x + y''(x), \quad 0 < x < 1, \]
TABLE 1. (Error estimates) Error = Exact solution - Series solution.

<table>
<thead>
<tr>
<th>$x$</th>
<th>Exact solution</th>
<th>Series solution</th>
<th>Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1.00000000000</td>
<td>1.00000000000</td>
<td>0.00000</td>
</tr>
<tr>
<td>0.1</td>
<td>1.105170918</td>
<td>1.10517233</td>
<td>-1.41E-6</td>
</tr>
<tr>
<td>0.2</td>
<td>1.221402758</td>
<td>1.221405446</td>
<td>-2.69E-6</td>
</tr>
<tr>
<td>0.3</td>
<td>1.349858808</td>
<td>1.349862509</td>
<td>-3.70E-6</td>
</tr>
<tr>
<td>0.4</td>
<td>1.491824698</td>
<td>1.49182905</td>
<td>-4.35E-6</td>
</tr>
<tr>
<td>0.5</td>
<td>1.648721271</td>
<td>1.648725849</td>
<td>-4.58E-6</td>
</tr>
<tr>
<td>0.6</td>
<td>1.822118800</td>
<td>1.822123158</td>
<td>-4.36E-6</td>
</tr>
<tr>
<td>0.7</td>
<td>2.013752707</td>
<td>2.013756415</td>
<td>-3.71E-6</td>
</tr>
<tr>
<td>0.8</td>
<td>2.225540928</td>
<td>2.225543623</td>
<td>-2.69E-6</td>
</tr>
<tr>
<td>0.9</td>
<td>2.459603111</td>
<td>0.459604528</td>
<td>-1.42E-6</td>
</tr>
<tr>
<td>1.0</td>
<td>2.718281828</td>
<td>2.7182830</td>
<td>2.00E-9</td>
</tr>
</tbody>
</table>

FIGURE 1. Graphical comparison between approximate solution and exact solution

with boundary conditions

\[
y(0) = 1, \quad y'(0) = 0, \quad y''(0) = -1, \quad y'''(0) = -2, \quad y^{(iv)}(0) = -3, \\
y(1) = 0, \quad y'(1) = -e, \quad y''(1) = -2e, \quad y'''(1) = -3e, \quad y^{(iv)}(1) = -4e.
\]

The exact solution of the problem is

\[
y(x) = (1 - x)e^x.
\]
Applying the convex homotopy method
\[ y^{(x)}_0(x) + py^{(x)}_1(x) + p^2 y^{(x)}_2(x) + \cdots = p \left( (y^{(x)}_0(x) + py^{(x)}_1(x) + p^2 y^{(x)}_2(x) + \cdots) - 8e^x \right). \]

Comparing the co-efficient of like powers of p

\[ p^{(0)}: \quad y_0(x) = 1, \]
\[ p^{(1)}: \quad y_1(x) = -8e^x + 8 + 8x + \frac{7}{2!} x^2 + x^3 + \frac{5}{4!} x^4 + \left( \frac{1}{15} + \frac{1}{5!} A \right) x^5 \]
\[ + \left( \frac{1}{90} + \frac{1}{6!} B \right) x^6 + \left( \frac{1}{630} + \frac{1}{7!} C \right) x^7 + \left( \frac{1}{7!} + \frac{1}{8!} D \right) x^8 \]
\[ + \left( \frac{1}{45360} + \frac{1}{9!} E \right) x^9, \]
\[ p^{(2)}: \quad y_2(x) = \frac{1}{518400} x^{10} + \frac{1}{6652800} x^{11} + \frac{1}{9580320} x^{12} \]
\[ + \left( \frac{1}{778377600} + \frac{1}{6227020800} A \right) x^{13} + \cdots , \]
\[ \vdots \]

The series solution is given by
\[ y(x) = 17 - 16e^x + 16x + \frac{15}{2!} x^2 + \frac{7}{3} x^3 + \frac{13}{4!} x^4 + \left( \frac{2}{15} + \frac{1}{15!} \right) x^5 \]
\[ + \left( \frac{1}{45} + \frac{1}{6!} B \right) x^6 + \left( \frac{1}{315} + \frac{1}{7!} C \right) x^7 + \left( \frac{2}{7!} + \frac{1}{8!} D \right) x^8 \]
\[ + \frac{1}{9!} (8 + E) x^9 + \frac{7}{10!} x^{10} + \frac{6}{11!} x^{11} + \frac{5}{12!} x^{12} + \cdots . \]

Imposing the boundary conditions at \( x = 1 \) yields
\[ A = -4.00002, \quad B = -4.99999999, \quad C = -6.00100, \quad D = -7.00000, \quad E = -8.010000. \]

The series solution is given by
\[ y(x) = 17 - 16e^x + 16x + \frac{15}{2!} x^2 + \frac{7}{3} x^3 + \frac{13}{4!} x^4 + 0.999999999997 x^5 \]
\[ + 0.15277791666666666666x^6 + 0.00198392857142857x^7 \]
\[ + 0.0002332142857142857x^8 - 2.75573192239859 \times 10^{-8} x^9 \]
\[ + \frac{7}{518400} x^{10} + \frac{1}{6652800} x^{11} + \frac{1}{9580320} x^{12} + \cdots . \]

Table 2 exhibits the exact solution and the series solution along with the errors obtained by using the homotopy perturbation method. It is obvious that the errors can be reduced further and higher accuracy can be obtained by evaluating more components of \( y(x) \).
TABLE 2. (Error estimates) Error = Exact solution - series solution.

<table>
<thead>
<tr>
<th>x</th>
<th>Exact solution</th>
<th>Series solution</th>
<th>Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>1.000000000000</td>
<td>1.000000000000</td>
<td>0.000000000000</td>
</tr>
<tr>
<td>0.10</td>
<td>.9946538263</td>
<td>.9946538263</td>
<td>0.000000000000</td>
</tr>
<tr>
<td>0.20</td>
<td>.9771222065</td>
<td>.9771222065</td>
<td>0.000000000000</td>
</tr>
<tr>
<td>0.30</td>
<td>.9449011653</td>
<td>.9449011648</td>
<td>0.000000000005</td>
</tr>
<tr>
<td>0.40</td>
<td>.8950948186</td>
<td>.8950948125</td>
<td>0.000000000061</td>
</tr>
<tr>
<td>0.50</td>
<td>.8243606354</td>
<td>.8243605909</td>
<td>0.00000000444</td>
</tr>
<tr>
<td>0.60</td>
<td>.7288475202</td>
<td>.7288472928</td>
<td>0.0000002274</td>
</tr>
<tr>
<td>0.70</td>
<td>.6041258122</td>
<td>.6041249068</td>
<td>0.0000009055</td>
</tr>
<tr>
<td>0.80</td>
<td>.4451081857</td>
<td>.4451051850</td>
<td>0.0000030006</td>
</tr>
<tr>
<td>0.90</td>
<td>.2459603111</td>
<td>.2459516717</td>
<td>0.0000086394</td>
</tr>
<tr>
<td>1.00</td>
<td>0.000000000000</td>
<td>-.000222586</td>
<td>0.0000222586</td>
</tr>
</tbody>
</table>

Figure 2 clearly indicates the accuracy of the proposed homotopy perturbation method (HPM).
Example 3.3 [39] Consider the following ninth order boundary value problem

\[ y^{(ix)} = -9e^x + y(x), \quad 0 < x < 1 \]

with boundary conditions

\begin{align*}
    y(0) &= 1, \quad y^{(1)}(0) = 0, \quad y^{(2)}(0) = -1, \quad y^{(3)}(0) = -2, \quad y^{(4)}(0) = -3, \\
    y(1) &= 0, \quad y^{(1)}(1) = -e, \quad y^{(2)}(1) = -2e, \quad y^{(3)}(1) = -3e.
\end{align*}

The exact solution of the problem is

\[ y(x) = (1 - x)e^x. \]

Applying the convex homotopy method

\[
    y^{(ix)}_0(x) + py^{(ix)}_1(x) + p^2 y^{(ix)}_2(x) + \cdots = p \left( (y_0(x) + py_1(x) + p^2 y_2(x) + \cdots) - 9e^x \right).
\]

Comparing the co-efficient of like powers of \( p \)

\[
    p^{(0)} : y_0(x) = 1, \\
    p^{(1)} : y_1(x) = -\frac{1}{2}x^2 - \frac{1}{3}x^3 - \frac{1}{8}x^4 + \frac{1}{5!}Ax^5 + \frac{1}{6!}Bx^6 + \frac{1}{7!}Cx^7 + \frac{1}{8!}Dx^8 \\
    -\frac{8}{9!}x^9 - \frac{9}{10!}x^{10} - \frac{10}{11!}x^{11} - \frac{11}{12!}x^{12} + \cdots,
\]

The series solution is given by

\[
    y(x) = 1 - \frac{1}{2}x^2 - \frac{1}{3}x^3 - \frac{1}{8}x^4 + \frac{1}{5!}Ax^5 + \frac{1}{6!}Bx^6 + \frac{1}{7!}Cx^7 + \frac{1}{8!}Dx^8 \\
    -\frac{8}{9!}x^9 - \frac{9}{10!}x^{10} - \frac{10}{11!}x^{11} - \frac{11}{12!}x^{12} + \cdots.
\]

Imposing the boundary condition at \( x = 1 \) gives

\[ A = -3.999992, \quad B = -5.00017, \quad C = -5.9985, \quad D = -7.005. \]

The series solution is given as

\[
    y(x) = 1 - \frac{1}{2}x^2 - \frac{1}{3}x^3 + \frac{1}{8}x^4 - 0.03333326667x^5 - 0.006944680556x^6 \\
    -0.00119017851x^7 - 0.000173735119x^8 \\
    -\frac{8}{9!}x^9 - \frac{9}{10!}x^{10} - \frac{10}{11!}x^{11} - \frac{11}{12!}x^{12} + \cdots.
\]

Table 3 exhibits the exact solution and the series solution along with the errors obtained by using the homotopy perturbation method. It is obvious that the errors can be reduced further and higher accuracy can be obtained by evaluating more components of \( y(x) \).

Figure 3 clearly indicates the accuracy of the proposed homotopy perturbation method (HPM).
Table 3. (Error estimates) \( \text{Error} = \text{Exact solution} - \text{Series solution} \).

