This is the third IMA summer program in mathematical modeling. The previous two programs were documented in IMA Preprint Series # 1021, 1992 and # 1254, 1994. A group of 30 students (consisting of graduate students and instructors) participated in the program, under the guidance of six instructors.

The program was divided into two periods of ten days each. For each period, the students were divided into teams of 7-10 people, each with one tutor in charge. Each team worked on a different problem for the whole period of 10 days, and presented final oral and written reports. The written reports were assembled in this Technical Report, without any additional editing. They are meant to convey the tentative and flexible environment that characterizes mathematical modeling. It is hoped that some of this material will serve instructors who wish to develop mathematical modeling courses for undergraduates.
Some remarks by Don Drew:

"The objective was to learn modeling rather than to completely solve the problems. When a model is devised and it does not explain the experimental results, you know something is missing in the model. But even if the model explains the experimental results – this does not yet mean that the model is correct. The thing to do is try to use this model to predict, and if it predicts well – then you know the model has more credibility."

"When you have a problem, break it down to simple submodels. Always be aware that with all submodels around, there is no need for one submodel to be more accurate than the other submodels."

"Nondimensional, so you can better gain insight."
List of Tutors

Danny Baker
Donald Drew
Greg Kriegsmann
Blaise Morton
David Ross
Colin Please

General Motors
Rensselaer Polytechnic Institute
New Jersey Institute of Technology
Honeywell
Eastman Kodak
Southampton University, England

List of Students

Adair Aguiar
Deborah Alterman
David R. Bainbridge
George F. Barrick
Patrick T. Bidigare
Mary E. Bond
Donna Calhoun
Xinflu Chen
Jonathan Drexler
Chaitan P. Gupta
Chaoeng Huang
Xianhua Huang
Aaron Johnson
Daniel Kern
Monica Lluis
Hector Lomeli
Tanya L. Leise
Karen Matthews
Ruth Michler
Saib Othman
Julie Plante
Esa Markus Rantanen
Svetlana Rudnaya
Linda B. Smolka
Christian Turner
Brian Von Dohlen
Patrick A. Worfolk
Daoqi Yang
Karen Zangara
Wei Zhang

University of Minnesota
University of Michigan
University of Michigan
Kent State University
University of Michigan
University of Kentucky
University of Washington
University of Pittsburgh
University of Minnesota-Duluth
University of Nevada
Wright State University
Delft University of Technology
University of Kansas
University of Illinois at Chicago
University of Minnesota
University of Colorado
Texas A&M University
University of Illinois at Chicago
University of North Texas
University of Iowa
Université de Montreal/CERCA
Pennsylvania State University
University of Minnesota
Pennsylvania State University
Texas A&M University
University of Michigan
University of Minnesota
Wayne State University
Virginia Tech
Michigan State University
A Fluid Mechanics Approach
to Mixing Problems

Adair Aguiar  
U of Minnesota

Donna Calhoun  
U of Washington

Jonathan Drexler  
U of Minnesota-Duluth

Tanya Leise  
Texas A&M U

Monica Lluis  
U of Minnesota & U of Colorado

Julie Plante  
U of Montreal

Svetlana Rudnaya  
U of Minnesota

Daoqi Yang  
Wayne State U

Tutor Donald Drew  
Rensselaer Polytechnic Institute

August 6, 1996
Abstract

We study the problem of mixing two fluids by assuming that the flow is generated by a flat paddle in a two dimensional steady state regime. To understand the underlying principles of mixing, we examine the special cases of potential flow and viscous flow. For miscible fluids a diffusion criterion is used while for immiscible fluids a model of droplets breakup is devised, which gives an estimate for when droplets are broken up to a size smaller than a given characteristic size. In both cases, a probabilistic model is used to determine the amount of mixing for a number of paddle passes through the fluid.
1 Introduction

The mixing of fluids is a complicated fluid dynamical problem. The motion of the fluid can take the form of stretching and folding strands of fluid or of the breakup of droplets. The goal of mixing is to reduce length scales (by stretching or breakup) to a sufficiently small size such that the resulting mixture is homogenized.

The mixing process depends greatly on the fluids involved. For miscible fluids, diffusion is the ultimate homogenization process and will dominate once the length scale is reduced. For immiscible fluids, diffusion is not the mechanism that provides the final homogenization. The thickness of stretched and folded strands of fluid, or the diameter of droplets could be used as a measure of how well mixed the fluids are in this case.

2 Fluid Mechanics

The mixing of two fluids occurs when the two materials are brought into contact and are allowed (or forced) to flow together. For simplicity, we assume the same basic properties for the two fluids (e.g., incompressible fluids with the same coefficient of viscosity). We study the process of mixing as appropriate to a tank of fluid with a mechanical mixer, which we assume is a paddle. We want to isolate the local effect of the paddle by examining the planar flow of two fluids around the paddle in an infinite region. We study two different flows, inviscid irrotational flow, also called potential flow, and a viscous flow in steady state. Potential flow is valid for high Reynolds numbers ("fast" stirring), while the second flow is valid for low Reynolds numbers ("slow" stirring).

2.1 Potential Flow

Let $B \subset \mathbb{R}^2$ be the region occupied by an incompressible fluid such that

$$\text{Div } V = u_x + v_y = 0 \quad \text{in } B,$$

(1)

where $u$ and $v$ are the components of the velocity field $V$ in an cartesian coordinate system fixed to the paddle. The Cauchy stress for a fluid that behaves linearly with respect to the velocity gradient $\nabla u$ is then given by

$$T = -p1 + 2\mu \nabla V,$$

(2)
where $\nabla V$ is the symmetric gradient of $V$, $p$ is the pressure field, and $\mu$ is the coefficient of viscosity. Under the additional assumptions of inviscid fluid, i.e., $\mu = 0$, and steady irrotational flow with uniform density $\rho$, the expression for the conservation of momentum

$$\text{Div} \mathbf{T} = \rho \frac{d u}{d t} \quad \text{in } B$$  \hspace{1cm} (3)

reduces to the special case of Bernoulli's equation

$$\frac{p}{\rho} + \frac{1}{2} |V|^2 = \text{const}. \hspace{1cm} (4)$$

The pressure $p$ can be determined from (4) once the velocity field $V$ is known.

To determine $V$ we use complex variables techniques. Let $V = u + iv$ be the complex representation of the velocity field of a particle of the fluid at any point $(x, y)$. Then we show in Appendix A that $V$ is the conjugate of the derivative of an analytic function $F(z)$ defined in $B$, i.e.,

$$V = F'(z). \hspace{1cm} (5)$$

The function $F(z)$ is called the complex potential of the flow.

Hence, the steady state problem for an irrotational flow of an incompressible and inviscid fluid occupying a simply connected domain consists of finding the complex potential $F(z)$ such that the fluid can slip on the boundary of the paddle and that the velocity field is uniform at infinity.

We found the complex potential $F(z)$ for the flow passed the flat plate (paddle) as the composition of the complex potential for the uniform flow on an infinite plate and the conformal mapping between the infinite plate and the flat plate. The resulting composition has the following form:

$$F(\xi) = A [z \cos \alpha - i\sqrt{z^2 - 4} \sin \alpha], \hspace{1cm} (6)$$

where the width of the flat plate is 4 and $A$ and $\alpha$ are, respectively, the magnitude and the direction of the uniform velocity field at infinity. The cross section of the paddle is the line segment joining the flow points $z = \pm 2$.

### 2.2 Viscous Flow

We next consider the flow generated by a paddle in an infinite medium occupied by a viscous fluid. Again we seek a steady-state solution for an
3 SOLVING FOR THE STREAMLINES

incompressible fluid, but now the fluid is viscous and vorticity may occur. For this, we consider the boundary value problem of finding the solution $u : B \rightarrow \mathbb{R}^2$ that satisfies

$$- \nabla p + \frac{1}{R} \Delta V = 0, \quad (7)$$

$$\text{Div} V = 0, \quad (8)$$

where $R = L A/\mu$ is the Reynolds number, given in terms of some characteristic length $L$ and velocity $A$ of the flow. In the present case $L$ is the width of the paddle and $A$ is the magnitude of the velocity at infinity. The solution must satisfy the boundary conditions

$$V(x, y) = (0, 0) \quad \text{for } (x, y) \in \mathbb{R}^2 - B, \quad (10)$$

$$V(x, y) \rightarrow (1, 0) \quad \text{as } x^2 + y^2 \rightarrow \infty \quad (11)$$

Notice that these boundary conditions differ from those given in the potential flow problem in a fundamental manner. In the potential flow, fluid slipped around the paddle; in the viscous flow, fluid sticks to the paddle and is pulled along with it. The difference is apparent in the streamlines of the two flows. In potential flow, only stretching occurs, while in viscous flow both stretching and folding occur.

For small Reynolds number, the viscous flow problem posed above has been solved analytically by Stuart [6] and is closely related to a similar problem in classical linear elasticity theory (see, for instance, Daniele [2] and Eschelby [3]). The solution is explained briefly in Appendix B. Unfortunately, this solution proved too unwieldy for practical numerical computation. To generate streamlines for the Stokes problem, we solve a corresponding variational problem in a finite domain by using a finite element method based on the Galerkin approximation. The conditions at infinity are replaced by $V = 0$ along the boundary of a large square. The square is then discretized by a regular mesh of biquadratic elements, i.e., rectangular elements with bilinear functions.

3 Solving for the streamlines

Once the velocities for the fluid flow have been determined, they can be used to obtain streamlines of the flow. Streamlines are the paths traced out by
SOLVING FOR THE STREAMLINES

A particle in the flow. In the case of potential flow, these stream lines can be easily obtained from the potential function. In this case, the streamlines are just the level sets of the function

$$\psi(x, y) = \text{Im}(F(z))$$

The contours of this function can be plotted using any standard graphics package.

More interesting, however, is to trace out the path that a particular initial configuration of particles will trace out. By doing this, one can see how shapes, such as circles, discs, or squares will be deformed by the flow. To do this, one must start with an initial configuration of particles, $$\{x_j\}_{j=0}^M$$ and evolve them according to the dynamical system given by:

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

$$\dot{y}(t) = v(x(t), y(t), t)$$

where $$(u, v)$$ are the velocities from the complex potential equations, or some other solution to the desired fluid equations.

In most cases, these equations must be solved numerically. In the case of potential flow, at least, one must also be sure to use a numerical routine that is accurate enough to keep the particles on the streamlines described by the streamline function $$\psi(x, y)$$. We used the numerical solver LSODE from ODEPACK.

Plots from an evolved circle and a square, for different flow fields, are shown in figure.

3.1 Measuring stretching

Once one has passed an initial configuration of particles pass the paddle, then the resulting configuration can be used as an initial configuration for a the flow field. Proceeding in this way, one can estimate how much stretching can occur in the flow field. In figure (3), one can see that stretching of the initial square of particles, and the accompanying area-preserving thinning occurs. Rough estimates yield a stretching factor of 1.5 and a thinning factor of 2/3. These numbers will be used to make an estimate of the time it will take for strands to reach the critical thickness and spacing at which diffusion finishes the mixing.
Figure 2: Viscous flow around a paddle.
Figure 1: Drop stretching and breaking in potential flow

Figure 2: Drop stretching and breaking in potential flow
3 SOLVING FOR THE STREAMLINES

Figure 3: Square stretching in potential flow

Figure 4: Deformation and stretching in potential flow
3.2 Efficiency

Ottino\(^1\) presents the idea of efficiency of mixing. Mixing is a combination of stretching and folding, where more stretching implies greater efficiency.

Define the length stretch parameter as:

\[
\lambda = \frac{||dz||}{||dX||} = \frac{\text{length of strand after deformation}}{\text{length of strand before deformation}} = \sqrt{F^TF} : nn
\]

where \(F\) is the deformation gradient \(\nabla_X x\) (\(X\) is the material point and \(x\) is its new spatial position after deformation) and \(n\) is the unit normal of the strand.

Efficiency can be defined as:

\[
\frac{D(\ln \lambda)}{Dt} = \frac{D}{Dt} \left( \frac{1}{2} \ln(F^TF : nn) \right) = Sym(\nabla V) : nn
\]

where \(Sym\) means the symmetric part. Note that \(\frac{D}{Dt}F = (\nabla V)^T \cdot F\). The eigenvalues of \(Sym(\nabla V)\) will indicate the magnitude of stretching and thinning caused by the mixing.

4 The final stages of mixing: Diffusion

The second important aspect of mixing for miscible fluids is that of diffusion. Diffusion acts on a molecular scale and is responsible for blending thin strands together.

If we consider thin strands, as opposed to droplets, it is reasonable to model diffusion using one dimensional equations. The one dimensional diffusion equation can be derived from random walk arguments and is given by

\[
C_t = DC_{xx}, \quad -\infty < x < \infty
\]

subject to the initial conditions

\[
C(x, 0) = f(x)
\]

where \(C(x, t)\) is the concentration of the substance to be mixed, and \(D\) is the diffusion coefficient.

\(^{\text{1}}\)See [4] Chp.4
The solution to this initial-value problem is well known \(^2\) and given by

\[
C(x, t) = \frac{1}{\sqrt{\pi Dt}} \int_{-\infty}^{\infty} e^{-(x-\xi)^2 / 4Dt} f(\xi) d\xi
\]

We modeled the initial conditions of the cross-sectional concentration profile of the strands by composing squares waves of the form

\[
g(x) = \begin{cases} 
1 & w_1 < x < w_2 \\
0 & \text{otherwise.}
\end{cases}
\]

The initial conditions are shown in figure(*)).

With this initial concentration for a strand from \(w_1\) to \(w_2\) in millimeter, the solution reduce to:

\[
C(x, t) = \frac{1}{2} \left[ \text{erf} \left( \frac{w_1 - x}{2\sqrt{Dt}} \right) - \text{erf} \left( \frac{w_2 - x}{2\sqrt{Dt}} \right) \right]
\]

We consider the two fluids mixed if the strands have diffused to a point where the concentration of the one fluid in the other is within 10% of being uniform. With this criteria, we need to find the time it takes to diffuse the two fluids, given the separation between the two strands and the width of the strands. We computed time values for numerous lengths and separations from 0.05 mm to 0.25 mm with the diffusion of alcohol in water and we used least square fitting technique to obtain the following model:

\[
T_{\text{stop}}(s, w) = 36.68 s + 33.96 w - 4.940
\]

We graphed it and compared it with the computed time values in figure titled Max Time \((s, w)\).

The error made by using the approximation \(T_{\text{stop}}\) is negligible when compared to realistic paddle speed.

### 4.1 Example

Suppose we mix two fluids (which are identical) in a container of radius 50mm with a 20mm paddle, in imitation of stirring coffee in a cup. We move the paddle twice per second, and the cup contains one 20mm wide strand of one fluid surrounded by the other fluid, e.g., a bit of cream in the coffee. Each paddle pass through the strand reduces the width by a

\(^2\)See [5]
third. To have a diffusion time less than the time for a paddle pass (half of a second) requires that

\[ T_{\text{stop}}(20(2/3)^t) < 0.5. \]

This occurs at approximately 7 seconds.

5 Immiscible fluids

In this section we present a model for the mixing of two immiscible fluids. Specifically, we will examine the mixing of a droplet A moving in fluid B by a flow created by the paddle structure for the inviscid fluid (potential flow). For the droplet we make the following assumptions:

**Assumption 5.1**

1) **Droplets which are nonspherical in shape will stabilize into a sphere before the next pass by the paddle**

2) **Droplets which have separated will not come back together**
3) The initial droplet is considered mixed when it has separated into droplets; all of which are below a given (threshold) radius, \( r_{tol} \)

We shall analyze two separate approaches for addressing the mixing of the droplets:

**Approaches 5.1**

1) *Probabilistic approach*

2) *Physical approach*

**5.1 Probabilistic approach**

Given a container of radius \( r_c \), we want to determine the length of time it takes to mix a droplet of radius \( r_{drop} \). We note from the study of the potential flow that a droplet is split in two if the droplet flows into the stagnation point of the paddle. If we consider the center of the container to be the origin, all droplets lie on some polar angle within the container. Thus, if our paddle takes a random path (of increasing angle) about the container, we assume the probability of a droplet being intersected by the stagnation point is:
Probability for droplet \( i = \frac{2r_{drop,i}}{r_c} \) (12)

For this model, let us assume:

Any droplet intersected by the stagnation point separates into two droplets of radius \( \frac{r_{drop}}{\sqrt{2}} \) (by conservation of area)

This model was encoded in FORTRAN using the following nominal parameters:

\[
\begin{align*}
   r_c & = 5\text{cm} = 50\text{mm} \\
   r_{drop} & = 1\text{cm} = 10\text{mm}
\end{align*}
\] (13)

These parameters correspond to a droplet of cream being mixed in a container the size of a coffee cup. Solving for various \( r_{tot} \) gives us the following set of data:

<table>
<thead>
<tr>
<th>( r_{tot} (mm) )</th>
<th>number of paddle passes necessary to mix</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.0</td>
<td>0</td>
</tr>
<tr>
<td>5.0</td>
<td>22</td>
</tr>
<tr>
<td>2.0</td>
<td>59</td>
</tr>
<tr>
<td>1.0</td>
<td>136</td>
</tr>
<tr>
<td>0.5</td>
<td>385</td>
</tr>
<tr>
<td>0.2</td>
<td>1196 ( 47000 droplets)</td>
</tr>
<tr>
<td>0.1</td>
<td>3142 ( 327000 droplets)</td>
</tr>
</tbody>
</table>

Noting that the number of necessary paddle passes necessary to mix goes to infinity as our tolerated radius goes to zero (and using the table we obtain the following graph:
This is an estimate of how many paddle passes are necessary to mix the droplets using the nominal parameters in (A). If we know the speed of the paddle (in paddle passes/unit time), we can estimate the amount of time necessary to mix the droplet. For example, if the stirring occurs at $2\pi$ radians per second, we conclude the mixing times are equal to the number of paddle passes graph (5.1) in seconds. More statistical testing with a more reliable random number generator is necessary to refine the graph (5.1). Since the parameters $r_c$ and $r_{drop}$ were chosen arbitrarily, we can use the same method to generate mixing time estimates for any parameters.

5.2 Physical approach

In this model, we assume that a droplet is made up of a fluid with similar properties to the surrounding fluid, except that surface tension becomes important as the droplet is stretched. We assume that a droplet stretches and thins as it moves near the paddle. We further assume that the location where the droplet is most likely to break is when it is even with the edge of the paddle, where the pressure and velocity fields are singular.
Figure 9: *Stretching in potential flow*

Given any spherical droplet, we can associate two stream lines with the edges if the droplet. When the droplet is even with the edge of the paddle, we can determine the total pressure inside the droplet as it passes.

From this diagram we make the following assumptions:

**Assumption 5.2**

1) *Droplets never lie on the stagnation line*

2) *The droplet always fills the entire distance, d, between the flowlines*

3) *The surface tension always directly opposes the velocity vector*

4) *The droplet breaks if the pressure force on the droplet exceeds the surface tension at the edges*

From our diagram (5.2) we obtain the following:

\[ dF = pdx \]

Integrating with respect to \( x \) yields:

\[ F = \int_{x_p=d}^{x_p} pdx \]

where \( \sigma \) is the droplets maximum distance from the edge and \( d \) is the stretched width of the droplet at the edge. Twice the surface tension \( \sigma \) equals the force from surface tension (by our diagram 5.2). The pressure
5 IMMISCIBLE FLUIDS

force depends on \( d \). If \( d \) is sufficiently large such that \( \int_{x = -d}^{x = d} p \, dx > 2\sigma \), we assume that the droplet will break. Thus, setting the following determines the critical size \( d^* \) for a droplet to break under the force of surface tension:

\[
2\sigma = \int_{x = -d^*}^{x = d^*} p \, dx \tag{14}
\]

If the droplet is small enough that it can pass in its entirety (without breaking), by our assumption, it will return to a spherical shape. If the droplet breaks, we assume it will break into droplets of radius \( d/2 \).

Now let us examine the pressure field to determine if and when a droplet breaks. Using Bernoulli's equation:

\[
p + \frac{1}{2} |V|^2 = \text{constant}
\]

we can determine the pressure at any point in the potential flow. Note that \( p \) is proportional to \( V^2 \), where \( V \) is the velocity of the paddle.

At points \((x, 0)\) even with the edge of the paddle we note \( V = 0 \) and \( u \) is its maximum value. Furthermore, we note that as the point approaches the edge of the paddle then \( u \to \infty \), and as \( x \) approaches infinity, \( u \to 0 \).

As \( x \) can be mapped to a specific initial condition \( x_0 \), we can determine the velocity of the point at the edge of the paddle from the initial condition. Furthermore, it looks like the following:

If we measure \( x_0 \) for a given droplet as the point furthest from the stagnation line, we can estimate the velocity of the droplet at the edge of the paddle. As the pressure depends on the velocity (and in turn the surface tension on the pressure), we can determine from formula (14) a critical diameter \( d^* \) for the droplet. Notice that if \( d > d^* \) the droplet will break into droplets of radius \( d/2 \), otherwise the droplets will not break.

Graphing the critical radius \( r^* \) vs. \( x_0 \), we obtain the graph (5.2). For a given mixing situation, we need to determine the velocity of the paddle in order to get the graph (5.2). From this graph we are able to determine a distance \( x_0^*(r_{tol}) \) such that all droplets within this area are considered to be broken (automatically). Thus from this data we are able to determine the following:

If \( x_0 \leq x_0^*(r_{tol}) \) then the droplet breaks

If \( x_0 > x_0^*(r_{tol}) \) we determine if it breaks using graph (5.2)

To improve our probabilistic model, in the future, we can incorporate the information in this section into our probabilistic model. This should improve our first assumption (5.2). In order to improve assumptions (2),
Figure 10: *Velocity versus initial position*

Figure 11: *Critical radius versus initial position*
(3), and (4) it would be necessary to track all boundaries of all droplets for all time. As the number of droplets during the mixing process can easily grow into the millions, this appears to be computationally infeasible. Last, as we are including surface tension, the assumption of similar fluids, which generated the flows, is no longer valid. An improved model might account for the change in the flow field generated by the surface tension.

We have developed a probabilistic model and determined, as expected, that mixing time increases exponentially as our tolerated radius, \( r_{tol} \), decreases. Furthermore, we have used a physical approach to incorporate the droplet breakup due to shearing (which was lacking in the probabilistic method). This information, in the future, can be incorporated to refine our current probabilistic model from which we can better predict the mixing time.

6 Conclusion

The mathematical models pertaining to mixing give insights into the fundamental process involved in mixing. We considered the mixing of two similar fluids, so that the dynamic of mixing could be determined by considering the flow of a fluid with given properties. We treated immiscible and miscible fluids in two different cases: the first one resulting in droplet formation and the second one forming strands.

The fluid dynamics of mixing is complicated; therefore, we concentrated on flows that had the fundamental aspects pertinent to mixing. We considered the two-dimensional flow around a flat plate, which models a moving paddle. We solved this problem for potential flow, which results in a reasonable velocity and pressure profile for fluids of small viscosity assuming no boundary layer separation on the paddle. Using the velocity field of the flow, we obtained the stretching of strands each time the paddle passes through the mixture. We could define effectiveness of mixing with these results using a probabilistic model for the paddle to intercept unmixed fluid. The strands were assumed to be stretched every time the paddle passes. We also examined Stokes flow around a moving paddle. The flow is qualitatively similar; there are some quantitative differences.

The stopping criteria for the immiscible model specifies a maximal acceptable radius of droplets. When all droplets are smaller than this specified value, the mixing is done. A physical model was developed to predict droplet breakup. The model is based on pressure on the surface of droplets and the
surface tension. From this, a critical radius is defined: if a droplet is larger
then this radius, it breaks into droplets of a size which depends on paddle
velocity and on surface tension.

The stopping criteria for the miscible fluids depends on diffusion. The
mixing is essentially complete when neighbouring strands can diffuse into
each other in a shorter time than it takes for a paddle pass. Two neighbour-
ing strands are diffused in each other if the difference in their concentration
over the distance that separates them is less than 10spiralizing effect in the
recirculation zone behind the paddle. This feature may enhance or reduce
mixing; further study is warrented. The model used for droplet breakup did
not include the shear stress that would develop in Stokes flows. In this case,
a more sophisticated physical model should be developed.

7 Appendix A

Let \( V = u + iv \) be the complex representation of the velocity of a particle
of a fluid at any point \( (x, y) \). The circulation of the fluid along any contour
\( C \) is defined as

\[
\int_C V_T(x, y) d\sigma,
\]

where \( V_T(x, y) \) is the tangential component of the velocity field along \( C \) and
\( \sigma \) is the arc length. Given a positively oriented simple closed countour \( C \)
lying in a simple connected domain containing no sources or sinks, we then
have by Green's Theorem (see [1]) that

\[
\int_C V_T(x, y) d\sigma = \int u(x, y) dx + v(x, y) dy \\
= \int \int_B [u_x(x, y) - v_y(x, y)] dx dy. \tag{15}
\]

The function

\[
\omega(x, y) = \frac{1}{2} [u_x(x, y) - v_y(x, y)] \tag{16}
\]

represents the limiting angular speed of a circular element of the fluid as
a circle shrinks to its center \( (x, y) \), the point at which \( \omega \) is evaluated. If
\( \omega(x, y) = 0 \) at each point in some simple connected domain, the flow is
called irrotational in that domain.
Since the flow is irrotational, \( i.e \, u_y = v_x \), we can introduce the velocity potential function

\[
\phi(x, y) = \int_{(x_0, y_0)}^{(x, y)} p(s, \tau) ds + q(s, \tau) d\tau,
\]

which is well defined in a simply connected domain \( \mathcal{D} \). The gradient of the velocity potential \( \phi(x, y) \) represents the velocity field

\[
V(x, y) = \nabla \phi(x, y)
\]

The level curves are equipotentials and the velocity field \( V \) is normal to an equipotential at any point where \( V \) is not zero. The incompressibility condition \( i.e \text{Div} \, V = 0 \) yields

\[
\Delta \phi = 0 \quad (x, y) \in \mathcal{D}.
\]

If, moreover, \( \psi(x, y) \) is the harmonic conjugate of \( \phi(x, y) \) in \( \mathcal{D} \), the velocity field is tangent to a curve \( \psi(x, y) = c \). These curves are called the streamlines of the flow, and the function \( \psi \) is the stream function. The analytic function

\[
F(z) = \phi(x, y) + i\psi(x, y)
\]

is then the complex potential of the flow and thus,

\[
V = \overline{F'(z)}.
\]

8 Appendix B

The closed form solution for three-dimensional viscous fluid flow around an ellipsoid is given by (see Daniele [2])

\[
\begin{align*}
u &= \alpha y \frac{\partial U_0}{\partial x} + \alpha_1 \frac{\partial^2 U_1}{\partial y^2}, \\
v &= \alpha_0 U_0 + \alpha y \frac{\partial U_0}{\partial y} + \alpha_1 \frac{\partial^2 U_1}{\partial y^2}, \\
w &= -\alpha_0 y \frac{\partial U_0}{\partial z} - \alpha_1 \frac{\partial^2 U_1}{\partial y \partial z},
\end{align*}
\]

where \( V = (u, v, w) \) is the velocity field and
REFERENCES

\[ U_0 = \frac{1}{2} \int_{s_0}^{\infty} \frac{ds}{R(s)}, \]
\[ U_1 = \frac{3}{4} \int_{s_0}^{\infty} \frac{E(s)}{R(s)} ds, \]
\[ R(s) = \sqrt{(a^2 + s)(b^2 + s)(c^2 + s)}, \]
\[ E(s) = 1 - \frac{x^2}{a^2 + s} - \frac{y^2}{b^2 + s} - \frac{z^2}{c^2 + s}. \]

The number \( s_0 \) depends on \((x, y, z)\) and is the greatest real solution to \( E(s) = 0 \). The coefficients can be determined as follows.

\[ \alpha = \frac{1}{hH - Ab^2} \]
\[ \alpha_0 = h\alpha \]
\[ \alpha_1 = \frac{\alpha b^2}{3} \]
\[ h = \frac{2}{k} - 1 \]
\[ H = \frac{1}{2} \int_{s_0}^{\infty} \frac{ds}{R(s)} \]
\[ A = \frac{1}{2} \int_{s_0}^{\infty} \frac{ds}{(a^2 + s)R(s)} \]

and where \( b \) is the axis of the ellipsoid in the \( y \)-direction and \( k \) is chosen very small. This solution is for elastic solids, but holds for slow viscous motion in the limit \( k \to 0 \).

References


REFERENCES


A mathematical model for the effect of surfactants in the production of photographic films.

Chaoeng Huang, Wright State University
Christian Turner, Texas A.&M. University
Saib Othman, Iowa State University
Linda Smolka, Pennsylvania State University
Xinfu Chen, University of Pittsburgh
George Barrick, Kent State University
Daniel Kern, University of Illinois, Chicago
David Ross (advisor), Eastman Kodak Co.

Institute for Mathematics & its Applications

August 6, 1996

1 Introduction

As part of the process of producing photographic films, an aqueous solution of photosensitive chemicals is applied to the film backing by allowing the solution to flow through a slit in a rectangular “pot”, creating a “curtain” of fluid which evenly coats the backing. Since the coating needs to be uniform and smooth, a very smooth non-turbulent flow is required in the curtain. If the solution has a high surface tension, the flow in the curtain will be uneven, and undesirable non-uniformities in the coating will result. Minimizing the surface tension of the fluid vastly reduces these difficulties.

In order to reduce the surface tension of the fluid, surfactants are added to the pot. When in contact with an air/liquid interface, the hydrophobic “tails” of the molecules stick
through the surface into the air. When no surface is exposed to the air (as when the fluid is still in the pot), molecules are uniformly distributed through the solution. As the fluid begins its fall from the pot, new surface is exposed, and chemical kinetic considerations urge the surfactant molecules to begin diffusing to the surface, thus reducing surface tension of the moving fluid. With further distance below the pot, more surfactant molecules diffuse to the surface, reducing surface tension yet more, until eventually the surface is saturated.

The choice of surfactant effects the reduction in surface tension. The longer the tail of a surfactant molecule species, the greater is its hydrophobicity. This notion correlates tightly with, and is usually measured in terms of, the critical micelle concentration (CMC) for the surfactant. Above this particular bulk concentration of individual surfactant molecules (monomers), the tails of the surfactant molecules begin to aggregate with one another rather than diffusing quickly to the surface, thus creating micelles almost immediately. Micelles are balls consisting of approximately 60 surfactant molecules, all with their hydrophobic tails directed in towards the center of the cluster. Micelles diffuse to the surface of the fluid much less rapidly than do monomers, though throughout any solution in which the surfactant solute is above CMC there always exists some steady-state exchange between micelles and monomers.

At first glance, it would appear that surfactant species having longer hydrophobic tails (and thus lower values for CMC) contribute directly to lower surface tension in the fluid. This is in fact the situation for fluids when enough time has elapsed for monomeric and micellular forms of the surfactant to reach equilibrium. However, experimental data (Fig. 1) gathered for fluid curtains at times shortly after the emergence of the fluid from the pot show that the relationship between surface tension and CMC is not always monotonic. The graph suggests that before equilibrium is attained, surface tension has a minimum value at approximately that place where the CMC and bulk concentration of surfactant coincide. The goal of this project was to develop a mathematical model capable of explaining the experimental data and also of predicting the behavior of surfactants based on their CMC.

2 The model

The process we seek to model here is simply the diffusion of monomers and micelles. The flow of fluid is assumed to be in the steady state, incompressible and laminar (with a constant
viscosity.) This means that the flow of the fluid will not significantly effect the process of surfactant molecules diffusing to the surface of the fluid, and will thus be ignored. We assume the fluid curtain is rectangular with a relatively small thickness, approximately 10cm x 10cm x 1mm. Across the width of the curtain, the diffusion process does not vary, though there is variability in the vertical direction due to different ages for the surface elements as they move (downward) away from the slit in the pot. In terms of an Eulerian reference frame, the diffusion process varies as the fluid falls away from the slit. In terms of a Lagrangian reference frame we can consider a particular packet of fluid (whose coordinate system we adopt), and use as our independent variables time after emergence from the slit and one spatial dimension giving the distance (depth) from the air/fluid interface.

2.1 A simplified mathematical model

We examine here those situations in which the bulk concentration of the surfactant is less than the CMC (as with short-tailed surfactants,) so that there are no micelles in the solution. Henceforward we will use the notations:

\[
\begin{align*}
t & = \text{ time (s)} \\
x & = \text{ distance from the air/fluid interface (cm)} \\
C(x,t) & = \text{ bulk concentration of monomers (mol/cm}^3) \\
\Gamma(t) & = \text{ surface concentration of monomers (mol/cm}^2) \\
\Gamma_{\text{max}} & = \text{ maximum surface concentration of monomers (mol/cm}^2) \\
\gamma(t) & = \text{ surface tension (dyne/cm}^2)
\end{align*}
\]

Whenever the solution is in equilibrium with its surface, the Langmuir isotherm equation is valid:

\[
K_\text{on} C(0_+,t) (\Gamma_{\text{max}} - \Gamma(t)) = K_{\text{off}} \Gamma(t) .
\] (2.1)

In this formula, \( K_\text{on} \) (cm\(^3\)/(mol \cdot s)) is a constant indicating the likelihood of an individual surfactant molecule colliding with and sticking to the surface. The constant \( K_{\text{off}} \) (s\(^{-1}\)) reflects the likelihood of monomers dissociating from the surface. Rewriting equation (1) gives

\[
\frac{(\Gamma_{\text{max}} - \Gamma(t))}{\Gamma(t)} C(0_+,t) = \frac{K_{\text{off}}}{K_\text{on}} \propto \frac{\text{mol}}{\text{cm}^3}
\]
or,

\[
\frac{1}{\Gamma_{\text{max}}} \cdot \Gamma(t) = 1 - \frac{1}{1 + \frac{1}{\kappa} C(0_+, t)} ,
\]

where \( \kappa := 10^{-3} \times \text{CMC} \). This provides us with a relationship between the surface concentration and the bulk concentration. The connection with surface tension appears in the Frumkin equation.

\[
\gamma(t) = \gamma_0 + R K \Gamma_{\text{max}} \ln \left( 1 - \frac{\Gamma(t)}{\Gamma_{\text{max}}} \right) ,
\]

where \( \gamma_0 \) is the surface tension of a clean surface (approximately 65 dyne/cm²), \( R \) is the Boltzmann gas constant (about \( 8 \times 10^7 \) dyne/(mol · K) ), and \( K \) is Kelvin temperature (300 K).

The classical diffusion equation is a partial differential equation (p.d.e.) of the form

\[
\frac{\partial}{\partial t} C(x, t) = D \frac{\partial^2}{\partial x^2} C(x, t) , \quad (x, t) \in (0, \infty) \times (0, \infty)
\]

where \( D \) is the molecular diffusivity constant (approximately \( 4 \times 10^{-6} \) cm²/s).

To arrive at a clear notion of what the boundary conditions for the problem ought to be, we further refine the physical model discussed earlier. As fluid begins to flow through the slit, fresh surface (with no embedded surfactant molecules) is formed, so \( \Gamma(0) = 0 \). For \( x = 0 \) and \( t = 0 \) this implies that \( C = 0 \) by (2). Thus

\[
C(x, 0) = \begin{cases} 
0 & , \quad x = 0 \\
C_0(x) & , \quad x > 0 
\end{cases}
\]

where \( C_0(x) \) is the initial bulk concentration of monomers (\( 0.3 \times 10^{-3} \) mol/cm³ in the physical experiments).

At the spatial extremes of the domain (i.e. at \( x = 0 \) and \( x = \infty \)), we need to consider conditions describing the flux of molecules into and out of the domain. For the left boundary condition (at \( x = 0 \)), the net flux should be equal to the rate of change for the surface concentration:

\[
D \frac{\partial C(0_+, t)}{\partial x} = \frac{d\Gamma(t)}{dt} .
\]
Differentiating (2) with respect to \( t \) gives an expression in terms of \( C \) itself for the right-hand side of this condition:

\[
D \frac{\partial C}{\partial x}(0_+, t) = \left( \frac{\Gamma_{\text{max}}}{\kappa} \right) \frac{\partial C}{\partial t}(0, t) \left[ 1 + \frac{1}{\kappa} C(0, t) \right]^2
\]

\[=: \phi(C(0, t)) \frac{\partial C}{\partial t}(0, t) \tag{2.6} \]

As for the condition at \( x = \infty \), we note that far from the fluid/air interface the gradient for the concentration tends toward zero for all \( t \):

\[
\lim_{x \to \infty} \frac{\partial C}{\partial x}(x, t) = 0 \quad , \quad t \in [0, \infty) \tag{2.7}
\]

Thus, the simple mathematical model proposed for situations in which the bulk concentration of the surfactant is less than the \( CMC \) consists of equations (4)-(7) which determine bulk concentration \( C \) along with the formulas (3) and (2) for calculating the resultant surface tension:

\[
\begin{align*}
\frac{\partial}{\partial t} C(z, t) &= D \frac{\partial^2}{\partial z^2} C(z, t) \quad , \quad (x, t) \in (0, \infty) \times (0, \infty) \tag{2.4} \\
C(z, 0) &= \begin{cases} 0 \quad , \quad z = 0 \\
C_0(z) \quad , \quad z > 0 \end{cases} \tag{2.5} \\
D \frac{\partial C}{\partial x}(0, t) &= \phi(C(0, t)) \frac{\partial C}{\partial t}(0, t) \quad , \quad t \in (0, \infty) \tag{2.6} \\
\lim_{x \to \infty} \frac{\partial C}{\partial x}(x, t) &= 0 \quad , \quad t \in [0, \infty) \tag{2.7}
\end{align*}
\]

2.2 The general mathematical model

Next to be considered is the situation in which bulk concentration of the surfactant molecules is greater than the \( CMC \). However, we can also reasonably expect that the previous model for solutions containing no micelles will be merely a special case of the one we now develop. The quantities involved will be:
\[ C(x, t) = \text{bulk concentration of monomers (mol/cm}^3) \]
\[ M(x, t) = \text{bulk concentration of monomers in the form of micelles (mol/cm}^3) \]
\[ T(x, t) = \text{total concentration of surfactant (mol/cm}^3) \]
\[ D_1 = \text{monomer diffusivity constant (cm}^2\text{/s}) \]
\[ D_2 = \text{micelle diffusivity constant (cm}^2\text{/s}) \]

Beginning with the mass conservation notion that \[ T(x, t) = C(x, t) + M(x, t) \] we can find \( C \) and \( M \) using the formulas:

\[
C(x, t) = \min(T(x, t), CMC) \\
M(x, t) = \max(T(x, t) - CMC, 0)
\]

Because of the varying rates of diffusion for monomeric versus micellar species in the solution, the diffusion equation takes on the modified form

\[
\frac{\partial T}{\partial t}(x, t) = \frac{\partial}{\partial x} \left( \hat{D}(T(x, t)) \frac{\partial T}{\partial x}(x, t) \right), \quad (x, t) \in [0, \infty) \times [0, \infty) \tag{2.8}
\]

where \( \hat{D}(T) = \begin{cases} D_1, & T \leq CMC \\ D_2, & T > CMC \end{cases} \)

The initial condition for the problem looks much the same as it did before:

\[
T(x, 0) = \begin{cases} 0, & x = 0 \\ T_0, & x > 0 \end{cases} \tag{2.9}
\]

where the initial total concentration \( T_0 = 10^{-3}\text{mol/cm}^3 \) in all of the situations of current interest. This merely reflects the notion that prior to being exposed to air (when it falls from the pot) the fluid’s surfactant concentration is uniform throughout the volume.

