Proceedings of the Eighth

PIMS Graduate Industrial Math Modelling Camp

May 7–11, 2005
University of Lethbridge

Co-sponsored by:

Pacific Institute for the Mathematical Sciences

Informatics Circle of Research Excellence

Alberta Innovation and Science

University of Lethbridge

Editors: Elena Braverman (University of Calgary)
and Daya Gaur (University of Lethbridge)

FOREWORD BY THE PIMS DIRECTOR

The annual PIMS **Graduate Industrial Mathematics Modelling Camp (GIMMC)** is held in one of the PIMS universities as part of the PIMS industrial programme. It is part of PIMS' commitment to providing training for young mathematical scientists who are either pursing careers in academia or in industry.

The goal of the GIMMC is to provide experience in the use of mathematical modelling as a problem-solving tool for graduate students in mathematics, applied mathematics, statistics and computer science. In addition, it helps prepare the graduate students for the **Industrial Problem Solving Workshop** which takes place the week after GIMMC.

At the workshop, students work together in teams under the supervision of invited mentors. Each mentor poses a problem arising from an industrial or engineering application and guides his or her team of graduate students through a modelling phase to a resolution.

The Eighth GIMMC was held May 7–11, 2005, at the University of Lethbridge in Alberta. This year it was co-sponsored by Informatics Circle of Research Excellence (iCORE), Alberta Innovation and Science, and the University of Lethbridge.

There were four problems posed, and a total of thirty-four students in attendance. The students came from across North America with seven from the United States.

My sincere appreciation and gratitude goes to all the people involved in this workshop. In particular I would like to thank the organizers (Elena Braverman, Hadi Kharaghani) and the mentors (C. Sean Bohun, Chris Bose, Lou Fishman, Daya Gaur).

I look forward to the 2006 GIMMC which will be held at Simon Fraser University.

Dr. Ivar Ekeland, Director
Pacific Institute for the Mathematical Sciences

EDITOR'S PREFACE

From May 7 through May 11, 2005, thirty-four graduate students and four mentors gathered at the University of Lethbridge for the Eighth Annual Graduate Industrial Mathematics Modelling Camp (GIMMC). The students came from all across Canada (from Newfoundland to Victoria) and the United States (from Massachusetts to California).

Both students and mentors mentioned that the University of Lethbridge was an excellent venue for the event.

Most of the participants arrived on Saturday, May 7, at the Calgary airport, and proceeded by bus to Lethbridge. The next day (Sunday, May 8) began with presentations of problems by mentors. Shortly after this, students broke into four teams to tackle the problems. For three and a half days students, guided by mentors, worked on problems, until the final presentations on May 11. After that students went to Calgary for the MITACS meeting (May 11-14). Some of the groups mentioned that they continued to work on the problem during the MITACS meeting. Some time after the workshop, proceedings papers were submitted for review, and these papers form the rest of this document.

A workshop could not be successful without the combined effort of many. First, we would like to thank mentors:

- Sean Bohun (Pennsylvania State University)
- Chris Bose (University of Victoria)
- Lou Fishman (MDF International)
- Daya Gaur (University of Lethbridge)

Feedback from students indicated that the mentors were greatly appreciated for their efforts. For example:

“*Our mentor's passion in research and treating us as fellow researchers were so inspiring.*”

“*The topic of the our GIMMC problem was awesome.*”

“*The problems were excellent, and the facilities were great as well.*”

We would also like to express our appreciation to Wolf Holzmann, Jim Liu and Amir Akbary (University of Lethbridge) for contributing to the organization of the conference. They were also involved in editing proceedings, as was Bill Sands (University of Calgary). The Administrative Assistant, Zoia Schacher (University of Lethbridge), and the PIMS Administrative Assistant, Marian Miles (University of Calgary), were invaluable. We also acknowledge PIMS Website Manager, Kelly Choo, for his support with the web site, and PIMS Publications and Communications Manager, Heather Jenkins, for her assistance with the editorial process.

Finally, we thank the local organizer Hadi Kharaghani (University of Lethbridge) for his efforts that led to a very successful modelling camp.

Elena Braverman (University of Calgary)
and Daya Gaur (University of Lethbridge)

October 2005

Contents

Chapter 1

Modelling a Stirling Engine

Participants: C. Sean Bohun (Mentor, Penn State University), Parisa Jamali (University of Western Ontario), Massih Khorvash (University of Lethbridge), Mahyar Mohajer (University of Calgary), Comron Nouri (University of Western Ontario), James Odegaard (University of Western Ontario), Peter Smith (Memorial University), Naveen Vaidya (York University)

PROBLEM STATEMENT: A Stirling engine is an externally heated engine developed in the 1800's as a safe alternative to the steam engine. The boilers of steam engines were known to explode due to a combination of the high pressure of the steam and the use of materials of insufficient strength in their construction.

The engine functions by repeatedly heating and cooling a sealed amount of gas. Since the gas is sealed, there are no intake or exhaust valves as one would find with other piston engines. In addition, the theoretical efficiency of an ideal Stirling engine cycle exceeds that of a steam engine. In fact, it is has the same efficiency as a Carnot heat engine for the same input and output temperatures. The sealed gas operates at a reduced pressure, consequently allowing the cylinders to be much less robust than their internal combustion counterparts. However, due to the size of the heat exchangers, the engines tend to be quite large. This is especially true for Stirling engines that run on a small temperature differential.

Stirling engines are also capable of working in reverse in that motion applied to the shaft will set up a temperature differential across the tube containing the sealed gas. Because of this effect, one of the modern uses of Stirling engines is as a refrigerator.

Traditionally the analysis of a Stirling engine begins by assuming that the hot and cold sides of the engine remain at a fixed temperature throughout the engine cycle. This is tantamount to assuming that the heat exchangers and regenerator are perfectly effective at maintaining the spatial temperature distribution within the engine. Rather than determining the heat transferred, work done, and efficiency of the engine, our primary goal is to determine the equation of motion describing the flywheel of the engine.

Our second goal is to attempt to understand the gas flow inside the sealed chamber. Again, we assume an ideal gas where the mass flow is driven by the temperature distribution along the sealed tube.

In realistic machines the working spaces are more accurately approximated with a adiabatic assumption rather than an isothermal assumption. This means that no heat (thermal energy) is transferred by the enclosed gas which is the opposite extreme of the isothermal case where heat transfer is maximized. Consequently, the next problem to be considered would be to use an adiabatic assumption allowing one to make insights into the design of the regenerator and the effectiveness of using various gases.

1.1 Introduction

In this report we begin modelling a Stirling engine in the α-configuration using a Lagrangian formalism to obtain the equation of motion for the system. By considering a one dimensional gas dynamics model for the trapped gas, we find that at the zeroth order of the gas dynamics equations, the pressure is uniform within the gas chamber. The gas dynamics model also describes a clear hierarchy of the higher order effects.

The following sections of this paper proceeds as follows: In Section 1.2 we derive the equations of motion for a particular Stirling engine with a w-linkage and illustrate the dynamical behaviour of the model. In Section 1.3 we use the theory of gas dynamics to derive a simplified set of equations of motion for the trapped gas. At zeroth order the gas dynamics model predicts that the pressure is uniform within the chamber, consistent with our Lagrangian model. In the final section of the paper we propose some future directions for research.

1.2 Equations of Motion

We consider an α-configuration Stirling engine as shown in Figure 1.1. The x-axis of the coordinate system is parallel to the length of the cylinder. The quantities $x_1(t)$ and $x_2(t)$ represent the positions from the inner side of each piston to the y-axis respectively. Each piston is connected to the corresponding triangle formed by rigid rods of various lengths. The applied temperature is assumed to be piecewise constant along the length of the chamber of the form

$$T(x) = \begin{cases} T_1, & x_1 \le x < 0 \\ T_2, & 0 \le x \le x_2. \end{cases}$$

The grey region separating the two sides of the chamber indicated in Figure 1.1 is considered ideal in that it maintains the temperature differential while not impeding the flow of the gas from one side to the other.

By assuming an ideal gas, an expression for the pressure is simply a result of the fact that the amount of gas in the chamber is constant. If A denotes the cross sectional area of the chamber and n the number of moles of contained gas then

$$P(x_1, x_2) = \frac{nRT_1T_2}{A(T_1x_2 - T_2x_1)} \tag{1.1}$$

where R is an appropriate ideal gas constant ($R = 8.314$ J/K/mol in MKS units).

The motion is constrained by the linkage and to derive expressions for the constraints we notice that

$$\overrightarrow{OP} = \langle r\sin x_3, h - r\cos x_3\rangle, \qquad \overrightarrow{OQ} = \langle d - a\cos\varphi_2, -a\sin\varphi_2\rangle, \tag{1.2}$$

$$\overrightarrow{OR} = \langle -d + a\cos\varphi_1, -a\sin\varphi_1\rangle. \tag{1.3}$$

Algebraic constraints due to the linkage result from the condition that $|\overrightarrow{PR}|^2 = |\overrightarrow{PQ}|^2 = l^2$ yielding

$$g_1(x_1, x_3) = (d - a\cos\varphi_1 + r\sin x_3)^2 + (h + a\sin\varphi_1 - r\cos x_3)^2 - l^2 = 0 \tag{1.4a}$$

and

$$g_2(x_2, x_3) = (-d + a\cos\varphi_2 + r\sin x_3)^2 + (h + a\sin\varphi_2 - r\cos x_3)^2 - l^2 = 0 \tag{1.4b}$$

where

$$\varphi_1 = \alpha - \cos^{-1}\left(\frac{d - \zeta + x_1}{b}\right), \qquad \varphi_2 = \alpha - \cos^{-1}\left(\frac{d - \zeta - x_2}{b}\right). \tag{1.4c}$$

To obtain the equation of motion we use the Lagrange equation

$$\frac{d}{dt}\left(\frac{\partial K}{\partial \dot{x}_3}\right) - \frac{\partial K}{\partial x_3} = Qr \tag{1.5}$$

where K is the kinetic energy of the system and Q consists of the forces acting on the flywheel due to the pressure in the chamber and the friction of the pistons.

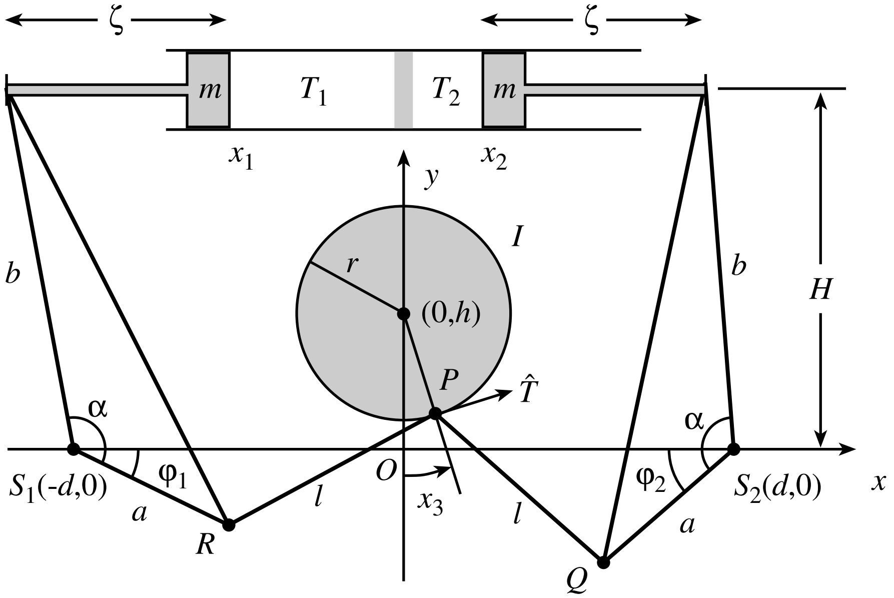


Figure 1.1: An ideal α-configuration Stirling engine.

The kinetic energy of the system is given by

$$K = \frac{m_p}{2}\left(\dot{x}_1^2 + \dot{x}_2^2\right) + \frac{I}{2}\dot{x}_3^2$$

where m_p is mass of the piston assemblies and $I = m_d r^2/2$ is the moment of inertia of the solid cylindrical flywheel of mass m_f and radius r. As a result of the expressions (1.4) the dynamics of the drive wheel are connected to the motion of the pistons through

$$\frac{\partial g_1}{\partial x_1}\dot{x}_1 + \frac{\partial g_1}{\partial x_3}\dot{x}_3 = 0, \qquad \frac{\partial g_2}{\partial x_2}\dot{x}_2 + \frac{\partial g_2}{\partial x_3}\dot{x}_3 = 0. \tag{1.6}$$

Using these relations, the kinetic energy becomes

$$K = \left[\frac{m_p}{2}\left(\frac{a_{13}^2}{a_{11}^2} + \frac{a_{23}^2}{a_{22}^2}\right) + \frac{I}{2}\right]\dot{x}_3^2 \tag{1.7}$$

with $a_{ij} = \partial g_i/\partial x_j$.