<table>
<thead>
<tr>
<th>( x )</th>
<th>Exact solution</th>
<th>Series solution</th>
<th>Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1.00000000</td>
<td>1.000000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>0.1</td>
<td>0.99465383</td>
<td>0.9946538264</td>
<td>-2.0E-10</td>
</tr>
<tr>
<td>0.2</td>
<td>0.97712221</td>
<td>0.9771222066</td>
<td>-2.0E-10</td>
</tr>
<tr>
<td>0.3</td>
<td>0.94490117</td>
<td>0.9449011654</td>
<td>-2.0E-10</td>
</tr>
<tr>
<td>0.4</td>
<td>0.89509482</td>
<td>0.8950948186</td>
<td>-2.0E-10</td>
</tr>
<tr>
<td>0.5</td>
<td>0.82436064</td>
<td>0.8243606355</td>
<td>-2.0E-10</td>
</tr>
<tr>
<td>0.6</td>
<td>0.72884752</td>
<td>0.7288475206</td>
<td>-6.0E-10</td>
</tr>
<tr>
<td>0.7</td>
<td>0.60412581</td>
<td>0.6041258131</td>
<td>-1.0E-9</td>
</tr>
<tr>
<td>0.8</td>
<td>0.44510819</td>
<td>0.4451081876</td>
<td>-2.0E-9</td>
</tr>
<tr>
<td>0.9</td>
<td>0.24596031</td>
<td>0.2459603145</td>
<td>-3.4E-9</td>
</tr>
<tr>
<td>1.0</td>
<td>0.00000000</td>
<td>0.00000000000</td>
<td>0.000000</td>
</tr>
</tbody>
</table>

Figure 3. Graphical comparison between approximate solution and exact solution

4. Conclusions

In this paper, we applied the homotopy perturbation method (HPM) for finding the solution of ninth and tenth-order boundary value problems. The method is applied in a direct way without using linearization, transformation, discretization or restrictive assumptions. It may be concluded that HPM is very powerful and efficient in finding the analytical solutions for a wide class of boundary value problems. The method gives more realistic series solutions that
converge very rapidly in physical problems. It is worth mentioning that the method is capable of reducing the volume of the computational work as compare to the classical methods while still maintaining the high accuracy of the numerical result. The fact that the HPM solves nonlinear problems without using the Adomian’s polynomials is a clear advantage of this technique over the decomposition method.

REFERENCES


A NEW UNDERSTANDING OF THE QR METHOD

CHOHONG MIN

DEPARTMENT OF MATHEMATICS, KYUNGHEE UNIVERSITY, SOUTH KOREA
E-mail address: chohongmin@gmail.com

ABSTRACT. The QR method is one of the most common methods for calculating the eigenvalues of a square matrix, however its understanding would require complicated and sophisticated mathematical logics. In this article, we present a simple way to understand QR method only with a minimal mathematical knowledge. A deflation technique is introduced, and its combination with the power iteration leads to extracting all the eigenvectors. The orthogonal iteration is then shown to be compatible with the combination of deflation and power iteration. The connection of QR method to orthogonal iteration is then briefly reviewed. Our presentation is unique and easy to understand among many accounts for the QR method by introducing the orthogonal iteration in terms of deflation and power iteration.

1. INTRODUCTION

Since its inception by Francis [4, 5] and Kublanovskaya [8], the QR method has been the most widely used and the most popular method for calculating the eigenvalues of a full matrix. It has been generalized to wider range of eigenvalue problems; QZ method, one of its variants, solves generalized eigenvalue problem $Ax = \lambda Bx$ [7], and a generalization of the QR method called GR method has been researched by Watkins et al [11]. For each specific application, the method has been tuned and upgraded; the restarted QR method is for comrade and fellow matrices [3], and the QR method with a balance technique is for finding the roots of polynomials [1].

Though its importance cannot be overstated, the QR method is normally deferred to graduate course, or would be just presented without enough explanation why it works. The enigma stems from the fact that the convergence proof is not trivial at all [6]. Even presenting a simple way to understand the QR method has been a research topic [10, 2]. The most accounts take the approach of explaining the orthogonal iteration and its connection to QR method [9, 6, 10], and so does this article. But our presentation is different in explaining the orthogonal iteration as successive application of the power iteration to deflated matrices. This article introduces a simple way to understand the QR method only with a minimal knowledge of mathematics.
2. Power Iteration and Deflated Matrix

Throughout this paper, we assume a matrix \( A \in \mathbb{C}^{n \times n} \) to have distinct eigenvalues \( \lambda_1, \lambda_2, \ldots, \lambda_n \) with their associated eigenvectors \( x_1, x_2, \ldots, x_n \). The eigenvalues are numbered in decreasing order, \( |\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0 \).

**Theorem 2.1.** (Power Iteration) Assume that \( v^0 \) in algorithm 1 is chosen randomly enough to have nonzero component of eigenvector \( x_1 \), then the sequence \( v^k \) satisfies

\[
\lim_{k \to \infty} \text{dist} \left( v^k, \text{span}(x_1) \right) = 0.
\]

**Proof.** By the assumption, \( v^0 = a_1x_1 + a_2x_2 + \cdots + a_nx_n \) with \( a_1 \neq 0 \). \( v^k \) in algorithm 1 is parallel to \( Av^k \) whenever \( k \geq 1 \). Repeating this argument leads to the fact that \( v^k \) is parallel to \( A^kv^0 \), and \( v^k = \frac{A^kv^0}{\|A^kv^0\|} \).

\[
A^kv^0 = a_1\lambda_1^kx_1 + a_2\lambda_2^kx_2 + \cdots + a_n\lambda_n^kx_n
\]

\[
\lim_{k \to \infty} v^k = \lim_{k \to \infty} \frac{a_1\lambda_1^kx_1 + a_2\lambda_2^kx_2 + \cdots + a_n\lambda_n^kx_n}{\|a_1\lambda_1^kx_1 + a_2\lambda_2^kx_2 + \cdots + a_n\lambda_n^kx_n\|}
\]

\[
= \lim_{k \to \infty} \left( \frac{\lambda_1}{\|\lambda_1\|} \right)^k x_1 + \frac{a_2}{a_1} \left( \frac{\lambda_2}{\|\lambda_1\|} \right)^k x_2 + \cdots + \frac{a_n}{a_1} \left( \frac{\lambda_n}{\|\lambda_1\|} \right)^k x_n
\]

\[
= \lim_{k \to \infty} \left( \frac{\lambda_1}{\|\lambda_1\|} \right)^k \frac{x_1}{\|x_1\|}
\]

As \( k \to \infty \), \( v^k \) belongs to the \( \text{span}(x_1) \), thus \( \lim_{k \to \infty} \text{dist} \left( v^k, \text{span}(x_1) \right) = 0 \). \( \square \)

**Algorithm 1** Power iteration : \( v = \text{power}(A) \)

1. **Input**: an \( n \times n \) matrix \( A \)
2. \( v^0 \) : randomly chosen
3. **for** \( k = 0, 1, 2, \cdots \) **do**
   1. \( v^{k+1} = \frac{Av^k}{\|Av^k\|} \)
4. **end for**
5. **Output** : \( v = v^k \) for some large \( k \)

The above theorem shows that the sequence of power iteration will eventually belong to \( \text{span}(x_1) \). Let us set a unit eigenvector \( v^1 \in \text{span}(x_1) \) as a vector \( v^k \) in the power iteration applied to \( A \) for some large \( k \). The next theorem suggests a way to calculate the next largest one.
Theorem 2.2. (Deflated Matrix) Let $v_1$ be a unit eigenvector of $A$ associated with $\lambda_1$, then the matrix $(I - v_1 v_1^T) A$, called the deflated matrix of $A$ with $v_1$, has eigenvalues $0, \lambda_2, \cdots, \lambda_n$ with corresponding eigenvectors $v_1, (I - v_1 v_1^T) x_2, \cdots, (I - v_1 v_1^T) x_n$.

Proof. Since $(I - v_1 v_1^T) A v_1 = \lambda_1 v_1 - \lambda_1 v_1 = 0$, 0 is an eigenvalue of the deflated matrix with eigenvector $v_1$. The deflated matrix annihilates the $v_1$ component, hence

$$(I - v_1 v_1^T) A (I - v_1 v_1^T) x_j = (I - v_1 v_1^T) A x_j = \lambda_j (I - v_1 v_1^T) x_j.$$  

Note that whenever $j \neq 1$, $x_j - (v_1 \cdot x_j) v_1$ cannot be a zero vector, otherwise it contradicts the linear independence of the eigenvectors belonging two different eigenvalues. □

Corollary 2.3. Let the power iteration, denoted by $v_2^k$, operate on the deflated matrix $(I - v_1 v_1^T) A$, then

$$\lim_{k \to \infty} \text{dist} \left( v_2^k, \text{span} (x_1, x_2) \right) = 0.$$  

Proof. The largest eigenvalue of the deflated matrix is $\lambda_2$ and its associated eigenvector is $x_2 - (v_1 \cdot x_2) v_1$. From theorem 2.1, as $k \to \infty$, $v_2^k$ belongs to the space $\text{span} (x_2 - (v_1 \cdot x_2) v_1)$ which is the subspace of $\text{span} (x_1, x_2)$. □

Power iteration obtains a unit eigenvector of the eigenvalue with the largest modulus. The deflation annihilates the largest eigenvalue, and exposes the next largest one for the power iteration to pick up. In this way, we set a unit vector $v_2 \in \text{span} (x_1, x_2)$ as $v_2^k$ in the power iteration applied to $(I - v_1 v_1^T) A$ for some large $k$. We repeatedly apply the power iteration on the deflated matrix to obtain unit length vectors $v_1, v_2, \cdots, v_n$ with the properties listed in Algorithm 2.

Algorithm 2 Successive power iterations on deflated matrices

| Input : an $n \times n$ matrix $A$ |
| $v_1 = \text{power} (A)$ |
| $v_2 = \text{power} \left( (I - v_1 v_1^T) A \right)$ |
| $\vdots$ |
| $v_n = \text{power} \left( (I - v_{n-1} v_{n-1}^T) \cdots (I - v_1 v_1^T) A \right)$ |
| Output : $v_1, v_2, \cdots, v_n$ |

Theorem 2.4. The vectors $v_1, v_2, \cdots, v_n$ in Algorithm 2 satisfy

$v_1 \in \text{span} (x_1)$  
$v_2 \in \text{span} (x_1, x_2)$  
$\vdots$  
$v_n \in \text{span} (x_1, x_2, \cdots, x_n)$
Proof. Let $A_i = (I - v_i v_i^T) \cdots (I - v_1 v_1^T) A$ be the $i$th deflated matrix in Algorithm 2. By Theorem 2.2, each deflation annihilates the largest eigenvalue of $A_i$ and adds a scalar multiple of the associated eigenvector to all other eigenvectors. Thus the largest eigenvalue in modulus of $A_i$ is $\lambda_i + 1$, and its associated eigenvector belongs to $\text{span}(v_1, \cdots, v_i, x_{i+1})$, for each $i = 0, 1, \cdots, n - 1$. For the matrix $A_i$, $v_{i+1}$ is an eigenvector associated with the eigenvalue of the largest modulus, and $v_{i+1} \in \text{span}(v_1, \cdots, v_i, x_{i+1})$ for each $i$. By mathematical induction on $i$, it is clear that $v_{i+1} \in \text{span}(x_1, \cdots, x_i, x_{i+1})$ for each $i$.

3. ORTHOGONAL ITERATION

Algorithm 2 sequentially computes all the eigenvectors. Even though the power iteration for $v_i$ is not completed, the temporary value should serve as a good approximation, and Algorithm 3 combines all the power iterations in one iteration by using the approximations $v_1^k, \cdots, v_n^k$ for $v_1, \cdots, v_n$, respectively. The two algorithms are just two formulations of the successive power iterations on deflated matrices. Practically, Algorithm 3 is more efficient than Algorithm 2 in a sense that the former can propose a good approximation for the full eigenvector system in a meantime of the iteration, while the latter cannot.