The flux condition at the air/liquid interface (i.e. the “left-hand boundary”) is a bit more complicated than it was for the simplified model. Since micelles can not enter the surface, the flux \( D_2 \frac{\partial M}{\partial x} = 0 \) for \( T > CMC \). At the same time, \( C = CMC \) when \( T > CMC \), so \( D_1 \frac{\partial C}{\partial x} = 0 \) as well. When \( T < CMC \), the flux condition is the same as for the earlier model. Thus we use

\[
\frac{\partial T}{\partial x}(0, t) = \begin{cases} \phi(T) \frac{\partial T}{\partial t}(0, t), & T < CMC \\ 0, & T > CMC \end{cases}, \quad t \in (0, \infty) \tag{2.10}
\]
Also, far from the surface of our liquid, the gradient for the total concentration should again tend toward 0 for all $t$:

$$\lim_{x \to \infty} \frac{\partial T}{\partial x} = 0. \quad (2.11)$$

This sets the right boundary condition.

As a final piece of the model, we consider the flux across the right boundary when $T = CMC$. To guarantee that the flux is continuous throughout any arbitrarily small neighborhood of the curve in $R^3$ where $T = CMC$, we would need to equate

$$D_1 \frac{\partial T}{\partial x} = D_2 \frac{\partial T}{\partial x} \quad (2.12)$$

for all possible values of $x$ and $t$. From a physical standpoint, the flow of monomers from one neighborhood must balance any flux of micelles from an adjacent one. This, however, is guaranteed whenever the flux is merely piecewise continuous, thus giving a solution function $T(x,t)$ which possesses a continuous second derivative. Because of this we merely make a minor adjustment to the left boundary condition which appears below.

Summarizing, the final form for our parabolic p.d.e. problem consists of the d.e. (8) along with its initial and boundary conditions (2.9)-(2.11):

$$\left\{ \begin{array}{l} \frac{\partial T}{\partial t} = \frac{\partial}{\partial x} \left( D(T) \frac{\partial T}{\partial x} \right), \quad (x,t) \in (0, \infty) \times (0, \infty) \\
T(0,x) = T_0 \chi_{(0,\infty)}(x), \quad x \in [0, \infty) \\
\frac{\partial T}{\partial x}(0,t) = \begin{cases} \phi(T) \frac{\partial T}{\partial t}(0,t), & T \leq CMC \\ 0, & T > CMC \end{cases}, \quad t \in (0, \infty) \\
\lim_{x \to \infty} \frac{\partial T}{\partial x}(x,t) = 0 \quad , \quad t \in [0, \infty) \end{array} \right. \quad (2.10)$$

where

$$\hat{D}(T) = \begin{cases} D_1, & T \leq CMC \\ D_2, & T > CMC \end{cases}, \quad (2.11)$$

$$\phi(T) = \frac{(\kappa T_{\max}/D_1)}{(\kappa + T)^2} ,$$

and
and \( T_0, D_1, D_2, \kappa, \Gamma_{\text{max}} \) and \( CMC \) are prescribed constants.

### 3 Non-dimensionalization of the problem

In order to mathematically encapsulate the problem in a way which avoids huge disparities in the scaling of variables and captures the relative sizes of all those quantities, we non-dimensionalize the variables. Since the elapsed time 0.1s of the experimental observations is so important to the problem, we choose the scaling \( t = \Theta \tau \) where \( \Theta = 0.1 \) s. In the case of the spatial variable \( x \), the characteristic length chosen will be \( L = \sqrt{D_1 \Theta} \) where \( D_1 = 4 \times 10^{-6} \text{cm}^2/\text{s} \), and we use \( x = L \xi \). This provides \( L \approx 6.2 \times 10^{-4} \text{cm} \) (or 6.2 \( \mu \text{m} \)), a scale which roughly corresponds to the depth of the region from which surfactant molecules can readily diffuse to the surface. As a non-dimensionalizing formula for the concentration, we offer \( T = T_0 \hat{T} \) where \( T_0 \) is the bulk initial concentration of surfactant in the solution \( (0.3 \times 10^{-3} \text{mol/cm}^3) \). For the \( CMC \) we set \( CMC = T_0 C_* \) so that the new quantity \( C_* \) gives \( CMC \) relative to \( T_0 \), and lastly the relation \( \hat{D} = D_1 \hat{D} \) which causes the new step-function \( \hat{D} \) to have 1 as its maximum value. In summary, the substitutions we use will be:

\[
\begin{align*}
  t &= \Theta \tau \quad \text{where } \Theta = 0.1 \text{s} \\
  x &= L \xi \quad \text{where } L = 6.32 \times 10^{-4} \text{cm} \\
  T &= T_0 \hat{T} \quad \text{where } T_0 \text{ is initial bulk concentration (mol/cm}^3) \\
  CMC &= T_0 C_* \quad \text{with } T_0 \text{ as above} \\
  \hat{D} &= D_1 \hat{D} \quad \text{where } D_1 \text{ is monomeric diffusivity (cm}^2/\text{s}).
\end{align*}
\]

In the statement of the p.d.e., the substitutions give the following replacements for the left- and right-hand sides respectively

\[
\begin{align*}
  \frac{\partial T}{\partial \tau} &= \left( \frac{T_0}{\Theta} \right) \frac{\partial \hat{T}}{\partial \tau} \\
  \frac{\partial}{\partial x} \left[ D(T) \frac{\partial T}{\partial x} \right] &= \left( \frac{D_1 T_0}{L^2} \right) \frac{\partial}{\partial \xi} \left[ \hat{D}(\hat{T}) \frac{\partial \hat{T}}{\partial \xi} \right].
\end{align*}
\]

Thus the non-dimensionalized version becomes

\[
\frac{\partial \hat{T}}{\partial \tau} = \left( \frac{D_1 \Theta}{L^2} \right) \frac{\partial}{\partial \xi} \left[ \hat{D}(\hat{T}) \frac{\partial \hat{T}}{\partial \xi} \right] \quad (3.1)
\]
where the diffusivity step-function is

\[
\bar{D}(T) = \begin{cases} 
1 , & \hat{T} \leq C_*, \\
\frac{D_2}{D_1} , & \hat{T} > C_* 
\end{cases}
\]

The new version of the initial condition for the problem is

\[
\hat{T}(0, \xi) = \chi_{(0, \infty)}(\xi)
\]

where \(\xi \in [0, \infty)\). The right boundary condition is virtually unchanged:

\[
\lim_{\xi \to \infty} \frac{\partial \hat{T}}{\partial \xi}(\xi, \tau) = 0
\]

where \(\tau \in [0, \infty)\). Some modification also appears in the left boundary condition since

\[
\frac{\partial T}{\partial x} = \left( \frac{T_0}{L} \right) \frac{\partial \hat{T}}{\partial \xi} , \quad \frac{\partial T}{\partial t} = \left( \frac{T_0}{\Theta} \right) \frac{\partial \hat{T}}{\partial \tau}
\]

\[
\hat{T}(0, \xi) = \phi(T_0 \hat{T}) = \frac{\left( \frac{\kappa \Gamma_{\text{max}}}{D_1} \right)}{\left( \kappa + T_0 \hat{T} \right)^2} = \frac{\left( \frac{\Gamma_{\text{max}}}{\kappa D_1} \right)}{\left( 1 + \frac{T_0}{\kappa} \hat{T} \right)^2}
\]

Thus the final form we arrive at for our parabolic p.d.e. in its non-dimensionalized form is:

\[
\begin{align*}
\frac{\partial \hat{T}}{\partial \tau} &= \frac{\partial}{\partial \xi} \left( \frac{\partial T}{\partial \xi} \right), \quad (\xi, \tau) \in (0, \infty) \times (0, \infty) \quad (3.1) \\
\hat{T}(0, \xi) &= \chi_{(0, \infty)}(\xi), \quad \xi \in [0, \infty) \quad (3.2) \\
\frac{\partial \hat{T}}{\partial \xi}(0_+, \tau) &= \phi(\hat{T}(0_+, \tau)) \frac{\partial \hat{T}}{\partial \tau}(0_+, \tau), \quad \hat{T}(0_+, \tau) \leq C_*, \quad \tau \in (0, \infty) \quad (3.3) \\
\lim_{\xi \to \infty} \frac{\partial \hat{T}}{\partial \xi}(\xi, \tau) &= 0, \quad \tau \in [0, \infty) \quad (3.4)
\end{align*}
\]

where
\[ \hat{\phi}(\hat{T}) = \left( \frac{1}{\alpha} \right) \left( 1 + \beta \hat{T} \right)^{-2}, \]

with \( \alpha = \frac{\kappa D_1 \Theta}{\Gamma_{\text{max}} L} \) and \( \beta = \frac{T_0}{\kappa} \). The constants \( T_0, D_1, D_2, \kappa, \Gamma_{\text{max}} \) and \( C_* \) are as described earlier.

4 A semi-discrete method for numerical solution

Several alternatives for the numerical solution of our problem were considered. Significant in these considerations was the fact that the physical system is meant to model a situation which changes continuously with time. Additionally, when \( x \in (0, \infty) \) is held fixed at, say \( x_j \) in our p.d.e., the system becomes:

\[
\begin{align*}
\frac{\partial T}{\partial t}(x_j, t) &= \frac{\partial}{\partial x} \left( \hat{D}(T(x_j, t)) \frac{\partial T}{\partial x}(x_j, t) \right), \quad t \in (0, \infty) \\
T(x_j, 0) &= 1
\end{align*}
\]

which (so long as the \( x \)-partial is known) is simply an o.d.e. initial value problem. Also, it should be mentioned here that we have returned to the notation \( T, x, \) and \( t \), though these now refer to the non-dimensionalized quantities described in the previous section.

Of course, the right-hand side of the d.e. is not generally known, but one way of approximating its effect on the problem is to find values for it at a succession of nodes \( \{x_j \mid j = 1, \ldots, M\} \) where \( x_1 = 0 \) and \( x_M \) are ”large”. The most straightforward scheme for treating this is to take advantage of the formulas

\[
\frac{\partial}{\partial x} \left( \hat{D}(T(x_j, t)) \frac{\partial T}{\partial x}(x_j, t) \right) = \left( \frac{1}{\Delta x} \right) \left\{ \hat{D}(x_{j+\frac{1}{2}, t}) \frac{T(x_{j+1}, t) - T(x_j, t)}{\Delta x} - \hat{D}(x_{j-\frac{1}{2}, t}) \frac{T(x_{j+1}, t) - T(x_{j-1}, t)}{\Delta x} \right\} + O(\Delta x^2)
\]

with
\begin{align*}
\bar{D}(x_{j+\frac{1}{2}}, t) &= \begin{cases} 
1, & \frac{1}{2}[T(x_{j+1}, t) + T(x_j, t)] \leq C_* \\
\frac{D_2}{D_1}, & \frac{1}{2}[T(x_{j+1}, t) + T(x_j, t)] > C_*
\end{cases}
\end{align*}

The plan is to solve the system of o.d.e.'s

\begin{align*}
\frac{dT_2}{dt}(t) &= \frac{1}{\Delta x} \left[ D_{\frac{1}{2}} \frac{T_3 - T_2}{\Delta x} - D_{\frac{3}{2}} \frac{T_2 - T_1}{\Delta x} \right] \\
\frac{dT_3}{dt}(t) &= \frac{1}{\Delta x} \left[ D_{\frac{1}{2}} \frac{T_4 - T_3}{\Delta x} - D_{\frac{3}{2}} \frac{T_3 - T_2}{\Delta x} \right] \\
\vdots \\
\frac{dT_{M-1}}{dt}(t) &= \frac{1}{\Delta x} \left[ D_{M+\frac{1}{2}} \frac{T_M - T_{M-1}}{\Delta x} - D_{M-\frac{1}{2}} \frac{T_{M-1} - T_{M-2}}{\Delta x} \right] 
\end{align*} \tag{4.3}

as an \( M \)-vector o.d.e. initial value problem using the data

\begin{align*}
T_j(0) &= \begin{cases} 
0, & \text{for } j = 0 \\
1, & \text{for } j = 1, \ldots, M
\end{cases}
\end{align*}

The one minor omission of course, is that we've failed to mention just what the first and the \( M \)th differential equations ought to be. The proposal which was implemented in our numerical model was

\begin{align*}
\frac{dT_1}{dt}(t) &= \begin{cases} 
\alpha (1 + \beta T_1)^2 \frac{T_2 - T_1}{\Delta x}, & T_1 \leq C_* \\
\left( \frac{D_2}{D_1} \right) \left( \frac{2}{\Delta x} \right) \frac{T_2 - T_1}{\Delta x}, & T_1 > C_*
\end{cases} \tag{4.4}
\end{align*}

The \( D_2 \) and \( D_1 \) used above are the original micellar- and monomeric-diffusivities stated in the general model, not the "averaged" nodal quantities, which take values of either 1 or \( D_2/D_1 \). This scheme is an \( O(\Delta x) \) approximation to the d.e. in the case where \( T_1 \leq C_* \), and is accurate to \( O(\Delta x^2) \) in the \( T_1 > C_* \) case. It should be mentioned that in the case where \( T_1 > C_* \), one tempting possibility was to have \( \frac{dT_1}{dt} = \frac{dT_2}{dt} \). This is of course an \( O(\Delta x) \) approximation to the p.d.e. at \( (x_1, t) \), but it turns out be a solution to the problem for which \( \frac{\partial^M T}{\partial x^M} = 0 \), which is not consistent with the current formulation of the problem. The \( M \)th
equation, which refers to the no-flux condition as \( x \) tends to infinity, is modeled by

\[
\frac{dT_M}{dt}(t) = D_{M-\frac{1}{2}} \left( \frac{2}{3 \Delta x} \right) \frac{T_M - T_{M-1}}{\Delta x}.
\] (4.5)

This last equation is an \( O(\Delta x^2) \) approximation at \( (x_M, t) \).

To numerically solve this \( M \)-vector o.d.e., an adaptive o.d.e. solver with step control was implemented to solve the system out to a heuristically chosen \( t_f \).

5 Mathematical analysis

5.1 Preliminaries

We next address the well-posedness, regularity and asymptotic behavior for the parabolic p.d.e. initial boundary value problem:

\[
\begin{align*}
[A(u)]_t &= u_{xx}, & x \in (0, 1), & t > 0, \\
u_x(1, t) &= 0, & t > 0 \\
u_x(0, t) &= [B(u(0, t))]_t, & t > 0 \\
u(x, 0) &= u_0(x), & x > 0,
\end{align*}
\] (5.1)

where \( A(\cdot) \) and \( B(\cdot) \) are two Lipschitz and monotonically increasing functions, and the initial condition has the form

\[
u_0(x) = \begin{cases} 
0 & \text{if } x > 0 \\
\alpha & \text{if } x = 0,
\end{cases}
\]

where \( \alpha \neq 0 \). Let it suffice to say that this minor generalization of the problem is valid whenever \( u(x, t) = T(x, t) \) as set forth in section 2.2, \( A(\cdot) \) is the diffusivity step-function and \( B(\cdot) \) is the left-hand boundary condition.

The parabolic system (5.1) raises several interesting questions. First of all, unlike usual initial boundary value problems, the initial condition here is discontinuous only at the end point \( x = 0 \). Hence, a solution must be defined properly so that it catches the physical essence and can be used to explain the experimental results. One question regards how the initial condition may be rigorously defined. It is also natural to ask whether the problem is mathematically well-posed. Second, when using an \( \varepsilon \)-perturbation of the initial function, mathematical intuition would make it seem that the limit as \( \varepsilon \to 0 \) should converge to a
solution of the problem in which the initial condition is replaced with \( u(x, 0) = T_0 \). From the physical point of view, since some monomers have migrated stably to the interface, the mass conservation law implies that this intuition can not be true. In fact, we have the following observation. Suppose that \( T_0 \) is a constant and that the system (5.1) with smooth initial condition

\[
T(x, 0) = T_0^e,
\]

where \( T_0^e \) is a \( \epsilon \)-perturbation of \( T(x, 0) \) in (5.1), in some sense admits a solution \( T^e \). Set \( u^e = T^e - T_0 \). Then \( u^e \) solves an analogous system. For any test function \( v(x, t) \) that satisfies

\[
v_x(1, 0) = 0, \quad v(x, \Theta) = 0,
\]

we have, from integration by parts,

\[
\int_{0}^{1} \int_{0}^{\Theta} u^e (v_t + v_{xx}) \, dx \, dt = -\int_{0}^{\Theta} (u^e v - \Phi(u^e)v_t)(0, t) \, dt - \Phi(-T_0)v(0, 0),
\]

where \( \Phi \) is an antiderivative of \( \phi \). Notice that an extra term \( \Phi(-T_0)v(0, 0) \) appears in the above weak formulation. By passing to the limit, we find that the limit \( T = \lim_{\epsilon \to 0} T^\epsilon \) is not identical to \( T_0 \). This observation implies that the single-point value \( T_0 \) is not negligible.

Third, the parabolic system actually involves a free boundary \( \Gamma_t = \{(x, t) \mid T(x, t) = C_*\} \). The diffusion process will stop when this free boundary reenters the interface. Physically, one is interested in determining when the concentration \( T \) exceeds \( C_* \) everywhere and thus is stabilized. Finally, despite the initial singularity at \( x = 0 \), one expects that there exists a classical solution in a certain sense, since the model is physically sound.

To study this model problem (5.1), we first rewrite the system as follows. We introduce the functions \( A, B, u \) by

\[
A(y) = F^{-1}(y + T_0) \\
B(y) = \begin{cases} 
\int_{0}^{y} \phi(A(z + T_0)) \, dz & \text{if } y < C_* - T_0 \\
\int_{C_*}^{CMC} \phi(A(z)) \, dz & \text{if } y > C_* - T_0
\end{cases} \\
u(x, t) = F(T(x, t) - T_0),
\]

where

\[
F(y) = \int_{0}^{y} D(z) \, dz
\]

is strictly increasing so that its inverse function \( A \) is well-defined. Since the initial concentration is near uniform, for simplicity we assume that \( T_0(x) = T_0 \) is a positive constant. Consider
the following problem (P):

\[
(P) \begin{cases}
[A(u)]_t = u_{xx}, & x \in (0,1), t > 0, \\
u_x(1,t) = 0, & t > 0 \\
u_x(0,t) = [B(u(0,t))]_t, & t > 0 \\
u(x,0) = 0, & x > 0 \\
\lim_{t \to 0^+} u(0,t) = \alpha,
\end{cases}
\tag{5.2}
\]

where \(\alpha = F(0) - F(C) < 0\) is a constant. It is easy to verify that the equation and the initial and boundary conditions in (3.4)-(3.7) are equivalent to the corresponding ones in (P), except at \((x,t) = (0,0)\).

One of the main purposes in the present paper is to study qualitatively as well as numerically the problem (P). We find that the problem can also be used to model other related industrial processes, and that it has its own mathematical interest. Therefore, instead of the above specifically defined functions \(A\) and \(B\), we shall consider more general functions. We will shortly define a weak solution for the problem (P) and establish uniqueness of that solution. Existence of a weak solution will be presented afterward. We'll adopt a semi-discrete Galerkin scheme as an approach to the problem. One advantage of this constructive method is that we can use the scheme to easily carry out numerical computation. Finally, we will prove some regularity results and consequently establish existence of a classical solution.

Throughout the paper, we shall always assume that

\[
\begin{cases}
A(\cdot), B(\cdot) \text{ are Lipschitz continuous and monotonically increasing,} \\
0 < a \leq A(\cdot), B(\cdot) \leq b, \\
A'(\cdot) \text{ and } B'(\cdot) \text{ are bounded, } A'(\cdot) \geq a, B'(\cdot) \geq 0
\end{cases}
\tag{5.3}
\]

Note that the above assumptions (5.3) hold for the data in the model problem (3.4)-(3.7). When \(B'(u) = 0\), the problem reduces to a simple Neumann problem.

We now introduce some notations used in the paper. Let \(X\) be any Banach space. We denote by \(L^p_{loc}((0,T), X)\) the space of all function that are locally square integrable in \((0,T)\). The space \(L^p((0,T), X)\) is defined similarly. For any integer \(m\), denote by \(H^m((0,1))\) the usual Sobolev spaces. We shall use \(\| \cdot \|_X\) to indicate the norm in the space \(X\). The notation \(\langle \cdot, \cdot \rangle\) shall always mean appropriate scalar products. Some common notations of functional spaces such as \(C^m\) (continuous up to \(m\)th derivatives) shall be adopted. We begin with the definition of a weak solution of (5.2).
Definition 5.1 A function \( u \in L^2_{loc}([0, \infty); H^1(0, 1)) \) is called a weak solution of (5.2) if for every \( T > 0 \) and every function \( v \in C^1([0, 1] \times [0, T]) \) with \( v(\cdot, T) = 0 \), the following equation holds:
\[
\int_0^T \int_0^1 \left[-A(u)v_t + u_xv_x \right] dx \, dt - \int_0^T B(u(t))v(0, t) \, dt = \int_0^1 A(u_0(z)) v(z, 0) \, dz + B(\alpha) v(0, 0).
\] (5.4)

Remark 5.1 In (5.4), there is an extra term \( B(\alpha)v(0, 0) \), comparing to usual weak formulation for initial boundary problems for parabolic equations. It makes significant difference between our problem (P) and the regular problem. In fact, if \( B(\alpha) = 0 \), the problem reduces to those that have been studied intensively.

Remark 5.2 In a reasonable functional space, our Definition 5.1 for weak solutions in \( L^2_{loc}([0, \infty); H^1((0, 1)) \) has been made in some senses the strongest possible. In fact, since the solution is not expected to have a limit at \((0, 0)\), there may not exist any solution in higher order Sobolev spaces.

The classical solution should be defined in consistence with the conditions (3.4)-(3.7).

Definition 5.2 A weak solution \( u(x, t) \) is called a classical solution if
\[
u, \ u_x, \ u_{xx}, \ u_t \in C([0, 1] \times [0, \infty) - (0, 0)),
\]
and if the limits
\[
\lim_{t \to 0} u_t(0, t), \ \lim_{t \to 0} u_{xx}(0, t), \ \lim_{t \to 0} u_t(0, t)
\]
exist.

We first establish uniqueness of weak solutions.

Theorem 5.1 Problem (P) admits at most one weak solution.

Proof. Assume that \( u_1 \) and \( u_2 \) are two solutions. Set \( w = u_1 - u_2 \). Then for any \( T > 0 \),
\[
\int_0^T \int_0^1 \left\{ [-A(u_2 + w) - A(u_2)]v_t + w_xv_x \right\} dx \, dt - \int_0^T [B(u_2 + w) - B(u_2)]v_t|_{x=0} dt = 0 \quad (5.5)
\]
for any \( v \) satisfying

\[
\begin{align*}
 v_t(x, \cdot) & \in C([0, 1], L^2([0, T])), \\
v_x & \in L^2((0, 1) \times [0, T]), \\
v(\cdot, T) &= 0.
\end{align*}
\]

Consider the function defined by \( v(x, t) := \int_t^T w(x, \tau) d\tau \). Since \( w \in L^1((0, T); H^1(0, T)) \), we have \( (v_t)_x = w_x \) so that \( v_t(x, \cdot) \in C((0, 1); L^2((0, T))) \). Therefore, setting this \( v \) in (5.5) we obtain

\[
\int_0^T \int_0^1 [A(u_2 + w) - A(u_2)] w + \int_0^T \int_t^T w_x \int_t^T w_x + \int_0^T [B(u_2 + w) - B(u_2)] w|_{x=0} dt = 0.
\]

Since \( A \) and \( B \) are monotonic, the first and third integral is non-negative. Also, the second integral can be written as

\[
\int_0^T \int_t^T w_x \int_t^T w_x = \int_0^1 dx \left( \int_0^T w_x(x, \tau) d\tau \right)^2 \geq 0.
\]

From this, we conclude that \( w \equiv 0 \) which implies that \( u_1 = u_2 \). \( \square \)

6 Existence of a weak solution.

In this section, we shall prove that problem (P) has a weak solution. The idea of the proof is as follows: First we construct approximate solutions via a semi-discrete finite difference scheme. Then we estimate the norm of the approximate solution in certain function spaces. Finally, we show that, as the mesh size approaches zero, a subsequence of the approximate solutions converges to a weak solution of (P).

6.1 The finite difference scheme.

Let \( n \geq 3 \) be the number of grid points, \( h = 1/n \) be the mesh size and \( x_j = jh, j = 0, \ldots, n \) be the grid points. We use \( U_j(t) \) to denote the value of an approximate solution at the grid point \( x_j \).
Consider the following semi-discrete finite difference scheme for the problem (P):

\[
\begin{align*}
\dot{U}_j(t) &= \frac{1}{A'_n(U_j)} \cdot \frac{U_{j-1}(t) - 2U_j(t) + U_{j+1}(t)}{h^2}, \quad j = 1, \ldots, n-1, \, t > 0 \\
U_n(t) &= U_{n-1}(t), \quad t \geq 0, \\
U_0(t) &= \frac{1}{B'_n(U_0)} \cdot \frac{U_1(t) - U_0(t)}{h}, \quad t > 0, \\
U_j(0) &= \frac{1}{h} \int_{x_{j-1}}^{x_j} u_0(x) \, dx, \quad j = 1, \ldots, n-1, \\
U_0(0) &= \alpha,
\end{align*}
\]

(P\textsuperscript{h})

where \(A_n\) and \(B_n\) are approximations of \(A\) and \(B\) with the following properties:

\[
\begin{align*}
A_n(\cdot), \quad B_n(\cdot) &\in C^\infty(R), \\
|A_n(u) - A(u)| &\leq h \quad \text{and} \quad |B_n(u) - B(u)| \leq 2h(1 + |u|) \quad \forall u \in R \\
\frac{1}{C} &\leq A'_n(u) \leq C \quad \text{and} \quad h \leq B'_n(u) \leq C \quad \forall u \in R.
\end{align*}
\]

Clearly, such \(A_n\) and \(B_n\) exist. In the sequel, we assume that they are fixed.

\textbf{Lemma 6.1} Assume that \(A_n\) and \(B_n\) satisfy (6.7). Then for every \(u_0 \in L^\infty((0,1))\), problem (P\textsuperscript{h}) has a unique solution for all \(t \in [0, \infty)\).

\textbf{Proof.} Observe that we can use the relation \(U_n = U_{n-1}\) to eliminate \(U_n\) from (P\textsuperscript{n}) to obtain an initial value problem for a system of \(n\) ordinary differential equations with \(n\) unknown functions \(U := (U_0, \cdots, U_{n-1})\). Since \(\frac{1}{A'_n(u)}\) and \(\frac{1}{B'_n(u)}\) are bounded smooth functions of \(u \in R\), the right-hand side of the differential equations for \(U_j\) \((j = 0, \cdots, n-1)\) is Lipschitz continuous and grows at most linearly with respect to \(U_0, \cdots, U_{n-1}\). Hence, by a standard theorem of ordinary differential equations [?], this system has a unique solution for all \(t \in [0, \infty)\). \(\square\)

\section{An a priori estimate}

To estimate the solution of (P\textsuperscript{h}), it is convenient to introduce the following notations:

\[
DU_{j+\frac{1}{2}}(t) := [U_{j+1}(t) - U_j(t)]h, \quad D^2U_j(t) := [U_{j+1} - 2U_j + U_{j-1}]h^2 = [DU_{j+\frac{1}{2}} - DU_{j-\frac{1}{2}}]h.
\]
Lemma 6.2 Assume the condition of Lemma 6.1. Also let \( U = (U_0, \ldots, U_n) \) be the unique solution of \((P^n)\). Then the following hold:

(i) For any \( j = 0, \ldots, n \) and any \( t \geq 0 \),
\[
\min\{U_0(0), \ldots, U_{n-1}(0)\} \leq U_j(t) \leq \max\{U_0(0), \ldots, U_{n-1}(0)\}; \tag{6.8}
\]

(ii) Let \( \hat{B}_n(u) \) and \( \hat{A}_n(u) \) be defined by
\[
\hat{A}_n(u) := \int_0^u sa_n'(s)ds, \quad \hat{B}_n(u) := \int_0^u sb_n'(s)ds. \tag{6.9}
\]
Then for any \( t > 0 \),
\[
\hat{B}_n(U_0(t)) + \sum_{j=1}^{n-1} h \hat{A}_n(U_j(t)) + \int_0^t \sum_{j=1}^n h |D U_j(t)|^2 dt = \hat{B}_n(\alpha) + \sum_{j=1}^{n-1} h \hat{A}_n(U_j(0)). \tag{6.10}
\]

(iii) For any \( t > 0 \),
\[
\int_0^t s \left\{ B'(U_0) \hat{U}_0^2 + \sum_{j=1}^{n-1} h \left[ A_n'(U_j) \hat{U}_j^2 + \frac{1}{A_n'(U_j)} |D^2 U_j|^2 \right] \right\} ds + t \sum_{i=1}^n h |D U_{j-\frac{1}{2}}(t)|^2 \\
+ \hat{B}_n(U_0(t)) + \sum_{j=1}^{n-1} h \hat{A}_n(U_j(t)) = \hat{B}_n(\alpha) + \sum_{j=1}^{n-1} h \hat{A}_n(U_j(0)). \tag{6.11}
\]

Proof. (i) We shall prove (6.8) by a contradiction argument.

Set \( M = \max\{U_0(0), \ldots, U_{n-1}(0)\} \). Assume that the second inequality in (6.8) is not true. Then there exist \( \varepsilon > 0 \) and \( \tau > 0 \) such that a) \( U_j < M + \varepsilon (1 + t) \) for all \( t \in [0, \tau) \) and \( j = 0, \ldots, n-1 \), and b) for some \( j^* \in \{0, \ldots, n-1\} \), \( U_{j^*}(\tau) = M + \varepsilon (1 + \tau) \). We show that this is impossible. In fact, that \( U_{j^*}(t) - [M + \varepsilon (1 + t)] \) obtains a minimum at \( t = \tau \) implies that \( \dot{U}_{j^*}(\tau) \geq \varepsilon > 0 \). On the other hand, if \( j^* = 0 \) then \( \dot{U}_{j^*}(\tau) = \frac{U_1(\tau) - U_0(\tau)}{h B(U_0(\tau))} \leq 0 \) which contradicts \( \dot{U}_{j^*}(\tau) > 0 \). Also, if \( j^* \in \{1, \ldots, n-1\} \), then \( \dot{U}_{j^*}(\tau) = [U_{j^*+1}(\tau) - 2U_{j^*}(\tau) + U_{j^*-1}(\tau)]/[A_n'(U_j)|h|^2 \leq 0 \) which is also impossible. Hence the second inequality in (6.8) must hold for all \( t \in [0, \infty) \). In a similar manner, we can show that the first inequality in (6.8) holds for all \( t \in [0, \infty) \). This proves the first assertion of the Lemma.

(ii) From the differential equation for \( U_j \) we have
\[
0 = \sum_{j=1}^{n-1} h U_j \left\{ A_n'(U_j) \hat{U}_j - [D U_{j+\frac{1}{2}} - D U_{j-\frac{1}{2}}]/h \right\}
\]
\[ = \sum_{j=1}^{n-1} h \frac{d}{dt} A_n(U_j) + \sum_{j=1}^{n} h |DU_{j-\frac{1}{2}}|^2 + U_0[U_1 - U_0]/h \]

where in the second equation, we have used the identity \( DU_{n-\frac{1}{2}} = 0 \). Using the equation for \( U_0 : B_n'(U_0)\dot{U}_0 = [U_1 - U_0]/h \) and the definition of \( \dot{A}_n \) and \( \dot{B}_n \) we then obtain the identity

\[
\frac{d}{dt} \left\{ \dot{B}_n(U_0(t)) + \sum_{j=1}^{n-1} h \dot{A}_n(U_j(t)) \right\} + \sum_{j=1}^{n} h |DU_{j-\frac{1}{2}}(t)|^2 = 0. \tag{6.12}
\]

Integrating this identity over \((0, t)\) we then obtain the second assertion of the Lemma.

(iii) Writing the differential equation for \( U_j \) \((1 \leq j \leq n - 1)\) as \([A_n'(U_j)]^{1/2}\dot{U}_j - [A_n'(U_j)]^{-1/2}D^2U_j = 0\) we obtain

\[
0 = \sum_{j=1}^{n-1} h \left( [A_n'(U_j)]^{1/2}\dot{U}_j - [A_n'(U_j)]^{-1/2}D^2U_j \right)^2
\]

\[
= \sum_{j=1}^{n-1} \left\{ h A_n'(U_j)\dot{U}_j^2 + \frac{h}{A_n'(U_j)}|D^2U_j|^2 - 2\dot{U}_j[DU_j + \frac{1}{h} - DU_{j-\frac{1}{2}}] \right\}
\]

\[
= \sum_{j=1}^{n-1} \left\{ h A_n'(U_j)\dot{U}_j^2 + \frac{h}{A_n'(U_j)}|D^2U_j|^2 + 2DU_{j-\frac{1}{2}}\frac{d}{dt}DU_{j-\frac{1}{2}} + 2\dot{U}_0DU_{\frac{1}{2}} \right\}.
\]

Using the relations \( DU_{n-\frac{1}{2}} = 0 \) and \( DU_{\frac{1}{2}} = B_n'(U_0)\dot{U}_0 \) we then obtain

\[
0 = \sum_{j=1}^{n-1} h \left\{ A_n'(U_j)\dot{U}_j^2 + \frac{1}{A_n'(U_j)}|D^2U_j|^2 \right\} + 2B_n'(0)\dot{U}_0^2 + \frac{d}{dt} \sum_{j=1}^{n} h |DU_{j-\frac{1}{2}}|^2.
\]

It then follows that

\[
\sum_{j=1}^{n-1} h \left\{ tA_n'(U_j)\dot{U}_j^2 + \frac{t}{A_n'(U_j)}|D^2U_j|^2 \right\} + 2tB_n'(0)\dot{U}_0^2 + \frac{d}{dt} \sum_{j=1}^{n} th |DU_{j-\frac{1}{2}}|^2
\]

\[
= \sum_{j=1}^{n} h |DU_{j-\frac{1}{2}}|^2
\]

\[
= -\frac{d}{dt} \left\{ \dot{B}_n(U_0(t)) + \sum_{j=1}^{n-1} h \dot{A}_n(U_j(t)) \right\}. \quad \text{(by (6.10))}
\]

Integrating the above relation over \((0, t)\) we obtain (6.11), thereby completing the proof of the lemma. \( \square \)
6.3 Existence

With the help of Lemma 6.1, we are now ready to prove the existence of a weak solution to (P).

**Theorem 6.3** Assume that $A$ and $B$ satisfies (5.3). Also assume that $u_0 \in L^\infty((0,1))$. Then problem $(P)$ admits a unique solution $u$. In addition, the solution $u$ has the following properties:

\[
\|u\|_{L^\infty((0,1) \times (0,\infty))} \leq \max\{\alpha, \|u_0\|_{L^\infty((0,1))}\},
\]

\[
\int_0^T \int_0^1 \left\{ tA'(u) u_t^2 + \frac{t}{A'(u)} u_{xx}^2 \right\} dx dt + \int_0^1 Tu_x^2(x,T) dx + \int_0^T tB'(u(0,t)) u_t^2(0,t) dt + \int_0^1 \hat{A}(u(x,T)) + \hat{B}(u(0,T)) \leq \int_0^1 \hat{A}_n(u_0(x)) + \hat{B}_n(\alpha).
\]

**Proof.** For every fixed integer $n \geq 3$, let $U_0, \ldots, U_n$ be the solution to problem $(P^n)$. We define functions $u^n(x,t)$ by

\[
u^n(x,t) = U_{j-1} + DU_{j-1} \left( x - jh \right) \quad \forall t \geq 0, x \in [(j-1)h, jh], j = 1, \ldots, n. \quad (6.13)
\]

Clearly, $u^n$ is Lipschitz continuous in $[0,1] \times [0,\infty)$. By (i) of Lemma 6.1,

\[
\|u^n\|_{C^0([0,1] \times [0,\infty))} \leq \max\{\alpha, \|u_0\|_{L^\infty((0,1))}\}.
\]

Also, by (ii) of Lemma 6.2,

\[
\int_0^T \int_0^1 \left( u^n_t \right)^2 dx dt = \int_0^T \sum_{j=1}^n h|DU_{j-1}|^2 \leq \hat{B}_n(\alpha) + \int_0^1 \hat{A}_n(u_0^2(x)) dx \leq \hat{C}
\]

where $\hat{C}$ is a constant independent of $n$. Furthermore,

\[
\int_0^1 |u^n_t|^2 = \frac{h}{3} \sum_{j=1}^n (U_{j-1}^2 + \dot{U}_{j-1} + \dot{U}_j^2)
\]

\[
= \frac{h}{2} (\dot{U}_0^2 + \dot{U}_n^2) + \sum_{i=1}^{n-1} h\dot{U}_j^2
\]

\[
\leq \min \left\{ \frac{h}{\min_{u \in R} B'_n(s)}, \frac{1}{\min_{s \in R} A'_n(s)} \right\} \left\{ B_n'(U_0) \dot{U}_0^2 + \sum_{i=1}^n A_n'(U_i) \dot{U}_i^2 \right\}.
\]

20
It then follows from (6.11) and the boundedness of \( A_n' \) and \( B_n' \) that
\[
\sup_{n \geq 2} \int_0^\infty t(u_n^2)^2 dx dt < \infty.
\]
Hence, the sequence \( \{u^n\}_{n=2}^\infty \) is a bounded sequence in \( L^2((0,T), H^1((0,1))) \) and in \( H^1((\delta,T); L^2((0,1))) \) for any \( 0 < \delta < T < \infty \). Therefore, there exists a function \( u \in L^2_{\text{loc}}([0,\infty); H^1((0,1))) \) and a subsequence \( \{n_i\}_{i=1}^\infty \) such that, as \( i \to \infty \), \( n_i \to \infty \) and
\[
u^n_i \rightharpoonup u \quad \text{for almost every } x \in (0,1), t \in (0,\infty),
\]
weakly in \( L^2((0,T); H^1((0,T))) \) \( \forall T > 0 \),
\[
u^n_i(0,t) \to u(0,t) \quad \text{for a. e. } t \in (0,\infty),
\]
\[
A_n(u^n(x,0)) \to A(u_0(x)) \quad \text{for a.e. } x \in (0,1).
\]
Now we show that \( u \) is a weak solution of problem (P).