There are two separate components for Q in equation (1.5). These include the forces due to the pistons acting tangentially to the flywheel and the frictional forces of the pistons. The former of these can be computed by determining the torque applied by each piston and then resolving the forces acting along $\overrightarrow{QP}$ and $\overrightarrow{RP}$ in the direction tangent to the flywheel. In detail, solving for the force along $\overrightarrow{QP}$

$$\overrightarrow{S_2Q} \times \vec{F}_{QP} = \frac{|\vec{F}_{QP}|}{l}\overrightarrow{S_2Q} \times \overrightarrow{QP} = A(P - P_0)H(-\hat{\mathbf{k}})$$

where P_0 is the ambient pressure outside the chamber and P is given by (1.1). Using (1.2),

$$|\vec{F}_{QP}| = \frac{A(P - P_0)Hl/a}{(h - r\cos x_3)\cos\varphi_2 + (d - r\sin x_3)\sin\varphi_2}$$

Linkage arms		Pistons/flywheel		Gas	
a	3.61 cm	h	2.86 cm	R	8.314 J/K/mol
b	7.72 cm	r	1.00 cm	n	1.20 mmol
α	143°	m_p	200 g	A	3.14 cm^2
l	5.00 cm	m_f	200 g	P_0	101325 Pa
d	6.39 cm	ζ	4.67 cm	T_1	293.15 K
H	7.61 cm	γ	1000 g/s	T_2	393.15 K

Table 1.1: Values of the parameters used for the simulation.

and resolving in a direction tangential to the flywheel one obtains

$$\begin{aligned} F_2 &= \operatorname{comp}_{\hat{\mathbf{T}}} \vec{F}_{QP} = |\vec{F}_{QP}| \frac{\overrightarrow{QP}}{l} \cdot \langle \cos x_3, \sin x_3 \rangle \\ &= A(P - P_0) \frac{H}{a} \frac{(h + a \sin\varphi_2) \sin x_3 - (d - a\cos\varphi_2)\cos x_3}{(h - r\cos x_3)\cos\varphi_2 + (d - r\sin x_3)\sin\varphi_2}. \end{aligned} \tag{1.8}$$

A similar argument for the force due to the piston on the left yields

$$F_1 = A(P - P_0) \frac{H}{a} \frac{(h + a \sin\varphi_1) \sin x_3 + (d - a\cos\varphi_1)\cos x_3}{(h - r\cos x_3)\cos\varphi_1 + (d + r\sin x_3)\sin\varphi_1}. \tag{1.9}$$

Combining these results together and using (1.5) gives the equation of motion

$$A(x_3)\ddot{x}_3 + B(x_3)\dot{x}_3^2 + C(x_3)\dot{x}_3 = D(x_3) \tag{1.10}$$

where

$$\left.\begin{aligned} A(x_3) &= m_p\left(\Gamma_1^2 + \Gamma_2^2\right) + I, \\ B(x_3) &= m_p\left(\Gamma_1 \frac{\partial\Gamma_1}{\partial x_3} + \Gamma_2 \frac{\partial\Gamma_2}{\partial x_3} - 2\Gamma_1^2 \frac{\partial\Gamma_1}{\partial x_1} - 2\Gamma_2^2 \frac{\partial\Gamma_2}{\partial x_2}\right), \\ C(x_3) &= -\gamma(\Gamma_1 + \Gamma_2)r, \\ D(x_3) &= (F_1 + F_2)r \end{aligned}\right\} \tag{1.11a}$$

and

$$\begin{aligned} \Gamma_1 = \frac{a_{13}}{a_{11}} &= \left[b^2 + (d - \zeta + x_1)^2\right]^{1/2} \frac{r}{a} \\ &\quad \times \frac{(h + a\sin\varphi_1)\sin x_3 + (d - a\cos\varphi_1)\cos x_3}{(h - r\cos x_3)\cos\varphi_1 + (d + r\sin x_3)\sin\varphi_1}, \end{aligned} \tag{1.11b}$$

$$\begin{aligned} \Gamma_2 = \frac{a_{23}}{a_{22}} &= -\left[b^2 + (d - \zeta - x_2)^2\right]^{1/2} \frac{r}{a} \\ &\quad \times \frac{(h + a\sin\varphi_2)\sin x_3 - (d - a\cos\varphi_2)\cos x_3}{(h - r\cos x_3)\cos\varphi_2 + (d - r\sin x_3)\sin\varphi_2}. \end{aligned} \tag{1.11c}$$

For any position x_3, the values of x_1 and x_2 are determined by (1.4). These values in turn define the pressure through equation (1.1) and as a result the values of F_1 and F_2 with (1.9) and (1.8) respectively.

Figure 1.2 illustrates the dependence of the coefficients in expression (1.10) with respect to x_3 using the data in Table 1.1. From these curves we see that this linkage has four equilibrium points $\{(x_3, \dot{x}_3)\} = (0,0), (0,\pi), (0, \pm\xi)$ with $\xi \simeq \pi/3$. By studying the phase portrait one can completely characterize the possible motion. Note that only $D(x_3)$ and hence the position of the equilibrium points can be controlled by modifying the pressure and temperature of the trapped gas. To the left of Figure 1.3 is the typical time dependence of the angular coordinate x_3. In the illustrated case, the flywheel starts at rest at an angle of $x_3(0) = 5°$ and the subsequent motion consists of

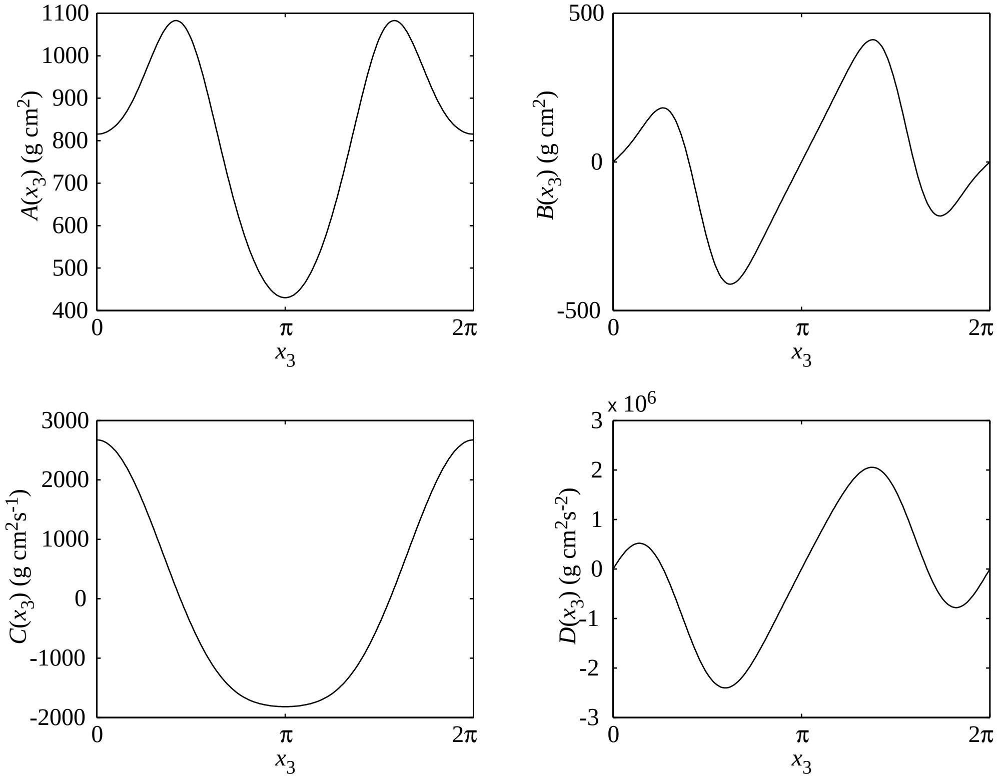


Figure 1.2: Angular dependence of the coefficients in expression (1.11a).

the flywheel oscillating until enough angular momentum is built up to complete one rotation. After this point the rotation becomes nearly uniform. The pressure of the trapped gas throughout one cycle is displayed to the right in Figure 1.3. Numerical experimentation suggests that the square of the angular frequency is proportional to the characteristic size of the ratio $\max_{x_3} |D/A|$. This relationship could be verified with an application of classical Floquet theory. See [4] for a modern treatment.

To extend our understanding of this model we next consider the motion of the trapped gas by developing a set of gas dynamics equations.

1.3 Gas Dynamics

Inside the chamber we consider a gas with macroscopic values of density, velocity, and energy of ρ, v, and ϵ respectively. These quantities are allowed to evolve both in time and with respect to position within the chamber.

If we assume that the gas evolves according to a Boltzmann equation with an underlying probability density function $f(x, v, t)$ then the macroscopic quantities can be obtained by computing the moments

$$\rho = \int f dv, \qquad \rho u = \int v f dv, \qquad \rho\epsilon = \frac{1}{2}\int |u - v|^2 f dv. \tag{1.12}$$

We further note that the pressure P is given by

$$P = \frac{1}{3}\int |u - v|^2 f dv = \frac{2}{3}\rho\epsilon.$$

Under suitable regularity assumptions, these quantities can be shown to satisfy the system of partial differential

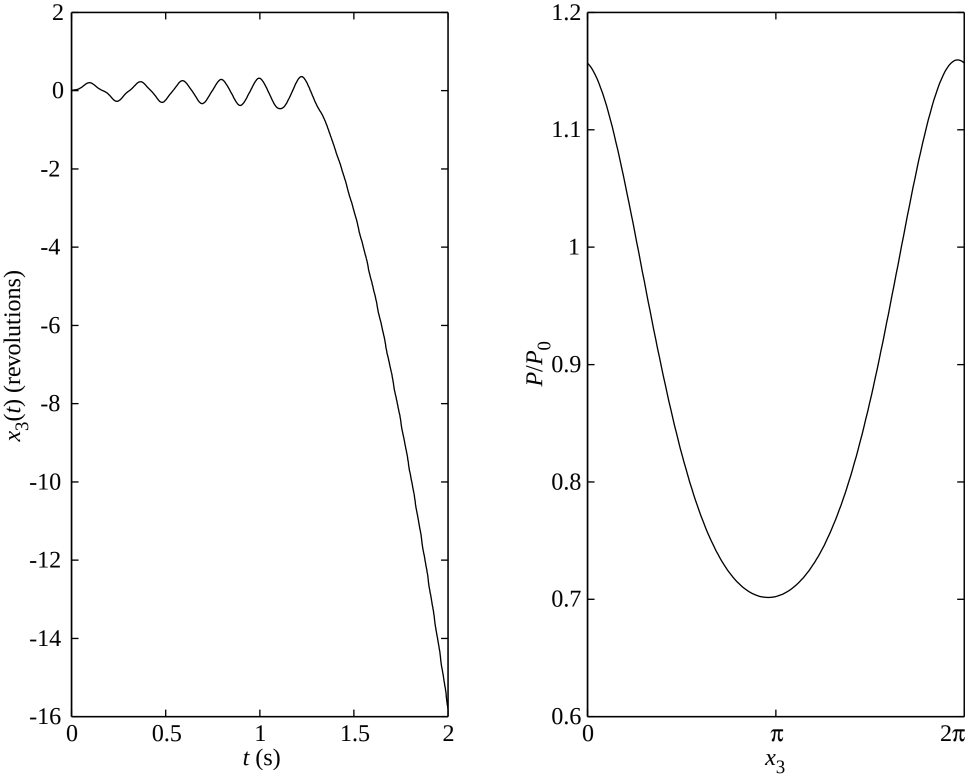


Figure 1.3: To the left is the time dependence of the rotation angle x_3 and to the right is the corresponding internal pressure of the trapped gas throughout one cycle.

equations [2]

$$\rho_t + (\rho u)_x = 0, \tag{1.13a}$$

$$(\rho u_i)_t + \sum_{j=1}^{3} (\rho u_i u_j + p_{ij})_{x_j} = 0, \tag{1.13b}$$

$$\left(\frac{1}{2}\rho|u|^2 + \rho\epsilon\right)_t + \sum_{j=1}^{3} \left[\rho u_j \left(\frac{1}{2}|u|^2 + \epsilon\right) + \sum_{l=1}^{3} u_l p_{jl} + q_j\right]_{x_j} = 0 \tag{1.13c}$$

where

$$p_{ij} = \int (u_i - v_i)(u_j - v_j) f dv, \qquad q_i = \frac{1}{2} \int (u_i - v_i)|u - v|^2 f dv. \tag{1.13d}$$

To determine the stress tensor p_{ij} and the heat flux q_i we require an explicit determination of the probability distribution f. At this point we assume that the gas chamber has an externally imposed temperature distribution which is radially independent. Furthermore, at any position within the chamber we assume that the gas behaves according to a local Maxwellian consistent with the temperature at that position. To this end we suppose that

$$f(x, v, t) = C_1 e^{-C_2(v-v_0)^2}$$

where C_1 and C_2 are functions of the position x. Using the definitions (1.12) we find that

$$\rho = C_1 \left(\frac{\pi}{C_2}\right)^{3/2}, \qquad \rho u = v_{0x} \left(\frac{\pi}{C_2}\right)^{3/2}, \qquad v_{0y} = 0, \qquad v_{0z} = 0, \tag{1.14a}$$

and

$$C_2 = \frac{3}{4}\left(\epsilon + \frac{1}{2}u^2\right)^{-1} \simeq \frac{3}{4\epsilon}. \tag{1.14b}$$

Computing p_{ij} and q_i according to (1.13d) we obtain the simplified set of expressions

$$\left.\begin{aligned} \rho_t + \alpha(\rho u)_x &= 0, \\ (\rho u)_t + \left(\rho u^2 - \frac{2}{3}\rho\epsilon\right)_x &= 0, \\ \left(\frac{1}{2}\rho u^2 + \rho\epsilon\right)_t + \left(\frac{2}{3}\rho\epsilon u - \frac{3}{2}\rho u^3\right)_x &= 0 \end{aligned}\right\} \tag{1.15}$$

for the enclosed gas.

1.3.1 Dimensional Analysis

We non-dimensionalize the model equations by defining

$$\hat{t} = t/\tau, \qquad \hat{x} = x/L, \qquad \hat{\rho} = \rho/R, \qquad \hat{u} = u/V, \qquad \hat{\varepsilon} = \epsilon/E$$

where $\tau = 1/3$ s, $L = 10$ cm, $V = 1$ cm/s, and $E = 1000$ J/kg are characteristic quantities for the process we are investigating. Transforming to the non-dimensional quantities and removing the hats we find

$$\left.\begin{aligned} \rho_t + \alpha(\rho u)_x &= 0, \\ (\rho\epsilon)_x - \alpha(\rho u)_t - \alpha\beta(\rho u^2)_x &= 0, \\ (\rho\epsilon)_t + \frac{2}{3}\alpha(\rho\epsilon u)_x + \frac{1}{3}\alpha\beta(\rho u^2)_t - \alpha^2\beta(\rho u^3)_x &= 0, \end{aligned}\right\} \tag{1.16}$$

where

$$\alpha = \frac{V\tau}{L} \simeq 0.03, \qquad \beta = \frac{3VL}{2E\tau} \simeq 5 \times 10^{-6}$$

are small parameters reflecting the drift speed and kinetic energy of the gas.