Algorithm 3 Combined power iterations on recursively deflated matrices

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

$v_{k+1}^1 = Av_1^k$
$v_{k+1}^2 = \left( I - v_1^{k+1} (v_1^{k+1})^T \right) A v_1^k$
$v_{k+1}^3 = \left( I - v_2^{k+1} (v_2^{k+1})^T \right) \cdots \left( I - v_1^{k+1} (v_1^{k+1})^T \right) A v_n^k$

end for

Note that the routine inside Algorithm 3 is nothing but the QR factorization that obtains the orthonormal basis $v_1^{k+1}, \cdots, v_n^{k+1}$ from the basis $Av_1^k, \cdots, A v_n^k$. Writing the vectors in columns, $V^k = [v_1^k, \cdots, v_n^k]$, Algorithm 4, called orthogonal iteration, is hence a simple restatement of Algorithm 3.

Algorithm 4 Orthogonal Iteration

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

$AV^k = V^{k+1} R^{k+1}$: QR factorization of the matrix $AV^k$

end for
Theorem 3.1. Let \((V^k)_{k \in \mathbb{N}}\) be the sequence in the orthogonal iteration, then

- \(\lim_{k \to \infty} \text{span}(v^k_1, \cdots, v^k_j) = \text{span}(x_1, \cdots, x_j)\) for \(j = 1, 2, \cdots, n\)
- \(\lim_{k \to \infty} v^k_i \cdot A v^k_j = 0\) if \(i > j\).

Proof. The orthogonal iteration is an implementation of the successive application of the power iteration on the deflated matrices. By theorem 2.4, \(\lim_{k \to \infty} \text{dist}(v^k_j, \text{span}(x_1, \cdots, x_j))\) for each \(j\), and the first argument follows.

As \(k \to \infty\), \(v^k_j\) belongs to \(\text{span}(x_1, \cdots, x_j)\), and \(A v^k_j\) belongs to the same space, since the eigenvectors are invariant. By the first argument, \(A v^k_j\) belongs to \(\text{span}(v^k_1, \cdots, v^k_j)\) when \(k \to \infty\). In the QR factorization, \(v^k_i\) is orthogonal to \(v^k_j\) whenever \(i > j\), and \(v^k_i \perp \text{span}(v^k_1, \cdots, v^k_j)\). Therefore \(\lim_{k \to \infty} v^k_i \cdot A v^k_j = 0\) whenever \(i > j\). \(\Box\)

4. QR Method

The previous section reveals the relation between the column vectors of the orthogonal iteration and the eigenvectors. To retrieve the eigenvalues of \(A\), let us set \(A^k = (V^k)^T A V^k\). Then Theorem 3.1 states that \(A^k\) becomes upper triangular as \(k \to \infty\). Since \(A^k\) is similar to \(A\), the eigenvalues of \(A\) should appear in the diagonal of the upper triangular matrix. Thus when \(A^k = (V^k)^T A V^k\) is inserted in the orthogonal iteration, the eigenvalues can be obtained. From \(A V^k = V^{k+1} R^{k+1}\),

\[
A^k = (V^k)^T A V^k = (V^k)^T V^{k+1} R^{k+1} = (V^{k+1})^T A V^{k+1} = R^{k+1} (V^k)^T V^{k+1}.
\]

Since \(V^k\) and \(V^{k+1}\) are orthogonal matrices, \((V^k)^T V^{k+1}\) is also orthogonal. Therefore the above equations can be simply written as

\[
A^k = Q^{k+1} R^{k+1},
A^{k+1} = R^{k+1} Q^{k+1},
\]

where \(A^k = Q^{k+1} R^{k+1}\) is the QR factorization of \(A^k\). The above recursive formula shows that the sequence \(A^k\) can be generated detached from the orthogonal iteration. Algorithm 5, called QR method, shows the complete procedure how to find the full eigenvalues of a general matrix. Since \(A^k\) is similar to \(A\), the eigenvalues are preserved each \(k\). By Theorem 3.1, the sequence \(A^k\) will eventually become upper triangular matrix, which reveals the eigenvalues on its diagonal.

Acknowledgments

This research was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD, Basic Research Promotion Fund) (KRF-2008-331-C00045).
Algorithm 5 QR Method

Input: an $n \times n$ matrix $A$

\[ A^0 = A \]

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

\[ A^k = Q^{k+1} R^{k+1} \]

\[ A^{k+1} = R^{k+1} Q^{k+1} \]

end for

Output: an upper triangular matrix $A^k$ for some large $k$

REFERENCES


A NOTE ON OPTIMAL RECONSTRUCTION OF MAGNETIC RESONANCE IMAGES FROM NON-UNIFORM SAMPLES IN $k$-SPACE

JUNE-YUB LEE

DEPARTMENT OF MATHEMATICS, EWHAN WOMANS UNIVERSITY, SEOUL 120-750, SOUTH KOREA
E-mail address: jyllee@ewha.ac.kr

ABSTRACT. A goal of Magnetic Resonance Imaging is reproducing a spatial map of the effective spin density from the measured Fourier coefficients of a specimen. The imaging procedure can be done by inverse Fourier transformation or backward fast Fourier transformation if the data are sampled on a regular grid in frequency space; however, it is still a challenging question how to reconstruct an image from a finite set of Fourier data on irregular points in $k$-space. In this paper, we describe some mathematical and numerical properties of imaging techniques from non-uniform MR data using the pseudo-inverse or the diagonal-inverse weight matrix. This note is written as an easy guide to readers interested in the non-uniform MRI techniques and it basically follows the ideas given in the paper by Greengard-Lee-Inati [10, 11].

1. INTRODUCTION

In Magnetic Resonance Imaging (MRI), the measured signal $s_m=s(t_m)$ induced by the proton density $\rho(r)$ is given by the signal equation

$$s(t_m) = \int_{\text{FOV}} \rho(x) e^{-2\pi i k(t_m) \cdot x} dx$$ (1.1)

where $\rho(x)$ is a $L_2$-function compactly supported on the field-of-view (FOV) and $k(t)$ is a trajectory in the Fourier domain, which can be controlled by the gradient waveforms. The original image can be obtained using an inverse Fourier Transform

$$\rho(x) = \int_{\mathbb{R}^2} \hat{\rho}(k) e^{2\pi i k \cdot x} dk$$ (1.2)

if one has a complete set of data $\hat{\rho}(k)$ covering whole $k$-space. Also a unique reconstructed image in a band-limited function space can be obtained via a backward discrete Fourier transformation if a MR machine is operated by the classical sampling method where $\{s_m\}_{m=1}^M$ are given on a regular grid points in the $k$-space.
For non-uniform sampling patterns, the FFT cannot be directly applied. And the image restoration process becomes finding an inverse of the ill-posed signal equation with the continuous-to-finite Fourier operator, $s = F\rho$. We present two linear reconstruction schemes, $\tilde{\rho} := F^+ W s$ in this paper. The first reconstruction method uses a pseudo-inverse for the weight matrix $W$,

$$\tilde{\rho}(r) := F^+ (FF^+)^+ s.$$  \hspace{1cm} (1.3)

where $+$ denotes a pseudo-inverse of $M \times M$ matrix and the second method does a diagonal weight matrix,

$$\tilde{\rho}(r) = \sum_{m=1}^{M} e^{2\pi i k_m \cdot r} w_m s_m.$$  \hspace{1cm} (1.4)

The purpose of this note is to provide an easy guide to graduate students and researchers who become interested in the non-uniform MRI techniques. A brief introduction to the principles of MRI is given in section 2 and section 3 explains mathematical properties of the continuous-to-discrete Fourier operator. The rest of the paper basically follows the basic ideas of the papers by Greengard-Lee-Inati-et al. [10, 11] but adds mathematical details to them. Section 4 provides a mathematical framework regarding error analysis and section 5 deals with some numerical implementation issues.

2. BASICS OF MAGNETIC RESONANCE IMAGING

Electro-magnetic signal $M(\vec{r}, t)$ generated by hydrogen atoms in a MRI machine is proportional to the proton density $\rho(r)$,

$$M(\vec{r}, t) = M_0 \rho(\vec{r}) e^{-i\omega t}$$ \hspace{1cm} (2.1)

where the Larmor frequency $\omega$ is a linear function of magnetic field strength $H$, $\omega = \gamma H$ with the gyromagnetic ratio $\gamma/2\pi = 42.57$ (MHz/Tesla) for $^1H$ atoms. The existence of the Larmor precession has been predicted by L. Landau and E. Lifshitz in 1935 and a technique known as Nuclear Magnetic Resonance (NMR) spectroscopy has been widely used since 1950s in order to exploit the magnetic properties of certain nuclei by isolating each of frequency components from the total signal captured by an antenna,

$$S(t) = \int_{FOV} M(\vec{r}, t) d\vec{r}$$

where FOV stands for the field of view.

Imaging technique which localizes proton density $\rho(\vec{r})$ in space, however, has not been known until late 1970s. A first successful imaging method known as the Echo-Planar Imaging (EPI) technique has been proposed by P. Lauterbur and P. Mansfield in 1977. A MRI machine implementing the EPI method first excites protons only on $z = z_0$ plane and then varies magnetic field strength $H(\vec{r}, t)$ in time and space, so that the angular velocity of the signal is no longer a constant,

$$M(\vec{r}, t) = M_0 \rho(\vec{r}|z_0) e^{-i\phi(t, \vec{r})}, \quad \frac{d\phi}{dt}(t, \vec{r}) = \gamma H(\vec{r}, t).$$
Suppose that the magnetic field consists of homogeneous background field and time-varying field induced by gradient coils,

$$H(t) = H_0 + \nabla G(t) \cdot \vec{r}. \quad (2.2)$$

Then the phase of the signal is a function of the gradient field \( \nabla G(t) \),

$$\phi(t, \vec{r}) = \gamma H_0 t + \gamma \int_0^t G_x(t')x + G_y(t')y \, dt'$$

and the total signal \( S(t) \) can be written as follows,

$$S(t) = M_0 \int_{\text{FOV}} \rho(\vec{r}|z_0) e^{-i\phi(t, \vec{r})} d\vec{r}$$

$$= M_0 e^{-i\omega_0 t} \int_{\text{FOV}} \rho(\vec{r}|z_0) e^{-i\gamma \int_0^t (G_x(t')x + G_y(t')y) \, dt'} \, dx \, dy,$$

$$= M'_0 \int_{\text{FOV}} \rho(\vec{r}|z_0) e^{-i\vec{k}(t)(x,y)} \, dx \, dy, \quad M'_0 = M_0 e^{-i\omega_0 t}$$

where

$$\vec{k}(t) = \gamma \int_0^t (G_x(t'), G_y(t')) \, dt'.$$  \hspace{1cm} (2.3)

Therefore, the signal \( S(t) \) can be considered as fourier coefficients of \( \rho(\vec{r}|z_0) \) at \( \vec{k}(t) \),

$$S(t) = M'_0 \sum_{m} \rho(\vec{r}|z_0) \left( \tilde{k}(t) \right)$$

and a MRI machine is able to accurately measure the fourier transformation of proton density \( \rho(\vec{r}|z_0) \) at any sampling points along \( k \)-space trajectory by controlling the gradient field \( \nabla G(t) \) in time. Readers interested in principles and histories of MRI, please refer to a classical text book by Liang and Lauterbur [14].

The \( k \)-space trajectory under the classical sampling method, which is commonly used in clinical MRI machines, covers the sampling points on a uniform regular grid. The reconstruction from data on regular grid is rather straightforward even though the data collection step is very slow. It is an active research area introducing a new trajectory on the \( k \)-space [1] with faster acquisition time and reconstructing the original proton density.