Let \( T > 0 \) be any given number and \( v \) be any given smooth function with \( v(\cdot, T) \equiv 0 \). Using the relations \( D^2 U_j = \left[ DU_{j+\frac{1}{2}} - DU_{j-\frac{1}{2}} \right]/h = \left[ u^n_x(x+h,t) - u^n_x(x,t) \right]/h \) for all \( x \in (x_{j-1}, x_j) \) and \( u^n_x(x,t) = DU_{n-\frac{1}{2}} = 0 \) for all \( x \in (1-h, 1) \), we obtain
\[
\sum_{j=1}^{n-1} [A_n(U_j)]_t \int_{(j-1)h}^{jh} v(x,t) dx = \sum_{j=1}^{n-1} D^2 U_j \int_{(j-1)h}^{jh} v(x,t) dx
\]
\[
= \sum_{j=1}^{n-1} \int_{(j-1)h}^{jh} \frac{u^n_x(x+h,t) - u^n_x(x,t)}{h} v(x,t) dx
\]
\[
= - \int_0^1 u^n_x v(x,t) - \frac{v(x-h,t)}{h} dx - DU_{\frac{1}{2}} h \int_{-h}^0 v(x,t) dx
\]
\[
= - \int_0^1 u^n_x v(x,t) - \frac{v(x-h)}{h} dx dt - B_n(u^n(0,t)) \frac{1}{h} \int_{-h}^0 v(x,t) dx.
\]
\[
= - \int_0^1 u^n_x v(x,t) - \frac{v(x-h)}{h} dx dt + \int_0^T B_n(u^n(0,t)) \frac{1}{h} \int_{-h}^0 v_i(x,t) dx dt + B_n(x) \frac{1}{h} \int_{-h}^0 v(x(t)) dt.
\]
To estimate the first term, we define \( \tilde{u}^n \) by
\[
\tilde{u}^n(x,t) = U_j(x,t), \quad t \geq 0, x \in [(j-1)h, jh), j = 1, \ldots, n.
\]
Then we can write
\[
\int_0^T \sum_{j=1}^{n-1} [A_n(U_j)]_t \int_{(j-1)h}^{jh} v(x,t) dx = \int_0^T \int_0^{1-h} A_n(\tilde{u}^n)_t v(x,t) dx dt
\]
\[
= - \int_0^T \int_0^{1-h} A h(\tilde{u}^n) v_i dx dt - \int_0^{1-h} A_n(\tilde{u}^n(0,t)) v(x,0) dx.
\]
Combining this with (6.14) we get

\[
\int_0^T \int_0^1 \left\{ A_n(\tilde{u}^n)u_t + u^n(x,t) \left( \frac{v(x,t) - v(x-h,t)}{h} \right) \right\} dx dt - \int_0^T B_n(u^n(0,t)) \frac{1}{h} \int_{-h}^0 v(x,t) dx
\]

\[
= \int_0^{1-h} A_n(\tilde{u}^n(x,0))v(x,0) dx + B(\alpha) \frac{1}{h} \int_{-h}^0 v(x,0)
\]

(6.16)

Observe that

\[
\int_0^T \int_0^1 |\tilde{u}^n - u^h|^2 = \frac{h^2}{3} \int_0^T \sum_{j=1}^n h|DU_j|^2 \to 0
\]

as \( n \to \infty \), so that as \( i \to \infty \), \( \tilde{u}^{n,i} \to u \) for a.e. \( x \in (0,1), t > 0 \). Thus, taking \( n = n_i \) in (??), sending \( i \) to \( \infty \), and using Lebesgue’s dominated convergence theorem, we obtain

\[
\int_0^T \int_0^1 [-A(u)v_t + u^2 v_x] - \int_0^T B(u(0,t))v_t(0,t) + B(\alpha)v(0,0) + \int_0^1 A(u_0(x))v(x,0) = 0.
\]

This shows that \( u \) is a weak solution of (P).

The estimate for \( u \) in (6.14) follows from Lemma 6.2 and the details are omitted. □

## 7 Results and Conclusions

The numerical data gives good qualitative results when compared to the experimental data (Fig. 2-5), but displays sensitivity to variations in certain parameters, especially when \( T \approx CMC \) (Fig. 6-7). The data for surfactants with \( T < CMC \) have a shallow gradient at first due to the initial condition. However, the diffusion coefficient and the condition of smoothness near \( T = CMC \) causes a steep gradient to occur. This is only mitigated as the saturation level of the air/fluid interface is approached. A low \( T \) to \( CMC \) ratio for monomers helps to explain why equilibrium is reached faster than with micelles, and why there is no steep gradient.

Other factors also appeared to effect the data noticeably. Changes in CMC itself, as well as bulk surfactant concentration occurred during our calculations. In examining the time versus surface tension data, there appears to be a high sensitivity to diffusion parameter, as well as the mesh size and scaling. The mesh sensitivity is especially obvious in figures 8 and 9, which compare 5 point and 20 point meshes.

The original mathematical model, including the numerical results, matches up well, at least qualitatively, when compared to our physical situation. From a practical standpoint,
the results can be used to choose a surfactant that will minimize the surface tension with a relatively small concentration. In the future, a larger mesh size and an adaptive grid, as well as knowing the precise diffusion coefficient, would help give more accurate quantitative results.
cmc = 10e-6

* numerical data point
+ experimental data point
Surface Age 0.1 seconds

- O experimental data
- * numerical results

Surface Tension (dynes/cm) vs. CMC (mol/cc)

10^-1 to 10^0 on the x-axis
10^-6 on the y-axis
2-Et Bu : cmc=1.4e-7
2-Et Bu: cmc=1.4e-7
n-C6, cmc=1e-7
$n-C4: \text{cmc}=2e^{-5}$

![Graph showing surface tension over time](image)
Mathematical Modeling of Electronic Rectifiers

Gregory Kriegsmann  Deborah Alterman  Mary Bond
(Tutor)  U of Michigan  U of Kentucky

Xianhua Huang  Héctor E. Lomelí  Karen Matthews
TU Delft  U of Colorado  U of Illinois at Chicago

Ruth Michler  Brian von Dohlen  Karen Zangara
U of North Texas  U of Michigan  Virginia Tech

August 6, 1996
Mathematical Modeling of Electronic Rectifiers

List of Figures

1  Half Wave Rectifier .............................................. 2
2  General Behavior of $V(t)$ ....................................... 5
3  Oscillations in the Tail of $V(t)$ ............................... 6
4  Half Wave Rectifier with Extra Resistor ...................... 9
5  Asymptotic and Numeric Graph of $\rho$ vs $w_0$ .............. 14
6  $v(t)$ vs. time for $\rho = 0.01$ .................................. 15
7  $v(t)$ vs. time for $\rho = 0.1$ .................................. 16
8  Three Periods of Oscillation for $v(t)$ vs. time ............... 17
9  Two Periods of $v_D$ vs. time ................................. 18
10 Full-Wave Rectifier .............................................. 19
11 $v(t)$ vs. time .................................................. 24
12 $v_1$ vs. time .................................................. 25
13 $i$ vs. time .................................................... 26
14 Oscillations of $v_1$ vs. time .................................... 27

Contents

1  Introduction ...................................................... 1

2  A Half-Wave Rectifier ......................................... 2

3  A Half-Wave Rectifier with an Extra Resistor ............... 9

4  A Full-Wave Rectifier ......................................... 13

1  Introduction

Rectifier circuits are prevalent in modern electronic equipment. These circuits convert the alternating current supplied by power companies into the direct current required by devices such as TVs, stereos, and computers. In this paper we analyze three such circuits. For each circuit, we derive, using Kirchoff’s laws, a system of ODEs. Due to the nonlinear nature of some of the components in our circuit, an exact solution to the ODEs is intractable. Instead, we find an asymptotic solution, using averaging and the method of multiple scales. This solution is well supported with numerical data. Some
slight discrepancies between the asymptotic formula and the numerical output can perhaps be resolved by observing that the ODEs are “stiff”, and the large time required to reach the steady-state can produce errors.

2 A Half-Wave Rectifier

Consider the half-wave rectifier given in Figure (1).

![Figure 1: Half Wave Rectifier](image)

First we will set up the equations governing the circuit and then analyze the equations both numerically and asymptotically. We will show in the end that both methods agree in their approximation of $V$, the voltage across the resistor, to an ord(1) approximation.

For this circuit, $V_I$ is the input voltage and is given by the equation

$$V_I = 117 \sin(\omega t),$$  \hfill (1)

where 117 volts is the input from the power source, $\omega = 2\pi f$, and $f = 60$ cycles per second. The transformer’s output voltage, $V_o$, is given by $V_o = MV_I$, where $M = \frac{N_o}{N_i}$ for $N_o$ and $N_i$ the number of turns of the coil on the output and input sides of the transformer, respectively. This system gives us the following set of equations:

$$V_o = M(117\text{volts})\sin(\omega t),$$  \hfill (2)
$$I_D = I_o(\epsilon^{\frac{V}{V_o}} - 1),$$  \hfill (3)
$$I_R = \frac{V}{R},$$  \hfill (4)
$$C\frac{dV}{dt} = I_C.$$  \hfill (5)
Equation (3) comes from the physics of the diode where $I_o$ is a constant dependent on geometry and the material, $\kappa$ is Boltzman's constant, $T$ is the temperature in degrees Kelvin, and $V_D$ is the voltage across the diode. Equation (4) is Ohm’s law with $V$, the voltage across the resistor, and finally in Equation (5), $C$ is the value of the capacitor.

From the equation for conservation of charge,

$$ I_D = I_C + I_R, \quad (6) $$

we see that we can substitute Equations (3)-(5) into (6) to obtain

$$ C \frac{dV}{dt} + \frac{V}{R} = I_o (e^{\frac{V}{RT}} - 1). \quad (7) $$

Finally we substitute $V_o = V_D + V$ into the last equation to obtain

$$ C \frac{dV}{dt} + \frac{V}{R} = I_o (e^{\frac{V_o - V}{RT}} - 1). \quad (8) $$

which has the initial condition $V(0) = 0$.

Next we will make Equation (8) dimensionless by letting $v = \frac{V}{117 MVolts}$ and $t = \omega t$. Upon substitution into (8) and recalling Equation (2) we obtain

$$ \frac{dv}{dt} + \frac{v}{\omega RC} = \frac{I_o}{117 M \omega C} (e^{\alpha (\sin t - v)} - 1), \quad (9) $$

where we define $\alpha = \frac{117 M Volts}{\kappa T}$. Setting $\varepsilon = \frac{1}{\omega RC}$ and $\lambda = \frac{R I_o}{117 M Volts}$ we rewrite Equation (9) as

$$ \frac{dv}{dt} + \varepsilon v = \varepsilon \lambda (e^{\alpha (\sin t - v)} - 1). \quad (10) $$

We have now simplified the equation with six parameters to one that has three dimensionless parameters, namely $\varepsilon$, $\lambda$, and $\alpha$. The initial condition for Equation (10) is $v(0) = 0$.

For the numerical analysis we took $\alpha = 40$, $\lambda = 1$, and $\varepsilon = 10^{-4}$ as typical values for the three parameters. First we transformed Equation (10) into an autonomous 3rd order system by letting $u = \sin t$ and $w = \cos t$ to get

$$ \frac{du}{dt} = w, \quad (11) $$
$$ \frac{dw}{dt} = -u, \quad (12) $$
$$ \frac{dv}{dt} + \varepsilon v = \varepsilon \lambda (e^{\alpha (u - v)} - 1), \quad (13) $$
with boundary conditions \( u(0) = 0 = w(0) = 1 \).

Using DSTOOL and a fourth order Runge-Kutta method with step size 0.01 we obtain the graph in Figure (2) for the system (13) as \( t \) runs between \( t = 0 \) and \( t = 2000 \). The computations took approximately 15 minutes.

Upon closer examination of the curve as it approaches its asymptote we note that the curve is not straight but oscillates as depicted in Figure (3). This resembles the sawtooth pictures associated with rectifiers as seen in textbooks.

To evaluate (10) asymptotically we employ the method of multiple scales by letting \( v = V(t, \tau) \) where \( \tau = \varepsilon t \) and \( \frac{dv}{dt} = \frac{\partial V}{\partial t} + \varepsilon \frac{\partial V}{\partial \varepsilon} \). Next we expand \( V \) in terms of \( \varepsilon \), the small parameter, keeping only terms up to \( \text{ord}(\varepsilon^2) \) so that

\[
V(t, \tau) = V_0(t, \tau) + \varepsilon V_1(t, \tau) + \varepsilon^2 V_2(t, \tau).
\]  

(14)

Substituting Equation (14) into (10) we find

\[
\begin{align*}
V_{0t} + \varepsilon V_{1t} + \varepsilon^2 V_{2t} + \varepsilon V_{0r} + \varepsilon^2 V_{1r} + \varepsilon V_0 + \varepsilon^2 V_1 & = \varepsilon \lambda (e^{\alpha (\sin t - V_0)} - e^{\alpha V_0}) - 1), \\
& = \varepsilon \lambda (e^{\alpha (\sin t - V_0)}(1 - \varepsilon \alpha V_0)(1 - \varepsilon^2 \alpha V_2) - 1).
\end{align*}
\]  

(15)

By equating terms of the same order in (15) we obtain

\[
\begin{align*}
\text{ord}(1) : V_{0t} & = 0, \\
\text{ord}(\varepsilon) : V_{1t} + V_{0r} + V_0 & = \lambda (e^{\alpha (\sin t - V_0)} - 1), \\
\text{ord}(\varepsilon^2) : V_{2t} + V_{1r} + V_1 & = \lambda (e^{\alpha (\sin t - V_0)} - \alpha V_1).
\end{align*}
\]  

(16) (17) (18)

From the ord(1) equation we see that \( V_0 \) depends only on \( \tau \) so that \( V_0 = V_0(\tau) \) and the boundary condition on \( V_0 \) is \( V_0(0) = 0 \). In order to determine \( V_0(\tau) \) completely we need to solve the \( \text{ord}(\varepsilon) \) equation by integrating with respect to \( t \) to get

\[
V_1 = -t(V_0 + V_{0r} + \lambda) + \lambda e^{-\alpha V_0} \int_0^t e^{\alpha \sin \tau \, ds} + A_1(\tau).
\]  

(19)

We need to ensure that as \( t \) gets large, \( V_1 \) is bounded. Thus, in (19) we show that the 1st two terms on the righthand side are bounded. Because of the periodic nature of the equation, we let \( t = 2\pi n \). Then (19) becomes:

\[
\begin{align*}
V_1 & = -2\pi n(V_0 + V_{0r} + \lambda) + \lambda e^{-\alpha V_0} \int_0^{2\pi n} e^{\alpha \sin \tau \, ds} + A_1(\tau), \\
& = 2\pi n\left(\frac{\lambda e^{-\alpha V_0}}{2\pi n} \int_0^{2\pi n} e^{\alpha \sin \tau \, ds} - (V_0 + V_{0r} + \lambda)) + A_1(\tau).
\end{align*}
\]  

(20)
Figure 2: General Behavior of $V(t)$
Figure 3: Oscillations in the Tail of $V(t)$
Mathematical Modeling of Electronic Rectifiers

The integral can be written as

$$\int_0^{2\pi n} e^{\alpha \sin s} ds = n \int_0^{2\pi} e^{\alpha \sin s} ds.$$  \hspace{1cm} (21)

We can then substitute the new integral into Equation (20) to obtain

$$V_1 = 2\pi n \left( \frac{\lambda e^{-\alpha V_0}}{2\pi} \int_0^{2\pi} e^{\alpha \sin s} ds - (V_0 + V_{0\tau} + \lambda) \right) + A_1(\tau).$$  \hspace{1cm} (22)

Since $n \to \infty$ we set

$$\frac{\lambda e^{-\alpha V_0}}{2\pi} \int_0^{2\pi} e^{\alpha \sin s} ds - (V_0 + V_{0\tau} + \lambda) = 0$$  \hspace{1cm} (23)

in (22). Let $I = \frac{1}{2\pi} \int_0^{2\pi} e^{\alpha \sin s} ds$ which can be estimated using Laplace's Method about $t = \frac{\pi}{2}$, which is the maximum of $\sin(s)$ over $[0, 2\pi]$, so that

$$I \approx \frac{1}{\sqrt{2\pi \alpha}} e^{\alpha}.$$  \hspace{1cm} (24)

Using Maple to solve for $I$ when $\alpha = 40$ we get $I = 0.9359 \times 10^{17}$, while the asymptotic approximation gives $I = 0.9329 \times 10^{17}$. The relative error is $0.32\%$.

Next we try to solve Equation (23) using the estimate from Laplace's Method so that the equation becomes

$$V_{0\tau} + V_0 = \lambda \left( \frac{e^{\alpha (1-V_0)}}{\sqrt{2\pi \alpha}} - 1 \right),$$  \hspace{1cm} (25)

with $V_0(0) = 0$. Solving for $\frac{dV_0}{d\tau}$ we get

$$\frac{dV_0}{d\tau} = \lambda \left( \frac{e^{\alpha (1-V_0)}}{\sqrt{2\pi \alpha}} - 1 \right) - V_0.$$  \hspace{1cm} (26)

At $V_0 = 0$ we obtain

$$\frac{dV_0}{d\tau} = \lambda \left( \frac{e^{\alpha}}{\sqrt{2\pi \alpha}} - 1 \right),$$  \hspace{1cm} (27)

which is positive. This means that the curve of $\tau$ versus $V_0(\tau)$ increases until it reaches the steady state and matches Figure (2) which was obtained numerically.
Next we approximate the steady state by letting \( \frac{dV_0}{dt} = 0 \) in (26) to find

\[
0 = \lambda \left( \frac{e^{\alpha(1-V_0)}}{\sqrt{2\pi \alpha}} - 1 \right) - V_0. \tag{28}
\]

Since \( \alpha \gg 1 \) we need \( V_0 \sim 1 \) to prevent a blow up, so this implies the steady state is approximately 1. We try to solve for \( V_0 \) by putting the exponential term on one side and taking the logarithm of both sides to obtain

\[
e^{\alpha(1-V_0)} = \frac{\sqrt{2\pi \alpha}}{\lambda}(\lambda + V_0), \tag{29}
\]

\[
\alpha(1-V_0) = \ln \left( \frac{\sqrt{2\pi \alpha}}{\lambda} \right) + \ln (\lambda + V_0) - \ln \lambda. \tag{30}
\]

On the right-hand side of (30), the \( \ln \sqrt{2\pi \alpha} \) is the largest term numerically so we drop the other terms and solve for \( V_0 \) to obtain

\[
V_0 = \frac{\alpha - \frac{1}{2} \ln \left( \frac{\sqrt{2\pi \alpha}}{\lambda} \right)}{\alpha}. \tag{31}
\]

as the approximation to the steady state. For \( \alpha = 40 \) the approximation to the steady state gives \( V_0 = .931 \) while the numerical approximation is .915.

Returning to equation (19) and substituting for \( \frac{dV_0}{dt} + V_0 + \lambda \) we get

\[
V_1 = A_1(\tau) + \lambda e^{-\alpha V_0} \left( -\frac{t}{2\pi} \int_0^{2\pi} e^{\alpha \sin s} ds + \int_0^t e^{\alpha \sin s} ds \right). \tag{32}
\]

Letting

\[
q(t) = -\frac{t}{2\pi} \int_0^{2\pi} e^{\alpha \sin s} ds + \int_0^t e^{\alpha \sin s} ds, \tag{33}
\]

we see that \( q(t) \) is \( 2\pi \) periodic, i.e. \( q(t + 2\pi) = q(t), \) and that we can substitute \( I \) into the first term to get

\[
q(t) = -tI + \int_0^t e^{\alpha \sin s} ds. \tag{34}
\]

Since the major contribution of the second term occurs at \( \frac{\pi}{2} \) we again use Laplace's Method on the second integral as \( \alpha \to \infty \) to obtain

\[
\int_0^t e^{\alpha \sin s} ds \approx e^{\alpha \frac{\sqrt{2}}{\sqrt{\alpha}}} \int_{-\infty}^{\frac{\pi}{2} - \frac{\pi}{2}} e^{-s^2} ds. \tag{35}
\]
Using this approximation (35) we obtain the following estimate for \( q(t) \):

\[
q(t) \approx e^a \sqrt{\frac{2}{\alpha}} \int_{-\infty}^{\sqrt{\frac{2I}{\alpha}(1-\frac{t}{\alpha})}} e^{-s^2} ds - tI. 
\]  

(36)

The \(-tI\) term represents the straight line in the sawtooth picture from the numerical calculations as shown in Figure (3). The integral has a negligible contribution for \( t < \frac{\pi}{2} \), makes most of its contribution near \( \frac{\pi}{2} \), and is approximately constant beyond \( \frac{\pi}{2} \). Therefore the asymptotic and numerical approximations for the circuit agree.

3 A Half-Wave Rectifier with an Extra Resistor

Next we will add into the problem another resistor before the diode in order to model loss in the transformer. Our circuit now looks like Figure (4):

![Figure 4: Half Wave Rectifier with Extra Resistor](image)

In the new system of equations, equations (2)-(6) are unchanged from our original system, but the sum of the voltages becomes

\[-V_v \sin (\omega t) + rI_D + V_D + V = 0\]  

(37)

with the addition of the second resistor. After substitution our second order system is

\[
C \frac{dV}{dt} + \frac{V}{R} = I_o(e^{\frac{V}{R}} - 1), \]

(38)

\[
rI_o(e^{\frac{V}{R}} - 1) + V_D + V = V_v \sin (\omega t). \]

(39)
Again we make the equations dimensionless by letting
\[ v = \frac{V}{V_o}, \]  
\[ v_D = \frac{V_D}{V_o}, \]  
\[ t = \omega t. \]

The system in its dimensionless form is
\[ \frac{dv}{dt} + \varepsilon v = \varepsilon \lambda (e^{\alpha v_D} - 1), \]  
\[ \lambda \rho (e^{\alpha v_D} - 1) + v_D + v = \sin t. \]

There are now four parameters \( \varepsilon, \lambda, \alpha, \) and \( \rho \), of which the first three are as in the original problem and \( \rho = \frac{v}{E} \). For a good transformer \( \rho \ll 1 \), that is, the resistance of the transformer is a small fraction of the load resistance. The initial conditions are \( v(0) = 0 = v_D(0) \).

As in the first problem we will use the method of multiple scales to study the voltage across the resistor, \( R \). We will let \( v \) and \( v_D \) depend on \( t \) and \( \tau \) and expand them both in terms of \( \varepsilon \). Expanding only to terms of \( \text{ord}(\varepsilon) \) we get
\[ v = w(t, \tau) = w_0(t, \tau) + \varepsilon w_1(t, \tau) + \varepsilon^2 w_2(t, \tau), \]  
\[ v_D = u(t, \tau) = u_0(t, \tau) + \varepsilon u_1(t, \tau) + \varepsilon^2 u_2(t, \tau). \]

After making the substitution we collect terms of similar order to get
\[ \text{ord}(1) : \quad w_{0t} = 0, \]  
\[ \lambda \rho (e^{\alpha u_0} - 1) + u_0 + w_0 = \sin t, \]  
\[ \text{ord}(\varepsilon) : \quad w_{1t} + w_{0\tau} + w_0 = \lambda (e^{\alpha u_0} - 1), \]  
\[ \lambda \rho \alpha u_1 e^{\alpha u_0} + w_1 + u_1 = 0. \]

From \( w_{0t} = 0 \) we find that \( w_0 = w_0(\tau) \), a function of \( \tau \) only. Next we integrate (49) with respect to \( t \) to solve for \( w_1 \), and obtain
\[ w_1 = -t(w_{0\tau} + w_0 + \lambda) + \lambda \int_0^t e^{\alpha u_0(s, \tau)} ds + A_1(\tau). \]

From (48) we see that \( u_0 \) is periodic so we let \( t = 2\pi n \) as \( n \to \infty \) in (51) to get
\[ w_1 = -2\pi n(w_{0\tau} + w_0 + \lambda) + \lambda \int_0^{2\pi n} e^{\alpha u_0(s, \tau)} ds + A_1(\tau). \]
We use (21) to simplify Equation (52) to obtain

\[ w_1 = -2\pi n (w_0 + w_0 + \lambda - \frac{\lambda}{2\pi} \int_0^{2\pi} e^{\alpha u_0(s,\tau)} ds) + A_1(\tau). \]  

(53)

As \( n \to \infty \), in order to bound \( w_1 \), we require

\[ w_0 + w_0 + \lambda = \frac{\lambda}{2\pi} \int_0^{2\pi} e^{\alpha u_0(s,\tau)} ds. \]  

(54)

Let \( I = \frac{1}{2\pi} \int_0^{2\pi} e^{\alpha u_0(s,\tau)} ds \) and let \( t_0 = \max_{[0,2\pi]}(u_0) \). Using Laplace's Method we get the following approximation for \( I \):

\[ I \approx \frac{e^{au_0^*}}{\sqrt{2\pi \alpha (u_0^*)''}}, \]  

(55)

where \( u_0(t,\tau) = u_0^* + p(u_0^*)' - \frac{p^2}{2}(u_0^*)'' \) and \( t = t_0 + p \). Since \( t_0 \) is a critical point then \( (u_0^*)' = 0 \) and \( (u_0^*)'' < 0 \). In order to find \( t_0 \) we differentiate (48) with respect to \( t \) to obtain

\[ \lambda \rho (e^{au_0} \alpha u_0) + u_0 + w_0 = \cos t. \]  

(56)

Recall that \( w_0 = 0 \) as \( w_0 = w_0(\tau) \) so that simplifying the above equation gives

\[ u_0 + \frac{\cos t}{\alpha \lambda \rho e^{au_0} + 1}. \]  

(57)

Since \( u_0 = 0 \) when \( \cos t = 0 \) then the critical points are \( t = \frac{\pi}{2}, \frac{3\pi}{2} \). Differentiating a second time, we obtain

\[ (1 + \alpha \lambda \rho e^{au_0}) \frac{d^2 u_0}{dt^2} + \alpha^2 \lambda \rho e^{au_0} \left( \frac{du_0}{dt} \right)^2 = -\sin t. \]  

(58)

We solve for the second derivative and recall that we will be evaluating it at the critical point. Thus the first derivative is zero which leads to the simplified form

\[ \frac{d^2 u_0}{dt^2} = \frac{-\sin t}{1 + \alpha \lambda \rho e^{au_0}}. \]  

(59)

From this we see that \( t = \frac{\pi}{2} \) is a maximum and \( t = \frac{3\pi}{2} \) is a minimum. Since the maximum occurs at \( t_0 = \frac{\pi}{2} \), which is the same as in the original problem, then

\[ I \approx \frac{e^{au_0^*} \sqrt{1 + \alpha \lambda \rho e^{au_0^*}}}{\sqrt{2\pi}}. \]  

(60)
where \( w_0^* = u_0^* (\frac{\tau}{\tau}, \tau). \) Therefore we would like to solve the system

\[
w_{0r} + w_0 = \lambda \left( e^{\alpha u_0^*} \frac{\sqrt{1 + \alpha \lambda \rho e^{\alpha u_0^*}}}{\sqrt{2\alpha \pi}} \right) - \lambda, \tag{61}
\]

\[
\lambda \rho (e^{\alpha u_0^*} - 1) + u_0^* = 1 - w_0. \tag{62}
\]

In order to find the steady state of the system let \( w_{0r} = 0 \) and solve (62) for \( w_0. \) Inserting this into (61) we obtain the following equation for \( u_0^*:\)

\[
K(u_0^*) \equiv 1 - u_0^* - \lambda - \lambda \rho (e^{\alpha u_0^*} - 1) + \lambda \frac{e^{\alpha u_0^*} \sqrt{1 + \alpha \lambda \rho e^{\alpha u_0^*}}}{\sqrt{2\alpha \pi}} = 0. \tag{63}
\]

Taking \( \lambda = 1 \) and \( \alpha = 40, \) as we did with the first circuit, we now try to find a first order approximation for \( w_0, \) i.e. the voltage across the resistor, for various values of \( \rho. \) To do this we need to substitute the root of \( K(u_0^*) = 0 \) into (62) to find \( w_0. \) Maple was used to solve for \( w_0 \) using the formulas from the asymptotic analysis and DSTOOL was used to find an approximation to \( w_0 \) from the original dimensionless equations for our circuit (43)-(44). In the numerical analysis a 4th order Runge-Kutta method was used with a step size of 0.01 and \( t \) ranged from 0 to 20,000. The final point was used as the numerical approximation. Table (1) is a comparison of the results found from the asymptotic and numerical analyses for \( \rho \) ranging from 0 to 1.

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>( w_0 ) (asymptotic)</th>
<th>( w_0 ) (numerical)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.91468</td>
<td>.91433</td>
</tr>
<tr>
<td>.01</td>
<td>.82762</td>
<td>.80506</td>
</tr>
<tr>
<td>.02</td>
<td>.77244</td>
<td>.73833</td>
</tr>
<tr>
<td>.05</td>
<td>.65903</td>
<td>.60418</td>
</tr>
<tr>
<td>.1</td>
<td>.53737</td>
<td>.46451</td>
</tr>
<tr>
<td>.15</td>
<td>.45292</td>
<td>.37021</td>
</tr>
<tr>
<td>.2</td>
<td>.38877</td>
<td>.30012</td>
</tr>
<tr>
<td>.3</td>
<td>.29588</td>
<td>.20157</td>
</tr>
<tr>
<td>.4</td>
<td>.23124</td>
<td>.13584</td>
</tr>
<tr>
<td>.5</td>
<td>.18374</td>
<td>.09059</td>
</tr>
<tr>
<td>.6</td>
<td>.14765</td>
<td>.05767</td>
</tr>
<tr>
<td>1</td>
<td>.06622</td>
<td>.00512</td>
</tr>
</tbody>
</table>

We can see from Figure (4) that both the numerical and asymptotic...
approaches led to graphs that have the same shape for \( \rho \) vs. \( w_0 \), except that the asymptotic solutions are higher than the numerical solutions. The estimates are close for \( \rho \) near 0 and get worse as \( \rho \) approaches 1. These errors may be due to the stiff ODEs (43)-(44). Figure (5) shows the data from Table (1), where the bottom curve represents the numerical data.

Note that it tends asymptotically to the limit \( w_0 \) discussed previously. As we can see, with \( \rho = .01 \), the values of \( v \) are larger than for the larger value \( \rho = .1 \). See Figure (6) and (7).

Upon zooming in on the tail of Figure (6) we see there are small oscillations on the time scale \( t \) rather than the overall behavior on time scale \( \tau \). Figure (8) represents three periods of this oscillation.

Figure (9) is a graph of two periods of \( v_D \) which was computed using Maple. While \( v \) changes on the slower \( \tau \) time scale, we see \( v_D \) oscillating on the fast \( t \) time scale starting at \( t = 1000 \).

### 4 A Full-Wave Rectifier

In the final problem that we analyzed, we added another capacitor \( (C_1) \), inductor \( (L) \), and diode to act as a filter to eliminate the ripple effect in the output voltage. In this fullwave rectifier, see Figure (10), \( L \) is known and \( E \) is half of the transformer's output voltage, where

\[
E = m v_0 \sin (\omega t). \tag{64}
\]

For two diodes, the current from diode \( j \) is given by

\[
I_{D_j} = I_0 (e^{\frac{V_{D_j}}{R_0}} - 1), \tag{65}
\]

for \( j = 1, 2 \). Similarly \( V_{D_j} \) is the voltage across the diode for \( j = 1, 2 \). By using Ohm's law and Kirchoff's equation we have

\[
\begin{align*}
C_1 \frac{dV_1}{dt} + \frac{V_1}{R} &= I_L, \tag{66} \\
C_1 \frac{dV_1}{dt} + I_L &= I, \tag{67} \\
I_{D_1} + I_{D_2} &= I, \tag{68} \\
L \frac{dI_L}{dt} &= V_1 - V, \tag{69} \\
-E + V_{D_1} + V_1 &= 0, \tag{70} \\
E + V_{D_2} + V_1 &= 0. \tag{71}
\end{align*}
\]
Figure 5: Asymptotic and Numeric Graph of $\rho$ v.s $w_0$
Figure 6: $v(t)$ vs. time for $\rho = 0.01$
Figure 7: $v(t)$ vs. time for $\rho = 0.1$
Figure 8: Three Periods of Oscillation for $v(t)$ vs. time
Figure 9: Two Periods of $v_D$ vs. time
Mathematical Modeling of Electronic Rectifiers

![Figure 10: Full-Wave Rectifier](image)

From the above equations, we have the following unknown parameters: \( V, V_1, V_{D_1}, V_{D_2}, I_{D_1}, I_{D_2}, I, \) and \( I_L. \)

By examining the differential equations (66), (67), and (69) we wish to combine them into three equations involving \( V_1, V, \) and \( I_L. \) Using Equations (68), (70), and (71) we see that the only equation that needs to be changed is (67) which becomes

\[
C_1 \frac{dV_1}{dt} = -I_L + 2I_o e^{\frac{-V_1}{2U}} \cosh \left( \frac{V_M}{kT} \sin (\omega t) \right) - 2I_o, \tag{72}
\]

where \( V_M = V_o M. \) We rescale the equation by making the following substitutions

\[
t = \omega t, \tag{73}
\]

\[
V_1 = V_M v_1, \tag{74}
\]

\[
V = V_M v, \tag{75}
\]

\[
I_L = I_o i, \tag{76}
\]

and let \( \alpha = \frac{V_M}{C_1}, \beta = \frac{R^2C}{L}, \) and \( \gamma = \frac{C}{C_1}, \) with \( \varepsilon \) and \( \lambda \) as defined in the first two circuits.

After simplification, the dimensionless differential equations are

\[
\frac{dv}{dt} = -\varepsilon v + \lambda \varepsilon i, \tag{78}
\]

\[
\frac{di}{dt} = \frac{\varepsilon \beta}{\lambda} (v_1 - v), \tag{79}
\]
\[ \frac{dv_1}{dt} = \gamma \lambda \varepsilon (-i + 2e^{-\alpha x_1} \cosh (\alpha \sin t) - 2). \] \tag{80}

Now we have reduced the unknown parameters to the following five: \( \varepsilon, \lambda, \alpha, \beta, \) and \( \gamma. \)

Once again we evaluate the system asymptotically using the method of multiple scales by letting \( v, i, \) and \( v_1 \) be functions of \( t \) and \( \tau \) and expanding in terms of \( \varepsilon \) so that
\[ v = x_0(t, \tau) + \varepsilon x_1(t, \tau) + O(\varepsilon^2), \] \tag{81}
\[ i = y_0(t, \tau) + \varepsilon y_1(t, \tau) + O(\varepsilon^2), \] \tag{82}
\[ v_1 = z_0(t, \tau) + \varepsilon z_1(t, \tau) + O(\varepsilon^2). \] \tag{83}

After substituting the above expansion we collect terms of similar order and find that
\begin{align*}
\text{ord}(1) & : x_{0t} = y_{0\tau} = z_{0t} = 0, \tag{84} \\
\text{ord}(\varepsilon) & : x_{\theta\tau} + x_{1t} = -x_0 + \lambda y_0, \tag{85} \\
 & \quad y_{\theta\tau} + y_{1t} = \frac{\beta}{\lambda} x_0 - x_0, \tag{86} \\
 & \quad z_{\theta\tau} + z_{1t} = \gamma \lambda (-y_0 + 2e^{-\alpha z_0} \cosh (\alpha \sin t) - 2). \tag{87}
\end{align*}

Thus we have from Equation (84) that \( x_0, y_0, \) and \( z_0 \) depend only on \( \tau, \) i.e. \( x_0 = x_0(\tau), \) \( y_0 = y_0(\tau), \) and \( z_0 = z_0(\tau). \) From Equations (85)-(87) we have, after integrating with respect to \( t, \) that
\begin{align*}
x_1 &= -(x_0 + x_{\theta\tau} - \lambda y_0)t + A_1(\tau), \tag{88} \\
y_1 &= \frac{\beta}{\lambda} (x_0 - x_0) - y_{\theta\tau} t + A_2(\tau), \tag{89} \\
z_1 &= \gamma \lambda t (-y_0 + \frac{2e^{-\alpha z_0}}{t} \int_0^t \cosh (\alpha \sin s) ds - 2 - \frac{z_{0\tau}}{\gamma \lambda} ) + A_3(\tau). \tag{90}
\end{align*}

Since we need \( x_1, y_1, \) and \( z_1 \) bounded, then by similar arguments used earlier, we find that
\begin{align*}
x_0 + x_{\theta\tau} - \lambda y_0 &= 0, \tag{91} \\
\frac{\beta}{\lambda} (x_0 - x_0) - y_{\theta\tau} &= 0, \tag{92} \\
-y_0 + \frac{2e^{-\alpha z_0}}{2\pi} \int_0^{2\pi} \cosh (\alpha \sin s) ds - 2 - \frac{z_{0\tau}}{\gamma \lambda} &= 0. \tag{93}
\end{align*}
From these equations we can find a fixed point solution \((x^*_0, y^*_0, z^*_0)\) by solving the equations

\[
\begin{align*}
z^*_0 &= x^*_0 = \lambda y^*_0, \\
z^*_0 &= 2\lambda e^{\alpha s^*_0}(K - 1),
\end{align*}
\]

where \(K = \int_0^{2\pi} \cosh(\alpha \sin s)ds\). Again the periodic behavior of the integral allows us to rewrite Equation \((93)\) as we have done in previous sections.

First we examine the local stability for this fixed point solution. We will linearize this system at \((x^*_0, y^*_0, z^*_0)\). The linear system has the following matrix.

\[
\begin{pmatrix}
-1 & \lambda & 0 \\
\frac{-\rho}{\lambda} & 0 & \frac{\rho}{\lambda} \\
0 & -\lambda \gamma & -\rho
\end{pmatrix},
\]

where \(\rho = 2K\alpha \lambda \gamma e^{-\alpha s^*_0}\). We can write the corresponding characteristic polynomial as

\[
p(\mu) = \mu^3 + (1 + \rho)\mu^2 + (\beta \gamma + \beta + \rho)\mu + \beta(\gamma + \rho).
\]

In order to decide the stability of the fixed point \((x^*_0, y^*_0, z^*_0)\), we need to show that the above cubic polynomial has only roots with negative real part.

**Lemma 1** For

\[
f(z) = z^3 + a_2 z^2 + a_3 z + a_4,
\]

with real, positive coefficients and \(a_3 > \frac{a_4}{a_2}\), \(f(z)\) has only roots with negative real part.

Since the coefficients in the cubic satisfy the Lemma, \((x^*_0, y^*_0, z^*_0)\) is asymptotically stable. We will now address the issue of global stability of the fixed point by constructing a suitable Liapunov function. We first shift the steady state solution to the origin by the following:

\[
\begin{align*}
x &= x^* + u, \\
y &= y^* + v, \\
z &= z^* + w.
\end{align*}
\]
Then we obtain

$$u_r = u - \lambda v, \quad (102)$$

$$v_r = \frac{\beta}{\lambda} (w - u), \quad (103)$$

$$w_r = \gamma \lambda (-v + 2K e^{-\alpha z^*} (e^{-\alpha w} - 1)). \quad (104)$$

We will use the following Liapunov function

$$H(u, v, w) = \frac{1}{2} (\gamma \beta u^2 + \lambda^2 \gamma v^2 + \beta w^2). \quad (105)$$

If \((u(\tau), v(\tau), w(\tau))\) is any solution of the system (102)-(104) then by direct calculation we have

$$\frac{dH}{d\tau} = \nabla H \cdot \begin{pmatrix} u_r \\ v_r \\ w_r \end{pmatrix}, \quad (106)$$

$$= -\gamma \beta u^2 + 2\beta \gamma \lambda K e^{-\alpha z^*} (w(e^{-\alpha w} - 1)). \quad (107)$$

We notice that \(\frac{dH}{d\tau} \leq 0\), and \(\frac{dH}{d\tau} = 0\) only if \((u, v, w) \in \{0\} \times R \times \{0\}\), which does not contain any other solution except for \(u = v = w = 0\). This implies the system is globally asymptotically stable. For more details on Liapunov method see [5].

Now we want to investigate the magnitude and period of the ripple. We obtain no ripple from integrating the equations \(x_1\) and \(y_1\). Recalling that \(x_1\) corresponds to the output voltage, \(V\), we observe that the ripple is significantly lower, i.e., \(\text{ord}(\varepsilon^2)\), then for the half-wave rectifier. This is the result of the additional diode and filtering components.