1.3.2 Zeroth Order

Keeping only the zeroth order terms of (1.16) one obtains the system

$$\rho_t = 0, \qquad (\rho\epsilon)_x = 0, \qquad (\rho\epsilon)_t = 0 \tag{1.17}$$

with a solution

$$\epsilon = \epsilon(x), \qquad \rho = \rho(x), \qquad \rho\epsilon = \text{const.} \tag{1.18}$$

This is consistent with the static situation where the externally imposed temperature profile determines the energy and the density changes in such a way to ensure that the internal chamber pressure remains constant.

1.3.3 First Order Corrections

At first order, the motion of the gas begins to play a role in the dynamics. In this case we consider the system

$$\left.\begin{aligned} \rho_t + \alpha(\rho u)_x &= 0, \\ (\rho\epsilon)_x - \alpha(\rho u)_t &= 0, \\ (\rho\epsilon)_t + \tfrac{2}{3}\alpha(\rho\epsilon u)_x &= 0 \end{aligned}\right\} \tag{1.19}$$

with the added boundary condition that the velocity of the gas at the piston heads must be that of the pistons themselves.

A preliminary analysis of these first order equations indicates that the density and velocity of the gas remained invariant but the energy of the gas slowly increased in the regeneration region (near $x = 0$) while decreasing at the face of the piston in the hot side of the chamber.

1.4 Conclusions and Future Work

An equation of motion describing the flywheel was derived and for the chosen simulation parameters, the rotation became nearly uniform after a few complete rotations.

Using a dimensional analysis technique, the relationships defining the behaviour of the trapped gas (1.16) were found to cascade into a natural hierarchy of importance. To leading order one recovers a spatially uniform pressure that is constant in time. A significant omission in the description of the enclosed gas are the moving boundary conditions at the ends of the gas tube. The next step in this work should be an analysis of (1.19) with an appropriate set of boundary conditions.

Other considerations of the current work include simplifying the linkage and realistically modelling the regeneration region as an extensible body with well defined thermal properties. As mentioned in the problem statement, the consideration of an adiabatic rather than isothermal assumption of the working spaces would allow insights to be made into the design of the regeneration region and the effectiveness of using various gases.

Bibliography

[1] Arnold, V.I. (1997). *Mathematical Methods of Classical Mechanics*. In Graduate Texts in Mathematics. Springer-Verlag: New York.

[2] Cercignani, C., Illner, R., & Pulvirenti, M. (1994). *The Mathematical Theory of Dilute Gases*. In Applied Mathematical Sciences, Volume 106. Springer-Verlag: New York.

[3] Goldstein, H., Poole, C.P., & Safko, J.L. (2002). *Classical Mechanics*, 3rd. ed. Addison Wesley.

[4] Moore, G. (2005). *Floquet Theory as a Computational Tool*. SIAM Journal of Numerical Analysis, **42**(6), pp. 2522-2568.

Chapter 2

A Dynamical Model of Drill Efficiency

Participants: Chris Bose (Mentor, University of Victoria), Mahmoud Akelbek (University of Regina), Amir Amiraslani (University of Western Ontario), Robin Clysdale (University of Calgary), Hui Huang (University of British Columbia), Xiaoping Liu (University of Regina), Nargol Rezvani (University of Western Ontario), Sarah Williams (University of California, Davis)

PROBLEM STATEMENT: The method used to bore holes in rock or other hard, brittle material is fundamentally different from that used to drill in softer, more malleable materials (like wood or metal). One particularly simple form of the rock drill is called the tri-cone drill head, wherein three conical ”gears” equipped with hardened chipping points are mounted in a single housing. Drilling is effected by the contact of the teeth of these gears along the bottom of the bore hole as the drill head is rotated – each time a tooth meets the rock, a piece of rock is broken off. The efficiency of this form of drilling depends on quite a number of factors including the applied force (the weight of the drill and rotating machinery), the geometry of the drill head and chipping points, and, what we are most interested in, the angular speed of rotation of the drill. The action of the drillhead has at least three qualitatively different modes depending on this angular speed.

For low rotational speeds there is essentially no drilling action as the gears ’walk’ along the rock. At moderate angular speeds there is very inefficient drilling as the motion of the rollers enters a ’periodic’ mode. For sufficiently high angular speeds, the gears contact the rock in a ’fully chaotic’ mode and effective drilling begins. Finally, as the angular speed is further increased, the drilling efficiency (advance per revolution) first peaks, then drops off to approximately 60% of maximum efficiency. In the field, operators try to identify this point of maximum efficiency and run the drillhead near this speed.

In this workshop our goal was to produce a simple model of the tri-cone drill which captures the above qualitative features. In order to deal with the ’chaotic’ mode of operation, we construct our model as a discrete-time dynamical system. While there is no equilibrium configuration in the usual sense for our discrete time system, there is an invariant measure which describes the asymptotic behaviour of the drillhead after transients have decayed. Once the invariant measure is known, the asymptotics may be computed (or estimated) and the qualitative features noted above may be derived.

2.1 The Drilling Model

We start by making a rather brutal geometric simplification of our drillhead compared to what was described in the Problem Statement. In our model we ignore multiple gears (and hence any possible gear-to-gear interaction); we assume regular gears (reasonable, but not always the case in modern applications); and we assume a flat bottomed bore hole. Finally we replace the circular geometry of the bore hole with a linear geometry. We are left with a model of the rock drill consisting of one regular gear in two dimensions, moving horizontally along a straight line (x-axis). We assume the upward vertical direction is the positive z-axis. The gear has N teeth and maximum radius r meters. We denote the velocity in the x-direction by c, and time (in seconds) by t. We assume that c is held constant by the drill operator, so that

$$c = \frac{x}{t}, \tag{2.1}$$

or, equivalently,

$$x = t \cdot c. \tag{2.2}$$

It is a simple exercise to verify that the distance T between consecutive 'teeth' on the gear is given by

$$T = 2r \sin\left(\frac{\pi}{N}\right), \tag{2.3}$$

Therefore, one rotation of the drill head in the bore hole of radius B corresponds to an increase by $2\pi B$ in the value of x and to a passing of teeth on the gear

$$\frac{\pi B}{r \sin\left(\frac{\pi}{N}\right)} \tag{2.4}$$

We assume that the origin $x = 0$ corresponds to the gear sitting with two consecutive teeth contacting the bottom of the bore hole. Then $x = 0.05T$ corresponds to one tooth pointing straight down and $x = T$ returns the gear to the position of two contacting teeth. Thus the configuration of the gear with respect to the bottom of the bore hole is periodic of period T. This is illustrated in Figure 2.1 for the simple case of $N = 5$ and $T = 1$.

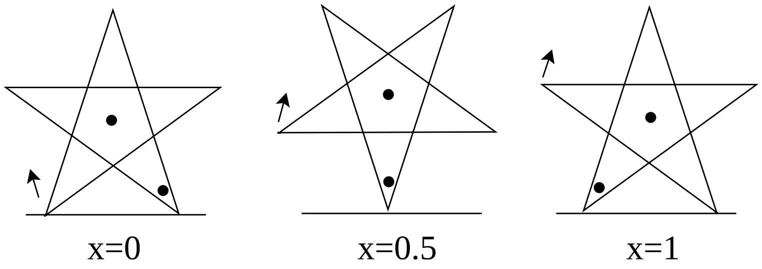


Figure 2.1: Drill configuration examples.

2.1.1 The Rolling Mode

The centre of a rolling drill traces the curve

$$b(x) = \sqrt{r^2 - (x - T/2)^2}, \tag{2.5}$$

where r is the maximum radius of the drill; that is, partial circles are traced as the drill rocks from one tooth to the next. We call this curve the base curve for the gear. You can clearly see the base curve in Figure 2.2.

Since the horizontal speed of the gear is held at the constant value c, using equation (2.2), the vertical acceleration of the gear moving along the base curve is $-c^2 b''(x)$, where the primes indicate differentiation with respect to the spatial variable x. We conclude that in order to maintain a rolling mode for the gear we must have

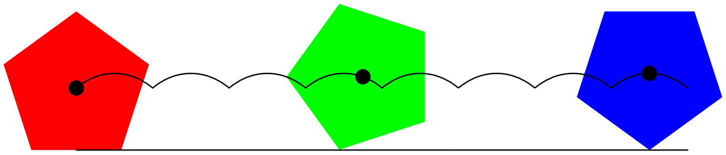


Figure 2.2: Path traced by the centre of a rolling drill ($b(x)$).

$$g \geq \frac{c^2}{r} \left(\frac{r}{b(x)} \right)^3 \tag{2.6}$$

for all $x \in [0, T]$, where g is the gravitational constant. Since the maximum on the right hand side is obtained when $x = 0$ (or $x = T$) the precise condition for the rolling mode is easily derived:

$$\left(1 - \left(\frac{T}{2r} \right)^2 \right)^{3/2} \geq \frac{c^2}{gr}. \tag{2.7}$$

The quantity $\omega^2 = \frac{c^2}{gr}$ is called Froude's number for the system. Note that $\omega \sim c$ for a gear of fixed radius.

2.1.2 The Bouncing Mode

From expression (2.6) we can also derive the critical value of c for the fully chaotic mode, that is, when

$$g \leq \frac{c^2}{r} \left(\frac{r}{b(x)} \right)^3$$

for all $x \in [0, T]$. Under this condition, whenever the gear makes contact with the rock it will immediately 'bounce' upward away from the rock due to the vertical acceleration imposed by the rocking of the gear over the contacting tooth. Since the minimum value of the right hand side is given when $x = \frac{T}{2}$, we record the bouncing condition as

$$\omega^2 = \frac{c^2}{gr} \geq 1. \tag{2.8}$$

Notice that this expression is independent of N, the number of teeth on the gear. A typical trajectory of a bouncing gear is shown in Figure 2.3.

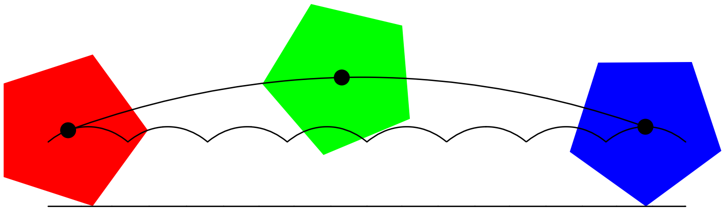


Figure 2.3: Path traced by the centre of a bouncing drill ($p(x)$).

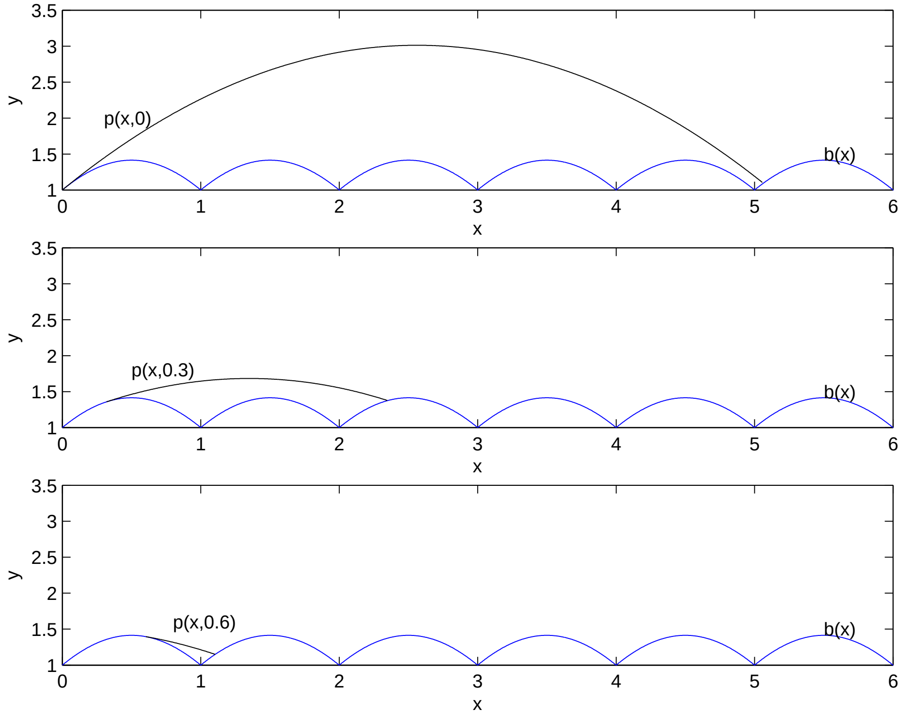


Figure 2.4: Paths traced by the centre of a rolling drill ($b(x)$) and a bouncing drill ($p(x)$).

2.1.3 The Mixed Mode

When $(1-(\frac{T}{2r})^2)^{3/2} \leq \omega^2 \leq 1$ the trajectory of the gear will pass through a combination of rolling and bouncing modes, depending on the configuration of the gear when it contacts the rock. In particular, the gear will leave the rock from contact configurations near $x = 0$ or $x = T$ and simply roll when the contact is at x near $\frac{T}{2}$. Since we are interested mainly in the chaotic motion of the drill, we will not discuss the mixed mode any further in this article.

2.1.4 Trajectory Map $\mathbb{T}$ for the Bouncing Mode

The centre of a bouncing drill launched from a contact configuration at $x = x_0$ traces the parabolic curve

$$p(x_0, x) = b(x_0) + b'(x_0)(x - x_0) - \frac{g}{2c^2}(x - x_0)^2. \tag{2.9}$$

A few of these curves are shown in Figure 2.4.