### 3. Continuous-to-discrete Fourier transformation

Let \( \rho(x) \) be an \( L_2(\text{FOV}) \) function defined on \( \text{FOV} := (-\frac{1}{2}, \frac{1}{2})^2 \subset \mathbb{R}^2 \), zero outside of the field of view. The task of image reconstruction is to produce an image of \( \rho(x) \) given a finite set of measurements \( \{s_m\}_{m=1}^M \) at the data acquisition points \( k_m \) on the \( k \)-space trajectory from the MRI signal equation,

$$s_m = (\mathcal{F}\rho)_m := \int_{\text{FOV}} \rho(x) e^{-2\pi i k_m \cdot x} \, dx$$  \hspace{1cm} (3.1)
where $\mathcal{F}$ is a continuous-to-discrete linear operator. The problem of inverting this linear system to find the continuous function $\rho(x)$ from the finite number of samples is inherently ill-posed, and restriction on the domain of the operator or some type of regularization is required.

It is worth to remark that there is a notably important case for this operator inversion problem. Suppose $\{k_m\}$ are regular grid points dense enough in $k$-space and $\rho(x)$ is a band-limited $L_2(\text{FOV})$ function, then $\mathcal{F}$ becomes a self-adjoint operator. Thus one can invert the signal equation without any loss using its adjoint

$$\mathcal{F}^* \{v_m\}(x) := \sum_{m=1}^{M} v_m e^{2\pi i k_m \cdot x}$$

(3.2)

or numerically using a discrete back-ward Fourier transformation. However if $\{k_m\}$ are irregularly sampled points, then there is no clear way to define a finite dimensional subspace of $L_2(\text{FOV})$ as a pre-image of the operator $\mathcal{F}$. Such a statement can be considered as a generalization of Shannon’s sampling theorem [19], which is yet to be studied [2].

This inversion process can be considered as an approximation of inverse Fourier transformation

$$\rho(x) = \int_{\mathbb{R}^2} \hat{\rho}(k) e^{2\pi i k \cdot x} dk$$

(3.3)

with finite set of Fourier coefficient data $\{s_m = \hat{\rho}(k_m)\}_{m=1}^{M}$. Since there is no clear way to define a proper domain of the operator $\mathcal{F}$, it is common practice to restrict the domain of $\mathcal{F}$ to the range of its adjoint operator $\mathcal{F}^*$,

$$\hat{\rho}(x) = \mathcal{F}^* v = \sum_{m=1}^{M} v_m e^{2\pi i k_m \cdot x} \in \text{range}(\mathcal{F}^*)$$

(3.4)

It is possible to develop a non-linear scheme which maps the input data $s := \{s_m\}$ to Fourier coefficients of the reconstruction image $v := \{v_m\}$, however, we want to focus our attention in this paper to linear schemes in the following form,

$$\hat{\rho}(x) = \sum_{m=1}^{M} e^{2\pi i k_m \cdot x} (Ws)_m$$

(3.5)

where $W$ is a $M \times M$ matrix. In the following sections, we will discuss mathematical properties of the linear reconstruction scheme with a pseudo-inverse and a diagonal-inverse for the weight matrix, $W$.

4. **Reconstruction Error of Pseudo-inverse and Diagonal-inverse**

Suppose we have a signal $s = \mathcal{F} \rho$ and try to apply an image restoration process defined by weight matrix $W$,

$$\hat{\rho} := \mathcal{F}^*Ws.$$  

(4.1)

A given image function $\rho$ can be uniquely decomposed as

$$\rho = \mathcal{F}^* v + \rho_0 \text{ where } v \in \text{Range}(F), \rho_0 \in \text{Null}(F)$$

(4.2)
and the reconstruction error in image space is
\[
\|\hat{\rho} - \rho\|_{L^2(\text{FOV})}^2 = \|\mathcal{F}^* W s - \mathcal{F}^* v\|_2^2 + \|\rho_0\|_2^2
\]
\[
= \|\mathcal{F}^*(W\mathcal{F}\mathcal{F}^* - I)v\|_2^2 + \|\rho_0\|_2^2.
\]  
(4.3)
Therefore, it is not possible to find a weight matrix \(W\) which matches the reconstructed \(\hat{\rho}\) to the original image \(\rho\) exactly for any \(\rho\) with \(\rho_0(\neq 0) \in \text{Null}(\mathcal{F})\). Instead of measuring the reconstruction error in image space, we measure the reconstruction error in signal space defined as follows
\[
\|\hat{\rho} - \rho\|_F^2 := \|\mathcal{F}(\hat{\rho} - \rho)\|_2^2 = \|\mathcal{F}\mathcal{F}^* W s - s\|_2^2
\]  
(4.4)
and the average reconstruction error for \(W\),
\[
E^2(W) := \frac{\int_{||s||_2=1} \|\mathcal{F}\mathcal{F}^* W s - s\|_2^2 ds}{\int_{||s||_2=1} ds} = \frac{1}{M} \|\mathcal{F}\mathcal{F}^* W - I\|_F^2.
\]  
(4.5)
where \(\|\cdot\|_F\) denotes the Frobenious norm for \(M \times M\) matrix.

A special case is taking \(W\) in the following form,
\[
W = M^+, \quad M_{mn} = (\mathcal{F}\mathcal{F}^*)_{mn} = \int_{\text{FOV}} e^{2\pi i (k_n - k_m) \cdot x} dx = \text{sinc}(k_n - k_m)
\]  
(4.6)
where sinc\((k) = \frac{\sin(\pi k_x)}{\pi k_x} \frac{\sin(\pi k_y)}{\pi k_y}\) and the superscript of \(M^+\) denotes the pseudo-inverse \([8]\) of the sinc-matrix \(M\). Note that the reconstructed pseudo-inverse solution \(\hat{\rho}(x) = \mathcal{F}^* M^+ s\) has the smallest reconstruction error in image space defined in (4.3),
\[
\|\hat{\rho} - \rho\|_{L^2(\text{FOV})}^2 = \|\rho_0\|_2^2 \leq \min_{\hat{\rho}} \{\|\hat{\rho} - \rho\| : \mathcal{F}\hat{\rho} = s\}
\]  
(4.7)
since
\[
(M^+ \mathcal{F}\mathcal{F}^* - I)v = 0 \quad \text{for} \quad v \in \text{Range}(\mathcal{F})
\]  
(4.8)
and the solution also has the minimum \(L_2\)-norm among all functions satisfying the data consistency requirement \(\mathcal{F}\hat{\rho} = s\),
\[
\|\hat{\rho}\|_{L^2(\text{FOV})}^2 = \|\mathcal{F}^* s\|_2^2 \leq \min_{\hat{\rho}} \{\|\hat{\rho}\| : \mathcal{F}\hat{\rho} = s\}.
\]  
(4.9)

The reconstruction error in signal space defined in (4.5) is always zero,
\[
\|\hat{\rho} - \rho\|_F^2 = \|\mathcal{F}\mathcal{F}^* s - s\|_2^2 = 0 \quad \text{for} \quad s \in \text{Range}(\mathcal{F})
\]  
(4.10)
although \(\|\mathcal{F}\mathcal{F}^* - I\|_F^2 > 0\) for a rank deficient matrix \(M\). The discrepancy is caused by the fact that we take the average of reconstruction error on the unit sphere \(\{s : ||s||_2 = 1\}\) which might be bigger than that on the actual signal space \(\{s : s = \mathcal{F}\rho, \rho \in L_2(\text{FOV})\}\).

Another special case of the linear reconstruction method (4.1) is choosing a diagonal weight matrix \(W\) which makes the reconstructed image in the follow form,
\[
(W)_{mn} = w_m \delta_{mn}, \quad \hat{\rho}(r) = \sum_{m=1}^{M} s_m e^{2\pi i k_m \cdot r} w_m.
\]  
(4.11)
where the quadrature weight $\{w_m\}$ should be fixed in order to minimize the average reconstruction error in signal space $E^2(W) = \frac{1}{M} \|M W - I\|_F^2$ defined in (4.5). We differentiate the error $\frac{M}{2} E^2(W)$ with respect to $w_m$,

$$
\frac{1}{2} \frac{\partial}{\partial w_m} \sum_{i,j} |M_{ij} w_j - \delta_{ij}|^2 = \sum_{i,j} (M_{ij}^2 w_j \delta_{jm} - \delta_{ij} M_{ij} \delta_{jm}).
$$

(4.12)

Therefore, $\frac{\partial}{\partial w_m} E^2(W) = 0$ implies, $\sum_i M_{im}^2 w_m = M_{mm}$, or equivalently,

$$
w_m = \frac{1}{\sum_n \text{sinc}^2(k_n - k_m)} \text{ for all } m.
$$

(4.13)

These are the optimum weights for density compensation in gridding reconstruction and the best weights to use in the quadrature reconstruction [3, 11, 18],

$$
\tilde{\rho}(r) = \sum_{m=1}^{M} s_m e^{2\pi i k_m \cdot r}
$$

(4.14)

in the Frobenious norm sense. The average reconstruction error with the optimal weights $\{w_m\}_{m=1}^{M}$ is given as

$$
\|M W - I\|_F^2 = \sum_{n,m} (M_{nm}^2 w_m^2 - 2 \delta_{nm} M_{nm} w_m + \delta_{nm})
$$

$$
= \sum_{m} ((\sum_n M_{nm}) w_m^2 - 2 w_m + 1) = \sum_{m} (1 - w_m).
$$

(4.15)

5. Numerical Implementation Issues

The computation of the sum (3.4)

$$
\tilde{\rho}(x) = \mathcal{F}^* v = \sum_{m=1}^{M} v_m e^{2\pi i k_m \cdot x}
$$

with $O(M)$ frequency data and $O(M)$ target points seems to require $O(M^2)$ operations, however, can be computed in $O(M \log M)$ operations using the nonuniform fast Fourier transform (NUFFT) which is now a relatively mature technology [4, 7, 16]. The evaluation of the reconstructed image on a regular grid points using the frequency data at irregular sampling points is referred as a Type-1 NUFFT in the paper by Greengard and Lee [9, 15].

The values $\{w_m\}$ in (4.11) can be considered quadrature weights, and the computation of all of the optimal weights in (4.13)

$$
\frac{1}{w_m} = \sum_n \text{sinc}^2(k_n - k_m).
$$

appears to require $O(M^2)$ operations, however, the fast $\text{sinc}^2$-transform described in [10] again reduces the computational cost to $O(M \log M)$. Therefore, the image reconstruction based on the optimal quadrature weights

$$
\tilde{\rho}(r) = \sum_{m=1}^{M} e^{2\pi i k_m \cdot r} w_m s_m
$$
can be computed in $O(M \log M)$ cost using a single application of sinc\textsuperscript{2}-transform followed by one Type-1 NUFFT evaluation.