However in Equation (93) we have

$$y_0 + 2 + \frac{z_0}{\gamma \lambda} = \frac{e^{-\alpha s}}{\pi} \int_0^{2\pi} \cosh (\alpha s) ds,$$  

$$\approx 2e^{-\alpha s} \frac{e^{\alpha}}{\sqrt{2\pi \alpha}}, \quad (108)$$

$$\approx 2e^{-\alpha s} \frac{e^{\alpha}}{\sqrt{2\pi \alpha}}, \quad (109)$$

where

$$\frac{1}{2\pi} \int_0^{2\pi} e^{\alpha \sin s} ds \approx \frac{e^{\alpha}}{\sqrt{2\pi \alpha}}, \quad (110)$$
as was shown previously. Substituting into (90) we obtain

\[ z_1 = \gamma \lambda t (-2 e^{-\alpha z_0} \frac{e^{\alpha t}}{\sqrt{2\pi \alpha}} + \frac{2e^{-\alpha z_0}}{t} \int_0^t \cosh (\alpha \sin s) ds), \] (111)

\[ = 2 \gamma \lambda e^{-\alpha z_0} (\int_0^t \cosh (\alpha \sin s) ds - \frac{e^{\alpha t}}{\sqrt{2\pi \alpha}}) + A_3(\tau). \] (112)

Let \( q_1(t) = \int_0^t \cosh (\alpha \sin s) ds - \frac{e^{\alpha t}}{\sqrt{2\pi \alpha}}. \)

We examine the above integral, similar to previous problems,

\[ \int_0^t \cosh (\alpha \sin s) ds = \frac{e^{\alpha \sqrt{3} (t - \frac{3}{2})}}{\sqrt{2\alpha}} \int_{-\infty}^{\sqrt{3} (t - \frac{3}{2})} e^{-s^2} ds \] (113)

\[ + \frac{e^{\alpha \sqrt{3} (t - \frac{3}{2})}}{\sqrt{2\alpha}} \int_{-\infty}^{\sqrt{3} (t - \frac{3}{2})} e^{-s^2} ds, \] (114)

\[ = \frac{e^{\alpha \sqrt{3} (J_1 + J_2)}}{\sqrt{2\alpha}}. \] (115)

Note that \( J_1 \) contributes near \( t = \frac{\pi}{2} \), and \( J_2 \) contributes near \( t = \frac{3\pi}{2} \) (similar to previous problems). Thus,

\[ q_1(t) = \frac{e^{\alpha \sqrt{3} (J_1 + J_2 - \frac{t}{\sqrt{3}})}}{\sqrt{2\alpha}}, \] (116)

and we have the periodic jumps occurring at \( \frac{\pi}{2} \) and \( \frac{3\pi}{2} \).

Figure (11) is a graph of \( v \) versus time. Notice that while the steady state is reached very slowly, there are no visible oscillations. If we used a higher value of \( \epsilon \) it would more closely approximate existing technologies but would not work as nicely with the asymptotics.

Figure (12) is a graph of \( v_t \) versus time. Note that it quickly reaches its limit, long before \( v \) stabilizes.

Figure (13) is a graph of \( i \) versus \( t \). We see that \( i \) achieves its steady state without many visible oscillations. Although the asymptotics do not preclude their existence, we could not discern them numerically.

So there exists a sawtooth pattern as before, but with half the amplitude and half the period as was previously observed in Parts 2 and 3. See Figure (14). Upon substituting back into equations (99)-(101) we see that \( x \) and \( y \) depend only on \( \tau \). Neither the output voltage nor the current involve a ripple. Thus by adding the extra capacitor and inductor we have eliminated the ripple in the output voltage and have reduced the amplitude of the ripple of the voltage across the diode by a factor of \( \frac{1}{2} \). These asymptotic results agree with the numerical data.
Figure 11: \( v(t) \) vs. time
Figure 12: $v_1$ vs. time
Figure 13: $i$ vs. time
Figure 14: Oscillations of $v_1$ vs. time
References


ERROR ANALYSIS OF A PENDULOUS INTEGRATING GYRO ACCELEROMETER

IMA Mathematical Modeling Workshop

David Bainbridge, Patrick Bidigare, Chaitan Gupta, Aaron Johnson
Esa Rantanen, Patrick Worfolk and Wei Zhang

Tutor: Blaise Morton

1. INTRODUCTION

Inertial navigation and guidance systems depend on the measurement of linear accelerations and subsequent integrations of these to obtain vehicle velocity and position. Linear acceleration is measured by precision accelerometers, which can be divided into two categories: Pendulous and non-pendulous accelerometers. Of the pendulous accelerometers, the pendulous integrating gyro accelerometer (PIGA) offers superior accuracy compared to floated and flexure-supported pendulums. This superiority is based on the measurement of the mechanical movement of the servo-driven turntable on which the accelerometer is mounted as opposed to the measurement of a weak current generated by the movement of a coil and an iron rotor mounted on the gimbal axis of floated and flexure-supported pendulums.

A PIGA consists of a pendulous gyro mounted on a gimbal, which in turn is mounted on a servo-driven turntable. A diagrammatic representation of the gyro, gimbal, and turntable assembly is depicted in Figure 1. The gimbal assembly, relative to the turntable on which it is mounted, has one degree of freedom in its motion, which is rotation about its horizontal axis. The gimbal assembly is essentially a pendulum since its center of mass does not lie on its axis of rotation. Thus an acceleration

Date: August 6, 1996.
IMA, University of Minnesota, Minneapolis, MN 55455.
of the PIGA apparatus can result in a torque being applied to the gimbal assembly. The PIGA is designed to be operated in a mode where the center of mass of the gimbal assembly lies in the horizontal plane containing the axis of rotation of the gimbal assembly. Ideally, only acceleration of the PIGA in the vertical direction will result in a torque being applied about the gimbal axis.

In order to maintain the gimbal in its horizontal position, a counter-torque must be applied. This is accomplished by a control system which rotates the turntable on which the gimbal assembly is mounted. The gyroscope, mounted transversally on the gimbal, reacts to this rotation, applying a new torque about the gimbal's axis. Thus, in principle, a careful motion of the turntable can result in the gimbal apparatus staying horizontal even when the PIGA is subject to accelerating motion. Net rotations of the turntable are a result of a change in velocity in the apparatus's vertical direction. Thus the angle of rotation of the turntable is used to estimate the velocity of the apparatus. In order to detect three-dimensional motion, three such devices mounted transversally are required.

The sensors for the gimbal rotation and the turntable position are not perfect, however. Mathematical models can be used to predict performance and improve sensor design. To formulate a more tangible goal for our modeling efforts, an objective of determining the design precision and accuracy to produce a PIGA that yields an error of at most $10^{-2}G$ over an operating range of 0 to 10 G, where G is force of gravity on Earth.

This report is organised in the following manner. We first present a nominal model of a PIGA. In this model we assume perfect knowledge and exact sensing of the dynamics. The model is linear and consequently amenable to mathematical analysis. A set of nominal parameter values for which the behavior of the nominal model falls within various operational guidelines were computed. These guidelines include, for example, accuracy of the velocity estimation and maximum rotation rate of the turntable. The design (i.e., the choice of parameters and control) of the nominal system results from these considerations and an analysis of the nominal model. Next, we present an off-nominal model. This model takes into account nonlinearities which
**Figure 1.** Sketch of a PIGA.
arise from not maintaining the gimbal assembly in a horizontal position, uncertainty in the physical parameters of the system, and limits in sensor accuracy. Numerical simulations are presented which indicate that the off-nominal model also has behavior which lies within the operational guidelines. We conclude this report with a discussion of other behaviors in the system that were not modeled and indicate areas of analysis where future work may be fruitful. Some of the mathematical results used in the text are included in the appendices.

2. The nominal model

In this section we present a nominal model for a PIGA. The components of the model are pictured diagrammatically in Figure 2. For this model we choose the simplest elements for each of the components. We shall now discuss each component of the model.

The gimbal dynamics are modeled by the linearized equations of motion of a pendulum. If $\theta$ is the angle of the gimbal apparatus measured away from the horizontal (and positive when the center of mass of the gimbal is below the horizontal plane), then we write

$$I_\psi \theta'' + C \theta' = Pa_x(t) - H \omega(t),$$

(1)

where $I_\psi$, $C$, $P$ and $H$ are physical parameters of apparatus, $a_x(t)$ is the acceleration in the vertical direction and $\omega(t)$ is the angular velocity of the turntable. The physical parameter $I_\psi$ is the moment of inertia of the gimbal apparatus about its axis of rotation, $C$ is the damping coefficient, $P$ is the pendulosity, and $H$ is the angular momentum of the gyroscope. The term $Pa_x(t)$ represents the torque applied to the gimbal by the acceleration of the apparatus in the vertical direction and is the zero-order approximation to $P(a_x(t) \cos(\theta) + a_y(t) \sin(\theta))$ for small $\theta$, where $a_y$ is acceleration of the apparatus in a direction transverse to both the gimbal's axis of rotation and the vertical. This transverse acceleration results in a torque about the gimbal's axis of rotation of $Pa_y(t) \sin(\theta)$. The term $H \omega(t)$ represents the torque due to the interaction of the rotation of the turntable and the gyroscope.
Figure 2. Diagram of the PIGA model.
Moving on to describe the second box in Figure 2, we assume for the nominal model that we have perfect sensing of the angle of the gimbal \( \theta \) and its rate of change \( \theta' \). Thus we assume that the exact values of these quantities are available to the controller of the turntable rotation. The controller is the third box and implements a proportional-integral-differential (PID) control scheme. The sole purpose of the control scheme is to attempt to keep the gimbal apparatus horizontal, i.e., \( \theta \approx 0 \).

The motivation for keeping \( \theta = 0 \) comes from the fact that if we succeed perfectly then the above differential equation becomes

\[
0 = Pa_x(t) - H\omega(t).
\]

Thus the change in velocity is given by

\[
v_x(t) - v_x(0) = \frac{H}{P} \int_0^t \omega(t) dt.
\]

So if \( \phi \) is the angle of the turntable at time \( t \) then \( \phi(t) = \phi(0) + \int_0^t \omega(t) dt \) and the change in velocity is given by

\[
v_x(t) - v_x(0) = \frac{H}{P} (\phi(t) - \phi(0)).
\]

Now when \( \theta \) is not identically zero, Equation (1)

\[
\omega(t) = (Pa_x(t) - I_v\theta'' - C\theta')/H.
\]

So

\[
\phi(t) - \phi(0) = \frac{P}{H} (v_x(t) - v_0(t)) - \frac{I_v}{H} (\theta'(t) - \theta'(0)) - \frac{C}{H} (\theta(t) - \theta(0))
\]

Since we estimate the change in velocity as \( \Delta v_x = \frac{H}{P} \Delta \phi \), where \( \Delta v = v(t) - v(0) \) and similarly, we have the error in estimated velocity at time \( t \) given by

\[
\Delta v_x - \Delta v_x = -\frac{I_v}{P} \Delta \theta' - \frac{C}{P} \Delta \theta.
\]

So here we explicitly see the dependence of the error in the calculated change in velocity as a function of how well we manage to keep the gimbal assembly horizontal.

Our PID controller is given by

\[
\omega_{\text{control}}(t) = K_0 \theta'(t) + K_1 \theta(t) + K_2 \int_0^t \theta(\tau) d\tau.
\]
The term \( K_1 \theta(t) \) is called the proportional control. The idea is that the greater the deflection of the gimbal angle \( \theta \) the faster we need to spin the turntable to provide a counter-torque.

The problem with only using proportional control lies in the solution of the differential equation

\[
I_y \theta'' + C \theta' = Pa_x - H \omega , \\
\omega = K_1 \theta .
\]

Here the steady state solution to a constant acceleration input has \( \theta''_{ss} \equiv 0 \), but \( \theta'_{ss} \neq 0 \), which leads to measurement errors in velocity as shown in Equation 2.

The term \( K_2 \int_0^t \theta(\tau) d\tau \) is the integral control. By adding this to the proportional control we get the following ODE

\[
I_y y''' + C y'' = Pa_x - H \omega , \\
\omega = K_1 y' + K_2 y ,
\]

where \( y' = \theta \). The solution of this ODE has the property that its asymptotic solution for a constant acceleration has \( \theta''_{ss} = \theta'_{ss} = 0 \).

The term \( K_0 \theta' \) is the differential control term. In a realistic system, the damping must be quite small in order to make the gimbal sensitive to small accelerations. This causes the transient effects to die out slowly. The differential control term can be thought of as introducing additional damping which is necessary to force the transient behavior to decay more rapidly.

Returning now to our discussion of the boxes in Figure 2, we next assume in the nominal model that the motor drives the turntable at the exact rotation rate requested by the controller, thus forcing the gyroscope to apply the desired torque. Finally, we assume that we can detect the rotation of the turntable exactly and use the linear factor \( H/P \) to convert the rotation to a change in velocity.
Putting this all together we get the nominal model equations:

\[
\theta'' + \frac{C + HK_0}{I_y} \theta' + \frac{HK_1}{I_y} \theta + \frac{HK_2}{I_y} y^+ = \frac{P}{I_y} a_x(t),
\]

\[
y' = \theta.
\]

If we write this as a third order equation in \(y\) the nominal model becomes

\[
y''' + \frac{C + HK_0}{I_y} y'' + \frac{HK_1}{I_y} y' + \frac{HK_2}{I_y} y = \frac{P}{I_y} a_x(t),
\]

(3)

where \(\theta = y'\). Furthermore, we can write the expression for the error in acceleration as a function of time as

\[
a_x \text{ measured}(t) - a_x(t) = \frac{H}{P} \omega(t) - a_x(t) = \frac{HK_0}{P} \theta' + \frac{HK_1}{P} \theta + \frac{HK_2}{P} y - a_x(t) = \frac{HK_0}{P} y'' + \frac{HK_1}{P} y' + \frac{HK_2}{P} y - a_x(t) = -\frac{I_y}{F} y''' - \frac{C}{F} y''.
\]

This can be integrated to get the accumulated error in velocity.

We shall now discuss the properties of the solutions of the inhomogeneous linear ODE which models the nominal system. As seen from the general solutions presented in Appendix 2, the dynamics of the homogeneous solution are essentially determined by the eigenvalues of the linear differential operator representing the ODE. Assuming that the eigenvalues (which are functions of the parameters) are given by \(\alpha, \beta \pm i \gamma\), then we need \(\alpha, \beta < 0\) in order for the homogeneous solution \(\theta_h = 0\) to be stable. If so, then the homogeneous solution will decay exponentially at a rate of \(e^{-\min(|\alpha|,|\beta|)t}\). The particular solution which depends on \(a(t)\) will give the nontransient behavior. We shall consider the explicit solutions for three different inputs: a constant acceleration, a linear acceleration, and a periodic acceleration.
We consider first a constant acceleration \( a_x(t) = a_0 \), then
\[
y_p(t) = \frac{Pa_0}{HK_2} \\
\theta(t) \to 0 \\
\theta'(t) \to 0
\]
\( a_x \text{ measured} - a_x \to 0 \).

For a linear acceleration \( a_x(t) = a_1 t \),
\[
y_p(t) = \frac{Pa_1 t}{HK_2} - \frac{PK_1 a_1}{K_2^2} \\
\theta(t) \to \frac{Pa_1}{HK_2} \\
\theta'(t) \to 0
\]
\( a_x \text{ measured} - a_x \to 0 \).

For a periodic acceleration \( a_x(t) = e^{i\Omega t} \), then
\[
y_p(t) = PI_y e^{i\Omega t} /(HK_2 - (C + HK_0)\Omega^2 + i(HK_1 \Omega - I_y \Omega^3)) \\
\theta(t) \to i\Omega PI_y e^{i\Omega t} /(HK_2 - (C + HK_0)\Omega^2 + i(HK_1 \Omega - I_y \Omega^3)) \\
\theta'(t) \to -\Omega^2 PI_y e^{i\Omega t} /(HK_2 - (C + HK_0)\Omega^2 + i(HK_1 \Omega - I_y \Omega^3))
\]
\( a_x \text{ measured} - a_x \to (I_y C\Omega^2 + iI_y^2 \Omega^3)e^{i\Omega t} /(HK_2 - (C + HK_0)\Omega^2 + i(HK_1 \Omega - I_y \Omega^3)) \).

3. THE OFF-NOMINAL MODEL

We now introduce the off-nominal model which takes into account the nonlinearity of the problem, the errors in sensing, and uncertainties in parameter values. To present this model we shall once again use the diagram in Figure 2 and simply replace the components we used for the nominal model with more complicated ones as necessary.

We begin by assuming there is some measurement error in our nominal parameter values:
Quantity | Measured value | True value
--- | --- | ---
Gimbal assembly moment of inertia | $I_y$ | $I_y(1 + \epsilon_1)$
Damping coefficient | $C$ | $C(1 + \epsilon_2)$
Pendulosity | $P$ | $P(1 + \epsilon_3)$
Gyro angular momentum | $H$ | $H(1 + \epsilon_4)$

Our differential equations becomes:

$$I_y(1 + \epsilon_1)\theta'' + C(1 + \epsilon_2)\theta' = P(1 + \epsilon_3)(a_x(t)\cos \theta + a_y(t)\sin \theta) - H(1 + \epsilon_4)\omega(t).$$

(4)

The parameters $\epsilon_1, \ldots, \epsilon_4$ are introduced to model the fact that the actual physical parameters of the system are only approximated by the estimated values $I_y, C, P, H$. The torque on the gimbal assembly due to acceleration is modeled by the nonlinear term $P(1 + \epsilon_3)(a_x(t)\cos \theta + a_y(t)\sin \theta)$ which takes into account a transverse acceleration $a_x$ and the fact that the gimbal is not always perfectly horizontal.

Sensors which measure the angle of the gimbal and its rate of change cannot be expected to behave perfectly. We model their errors with a scale error and a constant bias. Thus, the control will get inputs that approximate $\theta$ and $\theta'$ and are

$$\theta_{\text{measured}} = (1 + \epsilon_5)\theta + \epsilon_6$$
$$d\theta_{\text{measured}} = (1 + \epsilon_7)\theta' + \epsilon_8.$$

The controller will be as before, but including the sensing errors, resulting in a desired turntable rotation of

$$\dot{\omega}(t) = K_0 d\theta_{\text{measured}} + K_1 \theta_{\text{measured}} + K_2 y$$

with $y' = \theta_{\text{measured}}$.

We shall still assume that the motor can exactly control the turntable, but we need to introduce an error term into the turntable rotation sensor so we now write

$$\phi_{\text{measured}} = \int \omega(\tau)d\tau + \epsilon_9.$$
Then the measured change in velocity is given by

\[ V_{\text{measured}} = \frac{H}{P} \phi_{\text{measured}}. \]

Putting this all together, we get

\[ + C(1 + \epsilon_2) + H K_0 (1 + \epsilon_4)(1 + \epsilon_6) \frac{y''}{(1 + \epsilon_1)I_y} y'' + \frac{H K_1 (1 + \epsilon_4)(1 + \epsilon_6)}{(1 + \epsilon_1)I_y} y' + \frac{H K_2 (1 + \epsilon_4)(1 + \epsilon_6)}{I_y (1 + \epsilon_1)} y = \frac{P(a_x \cos \theta + a_y \sin \theta)(1 + \epsilon_5)(1 + \epsilon_3)}{I_y (1 + \epsilon_1)} \]

\[ - \epsilon_8 \frac{H K_0}{(1 + \epsilon_1)I_y} \]

where \( \theta \) is now written in terms of \( y \) as

\[ \theta(t) = \frac{y' - \epsilon_6}{1 + \epsilon_5}. \]

We present the results of simulation of this nonlinear model in a later section, but we can study a simplified linear version by once again replacing \( \cos \theta \) and \( \sin \theta \) with 1 and 0, respectively, in the above equations.

We now develop the formula for the error in acceleration as a function of time. In the physical apparatus we are of course getting only velocity readings. One can get the velocity error by integrating the acceleration error. This is accurate to within \( \frac{H}{P} \epsilon_9 \) which represents the error due to being unable to read the rotation of the turntable.
accurately. The equations describe the error in accelerations are then

\[
\begin{align*}
a_x \text{measured}(t) - a(t) &= \frac{H}{P} \omega(t) - a(t) \\
&= \frac{H}{P} \left( K_0 \frac{1 + \epsilon_7}{1 + \epsilon_5} y'' + K_1 y' + K_2 y + k_0 \epsilon_8 \right) - a(t) \\
&= \frac{I_y}{P} \left( 1 + \epsilon_1 \right) y'' \\
&\quad + \left( \frac{HK_0}{P} \frac{1 + \epsilon_7}{1 + \epsilon_5} \left( 1 - \frac{1 + \epsilon_4}{1 + \epsilon_3} \right) - \frac{C}{P} \frac{1 + \epsilon_7}{(1 + \epsilon_3)(1 + \epsilon_5)} \right) y'' \\
&\quad + \frac{HK_1}{P} \left( 1 - \frac{1 + \epsilon_4}{1 + \epsilon_3} \right) y' \\
&\quad + \frac{HK_2}{P} \left( 1 - \frac{1 + \epsilon_4}{1 + \epsilon_3} \right) y \\
&\quad + \frac{HK_0}{P} \epsilon_8 \left( 1 - \frac{1 + \epsilon_4}{(1 + \epsilon_3)(1 + \epsilon_5)} \right).
\end{align*}
\]

As in the nominal model, we can now use the solutions presented in Appendix 2 to find the asymptotic (or full) solutions for various inputs \(a_x(t)\) for the simplified linear off-nominal model.

We consider only a constant plus linear acceleration \(a_x(t) = a_0 + a_1 t\). In this case

\[
\begin{align*}
y_p(t) &= \frac{Pa_1}{HK_2} \frac{1 + \epsilon_3}{1 + \epsilon_4} \left( t - \frac{K_1}{K_2} \right) \\
&\quad + \frac{Pa_0}{HK_2} \frac{1 + \epsilon_3}{1 + \epsilon_4} - \frac{\epsilon_8}{1 + \epsilon_5} \frac{K_0}{K_2} \\
\theta(t) &\rightarrow \frac{Pa_1}{HK_2} \frac{1 + \epsilon_3}{(1 + \epsilon_4)(1 + \epsilon_5)} - \frac{\epsilon_6}{1 + \epsilon_5} \\
\theta'(t) &\rightarrow 0 \\
a_x \text{measured} - a_x &\rightarrow \frac{HK_0}{P} \frac{(\epsilon_3 - \epsilon_4) \epsilon_8}{(1 + \epsilon_3)(1 + \epsilon_5)} + \left( a_0 + a_1 t \right) \frac{\epsilon_3 - \epsilon_4}{1 + \epsilon_4}.
\end{align*}
\]
4. SELECTION OF PARAMETERS

Several system constraints were used in determining the parameter values. These constraints are summarized in Table 1 below.

TABLE 1: System constraints

<table>
<thead>
<tr>
<th>Constraint</th>
<th>Symbol</th>
<th>Magnitude/units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max. measurable acceleration</td>
<td>$a_{\text{max}}$</td>
<td>$10 \text{ G } = 9800 \text{ cm/sec}^2$</td>
</tr>
<tr>
<td>Max. turntable speed at 10 G</td>
<td>$\omega_{\text{max}}$</td>
<td>$2 \text{ Hz } = 12.57 \text{ rad/sec}$</td>
</tr>
<tr>
<td>Gyro speed</td>
<td>$\Omega$</td>
<td>$100 \text{ Hz } = 628.3 \text{ rad/sec}$</td>
</tr>
<tr>
<td>Turntable resolution</td>
<td></td>
<td>16384 positions $= 3.834 \cdot 10^{-4} \text{ rad/pos}$</td>
</tr>
<tr>
<td>Drive power limit</td>
<td></td>
<td>100 W</td>
</tr>
<tr>
<td>Gimbal angle stability</td>
<td></td>
<td>$\pm 10^{-4} \text{ rad}$</td>
</tr>
</tbody>
</table>

The gimbal angle stability constraint was imposed to ensure that the cross-coupling term of $\sin \theta a_y = a_y \theta$ was small enough that a 0.1 G lateral acceleration $a_y$ would not contribute more than $10^{-5}$ G error in acceleration, which is the desired system accuracy.

In the nominal case $T_i$ is the torque applied to the gimbal due to the acceleration of the instrument in the vertical direction. Expanding Figure 2 this is represented in the diagram in Figure 3. $T_i$ is expressed at $\theta = 0^\circ$ gimbal angle as

$$T_i = Pa$$

where $P$ is pendulosity and $a$ acceleration. As shown in Figure 1 the gyroscope is centered within the gimbal and thus contributes nothing to the pendulosity. If we view $M$ as a point mass, the pendulosity is given by

$$P = 4Mg \cdot cm$$
Figure 3: A block diagram of the PIGA gimbal inputs
The torque acting on the gimbal due to the rotation of the turntable is $T_3$. From Figure 3 we get

$$T_3 = H \omega$$

where $\omega$ is the rate of rotation of the turntable and $H$ is the angular momentum of the gyroscope about its axis of rotation.

We must first consider a bound on the ratio $\frac{\omega}{P}$. Acceleration is measured as $a = \frac{H}{P} \omega$. Since $\omega$ must be $\leq 12.57$ rad/sec but accelerations up to $9800 \text{ cm/sec}^2$ must be measured, we must have

$$\frac{P}{H} 9800 \text{ cm/sec}^2 = \omega_{\text{max}} \leq 12.57$$

Thus

$$\frac{H}{P} \geq 780 \text{ cm/sec}$$

It was assumed that the gyro consists of a stainless steel cylinder with a radius of 2 cm and a height of 4 cm. The density of stainless steel is 8.02 g/cm$^3$ (CRC Handbook of Chemistry and Physics). Angular momentum is the moment of inertia in the direction of gyro rotation times the angular velocity of the gyro. The moment of inertia of a body about an axis is the integral of the density times the square distance to the axis over the body. The moment of inertia is thus

$$8.02 \int_{\theta=0}^{2\pi} \int_{h=0}^{4} \int_{r=0}^{2} r^2 \text{d}r \text{d}h \text{d}\theta$$

$$= 2\pi \cdot 4 \cdot \frac{2^4}{4} \cdot 8.02$$

$$= 806.2 \text{ cm}^2 \cdot g$$

Thus, with a gyro speed of 628.3 rad/sec this gives us $H = 506600 \text{ g cm}^2 \cdot \text{rad/sec}$

Now we want
\[ 780 = \frac{H}{P} = \frac{506600}{4M} \]

Thus, the pendulous mass \( M \) must be 162.4 g to generate the correct pendulosity of \( P = 649.5 \text{ g} \cdot \text{cm} \).

The inertia of the gimbal \( I_y \) is the sum of the inertia of the gyro about the gimbal axis \( I_{gyro} \) and the inertia due to the pendulous mass \( M \).

\[
I_{gyro} = 8.02 \int_{x=-2}^{2} \int_{y=-2}^{2} \int_{z=-\sqrt{4-y^2}}^{\sqrt{4-y^2}} (x^2 + z^2) \, dz \, dy \, dx = 940.6 \text{ g} \cdot \text{cm}^2
\]

The inertia due to \( M \) is \( I_M = 4^2 M = 2598 \text{ g} \cdot \text{cm}^2 \). Thus \( I_y = I_{gyro} + I_M = 3539 \text{ g} \cdot \text{cm}^2 \).

The damping coefficient \( C \) is defined as proportionally constant between the angular velocity of the gimbal and and the torque acting on the gimbal axis due to damping friction. Thus

\[ T_c = C \dot{\theta} \]

and has units \( \text{g} \cdot \text{cm}^2 \cdot \text{rad/sec} \).

To calculate the value of the damping coefficient \( C \), it is assumed that the gimbal will lose about half of its angular velocity per minute when allowed to spin freely, i.e., with no torques than those generated by damping and inertia. The differential equation for this motion is

\[ I_y \ddot{\theta} + C \dot{\theta} = 0 \]

which has a solution

\[ \dot{\theta} = Me^{\frac{-C}{I_y}} \]

We want

\[ \frac{\dot{\theta}(60)}{\dot{\theta}(0)} = \frac{1}{2} \]
thus

\[-\frac{c_{(60)}}{I_f} = \frac{1}{2}\]

so we must choose

\[C = \frac{I_y \ln(2)}{60} \text{ g cm}^2 / \text{sec}\]

Using the value of \(I_y\) obtained above we get \(C = 40.88 \text{ g cm}^2 / \text{sec}\). Assuming that the turntable inertia is twice the gimbal inertia we get \(I_{\text{turntable}} = 7078 \text{ g cm}^2\).

The controller rotating the turntable to counter the torque on the gimbal assembly can be depicted, expanding Figure 1, as a block diagram in Figure 4 and expressed as

\[I_y \ddot{\theta} + C \dot{\theta} = Pa(t) - H(K_o \dot{\theta} + K_1 \theta + K_2 \int \theta)\]

where \(Pa(t) = T_1\), i.e., torque on the gimbal due to acceleration of the instrument and \(H(K_o \dot{\theta} + K_1 \theta + K_2 \int \theta) = T_3\), i.e., torque on the gimbal. Given the parameters \(C, I_y, H,\) and \(P\), eigenvalues of the linear system are chosen: \(a, b \pm i\gamma\).

The characteristic polynomial of our nominal ODE is obtained from Equation 3. Since this must also equal

\[(y - \alpha)(y - \beta + i\gamma)(y - \beta - i\gamma)\]

By equating coefficients we get the following values for \(K_1, K_2, K_3\)

\[K_0 = -\frac{([\alpha + 2\beta]I_y + C)}{H}\]

\[K_1 = \frac{I_y}{H} (\beta^2 + \gamma^2 + 2\alpha\beta)\]

\[K_2 = -\alpha \frac{I_y}{H} (\beta^2 + \gamma^2)\]
FIGURE 4: A block diagram of the PIGA turntable controller
We can now do a simulation and solve explicitly for the solutions of the linear ODE. However, we can explicitly solve only for particular choices of \( a(t) \), e.g., constant, periodic, etc.

To ensure \(|\theta| \leq 10^4 \) radians, the parameters \( K_0, K_1, \) and \( K_2 \) were calculated using the formula

\[
\|\theta\|_\infty \leq \frac{P}{I_y \alpha^2} \cdot \|\dot{a}\|_\infty
\]

where \( \theta(t) \) solves

\[
I_y \ddot{\theta} + (HK_0 - C) \dot{\theta} + (HK_1) \dot{\theta} + (HK_2) \theta = P \dot{a}(t)
\]

with \( \theta(0) = \dot{\theta}(0) = \ddot{\theta}(0) = 0 \) and \( \alpha, \beta + i \gamma, \beta - i \gamma \) are the eigenvalues of the above equation. If we assume that \( \alpha = \beta, \gamma = 0 \), then

\[
\|\theta\|_\infty \leq \frac{P}{I_y \alpha^3} \cdot \|\dot{a}\|_\infty = \frac{P}{HK_2} \|\dot{a}\|_\infty
\]

because

\[
-\alpha^3 = \frac{HK_2}{I_y}
\]

Using our values for \( P, H, \) and \( \|\dot{a}\|_\infty \) we get \( \|\theta\|_\infty \leq 10^4 \) radians if \( K_2 \geq 838 \text{ 1/sec}^2 \).

Using the formulas

\[
3\alpha = \left( \frac{HK_0 - C}{I_y} \right)
\]

\[
3\alpha^2 = \frac{HK_1}{I_y}
\]

\[
\alpha^3 = -\frac{HK_2}{I_y}
\]
Then we get $a = -49.3$ rad/sec. If we use $a = -63$ rad/sec instead, then $K_p = 1.32$, $K_i = 83$ rad/sec, and $K_2 = 1747$ rad/sec$^2$. The reason for choosing $a = -63$ rad/sec rather than $a = -50$ rad/sec is due to possible errors in $H$, $C$, $I_y$ and measurement errors in $\dot{\theta}$ or $\ddot{\theta}$. Any small perturbations in these values will move the eigenvalues slightly away from -63 in the complex plane. The choice of $a = -63$ ensures that with any reasonable errors in the above values we will still have $a, b < -50$ and thus $\|\theta\|_\infty \leq 10^{-4}$ radians. Our system parameters and their values are summarized in Table 2 below:

**Table 2: System parameters and their respective values**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pendulosity</td>
<td>$P$</td>
<td>649.4 g·cm</td>
</tr>
<tr>
<td>Gyro angular momentum</td>
<td>$H$</td>
<td>506600 g·cm$^2$/sec</td>
</tr>
<tr>
<td>Gimbal inertia</td>
<td>$T_y$</td>
<td>3539 g·cm$^2$</td>
</tr>
<tr>
<td>Damping coefficient</td>
<td>$C$</td>
<td>40.88 g·cm$^2$/sec</td>
</tr>
<tr>
<td>Turntable inertia</td>
<td>$I_n$</td>
<td>7078 g·cm$^2$</td>
</tr>
<tr>
<td>Differential feedback</td>
<td>$K_0$</td>
<td>1.32</td>
</tr>
<tr>
<td>Proportional feedback</td>
<td>$K_i$</td>
<td>83 rad/sec</td>
</tr>
<tr>
<td>Integral feedback</td>
<td>$K_2$</td>
<td>1747 rad/sec</td>
</tr>
</tbody>
</table>

Lastly, we wish to use the above values to obtain a set of bounds for the error terms in the non-nominal model. Our benchmark test for determining the tolerances on the $\varepsilon$'s was a ramped input acceleration going from 1 G to 5 Gs over 60 seconds. This benchmark simulates the acceleration that would be experienced in a launch vehicle as it exhausts its fuel supply and becomes lighter. We must ensure that the $\varepsilon$'s are small enough to meet the system design criteria $|\theta| < 10^{-4}$.
radians and \( la_{error} < 10^{-5} \) Gs = \( 9.8 \cdot 10^{-5} \) Gs cm/sec\(^2\). Using the values from Table 1 in the steady state solutions for \( \theta \) and \( a_{error} \) given in Equation 5 we obtain

\[
\theta = 4.795 \cdot 10^{-5} \frac{1 + \epsilon_4}{(1 + \epsilon_4)(1 + \epsilon_5)} - \frac{\epsilon_6}{(1 + \epsilon_5)}
\]

\[
a_{error} = \frac{1030(\epsilon_3 + \epsilon_4)\epsilon_5\epsilon_8}{(1 + \epsilon_3)(1 + \epsilon_5)} + \frac{(\epsilon_3 + \epsilon_4)(a_0 + a_1t)}{(1 + \epsilon_4)}
\]

At \( t = 60 \) seconds our expression for \( a_{error} \) is

\[
a_{error} = \frac{(\epsilon_3 + \epsilon_4)4900}{(1 + \epsilon_4)} + \frac{1030(\epsilon_3 + \epsilon_4)\epsilon_5\epsilon_8}{(1 + \epsilon_3)(1 + \epsilon_5)}
\]

We initially make the simplifying assumption that \( \epsilon_3, \epsilon_4, \) and \( \epsilon_5 \) are small enough that the terms \( 1 + \epsilon_3, 1 + \epsilon_4, \) and \( 1 + \epsilon_5 \) can essentially be replace by 1, so \( \theta = 4.795 \cdot 10^{-5} - \epsilon_6 \) and \( a_{error} = 1030(\epsilon_3 - \epsilon_4) \epsilon_5 \epsilon_8 + 4900(\epsilon_3 - \epsilon_4) \). To keep \( |\theta| < 10^{-4} \) we must ensure that \( |\epsilon_6| \leq 5 \cdot 10^{-5} \) radians. Now, if \( \epsilon_3, \epsilon_4, \) and \( \epsilon_8 \) are small, then \( a_{error} \equiv 4900(\epsilon_3 - \epsilon_4) \) so we need \( (\epsilon_3 - \epsilon_4) \leq 9.8 \cdot 10^{-3} / 4900 \) or \((\epsilon_3 - \epsilon_4) \leq 2 \cdot 10^{-6} \). Thus, we need, say \( |\epsilon_3|, |\epsilon_4|, |\epsilon_5| \leq 10^{-6} \). If we ensure that \( |\epsilon_3|, |\epsilon_4|, |\epsilon_5| < 10^{-6} \) then our assumptions above will be valid.

Note that the error terms \( \epsilon_3, \epsilon_4, \) and \( \epsilon_5 \) do not appear explicitly in in the equations for \( \theta \) or \( a_{error} \). The non-nominal equations for the control constants for given eigenvalues \( \alpha = \beta + \gamma t = \beta - \gamma t = \lambda \), \( K_0, K_1, \) and \( K_2 \), are given by
\[
K_0 = \frac{-3\lambda J_y (1 + \varepsilon_1) - C (1 + \varepsilon_2)}{H(1 + \varepsilon_4)} = 1.32 \frac{(1 + \varepsilon_1)}{(1 + \varepsilon_4)} - 8.07 \cdot 10^{-3} \frac{(1 + \varepsilon_2)}{(1 + \varepsilon_4)}
\]

\[
K_1 = \frac{I_y (1 + \varepsilon_1)}{H(1 + \varepsilon_4)} (3\lambda^2) = 83.16 \frac{(1 + \varepsilon_2)}{(1 + \varepsilon_4)}
\]

\[
K_2 = -\frac{I_y (1 + \varepsilon_1)}{H(1 + \varepsilon_4)} (\lambda^3) = 1714 \frac{(1 + \varepsilon_2)}{(1 + \varepsilon_4)}
\]

Since these depend on \(\varepsilon_1\) and \(\varepsilon_2\), these errors will affect the true locations of the eigenvalues of our system. Note, however, that only large error values for the \(\varepsilon\)'s will affect the values of \(K_0\), \(K_1\), or \(K_2\) significantly. If we ensure that \(|\varepsilon_1| < 10^{-3}\), \(|\varepsilon_2| < 10^{-4}\) then our true eigenvalues will be very close to the chosen values of \(\lambda = 63\).

The error bound on \(\varepsilon_2\) may seem large, but it should be noted that it is difficult to obtain a true value for a damping coefficient and hence a large error tolerance for \(C\) is desirable. The error bound for \(\varepsilon_1\) is not critical and we may impose a constraint \(|\varepsilon_1| < 10^{-2}\). The bounds on our non-nominal error terms which allow us to achieve the design goal of an accelerometer accurate to \(10^{-5}\) Gs are summarized in Table 3 below.
TABLE 3: Non-nominal error term bounds

<table>
<thead>
<tr>
<th>Error term</th>
<th>Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_y$ measurement error</td>
<td>$</td>
</tr>
<tr>
<td>$C$ measurement error</td>
<td>$</td>
</tr>
<tr>
<td>$P$ measurement error</td>
<td>$</td>
</tr>
<tr>
<td>$H$ measurement error</td>
<td>$</td>
</tr>
<tr>
<td>$\theta$ measurement scale error</td>
<td>$</td>
</tr>
<tr>
<td>$\theta$ measurement bias error</td>
<td>$</td>
</tr>
<tr>
<td>$\dot{\theta}$ measurement scale error</td>
<td>$</td>
</tr>
<tr>
<td>$\dot{\theta}$ measurement bias error</td>
<td>$</td>
</tr>
</tbody>
</table>
5. SIMULATION

We used MATLAB's ODE45 for our simulations. Our initial conditions were

\[
\begin{bmatrix}
\theta \\
\dot{\theta} \\
y \\
\Delta v
\end{bmatrix} = \begin{bmatrix}
0 \\
0 \\
0 \\
0
\end{bmatrix}
\]

where \( \dot{y} = \theta \). The system was solved for a time interval of 60 seconds. Beginning with the nominal model we tested our program with three types of input acceleration: constant, linear, and periodic. For each type of acceleration we looked at the results in four graphs. The top two graphs show the input as acceleration and velocity. The lower left graph is the estimate error in velocity. The lower right graph shows the phase space of \( \theta \). Our simulations included proportional, integral and differential control. For the off-nominal model we looked at maximum errors we could tolerate in the parameters and still meet specification requirements in \( \Delta v \) and \( \theta \).

5.1. **Nominal Model.** For the simulation with constant acceleration estimate error in velocity is essentially zero. This agrees with the previous analytical calculations with the exception of an initial transient phase lasting less than 1.2 seconds. The length of the transient is controlled by choice of \( k_0, k_1, k_2 \) and is not considered a source of error.