In order to calculate the bouncing trajectory we need to introduce the map $\mathbb{T} : [0,T] \to [0,T]$ that takes a launching configuration to the associated landing configuration. Therefore, we define $\mathbb{T}(x_0) = x \,(\mathrm{mod}\, T)$, where T is period and x is the smallest value $x > x_0$ such that

$$p(x_0, x) - b(x) = 0. \tag{2.10}$$

Clearly the map $\mathbb{T}$ is well defined. In the next section we will derive some of its geometric properties. For now, one should compare Figure 2.5 to Figure 2.4 and note that for a drill launched from a configuration greater than 0.5, only one tooth is advanced. On Figure 2.5, this corresponds to the right-most branch. The next branch to the left corresponds to two teeth advanced, and so on. In Section 4, we will show that there is a non-increasing, piecewise constant function $\kappa(x)$ which maps configuration x to number of teeth advanced between launch point

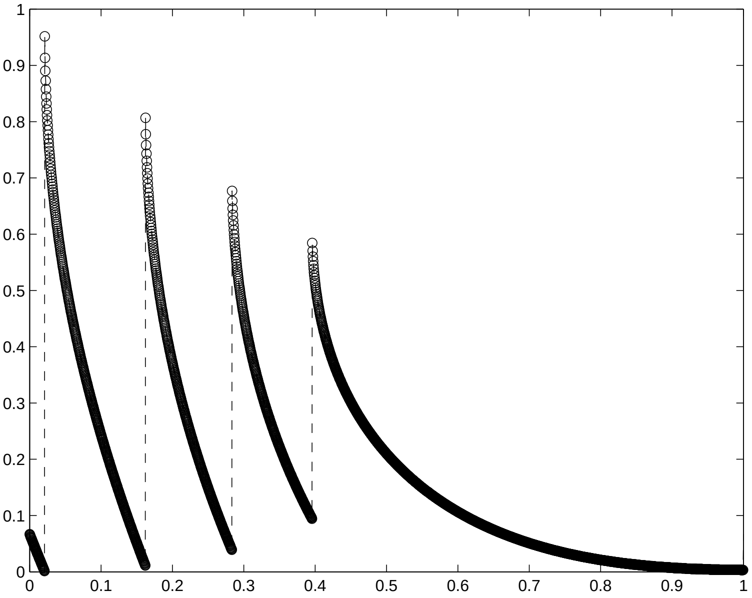


Figure 2.5: Map from launching configuration to landing configuration.

x and landing point $\mathbb{T}(x)$. This will allow us to reconstruct the true trajectory of the gear from only information about the launch and land configurations as elements in $[0, T]$.

2.2 Geometry of the Trajectory Map $\mathbb{T}$

Recall that equations (2.5) and (2.9) define the base curve $b(x)$ (gear contacting rock) and the bouncing trajectory $p(x_0, x)$ (with launch point x_0), respectively. We also remind the reader of the geometric parameters r, the radius of the gear and T, the distance between adjacent teeth on the gear. The base curve function $b(x)$ is periodic of period T and at this point it seems natural to non-dimensionalize the spatial variables as follows. Define

$$\mathrm{x} = \frac{x}{T}, \quad \mathrm{z} = \frac{z}{r}, \quad \mathrm{b(x)} = \frac{1}{r}b(T\mathrm{x}), \quad \mathrm{c} = \frac{c}{T}, \quad \mathrm{g} = \frac{g}{r}$$

so $\mathrm{x} \in [0, 1]$, $\mathrm{z} \geq 0$, and so-on.

The renormalized base curve becomes

$$\mathrm{b(x)} = \sqrt{1 - \frac{T^2}{r^2}\left(\mathrm{x} - \left(k + \frac{1}{2}\right)\right)^2}, \qquad k = 0, 1, 2, ...$$

Also, the critical velocity values, in non-dimensionalized form are easily derived:

Rolling condition:

$$w^2 \leq \left(1 - \left(\frac{T}{2r}\right)^2\right)^{\frac{3}{2}} \Leftrightarrow \frac{\mathrm{c}^2}{\mathrm{g}}\frac{T^2}{r^2} \leq \left(1 - \left(\frac{T}{2r}\right)^2\right)^{3/2} \Leftrightarrow \frac{\mathrm{c}^2}{\mathrm{g}} \leq \frac{T}{r}\left(\frac{r^2}{T^2} - \frac{1}{4}\right)^{3/2} = \mathrm{A} \tag{2.11}$$

Bouncing condition:

$$w^2 \geq 1 \Leftrightarrow \frac{\mathrm{c}^2}{\mathrm{g}} \geq \frac{r^2}{T^2}(> \frac{1}{4}) = \mathrm{B} \tag{2.12}$$

We are using A and B to denote the nondimensionalized critical values of velocity. Note also that

$$\sqrt{r^2 - \frac{T^2}{4}} \leq b(x) \leq r \Rightarrow \sqrt{1 - \frac{T^2}{4r^2}} \leq \mathrm{b(x)} \leq 1$$

$$\mathrm{b}'(\mathrm{x}) = \frac{T}{r} b'(T\mathrm{x}) = \frac{T^2(\frac{1}{2} - \mathrm{x})}{r^2 \mathrm{b(x)}}$$

$$\mathrm{b}''(\mathrm{x}) = \frac{T^2}{r} b''(T\mathrm{x}) = -\frac{T^2}{r^2 \mathrm{b(x)}^3}$$

$$g \leq -c^2 b''(x) \quad \text{for all } x \in [0,T] \Leftrightarrow \frac{\mathrm{g}r}{T^2\mathrm{c}^2} \leq -\frac{r}{T^2}\mathrm{b}''(\mathrm{x}) \Leftrightarrow \mathrm{b}''(\mathrm{x}) + \frac{\mathrm{g}}{\mathrm{c}^2} \leq 0$$

When in the bouncing mode, we need to use the non-dimensionalized trajectory

$$\mathrm{p}(\mathrm{x}_0, \mathrm{x}) = \frac{1}{r} p(T\mathrm{x}_0, T\mathrm{x}) \tag{2.13}$$

After renormalizing variables, the transformation $\mathbb{T}$ which is derived from the projectile map, launching from $\mathrm{x}_0 \in [0,1]$ to hit at $\mathrm{x} = \mathbb{T}(\mathrm{x}_0) \in [0,1]$ generates a graph on $[0,1] \times [0,1]$ (Figure 2.5). The map $\mathbb{T}$ is clearly well-defined, and from the geometry we expect it to be locally continuous. From the plot in Figure 2.5 it would appear to be in fact piecewise monotone decreasing. In order to derive these properties analytically we will appeal to the definition of $\mathbb{T}$ in implicit form: $\mathrm{x} = \mathbb{T}(\mathrm{x}_0)$ if and only if

$$G(\mathrm{x}_0, \mathrm{x}) = \mathrm{p}(\mathrm{x}_0, \mathrm{x}) - \mathrm{b}(\mathrm{x}) = 0 \tag{2.14}$$

We begin our analysis with

Lemma 1. Suppose $G(\mathrm{x}_0, \mathrm{x}) = 0$, Then

- $\frac{\partial G}{\partial \mathrm{x}}(\mathrm{x}_0, \mathrm{x}) = \mathrm{p}'(\mathrm{x}) - \mathrm{b}'(\mathrm{x})$
- $\frac{\partial G}{\partial \mathrm{x}}(\mathrm{x}_0, \mathrm{x}) < 0$ except at possibly finitely many pairs $(\mathrm{x}_0, \mathrm{x})$ corresponding to the ends of branches in the map $\mathbb{T}$ or landing points x where b' fails to exist.

Proof: A straightforward calculation gives (remember that all quantities are renormalized)

$$G_{\mathrm{x}}(\mathrm{x}_0, \mathrm{x}) = -\frac{\mathrm{g}}{\mathrm{c}^2}(\mathrm{x} - \mathrm{x}_0) + \mathrm{b}'(\mathrm{x}_0) - \mathrm{b}'(\mathrm{x})$$

and

$$\mathrm{p}(\mathrm{x}_0, \mathrm{x}) - \mathrm{b}(\mathrm{x}) = 0 \Leftrightarrow -\frac{\mathrm{g}}{\mathrm{c}^2}(\mathrm{x} - \mathrm{x}_0) = 2\left\{\frac{\mathrm{b}(\mathrm{x}) - \mathrm{b}(\mathrm{x}_0)}{\mathrm{x} - \mathrm{x}_0} - \mathrm{b}'(\mathrm{x}_0)\right\}$$

Combining, we get

$$G_{\mathrm{x}}(\mathrm{x}_0, \mathrm{x}) = 2\left\{\frac{\mathrm{b}(\mathrm{x}) - \mathrm{b}(\mathrm{x}_0)}{\mathrm{x} - \mathrm{x}_0} - \frac{\mathrm{b}'(\mathrm{x}) + \mathrm{b}'(\mathrm{x}_0)}{2}\right\}$$

Note that projectile $\mathrm{p}(\mathrm{x}_0, \mathrm{x})$ is parabolic arc and for parabolas, we know the slope of any chord = average value of two end point derivatives, i.e.

$$\frac{\mathrm{p}'(\mathrm{x}) + \mathrm{p}'(\mathrm{x}_0)}{2} = \frac{\mathrm{p}(\mathrm{x}) + \mathrm{p}(\mathrm{x}_0)}{\mathrm{x} - \mathrm{x}_0}$$

So, since $\mathrm{b}(\mathrm{x}) = \mathrm{p}(\mathrm{x})$ and $\mathrm{b}(\mathrm{x}_0) = \mathrm{p}(\mathrm{x}_0)$, we have

$$G_{\mathrm{x}} = 2\left(\frac{\mathrm{p}(\mathrm{x}) + \mathrm{p}(\mathrm{x}_0)}{\mathrm{x} - \mathrm{x}_0} - \frac{\mathrm{b}'(\mathrm{x}) + \mathrm{p}'(\mathrm{x}_0)}{2}\right) = 2\left(\frac{\mathrm{p}'(\mathrm{x}) + \mathrm{p}'(\mathrm{x}_0)}{2} - \frac{\mathrm{b}'(\mathrm{x}) + \mathrm{p}'(\mathrm{x}_0)}{2}\right) = \mathrm{p}'(\mathrm{x}) - \mathrm{b}'(\mathrm{x})$$

where $\mathrm{x} = \mathbb{T}(\mathrm{x}_0)$. The first claim has been shown. To see the second claim, observe first that for $1/2 \leq \mathrm{x}_0 \leq 1$ we have $0 \leq \mathrm{x} \leq 1/2$ and for such x, $\mathrm{b}'(\mathrm{x}) > 0$. On the other hand, when $\mathrm{b}'(\mathrm{x}) \leq 0$ one must have $\mathrm{p}'(\mathrm{x}) \leq \mathrm{b}'(\mathrm{x})$

at the landing point by definition of x. This will be strict inequality unless the contact between the two curves is tangential, in which case $(\mathrm{x}_0, \mathrm{x})$ is laying on the end of one of the monotone branches in the graph of $\mathbb{T}$ where the number of teeth passed takes a discrete jump. Conclude that $G_{\mathrm{x}} = \mathrm{p}'(\mathrm{x}) - \mathrm{b}'(\mathrm{x}) < 0$ except at finitely many points. □

Remark. Lemma 1 already verifies the local continuity of $\mathbb{T}$ away from points x_0 whose landing points x result in tangency of p and b. With a little effort, one can show that, provided we have $\omega^2 > \mathrm{B}$, the landing point cannot be an integer unless $\mathrm{x}_0 = 1$ in which case $\mathrm{x} = 0$; we are again dealing with the endpoint of a branch for the map (in this case the rightmost branch).

Now using Lemma 1, we can obtain the following conclusion about geometric features of each branch in the map $\mathbb{T}$.

Theorem 1. $\frac{d\mathbb{T}}{d\mathrm{x}} \leq 0$ at all but finitely many points in [0,1].

Proof: By the implicit function theorem, we have

$$\left.\frac{d\mathbb{T}}{d\mathrm{x}}\right|_{\mathrm{x}_0} = -\frac{G_{\mathrm{x}_0}(\mathrm{x}_0, \mathrm{x})}{G_{\mathrm{x}}(\mathrm{x}_0, \mathrm{x})}$$

Note that from Lemma 1, except for finitely many points, we have $G_{\mathrm{x}} = \mathrm{p}'(\mathrm{x}) - \mathrm{b}'(\mathrm{x}) < 0$. Since

$$G_{\mathrm{x}_0} = (\mathrm{x} - \mathrm{x}_0)(\mathrm{b}''(\mathrm{x}_0) + \frac{\mathrm{g}}{\mathrm{c}^2})$$

Using bouncing condition, we have $\mathrm{b}''(\mathrm{x}_0) + \frac{\mathrm{g}}{\mathrm{c}^2} \leq 0$. Obviously $\mathrm{x} - \mathrm{x}_0 > 0$, so we have $G_{\mathrm{x}_0} < 0$, then conclude $\frac{d\mathbb{T}}{d\mathrm{x}} \leq 0$ at all such points in [0,1]. □

Another interesting feature observed in Figure 2.5 is the tangency at $\mathrm{x} = 1$. This can also be verified analytically.

Theorem 2. $\lim_{\mathrm{x}_0 \to 1^{-1}} \frac{d\mathbb{T}}{d\mathrm{x}}(\mathrm{x}_0) = 0$.