There are many attempts to solve the signal equation (3.1) directly with some additional constrains or regularization techniques [5, 13, 17]. The minimum-norm least-squares solution to this problem, denoted by $\tilde{\rho}(x)$, can be found by applying the pseudo-inverse of the finite-discrete operator $\mathcal{F}$ to the signal. Following the discussion of [8, 20], we write the pseudo-inverse solution $\tilde{\rho}(x)$ defined in (4.6) with the pseudo-inverse of the sinc-matrix,

$$
\tilde{\rho}(r) = \mathcal{F}^+ s = \mathcal{F}^+ \mathcal{M}^+ s, \quad (5.1)
$$

can be computed in two steps,

$$
\mathcal{M}a = s \quad \text{followed by} \quad \tilde{\rho}(r) = \mathcal{F}^* a \quad (5.2)
$$

where the second step can be done by a single application of the Type-1 NUFFT evaluation. The matrix $\mathcal{M}$ is symmetric semi-positive definite, however, may be ill-conditioned [20]. Therefore, computation of

$$
a = \mathcal{M}^+ s \quad (5.3)
$$

can be done using the singular value decomposition (SVD) with some regularization techniques at a cost of $O(M^3)$ work or solved iteratively using the conjugate gradient method as suggested in [5]. With aid of the fast sinc-transform described in [10], an iterative solution of (5.3) can be obtained in $O(J \cdot M^2)$ computational cost where $J$ denotes the number of iterations and $M$ does the number of data points in $k$-space.

Image reconstruction process using either pseudo-inverse or diagonal-inverse requires a Type-1 NUFFT application for the final image generation and both benefit from the fast sinc\textsuperscript{2}-transformation for optimal weights and the fast sinc-transformation for the iterative solution of the signal equation.

REFERENCES


NUMERICAL PROPERTIES OF GAUGE METHOD FOR THE INCOMPRESSIBLE NAVIER-STOKES EQUATIONS

JAE-HONG PYO

DEPARTMENT OF MATHEMATICS, KANGWON NATIONAL UNIVERSITY, SOUTH KOREA
E-mail address: jhpyo@kangwon.ac.kr

ABSTRACT. The representative numerical algorithms to solve the time dependent Navier-Stokes equations are projection type methods. Lots of projection schemes have been developed to find more accurate solutions. But most of projection methods [4, 11] suffer from inconsistency and requesting unknown datum. E and Liu in [5] constructed the gauge method which splits the velocity \( u = a + \nabla \phi \) to make consistent and to replace requesting of the unknown values to known datum of non-physical variables \( a \) and \( \phi \). The errors are evaluated in [9]. But gauge method is not still obvious to find out suitable combination of discrete finite element spaces and to compute boundary derivative of the gauge variable \( \phi \). In this paper, we define 4 gauge algorithms via combining both 2 decomposition operators and 2 boundary conditions. And we derive variational derivative on boundary and analyze numerical results of 4 gauge algorithms in various discrete spaces combinations to search right discrete space relation.

1. INTRODUCTION

Given an open bounded polygon \( \Omega \) in \( \mathbb{R}^d \) with \( d = 2 \) or \( 3 \), we consider the time dependent Navier-Stokes Equations [NSE]:

\[
\begin{align*}
\mathbf{u}_t + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p - \mu \Delta \mathbf{u} &= \mathbf{f}, & \text{in } & \Omega, \\
\mathop{\text{div}}\nabla \mathbf{u} &= 0, & \text{in } & \Omega, \\
\mathbf{u}(\mathbf{x}, 0) &= \mathbf{u}^0, & \text{in } & \Omega,
\end{align*}
\]

(1.1)

with vanishing Dirichlet boundary condition \( \mathbf{u} = 0 \) on \( \partial \Omega \) and pressure mean-value \( \int_\Omega p = 0 \).

The unknowns are vector function \( \mathbf{u} \) (velocity) and the scalar function \( p \) (pressure). And \( \mu = Re^{-1} \) is the reciprocal of the Reynolds number.

Pressure \( p \) can be viewed in (1.1) as a Lagrange multiplier corresponding to the incompressibility condition \( \mathop{\text{div}}\nabla \mathbf{u} = 0 \). This coupling is responsible for compatibility conditions between the spaces for \( \mathbf{u} \) and \( p \), characterized by the celebrated inf-sup condition, and associated numerical difficulties [1, 7]. On the other hand, projection methods were introduced independently

Received by the editors March 6 2010; Accepted March 9 2010.
2000 Mathematics Subject Classification. 65M12, 65M15, 76D05.
Key words and phrases. Projection method, Gauge method, Navier-Stokes equation, incompressible fluids.
The author was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD, Basic Research Promotion Fund) (KRF-2007-331-C00043).
by Chorin [4] and Temam [15, 16] in the late 60’s to decouple $u$ and $p$ and thus reduce the computational cost. However, some projection methods impose artificial boundary and initial conditions on $p$, which leads to boundary layers and reduced convergence rates for $p$ [6, 12]. We will introduce Chorin and Chorin-Uzawa method in §2 to discuss inconsistency and to compare with the gauge method which is studied in §3. Also we collect theoretical estimations which were proved in [2, 6, 14].

E and Liu in [5] introduced the gauge method which splits the velocity $u = a + \nabla \phi$ in terms of non-physical variables $a$ and $\phi$. The gauge method impose initial and boundary condition on gauge variable $\phi$ but pressure $p$. Moreover this scheme doesn’t include inconsistency. So we can say it is more natural method than any other projection type method in PDE level. But troubles are to compute boundary derivation on the discrete space and to find out suitable combinations of discrete finite element space of each function. The former limits their application on 2d and the latter make calculation heavy. The goal of this paper is to implement the gauge method using finite element method without losing advantages of the method. In order to discuss about the drawback of the classical projection method, we introduce Chorin method and Chorin-Uzawa method in §2. And We will construct 4 time discrete gauge algorithms to solve the difficulties on boundary differentiation via using both stream line functions and boundary properties in §3. We construct variational approach of the boundary derivation in §4, and we compute error decay on various discrete space combinations to analyze stability condition among finite element spaces in §5.

In whole this paper, $\nu$ and $\tau$ are the unit vectors in normal and tangential direction, respectively. And $\tau$ designates the time step. And we indicate with $\| \cdot \|$ the norm in $H^s(\Omega)$.

2. Review of Projection Methods

The main strategy of the projection type method is to find an artificial velocity $\tilde{u}$ via solving the momentum equation including transformed pressure term without divergence free constraint. And then we project $\tilde{u}$ to solenoidal space using the following Helmholtz decomposition lemma in [7]:

**Lemma 2.1 (Helmholtz decomposition theorem).** Let

$$H = \{ \nu \in L^2(\Omega)^d : \nabla \cdot \nu = 0 \text{ and } \nu \cdot \nu = 0 \text{ on } \partial \Omega \}.$$ 

Then we have the decomposition

$$L^2(\Omega)^d = H \oplus H^\perp,$$

where $H^\perp$ is defined as

$$H^\perp = \{ \nabla q \in L^2(\Omega) : q \in L^2(\Omega) \}.$$ 

Equivalently, all $\tilde{u} \in L^2(\Omega)^d$ can be written by

$$\tilde{u} = u + \nabla q,$$ (2.1)
where \( u \in H \) and \( \nabla q \in H^\perp \). The classical projection method impose divergence operator in (2.1) to compute \( u \) and \( q \), which is

\[
\triangle q = \text{div} \tilde{u}, \quad \text{in } \Omega,
\]

\[
\partial_\nu q = 0, \quad \text{on } \partial \Omega,
\]

and then we can obtain divergence free velocity via adding known 2 functions

\[
u = \tilde{u} + \nabla q.
\]

We now introduce a classical projection method by Chorin [4, 6, 11]:

**Algorithm 1 (Chorin method).** Start with \( u^0 = u(0) \).

**Step 1: (Momentum equation)** Find \( \tilde{u}^{n+1} \) as the solution of

\[
\frac{\tilde{u}^{n+1} - u^n}{\tau} + (u^n \cdot \nabla)\tilde{u}^{n+1} - \mu \triangle \tilde{u}^{n+1} = f(t_{n+1}), \quad \text{in } \Omega,
\]

\[
\tilde{u}^{n+1} = 0, \quad \text{on } \partial \Omega.
\]

**Step 2: (Projection step)**

\[
\frac{u^{n+1} - \tilde{u}^{n+1}}{\tau} + \nabla p^{n+1} = 0, \quad \text{in } \Omega,
\]

\[
\text{div } u^{n+1} = 0, \quad \text{in } \Omega,
\]

\[
u^{n+1} \cdot \nu = 0, \quad \text{on } \partial \Omega.
\]

In virtue of (2.2) and (2.3), \( \tilde{u}^{n+1} \) in (2.5) can be split into its solenoidal and irrotational parts by solving

\[
\triangle p^{n+1} = \frac{1}{\tau} \text{div } \tilde{u}^{n+1}, \quad \text{in } \Omega,
\]

\[
\partial_\nu p^{n+1} = 0, \quad \text{on } \partial \Omega,
\]

and then adding 2 functions

\[
u^{n+1} = \tilde{u}^{n+1} - \tau \nabla p^{n+1}.
\]

**Remark 2.2 (Artificial boundary condition).** In the view of \( u^{n+1} \cdot \nu = 0 \), pressure \( p \) automatically satisfies the non-physical Neumann boundary condition \( \partial_\nu p^{n+1} = 0 \) on \( \partial \Omega \). This artificial boundary condition is responsible for a non-physical boundary layer for \( p \).

**Remark 2.3 (Inconsistency).** Upon plugging (2.6) into (2.4), we also discover an inconsistency in the momentum equation

\[
\frac{u^{n+1} - u^n}{\tau} + (u^n \cdot \nabla)u^{n+1} - \mu \triangle u^{n+1} + \nabla p^{n+1} - \mu \tau \nabla \nabla p^{n+1} = f(t_{n+1}),
\]

where \( -\mu \tau \nabla \nabla p^{n+1} \) is the inconsistent term.
There are several publications concerning error estimates for Chorin Algorithm 1. The most relevant for us is Prohl [11] who employs a variational approach with some reasonable assumptions. If \( \sigma(t) = \min\{t, 1\} \), then
\[
|u(t^{n+1}) - u^n|_0 + \sigma(t^{n+1})|p(t^{n+1}) - p^{n+1}|_1 \leq C\tau,
\]
\[
|u(t^{n+1}) - u^n|_1 + \sqrt{\sigma(t^{n+1})}|p(t^{n+1}) - p^{n+1}|_0 \leq C\sqrt{\tau}.
\]

The second paper of interest [6] is by E and Liu, who derive error estimates via an asymptotic expansion approach: If the exact solution \((u(t), p(t))\) of (1.1) is smooth enough, then
\[
|u(t^{n+1}) - u^n|_0 + \sqrt{\tau}|p(t^{n+1}) - p^{n+1}|_0 \leq C\tau.
\]

This result requires regularity which is often not valid for realistic incompressible flows.