The linear or “ramp” acceleration consists of two parts. First within the interval \( t = [0, 1] \) seconds the acceleration is constant at \( 1g \). The analogy is a launch vehicle sitting at a platform on earth ready for takeoff. After 1 second the system encounters an linear acceleration lasting the next 59 seconds with max acceleration of \( 4g \) at \( t = 60 \). The equation for the acceleration force for \( t = [1, 60] \) is

\[
a(t) = 980 + (t - 1) \times (4 \times 980) / 60.
\]

For this type of input the estimate in velocity error incurs a constant bias error. This error occurs at the beginning of the ramp, \( t = 1 \), and remains constant thereafter.

For a periodic acceleration we used \( \sin \pi t \). Here the error estimate in velocity is not constant. The resulting error is best seen in the phase space as the error in \( \theta \) and \( d\theta \) centered about the origin. The error for periodic acceleration of fixed amplitude
Figure 3. Constant input acceleration
**Figure 4.** Ramp input acceleration
is bounded. In the first two accelerations only the transient incurs large change in $\theta$ whereas for periodic acceleration this change in $\theta$ does not go to zero.

5.2. Off-nominal Model. After verifying the nominal simulation with the analytical results we added error terms to the parameters to create a off-nominal model. We noticed that $\pm \epsilon$ had to be considered in order to avoid cancellation of error. We looked for the $\epsilon$ giving max error allowed to stay within specifications.

\[
\begin{align*}
I_y & \quad \epsilon_1 \\
C & \quad \epsilon_2 \\
H & \quad \epsilon_3 \\
P & \quad \epsilon_4 \\
\theta_{\text{scale}} & \quad \epsilon_5 \\
\theta_{\text{bias}} & \quad \epsilon_6 \\
d\theta_{\text{scale}} & \quad \epsilon_7 \\
d\theta_{\text{bias}} & \quad \epsilon_8
\end{align*}
\]

These parameters are described in a previous section. For these tests we used a ramp acceleration. The simulation reveals that the most sensitive parameters are $H, P$ and $\theta_{\text{bias}}$. We looked at $|\epsilon_i| < .001$. Our criteria for $H, P$ was $\Delta v < .294$. In order to achieve this accuracy in $\Delta v$ we require $\epsilon_3 < 8.5e - 7$ and $\epsilon_4 < 8.5e - 7$. In order to avoid cancellation in the magnitudes of $\epsilon_3$ and $\epsilon_4$ opposite signs were used. The maximum $\Delta v$ incurred by these parameter variations was 0.2933.

For the error in $\theta_{\text{scale}}$ we used the criteria $\theta_{\text{drift}} < .0001$. This error is the distance from the origin in phase space. To obtain this tolerance $\epsilon_6 < 5e - 5$.

Changes in the other parameter magnitudes $|\epsilon_i| < .001$ showed no increase in the error criteria we observed.
FIGURE 5. Periodic input acceleration
6. CONCLUSIONS

We began modeling our accelerometer with a description of a PIGA system as well as performance criteria we wished to meet. We put realistic operating constraints on the system for its operation over a range of accelerations bounded by 10G with maximum jerk 4G per minute. These constraints included bounds on how far the gimbal rotated from the horizontal, accuracy and resolution of the gimbal rotation, size and spin rate of the gyroscope, resolution of the turntable rotation pickoff, accelerometer sensitivity, maximum turntable rotation rate and power constraints.

We studied the PIGA design until we had a firm understanding of the operational principles of the system. We then developed a nominal mathematical model for the system. We extended this model to include off-nominal effects to account for manufacturing tolerances and gimbal sensor errors. We then selected a reasonable set of nominal parameters for our system. This selection was done to insure consistency with the performance criteria. We analyzed the effects of introducing the off-nominal parameters, both by simulation and analytically. At this point we determined our control gains and were able to establish a set of manufacturing tolerances and gimbal sensing tolerances which allowed us to meet our performance goals.

We conclude this it is possible to design a PIGA which will meet our performance criteria, provided we can achieve certain manufacturing and sensing tolerances.

The PIGA accelerometer is more accurate than the simpler designs that use an electromagnetic device to provide the control torque on the gimbal. This is because we can achieve much finer control of the gimbal torque indirectly by controlling the speed of the turntable than can be achieved using a circuit.

In this analysis we did not consider a number of off-nominal effects which would tend to degrade performance. Effects such as inaccuracies due to thermal expansion, turntable dynamics (delays, lags, etc...) as well as coning effects from sinusoidal transverse accelerations were not considered. We did, however, consider all the dominant effects which are considered by engineers to be the major sources of error in this system.
7. Bibliography

APPENDIX A: bounds for $\theta$

Differentiating $I_y \ddot{\theta} + C\dot{\theta} = Pa(t) - (K_0 \dot{\theta} + K_1 \theta + K_2 \int \theta)$ we get

$$I_y \ddot{\theta} + (C + K_0) \dot{\theta} + K_1 \dot{\theta} + K_2 \theta = Pa(t)$$

(1)

Let $\theta(t)$ be the solution to (1) with $\theta(0) = \dot{\theta}(0) = \ddot{\theta}(0) = 0$. Then if $|\dot{\theta}(t)| \leq M$ for all $t > 0$, we have

$$|\theta(t)| \leq \frac{P}{I_y} \frac{M}{|Re \lambda_1||Re \lambda_2||Re \lambda_3|}$$

(2)

for all $t > 0$, where $\lambda_1$, $\lambda_2$, $\lambda_3$ are eigenvalues of the equation (1) and assuming $Re \lambda_1$, $Re \lambda_2$, $Re \lambda_3 < 0$. Dividing (1) by $I_y$ yields

$$\ddot{\theta} + \left(\frac{C + K_0}{I_y}\right) \dot{\theta} + \frac{K_1}{I_y} \dot{\theta} + \frac{K_2}{I_y} \theta = \frac{P}{I_y} \dot{\theta}(t)$$

(3)

which is the same as

$$\left(\frac{d}{dt} - \lambda_1\right) \left(\frac{d}{dt} - \lambda_2\right) \left(\frac{d}{dt} - \lambda_3\right) \theta(t) = \frac{P}{I_y} \dot{\theta}(t)$$

(4)

we solve

$$\left(\frac{d}{dt} - \lambda_1\right) \Phi_1(t) = \frac{P}{I_y} \dot{\theta}(t)$$

(5)

$$\left(\frac{d}{dt} - \lambda_2\right) \Phi_2(t) = \Phi_1(t)$$

(6)

$$\left(\frac{d}{dt} - \lambda_3\right) \theta(t) = \Phi_2(t)$$

(7)
The solution to (5) is

\[
\Phi_1(t) = e^{\lambda_t t} \int_0^t e^{-\lambda_s s} \frac{P}{I_y} \hat{a}(s) ds
\]  

(6)

\[
|\Phi_1(t)| = e^{(Re \lambda_1) t} \int_0^t e^{-(Re \lambda_1) s} \frac{P}{I_y} |\hat{a}(s)| ds \leq e^{(Re \lambda_1) t} \int_0^t e^{-(Re \lambda_1) s} \frac{P}{I_y} M ds
\]

(7)

\[
\int_0^t e^{-(Re \lambda_1) s} \frac{P}{I_y} M ds = e^{(Re \lambda_1) t} \left( \frac{P}{I_y} M \right) \frac{1}{-Re \lambda_1} e^{-(Re \lambda_1) t} |0|
\]

\[
= \left( \frac{P}{I_y} M \right) \frac{1}{-Re \lambda_1} (1 - e^{-(Re \lambda_1) t})
\]

Since \( Re \lambda_1 < 0 \) and \( t > 0 \), \( 1 - e^{-(Re \lambda_1) t} < 1 \). So,

\[
|\Phi_1(t)| \leq \left( \frac{P}{I_y} M \right) \frac{1}{Re \lambda_1} = \frac{P}{I_y} \frac{1}{|Re \lambda_1|} M
\]

(8)

We solve (6) and (7) in a similar manner:

\[
\Phi_2(t) = e^{\lambda_t t} \int_0^t e^{-\lambda_s s} \Phi_1(s) ds
\]

\[
\theta(t) = e^{\lambda_t t} \int_0^t e^{-\lambda_s s} \Phi_2(s) ds
\]

Using the same kind of estimate we get
\[ |\Phi_2(t)| \leq \frac{1}{\text{Re} \lambda_2} \left( \frac{P}{I_y} \right) \frac{M}{|\text{Re} \lambda_1|} = \frac{P}{I_y} \frac{M}{|\text{Re} \lambda_1||\text{Re} \lambda_2|} \]  

(9)

\[ |\theta(t)| \leq \frac{1}{\text{Re} \lambda_2} \left( \frac{P}{I_y} \right) \frac{M}{|\text{Re} \lambda_1||\text{Re} \lambda_2|} = \frac{P}{I_y} \frac{M}{|\text{Re} \lambda_1||\text{Re} \lambda_2||\text{Re} \lambda_3|} \]  

(10)

Using these solution formulas ensures \( \theta(0) = \dot{\theta}(0) = \ddot{\theta}(0) = 0. \)
APPENDIX B. THE SOLUTION OF THE THIRD ORDER ODE

Consider the third order linear nonhomogeneous ODE,

\[ y''' + ay'' + by' + cy = f(t). \]

It’s characteristic polynomial is \( \lambda^3 + a\lambda^2 + b\lambda + c = 0 \). If we assume the roots are given by \( \alpha, \beta \pm i\gamma \), then the homogeneous solution of the differential equation is given by

\[ y_h = Ae^{\alpha t} + Be^{\beta t} \cos \gamma t + Ce^{\beta t} \sin \gamma t. \]

We compute

\[
\begin{align*}
y'_h &= \alpha Ae^{\alpha t} + (B\beta + C\gamma)e^{\beta t} \cos \gamma t + (C\beta - B\gamma)e^{\beta t} \sin \gamma t \\
y''_h &= \alpha^2 Ae^{\alpha t}(B\beta^2 + 2C\gamma\beta - B\gamma^2)e^{\beta t} \cos \gamma t + (C\beta^2 - 2B\gamma\beta - C\gamma^2)e^{\beta t} \sin \gamma t.
\end{align*}
\]

The initial conditions are given by,

\[
\begin{pmatrix}
y_h(0) \\
y'_h(0) \\
y''_h(0)
\end{pmatrix} =
\begin{pmatrix}
1 & 1 & 0 \\
\alpha & \beta & \gamma \\
\alpha^2 & \beta^2 - \gamma^2 & 2\beta
\end{pmatrix}
\begin{pmatrix}
A \\
B \\
C
\end{pmatrix}.
\]

Let \( d = \gamma(\alpha^2 - 2\alpha\beta + \beta^2 + \gamma^2) \), then we write the arbitrary parameters \( A, B, C \) as functions of the initial conditions by

\[
\begin{pmatrix}
A \\
B \\
C
\end{pmatrix} = \frac{1}{d}
\begin{pmatrix}
\gamma(\beta^2 + \gamma^2) & -2\gamma\beta & \gamma \\
\alpha\gamma(\alpha - 2\beta) & 2\gamma\beta & -\gamma \\
\alpha(\beta^2 - \gamma^2 - \beta\alpha) & \alpha^2 - \beta^2 + \gamma^2 & \beta - \alpha
\end{pmatrix}
\begin{pmatrix}
y_h(0) \\
y'_h(0) \\
y''_h(0)
\end{pmatrix}.
\]

We now present particular solutions for various forcing functions:

\[
\begin{align*}
f(t) &= f_0 \quad y_p = \frac{f_0}{c}, \\
f(t) &= f_1 t \quad y_p = \frac{f_1}{c}(t - \frac{\beta}{c}), \\
f(t) &= e^{i\omega t} \quad y_p = \frac{1}{-i\omega^3 - a\omega^2 + bi\omega + c}e^{i\omega t}.
\end{align*}
\]
APPENDIX C. MATLAB CODE

The code we used for the off-nominal model includes $\epsilon$ parameter variations and nonlinearity terms. The model agrees with the nominal simulation when nonlinearity is zeroed out and all $\epsilon_i = 0$. We include the wrapper script run.m which defines all the global variables used by the integrator.

C.1. run.m.

```
% # code to do analysis of PIGA simulation
%
global Iy C P H K0 K1 K2 eps accel_function alpha beta gamma nlin_on;

% set up the nominal parameters
%
Iy = 3539;
Itturn = 2*Iy;
C  = 40.88;
P  = 649.5;
H  = 506600;
K0 = -((alpha+2*beta)*Iy+C)/H;
K1 = Iy*(beta*beta+gamma*gamma+2*alpha*beta)/H;
K2 = -alpha*Iy*(beta*beta+gamma*gamma)/H;

% set the following to 0 to forget nonlinearity
nlin_on = 0.0;

% set up the error terms
%
%eps_mag = 1e-2;
eps_mag = 1.0;
```
eps = eps_mag * zeros(1,9);
eps(1) = 1;

% eigenvalues of linear nominal system
scale = 1.0;
alpha = -63;
beta = alpha;
gamma = 0.0;

% choose an acceleration function

%accel_function = 'accel_constant';
accel_function = 'accel_ramp';
%accel_function = 'accel_periodic';

% choose a model
model_function = 'nlin';

% choose initial conditions and integrate
Y0 = zeros(4,1);
t0 = 0;
tf = 60; % run for tf seconds
[t,Y] = ode45(model_function, t0, tf, Y0);

% now plot some output
%

n=3;
m=2;