Proof: For $\frac{1}{2} \ll \mathrm{x}_0 < 1$, we have

$$\frac{d\mathbb{T}}{d\mathrm{x}} = \frac{(\mathrm{x} - \mathrm{x}_0)(\mathrm{b}''(\mathrm{x}_0) + \frac{\mathrm{g}}{\mathrm{c}^2})}{\mathrm{b}'(\mathrm{x}) - \mathrm{p}'(\mathrm{x})}$$

Since $\sqrt{1 - \frac{T^2}{4r^2}} \leq \mathrm{b}(\mathrm{x}) \leq 1$ and $\mathrm{b}''(\mathrm{x}) = -\frac{T^2}{r^2 \mathrm{b}(\mathrm{x})^3}$, $\exists M$, positive constant such that

$$-M < \mathrm{b}''(\mathrm{x}_0) + \frac{\mathrm{g}}{\mathrm{c}^2} < 0 \quad \text{(bounded)}$$

From the projectile map, we know for $\forall\ \mathrm{x}_0 \sim 1$, $\mathrm{b}'(\mathrm{x}) > 0$ and $\mathrm{p}'(\mathrm{x}) < 0 \Rightarrow \mathrm{b}'(\mathrm{x}) - \mathrm{p}'(\mathrm{x}) > \mathrm{b}'(\mathrm{x})$. Since we have $\mathrm{b}'(\mathrm{x}) = \frac{T^2(\frac{1}{2} - \mathrm{x})}{r^2 \mathrm{b}(\mathrm{x})} \geq \frac{T^2(\frac{1}{2} - \mathrm{x})}{r^2}$ and $\mathrm{x}_0 \to 1^-$, $\exists\ \frac{1}{2} > \delta > 0$ such that

$$0 < \mathrm{x} - 1 < \delta \Rightarrow \mathrm{b}'(\mathrm{x}) > \frac{T^2(\frac{1}{2} - \delta)}{r^2} > 0 \quad \text{(bounded below)}$$

Clearly $\mathrm{x} - \mathrm{x}_0 \to 0$ as $\mathrm{x}_0 \to 1^-$, so the result is proved. □

2.3 Drilling Efficiency

2.3.1 Drill-Rock Interaction

The major mechanism for the drill breaking rock is the transfer of kinetic energy as the drill strikes the surface. As the horizontal kinetic energy K_x is maintained by the constant x-velocity c, only the vertical component K_z contributes to rock removal. The vertical kinetic energy is

$$K_z = \frac{m}{2} V_z^2, \tag{2.15}$$

where m is the mass of the drillhead (we normally would include also the borehole machinery and above ground equipment attached to the drillhead in the calculation of m).

Vertical velocity V_z is

$$V_z = c\frac{dp(x_0, x)}{dx}. \tag{2.16}$$

An empirical parameter σ, including the material properties of the rock, drill geometry and more, must be determined experimentally. This parameter is used to convert the kinetic energy into a volume measure of rock broken. The amount of rock broken due to a single impact, in appropriate units is

$$R(x_0) = \frac{\sigma m}{2}\left(c\frac{dp(x_0, x)}{dx}\right)^2 \tag{2.17}$$

and this clearly depends on the landing and launching configurations x and x_0 respectively.

2.3.2 Efficiency Calculation

We choose the following notion of drilling efficiency as a function of speed c, namely, the ratio

$$E(c) = \frac{\text{total rock broken}}{\text{number of revolutions}}. \tag{2.18}$$

Here the assumption is that these quantities are calculated over a long time interval.

If we consider the action of the drill over many bounces (say K bounces), rather than during an interval of time, we can rewrite Equation (2.18) in a more amenable form

$$E(c) = \frac{\text{amount of rock broken after } K \text{ bounces}}{(r/NB)\cdot \text{ number of teeth advanced after } K \text{ bounces}}. \tag{2.19}$$

Here, B is the radius of the bore hole so the factor in the denominator is the inverse of the number of teeth advanced per revolution around the bore hole. By ”teeth advanced per bounce” we mean that, since c is held constant, while the drill is mid-bounce it continues to rotate. The duration of the bounce determines how much rotation will be achieved, or how many teeth will advance during that bounce. In this sense, the duration of a bounce depends on the configuration of the drill when it is launched (see Figure 2.4). Averages in this expression mean with respect to the discrete time events of K launches.

Figure 2.6 shows a number of tooth impacts of a five-toothed drill. The large point is at x_0 = 0.25 and there are 5 impacts at the small points, as 39 teeth pass by, so here $K = 35$.

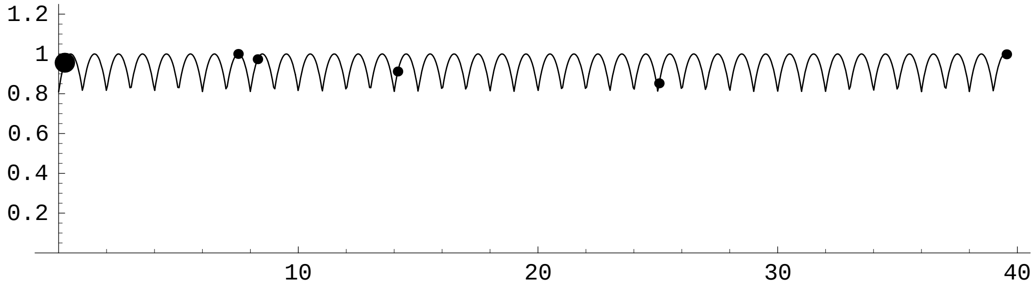


Figure 2.6: Five subsequent impacts from x_0 = 0.25.

The i^{th} impact at x_i is

$$x_i = T^i x_0. \tag{2.20}$$

For K impacts, the average amount of rock broken is

$$\frac{1}{K}\sum_{i=1}^{K} R(T^i x_0). \tag{2.21}$$

Note that the expression is the ergodic average. By Birkhoff's Ergodic Theorem [1], for large K we know this time average is nearly equal to the spatial average

$$\mathrm{Avg}(R) = \int_0^1 R(x_0)\phi(x_0)dx_0. \tag{2.22}$$

The term $\phi(x_0)dx_0 = dm_\phi(x_0)$ represents integration with respect to the invariant measure for the transformation $\mathbb{T}$, so the amount of rock broken after K contacts is estimated by

$$K\int_0^1 R(x)dm_\phi(x). \tag{2.23}$$

Remark. An assumption is being made here that the invariant measure for $\mathbb{T}$ is computed with respect to a density ϕ on $[0, 1]$, i.e., that it is an absolutely continuous invariant measure. We will have more to say about this assumption in the next section.

A similar approach may be taken for denominator in (2.19). In Figure 2.6 it is seen that the number of teeth to pass will depend on the configuration at the launch point. Let $\kappa(x)$ denote the number of teeth to pass between under the trajectory launched at x. Figure 2.7 shows how κ corresponds to the branches in map $\mathbb{T}$.

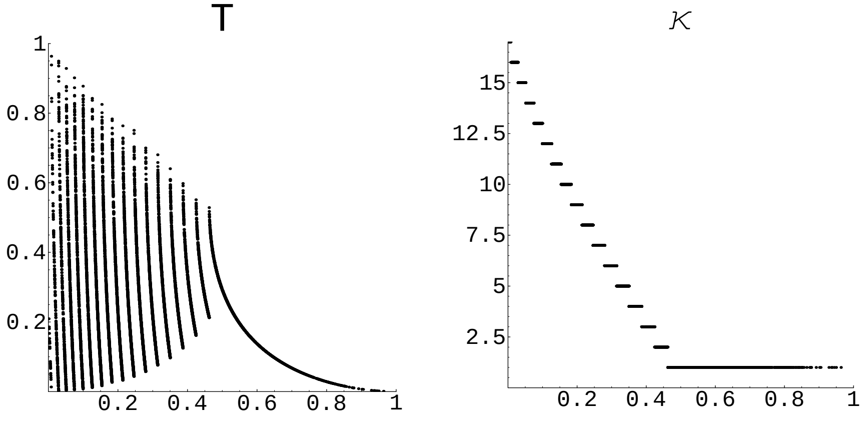


Figure 2.7: Relationship between T and κ.

Once again, using Birkhoff's Ergodic Theorem, the number of revolutions of the drill for K impacts is

$$\frac{K\,\mathrm{Avg}(\kappa)}{N\frac{B}{r}} = \frac{K\int_0^1 \kappa(x)dm_d}{N\frac{B}{r}}, \tag{2.24}$$

where, again, $N\frac{B}{r}$ is the number of teeth that pass for one revolution of the drillhead in the bore hole. Now the efficiency E is

$$E(c) = \frac{NB}{r}\frac{\int_0^1 R(x)dm_d}{\int_0^1 \kappa(x)dm_d}. \tag{2.25}$$

2.4 Dynamical Systems

We are given a map $\mathbb{T} : [0, 1] \to [0, 1]$ and from the previous section we see that it is crucial to work out an invariant measure for $\mathbb{T}$, that is, a measure m, such that for every measurable subset $I \subseteq [0, 1]$

$$m \circ \mathbb{T}^{-1} = m.$$

In good situations we also hope that the measure is absolutely continuous with respect to the Lebesgue measure dx, that is $m = m_\phi$ where

$$m_\phi(I) = \int_I \phi(x)dx,$$

”ϕ is called the (invariant) density for m_ϕ”.

In [2] there is a theorem which ensures that a broad class of mappings (like our $\mathbb{T}$) will have absolutely continuous invariant measures. We note, however, that one of the conditions in the theorem is that $|\mathbb{T}'| \geq \alpha > 1$ and this does not appear to be true for our $\mathbb{T}$ near $x = 1$. However, we can apply the result to a power of $\mathbb{T}$ to obtain the condition, or we can induce the mapping on a subinterval where the derivative condition holds. Since we did not have time to explore these technical issues in detail, we will assume that our map has an invariant density.

We wish, then, to approximate a histogram of the invariant density ϕ and hence to estimate m_ϕ for the given data. We can again use Birkhoff's Ergodic Theorem to conclude that the histogram density value over a bin is proportional to the number of visits of an orbit of the map to that bin. So we should find a suitable interpolation of the map $\mathbb{T}$ in order to compute the long orbit for Birkhoff's Ergodic Theorem.

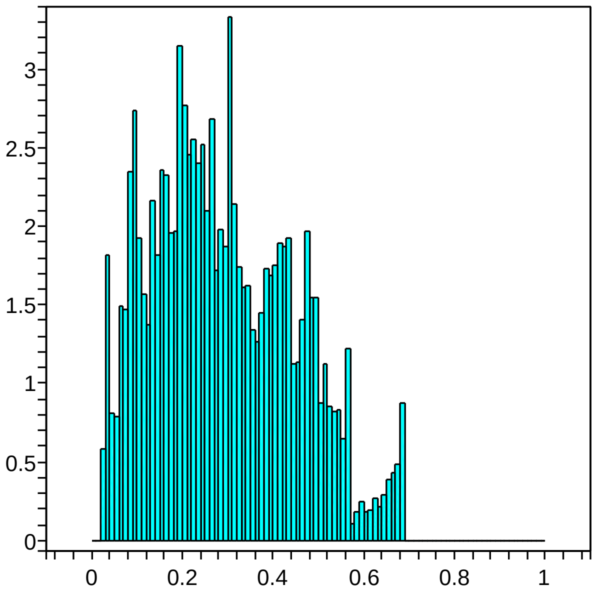


Figure 2.8: Density histogram for N=4.00, r=1.41, $c \cdot T$=10.

Cubic spline interpolation was used in a simple code to produce the histograms and the results were plotted in Figures 2.8 and 2.9. In the first case, where the speed is $c = 10$ and there are 4 teeth we appear to get a density for the map $\mathbb{T}$. In the second plot, another calculation at speed $c = 6$ was made and we notice that the density peaks in a single bin. In this case there are also 4 teeth. It would appear that, for this lower speed value, there is an attracting fixed point for the dynamics near 0.15 so ϕ is a delta function. In fact there is no absolutely continuous invariant measure. Point mass at the fixed point is the only relevant measure and we appear to have the case of non-productive drilling mentioned in the introduction. In Figure 2.10, we can see the number of the jumps in the data($\kappa(x)$).

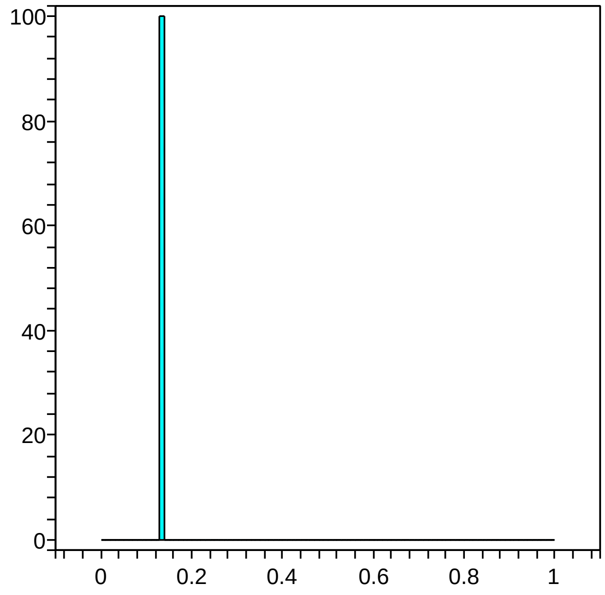


Figure 2.9: Density histogram for N=4.00, r=1.41, $c \cdot T$=6.

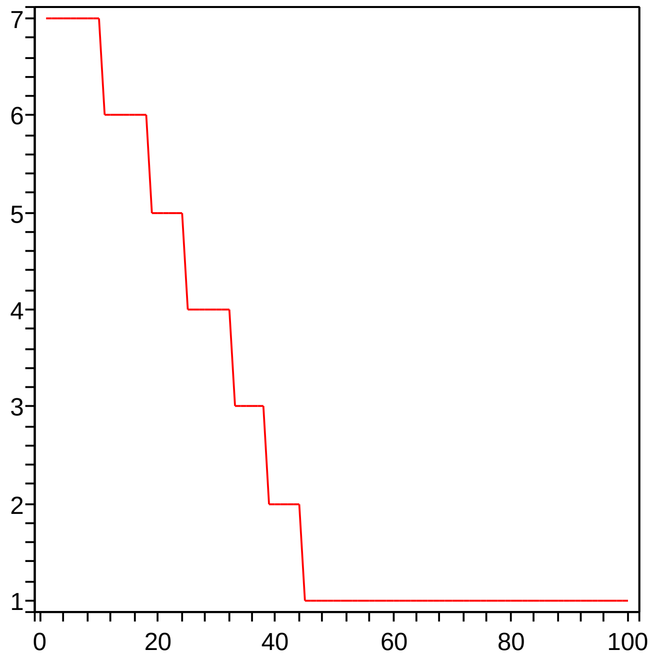


Figure 2.10: The function $\kappa(x)$ for N=4.00, r=1.41, $c \cdot T$=10.