One of the famous projection method is the Chorin-Uzawa method which has been introduced by Prohl in [11] to get rid of the boundary layer and inconsistency of Chorin method:

**Algorithm 2 (Chorin-Uzawa method).** Start with given data \((u^0, p^0, \tilde{p}^0)\) such that
\[
|u(0) - u^0|_0 + \sqrt{\tau}|p(0) - p^0|_0 \leq C\tau, \quad \tilde{p}^0 = 0.
\]

**Step 1:** *(Momentum equation)* Find \(\tilde{u}^{n+1}\) as the solution of
\[
\frac{\tilde{u}^{n+1} - u^n}{\tau} + (u^n \cdot \nabla)\tilde{u}^{n+1} - \mu \Delta \tilde{u}^{n+1} + \nabla (p^n - \tilde{p}^n) = f(t_{n+1}), \quad \text{in } \Omega,
\]
\[
\tilde{u}^{n+1} = 0, \quad \text{on } \partial\Omega.
\]

**Step 2:** *(Projection step)*
\[
\frac{u^{n+1} - \tilde{u}^{n+1}}{\tau} + \nabla p^{n+1} = 0, \quad \text{in } \Omega,
\]
\[
\text{div } u^{n+1} = 0, \quad \text{on } \Omega.
\]
\[
\tilde{u}^{n+1} = 0, \quad \text{on } \partial\Omega.
\]

**Step 3:** *(Pressure step)*
\[
p^{n+1} = p^n - \alpha \mu \text{div } \tilde{u}^{n+1}, \quad 0 < \alpha < 1.
\]

The Chorin-Uzawa method is a combination of Chorin Algorithm 1 and Uzawa Algorithm which is an iterative solver of the stationary Stokes equations [1, 7, 8]. The condition of relaxation parameter, \(0 < \alpha < 1\), is necessitated to prove convergence of the Uzawa algorithm, but it is proved that the optimal \(\alpha = 1\) and that its convergence range is \(0 < \alpha < 2\) in [8]. So \(\alpha\) can be chosen as 1 simply.

In the projection step (2.9), we split \(\tilde{u}^{n+1}\) into \(u^{n+1}\) and \(\nabla \tilde{p}^{n+1}\) by the same manner with (2.2) and (2.3). Note the presence of the auxiliary pressure \(\tilde{p}^n\) with artificial boundary value \(\partial_{\nu} \tilde{p}^n = 0\) in (2.9). No boundary condition is imposed on pressure \(p^n\) any longer. Regardless of this improvement, Chorin-Uzawa exhibits the following pitfalls:
Remark 2.4 (Initial pressure). The initial value \( p^0 \) cannot be chosen arbitrarily, because of the initial condition \( \| p(0) - p^0 \| \) in (2.7). So it requires estimating initial pressure or choosing small time distance \( \tau \) at initial steps.

Remark 2.5 (Inconsistency). Upon plugging \( \tilde{u}^{n+1} = u^{n+1} + \tau \nabla \tilde{p}^{n+1} \) from (2.9) into (2.8), we see that

\[
\frac{u^{n+1} - u^n}{\tau} + (u^n \cdot \nabla) \tilde{u}^{n+1} + \nabla \left( p^n - \alpha \mu \tau \Delta \tilde{p}^{n+1} \right) - \mu \Delta u^{n+1} \]

\[
+ (\alpha - 1) \mu \tau \nabla \Delta \tilde{p}^{n+1} + \nabla \left( \tilde{p}^{n+1} - \bar{p}^n \right) = f(t_{n+1}).
\]

Since (2.9) and (2.10) imply \( \bar{p}^{n+1} = p^n - \alpha \mu \tau \Delta \tilde{p}^{n+1} \), we end up with

\[
\frac{u^{n+1} - u^n}{\tau} + (u^n \cdot \nabla) \tilde{u}^{n+1} + \nabla \bar{p}^{n+1} - \mu \Delta u^{n+1} \]

\[
+ (\alpha - 1) \mu \tau \nabla \Delta \tilde{p}^{n+1} + \nabla \left( \tilde{p}^{n+1} - \bar{p}^n \right) = f(t_{n+1}).
\]

Here \( (\alpha - 1) \mu \tau \nabla \Delta \tilde{p}^{n+1} + \nabla \left( \tilde{p}^{n+1} - \bar{p}^n \right) \) are the inconsistency terms. If we choose \( \alpha = 1 \), then the first term disappears but the second term still remained.

The following a priori error bound is stated by Prohl [11]:

\[
\left\| u(t^{n+1}) - u^{n+1} \right\|_1 + \sqrt{\tau} \left\| p(t^{n+1}) - p^{n+1} \right\|_0 \leq C \tau.
\]

3. GAUGE METHODS

As we reviewed in §2, projection methods suffer from inconsistencies. E and Liu in [5] construct gauge method which is a consist projection type method hiring (2.2) and (2.3) which is called divergence operator:

**Algorithm 3** (Gauge method with div operator and Neumann boundary condition). Start with initial values \( \phi^0 = 0 \) and \( a^0 = u^0 = u(x, 0) \).

**Step 1:** Find \( a^{n+1} \) as the solution of

\[
\frac{a^{n+1} - a^n}{\tau} + (u^n \cdot \nabla) u^n - \mu \Delta a^{n+1} = f(t_{n+1}), \quad \text{in } \Omega,
\]

\[
a^{n+1} \cdot \nu = 0, \quad a^{n+1} \cdot \tau = -\partial_{\nu} \phi^n, \quad \text{on } \partial \Omega.
\]

**Step 2:** Find \( \phi^{n+1} \) as the solution of

\[
-\Delta \phi^{n+1} = \text{div } a^{n+1}, \quad \text{in } \Omega,
\]

\[
\partial_{\nu} \phi^{n+1} = 0, \quad \text{on } \partial \Omega,
\]

**Step 3:** Find

\[
u^{n+1} = a^{n+1} + \nabla \phi^{n+1}, \quad \text{in } \Omega.
\]
One may compute the pressure whenever necessary as
\[ p^{n+1} = -\frac{\phi^{n+1} - \phi^n}{\tau} + \mu \Delta \phi^{n+1}. \] (3.2)

We can check easily that Algorithm 3 consists to (1.1), and a priori bound for Algorithm 3 is proved in [9]:
\[ \tau \sum_{n=0}^{N} \left( \| u(t^{n+1}) - u^n \|^2_0 + \tau \| p(t^{n+1}) - p^n \| \right) \leq C \tau^2. \] (3.3)

But one difficulty in implementation is to compute boundary differentiation in (3.1). We will discuss about the variational calculation on boundary in §4. On the other hand, we can avoid the difficult boundary differentiation, provided we know \( \nabla \phi^n \) in (3.1). So we can consider to use stream line function in [7] instead of using (2.2) and (2.3).

**Lemma 3.1 (Stream line function).** A function \( v \) is in 2-dimension \( H \) if and only if there exists a stream function \( \psi \in H^1(\Omega) \) such that
\[ u = \text{curl} \ \psi. \] (3.4)

Since \( u \) is in \( H \), there exists a stream function (3.4). And the stream function hold Dirichlet boundary condition because of \( u \cdot \nu = 0 \). Owing Lemma 2.1, \( a \) can be rewritten by
\[ a = u - \nabla \phi = \text{curl} \ \psi - \nabla \phi. \] (3.5)

If we impose rot in (3.5), then we arrive at
\[-\Delta \psi = \text{rot} \ a, \quad \text{in } \Omega, \]
\[ \psi = 0, \quad \text{on } \partial \Omega, \]
We now have \( \psi \) and easily obtain \( u \) by solving (3.4) and then get \( \nabla \phi \) by computing
\[ \nabla \phi = u - a. \]
So we do not need to compute the boundary derivative in (3.1), because we know \( \nabla \phi \) already. Finally, we are ready to define the gauge method to using rotational operator.

**Algorithm 4 (Gauge method with rot operator and Neumann boundary condition).** Start with initial values \( \phi^0 = 0 \) and \( a^0 = u^0 = u(x, 0) \).

**Step 1:** Find \( a^{n+1} \) as the solution of (3.1)

**Step 2:** Find \( \psi^{n+1} \) as the solution of
\[ -\Delta \psi^{n+1} = \text{rot} \ a^{n+1}, \quad \text{in } \Omega, \]
\[ \psi^{n+1} = 0, \quad \text{on } \partial \Omega. \]

**Step 3:** Find
\[ u^{n+1} = \text{curl} \ \psi^{n+1}, \quad \text{in } \Omega, \]
\[ \nabla \phi^{n+1} = u^{n+1} - a^{n+1}, \quad \text{in } \Omega. \]
We can compute pressure via (3.2) whenever necessary. Also we can obtain a priori error bound (3.3) for Algorithm 4 because it is equivalent to Algorithm 3.

In the view of (3.5) and \( \mathbf{u} = 0 \) on \( \partial \Omega \), we can also take boundary condition \( \phi = 0 \) and \( \mathbf{a} = -\partial_{\nu} \phi \). Therefore we can define a gauge method via imposing \( \text{div} \) operator and Dirichlet boundary condition:

**Algorithm 5** (Gauge method with \( \text{div} \) operator and Dirichlet boundary condition). *Start with initial values* \( \phi^0 = 0 \) and \( \mathbf{a}^0 = \mathbf{u}^0 = \mathbf{u}(\mathbf{x}, 0) \).

**Step 1:** Find \( \mathbf{a}^{n+1} \) as the solution of
\[
\frac{\mathbf{a}^{n+1} - \mathbf{a}^n}{\tau} + (\mathbf{u}^n \cdot \nabla)\mathbf{u}^n - \mu \nabla \mathbf{a}^{n+1} = f(t_{n+1}), \quad \text{in } \Omega,
\]
\[
\mathbf{a}^{n+1} \cdot \nu = -\partial_{\nu} \phi^n, \quad \mathbf{a}^{n+1} \cdot \tau = 0, \quad \text{on } \partial \Omega.
\] (3.6)

**Step 2:** Find \( \phi^{n+1} \) as the solution of
\[
-\Delta \phi^{n+1} = \text{div} \mathbf{a}^{n+1}, \quad \text{in } \Omega,
\]
\[
\phi^{n+1} = 0, \quad \text{on } \partial \Omega.
\]

**Step 3:** Find
\[
\mathbf{u}^{n+1} = \mathbf{a}^{n+1} + \nabla \phi^{n+1}, \quad \text{in } \Omega.
\]

The following a priori bound for Algorithm 5 which is imposed Dirichlet boundary condition is proved in [9]
\[
\tau \sum_{n=0}^N \| \mathbf{u}(t_{n+1}) - \mathbf{u}^{n+1} \|_0^2 \leq C \tau.
\] (3.7)

Also we define a gauge method with rotational operator and with Dirichlet boundary condition.

**Algorithm 6** (Gauge method with \( \text{rot} \) operator and Dirichlet boundary condition). *Start with initial values* \( \phi^0 = 0 \) and \( \mathbf{a}^0 = \mathbf{u}^0 = \mathbf{u}(\mathbf{x}, 0) \).

**Step 1:** Find \( \mathbf{a}^{n+1} \) as the solution of (3.6).

**Step 2:** Find \( \psi^{n+1} \) as the solution of
\[
-\Delta \psi^{n+1} = \text{rot} \mathbf{a}^{n+1}, \quad \text{in } \Omega,
\]
\[
\partial_{\nu} \psi^{n+1} = 0, \quad \text{on } \partial \Omega.
\]

**Step 3:** Find
\[
\mathbf{u}^{n+1} = \text{curl} \psi^{n+1}, \quad \text{in } \Omega,
\]
\[
\nabla \phi^{n+1} = \mathbf{u}^{n+1} - \mathbf{a}^{n+1}, \quad \text{in } \Omega.
\]

We can impose the error bound (3.7) for Algorithm 6 because of equivalent to Algorithm 5.