% plot theta vs theta_dot
```matlab
cmplot(n,m,1);plot(Y(:,1),Y(:,2),'r'); title('Phase space');
xlabel('theta'); ylabel('d(theta)/dt');

% plot time series for error in velocity
subplot(n,m,2);plot(t,Y(:,4),'g.'); title('Error in estimated velocity');
xlabel('time'); ylabel('velocity error');

% plot time series for turntable rotation rate
dtheta_meas = (1+eps(7))*Y(:,2) + eps(8)*ones(size(Y(:,1)));
theta_meas = (1+eps(5))*Y(:,1) + eps(6)*ones(size(Y(:,1)));
omega = -K0*dtheta_meas - K1*theta_meas - K2*Y(:,3);
subplot(n,m,3);plot(t,omega); title('Turntable');
ylabel('rotation rate'); xlabel('time');

yyy = (P.*feval(accel_function,t)+H.*omega - C.*Y(:,2))/Iy;

omega_dot = -K0*(1+eps(7))*yyy - K1*(1+eps(5))*Y(:,2) - K2*Y(:,1);
subplot(n,m,4);
plot(t,omega_dot);

subplot(n,m,5);
plot(t,Iturn.*omega.*omega_dot);

K=[K0,K1,K2],
Y(length(Y),1),

C.2. nlin.m.

function [Yprime] = nlin(t,Y)
%
% nlin is the right hand side of a system
% of nonlinear ODEs modelling a PIGA.
%

% the parameters for this model are global variables
global Iy C P H K0 K1 K2 eps accel_function alpha beta gamma nlin_on;

% construct other parameters
Iy_true = Iy * (1+eps(1));
C_true = C * (1+eps(2));
P_true = P * (1+eps(3));
H_true = H * (1+eps(4));

% the input Y is [theta u y dv]
theta = Y(1);
u = Y(2); % u = theta_dot
y = Y(3);
dv = Y(4);

% calculate some intermediate quantities
a_y = 0.0; % skip these effects for now
a_x = feval(accel_function, t);

theta_meas = (1+eps(5)) * theta + eps(6);
dtheta_meas = (1+eps(7)) * u + eps(8);
omega = - K0 * dtheta_meas - K1 * theta_meas - K2 * y;

torque_acceleration = P_true / Iy_true * ...
(a_x * cos(nlin_on*theta) + a_y * sin(nlin_on*theta));
torque_turntable = H_true / Iy_true * omega;
% now work out the right hand side
%

% The dynamics of the gimbal:
theta_dot = u;
u_dot = - C_true / Iy_true * theta_dot ... 
+ torque_turntable + torque_acceleration;

% the dynamics of the control
y_dot = theta_meas;

% the dynamics of the error in velocity
dv_dot = - H / P * omega - a_x;

Yprime = [theta_dot; u_dot; y_dot; dv_dot]';

C.3. accel-constant.m.
function [a] = accel_constant(t)
%
% constant acceleration
% g = 9.8 m/s
%
a=9.8*100*ones(size(t)); % cm/sec

C.4. accel-ramp.m.
function [a] = accel_ramp(t)
%
% acceleration ramp of 4g/minute
% g = 9.8 m/s
% 4g = 39.2 m/s
for i=1:length(t),
    if (t(i)<1)
        a = 980;
    else
        a = 980 + (t(i)-1)*(4*980)/60;
    end
end

C.5. accel-periodic.m.

function [a] = accel_periodic(t)
    
    %
    % periodic acceleration
    a = sin(pi*t);
# A Mathematical Model for Ore Stack Drying

**Group 1**

**Tutor:** Colin Please  
Southampton University, England

<table>
<thead>
<tr>
<th>Group Members</th>
<th>Institutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deborah Alterman</td>
<td>U of Michigan</td>
</tr>
<tr>
<td>Patrick T. Bidigare</td>
<td>U of Michigan</td>
</tr>
<tr>
<td>Jonathan Drexler</td>
<td>U of Minnesota-Duluth</td>
</tr>
<tr>
<td>Chaoeng Huang</td>
<td>Wright State University</td>
</tr>
<tr>
<td>Aaron Johnson</td>
<td>U of Kansas</td>
</tr>
<tr>
<td>Daniel Kern</td>
<td>U of Illinois at Chicago</td>
</tr>
<tr>
<td>Monica Lluis</td>
<td>U of Minnesota &amp; U of Colorado</td>
</tr>
<tr>
<td>Svetlana Rudnaya</td>
<td>U of Minnesota</td>
</tr>
<tr>
<td>Linda Smolka</td>
<td>Pennsylvania State University</td>
</tr>
<tr>
<td>Karen Zangara</td>
<td>Virginia Tech.</td>
</tr>
</tbody>
</table>
1 Introduction

The procedure for extracting iron ore from a mine follows the next steps: First the extraction which is basically digging and blasting, then the processing which includes crushing and washing the rocks and finally the storage of the rocks in stockpiles. The stockpiles are subsequently transported by rail to the place of final destination. We concentrate our attention on the storage within the stockpiles. See figures (4) and (1).

During the mining of iron ore it is common to use water to facilitate the extraction and crushing. But for transportation it is preferable that the ore have as low a water content as possible. However, the water content must not be too low or too much dust will be created. For iron ore the storage period may last from a few days up to two weeks.

![Image](image_url)

Figure 1: *Pictures of how the rock are moved by the stacker to form the stockpiles*

A typical stockpile is 20 m high, 1000 m long and 40 m wide and contains particles from 10 cm to 1 mm in size. Iron ore leaving the washery contains about 40% by volume of water. The desirable percentage of water in the stockpiles before the transportation is 16% by volume.

The aim is to develop a model of moisture movement in stockpiles to ore management to accelerate the drying process.

We consider three mechanisms that would contribute to model the moisture loss:

- Evaporation
- Inhomogeneous stockpile formation
based upon Poiseuille's Law. Section 6 uses Darcy's Law to analyze a more complicated model of the flow of the water within the stockpiles. Finally a steady state solution is given in the last section.

2 Rain Analysis

It is reasonable to believe that rainfall will affect the overall volume of water in the iron ore pile. Recall that after iron ore leaves the washery it is stacked into piles with the following characteristics:

- height = 20m
- length = 1000m
- width = 40m
- water = 40% (by volume)

Based on these parameters, the volume of the iron ore is approximately 400,000$\text{m}^3$. As the water volume is 40% (160,000$\text{m}^3$) of the pile and our goal for water volume is 16% (64,000$\text{m}^3$) of the pile, we wish to remove approximately 100,000$\text{m}^3$ of water during the storage period. In order to determine if rain is a primary contributor to the amount of water in the pile, we shall analyze the effect of exceptionally heavy rains during the two week storage period.

Assuming it rains one inch (.025m) we will accrue 1000$\text{m}^3$ of water assuming no water runs off the sides of the pile. In order for the rainfall to become significant (10% = 10,000$\text{m}^3$), it would need 10 inches of rain over the two week period. As it would require extreme (and rare) conditions for mining areas to receive this amount of rain, we infer that rainfall is not a primary contributor to the overall water volume. Thus, we will not include rain as a factor in our model.

3 Evaporation Study

In this section we consider the effect of evaporation on loss of water in our ore pile. We wish to show that evaporation of water from the pile has a negligible effect on total water loss.

Evaporation can occur in two ways, by diffusion and by advection. Diffusion occurs in the absence of wind as water vapor is evaporated from a
1 INTRODUCTION

Figure 2: *Top view of how the rock are moved by the stacker to form the stockpiles*

- Drainage

and try to answer the questions:

- How quickly does a stockpile dry?
- What is the water distribution in a stockpile?
- How the stockpiles should be made so they dry more quickly?

We will model the different aspects of the problem to decide how much some of them will contribute to the problem.

In section 2 a simple model was designed to determinate that the effect of the rain is not important to our problem. In section 3 a model using fluid dynamics was implemented to calculate the amount of water that evaporates for the worse case scenario. The next section studies the model of formations of the stockpiles. It is important to include and model the distribution of the stockpiles since it has been observed that the lower edges of the stockpile are significantly drier than the inner and the top part. This section gives information of the porosity and diameter of void space within the pile. This data are needed for studying how the water runs inside the stockpile. Section 4 models the distribution of the water in the stockpiles. This model is
surface and diffuses into the surrounding air. This process is very slow. The rate of evaporation increases dramatically when one considers advection. In this process, water vapor evaporates from a surface and is carried away by moving air, leaving dry air into which water evaporates much more quickly.

We wish to present a worst-case analysis of evaporation, thus we assume that evaporation occurs via advection as wind-forced air passes into one face of the ore pile, is humidified, and passes out the other face. To maximize this effect we assume that the wind is blowing perpendicular to the ridgeline as shown in Figure 3 below.

We model our ore pile as being permeated by many thin long pipes through which air may pass. We assume that all of the air that passes through the pile has become saturated with water (100% relative humidity).

For the purposes of modeling, we assume some worst-case conditions on temperature, wind speed and relative humidity shown in Table 1.

In order to determine the amount of water removed from the pile via evaporation, we need to calculate the amount of water in 1 cubic meter of
Table 2: Parameters in Ideal Gas Equation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Parameter</th>
<th>Value</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$</td>
<td>Pressure of Water Vapor</td>
<td>$7.381410^3$</td>
<td>$J/m^3$</td>
</tr>
<tr>
<td>$T$</td>
<td>Temperature</td>
<td>313</td>
<td>$K$</td>
</tr>
<tr>
<td>$R$</td>
<td>Gas Law Constant</td>
<td>8.31441</td>
<td>$J/K \cdot mol$</td>
</tr>
<tr>
<td>$V$</td>
<td>Volume of Water Vapor</td>
<td>1</td>
<td>$m^3$</td>
</tr>
<tr>
<td>$n$</td>
<td>Moles of $H_2O$</td>
<td></td>
<td>$mol$</td>
</tr>
</tbody>
</table>

air at 100% relative humidity. The ideal gas model tells us that

$$n = \frac{PV}{RT}$$  

(1)

The meaning of the variables $P$, $V$, $T$, $R$ and $n$ are given in Table 2.

We found that at $40^\circ C$, the vapor pressure of water is $P = 7.3814 \times 10^3 J/m^3$ thus the number of moles of water in 1 cubic meter of water vapor is obtained from (1) to be $n = 2.836 mol$. Using the atomic weight 18 of water and it’s density as a liquid $1000 Kg/m^3$ we find that the volume of water (as a liquid) which will evaporate to give one cubic meter of water vapor at the above temperature is

$$51 \times 10^{-6} m^3 H_2O \text{ liquid}/m^3 \text{ vapor}$$

Actually, the amount of water in one cubic meter of saturated air will be smaller than this because of the presence of the air molecules, however this number gives us a good worst-case estimate.

We will now use fluid dynamics to study the flow of air through a single capillary. This will be modeled by the basic equations of conservation of mass and momentum.

$$\frac{d\rho}{dt} + \nabla \cdot (\rho V) = 0$$  

(2)

$$\rho \left( \frac{\partial V}{\partial t} + V \cdot \nabla V \right) = -\nabla p + \nabla (\mu \nabla \cdot V) - \rho g \hat{j}$$  

(3)

where $p$ is the air pressure, $\mu$ is the viscosity of the air (considered constant), $\rho$ is the density of the air (also constant in the incompressible case), and $V$ is the vector velocity of the wind.

For the steady state solution, we have that $\frac{\partial V}{\partial t} = 0$. Therefore the equations (2) and (3) can be written as:
\[ \nabla \cdot (\rho V) = 0 \]  
(4)

\[ \rho \left( V \cdot \nabla V \right) = -\nabla p + (\mu \nabla^2 \cdot V) \]  
(5)

Now, we proceed to make dimensionless equations by the following change of variables:

\[ x = L \bar{x} \]
\[ y = A \bar{y} \]
\[ p = p_I + (p_0 - p_I) \bar{p} \]

where \( A \) is the typical radius of the pipe, \( p_I \) is the pressure at the entrance of the pipe and \( p_0 \) is the pressure at the end and \( L \) is the length of the pipe.

\[ \bar{u}_x + \bar{u}_y = 0 \]  
(6)

\[ \rho \left( \frac{p_0 - p_I}{L^2 \mu^2} \right) A^4 \left( \bar{u} u_x + \bar{v} u_y \right) = -\bar{p}_x + \bar{u}_{yy} + \left( \frac{A}{L} \right)^2 \bar{u}_{xx} \]  
(7)

\[ \rho \left( \frac{p_0 - p_I}{L^4 \mu^2} \right) A^6 \left( \bar{u} u_x + \bar{v} u_y \right) = -\bar{p}_y + \left( \frac{A}{L} \right)^2 \bar{v}_{yy} + \left( \frac{A}{L} \right)^2 \bar{v}_{xx} \]  
(8)

We now consider \( A \ll L \) and viscosity \( \mu \) as a large value compared to inertia, thus we have the following:

\[ \bar{u}_x + \bar{u}_y = 0 \]  
(9)

\[ 0 = -\bar{p}_x + \bar{u}_{yy} \]  
(10)

\[ 0 = -\bar{p}_y \]  
(11)

Now let \( r = f(z) \) the radius of our pipe at position \( z \). We impose the "no-slip" boundary conditions \( \bar{u}(f(z)) = \bar{v}(f(z)) = 0 \) for all \( z \). Changing the above equations to polar coordinates and eliminating the bars we have:

\[ u_x + u_r = 0 \]  
(12)

\[ -p_z + \frac{1}{r} \frac{d}{dr} \left( ru_r \right) = 0 \]  
(13)
$-p_r = 0 \quad (14)$

The following equation can then be found for pressure and velocity:

$$p(z) = p_1 + \left(p_0 - p_1\right) \frac{\int_0^z f^{-3}(s)ds}{\int_0^L f^{-3}(s)ds} \quad (15)$$

$$u(z) = \frac{1}{4\mu} p'(z) \left(r^2 - f^2(z)\right) \quad (16)$$

Using Bernoulli’s equation where $u_0$ is the initial velocity, $p_0$ is the initial pressure,

$$u_0^2 = \frac{2(p_0 - p)}{\rho} + u^2 \quad (17)$$

we can consider $u$ small and solve for the volume flux using the assumption that we have a constant radius $R$. Thus $f(z) = R$

$$V_f = \int_0^{2\pi} \int_0^f r u(r, z) drd\theta \quad (18)$$

$$V_f = \frac{\pi \rho u^2 R^4}{16\mu L} \quad (19)$$

Since this is the volume flux for one pipe passing through the stockpile, we now need to estimate the number of pipes. Since we have 40% water, we take the largest cross-sectional area to find a worst case scenario for the number of pipes, $N$.

$$N = \frac{8 \cdot 10^3}{\pi R^2} \quad (20)$$

The total volume of air ($V_a$) passing through the stockpile is now computed

$$V_a = \frac{8 \cdot 10^3 \rho u^2 R^2}{16L\mu} \quad (21)$$

Using the previous calculations, we have $51 \times 10^{-6} m^3$ of water in $1 m^3$ of completely saturated air. We will now examine the worst case scenario with the following conditions:

- Wind speed $u = 30 mph = 13.41 \frac{m}{s}$
4 INITIAL INVESTIGATION OF WATER FLOW

- Temperature $T = 40^\circ C$
- Constant viscosity of the air $\mu = 1.9 \cdot 10^{-5} \frac{kg}{m \cdot s}$
- Average length of the pipe $L = 20m$
- Constant density of the air $\rho = 1 \frac{kg}{m^3}$
- Average radius of each pipe $R = 1mm = 10^{-3}m$
- Time interval $t = 2$ weeks $= 1209600s$

With all of these considerations the volume of water removed from the stockpile by evaporation is $14,620m^3$ which represents 9% of the initial amount of total water in the stock piles. This appears to be a sizable amount of water removed. However, we must point out the following unrealistic assumptions used in our calculations. First of all, the actual channels will have an irregular twisting shape which will greatly reduce the velocity of the wind. Thus, the wind will not be able to pass straight through all of the capillaries. Furthermore, the air is not likely to become fully saturated with water vapor after “passing” through the stockpile. In addition, wind speeds of $30mph$ lasting for two weeks is highly unlikely, and due to the wind flow, pressure inside the channels will be smaller. When all of the realistic observations are taken into account, it appears that evaporation will decrease the water amount by much less than 9%. Thus, we assert that evaporation is not the dominate contributor to water removal and will only play a factor in extreme conditions.

4 Initial Investigation of Water Flow

In this section we will begin to analyze the flow of water through the stockpile. For this initial investigation, we use the simple model of straight pipes with uniform radius passing vertically through the stock pile. This model is based upon Poiseuille's Law which describes the steady flow of a fluid in a straight pipe.

4.1 Model Description and Assumptions

We are assuming we have a steady, incompressible fluid with no slip conditions on the boundaries. Furthermore, the atmospheric pressure will remain
constant throughout the capillaries, thus the pressure gradient will be zero in this model. (See figure)

Using Poiseuille's Law (reference Currie) for steady flow in straight pipes, we have the following:

\[ \rho \left( u \cdot \nabla \right) u = -\nabla p + \mu \nabla^2 u - g \rho \]  \hspace{1cm} (22)

where \( \rho \) is the water density, \( u(y) \) is the velocity, \( \mu \) is the dynamic viscosity of water, \( g \) is the acceleration due to gravity and \( p(y) \) is the pressure. Equation (22) has the following components:

- \( \rho \left( u \cdot \nabla \right) \) represents the inertia
- \( -\nabla p \) is the pressure gradient
• $\mu \nabla^2$ accounts for the viscosity in the fluid

• $g\rho$ is the gravitational term

Since we have an incompressible fluid, we have $\nabla \cdot u = 0$. We will now compare the remaining components inside equation (22) to the gravitational term, which must be included in this model. When comparing the inertial and gravitational components, we find

$$u \sim \sqrt{Lg} \sim 14 \frac{m}{sec}$$

This implies that the velocity will be unrealistically large if inertia plays a dominant role in the equation (22). However, when we compare the viscous component with the gravitational term, we have for a given diameter, $d$, of 1 mm:

$$u \sim \frac{gd^2}{\mu} \sim 10^{-4} \frac{m}{sec}$$

which is a fairly reasonable approximation for the velocity of the fluid. Thus we know for a small velocity, the inertial component is negligible (i.e. $\rho(u \cdot \nabla) = 0$).

We can now rewrite equation (22) as

$$0 = \mu \nabla^2 u - g\rho \quad (23)$$

$$0 = \mu (u_{zz} + u_{yy}) - g\rho \quad (24)$$

$$u_{zz} + u_{yy} = \frac{g\rho}{\mu} \quad (25)$$

After converting the above equation into polar coordinates (assuming an axisymmetric pipe), we will have

$$\frac{1}{R} \frac{d}{dR} \left( R \frac{du}{dR} \right) = \frac{g\rho}{\mu} \quad (26)$$

$$u(R) = \frac{g\rho}{4\mu} \left( a^2 - R^2 \right) \quad (27)$$

Where $a$ is the constant radius of the capillary and $r$ is the radial coordinate where velocity is being analyzed.
Averaging the velocity function over a cross sectional area, we have

\[
\bar{u} = \frac{g\rho}{\pi 4\mu a^2} \int_0^{2\pi} \int_0^a r \left( a^2 - r^2 \right) dr d\theta
\]  

(28)

\[
\bar{u} = \frac{g\rho a^2}{8\mu}
\]  

(29)

We know the stockpile is usually ready to ship in 10 days. Thus we assume that the worst case scenario for the average velocity of the water is approximately:

\[
\bar{u} = 1.6 \cdot 10^{-5} \text{ m/sec}
\]  

(30)

By equating (29) and (30) we can find the approximate maximum radius for each capillary as

\[
a = 10^{-6} \text{ m} = 1 \mu\text{m}
\]  

(31)

This appears to be unrealistically small for the radius of the capillaries. Thus we can conclude that pressure inside the capillaries needs to be accounted for in our model. At this point we attempt to construct a more realistic model for fluid flow by analyzing the pressure acting on the water inside the capillaries.
5 Stockpile Formation

The motivation behind our study of stockpile formation developed from the group's hypothesis that composition of the pile influenced the flow of water within the pile. Our hypothesis, based on observations that the lower edges of the stockpile were significantly drier than the core, assumes the distribution of rock size and void space are not homogeneous throughout the cross-section of the stockpile. To determine whether our hypothesis is true a model which simulates pile formation was developed. From our study we hope to make conclusions about the distribution of porosity and diameter of void space within the pile.

5.1 Background

Our approach was based on the work of several authors. The principle theory in pile formation was developed by Bak, Tang and Weisenfeld[1], henceforth referred to as (BTW). Studies done by (BTW) on avalanches, simulated using sandpiles, were performed to understand the self-critical organizational nature of piles. (BTW) asserted that when the slope of a pile exceeded a critical angle an avalanche would result. To simulate such a situation a method known as cellular automata was employed. A cellular automata model involves dropping “cells” over random positions of a pile. A set of rules are established to define when these cells are unstable, normally expressed in the form of slope criteria, then rules for reorganization are imposed to establish a stable pile configuration. Specifically the rules to our model were based on the work of L. P. Kadanoff, et al[5].

An example of a one-dimensional cellular automaton model, similar to one provided by H. M. Jaeger and S.R. Nagel[7], is given in (fig.1). In this model the cells are of equal size. Here the local rules which govern a “simple avalanche” are based on the height of a column relative to neighboring columns. Height is determined by the number of cells stacked in a column. If the height difference is greater than one, then a shift of the cell will occur. Initially the pile is stable (fig. 1.a). The shaded block is dropped in a random position. Since the critical height has been exceeded, the cell moves down and to the right until it has settled to a stable position (fig. 1.b-d). This model takes into account only local instability behavior, since it does not allow for global shifting in the pile. Hence once a cell becomes stable in the pile, it remains stable, a gross simplification.

In the following sections we introduce a cellular automata model developed to simulate “simple avalanche” dynamics for the iron-ore stockpile. We refer to this as a “simple avalanche” model since dynamics are controlled locally and discretely, as in the above example, rather than in a global framework, with the intent of defining the global structure of the stockpile. We look to our results to determine whether the model employed displays basic features of self-organized criticality.
Figure 1: Example falling for a size one block.

5.2 Assumptions of Model

- The size of rocks were specified in the range of 10 cm to 1 mm. However, in our model we consider rocks, which are approximated as square blocks of two sizes, the sides of which are of ratio 2:1.

- We will then approximate the rocks as circles contained within the square block in order to represent void space between rocks in specific regions of the stockpile.

- For simplicity we consider any side elevation cross-section of the stockpile to be similar in composition. For this assumption to be valid we require the stacker to sweep uniformly back and forth depositing iron-ore, rather than dumping rock in one position for an extended period of time before moving to a new position. This method will produce a ridge of constant height as viewed from the front elevation.

- Since we view the composition of the pile as a cross-sectional area, we define percentage of rock in terms of area. In case(A) we assume that by cross-sectional area 50% of the stockpile will be composed of large blocks (2 x 2) and 50% by small blocks (1 x 1). In case(B) we let 80% of the cross-sectional area be composed of large blocks and 20% of small blocks.

- In the model one block drops at a time. The stockpile settles in a stable configuration before a new block is dropped.

5.3 The model

A matrix is defined to represent the cross-sectional area of the stockyard. The dimensions of the actual stockyard are preserved by the choice of rows and columns of the matrix. A coordinate \((m, n)\) in the matrix represents the column and row of the cellular model, respectively. A 1 x 1 block is represented as a 1 in the matrix with position \((m, n)\). A 2 x 2 block is represented as a quadrant of 2's with the lower left corner of the block's position denoted by \((m, n)\). The height of a stack of blocks are represented by the function \(H(m)\), where \(m\) specifies the column in which the blocks reside.
In simplified form the model has four key steps. Once the matrix and heights are initialized the size of a block to be dropped is determined. A random number generator is employed to insure the correct distribution of large to small blocks are dropped for case(A) and case(B). For instance in case(A), there is a 50% distribution of large blocks to small blocks in terms of area. This requires 20% of the rocks dropped to be large, 80% small. Next the drop column of the block is determined. To simulate the stacker, an array of 5 positions is established from which a block may be dropped. To assure randomness in the model we again use a random number generator to determine from which of the five positions the block starts. Once dropped, the stability of the block is checked. If the block is unstable, it will be shifted to a new position, at which point stability is again checked. After the block reaches a stable position, the process repeats for a new block.

5.3.1 Stability Requirements

The stability requirements are specific to each block size.

1 × 1 case: In the case of the small block height instability occurs only when a maximum slope is exceeded, similar to the model of L.P. Kadanoff, et al[5]. When a small block lands in its initial position the height function is incremented by

\[ H(m) = H(m) + 1 \]

If the slope to the left

\[ \sigma_L = H(m) - H(m - 1) \]

is greater than 1, then the block is shifted in the appropriate direction, one unit over and one unit down. (see fig 1) The new position is reevaluated until the block is stable, at which point a 1 is placed into the matrix.

2 × 2 case: In the case of a large block, instability can occur in one of two ways. The first instability evaluated is called a ledge instability. In this case a large block lands on a small block such that one half of the block is unsupported (see fig. 2) and is thus on a ledge. The block will reposition itself if there is room in the respective direction that it is unsupported to move. Since the large block has sides of length two, the block will move over two blocks and down one, if these positions in the matrix are void. If not, a large block will remain on the ledge, introducing void space into the pile. (see fig. 3)

Once a block is ledge stable the model checks for height stability. When a large block lands in a column the height function is incremented by

\[ H(m) = H(m) + 2 \]
Figure 2: Shifting of size two block due to ledge instability.

Figure 3: A large block lands on a ledge but has no room to shift.

If the slope to the left or right, similar to the $1 \times 1$ case, exceeds 3 units and the appropriate space exists, then the block will shift two units over in the direction of height instability and one unit down. The model continues the process of checking for instability, in the order specified, until the block is stable. Once stable a quadrant of 2's is placed in the matrix and a new block is dropped.

Figure 4: Shifting of size two block due to height instability.

5.4 Results

To model the stockpile our code drops two different sizes of blocks within a given region. These blocks then settle onto the pile following the rules described above. We ran 15 trials of the pile formation under four sets of conditions. These 15 trials were averaged together, point by point, to get a temporal average of the cross section of the pile. The temporal averaging enabled us to understand
general trends in the evolution of the pile structure. Tendencies that recurred through all runs of the model emerged more strongly.

Figure 5: Pile block distributions in terms of area.

We examined two different ratios by area of large blocks to small blocks. In case(A) the ratio of the area of large to small blocks is 1:1, hence there will be four times as many small blocks as large blocks. (fig. 5.A) In case(B) we considered a 4:1 ratio by area of large blocks to small blocks, hence the number of each is equal. (fig. 5.B)

In fig.6 are temporal averages of two piles each with 5,000 blocks, representing case(A) and case(B). NOTE: The heavy lines along the slopes of the pile are artifacts of the temporal averaging, created by the irregular shape of the slopes on a fine level. These heavy lines do not affect the global behavior which has been analyzed. In case(B) where there are greater numbers of large blocks, we see sorting within the pile. The large white regions in figure (6) at the two lower edges of the pile correspond to areas where the temporal average is greater than 1.5. This corresponds to an area composed mainly of large blocks. As specified, there are more stability criteria that must be met for a large block to come to rest on the pile. When we have a substantial fraction of the pile comprised of these large blocks, they fall to the edges of the pile where they mix less with the smaller blocks.

Figure 6: A pile of 5,000 blocks with ratio expressed in terms of area of large to small blocks.

When the large blocks make up a smaller fraction of the pile by numbers, as in case(a), the distribution of block sizes is homogeneous.
The next figure (7) shows temporal averages of piles of 10,000 blocks. While the overall pile size increases, the structural patterns of large and small block sorting are qualitatively the same, as above.

Figure 7: A pile of 10,000 blocks with ratio expressed in terms of area of large to small blocks.

Hence these results support our hypothesis that sorting by self-organized criticality occurs within the pile only if the number of large blocks exceeds the number of small blocks in the pile.
6 Saturated & Unsaturated Flow Model

In this section we analyze a more sophisticated model for fluid flow within the ore pile. This model is based on the law of conservation of mass as well as a crucial relationship called Darcy's Law which relates the velocity of fluid through a porous media to the pressure of the liquid.

6.1 Model Description and Assumptions

Our pile of iron ore will consist of many different blocks with varying shapes and diameters piled up in such a way that there is a complicated system of water pathways formed from the spaces around the ore particles. Rather than try to model this network of flows around the rocks, we treat the entire pile as a continuous porous media through which water can flow. The ability of the porous material to conduct the flow of water is called the hydraulic conductivity and denoted by $K$.

In this model we do not assume that the hydraulic conductivity is constant throughout the pile (so it is a heterogeneous media), but we do assume that at a specific point in the media, the resistance to water flow does not depend on the direction of the flow (so the media is isotropic).

One critical assumption that we make concerns the way in which the ore pile is built. Since the pile is being formed as the ore stacker moves along the ridge and deposits ore atop the pile, it is reasonable to assume that every cross section of the hill (perpendicular to the ridge-line) is essentially the same, thus fluid flow, water pressure, hydraulic conductivity, etc... do not depend on the position along the ridge. This assumption allows us to reduce our problem to a two dimensional model and consider fluid flow in a cross section of the ore pile. This is illustrated in Figure 6.1.

Our model assumes both the iron ore and the water are incompressible substances, thus the density $\rho$ of water is constant. We also assume that no settling of ore particles occurs as fluid flow takes place (thus the hydraulic conductivity is constant with respect to time).

6.2 Conservation of Mass

Consider a small square of porous media as shown in Figure 6.1. Since the media is porous, it will not be completely filled by iron ore. The area in the square which is not occupied by ore is called the void space. The voids may be completely filled with water (in which case we say they are saturated)
or may be partially filled with water and partially filled with air (and are called unsaturated). The porosity $n$ is defined as the ratio of the volume of the void space to the total volume. It is a percentage of the total volume. The moisture content $\theta$ is defined to be the ratio of the volume of the void space filled with water to the total volume of the square. We see that for saturated flow we have $\theta = n$ but for unsaturated flow we have $\theta < n$. It will be useful to define the degree of saturation $S$ to be $S = \theta/n$. Since we assume our ore media to be incompressible, $n$ will not change with time.

We denote by $v = (v_x, v_y)$ the velocity of the water at any point in the porous media. Note that $\rho v$ represents the mass flow of water (per unit length) in the direction of the fluid flow. Further, the product $\rho \theta$ is the total mass per unit area. The law of conservation of mass requires that the net rate of fluid mass flow into square of porous media be equal to the time rate of change of fluid mass within the square. Thus we have the equation:

$$
-\frac{\partial}{\partial x} \rho v_x - \frac{\partial}{\partial y} \rho v_y = \rho \frac{\partial \theta}{\partial t}
$$

(32)
\[ \rho v_y - \frac{\partial \rho v_y}{\partial y} \]

\[ \rho v_x - \frac{\partial \rho v_x}{\partial x} \]

\[ \rho v_y \]

Figure 6: Diagram of fluid flow in small square of porous media
6.3 Darcy’s Law

We next wish to determine a relationship between velocity and pressure. To begin with, we make a distinction between real and gauge pressures. The real pressure of a liquid is the pressure including the atmospheric pressure $p_a$. The gauge pressure refers to the pressure of the liquid above atmospheric pressure. While we cannot have a negative real pressure, it is common to encounter negative gauge pressures. In this paper when we speak of pressure $p$ we are referring to gauge pressure, not real pressure.

There are three forces acting upon the water in the ore pile. First, pressure differences can cause water to move from an area where pressure is higher to an area of lower pressure. Second, gravity tends to cause the water to move downward. Third, the viscous friction of the water in the porous media exerts a force opposite to the direction of the fluid flow. These three effects are incorporated in an empirical law called Darcy’s law. Darcy’s law comes from much experimental evidence and has shown to be a very accurate tool for modeling fluid flow through a porous media. Darcy’s law in our case is given by:

$$
\begin{align*}
v_x &= -K \frac{\partial p}{\rho g \partial x} \\
v_y &= -K \left( \frac{1}{\rho g} \frac{\partial p}{\partial y} + 1 \right)
\end{align*}
$$

Where $K$ is our hydraulic conductivity, and $g$ is the gravitational acceleration constant. These equations relate the velocity of the water in the media to the pressure gradient.

In our model we will use Darcy’s law to develop our boundary value problem. By substituting Darcy’s law into the equation for conservation of mass, multiplying through by $g$, moving some terms around and applying the chain rule we get:

$$
\frac{\partial}{\partial x} \left( K \frac{\partial p}{\partial x} \right) + \frac{\partial}{\partial y} \left( K \frac{\partial p}{\partial y} \right) = \rho g \left( \frac{\partial \theta}{\partial p} \frac{\partial p}{\partial t} - \frac{\partial K}{\partial y} \right)
$$

6.4 Functional Relationship Modeling

In the equation above, the hydraulic conductivity $K$ and the moisture content $\theta$ will depend heavily on the water pressure. Before any solution to the above PDE can be obtained we need to model the functional relationships between $\theta$ and $p$ as well as $K$ and $p$. 
6.4.1 Moisture Content Dependency on Pressure

When the gauge pressure of the water in our porous media is positive, no air can enter the media and the moisture content will simply be the porosity ($\theta = n$). In this saturated condition, $\theta$ will be constant with respect to $p$. When the size of the ore particles is relatively small, there will actually be a small negative water pressure $p_s$ for which our media still remains saturated. This is called the tension saturated pressure, and is illustrated in Figure 6.4.1. For pressures below $p_s$ we need some experimental results to determine the relationship between $\theta$ and $p$.

Recall that $S = \theta/n$ is the degree of saturation of our media. Because of hydrostatic attraction between the water and the media, even if we allow our media to completely drain, a small amount of water will remain in the ore stuck to the surfaces within. This final saturation value is called the irreducible wetting fluid saturation and is denoted $S_0$. From this quantity we can define another type of saturation $S_e$ called the effective saturation which is given by:

$$S_e = \frac{S - S_0}{1 - S_0}$$

Examples documented in ?? have shown that if pressure is graphed as a function of effective saturation on log-log paper, a straight line is obtained except for $S_e$ close to unity. The slope of this line $\lambda$ is called the pore-size distribution. Mathematically we have

$$\log_{10} \left( \frac{-p}{\rho g} \right) = -\lambda \log_{10}(S_e) + b$$

Thus

$$S_e = \left( \frac{-p}{B \rho g} \right)^{-1/\lambda}, \quad B = 10^b$$

When $p = p_s$ we have $\theta = n$, thus $S_e = 1$. Solving for $B$ in the above equation we get:

$$B = \frac{-p_s}{\rho g}$$

This gives us an elegant model for the effective saturation as a function of pressure:
\begin{equation}
S_e = \begin{cases} 
1 & \text{if } p \geq p_s \\
\left(\frac{p}{p_s}\right)^{-1/\lambda} & \text{if } p < p_s 
\end{cases}
\end{equation}

Converting this to an expression for $\theta$ as a function of $p$ we get

\begin{equation}
\theta = \begin{cases} 
n & \text{if } p \geq p_s \\
n\left(\frac{p}{p_s}\right)^{-1/\lambda}(1 - S_0) + S_0 & \text{if } p < p_s 
\end{cases}
\end{equation}

The graph of this function is shown in Figure 6.4.1.

6.4.2 Hydraulic Conductivity Dependency on Pressure

The hydraulic conductivity exhibits two different behaviors depending on the pressure in the porous media. When our media is saturated with water (i.e. for pressures $p > p_s$), the hydraulic conductivity $K$ does not depend on pressure. We call this the *saturated hydraulic conductivity* and denote it by $K_{SAT}$. Using the fluid dynamical results and assumptions from the previous section we may obtain a formula for $K_{SAT}$. 
\[ K_{\text{SAT}} = \frac{C a^2 \rho g}{8 \mu} \]

Here \( a^2 \) is the mean square radius of the pores at a particular place in the porous media and \( \mu \) is the viscosity of water.

Our current model, unlike the model of the previous section, takes into account that the path the water follows varies in diameter and does not go straight down. Here \( C \) is a constant which depends on various properties of the soil such as porosity, roundness and size of particles, etc.

When our porous media becomes unsaturated (when \( p < p_s \)), the hydraulic conductivity becomes dependent on pressure. Experimental evidence suggests that we have:

\[ K(S) = K_{\text{SAT}} S_e^3 \]

Where \( S_e \) is the effective saturation defined above. From equation 35 we obtain an expression for \( K \) as a function of \( p \).

\[ K = \begin{cases} 
K_{\text{SAT}} & \text{if } p \geq p_s \\
K_{\text{SAT}} \left( \frac{p}{p_s} \right)^{-3/\lambda} & \text{if } p < p_s 
\end{cases} \quad (37) \]

### 6.5 Boundary and Initial Conditions

At this point we have a complete partial differential equation describing the water pressure within the ore pile. We now need to determine the boundary conditions on the problem.

Our ore pile will have three bounding lines, namely the bottom and the two sides (refer to Figure 6.1). Consider first the boundary condition on the bottom of the pile. We assume that the ore pile sits on a non-porous rock bed, and thus no water can leave the pile from the bottom. Mathematically this says that

\[ u_y = 0 \quad \text{for } y = 0 \]

Applying Darcy's law we get a boundary condition on the pressure at the bottom of the pile.

\[ \frac{\partial p}{\partial y} = -\rho g \quad \text{for } y = 0 \quad (38) \]
The boundary conditions on the sides of the ore pile are a bit more complicated. The water table is the pressure level curve where $p = p_s$. This is the dividing line between saturated and unsaturated flow. This is illustrated in Figure 6.5.

In an unsaturated region, no water can leave the ore pile from the sides, thus along the sides, the component of velocity in the direction of the normal vector $n = (\pm H, L/2)$ must be zero. Applying Darcy’s law to the expression $v \cdot n = 0$ we get the following boundary conditions:

$$-H \frac{\partial p}{\partial x} + \frac{L}{2} \frac{\partial p}{\partial y} + \frac{L \rho_{aq}}{2} = 0 \quad \text{Left Unsaturated Boundary}$$

$$H \frac{\partial p}{\partial x} + \frac{L}{2} \frac{\partial p}{\partial y} + \frac{L \rho_{aq}}{2} = 0 \quad \text{Right Unsaturated Boundary}$$

In the saturated region, we see that while water will flow out of the boundary, the pressure at the boundary will be 0 (atmospheric pressure), thus we get the boundary equations:

$$p = 0 \quad \text{Left and Right Saturated Boundaries}$$
Notice at all points along the left and right boundaries, one of these boundary conditions will always hold, i.e. one of the equations will always be zero. Because of this we can multiply these together and obtain a boundary condition which holds over the entire side. This is a much better condition, because it does not require that we know where the unsaturated flow becomes saturated. The boundary conditions are given below.

Left Side:

\[
p \left( -H \frac{\partial p}{\partial x} + \frac{L \partial p}{2 \partial y} + \frac{L \rho g}{2} \right) = 0
\]

(39)

Right Side:

\[
p \left( H \frac{\partial p}{\partial x} + \frac{L \partial p}{2 \partial y} + \frac{L \rho g}{2} \right) = 0
\]

(40)

As for initial conditions, we assume that to start with, our ore pile is completely saturated with water, thus the water pressure within the pile must be greater than \( p_s \) everywhere. That is

\[p > p_s \quad \text{for } t = 0\]

(41)

### 6.6 Non-dimensional Analysis

In this section we perform a non-dimensional analysis of the partial differential equation developed to model the pressure within the ore pile. This analysis allows us to determine which terms in the equation contribute significantly to the solution.

#### 6.6.1 Non-dimensional Equations

We make the following substitutions in order to obtain a set of non-dimensional variables:

\[
x = L \bar{x} \\
y = L \bar{y} \\
t = T \bar{t} \\
p = -p_s \bar{p} \\
K = K_{SAT} \bar{K} \\
\theta = n S_0 + n(1 - S_0) \bar{\theta}
\]

(42)
With these substitutions our non-dimensional version of equation 34 becomes:

\[ \frac{\partial}{\partial x} \left( \bar{K} \frac{\partial \bar{p}}{\partial x} \right) + \frac{\partial}{\partial y} \left( \bar{K} \frac{\partial \bar{p}}{\partial y} \right) = Q_1 \left( \frac{\partial \bar{\theta}}{\partial \bar{t}} \right) - Q_2 \frac{\partial \bar{K}}{\partial y} \]  

(43)

We should point out that the terms on the right hand side of the above equation represent the time-dependency and gravitational effects for the model. The relative size of the coefficients \( Q_1 \) and \( Q_2 \) will tell us the importance of each of these terms.

The coefficients \( Q_1 \) and \( Q_2 \) are given by:

\[ Q_1 = -\frac{\rho g n (1 - S_0) L^2}{TK_{SAT} p_s} \]  

(44)

\[ Q_2 = -\frac{\rho g L}{p_s} \]  

(45)

The equations for the non-dimensionalized hydraulic conductivity and moisture content become:

\[ \bar{K}(\bar{p}) = \begin{cases} 
1 & \text{if } \bar{p} \geq -1 \\
(-\bar{p})^{-3/\lambda} & \text{if } \bar{p} < -1 
\end{cases} \]  

(46)

\[ \bar{\theta}(\bar{p}) = \begin{cases} 
1 & \text{if } \bar{p} \geq -1 \\
(-\bar{p})^{-1/\lambda} & \text{if } \bar{p} < -1 
\end{cases} \]  

(47)

The bottom boundary condition becomes:

\[ \frac{\partial \bar{p}}{\partial \bar{y}} = \frac{\rho g L}{p_s} \text{ for } \bar{y} = 0 \]  

(48)

The left and right boundary conditions are

Left Side:

\[ \bar{p} \left( -\frac{\partial \bar{p}}{\partial x} + \frac{\partial \bar{p}}{\partial y} - \frac{L \rho g}{p_s} \right) = 0 \]  

(49)

Right Side:

\[ \bar{p} \left( +\frac{\partial \bar{p}}{\partial x} + \frac{\partial \bar{p}}{\partial y} - \frac{L \rho g}{p_s} \right) = 0 \]  

(50)

And our initial condition becomes

\[ \bar{p} > -1 \text{ for } t = 0 \]  

(51)
### Table 3: Typical Parameter Values

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Typical Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Porosity</td>
<td>$n$</td>
<td>$0.4 = 40%$</td>
</tr>
<tr>
<td>Mean Pore Size</td>
<td>$a$</td>
<td>$10^{-3} m$</td>
</tr>
<tr>
<td>Water Density</td>
<td>$\rho$</td>
<td>$1000 Kg/m$</td>
</tr>
<tr>
<td>Gravitational Acceleration</td>
<td>$g$</td>
<td>$9.8 m/s$</td>
</tr>
<tr>
<td>Water Viscosity</td>
<td>$\mu$</td>
<td>$1.15 \times 10^{-4} Kg/m \cdot s$</td>
</tr>
<tr>
<td>Wetting Fluid Saturation</td>
<td>$S_0$</td>
<td>0.4</td>
</tr>
<tr>
<td>Ore Pile Height &amp; Length</td>
<td>$L$</td>
<td>20m</td>
</tr>
<tr>
<td>Tension Saturated Pressure (sand)</td>
<td>$p_s$</td>
<td>$-2940 Kg/m \cdot s$</td>
</tr>
<tr>
<td>Pore Size Distribution (sand)</td>
<td>$\lambda$</td>
<td>7.3</td>
</tr>
</tbody>
</table>

### 6.6.2 Typical Parameter Values

The following table gives a list some typical values for the parameters used in this non-dimensional analysis. These parameter values will be used in the next section to determine the relative importance of the terms in our non-dimensionalized equations above.

Using the values in the table above we can obtain values for the non-dimensional coefficients $Q_1$ and $Q_2$.

$$Q_1 = \frac{4 \times 10^6}{T}$$

$$Q_2 = 67$$

### 6.6.3 Non-dimensional Analysis

Consider now the coefficients $Q_1$ and $Q_2$. We see that over a small time scale (say an hour, $T = 3600s$), the time-dependent term will dominate the gravity term. Physically this means that initially the drainage from the pile controls the pressure within the pile. On a much larger time scale (say one week, $T = 604800s$) $Q_1$ will be two orders of magnitude smaller than $Q_2$. Physically this says that drainage no longer significantly affects the water pressure in the pile and that we have essentially achieved steady state.

In general our ore pile will be held in storage for over a week. What is crucial to know is how much water remains in the pile once it is time to ship it off for further processing. Our next step, then, is to determine a steady state solution for the ore pile (which we will essentially achieve in a weeks time) and determine the amount of water still present in the pile.
6.7 Steady State Solution

It was stated in the last section, that after approximately one week we reached a steady state condition where no more water drains from the stockpile. The amount of water left in the stockpile is completely determined at this point by the steady state solution of the problem. Neglecting the time dependent term we get the following steady state equation:

\[
\frac{\partial}{\partial x} \left( K \frac{\partial p}{\partial x} \right) + \frac{\partial}{\partial y} \left( K \frac{\partial p}{\partial y} \right) = -Q_2 \frac{\partial K}{\partial y} \tag{52}
\]

It is straightforward to verify that the equation

\[
\bar{p}(\bar{x}, \bar{y}) = -Q_2 \bar{y} - 1 \tag{53}
\]

satisfies both the steady state differential equation above as well as the boundary conditions given in equations (48), (49) and (50).

The dimensional solution is thus given by:

\[
p = -\rho g y + p_s
\]

We now use equation (36) to obtain an expression for water content as a function of height.

\[
\theta(y) = n \left( \left( \frac{-\rho g y}{p_s} + 1 \right)^{-1/\lambda} \right)^{(1 - S_0) + S_0} \tag{54}
\]

A graph of this function (for \(\lambda = 7.3\)) is shown in Figure 6.7.

As one can see, the moisture content left in steady state is still very high (around 30%). This is a pessimistic result from the standpoint of shipping because an optimal water content for shipping is 16%.