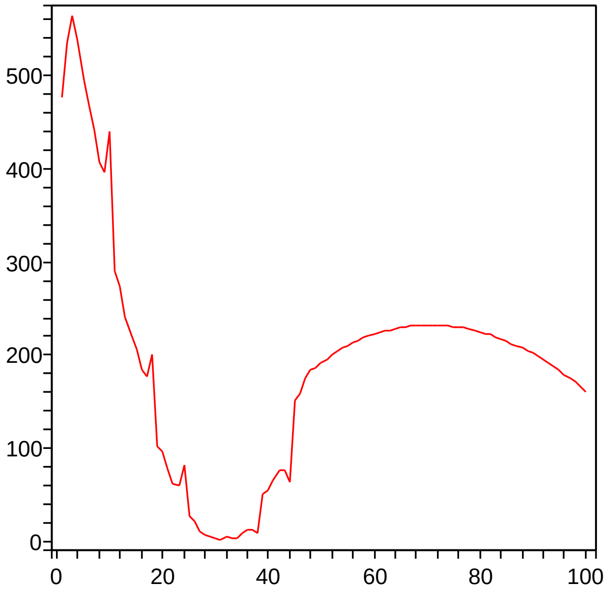


Figure 2.11: The function R(x) for N=4.00, r=1.41, $c \cdot T$=10.

We may plot the amount of rock broken by

$$R(x) = \frac{2\sigma}{m}\left[-\frac{g}{T}\left(f(x) + \kappa(x) - x\right) + c\sqrt{1 - \left(\frac{T}{R}(x - 0.5)\right)^2}\right]^2,$$

where we have

$$T = 1.994041122,$$
$$N = 4.00,$$
$$r = 1.41,$$
$$g = 9.8/r = 6.950354610,$$
$$c = 10/T = 5.014941713,$$
$$\sigma = 0.25,$$
$$B = 10.0,$$
$$m = 100.$$

We should keep in mind that N is the number of teeth, r is the radius of the drill and c is the (renormalized) speed. This is a very big drill! We may approximate the average of rock breaking by

$$\mathrm{Avg}(R) \cong \int_0^1 R(x)\phi(x)dx \cong \sum_{i=1}^{100} R_i d_i,$$

which will lead us to the number 14839.89. And we can also approximate the number of revolutions (teeth) of drill by

$$\mathrm{Avg}(\kappa) \cong \int_0^1 \kappa(x)\phi(x)dx \cong \sum_{i=1}^{100} \kappa_i d_i,$$

which will lead us to the number 394.83. Dividing the two results the efficiency will be found as

$$E(c) = \frac{r}{NB} \frac{\int_0^1 R(x)\phi(x)dx}{\int_0^1 K(x)\phi(x)dx} = 1.32.$$

In order to explore the asymptotic behaviour of this drill, many more runs at different values of c would need to be made and there was insufficient time during the workshop to do this.

2.5 Conclusion

This paper reports on the progress made during the 2005 PIMS Graduate Industrial Mathematics Modelling Camp held during 4 days in May at the University of Lethbridge. The idea behind this model dates back to an old, but obscure publication by A. Lasota and P. Rusek [3] in which a dynamical model of the hard rock drill was presented and its analysis compared to field data for qualitative similarities. In many respects the work of Lasota and Rusek was ad-hoc — nevertheless it gave an interesting application of ergodic theory to a concrete problem and the qualitative agreement with field observations was striking.

The work in this current project mostly retraces the path of Lastoa and Rusek. However, the observations in Section 2 (Geometry Section) are new results that have not been observed before. The numerical work in this project could be extended in at least two directions. First, the codes to estimate the invariant density histogram need to be made faster and more accurate. The cubic spline interpolation, while natural, is likely to introduce too much error for long orbit calculations in expansive systems such as this. Second, the use of Birkhoff's Ergodic theorem to compute the histogram cannot be rigorously justified, even if it is carried out to a very high degree of accuracy. (This is a valid criticism even for Lasota and Rusek's work.) Other methods based on finite-dimensional approximations to the action of $\mathbb{T}$ on the function space L^∞ are robust and quick to converge. In some parts of applied math these are known as Galerkin-type methods. In dynamics they are known as Ulam's method for computing the invariant density. Although these ideas were discussed during the workshop, there was insufficient time to explore a numerical implementation of Ulam's method.

In [4], [5], [6] one can find other treatments of the dynamical drill model. In [7], and [8] one can find state-of-the-art applications of Ulam's method to determination of invariant densities.

Bibliography

[1] P. Walters, An introduction to ergodic theory. Graduate Texts in Mathematics, **79** (1982), Springer Verlag, New York/Berlin.

[2] A. Lasota, J. Yorke, On the existence of invariant measures for piecewise monotone transformations, Trans. AMS, **186** (1973), pp. 481–488.

[3] A. Lasota, P. Rusek, An application of ergodic theory to the determination of the efficiency of cogged drilling bits. Arch. Górnictwa **19** (1974), pp. 281–295.

[4] A. Davidova, The dynamics of a rock drill. Thesis, University of Victoria, 2001.

[5] G. Chakvetadze, A. Stepin, On the dynamical model of drilling. Int. J. of Bifurc. and Chaos. September, 1999.

[6] A. Boyarsky, P Góra, Laws of chaos. Probability and its applications. Birkhauser, 1997.

[7] R. Murray, Existence, mixing and approximation of invariant densities for expanding maps on $\mathbb{R}^r$. Nonlinear Anal. **45** (2001), pp. 37–72.

[8] C. Liverani. Rigorous numerical investigation of the statistical properties of piecewise expanding maps. A feasibility study. Nonlinearity **14** (2001), pp. 463–490.

Chapter 3

Symbols “R” Us
Seismic Imaging, One-Way Wave Equations, Pseudodifferential Operators, Path Integrals, and all that Jazz

Participants: Lou Fishman (Mentor, MDF International), Ojenie Artoun (Concordia University), Diana David-Rus (Rutgers University), Matthew Emmett (University of Calgary), Sandra Fital (University of Regina), Chad Hogan (University of Calgary), Jisun Lim (University of Colorado), Enkeleida Lushi (Simon Fraser University), Vesselin Marinov (Rutgers University)
a.k.a. Lou and the Ψseudo-eight

PROBLEM STATEMENT: In this report we summarize an extension of Fourier analysis for the solution of the wave equation with a non-constant coefficient corresponding to an inhomogeneous medium. The underlying physics of the problem is exploited to link pseudodifferential operators and phase space path integrals to obtain a marching algorithm that incorporates the backward scattering into the evolution of the wave. This allows us to successfully apply single-sweep, one-way marching methods in inherently two-way environments, which was not achieved before through other methods for this problem.

3.1 Introduction

Understanding wavefield propagation and the construction of efficient, approximate numerical algorithms have been important issues in many industrial settings for many years. Mathematicians, physicists, and other applied scientists, in fields as diverse as ocean acoustics, electromagnetics, medical imaging, optical design, and seismic exploration, have struggled with these issues. Wave propagation in complex environments is a classical problem that has been attacked by a variety of approaches over the years, including (1) direct wavefield approximations (e.g., perturbation theory, asymptotic ray theory, Gaussian beams) and (2) computational partial differential equation (PDE) methods (e.g., finite differences, finite elements, spectral methods, wavelets). However, these approaches can often be problematic. Approximation methods often offer some physical insight, but they may be limited to specific parameter regimes, not properly accounting for the underlying physics outside of these regimes. Computational PDE methods, in the frequency domain, in particular, can often be prohibitively expensive and time consuming for large scale problems involving large data sets. Moreover, the computational PDE methods do not, in general, reveal the underlying physics in a transparent manner.

Rather than considering a wide range of wave propagation formulations in this project, we will focus on seismic imaging applications. The problem is illustrated in Figure 3.1, which shows the Marmousi synthetic velocity model for an area off the coast of West Africa. Even though this is a relatively simple, two-dimensional, constant density, scalar model, there is complex fine layering, fault lines, velocities ranging from $1500\ m/s$ to almost $6000\ m/s$, regions with very strong gradients, a reservoir, and regions exhibiting pronounced focusing and defocusing phenomena. The idea is to set off explosive sources along the surface, and from the collected reflection data along the surface, construct an appropriate image of the substructure. In other words, we want to construct images of the Marmousi model, from the synthetic data, that accurately reflect the complex model structure (at least to the extent that this inverse process will allow). In the seismic industry, of course, these reconstructions are done with actual field data sets, which inherently contain limited and noisy data. An image of this model computed in 1998 is illustrated in Figure 3.2.

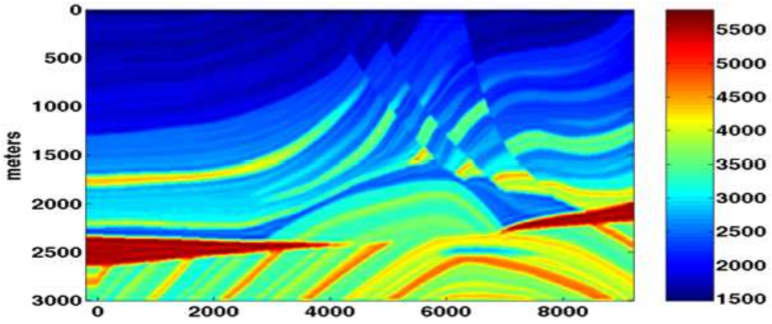


Figure 3.1: Marmousi synthetic velocity geological model.

One of the most commonly used techniques for seismic imaging is seismic depth migration. Kirchhoff migration is based on applying asymptotic ray theory for the wave propagation process. While the workhorse of the seismic industry, and very appealing from a physics standpoint, this high-frequency asymptotic method is inadequate for modelling wave propagation in Marmousi-type environments at the lower end of typical experimental seismic bandwidths (say between $10 - 30\ Hz$). This can be contrasted with wave equation migration methods based on computational PDE methods and the locally-homogeneous-medium (LHM) approximation to the wavefield extrapolator (the Generalized Phase Shift Plus Interpolation (GPSPI) algorithm). Computational PDE methods can often result in excessively large linear systems. The GPSPI algorithm is based on the industry belief that the LHM approximation to the wavefield extrapolator is an exact statement of the global propagation process (at least for range-independent models). This, unfortunately, is not the case outside of a strictly homogeneous medium.

The specific problem for this GIMMC is the following. Construct a wave propagation model, at the level of the fixed-frequency, scalar Helmholtz equation, which can form the basis for wave equation migration algorithms. This wave propagation model should be a full-wave model, so as to be appropriate over the entire seismic

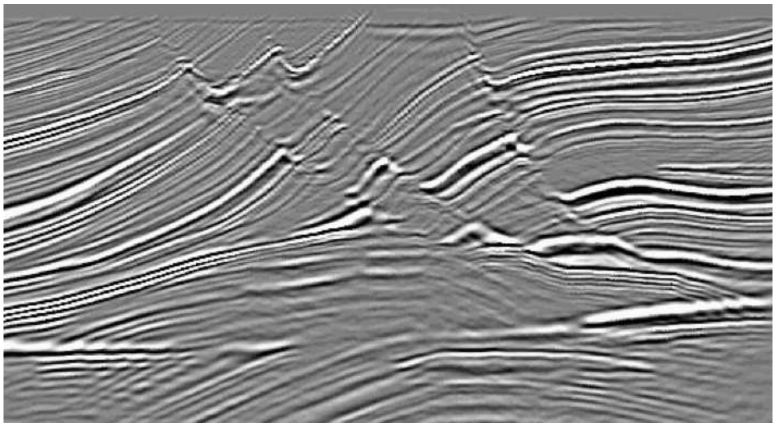


Figure 3.2: Image of Marmousi model computed in 1998. Curtis Ober. 1998, from website (http://www.cs.sandia.gov/ ccober/).

bandwidth, and computationally faster and physically more transparent than PDE-based methods, all the while providing a complete mathematical framework in which to examine, correct, and extend the GPSPI algorithm. It would also be very convenient if the computational realization of this model ran (in principle) exactly in the same manner as the GPSPI algorithm. In other words, we want to combine the physical insight and relative computational speed of GPSPI with the numerical accuracy of a complete PDE-solver-based algorithm. Essentially, we want the best of both worlds for a seismic migration algorithm.

We will approach this problem through the application of what is loosely referred to as “phase space and path integral” methods, which were developed primarily in the quantum physics and theoretical mathematics (PDE) communities. The principal aims of this approach are to (1) exploit well-posed, one-way marching methods in inherently two-way global problems, (2) exploit the correspondences between classical wave propagation, quantum physics, and microlocal analysis (modern mathematical asymptotics), and (3) extend Fourier analysis to inhomogeneous environments. Since this is a brief report, and these “phase space and path integral” methods have a rather involved and extensive literature, we will introduce them essentially through the homogeneous medium case, and rely on the citations in the References section for the detailed story.

3.2 Homogeneous Medium

We begin with the full wave equation which models the propagation of waves in a medium. For simplicity we will work in two dimensions (everything can be extended to arbitrary dimension).

$$\left(\partial_x^2 + \partial_z^2 - \frac{1}{c^2(\vec{x})}\partial_t^2\right)\psi(\vec{x},t) = S(\vec{x}-\vec{x}_s, t-t_0), \tag{3.1}$$

where x is the propagation direction, z is the transverse direction, $c(\vec{x})$ is the speed of sound in the medium, $S(\vec{x},t)$ is a source term, t is time, and $\vec{x} = (x,z)$.