**Remark 3.2** (Consistency). *We can see consistency easily by inserting* \( \mathbf{a}^{n+1} = \mathbf{u}^{n+1} - \nabla \phi^{n+1} \) *and pressure equation* (3.2) *into the gauge discrete momentum equation* (3.1) *or* (3.6).
Remark 3.3 (Compatibility condition). We uncover $\phi^n$ for Algorithms 5-6 doesn’t satisfy compatibility condition in [9]. Since pressure $p^n$ and $\phi^n$ are linked via (3.2), we cannot expect convergence of $p^n$ to $p(t^{n+1})$.

Remark 3.4 (Application in 3 dimension). Since the rot operator can be defined on only 2 dimension, Algorithms 4-6 are not applicable on 3 dimension. So the only Algorithm 3 is applicable to compute pressure on 3 dimension in conjunction with above Remark 3.3.

To apply gauge algorithms on finite element method, we consider stability relation between space of each variable.

Remark 3.5 (Finite element space stability). Since the gradient of $\phi$ is the addition of $a^{n+1}$ and $u^{n+1}$, we can expect that the gauge variable $\phi$ is necessary in one higher degree space than those of $a^{n+1}$ and $u^{n+1}$. For example, if we consider Taylor-Hood family which is degree 2 for velocity and degree 1 for pressure, then degree 3 is required for $\phi$. But this combination requests too heavy computation by hiring high resolution for non-concerning variable $\phi$. It will be examined in §5 by comparing numerical results on several combinations.

4. VARIATIONAL APPROACH TO COMPUTE DIFFERENTIATION ON BOUNDARY

A key difficulty in actual computations with gauge methods is to provide accurate approximation of boundary derivatives $\partial_\nu \phi^n + 1$ or $\partial_\tau \phi^n + 1$ on $\partial \Omega$. We recall now a variational approximation of boundary derivatives. First we consider the Laplace equation

$$
-\Delta \phi = f, \quad \text{in } \Omega,
\phi = 0, \quad \text{on } \partial \Omega,
$$

(4.1)

and approximation of the normal derivative $\partial_\nu \phi$. Integrating (4.1) by parts against $\psi \in H^1(\Omega)$, we find the variational expression

$$
- \int_{\Omega} \Delta \phi \psi d\mathbf{x} = - \int_{\partial \Omega} \partial_\nu \phi \psi d\Gamma + \int_{\Omega} \nabla \phi \nabla \psi d\mathbf{x}
\text{or}
\int_{\partial \Omega} \partial_\nu \phi \psi d\Gamma = - \int_{\Omega} f \psi d\mathbf{x} + \int_{\Omega} \nabla \phi \nabla \psi d\mathbf{x}.
$$

(4.2)

where the unit normal $\nu$ is well defined except at corners. Equality (4.2) defines $\partial_\nu \phi \in H^{-1/2}(\partial \Omega)$ uniquely as a linear functional in $H^{1/2}(\partial \Omega)$ (Trace space of $H^1(\Omega)$). One goal is to use a similar expression to defined the discrete counterpart. To this end, we follow Pehlivanov et al [3, 10]. The first issue is the concept of normal derivative, at a corner. Since $\phi = 0$ on $\partial \Omega$, the tangential derivatives vanish, and so does $\nabla \phi$, at a corner (see (a) in Figure 1). We thus impose

$$
\partial_\nu \phi = 0 \quad \text{at corners of } \partial \Omega.
$$

Let $\mathcal{T} = K$ be a shape-regular quasi-uniform partition of $\Omega$. Let $\mathbb{B}_h$ be a conforming finite element space containing piecewise linear and let $\mathbb{B}^b_h$ be the boundary finite element space

$$
\mathbb{B}^b_h = \{ w_h \in \mathbb{B}_h : w_h = 0 \text{ at the interior and corner nodes of } \Omega \}. 
$$
We also define
\[ B^0_h = \{ w_h : w_h \in H^1_0(\Omega) \} \]
Let \( \phi_h \in B^0_h \) be the finite element solution of (4.1), namely,
\[ \phi_h \in B^0_h : \int_\Omega \nabla \phi_h \nabla \psi_h dx = \int_\Omega f \psi_h dx, \quad \forall \psi_h \in B^0_h. \]
In view of (4.2), we define the approximate normal derivative \( \partial_\nu \phi_h \) to be:
\[ \phi_h \in B^0_h : \int_{\partial \Omega} \partial_\nu \phi_h \psi_h d\Gamma = -\int_\Omega \text{div} \, a \phi_h \psi_h dx + \int_\Omega \nabla \phi_h \nabla \psi_h dx, \quad \forall \psi_h \in B^b_h. \]
The following lemma was proved in [10]

**Lemma 4.1.** If \( f \in H^2(\Omega) \) and \( \phi \in H^3(\Omega) \), then
\[ \| \partial_\nu \phi - \partial_\nu \phi_h \|_{0,\Gamma} \leq Ch^{\frac{3}{2}}(\| \phi \|_{3,\Omega} + \| f \|_{2,\Omega}). \]
Thus derivative \( \partial_\nu \phi^n_h \) can be calculated by the variational formula:
\[ \partial_\nu \phi^n_h \in B^0_h : \int_{\partial \Omega} \partial_\nu \phi^n_h \psi_h d\Gamma = -\int_\Omega \text{div} \, a \phi^n_h \psi_h dx + \int_\Omega \nabla \phi^n_h \nabla \psi_h dx, \quad \forall \psi_h \in B^b_h. \]
Now we consider the approximation of tangential derivative \( \partial_\tau \phi \) on \( \partial \Omega \) provided \( \phi \) does no longer vanish on \( \partial \Omega \). Integration by parts of \( -\Delta \phi = f \) yields for all \( \psi \in H^1(\Omega) \)
\[ \int_{\partial \Omega} \partial_\tau \phi \psi d\Gamma = \int_\Omega \nabla \phi \text{curl} \psi dx. \quad (4.3) \]
If \( \phi \) satisfies Neumann boundary condition, the tangential derivative on each corner is 0 by (b) in Figure 1. If \( \phi^n \in \mathbb{B}_h \) is the finite element approximation in gauge Algorithms 3-6, then \( \phi^n_r = \phi^n_t = 0 \) and the discrete of (4.3) reads:

\[
\int_{\partial \Omega} \partial_t \phi^n_r h \psi_h d\Gamma = \int_{\Omega} \nabla \phi^n_r \text{curl} \psi_h dx, \quad \forall \psi_h \in H^1(\Omega).
\]  

(4.4)

Formula (4.4) can be used to approximate \( \partial_t \phi^n_r \) in 2d. However, we have 2 orthogonal tangential differentiations in 3d, and we can calculate \( \mathbf{a}^{n+1} \cdot \mathbf{v} = 0 \).

5. Numerical Experiments

We, in this section, analyze and compare numerical results of projection methods which are Chorin, Chorin-Uzawa, gauge, and Gauge-Uzawa methods, with both smooth and singular solutions. The first Experiment comes from multiplication time function \( \cos(t) \) and the example of Prohl in [11]: the computational domain is \( \Omega = [0, 1] \times [0, 1] \) and \( \mu = 1 \). We choose the following exact solution of (1.1) and determine the corresponding force term \( \mathbf{f} \)

\[
\begin{align*}
    u(x, y, t) &= \cos(t)(x^2 - 2x^3 + x^4)(2y - 6y^2 + 4y^3) \\
    v(x, y, t) &= -\cos(t)(y^2 - 2y^3 + y^4)(2x - 6x^2 + 4x^3) \\
    p(x, y, t) &= \cos(t) \left( x^2 + y^2 - \frac{2}{3} \right).
\end{align*}
\]

Remark 5.1 (Distorted mesh). In order to avoid super convergence due to mesh uniformity and symmetry, we choose the distorted quasi-uniform mesh Figure 2. Mesh distortion is crucial to uncover numerical difficulties that may go unnoticed otherwise. For instance, Gauge method is insensitive to the discrete inf-sup condition for uniform mesh.

![Figure 2. The Computational mesh for experiments.](image)
following combinations of discretization parameters and polynomial degrees ($K_v =$ polynomial degree of velocity, $K_p =$ polynomial degree of pressure, and $K_\phi =$ polynomial degree of $\phi$):

**Combination 1:** $K_v = 1$, $K_p = 1$, $K_\phi = 1$.

**Combination 2:** $K_v = 2$, $K_p = 1$, $K_\phi = 1$.

**Combination 3:** $K_v = 1$, $K_p = 1$, $K_\phi = 2$.

**Combination 4:** $K_v = 2$, $K_p = 1$, $K_\phi = 2$.

**Combination 5:** $K_v = 2$, $K_p = 1$, $K_\phi = 3$.

The gauge methods show different dependence on these combinations. As we know, the finite element spaces of Combinations 2 and 4-5 correspond to the Taylor-Hood family $P^2 - P^1$ ($K_v = 2$, $K_p = 1$) which satisfies the discrete inf-sup condition. In contrast, the finite element pairs $P_1 - P_1$ ($K_v = 1$, $K_p = 1$) of Combinations 1 and 3 do not satisfy the discrete inf-sup condition. Since gauge Algorithms 3-6 have convergence order 1, we compute with relation $\tau = h^2$ to get same order 2 for space and time contributions. If $\tau = h$ and the space errors are $O(h^{\kappa+1})$ with $\kappa \geq 1$, then the time error $O(\tau)$ dominates the calculation. All numerical results in this paper are computed by ALBERT which is a finite element toolbox [13].

We note that Algorithms 5-6 do not necessary the error of pressure to decrease (see Remark 3.2). We use a pair $(\alpha, \beta)$, where $\alpha$ and $\beta$ are convergence orders in $L^2(\Omega)$ and $L^\infty(\Omega)$ spaces, respectively. We first consider Figure 3 which is error decay of Combination 1. As we know that Combination 1 doesn’t satisfy inf-sup condition, so pressure $p^{n+1}$ doesn’t need to converge $p(t^{n+1})$. But the pressure in Algorithm 4 has (1.6, 1.0) convergence order and the velocity has

![Figure 3. Error decay of gauge methods with $\tau = h^2$ and $P_1 - P_1 - P_1$ elements.](image-url)
convergence order \((2.0, 1.0)\). The reason of losing order in \(L^\infty(\Omega)\) space seems like due to the broken inf-sup condition. Combinations 2 satisfies inf-sup condition, but pressure in \(L^\infty(\Omega)\) doesn’t converge to exact solution for all gauge methods in Figure 4. Moreover velocity has only order \((1.0, 1.0)\). So we conclude it is not stable combination. We now see Figures 5-7. The velocities of Algorithms 3-4 have convergence order \((2.0, 2.0)\) which is optimal error decay. But pressures of both algorithms have order \((1.5, 1.0)\) for Combination 3, \((2.0, 1.8)\) for Combinations 4-5. The size of errors are also similar for both Combinations 4-5, but Combination 5 requests much higher computational cost than Combination 4. So we conclude, in this experiment, Combination 4 are the best family to imply Algorithms 3-4. Also we note Combination 3 is also acceptable family, if we have relatively big tolerance for pressure. In other word, Algorithms 5-6 have order \((2.0, 1.0)\) for velocity and not computable for pressure on Combinations 3-5.