The physical reason for our high water content is that most of the water, held in place by static pressure, never drains from the stack. The value of \(p_s\) and \(\lambda\) we used were for fine sand. Although we were not able to get true values for \(p_s\) or \(\lambda\) for our ore pile, we can use the values for silt which are \(p_s = 6860Kg/m \cdot s\) and \(\lambda = 1.82\). The steady state moisture content graph for these values is shown in Figure 6.7.

As can be seen from this graph, our ore pile had drained out to about 18% water content, close to the value we want for shipping.
Figure 9: Moisture Content vs. Height for $\lambda = 7.3$
Figure 10: Moisture Content vs. Height for $\lambda = 1.82$
7 Conclusions

We have found several interesting results about water flow through a pile of iron ore. We have shown that under normal conditions evaporation and rain do not significantly affect the moisture content of the pile. Furthermore, an accurate $\lambda$, pore-size distribution, is crucial to obtaining the correct moisture content of the pile. Physically, it is difficult to determine the pore-size distribution. One method is to analyze grain-size distribution using statistical analysis from cross-sections of the pile and drawing tentative conclusions about $\lambda$. Another approach to finding $\lambda$ is to use the computer model of the stockpile formation.

If one uses the computer model with equal numbers of large and small blocks, then the lower outer regions are composed almost exclusively of larger blocks. We hypothesize that $\lambda$ will be slightly smaller in these regions. Furthermore, $p_3$ will be closer to atmospheric pressure. Both of these factors imply the steady state solution for water content will be less in these regions. This explains the experimentally seen phenomena in piles.

When the computer model has four small blocks for each large block, then we have a well mixed pile. This produces a small $\lambda$ over the entire pile, which makes the water content relatively small overall. Similarly, porosity could also be found from the computerized model, which would also help in determining the accurate water content.

Further investigations into the structure of the pile formation would also improve our model. In particular, we would like to examine the role of different particle size ratios on pile formation. When the large particles are substantially larger than the small ones, we expect to find more void space. There will be more complex stability criteria that must be met for the larger block to come to rest. These new configurations will allow more opportunities for void space to be created under the large blocks. It would also be interesting to consider modelling stockpile formation with three different sizes of blocks. It would further be useful to model the pile structure that occurs when the stacker is moved to create two peaks in each cross-section. Modeling this interaction may help us to suggest improvements to pile creation methods that would optimize heterogeneity within the pile.
Appendix A - Definitions

**bulk volume** - Volume of entire pile of iron ore, including void space.

**hydraulic conductivity** - Denoted by \( K \). Indicates the ability of the material to conduct a fluid through it. Hydraulic conductivity depends upon the porosity of the material and the fluid which is draining.

**irreducible wetting water saturation** - Denoted by \( S_{w0} \). The percentage of water present in a substance when drainage is the only factor involved. For water in iron ore, \( S_{w0} = .16 \)

**gauge pressure** - Pressure above atmospheric pressure.

**pore-size distribution** - Denoted by \( \lambda \). The fraction of the total pore volume which contains pore diameters between \( \delta \) and \( \delta + d\delta \) where \( \delta \) is the diameter of the largest pore in a given region. The larger the disparity in pore sizes, the smaller the pore-size distribution. Often grain-size distribution is used to determine pore-size distribution.

**porosity** - Denoted by \( n \). The ratio of the volume of void space to bulk volume:

\[
n = \frac{\text{volume of void space}}{\text{volume of bulk}}
\]

**saturated zone** - Where the capillaries are completely filled with water.

**saturation** - Denoted by \( S \). The ratio of the volume of fluid to the volume of voids, taken over a small volume of bulk:

\[
S = \frac{\text{volume of fluids}}{\text{volume of voids}}
\]

**static pressure** - Denoted by \( p_e \). Pressure level required to force the fluid through the pore - negative quantity.

**steady flow** - Fluid velocity at any given point is constant in time.

**unsaturated zone** - Where the capillaries contain both water and gas.
void space - The portion of the pile not occupied by solid matter. This space includes water and/or air.

water content - (also moisture content) Denoted by \( \theta \) ratio of weight of water to weight of solids within a sample.

\[
\theta = \frac{\text{weight of water}}{\text{weight of solids}}
\]

water table - The zone inside the pile where the pressure is static.
References


A mathematical model for an exhaust oxygen sensor.

Patrick Worfolk, University of Minnesota
Tanya Leise, Texas A.&M. University
Brian von Dohlen, University of Michigan
Xinfu Chen, University of Pittsburgh
Saib Othman, University of Iowa
George Barrick, Kent State University
Donna Calhoun, University of Washington
Xianhua Huang, Delft University of Technology
Hector Lomeli, University of Colorado
Ruth Michler, University of North Texas
Daniel Baker (advisor), General Motors Corp.

Institute for Mathematics & its Applications

August 15, 1996

1 Introduction

By federal and state regulations the U.S. auto industry is forced to develop strategies to control emission whereas rising gas prices demand higher fuel economy. An important mechanism that is used in this process of control is the exhaust oxygen sensor.

An oxygen sensor is a chemical device that is constantly making a comparison between the level of oxygen inside the exhaust manifold and air from outside the engine.
The result of this comparison is detected with voltage that is generated at all times. The oxygen sensor is active any time the engine is hot enough, typically with a working temperature of 700°C.

![Figure 1: Schematic diagram of an oxygen Sensor](image)

During combustion the so-called fuel air ratio ($A/F$) ratio determines the different concentrations of the exhaust gases. The $A/F$ ratio is defined to be the ratio of the mass of air to the mass of fuel going into the engine. To good approximation, this ratio is independent of the engine speed. In the fuel-rich region there is not enough oxygen to burn off all the combustible gases. For instance, if not all the gases are burned, a high concentration of $CO$ (a dangerous pollutant) is generated and and fuel is wasted. On the other hand, in the fuel-lean region there is too much oxygen left over which has been shown to oxidize nitrogen at high temperatures (again a pollutant.) Hence it is important to keep the $A/F$ ratio at the optimal point, where there is exactly enough oxygen to burn off all the fuel, the so-called stoichiometric point.
In this paper we deal with a mathematical model for an oxygen sensor that is widely used in the automotive industry. A technical description follows. The sensor produces a voltage difference that essentially depends on the difference in oxygen concentration on the two different chambers of exhaust gases and air. A schematic diagram of the oxygen sensor is shown in figure 1.

The oxygen sensor uses a voltage drop to detect when the $A/F$ ratio changes from too rich to too lean. This information is used by the engine to adjust the air/fuel ratio. Experiments have shown that the sensor displays a voltage drop at a point slightly in the fuel lean region. See figure 2

![Diagram](image)

Figure 2: Voltage response of an oxygen Sensor

**Technical Description of the sensor**

The oxygen sensor consists of four different layers, two of which serve as electrodes. The important layers are: a solid electrolyte layer that helps in the creation of the voltage drop, two porous platinum layers that serve as catalysts of the reactions and as electrodes, and a porous spinel layer on one electrode that protects the device from particles in the
exhaust gases. The other electrode is directly exposed to air. The two platinum layers are in an open circuit configuration and connected to a device capable of measuring voltage. See figure 3.

![Porous Spinel Layer Diagram](image)

Figure 3: Different layers of an oxygen Sensor

Certain level of diffusion takes place in the porous spinel layer. It is possible to assume that the effect of the chemical reaction away of the platinum layer is negligible compared with the effects due to this diffusion and the reaction close to the platinum.

The important chemical reactions take place very close to the platinum layer in contact with the exhaust gases. On the other hand, the layers of the sensor are very thin compared to the height of the device. These two facts, will allow us to assume that locally the chemical reactions and diffusive transport don’t feel the effect of curvature and so we can use a one dimensional model.

2 Background for our model

2.1 Basic Chemistry

Although there are many gases in the exhaust the main components that influence voltage response are $N_2$, $O_2$, $CO$, $CO_2$, $H_2$ and $H_2O$. The outside air consists of roughly 21% oxygen
and 79% nitrogen, whereas the nitrogen concentration in the exhaust tends to be around 60-68% with less than 4% oxygen.

The exhaust gases react with each other according to the following schemes:

The platinum electrode is prone to develop so-called vacancies which we denote by the symbol $S$. Carbon Monoxide, hydrogen and oxygen molecules attach themselves to these vacancies and become adsorbed. We have the following adsorption reactions. (The reverse reaction is called a desorption reaction.)

\[
O_2 + 2S \rightleftharpoons 2O_{ad}
\]
\[
H_2 + 2S \rightleftharpoons 2H_{ad}
\]
\[
CO + S \rightleftharpoons CO_{ad}.
\]

At the platinum electrode the adsorbed carbon monoxide and hydrogen and the oxygen react to produce carbon dioxide and water in gaseous form.

\[
CO_{ad} + O_{ad} \rightleftharpoons CO_2 + 2S \uparrow
\]
\[
2H_{ad} + O_{ad} \rightleftharpoons H_2O + 3S \uparrow
\]

As a starting point, we will assume that the high air temperature in the exhaust pipes and the platinum coating (a catalyst) have ensured that all the above reactions are in equilibrium. Hence we can combine them into the two reactions:

\[
\frac{1}{2}O_2 + CO \rightleftharpoons CO_2
\]
\[
H_2 + \frac{1}{2}O_2 \rightleftharpoons H_2O.
\]

At equilibrium we have that the mole fractions of the gases obey:

\[
X_{O_2}^{1/2}X_{CO} = K_I P^{-1/2}X_{CO_2},
\]
\[
X_{O_2}^{1/2}X_{H_2} = K_{II} P^{-1/2}X_{H_2O},
\]
where $K_I, K_{II}$ are equilibrium constants and $P$ is the pressure. The electrochemical reaction:

$$O_{ad} = O^{2-} + S$$

is responsible for the voltage response. The adsorbed oxygen adsorbs two electron from the platinum electrode and enters the electrolyte. This reaction results in a positive electrochemical potential at the electrodes. The magnitude of that electrochemical potential depends the mole fraction. The greater the mole fraction of oxygen at the electrode the greater is the positive electrochemical potential. In what follows we denote the electrode in the exhaust air by electrode 1 and the (reference) electrode in fresh air by electrode 2. Hence there is a greater positive electrochemical potential at electrode 2 than at electrode 1. The difference in electrochemical potentials between electrode 2 and electrode 1 is the voltage measured. This difference is positive in our case. There is no current between the two electrodes. Since the oxygen concentration in the outside air is constant the voltage observed measures the mole fraction of oxygen in the exhaust gas. In our reaction 4 electrons per $O_2$ molecule are being exchanged. Hence in our setting a Nernst type equation computes the voltage $V$ by:

$$V := -(RT/4F)[\ln(\frac{[X_{O_2}]}{.21})].$$

Here $F = 96487$ Coulomb/mol is Faraday's constant. This description assumes that the number of vacancies $S$ in the platinum is large and the ratio of vacancies at electrode 1 and vacancies at electrode 2 is near 1. In fact the observed voltage will depend on the mole fraction of the adsorbed oxygen and hence is only equal to the mole fraction of the gaseous oxygen in equilibrium. The exact mechanics of that electrochemical process are still an area of research. It has been observed experimentally that the voltage jumps near the so-called stoichiometric point. This can be explained by the fact that the oxygen concentration is very close to zero (on the order of $10^{-20}$) in the fuel-rich region near the stoichiometric point. The potential difference is greater when the oxygen mole fraction at electrode 1 is smaller i.e. in the fuel-rich region than when there is more oxygen i.e. the fuel-lean region. The stoichiometric condition is for the exhaust gases is:

$$2X_{O_2} = X_{CO} + X_{H_2}.$$
We will assume that this relation is satisfied when \( \lambda = 1 \).

We use our model to explain why and where this voltage jump occurs. We also determine what influences the size and location of the jump and why the voltage jump does not occur at the stochiometric point. A good model will be of use in future designs of oxygen.

2.2 Description of the model

In our model we made the simplifying (starting) assumption that all chemical reactions are in equilibrium and that the exhaust gases are obeying the ideal gas law:

\[
pV = nRT,
\]

where \( R = 8.314 \text{Joule/mole.K} \) (universal Gas constant.) We assume that the (total) pressure of the systems stays constant and denote that constant by \( P \). We also assume that the temperature of the outside air is equal to the temperature of the exhaust and stays constant at temperature \( T = 973K \). Instead of partial pressures or partial concentrations we will use mole fractions of the 6 gases. The main advantage of mole fractions is that they are dimensionless and we have:

\[
X_{\text{H}_2} + X_{\text{O}_2} + X_{\text{H}_2\text{O}} + X_{\text{CO}} + X_{\text{CO}_2} + X_{\text{N}_2} = 1.
\]

1 mole is equal to the number of carbon atoms in 12g of \( C^{12} \) or:

\[
1 \text{mol} = 6.02205 \times 10^{23} \text{ atoms}.
\]

Let \( n_i \) denote the density of molecules of gas \( i \), and let \( X_i \) denote the mole fraction of the gas \( i \). Note that \( X_i := \frac{N_i}{C} \) where \( C \) is the total concentration, that is the total number of molecules per unit volume. We have \( 0 \leq X_{\text{O}_2}, X_{\text{CO}}, X_{\text{CO}_2}, X_{\text{H}_2}, X_{\text{N}_2}, X_{\text{H}_2\text{O}} \leq 1 \). We have that

\[
X_{\text{N}_2} = 1 - (X_{\text{O}_2} + X_{\text{CO}} + X_{\text{CO}_2} + X_{\text{H}_2\text{O}} + X_{\text{H}_2}).
\]

We decide to focus our attention on the mole fraction of oxygen. There two processes that we decided to be the main influences.
A. The diffusion across the porous layer: While the main function of the porous layer is to protect the platinum electrode it does "slow the exhaust gases down". Since the 6 different gases diffuse at slightly different speeds through the porous layer, the mole fraction of the gases at the electrode will be different from the initial mole fractions in the exhaust.

We will assume that the mole fractions \( X_{N_2}, X_{O_2}, CO, CO_2, X_{H_2}, X_{H_2O} \) at the surface of the electrode are in steady state hence there is no time dependence. The diffusion through the porous layer, also called a transport problem, is (under the assumption of steady state) described via the diffusion equation:

\[
\nabla^2 X_{O_2} = \nabla^2 X_{H_2} = \nabla^2 X_{CO} = \nabla^2 X_{CO_2} = \nabla^2 X_{H_2O} = 0.
\]

We will also use the Fick's Law which is valid near equilibrium to describe the flux \( NX \) of the various exhaust gases \( X \) through the porous layer:

\[
N_{O_2} = -\frac{CD_{O_2}}{\tau} \nabla X_{O_2},
\]

\[
N_{CO} = -\frac{CD_{CO}}{\tau} \nabla X_{CO},
\]

\[
N_{CO_2} = -\frac{CD_{CO_2}}{\tau} \nabla X_{CO_2},
\]

\[
N_{H_2} = -\frac{CD_{H_2}}{\tau} \nabla X_{H_2},
\]

\[
N_{H_2O} = -\frac{CD_{H_2O}}{\tau} \nabla X_{H_2O}.
\]

The diffusion coefficients \( D_i \) of the various gas components \( i \) depend on the physical layout of the molecules and are computed using statistical mechanics. Since all gases are dilute in the nitrogen, we will use diffusion coefficient of the various gases in nitrogen. To see whether this is a good assumption, we used the Stefan - Maxwell equations as well and found good agreement. The diffusion coefficients also depend on the physical properties of the porous layer, such as tortuosity \( \tau \) and width \( \ell \). In section 3 we will carefully study the dependence of these constants on the physical lay-out of the sensor. This gives design constraints.
B. The chemical reactions between the non nitrogen gases yield different equilibrium concentrations according to different values of \( \lambda \). Moreover the equilibrium constants depend on the pressure and temperature. We assume that the porous layer has sufficiently many holes so that the adsorption reactions of hydrogen and oxygen and carbon monoxide are independent of each other and there is no competition for vacant sites. E. Logothesis [2] claims that the rather quick diffusion of the hydrogen in relation to the oxygen leads to the shift of the voltage jump into the fuel-lean region. In order to identify the main reason for the deviation we propose to look at a simplified system involving only \( \text{O}_2, \text{CO}, \text{CO}_2 \).

3 A simplified model

As an example of how the exhaust sensor works, we suppose that the exhaust contains only four species of gas molecules: \( \text{O}_2 \), \( \text{CO} \), \( \text{CO}_2 \) and \( \text{N}_2 \). It is assumed that the nitrogen is inert. Very close to the electrode, platinum catalyzes the reaction

\[
\text{CO} + \frac{1}{2} \text{O}_2 \rightleftharpoons \text{CO}_2.
\]

Far enough away from the electrode, at the exhaust, we assume that the gas mixture remains fixed. Specifically, it's taken that the mole fraction of each species depends on the fuel to air ratio supplied to the engine. This ratio is parameterized by \( \lambda \), a quantity selected so that stoichiometry (a situation where, for this model, oxygen concentration is half that of carbon monoxide) occurs at \( \lambda = 1 \). Data points for the mole fractions of each chemical species far from the plate are taken from [2] with an offset to satisfy our simplified assumption.

For simplicity, a one-dimensional cross section of the sensor is considered, with \( x = 0 \) at the left edge of the porous layer and \( x = \ell \) at the platinum electrode. The width \( \ell \) of the porous ceramic layer is assumed to be much greater than the width of the zone in which platinum catalyzes reactions.

The partial differential equations describing the various gas concentrations in the
porous ceramic layer of the device are

\[
\frac{\partial (cX_i)}{\partial t} = - \frac{\partial N_i}{\partial x}
\]

where \( X_i [\cdot] \) is the molar fraction of total gas concentration for species \( i \), \( N_i [\cdot/cm^2s] \) is molar flux and \( c [mol/cm^3] \) is the total concentration of all gas species. Fick's law of chemical diffusion states that in situations where the diffusing chemical species are essentially unaffected by one another we have the flux \( N_i \) proportional to the gradient for the molar concentration:

\[
N_i = - \frac{D_i}{\tau} \cdot \frac{\partial (cX_i)}{\partial x}
\]

where \( D_i [cm^2/s] \) is chemical diffusivity of gas \( i \) and \( \tau [\cdot] \) is tortuosity of the porous ceramic layer. This gives a system of diffusion equations

\[
\frac{\partial X_i}{\partial t} = \frac{\partial}{\partial x} \left( \frac{D_i}{\tau} \cdot \frac{\partial X_i}{\partial x} \right).
\]

We seek a steady state solution for the differential equations, and because we are working with a one dimensional model, it is assumed that \( N_i \) is constant for each species. Since we assume that \( c \) is constant, the differential equations become

\[
\frac{\partial X_i}{\partial x} = \text{constant}.
\]  

(3.1)

The concentrations for the chemical species at \( x = 0 \) are denoted \( X_{CO_2}^{\text{exh}} \), \( X_{CO}^{\text{exh}} \), and \( X_{O_2}^{\text{exh}} [mol/cm^3] \) and are obtained as known functions of \( \lambda [1] \).

At the platinum electrode, we make the steady state assumption that the rate at which \( CO_2 \), \( CO \) and \( O_2 \) are being created or destroyed must be balanced by the incoming fluxes of these species. That is,

\[
\tau \left( K_i \sqrt{\rho} X_{CO} X_{O_2}^{\frac{1}{2}} - X_{CO_2} \right) = \frac{c D_{CO_2}}{\tau} \cdot \frac{\partial X_{CO_2}}{\partial x}
\]

\[
\tau \left( K_i \sqrt{\rho} X_{CO} X_{O_2}^{\frac{1}{2}} - X_{CO_2} \right) = - \frac{c D_{CO}}{\tau} \cdot \frac{\partial X_{CO}}{\partial x}
\]  

(3.2)

\[
\frac{\tau}{2} \left( K_i \sqrt{\rho} X_{CO} X_{O_2}^{\frac{1}{2}} - X_{CO_2} \right) = - \frac{c D_{O_2}}{\tau} \cdot \frac{\partial X_{O_2}}{\partial x}.
\]
To obtain tractable boundary conditions for the equation describing the diffusion of species through the porous layer, first note that (3.2) leads to

\[
D_{\text{CO}_2} \frac{\partial X_{\text{CO}_2}}{\partial x} = -D_{\text{CO}} \frac{\partial X_{\text{CO}}}{\partial x}
\]

\[
D_{\text{CO}_2} \frac{\partial X_{\text{CO}_2}}{\partial x} = -2D_{\text{O}_2} \frac{\partial X_{\text{O}_2}}{\partial x} .
\]

Now choosing one of the equations from (3.2) and dividing through by \( r \), obtain

\[
(K_I \sqrt{p} X_{\text{CO}} X_{\text{O}_2}^{\frac{1}{2}} - X_{\text{CO}_2}) = \frac{cD_{\text{CO}_2}}{r \tau} \frac{\partial X_{\text{CO}_2}}{\partial x} .
\]

(3.4)

If we make the substitution \( x = \ell \xi \) we get the non-dimensional group of parameters

\[
\frac{cD_{\text{CO}_2}}{r \tau \ell} .
\]

The right hand side of (3.4) will be negligible if either the diffusivity is small relative to the other parameters, or if \( \ell \) or \( \tau \) are large relative to other parameters. Physically, this means that the reaction at the platinum electrode will be in equilibrium if the diffusion of species to the plate is very slow relative to the rate at which the species are being created or destroyed. Similarly, if the tortuosity \( \tau \) is large, molecules have a difficult time making it through the porous layer, and if the thickness \( \ell \) of the porous layer is large, the flux through the porous layer is low and species will be diffusing into the electrode slowly.

For the purposes of our model, we will simply assume that the parameter group

\[
\frac{cD_{\text{CO}_2}}{r \tau \ell} \ll 1 .
\]

From the discussion above, the equilibrium condition,

\[
K_I \sqrt{p} X_{\text{CO}} X_{\text{O}_2}^{\frac{1}{2}} - X_{\text{CO}_2} = 0 ,
\]

or

\[
\frac{X_{\text{CO}_2}}{X_{\text{CO}} X_{\text{O}_2}^{\frac{1}{2}}} = K_I \sqrt{p}
\]

(3.5)

holds. This, in conjunction with (3.3) gives us the three equations to solve for \( X_{\text{O}_2} \), \( X_{\text{CO}} \) and \( X_{\text{CO}_2} \) at the right edge of our domain. All that remains is some non-dimensionalization and algebraic simplifications.
First, given that the derivatives of the molar fractions are constant, we know that all the $X_i$ are linear functions in $x$, and so we can write (3.3) in terms of just the exhaust and electrode values of the $X_i$

\[
\frac{D_{CO_2}}{l} (X_{CO_2} - X_{CO_2}^{exh}) = -\frac{D_{CO}}{l} (X_{CO} - X_{CO}^{exh})
\]

\[
\frac{D_{CO_2}}{l} (X_{CO_2} - X_{CO_2}^{exh}) = -2\frac{D_{O_2}}{l} (X_{O_2} - X_{O_2}^{exh}) .
\]

(3.6)

We can also introduce a non-dimensionalizaton for the diffusivities by dividing each diffusion coefficient by a nominal diffusivity, $D_0$:

\[
D_i = \frac{D_i}{D_0} .
\]

$D_0$ may be chosen to be one of the diffusion coefficients in the system.

It is known that $K \gg 1$ (in fact $K \sim 10^{10}$) so we introduce

\[
\epsilon = \frac{1}{K \sqrt{p}} .
\]

The molar fractions are non-dimensional and the system of equations can be written in non-dimensional form

\[
X_{CO} X_{O_2}^{1/2} = \epsilon X_{CO_2} .
\]

\[
D_{CO_2} (X_{CO_2} - X_{CO_2}^{exh}) = -D_{CO} (X_{CO} - X_{CO}^{exh})
\]

\[
D_{CO_2} (X_{CO_2} - X_{CO_2}^{exh}) = -2D_{O_2} (X_{O_2} - X_{O_2}^{exh}) .
\]

(3.7)

4 Analysis of the simple model

Given the steady-state equations (3.7) describing diffusion in the porous layer and equilibrium at the platinum plate for the simplified Fick's law model, we derive an asymptotic expansion in terms of the small parameter $\epsilon$. Two different expansions arise, corresponding to the relative abundances of oxygen and carbon monoxide. If the exhaust contains excess carbon monoxide relative to oxygen, then the chemical reactions at the platinum
excess carbon monoxide relative to oxygen, then the chemical reactions at the platinum will nearly deplete the oxygen. This situation corresponds to a rich fuel to air ratio, i.e., $\lambda < 1$. In this case the leading order terms of the asymptotic expansion are

$$X_{\text{CO}} \sim X_{\text{CO}}^{\text{rich}} \equiv X_{\text{CO}}^{\text{exh}} - \frac{2D_{\text{O}_2}}{D_{\text{CO}}} X_{\text{O}_2}^{\text{exh}}$$

$$X_{\text{CO}_2} \sim X_{\text{CO}_2}^{\text{rich}} \equiv X_{\text{CO}_2}^{\text{exh}} + \frac{2D_{\text{O}_2}}{D_{\text{CO}_2}} X_{\text{O}_2}^{\text{exh}}$$

$$X_{\text{O}_2} \sim X_{\text{O}_2}^{\text{rich}} \equiv \varepsilon^2 \left( \frac{X_{\text{CO}_2}}{X_{\text{CO}}} \right)^2.$$

The other possible solution describes the situation where the exhaust contains abundant oxygen relative to carbon monoxide, so that the carbon monoxide is depleted while some oxygen remains. This corresponds to a lean fuel to air ratio, or $\lambda > 1$. The leading order terms of the asymptotic expansion are

$$X_{\text{O}_2} \sim X_{\text{O}_2}^{\text{lean}} \equiv X_{\text{O}_2}^{\text{exh}} - \frac{D_{\text{CO}}}{2D_{\text{O}_2}} X_{\text{CO}}^{\text{exh}}$$

$$X_{\text{CO}_2} \sim X_{\text{CO}_2}^{\text{lean}} \equiv X_{\text{CO}_2}^{\text{exh}} + \frac{D_{\text{CO}}}{D_{\text{CO}_2}} X_{\text{CO}}^{\text{exh}}$$

$$X_{\text{CO}} \sim X_{\text{CO}}^{\text{lean}} \equiv \varepsilon \frac{X_{\text{CO}_2}}{\sqrt{X_{\text{O}_2}}}.$$

These asymptotic expressions agree very well with numerical solutions (error of order $10^{-6}$) away from the steep intermediate region. Compare Figure 4 (mole fractions of gases in exhaust) with Figures 5 and 6 (mole fractions of gases near the platinum plate).
Figure 4: Comparison of mole fractions of $O_2$ and $CO$ in the exhaust.

We define the switching point as the point separating the regions where the different asymptotic expansions are valid. This switching point occurs when $\lambda$ satisfies the equation

$$D_{O_2}X_{O_2}^{exh} = \frac{1}{2} D_{CO} X_{CO}^{exh}.$$  

Notice the similarity of this equation to the equation for stochiometry

$$X_{O_2}^{exh} = \frac{1}{2} X_{CO}^{exh}.$$  

While stochiometry occurs at $\lambda = 1$, the switching point between depleted oxygen and excess oxygen is shifted by diffusion to $\lambda = 0.999433$. The relative magnitudes of the diffusion constants $D_{O_2}$ and $D_{CO}$ affects the position of the switching point. If $D_{O_2}$ were increased, this shift would be pushed farther to the left (recall that $\lambda < 1$ corresponds to a rich fuel to air ratio). If $D_{CO}$ were increased, the shift would be pushed to the right. Note that for this simple model the switching point lies to the left of stochiometry, while in the actual data it lies to the right.
Figure 5: Mole fraction of O\textsubscript{2} for the simplified Fick's law model as \( \lambda \) varies.

The existence of the two different asymptotic expansions explains another behavior of the oxygen sensor: two regions of fairly constant voltage separated by a region of very rapidly changing voltage. The voltage for the region \( \lambda < 0.999433 \) is approximately \( V(X_{\text{rich}}^{\text{rich}}) \approx 0.9 \), while the voltage for the opposite region is roughly \( V(X_{\text{lean}}^{\text{lean}}) \approx 0.05 \). The steepness of the intermediate region is due to the very small order of \( \varepsilon \) (recall \( \varepsilon \approx 10^{-10} \)). Since the reactions at the platinum plate are in equilibrium, either nearly all of the oxygen or of the carbon monoxide will be depleted, depending on their relative concentrations in the exhaust. The mole fraction of oxygen is relatively low for \( \lambda < 1 \) since the fuel to air ratio is rich, while the mole fraction for carbon monoxide is large. Hence at the platinum plate the oxygen will be depleted but the excess carbon monoxide will remain present. For \( \lambda > 1 \) exactly the opposite situation happens. See Figure 7.
Figure 6: Mole fraction of CO for the simplified Fick's law model as \( \lambda \) varies.

Figure 7: Voltage across the oxygen sensor for the simplified Fick's law model.
Figure 8: Initial concentrations of exhaust gases as a function of $\lambda$.

5 A more general model

We now add hydrogen and water to our system. This forces us to introduce a second reaction to our model, with equilibrium constant $K_{II}$. The initial concentration of all the gas species is shown in figure 8.

The following steady-state equations describe diffusion in the porous layer and equilibrium at the platinum plate for this six-species Fick's law model:

\[
\begin{align*}
D_{CO_2}(X_{CO_2} - X_{CO_2}^{exh}) &= -D_{CO}(X_{CO} - X_{CO}^{exh}) \\
D_{H_2O}(X_{H_2O} - X_{H_2O}^{exh}) &= -2D_{H_2}(X_{H_2} - X_{H_2}^{exh})
\end{align*}
\]
\[ D_{O_2}(X_{O_2} - X_{O_2}^{\text{exh}}) = \frac{1}{2} D_{CO}(X_{CO} - X_{CO}^{\text{exh}}) + \frac{1}{2} D_{H_2}(X_{H_2} - X_{H_2}^{\text{exh}}) \]
\[ X_{CO} \sqrt{X_{O_2}} = \varepsilon X_{CO_2} \]
\[ X_{H_2} \sqrt{X_{O_2}} = a \varepsilon X_{H_2O} \]

(where \( a = \frac{K_{I^L}}{K_{II}} = 1.4668 \)).

Using an asymptotic expansion in terms of the small parameter \( \varepsilon = \frac{1}{K_{III}} \), we find that two solutions exist, corresponding to the relative abundance of oxygen and carbon monoxide. If carbon monoxide and hydrogen are in excess, then the chemical reactions at the platinum electrode will nearly deplete the oxygen. This situation corresponds to a rich fuel to air ratio, i.e., \( \lambda < 1 \). In this case the leading order term of the asymptotic expansion of oxygen is:

\[ X_{O_2} \sim X_{O_2}^{\text{rich}} \equiv \zeta^2 \varepsilon^2 \]

where \( \zeta \) is obtained by solving:

\[ D_{CO} \left( \frac{D_{CO} X_{CO}^{\text{exh}} + D_{CO} X_{CO}^{\text{exh}}}{D_{CO} + D_{CO2} \zeta} \right) + D_{H_2} \left( \frac{D_{H_2} X_{H_2}^{\text{exh}} + D_{H_2O} X_{H_2O}^{\text{exh}}}{D_{H_2} + D_{H_2O} a \zeta} \right) \]

\[ = 2 \left( \frac{D_{CO} X_{CO}^{\text{exh}}}{2} + \frac{1}{2} D_{H_2} X_{H_2}^{\text{exh}} - D_{O_2} X_{O_2}^{\text{exh}} \right) \]

The other possible solution describes the situation where the exhaust contains relatively abundant oxygen relative to carbon monoxide and hydrogen, so that these gases are depleted while some oxygen remains. This corresponds to a lean fuel to air ratio, or \( \lambda > 1 \). In this case, the leading order term of the asymptotic expansion of oxygen is:

\[ X_{O_2} \sim X_{O_2}^{\text{lean}} \equiv X_{O_2}^{\text{exh}} - \frac{1}{2} D_{CO} X_{CO}^{\text{exh}} - \frac{1}{2} D_{H_2} X_{H_2}^{\text{exh}} \]

These asymptotic expressions give voltages (shown in figure 10) that agree very well with numerical solutions.
Figure 9: Asymptotic $O_2$ concentration as a function of $\lambda$, for the more general model.

The switching point between these two solutions occurs when $\lambda$ satisfies the equation

$$D_{O_2}X_{O_2}^{exh} = \frac{1}{2}D_{CO}X_{CO}^{exh} + \frac{1}{2}D_{H_2}X_{H_2}^{exh}$$

Notice the similarity of this equation to the equation for stochiometry

$$X_{O_2}^{exh} = \frac{1}{2}X_{CO}^{exh} + \frac{1}{2}X_{H_2}^{exh}.$$  

While stochiometry occurs at $\lambda = 1$, the switching point between depleted oxygen and excess oxygen is shifted by the diffusion constants to $\lambda = 1.01797$. In figure 9 we see the asymptotic $O_2$ concentration. Note the switch at $\lambda \approx 1.02$.

The relative magnitudes of the diffusion constants $D_{O_2}$, $D_{H_2}$ and $D_{CO}$ affects the
Figure 10: Asymptotic voltage as a function of $\lambda$, for the more general model.

position of the switching point. If $D_{O_2}$ were increased, the above shift would be pushed to the left, i.e., a "rich" shift (recall that $\lambda < 1$ corresponds to a rich fuel to air ratio). If $D_{CO}$ or $D_{H_2}$ were increased, the shift would be pushed to the right, i.e., a "lean" shift. Note that for this model, as opposed to the simple model, there is a lean shift, as is seen in the experimental data (shown as dots in figure 10) [1].

The existence of the two different asymptotic expansions explains another behavior of the oxygen sensor: two regions of fairly constant voltage separated by a region of very rapidly changing voltage. (See figure 10)

The voltage for the region $\lambda < 1.01797$ is approximately $V(X_{O_2}^{\text{rich}}) \approx 0.9$, while the voltage for the opposite region is roughly $V(X_{O_2}^{\text{lean}}) \approx 0.05$ (the values of lean and
rich oxygen formulas depend on \( \lambda \), so the voltage varies somewhat as \( \lambda \) does). Since the reactions at the platinum plate are in equilibrium, almost all of the oxygen will be depleted, or both the carbon monoxide and the hydrogen will be depleted, depending on their relative concentrations in the exhaust. The mole fraction of oxygen is relatively low for \( \lambda < 1 \) since the air to fuel ratio is rich, while the mole fraction for carbon monoxide and hydrogen is large. Hence at the platinum plate the oxygen will be depleted but the excess carbon monoxide and hydrogen will remain. For \( \lambda > 1 \), the opposite occurs.

6 A new transport law

6.1 The Stefan-Maxwell equations

In the previous models, we used Fick’s law to relate the flux to the gradient of the concentration of the species. This law unfortunately introduces physical inconsistencies into the model. One manifestation of this arises when integrating the differential equations from Fick’s law and finding that the sum of the mole fractions does not always equal one. The Stefan-Maxwell equations provide a better physical model for transport (i.e., diffusion) of gas species which remedies this problem [3, p. 570]. These equations account for the fact that gases diffuse at different rates depending on the presence of other gases and are given by:

\[
\nabla X_i = \sum_{j=1}^{n} \frac{1}{c D_{ij}} (X_i N_j - X_j N_i),
\]

(6.8)

where the \( X_i \) are the mole fractions of the species, \( N_i \) are the fluxes, \( c \) is the total mole concentration, \( D_{ij} \) is the diffusivity of the pair \( i - j \) in a binary mixture, and \( \nabla \) is the spatial gradient. (In our current case, \( \nabla = \frac{\partial}{\partial \xi} \).

We list a few properties of the Stefan-Maxwell equations:

1. In the appropriate limit, the Stefan-Maxwell equations become Fick’s law. Assume a binary gas mixture where \( X_1 = O(\epsilon) \) and \( X_2 = 1 - X_1 \). Also assume that the averaged mass flow is zero, i.e., \( \sum_i m_i X_i = 0 \). We write \( X_1 = \epsilon \tilde{X}_1 \) where \( \tilde{X}_1 = O(1) \),
and the Stefan-Maxwell equations give
\[ \epsilon \partial X_i / \partial \xi = \frac{1}{c D_{12}}(\epsilon \tilde{X}_1 N_2 - (1 - \epsilon \tilde{X}_1) N_1). \]

Taking the limit of the above equation as \( \epsilon \to 0 \) we find that \( N_1 = O(\epsilon) \). Using this, the averaged mass flow equations tells us that \( N_2 = O(\epsilon) \). So writing \( N_1 = \epsilon \tilde{N}_1 \) and \( N_2 = \epsilon \tilde{N}_2 \) we get
\[ \partial \tilde{X}_1 / \partial \xi = \frac{1}{c D_{12}}(\epsilon \tilde{X}_1 \tilde{N}_2 - (1 - \epsilon \tilde{X}_1) \tilde{N}_1). \]

Taking the limit as \( \epsilon \to 0 \) we find
\[ \partial \tilde{X}_1 / \partial \xi = \tilde{N}_1 / (c D_{12}) \]

which (after multiplying by \( \epsilon \)) is Fick’s law for the diffusion of \( X_1 \).

2. Since \( D_{ij} = D_{ji} \) for all \( i \) and \( j \), adding all the equations in (6.8) yields
\[ \nabla \sum_{i=1}^{n} X_i = \sum_{i,j=1}^{n} \frac{1}{c D_{ij}} (X_i N_j - N_j X_i) = 0. \tag{6.9} \]

It follows that \( \sum_{i=1}^{n} X_i \) is a constant ( = 1); that is, solutions to the Stefan-Maxwell equations have the property that the total mole concentration is space invariant, in agreement with our assumption that the temperature and pressure are space invariant.

3. If we write the Stefan-Maxwell equation as \( \nabla \bar{X} = A(\bar{X}) \tilde{N} \) where \( \bar{X} = (X_1, \ldots, X_n)^T \) and \( \tilde{N} = (N_1, \ldots, N_n)^T \) are two vectors, and \( A(\bar{X}) \) is an \( n \times n \) matrix with entries depending on \( \bar{X} \). From (6.9), one sees that \( A \) is degenerate. Also, since all \( D_{ij} \) are positive and all \( X_i \) are non-negative with \( \sum_{i=1}^{n} X_i = 1 > 0 \), one can show that \( A \) has rank \( n - 1 \). Hence, we can solve for \( \tilde{N} \) from the equation \( \nabla \bar{X} = A \tilde{N} \). (assuming \( \sum_{i=1}^{n} X_i = 1 \)). There are infinitely many solutions for \( \tilde{N} \). The general solution can be written as
\[ \tilde{N} = C \bar{X} + \tilde{N}^*(\bar{X}, \nabla \bar{X}), \tag{6.10} \]

where \( \tilde{N}^* \) is any special solution and \( C \) is an arbitrary constant. To relate the flux uniquely with the fractional molar concentration of species, one needs an additional condition, besides the Stefan-Maxwell equation. Hence, we can enforce the
conservation of mass condition:

$$\sum_{i=1}^{n} m_i N_i = 0,$$

(6.11)

where $m_i$ is the molar weight of a molecule of gas species $i$.

6.2 Applications: Simple case

We now develop a model for the oxygen sensor which is based on the Stefan-Maxwell equations instead of Fick’s law. For simplicity, we present the model where the only reactive agents are CO, O$_2$, and CO$_2$. The model equations are as follows

- Stefan-Maxwell:

$$\frac{\partial X_{CO}}{\partial x} = \frac{\tau}{cD_{CO,CO}} (X_{CO} N_{O_2} - X_{O_2} N_{CO}) + \frac{\tau}{cD_{CO,CO_2}} (X_{CO} N_{CO_2} - X_{CO_2} N_{CO})$$

$$- \frac{\tau}{cD_{CO,N_2}} X_{N_2} N_{CO},$$

$$\frac{\partial X_{O_2}}{\partial x} = \frac{\tau}{cD_{O_2,CO}} (X_{O_2} N_{CO} - X_{CO} N_{O_2}) + \frac{\tau}{cD_{O_2,CO_2}} (X_{O_2} N_{CO_2} - X_{CO_2} N_{O_2})$$

$$- \frac{\tau}{cD_{O_2,N_2}} X_{N_2} N_{O_2},$$

$$\frac{\partial X_{CO_2}}{\partial x} = \frac{\tau}{cD_{CO_2,CO}} (X_{CO_2} N_{CO} - X_{CO} N_{CO_2}) + \frac{\tau}{cD_{CO_2,O_2}} (X_{CO_2} N_{O_2} - X_{O_2} N_{CO_2})$$

$$- \frac{\tau}{cD_{CO_2,N_2}} X_{N_2} N_{CO_2},$$

$$\frac{\partial X_{N_2}}{\partial x} = \frac{\tau}{cD_{N_2,CO}} X_{N_2} N_{CO} + \frac{\tau}{cD_{N_2,O_2}} X_{N_2} N_{O_2} + \frac{\tau}{cD_{N_2,CO_2}} X_{N_2} N_{CO_2}.$$

- The equilibrium equation at the electrode:

$$\frac{X_{CO_2}}{X_{CO} X_{O_2}^{1/2}} = K_f \sqrt{P} \text{ at } x = \ell.$$

- Mass conservation at the electrode:

$$\begin{cases} N_{CO} = -N_{CO_2} \\ N_{CO} = 2N_{O_2} \\ N_{N_2} = 0 \end{cases} \text{ at } x = \ell.$$
There are eight unknown quantities in these equations: four mole fractions $X_i$ and four fluxes $N_i$. One way to think about this model is as a two point boundary value problem for a set of eight ordinary differential equations. Since we are considering steady state in a one-dimensional diffusion problem, the differential equations which describe the evolution of the $N_i$ are trivially given by $\partial N_i / \partial x = 0$. Now we need eight boundary conditions. Four are given by specifying the composition of the gas in the exhaust corresponding to values of $X_i$ at $x = 0$. The other four are given at the electrode ($x = \ell$) and are the equilibrium equation and mass conservation equations. In the differential equations given above we have already incorporated the knowledge that $N_{N_2} \equiv 0$ to eliminate some terms.

We can make the above equations dimensionless by using the transformations

$$K_I \sqrt{P} = 1 / \epsilon,$$

$$x = \ell \xi,$$

$$N_i = \frac{cD_i}{\ell \tau} \tilde{N}_i,$$

$$D_{ij} = D D_{ij}.$$

The quantity $1 / \epsilon$ was introduced in order to indicate the expected large value of $K_I \sqrt{P}$ and $D$ is a reference diffusivity which we will choose to be $D_{O_2,N_2}$. Thus, in dimensionless form and dropping the tildes on the fluxes, the model equations are

- Stefan-Maxwell:

$$\begin{align*}
\partial X_{CO}/\partial \xi &= \frac{1}{D_{CO,O_2}}(X_{CO}N_{O_2} - X_{O_2}N_{CO}) + \frac{1}{D_{CO,CO_2}}(X_{CO}N_{CO_2} - X_{CO_2}N_{CO}) \\
&\quad - \frac{1}{D_{CO,N_2}}X_{N_2}N_{CO}, \\
\partial X_{O_2}/\partial \xi &= \frac{1}{D_{O_2,CO}}(X_{O_2}N_{CO} - X_{CO}N_{O_2}) + \frac{1}{D_{O_2,CO_2}}(X_{O_2}N_{CO_2} - X_{CO_2}N_{O_2}) \\
&\quad - \frac{1}{D_{O_2,N_2}}X_{N_2}N_{O_2}, \\
\partial X_{CO_2}/\partial \xi &= \frac{1}{D_{CO_2,CO}}(X_{CO_2}N_{CO} - X_{CO}N_{CO_2}) + \frac{1}{D_{CO_2,O_2}}(X_{CO_2}N_{O_2} - X_{O_2}N_{CO_2}) \\
&\quad - \frac{1}{D_{CO_2,N_2}}X_{N_2}N_{CO_2},
\end{align*}$$

24
\[ \frac{\partial X_{N_2}}{\partial \xi} = \frac{1}{D_{N_2,CO}} X_{N_2} N_{CO} + \frac{1}{D_{N_2,O_2}} X_{N_2} N_{O_2} + \frac{1}{D_{N_2,CO_2}} X_{N_2} N_{CO_2} . \]

- The equilibrium equation at the electrode:
  \[ \frac{X_{CO_2}}{X_{CO} X_{O_2}^{1/2}} = \frac{1}{\varepsilon} \text{ at } \xi = 1 . \]

- Mass conservation at the electrode:
  \[ \begin{align*}
  N_{CO} &= -N_{CO_2} \\
  N_{CO} &= 2N_{O_2} \\
  N_{N_2} &= 0
  \end{align*} \]
  \[ \text{at } \xi = 1 . \]

Note that the only parameters that remain in the problem are \( \varepsilon \) and the diffusivity ratios \( D_{ij} \). Once again, the thickness of the porous layer and its tortuosity do not directly impact the solution, but we have implicitly assumed when writing these equations that the diffusion is slow enough so that equilibrium of the gases is attained at the platinum electrode.

6.3 Application: The general case.

A. Problem Setting.

Now we go back to our sensor problem. We use 1, 2, 3, 4, 5, 6 to denote, in the order given, the species \( O_2, CO, CO_2, H_2O, H_2, N_2 \). We use \( p \) and \( q \) to denote the flux of \( CO \) and \( H_2 \), respectively. Then we have, by the discussion in the previous sections,

\[ \begin{align*}
  N_1 &= N_{O_2} = \frac{1}{2}(p + q), \\
  N_2 &= N_{CO} = p, \\
  N_3 &= N_{CO_2} = -p, \\
  N_4 &= N_{H_2O} = -q, \\
  N_5 &= N_{H_2} = q, \\
  N_6 &= N_{N_2} = 0.
  \end{align*} \]  

(6.12)
Here we remark that the last equation \( N_6 = 0 \) is consistent with the equation (6.11). In fact, we can calculate
\[
\sum_{i=1}^{6} m_i N_i = p(m_{CO} + \frac{1}{2} m_{O_2} - m_{CO_2}) + q(m_{H_2} + \frac{1}{2} m_{O_2} - m_{H_2O}) + m_{N_2} N_6
\]
\[= m_{N_2} N_6,
\]
by using the fact that the molar weights satisfy \( m_{CO} + \frac{1}{2} m_{O_2} = m_{CO_2} \) and \( m_{H_2} + \frac{1}{2} m_{O_2} = m_{H_2O} \). Thus, the condition \( N_6 = 0 \) is equivalent to (6.11).

Now using \( X_6 = 1 - \sum_{i=1}^{5} X_i \) to eliminate \( X_6 \), we can write the Stefan-Maxwell equations in the following (non-dimensional) form:
\[
x'(\xi) = b(p, q) + G(p, q)x, \quad \xi \in (0, 1), \tag{6.13}
\]
where \( x = (X_1, X_2, X_3, X_4, X_5)^T \), \( b = (b_1(p, q), b_2(p), b_3(p), b_4(q), b_5(q))^T \) is a known linear vector function of \( p \) and \( q \), and \( G \) is a known \( 5 \times 5 \) matrix with entries which are linear functions of \( p \) and \( q \).

**Remark.** In the Fick's law formulation, we took the diffusion coefficients \( D_i \) as \( D_{i,6} \), the diffusion coefficients of species \( i \) in \( N_6 \). It is then an easy exercise to check that with the notation of (6.13), Fick’s law gives the equation
\[
x'(\xi) = b(p, q). \tag{6.14}
\]
Consequently, if the magnitude of \( x \) is small, Fick’s law (6.14) and the Stefan-Maxwell law (6.13) are not very different.

With the initial condition
\[x(0) = x_0\]
the solution of (6.13) can be expressed as
\[
\phi(\xi) = \exp[G(p, q)\xi] x_0 + G^{-1}[p, q]\left(\exp[G(p, q)\xi] - I\right)b(p, q)
\]
where \( I \) is the identity \( 5 \times 5 \) matrix.
Finally, using the equilibrium conditions \( \frac{X_{CO}}{X_{O_2}} = \varepsilon X_{CO_2} \) and \( \frac{X_{H_2O}}{X_{O_2}} = a \varepsilon X_{H_2O} \), we obtain the algebraic equations for \( p, q \):

\[
\begin{align*}
    x_2(p, q)\sqrt{x_1(p, q)} - \varepsilon x_3(p, q) &= 0, \\
    ax_2(p, q)x_4(p, q) - x_5(p, q)x_3(p, q) &= 0.
\end{align*}
\] (6.15)

where \( x_i(p, q) \) are given by

\[
(x_1, x_2, x_3, x_4, x_5)^T = \exp[G(p, q)\varepsilon]x_0 + G^{-1}\left[\exp[G(p, q)\varepsilon] - I\right]b(p, q).
\]

B. Numerical Technique.

There are two ways to solve (6.15), a system of two unknowns with two algebraic equations.

The first way is to use Mathematica by calling the command

\[
\text{NSolve[ \{equation1, equation2\}, \{p, q\}]}
\]

In such an application, one does not even need to know how the Mathematica program deals with such a complicated system.

The second way to solve (6.15) is an iteration method briefly described as follows:

Step 1. Pick the solution from Fick’s law as our initial guess; we call it \((p^{old}, q^{old})\).

Step 2. Calculate \( x(p, q, \varepsilon) \) with \((p, q) = (p^{old}, q^{old})\).

Step 3. Observe that in leading order, for any small \( \delta p, \delta q \),

\[
x(p^{old} + \delta, q^{old} + \delta q, \varepsilon) \sim x(p^{old}, q^{old}, \varepsilon) + [b_p \delta p + b_q \delta q] \varepsilon
\]

where \( b_p := \frac{\partial x}{\partial p} \) and \( b_q := \frac{\partial x}{\partial q} \) are constant vectors. Hence, assuming that \( p = p^{old} + \delta p \) and \( q = q^{old} + \delta q \) we can solve for \( \delta p \) and \( \delta q \) from the system

\[
\begin{align*}
    \left(x_2(p^{old}, q^{old}) + b_p^2 \delta p\right)\sqrt{x_1(p^{old}, q^{old}) + b_p^1 \delta p + b_p^3} &= \varepsilon \left(x_3(p^{old}, q^{old}) + b_p^3 \right), \\
a\left(x_2(p^{old}, q^{old}) + b^2 \delta p\right)\left(x_4(p^{old}, q^{old}) + b^4 \delta q\right) &= \left(x_5(p^{old}, q^{old}) + b^5 \delta q\right)\left(x_3(p^{old}, q^{old}) + b^3 \delta p\right).
\end{align*}
\]
Step 4. Update \((p^{\text{old}}, q^{\text{old}})\) by \((p^{\text{old}} + \delta p, q^{\text{old}} + \delta q)\) and goto step 1 or exit the process if \(\delta p\) and \(\delta q\) are small enough.

C. Numerical Result.

We implemented the second scheme mentioned. The rate of convergence is extremely fast. In three to four iterations, \(\delta p\) and \(\delta q\) will be smaller than \(10^{-7}\). The solution is depicted in the following figures.

Figure 11: Comparision of Fick’s law (solid curve) and Stefan-Maxwell’s law (dots).
7 Conclusions

In sections 3, 4 and 5, models using Fick’s law are given and the results are presented; in section 6, a model implementing the Stephan-Maxwell law is discussed. In section 5, where the solution of the simple model is presented, an important part of the results is figure 10. From this graph, we note that the outcome of our asymptotic analysis differs slightly from Logothetis’ [2] experimental results. In fact, our tutor advised us that Logothetis’ results are not accurate on the lean side, but they are reasonable elsewhere. On the other hand, our results seem reasonable on the lean side when comparing them to experimental data from other sensors. So we are left to explain the inaccuracy in our results on the rich side and at the stoichiometric point $\lambda = 1$.