We consider a wavefield $\psi(\vec{x},t) = \phi(\vec{x})e^{-i\omega t}$, where ω is the frequency, and assume that the source $S(\vec{x},t)$ is a delta distribution in both position and time. By taking a Fourier transform in the time variable we obtain the Helmholtz equation, which describes the propagation of a single frequency wave:

$$\left(\partial_x^2 + \partial_z^2 + \bar{k}^2 K^2(\vec{x})\right)\phi(\vec{x}) = -\delta(\vec{x}-\vec{x}_s), \tag{3.2}$$

where $K(\vec{x})$ is the refractive index of the medium at the point $\vec{x}$, and $\bar{k} = \omega/c_0$.

To gain physical insight, we consider a homogeneous medium in which $K(\vec{x}) = K_0$, and solve (3.2) away from the source:

$$\left(\partial_x^2 + \partial_z^2 + \bar{k}^2 K_0^2\right)\phi(x,z) = 0. \tag{3.3}$$

We can formally separate the solution into two non-interacting waves: the right-going wave ϕ^+, and the left-going wave ϕ^-:

$$\left(\frac{i}{\bar{k}}\partial_x + \mathbf{B}\right)\phi^+(x,z) = 0, \tag{3.4}$$

and

$$\left(-\frac{i}{\bar{k}}\partial_x + \mathbf{B}\right)\phi^-(x,z) = 0, \tag{3.5}$$

where

$$\mathbf{B} \equiv \left({K_0}^2 + \frac{1}{\bar{k}^2}\partial_z^2\right)^{1/2}. \tag{3.6}$$

The equations above are a formal factorization of expression (3.3). To make sense of $\mathbf{B}$, consider the Fourier transform of (3.3). The equation for $\phi^+(x,z)$ is as follows:

$$\frac{i}{\bar{k}}\partial_x\phi^+(x,z) + \int_R dp(K_0^2 - p^2)^{1/2}e^{i\bar{k}pz}\hat{\phi}^+(x,p) = 0. \tag{3.7}$$

where $\hat{\phi}^+(x,p)$ is the Fourier transform of $\phi^+(x,z)$. Similarly, we can write an equation for $\phi^-(x,z)$.

It is well known that the fundamental solution to this one way wave equation can be represented essentially as a first order Hankel function of the first kind, which can be expressed by the following identity:

$$\begin{aligned} G^+(x,z|0,z') = \lim_{N\to\infty}\int_{R^{2N-1}} \prod_{j=1}^{N-1} dz_j \prod_{j=1}^{N} \frac{\bar{k}}{2\pi}dp_j \\ \cdot \exp\left[i\bar{k}\sum_{j=1}^{N}\left[p_j(z_j - z_{j-1}) + \frac{x}{N}(K_0^2 - p_j^2)^{1/2}\right]\right]. \end{aligned} \tag{3.8}$$

The fundamental solution representation in (3.8) may appear rather involved for what is essentially a Hankel function identity. However, it is exactly in the form of a phase space path integral representation from quantum physics, and it is this form that will be used to represent the wavefield solutions in the more general inhomogeneous cases. Moreover, the form of (3.8) immediately results in a marching computational algorithm.

We can use the above expression to construct a marching algorithm, as follows:

$$\phi^+(x+\Delta x, z) = \int_R dp e^{i\bar{k}pz}\left[e^{i\bar{k}\Delta x(K_0^2-p^2)^{1/2}}\hat{\phi}^+(x,p)\right]. \tag{3.9}$$

We recognize this as a formula involving forward and inverse Fourier transforms:

$$\phi^+(x+\Delta x, z) = F^{-1}\left[e^{i\bar{k}\Delta x\,(K_0^2-p^2)^{1/2}}F\left[\phi^+(x,z),p\right],z\right]. \tag{3.10}$$

This algorithm is simple, well-posed, and incorporates all the physics of the homogeneous problem.

3.3 Transversely-inhomogeneous Medium

Now we consider a slightly more complex case than the completely homogeneous medium. We introduce dependence of K along the z axis, the transverse direction. Solving this problem will naturally lead us to the solution of the general problem in which we also have dependence on x, the propagation direction.

As in the homogeneous case, right- and left-going waves are decoupled. In both the transversely-inhomogeneous and homogeneous medium cases, there are physical, right- and left-travelling wave fields, as follows from simple separation of variables arguments. An explicit representation of (3.4), written for the transversely-inhomogeneous case, is provided by the theory of pseudodifferential operators from the modern mathematical theory of asymptotics, in addition to more formal operator construction methods developed in the quantum physics community.

These developments enable the explicit construction of nontrivial functions of non-commuting operators, such as the formal square-root operator in (3.6) with the constant index of refraction replaced by its local, transversely-inhomogeneous value. The equation for the right-going wave is given in the Weyl pseudodifferential operator calculus by:

$$\begin{aligned}&\frac{i}{\bar{k}}\partial_x\phi^+(x,z) + \\ &\qquad \frac{\bar{k}}{2\pi}\int_{R^2} dp\,dz'\,\Omega_{\mathbf{B}}\left(p,\frac{z+z'}{2}\right)\exp(i\bar{k}p(z-z'))\phi^+(x,z') = 0,\end{aligned} \tag{3.11}$$

where the operator symbol $\Omega_{\mathbf{B}}$ satisfies the composition equation:

$$\begin{aligned}\Omega_{\mathbf{B}^2}(p,q) &= K^2(q) - p^2 \\ &= \left(\frac{\bar{k}}{\pi}\right)^2 \int_{R^4} dt\,ds\,dy\,dw \\ &\quad \cdot\,\Omega_{\mathbf{B}}(t+p,s+q)\Omega_{\mathbf{B}}(y+p,w+q)\exp(2i\bar{k}(sy-tw)).\end{aligned} \tag{3.12}$$

The form of (3.11) is exactly the same as in the homogeneous case, except that the square root function is replaced by $\Omega_{\mathbf{B}}$. In the homogeneous medium limit, $\Omega_{\mathbf{B}}$ reduces to $\left(K_0^2 - p^2\right)^{1/2}$. This is a representation of the formal square root operator from the Weyl pseudodifferential calculus. The operator is defined through its square, as is the case in functional analysis.

It may seem strange that we are effectively replacing a linear partial differential equation with a quadratically nonlinear, nonlocal composition equation (3.12). However, it will turn out that these symbols $\Omega_{\mathbf{B}}$ can actually image the environments directly. Moreover, approximations made at the level of the symbol will have a far greater range of validity than the corresponding approximations made at the level of the wavefield. These points are discussed in the literature citations.

We have the same path integral and algorithm as the one we described for the homogeneous case. The difference is that $(K_0^2 - p^2)^{1/2}$ is replaced by:

$$h^s_{\mathbf{B}}(p,q) = \left(\frac{\bar{k}}{\pi}\right)\int_{R^2} ds\,du\,\Omega_{\mathbf{B}}(s,u)e^{-2i\bar{k}(q-u)(p-s)}. \tag{3.13}$$

While exact solutions for specific profiles $K^2(z)$ have been constructed, applications to seismo-acoustic wave propagation, imaging, and inversion will depend upon uniform asymptotic expansions of the operator symbol. For $\Omega_{\mathbf{B}}(p,q)$ this takes the general form:

$$\Omega_{\mathbf{B}}(p,q) \sim \left(K^2(q) - p^2\right)^{1/2} + \text{uniform terms through} O(1/\bar{k}^2), \tag{3.14}$$

in the high-frequency ($\bar{k} \to \infty$) limit. The details of the exact solution constructions and the explicit form and derivation of (3.14) are best left to the literature citations. The most important point is the following. While the modern mathematical theory of asymptotics and constructions from quantum physics provide the necessary mathematical framework to explicitly derive our equations, the asymptotic solution of the composition equation (3.12) lies outside of both of these areas. The reason for this is, essentially, that these theories and examples from quantum physics are based on propagation of singularities arguments (think time domain wave propagation), while the Helmholtz equation is a smoothing problem (think frequency domain wave propagation). Constructions like (3.14) require going beyond the modern mathematical theory of asymptotics and examples from quantum physics.

Assuming that we can compute $\Omega_{\mathbf{B}}$ numerically, or asymptotically approximate it, then we can use it in the following marching algorithm:

$$\phi^+(x+\Delta x,z) \approx \int_R dp\,\exp(i\bar{k}pz)\left[\exp(i\bar{k}\Delta x\,h^s_{\mathbf{B}}(p,z))\hat{\phi}^+(x,p)\right], \tag{3.15}$$

or, in Fourier transform notation:

$$\phi^+(x+\Delta x,z) \approx F^{-1}\left[\exp(i\bar{k}\Delta x\,h^s_{\mathbf{B}}(p,z))F\left[\phi^+(x,z),p\right],z\right]. \tag{3.16}$$

3.4 Generally-inhomogeneous Medium

In the homogeneous and transversely-inhomogeneous media we were able decouple the right- and left-going waves. We were able to formally factor the Helmholtz equation. However, in a generally-inhomogeneous medium, in which K depends on both the transverse and propagation directions, the right- and left-going waves are inherently coupled. In a general inhomogeneous medium, backscattering may be present – the problem becomes global in nature (see Figure 3.3) . The problem is inherently two-way. It is straightforward to construct one-way wave equations in the transversely-inhomogeneous and homogeneous cases since the physics already gives us independent one-way wave equations. But how can we construct one-way wave equations in this general case? The answer lies in considering the general scattering problem for the geometry of Figure 3. Applying invariant embedding methods enables the construction of the reflection and transmission operators associated with the generally-inhomogeneous block. Transforming this scattering picture into a boundary-value picture, in terms of Dirichlet-to-Neumann operators and the total wavefield and its normal derivative, produces the desired result. We get a one-way equation to construct the operator, and a one-way wave equation, in terms of the operator, governing the total wavefield. The detailed treatment can be found in the literature citations.

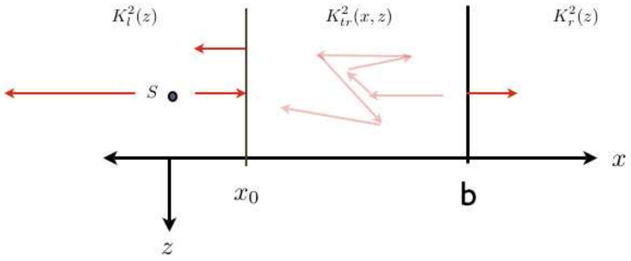


Figure 3.3: Schematic of inhomogeneous block.

In a similar manner to the homogeneous and transversely-inhomogeneous models, the right-going field is governed by:

$$\left(\frac{1}{\bar{k}}\partial_x + \mathbf{\Lambda}^+(x,b)\right)\phi(x,z) = 0, \tag{3.17}$$

(and similarly for the left-propagating field). From (3.17) we note that $\mathbf{\Lambda}^+$ is the Dirichlet-to-Neumann operator. In fact the square root operator $\mathbf{B}$ is the transversely-inhomogeneous DtN operator as well.

The one-way wave equation is:

$$\begin{aligned}&\frac{1}{\bar{k}}\partial_x\phi(x,z) + \\ &\qquad \frac{\bar{k}}{2\pi}\int_{\mathbb{R}^2} dp\,dz'\,\Omega_{\mathbf{\Lambda}^+}\left(x,b;p,\frac{z+z'}{2}\right)\exp(i\bar{k}p(z-z'))\phi(x,z') = 0\end{aligned} \tag{3.18}$$

where $\Omega_{\mathbf{\Lambda}^+}$ is the symbol in the Weyl calculus corresponding to $\mathbf{\Lambda}^+$.

The $\mathbf{\Lambda}^+(x,b)$ operator encapsulates the two-way scattering behaviour of the block between x and b. That is, $\mathbf{\Lambda}^+(x,b)$ contains all information about the block's internal reflections and transmissions between x and b. By solving the series of internal scattering problems that the wavefield encounters as it propagates into the block, the inherently coupled problem is decoupled into independent wavefield components, resulting in the desired one-way wave equation. We march one way (from b to x) to construct $\mathbf{\Lambda}^+$ (which describes the two-way wave behaviour), and then we march the other way to obtain the total wavefield.

A short, formal derivation of the one-way equation defining the DtN operator is the following. Taking the partial derivative with respect to x of equation (3.17), applying (3.17), and using the original Helmholtz equation result in the marching equation for $\mathbf{\Lambda}^+$:

$$\frac{1}{\bar{k}}\partial_x \mathbf{\Lambda}^+(x,b) = \left(\mathbf{\Lambda}^+(x,b)\right)^2 + \mathbf{B}^2(x), \tag{3.19}$$

with the initial condition: $\mathbf{\Lambda}^+(b,b) = -i\mathbf{B}(b)$.

The composition equation for $\Omega_{\mathbf{\Lambda}^+}$ is:

$$\begin{aligned}&\frac{1}{\bar{k}}\partial_x \Omega_{\mathbf{\Lambda}^+}(x,b;p,q) = \\ &\quad \left(\frac{\bar{k}}{\pi}\right)^2 \int_{\mathbb{R}^4} dt\, ds\, dy\, dw\, \Omega_{\mathbf{\Lambda}^+}(x,b;t+p,s+q) \\ &\quad \cdot \Omega_{\mathbf{\Lambda}^+}(x,b;y+p,w+q)\exp(2i\bar{k}(sy-tw)) + K^2(x,q) - p^2,\end{aligned} \tag{3.20}$$

with initial condition: $\Omega_{\mathbf{\Lambda}^+}(b,b;p,q) = -i\Omega_{\mathbf{B}}(b;p,q)$. Equation (3.20) is just the expression of (3.19) in the Weyl pseudodifferential operator calculus.

As in the transversely-inhomogeneous case, we avoid solving the composition equation (3.20) by applying uniform asymptotics. While computational algorithms exist that solve nonlocal Riccati-type equations like (3.20), the uniform asymptotic approximation of the operator symbol is the key to a computationally feasible algorithm. More details can be found in the literature citations.