![Figures 4](image.png)

**Figure 4.** Error decay of gauge methods with \(\tau = h^2\) and \(P_2 - P_1 - P_1\) elements.

**References**

Figure 5. Error decay of gauge methods with $\tau = h^2$ and $P_1 - P_1 - P_2$ elements.

Figure 6. Error decay of gauge methods with $\tau = h^2$ and $P_2 - P_1 - P_2$ elements.

Figure 7. Error decay of gauge methods with $\tau = h^2$ and $P_2 - P_1 - P_3$ elements.

ABSTRACT. In this paper, we develop the Automatic Number Plate Recognition (ANPR) System. ANPR is generally composed of the following four steps: i) The acquisition of the image; ii) The extraction of the region of the number plate; iii) The partition of the number and iv) The recognition. The second and third steps incorporate image processing technique. We propose to resolve this by using Partial Differential Equation (PDE) based segmentation method. This method is computationally efficient and robust. Results indicate that our methods are capable to recognize the plate number on difficult situations.

1. INTRODUCTION

A rapid motorization by an economic growth has brought not only positive aspects including convenience of living but also many negative aspects. The number of cars became increasingly faster than the extension of roads and parking spaces. This resulted in a traffic congestion, growth of the rate of car accidents, violation of the traffic regulations such as speeding and illegal parking, and a crime like car theft. To resolve these problems, we need the intelligent Transport System (ITS) to improve the environment and security of transportation by building intelligent system which can effectively operate transportation. The car recognition system is one of the essential components which are needed for the construction of the ITS. This is the system to intelligently obtain and analyze cars’ information through automatically classifying and recognizing features, shapes and number plates. Among various car’s information obtained by the car recognition system, the car number plate is the most distinguishing element which assigns a distinct identity to each car. For this reason, the Automatic Number Plate Recognition (ANPR) is the most fundamental requisite for the car recognition system. Figure 1 shows the general ANPR system which is composed of the following four steps.

\begin{enumerate}
  \item[i)] The acquisition of the image
  
  We obtain a car image from a high resolution camera or CCTV. The image deteriorates the resolution quality because of equipments and environmental influences.
\end{enumerate}
ii) The extraction of the region of the number plate
In this step, we use the intensity and the property of the location of the number plate considering the front or back sides of the car in the image.

iii) The partition of the number
The image processing is applied to separate the region of objects (letters and numbers) and backgrounds.

iv) The recognition
We recognize letters and numbers by the method of the pattern recognition from the partitioned numbers.

Figure 1. The outline of the Automatic Number Plate Recognition (ANPR).

ANPR could be used in various traffic and security applications, such as automatic charging system, automatic access and border control system, tracking of regulation-violating car, a criminal investigation and an intelligent traffic control system. We can expect the effect reducing costs and saving human resources. In spite of these advantages, ANPR system has been used only in the specific fields. This is, because the present technology can be applied to the high quality image. To get the high quality image, we need high-priced image equipments and the broadband wire communication network. Ultimately, it increases the investment-cost of facilities and equipments. Especially, most of investment-costs of extending the ANPR system is the cost of building the wire communication network to transmit the image data stably. The more insufficient the infra is, such as the road in rural areas and a detached house, the higher the investment-cost is.

If we transmit the acquired image data from image-equipment through wireless communication network, we can save the investment-cost for constructing the wire communication network. However, there are several problems for building ANPR system using the wireless communication network. First of all, it is difficult to send the high resolution data through the
wireless communication network in real time. This is, because the wire communication network transmits the data using the wave. Wireless communication network has lower bandwidth and higher possibility contaminated with noise and interference than wire communication network. And if we use the wave in high frequency, we can send more data and transmit-distance can be longer. But the wave is weakened by the obstacles between a transmitting station and a receiving station having the lower diffractive property. So we need to use the suitable bandwidth, eventually, the data amount we can transmit are limited. There are two approaches to solve this problem. One is developing the wireless communication technology to transmit the high quality image data in real time. And the other is that we transmit the low quality data using the present wireless communication technology and equipment and then, make it the high quality data using the image processing technique properly. If, by using image processing technique, we can acquire the image data up to the level we can analysis, we don’t need to try to get the high quality data and transmit it as it is. After we acquire image data using low resolution equipment, we transmit low quality data through wireless communication network. So after restoring the high quality data from the low ones using the image processing, we can build the system to analyze for the needed information. The above system saves the costs of constructing ANPR system and eventually, ANPR system can be applied to various application fields.

In this paper, we propose the method to improve the image quality. This method makes ANPR system more efficient. Using the low resolution equipment such as CCTV for road traffic information and crime prevention, and wireless communication network we can construct ANPR system that is useful and robust in the environment condition, such as light and atmosphere.

2. The Automatic Number Plate Recognition (ANPR)

The algorithm of the number plate recognition should not be sensitive to noise and variation because the obtained image is affected by environment. However, it is less complicated than the general recognition of images and letters since the number plate has the standard sort and size. We implemented for new cars' number plates among revised number plates of a private car (standard) but our experiments could be applied to other number plates. The size of number plate is 110 mm by 520 mm, so the proportion of length and width is about 1:5 and it plays an important role in the procedure to extract the region of number plate.

2.1. The preprocessing. In our experiment, we first dealt with images without noise and will expand this conduction for noisy images later. When we preprocess the noisy image through using filters such as mean filter, median filter, and Gaussian filter to remove noise, the image rather deteriorate due to blurring effects. Later, we want to devise a method to minimize these preprocessing so that the noisy image doesn’t worsen. We expect to resolve this by using Partial Differential Equation (PDE) based segmentation method.

2.2. The extraction of vertical components. An ordinary car image rarely has vertical components except the region of the number plate. So we take the convolution of the following
mask M and the image to get only vertical components in the original image. Figure 3 shows the binary image obtained from gray-scale threshold for the image after the convolution.

2.3. **The projections of x and y axis.** Except in the case of images having many decorations around the number plate or complex designs, the region of the number plate can be easily extracted by the projections of x and y axis. Figure 5 shows the projections of x and y axis for the binary image. As you can see in this figure, the maxima of the red graph which represent the projection of y axis is located near the number plate.
2.4. **The extraction of the region of the number plate.** This is the most important part in the process of the number plate recognition. To extract the region of the number plate, we sort differential projected values of y axis in descending order. We choose the largest two sorted values and the coordinates are assigned to the criteria \( y_1 \), \( y_2 \) of y axis. For the projection of x axis, we carry out in similar ways. When we choose \( x_1 \) as the criterion of x axis according to the rule that difference of large \( x_1 \) and its adjacent values is greater than 6 times \( |y_1 - y_2| \) and less than 7 times \( |y_1 - y_2| \). Since \( y_1, y_2 \) are very close to numbers in the number plate, \( |y_1 - y_2| \) is less than the vertical length of the number plate. From this and the information that the proportion of length and width of the number plate is about 1:5, we can choose \( x_1 \) and \( x_2 = x_1 + 6 \cdot \text{vertical length} \). Figure 6 shows the extracted number plate by \( x_1, x_2, y_1, \) and \( y_2 \). As stated above, the information about the size of the number plate is used. Since the extraction method like this is very sensitive to the image, if the image has the complex background near the number plate, we cannot correctly extract the desired region. Therefore we want to devise the method which is less sensitive to the image by mathematical modeling.

2.5. **The segmentation of numbers.** From this step, we progress only for the small image which is obtained after the extraction procedure. Without any preprocessing like threshold, we
segment the image by using Gibou-Fedkiw algorithm. This algorithm is considered as 2-means clustering method by using level set function with $c_p$, $c_n$, where $c_p$ is the average of positive level set values and $c_n$ is the average of negative level set values. If we let $\phi$ be a level set function and $I$ be the image which represent only the region of the number plate, the idea of Gibou-Fedkiw is as follows.


c_p(t) = \text{average of } \{I(x,y) : \phi(t,x,y) > 0\}

c_n(t) = \text{average of } \{I(x,y) : \phi(t,x,y) < 0\}

\frac{\partial \phi}{\partial t}(t,x,y) = -\lambda_1[I(x,y) - c_p(t)]^2 + \lambda_2[I(x,y) - c_n(t)]^2

(2.1)

The Ordinary Differential Equation can be simply solved through the following algorithm.

**Algorithm**

i) Set initial $\phi = \pm 1$.

ii) $c_p = \frac{\Sigma I(\phi+1)}{\Sigma(\phi+1)}$, $c_n = \frac{\Sigma I(\phi-1)}{\Sigma(\phi-1)}$.

iii) $F = -[I(x,y) - c_p(t)]^2 + [I(x,y) - c_n(t)]^2$

$\phi = 1$ if $F > 0$ or $\phi = -1$ if $F < 0$.

iv) Repeat ii) and iii) until $c_p$ and $c_n$ are not changed.

In Figure 7, the red line in the first image denotes zero-level set of the initial level set function and the second image shows the level set values $\phi = \pm 1$ used in the first image.

Figure 8 shows the zero-level set and the level set function after applying the Gibou-Fedkiw algorithm.
2.6. **The extraction of the region of the each number.** We use the projection of x-axis to extract the region of the number plate as the first step to extract each number. Since the maxima of the projected values of x axis are located near numbers as you can see in Figure 9, we can get information about the locations of starting and ending points for each number from these maxima. Figure 10 shows the result for the extraction of the region of the each number.

2.7. **The segmentation of the each number and the acquisition of number images.** We segment each number in each region by using the Gibou-Fedkiw algorithm and then we get the binary images as seen in Figure 11.
2.8. The normalization of number image. We prune the up, down, left and right margin to recognize the number efficiently. Since the sizes of number images extracted from number plate image are different, their pixel sizes should be normalized. Each number is scaled to $42 \times 24$ pixel image. Figure 12 shows the number image pruned out the margin and normalized.

2.9. Number Recognition. Since the car number plate system has single-font and fixed size number/character, the template matching technique is useful. Number plate is composed of number and Korean character. As seen in Figure 11, among segmented images in the number plate, third image is a Korean character and six other images are numbers. Thus, when the image we want to recognize is a number image, we use the number-template and choose the most similar one. In the case of Korean character, we compare segmented Korean character image with the character-template. The recognition process is based on the computation of
correlation values for each template over the extracted image. The Korean language made up of 28 phonetic signs (11 vowels and 17 consonants). The Korean characters are built up from two or more symbol. Some research has been done into segmenting the Korean character into individual symbols. But results have shown that this method has not produced results any better than those systems that don’t segment the character into symbols. So segmentation of the character was not implemented. Since Korean character is composed of two or three symbols, there are many possibilities. And many of these combinations look remarkably similar. So Korean character recognition is more challengeable than number recognition. Although the Korean character used in car number plate system is limited, the recognition rate corresponding to the Korean character is much lower than one to number. So in the following study, we try to improve the recognition rate for Korean character.
3. CONCLUSION

In this paper, we applied the segmentation method using level sets to the automatic number plate recognition for simple cases. The results are quite good but this approach is very limited. In practice, we have various images such as blurry, noisy, and transformed images. For the ANPR system, the extraction of the region of the number plate is a very important procedure. Therefore, we expect better and stable results by replacing the use of the projections of x and y axis in this procedure with an approach to find feature points of the number plate.

ACKNOWLEDGMENTS

This work was supported by the Korea Sanhak Foundation (2009).

REFERENCES