As a whole, our investigations provide a good understanding how the oxygen sensor works. We do observe the jump in the graph of voltage versus $\lambda$ as expected. We’ve also learned what tools can be used to improve the model for future work.

Fick’s law is used to establish the relation between the flux and the concentration gradients for the various chemical species. The accuracy of this law increases as the concentration of nitrogen becomes greater relative to the concentrations of other gases. Unfortunately the application of Fick’s law does not preserve the total mole fractions. The Stephan-Maxwell law does preserve the mole fractions of the reactant species as well as maintaining other basic assumptions of electrochemical reactions. However, our computational results show little difference between the two types of model. Thus it seems that the changes in mole fractions of the reactants as computed when using Fick’s law are indeed negligible. What’s more, our asymptotic analysis shows that computational precision of both models is sufficient for comparison with the real experimental data. Hence the inaccuracy in the models arises from other factors.

The equilibrium assumption states that a certain ratio of the concentrations of reactants to the concentration of product is a known constant (3.5). In fact this constant can be found in the literature. To say this in less technical words, recall that in the sensor the gases first diffuse through the porous layer before reaching the platinum where the reaction happens. In the equilibrium assumption we take it that the reaction happens
very rapidly relative to the gas diffusion. Thus the concentrations of gas species at the platinum electrode depend on the porosity of the ceramic layer and not on the flux of gas species spawned by the reactions occurring at the electrode.

References


Modeling UV-Damage to E. Coli Bacteria

Adair Aguiar
University of Minnesota

David Bainbridge
University of Michigan

Mary Bond
University of Kentucky

Chaitan Gupta
University of Nevada

Karen Matthews
University of Illinois at Chicago

Julie Plante
University of Montreal

Esa Rantanen
Pensylvania State University

Christian Turner
Texas A & M University

Daoqi Yang
Wayne State University

Wei Zhang
Michigan State University

Tutor Donald Drew
Rensselaer Polytechnic Institute

August 15, 1996
1 Introduction

Food poisoning is a very serious condition that may affect large populations and will often result in multiple deaths. Of the several microbes responsible for widespread food poisoning epidemics, the most common are strains of *E. Coli*, *Salmonella*, and *Botulinum*. Even in the industrialized world, where there exist high standards of hygiene, outbreaks of these bacteria occur frequently. The most recent example is the outbreak of *E. Coli* in Japan this August, which resulted in the death of a 10-year old schoolgirl and various ailments for over 6,000 other community members (Associate Press, 1996). The handling, storage, and distribution of food products in the modern world offers ample opportunities and time for the above mentioned microbes to grow and infest the foods. Consequently the sterilization of the food products prior to transportation and storage is imperative. The microbes can be effectively destroyed by heat, however, there are many fresh food products, e.g., fresh vegetables, fruit, and meat, that are heat-sensitive. These foods must therefore be treated by alternative methods.

Ultraviolet (UV) light is known to have bactericidal properties. Sunlight, due to its content of UV rays, has an important role in the sterilization that occurs under natural conditions ([1]). Due to these properties, UV light radiation has found applications in control of bacteria in indoor air, water supplies, and many heat-sensitive food products, e.g., meat.

1.1 Statement of the Problem

A small company involved in meat packing uses UV radiation to sterilize the meat. The company has experimented with pulsed UV radiation versus continuous UV radiation, measuring the effectiveness of the treatment as the number of bacteria killed at varying amounts of radiation. An experimental procedure to validate the effectiveness of the UV light system involves irradiation of *E. Coli* culture in an aqueous solution on a petri dish by UV light. The preliminary results indicate that the bactericidal effect of pulsed UV radiation is superior to continuous, i.e., pulsed radiation achieved a higher kill ratio at lower power than continuous radiation, up to a point, after which increase in the amount of radiation resulted in a higher kill ratio using continuous radiation rather than pulsed. This phenomenon is illustrated in Fig. 1. Experimental results have indeed shown that the crossing of the graphs (continuous and pulsed) occurs ([10]).

The purpose of this study was to develop a mathematical model that
would explain the phenomenon described above and predict the bactericidal effect of UV radiation given the various physical characteristics of the radiation equipment and sterilization processes. The modeling efforts were restricted to experimental conditions, i.e., irradiation of E. Coli in an aqueous solution in a petri dish.
2 Effect of UV Radiation on Organisms

Radiation affects molecules of bacteria through absorption, which promotes photochemical reactions. UV light is absorbed by purines and pyrimidines, which are major components of the deoxiribonucleic acid (DNA). As they absorb photons from the light source, the molecules receive energy in discrete units called photons, one photon at a time. The energy of a photon is inversely proportional to its wavelength. The total energy a molecule absorbs is proportional to the product of the duration and intensity of radiation as well as the absorption coefficient of the material being irradiated ([1]). Radiation, i.e., absorption of photons, elevates the energy level of the molecule, which may use the energy for chemical changes, lose it as heat or fluorescence, or, if the photon has sufficient energy, the atoms of the molecule may lose an electron and become ionized. Although high-intensity UV light can ionize atoms ([9]; [6]), low-intensity UV light does not have sufficient energy to be ionizing ([1]). Since each photon has the same energy, we interpret this to mean that ionization requires the absorption of more than one photon. The likelihood of absorbing a second photon, before the energy the molecule has received by being "hit" for the first time dissipates and the molecule returns to a lower energy level, increases with radiation intensity.

The absorption rate of the nucleic acid and the bactericidal effect of UV light are functions of wavelength, which achieve their maxima at the same wavelength. The absorption rate of the nucleic acid is independent of the nature of the bacteria (Koller, 1965). The absorption maximum of DNA occurs at a wavelength of about 260 nm which subsequently causes maximum damage to the DNA structure.

While the absorption of UV light is the primary underlying mechanism of its bactericidal properties, it also poses some serious limitations to its use. Because UV light is readily absorbed, it does not penetrate into solids and penetrates into liquids only slightly ([1]). This fact is aggravated by the typically low energy of UV radiation. Another factor that limits the sterilizing effectiveness of UV radiation is the ability of the irradiated bacteria to repair the damage inflicted on their DNA strands.

2.1 Heat

There is widespread consensus in the microbiological literature that the primary mechanism behind the bactericidal properties of UV light is damage
to the DNA of the bacteria. However, there was some anecdotal evidence that the very high heat (3000 °C, as claimed) resulting from absorption of photons from high-intensity, pulsed, UV radiation would contribute to the destruction and demise of the bacteria. To rule out this possibility and to allow models to be developed on a sound biological basis, the temperature rise of a single *E. Coli* cell, surrounded by an aqueous medium (i.e., water) when hit by a 200 ms pulse of UV laser was computed. The resulting temperature rise was only about 0.01 °C. If the bacterium is completely surrounded by air, however, the temperature rise under similar radiation is 154.54°C, which is sufficient to obliterate it. If the bacterium is assumed to be on the surface of a solid, e.g., meat, the rise in its temperature is about 60 °C. It is further shown that while a very powerful UV-laser light heat exposure may kill bacteria on the surface of a medium, it is ineffective at any depth. However, since the modeling effort was concentrated on explaining the data obtained from experimental studies, heat as a mechanism of devastation was given no further consideration, (see Appendix for computations).

### 2.2 Damage to the DNA

As organisms grow by division of individual cells, the genetic information is transmitted from a generation of cells to the next by splitting of the DNA ([5]). DNA consists of two very long chains of molecules twisted together. Molecules are grouped in four major building blocks, or bases, identified as T (Thymine), C (Cytosine), G (Guanine) and A (Adenine) to encode information. A's on one strand must face T's on the other strand, and Cs must face Gs. When a cell divides, the two strands are separated, both containing all the information needed to build a new cell. A second strand is then completed following the pairing rule above. This redundancy allows cells to correct defects in their DNA and is essential to the process of cell division.

If UV light hits a DNA chain, it will most likely be absorbed by purine and pyrimidine molecules present in the chain. This can interfere with the normal base pairing of DNA either as altered bases or single-strand breaks. Strand breaks require large amounts of energy, however, and double-strand breaks would require large amounts of ionizing energy ([7]). The most likely consequence of UV radiation is the formation of cyclobutane pyrimidine dimers ([9]) which will interfere with the reproduction of the cell.

In the simplified approach adopted here only two major damage types found in dead irradiated cells are considered: formation of dimers and strand
breaks. Pyrimidine dimer can be understood as a chemical change on a pair of the DNA base pairs that confuses the division process and thus stops the reproduction of the cell. When the cell loses its ability to reproduce, it is functionally dead. The amount of nutrients the cell needs to survive is proportional to its volume, but the amount of nutrients it can take in is proportional to its surface area. If the cell cannot reproduce, it continues to grow until it can no longer feed itself and dies. The formation of a Pyrimidine dimer requires the energy of one photon. If the damage is done when the cell has just finished a reproduction cycle, it most likely will repair itself before the next splitting occurs. However, if the damage is inflicted on a cell that has just started splitting or doubling the DNA, it will not be able to complete the cycle and will subsequently die.

A strand break represents a more severe damage to DNA. A chemical bond in one of the two strands is broken. A certain amount of energy is required to break this bond. The energy of one photon is not enough to do this, but the energy of two photons on the same site is sufficient. We include in this damage mechanism crosslinks between two adjacent pairs and denatured sites since those are also two photon reactions and are much more rare than single-strand breaks. After intercepting a photon, the strand is in a higher energy state. If the strand loses this extra energy before another photon hits the strand site again, the second photon does not destroy the DNA. Instead, it puts the strand back into the excited state, waiting for another photon to destroy it.

2.3 Repair Mechanisms

The *E. Coli* bacteria has several strategies for repair of various types of damage. The repair mechanism for the most likely damage resulting from exposure to UV light, the formation of cyclobutane pyrimidine dimers, is a chemical reaction with DNA photolyase. Repair mechanisms can be as fast as picoseconds (1 picosecond = 10^{-12} s) after a photon hits a base or as slow as 2 seconds to repair a severe damage to a strand ([7]). These are complex processes that are not fully understood, so we were unable to incorporate fine details of the repair mechanism into the model. Instead, we assume that a return to a lower energy level after being hit by a photon allows the bacteria to repair the damage to the DNA. Repairs of strand breaks are assumed not to occur.
3 Model Development

The model simulates the effect of UV light on a solution of bacteria in a petri dish. The solution is assumed to be spread in a layer 2 mm thick. The irradiation is done in two ways, (1) by a continuous light or (2) by a pulsed light over the petri dish. The objective of the model is to explain the differences between the two techniques as well as the differences in bactericidal effectiveness.

The data to which the models were compared came from experiments made by [9]. The pulsed irradiation was done with a laser giving a 15 ns pulse light 5 times per second at 0.42 J/m² per pulse. The continuous irradiation was done with a UV lamp at 1 J/m² per second of exposure. Both were using a wavelength very near the maximal absorption spectrum of DNA of 260nm.

3.1 Definition of Variables

Intensity is effectively the energy density of the beam of light. When light is shined on a surface, the intensity is the rate at which energy hits a unit of area. Its unit is J/m²/s. In the continuous irradiation, the intensity is constant in time while in the pulsed irradiation the intensity is a step function: high intensity during the flash of the light and no intensity when the light is off. In order to compare the killing rates, the total number of photons received per second has to be the same in both irradiations. The type of damage inflicted on the DNA is dependent on the intensity.

A photon is a light particle. The energy of one photon is equal to $hc/\lambda$, where $h$ is Planck’s constant ($6.63 \times 10^{-34}$ J s), $c$ is the speed of light ($3.00 \times 10^8$ m/s), and $\lambda$ is the wavelength of the light. Considering a particle having a certain wavelength may be confusing. This is because particles and waves are both models for the behavior of light, but neither is sufficient to explain it entirely. The idea of light as an electro-magnetic wave is necessary to explain the interference phenomenon of light ([2]). The model of light as a particle is necessary to explain Planck’s radiation law ([3]) and the photoelectric effect ([2]).

If we use UV light with a wavelength of 260 nm, the energy of one photon is $7.65 \times 10^{-19}$ J/photon. Therefore, if the intensity of the light is $I$, the rate at which photons hit a unit area as $I/7.65 \times 10^{-19} = 1.31 \times 10^{18} \times I$. 
4 Two-Hit Model

In the simplest model, we consider only one kind of damage to the DNA, assuming it is a two-photon reaction. Moreover, it is assumed that the second photon could be absorbed anywhere on the DNA to be effective in killing the cell. This means that the probability of the first hit and the probability of the second hit are the same. We also allow a cell to repair damage done by a single photon at a given rate.

After a solution of bacteria has been irradiated, a number of healthy cells, i.e., cells with no damage, remain. Their concentration is denoted by \( N(t) \). This quantity varies as a function of time of exposure. The time of exposure times the intensity is the total energy per area received by the solution of bacteria. The concentration of cells with one hit in the irradiated solution is denoted by \( S(t) \). We call them sick cells. In order to kill sick cells, the cells have to receive a second photon. The concentration of dead cells, i.e., cells that received two photons during the irradiation, at time \( T \) is computed as:

\[
N(0) - N(T) - S(T).
\]

The mathematical description of those concentrations in time is done by two ordinary differential equations. To derive them, consider the change of concentration for the two families.

The concentration of healthy cells changes because both healthy cells are hit by photons and become sick cells and sick cells repair themselves and become healthy cells. The concentration of sick cells changes in the following three ways: sick cells are hit by a second photon and die, sick cells repair themselves and become healthy cells, and healthy cells are hit by a photon and become sick cells. We then have the following system:

\[
\begin{bmatrix}
\frac{dN}{dt} \\
\frac{dS}{dt}
\end{bmatrix} =
\begin{bmatrix}
-C_1 & C_2 \\
C_1 & -(C_1 + C_2)
\end{bmatrix}
\begin{bmatrix}
N \\
S
\end{bmatrix}
\]

(2)

where \( C_1 \) is the rate at which cells received a photon and \( C_2 \) is the rate at which sick cells repair themselves.

If \( C_1 \) is a constant in time (which is the case for continuous irradiation), this system is easily solved:
4 TWO-HIT MODEL

\[
\begin{bmatrix}
N(t) \\
S(t)
\end{bmatrix} = e^{At} \begin{bmatrix}
N(0) \\
S(0)
\end{bmatrix},
\]

where

\[
A = \begin{bmatrix}
-C_1 & C_2 \\
C_1 & -(C_1 + C_2)
\end{bmatrix}.
\]

We assumed we start with all healthy cells, the initial condition of the system. If \(C_1\) is a step function in time, being \(C_1\) for the pulse duration, \(\tau_p = 15 \times 10^{-9}\) s, and zero for the rest of the pulse, \(\tau_r = 0.2\) s, we compute the solution to this system in two steps, one with the UV-laser on, and one with the UV-laser off. We obtain the following solution for the effect of \(K\) pulses:

\[
\begin{bmatrix}
N(K) \\
S(K)
\end{bmatrix} = (e^{A\tau_p} e^{B\tau_r})^K \begin{bmatrix}
N(0) \\
S(0)
\end{bmatrix},
\]

where \(B\) represents the matrix of the system when the light of the UV-laser is turned off, i.e.,

\[
B = \begin{bmatrix}
0 & C_2 \\
0 & -C_2
\end{bmatrix}.
\]

4.1 Analysis of the Model

The simplest model only incorporated one type of damage to the DNA. We assumed that two photons hitting at various locations on a DNA strand was sufficient to prevent the cell from reproducing. We observed from mathematical analysis that the continuous light model and the pulsed light model produced identical graphs corresponding to the fraction of bacteria killed as a function of the rate at which the bacteria was hit by photons. This conflicted with the experimental data graph which exhibits two separate, intersecting curves (Fig. 1). Since the graph resulting from the numerical analysis should exhibit the same intersection which is present in the graphs of the experimental data, we attempted to refine our mathematical model. We considered an N-hit model, i.e., a model for which N photons are required to prevent reproduction. For various values of N, the graphs again
Figure 2: *One-Hit Model: Continuous versus Pulse UV Light.*

showed that the continuous and the pulse cases were identical. Upon further research, we discovered two errors with this model. First, we found that in order to effect damage on the DNA, the two photons must hit at the same location on a single strand. Second, it is not sufficient to consider only strand damage.
Figure 3: Two-Hit Model: Continuous versus Pulse UV Light.

Figure 4: Twenty-Hit Model: Continuous versus Pulse UV Light.
5 The Two-Mechanism Model

In the second model, we consider the two different mechanisms of damaging the DNA. We have four kinds of cells in a solution of irradiated bacteria: (1) Healthy cells, their concentration noted \( N(t) \), (2) cells with one hit on a strand, their concentration noted \( S_2(t) \), (3) cells that had one hit on a base site, their concentration noted \( S_1(t) \), and (4) cells with one hit on a strand and one hit on a base, their concentration noted \( S_{12}(t) \).

In order to kill cells having one hit on a base, the cells have to go under splitting process. In order to kill cells having one hit on a strand, the cells have to receive a second photon on the site already hit. The killed cell concentration at time \( T \) is computed as:

\[
N(0) - N(T) = S_1(T) - S_2(T) - S_{12}(T).
\]

Following the concentration of the different kinds of cells will gives us:

\[
\begin{bmatrix}
\frac{dN}{dt} \\
\frac{dS_1}{dt} \\
\frac{dS_2}{dt} \\
\frac{dS_{12}}{dt}
\end{bmatrix}
= \mathbf{A}
\begin{bmatrix}
N \\
S_1 \\
S_2 \\
S_{12}
\end{bmatrix},
\]

where

\[
\mathbf{A} =
\begin{bmatrix}
-(C_1 + C_2) & C_{hF} & C_{hS} & 0 \\
C_1 & -(C_e + C_2 + C_{hF}) & 0 & C_{hS} \\
C_2 & 0 & -(C_3 + C_1 + C_{hS}) & C_{hF} \\
0 & -C_2 & +C_1 & -(C_{hF} + C_{hS} + C_e + C_3)
\end{bmatrix}.
\]

We also have that \( C_1 \) is the rate at which healthy cells receive a photon on a base, \( C_2 \) is the rate at which healthy cells receive a photon on a strand, \( C_3 \) is the rate at which \( S_2 \) cells receive a second lethal photon (those rates are proportional to the intensity), \( C_e \) is the rate at which \( S_1 \) cells divide, \( C_{hF} \) is the rate at which \( S_1 \) cells repair dimers, and \( C_{hS} \) is the rate at which
5 THE TWO-MECHANISM MODEL

$S_2$ cells repair damaged strands. For this model we set $C_{hS} = 0$ since it is almost impossible for the cell to return to a healthy state where there is no excitation of the strands. Both the continuous and the pulsed solutions of this system have the same form of, respectively, the solutions (3, 5), where, in the present case, the matrix $B$ is given by

$$
B = \begin{bmatrix}
0 & C_{hF} & C_{hS} & 0 \\
0 & -(C_c + C_{hF}) & 0 & C_{hS} \\
0 & 0 & -C_{hS} & C_{hF} \\
0 & 0 & 0 & -(C_{hF} + C_{hS} + C_c)
\end{bmatrix}.
$$

(10)

5.1 Analysis of the Model

For the second model, we refined two of the assumptions made in the basic model. For the first refinement, we assumed that two photons must hit at the same location on a DNA strand to prevent the cell from reproducing. Second, we considered damage to the building blocks of the DNA. Unfortunately, our second model also failed to produce two separate, intersecting curves as observed in the experimental data graphs. Once again, the two models exhibited identical behavior.

Assuming that repair mechanisms cannot handle a very large number of damages at the same time, we included a maximum number of repairs that could take place simultaneously in our pulsed UV light, two photon model. This was implemented using the following assumptions. First, all damages were given equal probability for occurring. Second, we assumed that there is no repair taking place while the light is on, a period of only 15 ns per pulse. A final modification was made concerning the healing rate of the bacterial cells. If the number of cells with one hit was less than a chosen fixed value, then the number of repairs was proportional to the number of sick cells, as in previous models. However, when the number of cells which have absorbed one photon exceeds the chosen fixed value, there is a constant maximal number of repairs which may take place. We did not find a significant difference with this modification and the experimental numerical behavior was not observed in this model.
Figure 5: Base Molecule and Strand-Break Model: Continuous versus Pulse UV Light.
6 A Nonlinear Model

Since the biochemical mechanisms affecting the repair of damaged DNA is limited, we assumed that the rate at which cells could heal themselves was limited. An examination of a similar chemical mechanism ([11]) shows that the repair rate could be modeled as $C_2/(1 + kS_1)S_1$, where $k$ is a constant. This form shows a linear behavior (as assumed in the Two-Mechanism model) at low concentrations of $S_1$. We implemented this model, but further studies need to be done because the results were not significantly different from the results of the previous models.

7 Summary and Conclusions

To study the bactericidal properties of ultraviolet light, we considered the effect of heat from the ultraviolet laser on the bacteria and three models of DNA damage due to UV photons. From our models, we calculated that heat was not a factor in destroying the bacteria. The first of the DNA damage models was the two-hit model, where it takes two photon hits on the DNA to destroy it, and the DNA can repair damage inflicted by one photon. Our second model incorporated the effects of two types of DNA damage inflicted by the photons. The third model took into account the idea that the healing processes in the first model have a maximum rate. None of these were able to predict the intersection of the graphs for continuous and pulsed UV radiation seen in the experimental data.

The two-hit model involves a system of two first-order linear ODEs. We were able to solve this analytically and found that no matter how we varied the intensity or the frequency of the pulses, the graphs of the number of kills for pulsed and continuous radiation could never intersect. We changed this to an N-hit model, where it takes N photon hits to break the DNA strand and solved it numerically for $N = 1, 2, \text{ and } 20$. This change still did not result in crossing of the curves.

The second model we used involves two mechanisms for DNA damage. Photons could break the DNA base pairs with one hit, or they could break a DNA strand with two hits. Each type of damage had its own repair mechanism. The base pair could be healed by a chemical reaction within the cell, while a site which previously absorbed a photon could drop out of the higher energy state before being hit a second time, effectively "healing" the damage caused by the first hit. In this model, the graphs still did not
intersect.

The third model was a variation on the first. In the previous models, the rate at which the DNA healed itself was fixed, but it is possible that the healing rate slows down when many base pairs are damaged due to limitations in the healing mechanisms. After incorporating this possible effect into the model, we plotted the solutions for various values of the UV intensity and pulse time, but the results were no different than those for the first model.

There are other effects that one might consider in further modeling how ultraviolet light kills bacteria. For example, we didn't consider "shadowing", which is when bacteria on the top of the petri dish block the rays from hitting bacteria on the bottom. There is a significant drop in the intensity of the UV light even in a 1 mm thick layer where the \emph{E. coli} occupies 1 of $4.544 \times 10^{-5}$ cm for the radius of an \emph{E. coli} bacterium, we found that the intensity at the bottom is about $10^{-7}$ of the original intensity. Thus, the high intensity pulses should be much more effective in killing the bacteria at the bottom than the continuous beam. It may be that this could account for the intersection of the two graphs, but we were unable to incorporate this into our model.

Another consideration is the distribution of DNA strand sites in the higher energy states after being hit by one photon. If the DNA strand intercepts one photon, that part of the molecule has a higher energy. The strand breaks when a second photon hits a part of the strand in the high energy state, providing enough energy to break the molecule. The probability of a cell's DNA being broken by the second photon depends on how many pieces of the strand are excited, i.e., if only a few strand sites have high energy, the probability that a second photon hits one of those sites is very low. If most of the strand sites are excited, the probability of being broken is very high. In our second model, the probability of getting the second hit was constant. We tried to get some idea of the distribution of excited DNA strand sites, but failed to get any answers analytically or numerically. A model which takes this distribution into account would certainly be more accurate.
References


Appendix

We study the heating effects of irradiating food with a ultraviolet light laser in order to see if it is a meaningful sterilization method. We make the following assumptions:

1. We consider a 200 ms pulse of an ultraviolet laser which emits an energy density of 3.8 $J/cm^2$.
2. We shall assume that the bacterium and the medium it is in closely resemble water in regard to their physical characteristics.

Physical Parameters and variables of the model.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$</td>
<td>0.99821 $g/(cm)^3$</td>
<td>density of water</td>
</tr>
<tr>
<td>$D_1$</td>
<td>$598.4 \ mW/K \ m$</td>
<td>thermal conductivity of water</td>
</tr>
<tr>
<td></td>
<td>$59840 \ cm \ g \ s^{-3} /K$</td>
<td></td>
</tr>
<tr>
<td>$D_2$</td>
<td>$2.7882 \times 10^{-5} \ J/m \ s \ K$</td>
<td>thermal conductivity of air</td>
</tr>
<tr>
<td></td>
<td>$2.7882 \ cm \ g \ s^{-3} /K$</td>
<td></td>
</tr>
<tr>
<td>$c$</td>
<td>$4.1818 \ J/g \ K$</td>
<td>specific heat of water</td>
</tr>
<tr>
<td></td>
<td>$4.1818 \ (cm)^2 \ s^{-3}/K$</td>
<td></td>
</tr>
<tr>
<td>Vol. of</td>
<td>$3.93 \times 10^{-13} \ (cm)^3$</td>
<td>assumed spherical</td>
</tr>
<tr>
<td>a bacterium</td>
<td>$4.544 \times 10^{-5} \ cm$</td>
<td>radius of a bacterium</td>
</tr>
<tr>
<td>$q$</td>
<td>$J/cm^3 \ s \ K$</td>
<td>energy per unit volume of UV-laser</td>
</tr>
<tr>
<td>$T$</td>
<td>= temperature [K]</td>
<td></td>
</tr>
</tbody>
</table>

MODEL 1

We first model how hot a bacterium in a vacuum gets when exposed to a celsius in vacuum when exposed to ultraviolet laser. Let $T(t)$ denote the temperature of the bacterium at time $t$-seconds. We first need to convert the energy density of the laser to energy density per unit volume. Now a laser emitting an energy density of 3.8 $J/cm^2$ is equal to a laser emitting $15000 \rho c \ J/cm^3 \ s \ K$. Since the heat conductivity in a vacuum is zero, $T(t)$ satisfies the differential equation

$$\rho c \frac{dT}{dt} = 15000 \rho c,$$
$T(0) = 0$,

where the temperature $T$ is measured in Celcius and time $t$ is measured in seconds. Solving the above differential equation we obtain

$T = 15000t$.

Since the duration of the pulse is 200 milliseconds, it follows that the temperature of the bacterium reaches $T(.2) = 3000^\circ C$.

**MODEL 2**

We shall denote, again, by $T$ the temperature at a point in the medium in the following.

Next, we model the heat effect of exposing a bacterium in an infinite medium of water by a 200 milli-second pulse of our laser by the radially-symmetric steady state heat-conduction equation as follows:

$$D_1 \left( \frac{d^2 T}{dr^2} + \frac{2}{r} \frac{dT}{dr} \right) + q = 0, \quad r < r_b,$$

$$\frac{d^2 T}{dr^2} + \frac{2}{r} \frac{dT}{dr} = 0, \quad r > r_b,$$

$$\frac{dT}{dr}(0) = 0,$$

$$T(\infty) = 0,$$

$$T(r_b-) = T(r_b+),$$

$$(D_1 \frac{dT}{dr})(r_b-) = (D_1 \frac{dT}{dr})(r_b+).$$

Now, for $r > r_b$ we have

$$\frac{d^2 T}{dr^2} + \frac{2}{r} \frac{dT}{dr} = 0,$$

$$\frac{d}{dr} (r^2 \frac{dT}{dr}) = 0,$$

$$r^2 \frac{dT}{dr} = A \text{ (a constant)},$$

$$T = -\frac{A}{r} + A_1,$$

$$T(\infty) = 0 \Rightarrow A_1 = 0,$$

$$T = -\frac{A}{r}.$$
Next, for $r < r_b$ we have

\[ D_1 \left( \frac{d^2T}{dr^2} + \frac{2}{r} \frac{dT}{dr} \right) + q = 0, \]

\[ D_1 \frac{d}{dr} \left( r^2 \frac{dT}{dr} \right) + qr^2 = 0, \]

\[ D_1 \frac{d^2T}{dr^2} + \frac{1}{3} qr^3 = B \text{ (a constant),} \]

\[ \frac{dT}{dr}(0) = 0 \implies B = 0, \]

\[ D_1 \frac{dT}{dr} + \frac{1}{3} qr = 0, \]

\[ D_1 \frac{d}{dr} \left( r^2 \frac{dT}{dr} \right) + qr^2 = B \text{ (a constant).} \]

We now apply the continuity conditions when $r = r_b$ to obtain

\[ \frac{1}{D_1} (B - \frac{1}{6} qr_b^2) = -\frac{A}{r_b}, \]

\[ D_1 \frac{A}{r_b^2} = -\frac{q}{3} r_b, \]

\[ \frac{1}{D_1} (B - \frac{1}{6} qr_b^2) = \frac{q}{3D_1} r_b^2, \]

\[ B = \frac{q}{2} r_b^2. \]

We thus obtain the temperature at the center of the bacterium to be

\[ T(0) = 0.0108^\circ C. \]

**Conclusion:** The temperature of the bacterium barely rises. The explanation is that the heat is conducted away by the surrounding medium. Note that we have assumed that the bacterium and the surrounding medium have the same thermal characteristics.

If we assume that the bacterium is surrounded by air then we need to solve the analogous model given by

\[ D_1 \left( \frac{d^2T}{dr^2} + \frac{2}{r} \frac{dT}{dr} \right) + q = 0, \quad r < r_b, \]

\[ \frac{d^2T}{dr^2} + \frac{2}{r} \frac{dT}{dr} = 0, \quad r > r_b, \]

\[ \frac{dT}{dr}(0) = 0, \]
REFERENCES

\[ T(\infty) = 0, \]
\[ T(r_s) = T(r_s^+), \]
\[ (D_1 \frac{dT}{dr})(r_s^-) = (D_2 \frac{dT}{dr})(r_s^+). \]

Solving this model, as done above, we obtain the temperature at the center of the bacterium to be

\[ T(0) = 154.54^\circ C. \]

MODEL 3

It is known among food scientists that UV-laser light radiation as a method of sterilization for food is effective only when the bacterium is present on the surface of the food and the light source is very near to the food, something on the order of 6 inches.

We shall develop a mathematical model which confirms this.

We model this problem as a one dimensional heat diffusion problem, with the depth into the medium as the only meaningful space variable. All points in a plane parallel to a horizontal plane being assumed to be at the same temperature.

We take \( q_1 \), the energy of the laser to be 3.8 \( J/cm^2 \). Let \( T(t, y) \) denote the temperature of the food at time \( t \) and at a depth of \( y \). We measure time in seconds and depth in centimeters. The differential equations of our model are:

\[ \rho c \frac{\partial T}{\partial t} = D_1 \frac{\partial^2 T}{\partial y^2}, \]
\[ T(0, y) = 0, \ T(t, \infty) = 0, \ y \geq 0, \ t \geq 0, \]
\[ -D_1 \frac{\partial T}{\partial y}(t, 0) = q_1, \ t \geq 0. \]

We note that the units of \( q_1 \) are \( J/cm^2 \ s \ K \). We set

\[ D = \frac{D_1}{\rho c}, \]
\[ \bar{q} = \frac{q_1}{\rho c}. \]
REFERENCES

We see that the values of $D$, $\tilde{q}$ are given by

$$D = \frac{D_1}{\rho c} = \frac{59840}{4.1818 \times .9981 \times 10^7} = .001433686748,$$

$$\tilde{q} = \frac{q_1}{\rho c} = \frac{3.8}{\rho c} = .910429418938.$$

In terms of $D$ and $\tilde{q}$ the differential equations of the model become

$$\frac{\partial T}{\partial t} = D \frac{\partial^2 T}{\partial y^2},$$

$$T(0, y) = 0, T(t, \infty) = 0, y \geq 0, t \geq 0,$$

$$-D \frac{\partial T}{\partial y}(t, 0) = \tilde{q}, t \geq 0.$$ 

Now, $T(t, y) = \sqrt{t} f\left(\frac{y}{\sqrt{D}t}\right)$ is a solution of the above equations, where $f(x)$ is a function of one variable $x$ and satisfies the following differential equation

$$f''(x) + \frac{1}{2} f'(x) - \frac{1}{2} f(x) = 0,$$

$$f(\infty) = 0,$$

$$f'(0) = -\frac{\tilde{q}}{D}.$$ 

It is easy to see (using Maple) that

$$f(x) = C_1 x + C_2 \left[ e^{-\frac{x^2}{4}} + \frac{1}{2} \sqrt{\pi} \text{erf}\left(\frac{1}{2} x\right) \right]$$

is the general solution of the above differential equations. Using the boundary conditions, we obtain that

$$C_1 = -\frac{\tilde{q}}{\sqrt{D}},$$

$$C_2 = \frac{2}{\sqrt{\pi} \sqrt{D}} 0.032.$$ 

Thus,

$$T(t, y) = \sqrt{t} f\left(\frac{y}{\sqrt{D}t}\right),$$

$$= \frac{\tilde{q}}{\sqrt{D}} \sqrt{t} \left[ -\frac{y}{\sqrt{D}t} + \frac{2}{\sqrt{\pi}} \left( e^{-\frac{x^2}{4}} + \frac{1}{2} \sqrt{\pi} \sqrt{D} \text{erf}\left(\frac{y}{2\sqrt{D}t}\right) \right) \right].$$
REFERENCES

It follows that the temperature at the center of the bacterium which is at a depth of \( r_b \) after 200 millisecond of UV-laser exposure is given by

\[
T(2, r_b) = \frac{\tilde{q}}{\sqrt{D}} \left[ -\frac{r_b}{\sqrt{D}} + \frac{2\sqrt{\pi}}{\sqrt{D}} e^{-\frac{r_b^2}{4D}} + \frac{1}{2} \sqrt{\pi} \frac{r_b}{\sqrt{2D}} \text{erf}(\frac{r_b}{2\sqrt{2D}}) \right], \\
= \frac{\tilde{q}}{\sqrt{D}} \left[ -\frac{r_b}{\sqrt{D}} + \frac{2\sqrt{\pi}}{\sqrt{D}} e^{-\frac{r_b^2}{4D}} + \frac{r_b}{\sqrt{D}} \text{erf}(\frac{r_b}{2\sqrt{2D}}) \right], \\
= 12^\circ C \text{ approximately.}
\]

Conclusion: Since we assume that the food is at room temperature of 20\(^\circ\)C the UV-laser exposure of 200 ms will cause the food temperature to rise about 32\(^\circ\)C. It can be seen from our work that if we use a laser which emits an energy density of 19J/cm\(^2\), a laser five times more powerful, then the temperature of the bacterium will rise to approximately 160\(^\circ\)F, an FDA standard.
<table>
<thead>
<tr>
<th>#</th>
<th>Author/s</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1339</td>
<td>R. Lipton</td>
<td>Reciprocal relations, bounds and size effects for composites with highly conducting interface</td>
</tr>
<tr>
<td>1340</td>
<td>H.A. Levine &amp; J. Serrin</td>
<td>A global nonexistence theorem for quasilinear evolution equations with dissipation</td>
</tr>
<tr>
<td>1341</td>
<td>A. Boutet de Monvel &amp; R. Purice</td>
<td>The conjugate operator method: Application to DIRAC operators and to stratified media</td>
</tr>
<tr>
<td>1342</td>
<td>G. Michele Graf</td>
<td>Stability of matter through an electrostatic inequality</td>
</tr>
<tr>
<td>1343</td>
<td>G. Avalos</td>
<td>Sharp regularity estimates for solutions of the wave equation and their traces with prescribed Neumann data</td>
</tr>
<tr>
<td>1344</td>
<td>G. Avalos</td>
<td>The exponential stability of a coupled hyperbolic/parabolic system arising in structural acoustics</td>
</tr>
<tr>
<td>1345</td>
<td>G. Avalos &amp; I. Lasiecka</td>
<td>A differential Riccati equation for the active control of a problem in structural acoustics</td>
</tr>
<tr>
<td>1346</td>
<td>G. Avalos</td>
<td>Well-posedness for a coupled hyperbolic/parabolic system seen in structural acoustics</td>
</tr>
<tr>
<td>1347</td>
<td>G. Avalos &amp; I. Lasiecka</td>
<td>The strong stability of a semigroup arising from a coupled hyperbolic/parabolic system</td>
</tr>
<tr>
<td>1348</td>
<td>A.V. Fursikov</td>
<td>Certain optimal control problems for Navier-Stokes system with distributed control function</td>
</tr>
<tr>
<td>1349</td>
<td>F. Gesztesy, R. Nowell &amp; W. Pötz</td>
<td>One-dimensional scattering theory for quantum systems with nontrivial spatial asymptotics</td>
</tr>
<tr>
<td>1350</td>
<td>F. Gesztesy &amp; H. Holden</td>
<td>On trace formulas for Schrödinger-type operators</td>
</tr>
<tr>
<td>1351</td>
<td>X. Chen</td>
<td>Global asymptotic limit of solutions of the Cahn-Hilliard equation</td>
</tr>
<tr>
<td>1352</td>
<td>X. Chen</td>
<td>Lorenz equations, Part I: Existence and nonexistence of homoclinic orbits</td>
</tr>
<tr>
<td>1353</td>
<td>X. Chen</td>
<td>Lorenz equations Part II: “Randomly” rotated homoclinic orbits and chaotic trajectories</td>
</tr>
<tr>
<td>1354</td>
<td>X. Chen</td>
<td>Lorenz equations, Part III: Existence of hyperbolic sets</td>
</tr>
<tr>
<td>1356</td>
<td>C. Liu</td>
<td>The Helmholtz equation on Lipschitz domains</td>
</tr>
<tr>
<td>1357</td>
<td>G. Avalos &amp; I. Lasiecka</td>
<td>Exponential stability of a thermoelastic system without mechanical dissipation</td>
</tr>
<tr>
<td>1358</td>
<td>R. Lipton</td>
<td>Heat conduction in fine scale mixtures with interfacial contact resistance</td>
</tr>
<tr>
<td>1359</td>
<td>V. Odisharia &amp; J. Peradze</td>
<td>Solvability of a nonlinear problem of Kirchhoff shell</td>
</tr>
<tr>
<td>1360</td>
<td>P.J. Olver, G. Sapiro &amp; A. Tannenbaum</td>
<td>Affine invariant edge maps and active contours</td>
</tr>
<tr>
<td>1361</td>
<td>R.D. James</td>
<td>Hysteresis in phase transformations</td>
</tr>
<tr>
<td>1362</td>
<td>A. Sei &amp; W. Symes</td>
<td>A note on consistency and adjointness for numerical schemes</td>
</tr>
<tr>
<td>1363</td>
<td>A. Friedman &amp; B. Hu</td>
<td>Head-media interaction in magnetic recording</td>
</tr>
<tr>
<td>1364</td>
<td>A. Friedman &amp; J.J.L. Velázquez</td>
<td>Time-dependent coating flows in a strip, part I: The linearized problem</td>
</tr>
<tr>
<td>1365</td>
<td>X. Ren &amp; M. Winter</td>
<td>Young measures in a nonlocal phase transition problem</td>
</tr>
<tr>
<td>1366</td>
<td>K. Bhattacharya &amp; R.V. Kohn</td>
<td>Elastic energy minimization and the recoverable strains of polycrystalline shape-memory materials</td>
</tr>
<tr>
<td>1367</td>
<td>G.A. Chechkin</td>
<td>Operator pencil and homogenization in the problem of vibration of fluid in a vessel with a fine net on the surface</td>
</tr>
<tr>
<td>1368</td>
<td>M. Carme Calderer &amp; B. Mukherjee</td>
<td>On Poiseuille flow of liquid crystals</td>
</tr>
<tr>
<td>1369</td>
<td>M.A. Finsky &amp; M.E. Taylor</td>
<td>Pointwise Fourier inversion: A wave equation approach</td>
</tr>
<tr>
<td>1370</td>
<td>D. Brandon &amp; R.C. Rogers</td>
<td>Order parameter models of elastic bars and precursor oscillations</td>
</tr>
<tr>
<td>1371</td>
<td>H.A. Levine &amp; B.D. Sleeman</td>
<td>A system of reaction diffusion equations arising in the theory of reinforced random walks</td>
</tr>
<tr>
<td>1372</td>
<td>B. Cockburn &amp; P.-A. Gremaud</td>
<td>A priori error estimates for numerical methods for scalar conservation laws. Part II: Flux-splitting monotone schemes on irregular Cartesian grids</td>
</tr>
<tr>
<td>1373</td>
<td>B. Li &amp; M. Luskin</td>
<td>Finite element analysis of microstructure for the cubic to tetragonal transformation</td>
</tr>
<tr>
<td>1374</td>
<td>M. Luskin</td>
<td>On the computation of crystalline microstructure</td>
</tr>
<tr>
<td>1375</td>
<td>J.P. Matos</td>
<td>On gradient young measures supported on a point and a well</td>
</tr>
<tr>
<td>1376</td>
<td>M. Nitsche</td>
<td>Scaling properties of vortex ring formation at a circular tube opening</td>
</tr>
<tr>
<td>1377</td>
<td>J.L. Bona &amp; Y.A. Li</td>
<td>Decay and analyticity of solitary waves</td>
</tr>
<tr>
<td>1378</td>
<td>V. Isakov</td>
<td>On uniqueness in a lateral cauchy problem with multiple characteristics</td>
</tr>
<tr>
<td>1379</td>
<td>M.A. Kouritzik</td>
<td>Averaging for fundamental solutions of parabolic equations</td>
</tr>
<tr>
<td>1380</td>
<td>T. Aktosun, M. Klaus &amp; C. van der Mee</td>
<td>Integral equation methods for the inverse problem with discontinuous wavespeed</td>
</tr>
<tr>
<td>1381</td>
<td>P. Morin &amp; R.D. Spies</td>
<td>Convergent spectral approximations for the thermomechanical processes in shape memory allows</td>
</tr>
<tr>
<td>1382</td>
<td>D.N. Arnold &amp; X. Liu</td>
<td>Interior estimates for a low order finite element method for the Reissner-Mindlin plate model</td>
</tr>
<tr>
<td>1383</td>
<td>D.N. Arnold &amp; R.S. Falk</td>
<td>Analysis of a linear-linear finite element for the Reissner-Mindlin plate model</td>
</tr>
<tr>
<td>1384</td>
<td>D.N. Arnold, R.S. Falk &amp; R. Winther</td>
<td>Preconditioning in $H(\text{div})$ and applications</td>
</tr>
<tr>
<td>1385</td>
<td>M. Laven'tieiev</td>
<td>Nonlinear parabolic problems possessing solutions with unbounded gradients</td>
</tr>
</tbody>
</table>
1386 O.P. Bruno & P. Laurence, Existence of three-dimensional toroidal MHD equilibria with nonconstant pressure
1387 O.P. Bruno, F. Reitich, & P.H. Leo, The overall elastic energy of polycrystallin martensitic solids
1388 M. Fila & H.A. Levine, On critical exponents for a semilinear parabolic system coupled in an equation and a boundary condition
1390 J.M. Berg & H.G. Kwany, Unfolding the zero structure of a linear control system
1391 A. Sei, High order finite-difference approximations of the wave equation with absorbing boundary conditions: A stability analysis
1392 A.V. Coward & Y.Y. Renardy, Small amplitude oscillatory forcing on two-layer plane channel flow
1393 V.A. Pliss & G.R. Sell, Approximation dynamics and the stability of invariant sets
1394 J.G. Cao & P. Roblin, A new computational model for heterojunction resonant tunneling diode
1395 C. Liu, Inverse obstacle problem: Local uniqueness for rougher obstacles and the identification of a ball
1396 K.A. Pericak-Spector & S.J. Spector, Dynamic cavitation with shocks in nonlinear elasticity
1397 G. Avalos & V. Lasička, Exponential stability of a thermoelastic system without mechanical dissipation II: The case of simply supported boundary conditions
1398 B. Brighi & M. Chipot, Approximation of minima in the calculus of variations
1399 G. Avalos, Concerning the well-posedness of a non-linearly coupled semilinear wave and beam-like equation
1400 R. Lipton, Variational methods, bounds and size effects for composites with highly conducting interface
1401 B.T. Hayes & P.G. LeFloh, Non-classical shock waves in scalar conservation laws
1402 K.T. Joseph & P.G. LeFloh, Boundary layers in weak solutions to hyperbolic conservation laws
1403 Y. Diao, C. Ernst, & E.J.J. Van Rensburg, Energies of knots
1404 Xiaofeng Ren, Multi-layer local minimum solutions of the bistable equation in an infinite tube
1405 Vlastimil Pták, Krylov sequences and orthogonal polynomials
1406 T. Aktosun, M. Klaus, & C. van der Mee, Factorization of scattering matrices due to partitioning of potentials in one-dimensional Schrödinger-type equations
1408 D.N. Arnold, R.S. Falk, & R. Winther, Preconditioning discrete approximations of the Reissner-Mindlin plate model
1409 M.A. Kouritzin, On exact filters for continuous signals with discrete observations
1410 R. Lipton, The second Stekloff eigenvalue and energy dissipation inequalities for functionals with surface energy
1411 R. Lipton, The second Stekloff eigenvalue of an inclusion and new size effects for composites with imperfect interface
1412 W. Littman & B. Liu, The regularity and singularity of solutions of certain elliptic problems on polygonal domains
1413 C.R. Collins, Spurious oscillations are not fatal in computing microstructures
1414 M.A. Horn, Sharp trace regularity for the solutions of the equations of dynamic elasticity
1415 A. Friedman, B. Hu & Y. Liu, A boundary value problem for the Poisson equation with multi-scale oscillating boundary
1416 P. Baumann, D. Phillips & Q. Tang, Stable nucleation for the Ginzburg-Landau system with an applied magnetic field
1417 J.M. Berg, A strain profile for robust control of microstructure using dynamic recrystallization
1418 P. Klouček, Toward the computational modeling of nonequilibrium thermodynamics of the Martensitic transformations
1419 S. Chawla & S.M. Lenhart, Application of optimal control theory to in situ bioremediation
1420 B. Li & M. Luskin, Nonconforming finite element approximation of crystalline microstructure
1421 H. Kang & J.K. Seo, Inverse conductivity problem with one measurement: Uniqueness of balls in $\mathbb{R}^3$
1422 Avner Friedman & Robert Gulliver, Organizers, Mathematical modeling for instructors, July 29 – August 16. 1996
1423 G. Friesecke, Pair correlations and exchange phenomena in the free electron gas
1424 Y.A. Li & P.J. Olver, Convergence of solitary-wave solutions in a perturbed Bi-Hamiltonian dynamical system
   1. Compactons and Peakons  II. Complex Analytic Behavior  III. Convergence to Non-Analytic Solutions
1425 C. Huang, On boundary regularity of vortex patches for 3D incompressible euler systems
1426 C. Huang, A free boundary problem with nonlinear jump and kinetics on the free boundaries
1427 X. Chen, C. Huang & J. Zhao, A nonlinear parabolic equation modeling surfactant diffusion
1428 A. Friedman & B. Hu, Optimal control of chemical vapor deposition reactor
1429 A. Friedman & B. Hu, A non-stationary multi-scale oscillating free boundary for the Laplace and heat equations
1430 X. Chen, Existence, uniqueness, and asymptotic stability of traveling waves in nonlocal evolution equations
1431 J. Yong, Finding adapted solutions of forward-backward stochastic differential equations – Methods of continuation
1432 J. Yong, Linear forward-backward stochastic differential equations
1433 D.A. Dawson & M.A. Kouritzin, Invariance principles for parabolic equations with random coefficients
1434 R. Lipton, Energy minimizing configurations for mixtures of two imperfectly bonded conductors