We have extended the homogeneous Fourier analysis to the general inhomogeneous case. The path integral representation of the fundamental solution is:

$$\begin{aligned}G^+(x,z|x_0,z') = \lim_{N\to\infty}\int_{\mathbb{R}^{2N-1}} \prod_{j=1}^{N-1} dz_j \prod_{j=1}^{N} \frac{\bar{k}}{2\pi} dp_j \\ \cdot \exp\left[i\bar{k}\sum_{j=1}^{N}\left[p_j(z_j - z_{j-1}) + i\frac{\Delta x}{N} h^s_{\mathbf{\Lambda}^+}(x_j,b;p_j,z_j)\right]\right],\end{aligned} \tag{3.21}$$

where

$$h^s_{\mathbf{\Lambda}^+}(x,b;p,z) = \frac{\bar{k}}{\pi}\int_{\mathbb{R}^2} ds dt \Omega_{\mathbf{\Lambda}^+}(x,b;s,t)\exp\left(-2i\bar{k}(q-t)(p-s)\right). \tag{3.22}$$

We use $h^s_{\mathbf{\Lambda}^+}(x,b;p,z)$, the symbol in the standard calculus, because it is computationally more efficient. Again, the marching computational algorithm takes the same form as the homogeneous medium construction.

This concludes our approach to modelling the two-way wave propagation problem using a one-way marching algorithm.

3.5 Discussion

We have introduced one-way methods into an inherently two-way problem. In a generally-inhomogeneous medium, even though the source wave is a one-way wave, once it enters an inhomogeneous environment, internal scattering produces a two-way wavefield. Our model describes these complex behaviours through a pseudodifferential operator, which allows us to employ a one-way marching algorithm to numerically solve for the wavefield.

Throughout the theory examples from various fields of physics have guided us. Wavefield splitting, invariant embedding, and phase space (Weyl pseudodifferential, Fourier integral operator and path integral) methods were used to reformulate the problem in terms of an operator that includes all scattering physics. There are both physically insightful and mathematically succinct ways to derive governing equations for this scattering operator. We have chosen to present this material as reasonable generalizations of the easily derived results for a homogeneous medium. The results, in effect, provide an extension of Fourier methods to inhomogeneous environments.

The methods and algorithms, especially the uniform asymptotic expansion of the DtN operator, mentioned in this report, are broadly applicable in many different areas of industry as outlined previously.

The developments in this paper provide the framework for seismic depth migration imaging. The uniform asymptotic approximations of the operator symbols extend the well known GPSPI algorithm. Figure 3.4 illustrates the Marmousi image produced with the leading-order square root function term alone (GPSPI). The inclusion of the uniform terms implied in (3.14) will result in even greater resolution.

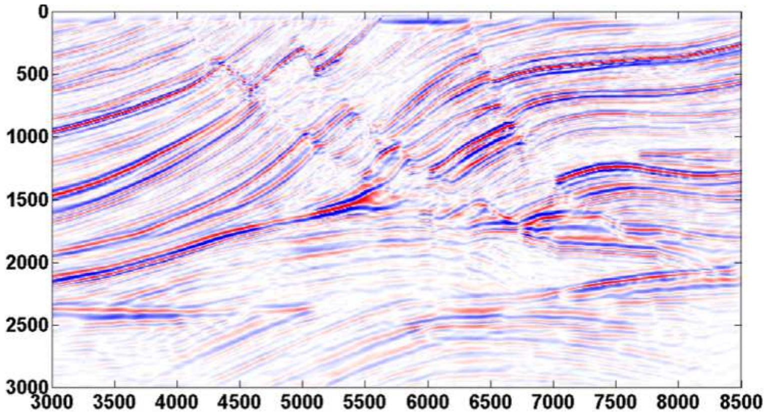


Figure 3.4: Image of Marmousi model computed using the leading term of the asymptotic approximation. Hugh Geiger, personal communication.

3.6 Acknowledgements

This work was done with the support of the Pacific Institute for the Mathematical Sciences (PIMS) at the 8th PIMS Graduate Industrial Mathematics Modelling Camp (GIMMC) at the University of Lethbridge, Alberta. The Ψseudo-eight would like to thank: PIMS, the University of Lethbridge, our excellent mentors, especially Lou Fishman, Hadi Kharaghani, Elena Braverman, Marian Miles, Gary Margrave and Hugh Geiger.

Bibliography

[1] Fishman, L., Gautesen, A. K., and Sun, Z., September 1997, Uniform high-frequency approximations of the square root Helmholtz operator symbol: Wave Motion, **26**, 127–161.

[2] Fishman, L., de Hoop, M. V., and van Stralen, M. J. N., July 2000, Exact constructions of square-root Helmholtz operator symbols: The focusing quadratic profile: Journal of Mathematical Physics, **41**, no. 7, 4881–4938.

[3] Fishman, L., May 1992, Exact and operator rational approximate solutions of the Helmholtz, Weyl composition equation in underwater acoustics – the quadratic profile: Journal of Mathematical Physics, **33**, no. 5, 1887–1914.

[4] Fishman, L., 2002, Applications of directional wavefield decomposition, phase space, and path integral methods to seismic wave propagation and inversion: Pure and Applied Geophysics, **159**, 1637–1679.

[5] Gochiocco, E. L. M., and Brzostowski, M., June 2005, Special section on migration: The Leading Edge, pages 601–654.

[6] Margrave, G. F., and Ferguson, R. J., July-August 1999, Wavefield extrapolation by nonstationary phase shift: Geophysics, **64**, no. 4, 1067–1078.

Chapter 4

Problems in Facility Location Optimization

Participants: Daya Gaur (Mentor, University of Lethbridge), Zac Friggstad (University of Lethbridge), John Gonzalez (Northeastern University), Chong Liu (University of Victoria), Maryam Mizani (University of Victoria), Hatesh Radia (University of Massachusetts, Lowell), Jihong Ren (University of British Columbia), Dallas Thomas (University of Lethbridge), Pengpeng Wang (Simon Fraser University), Liang Xu (University of Washington), Oulu Xu (York University), Yaling Yin (University of Saskatchewan), Zhidong Zhang (University of Saskatchewan)

PROBLEM STATEMENT: In the general setting of the Facility Location (FL) problem, we have a collection of N facilities $\mathcal{F}$ with opening cost f_i for facility i, and a collection of M clients $\mathcal{C}$. Client j also has a connection cost c_{ij} to facility i for being served. Alternatively, we use the notation of $i \in \mathcal{F}$ ($j \in \mathcal{C}$) to denote facility i (client j). The facility location problem refers to opening a subset of facilities and assigning each client with at least one open facility such that the total cost, the sum of the facility opening costs and the connection costs corresponding to the facility client assignments, is minimum. In [1], for a special case of FL with the metric assumption[1] for connection costs, Jain and Vazirani used a primal dual schema based algorithm to achieve an approximation ratio of 3. In this modelling camp, we considered several of these variants while relaxing metric constraints, through application and performance analysis based on algorithms utilizing the primal dual schema. The first problem addressed the case where the connection costs are within a certain range. From there focus shifted to where connection costs are either zero or infinity and opening costs are one. Lastly, we compared the performances of these algorithms with heuristic based algorithms such as local search.

List of Acronyms:

CSC — Complement Slackness Conditions
DLP — Dual Linear Program
FL — Facility Location Problem
IP — Integer Program
LP — Linear Program

[1] Metric assumption here refers to that for every two facilities i and i' and two clients j and j', the connection costs c_{ij}, $c_{i'j}$, $c_{i'j'}$ and $c_{ij'}$ satisfy $c_{ij} + c_{i'j} + c_{i'j'} \geq c_{ij'}$.

4.1 Introduction

In the general setting of the Facility Location (FL) problem, we have a collection of N facilities $\mathcal{F}$ with opening cost f_i for facility i, and a collection of M clients $\mathcal{C}$. Client j also has a connection cost c_{ij} to facility i for being serviced. Alternatively, we use the notation of $i \in \mathcal{F}$ ($j \in \mathcal{C}$) to denote facility i (client j). The facility location problem refers to opening a subset of these facilities and assigning each client with at least one open facility such that the total cost, the sum of the facility opening costs and the connection costs corresponding to the facility client assignments, is minimum. With the introductions of the integral binary variables y_i to denote whether to open facility i ($y_i = 1$) or not ($y_i = 0$), and x_{ij} to denote whether facility i is assigned to client j ($x_{ij} = 1$) or not ($x_{ij} = 0$), the Integer Program (IP) formulation, of the facility location problem is given by:

$$\begin{aligned}
\textbf{IP}\text{: Minimize} \quad & \sum_{i\in\mathcal{F}} f_i y_i + \sum_{i\in\mathcal{F},j\in\mathcal{C}} c_{ij}x_{ij} \\
\text{Subject to:} \quad & \forall j \in \mathcal{C}: \ \sum_{i\in\mathcal{C}} x_{ij} \geq 1 \\
& \forall j \in \mathcal{F}, \forall i \in \mathcal{C}: \ \ y_i \geq x_{ij} \\
& y_i, x_{ij} \in \{0,1\}
\end{aligned} \tag{4.1}$$

FL is a classic combinatorial optimization problem and as stated in [1], received much exposure as early as the 1960's. The NP-hardness [2] nature of this problem prevents a fast (in polynomial time) algorithm for an optimal solution. Therefore efforts were made for a fast approximation algorithm. (*Approximation* here refers to the algorithmic performance ratio with respect to the optimal solution in the worst case.) In the seminal paper [1], Jain *et. al.* used a Primal-dual schema based algorithm (running in polynomial time) to achieve a 3 approximation ratio for the Metric Facility Location problems, assuming connection costs satisfied triangle inequality properties. The idea was to first relax the IP formulation to a linear program (LP) and derive its dual linear program (DLP). Working in DLP, while maintaining dual feasibility, a feasible primal solution can be derived using the Complementary Slackness Conditions (CSC). The metric assumption is then used to prove the performance ratio of 3. Note that the metric assumption is used heavily in their analysis.

In this workshop, we investigated the power of the primal dual schema without the metric assumptions, i.e., we applied the primal dual schema based algorithms to two variants of FL problem without metric assumptions and analyzed the performances of the solutions. Specifically, first we looked into the case where the connection costs are bounded, or within a certain range. We show that in this special case, even under the condition that the triangle inequality property is violated, the primal-dual algorithm can still be applied and achieve a constant ratio performance provided that some suitably defined ratio is bounded. Second, motivated by industrial applications, we looked into another special case of FL where the connection costs are either 0 or ∞, i.e., either instantly connected or disconnected. For this we show that the primal-dual algorithm of Jain et. al. [1] cannot achieve a constant performance ratio via a counterexample. Third, we empirically compared the performance of the primal dual algorithm with a simple heuristic based algorithm, i.e., the local search algorithm, for some difficult problem instances. We show that for these instances, the local search algorithm always performs better. This also validates the popular usage of local search in common practices.

In the following sections, we discuss the problems, the primal dual algorithm from [1], and the analysis of the algorithm on our classes of problems.

4.2 Bounded Connection Costs

As described in the introduction, this case corresponds to connection costs within a certain range, $[c_{min}, c_{max}]$, while relaxing the metric assumption of triangle inequality. By definition, we have
$c_{min} = \min_{i,j}\{c_{ij}\}$ and $c_{max} = \max_{i,j}\{c_{ij}\}$. In the following, we give the formal problem formulation, the description of the applied primal dual algorithm, and an analysis of performance ratio.

[2] It is generally believed that NP-hard problems do not admit polynomial time algorithms.

4.2.1 Problem Definition

In the integer programming (IP) problem form, this variant FL problem is given as follows. Let y_i be 1 if facility i is open and 0 otherwise. Also, let x_{ij} be 1 if facility i is connected to client j and equal to 0 otherwise. Then this problem is formulated as:

$$\begin{aligned}
\textbf{IP}\text{: Minimize} \quad & \sum_{i\in\mathcal{F}} f_i y_i + \sum_{i\in\mathcal{F},j\in\mathcal{C}} c_{ij}x_{ij}, \quad c_{ij}\in[c_{min},c_{max}], \forall i\in\mathcal{F}, j\in\mathcal{C} \\
\text{Subject to:} \quad & \forall j\in\mathcal{C}: \sum_{i\in\mathcal{C}} x_{ij} \geq 1 \\
& \forall j\in\mathcal{F}, \forall i\in\mathcal{C}: \ y_i \geq x_{ij} \\
& y_i, x_{ij} \in \{0,1\}
\end{aligned} \tag{4.2}$$

Note that the only difference (in formulation) between this bounded case, (4.2), and the general form, (4.1) in the introduction, is the condition $c_{ij}\in[c_{min},c_{max}], \forall i\in\mathcal{F}, j\in\mathcal{C}$ in (4.2).

As mentioned, we relax this IP formulation into a LP setting by relaxing y_i and x_{ij} to be positive reals, and the LP formulation is given as:

$$\begin{aligned}
\textbf{LP}\text{: Minimize} \quad & \sum_{i\in\mathcal{F}} f_i y_i + \sum_{i\in\mathcal{F},j\in\mathcal{C}} c_{ij}x_{ij}, \quad c_{ij}\in[c_{min},c_{max}], \forall i\in\mathcal{F}, j\in\mathcal{C} \\
\text{Subject to:} \quad & \forall j\in\mathcal{C}: \sum_{i\in\mathcal{C}} x_{ij} \geq 1 \\
& \forall j\in\mathcal{F}, \forall i\in\mathcal{C}: \ y_i \geq x_{ij} \\
& y_i, x_{ij} \geq 0
\end{aligned} \tag{4.3}$$

Using α_j to denote the dual variable associated with each client and β_{ij} associated with each connection between $i\in\mathcal{F}$ and $j\in\mathcal{C}$, corresponding to each constraint above, the dual linear program (DLP) is given as:

$$\begin{aligned}
\textbf{DLP}\text{: Maximize} \quad & \sum_{j\in\mathcal{C}} \alpha_j \\
\text{Subject to:} \quad & \forall i\in\mathcal{F}, \forall j\in\mathcal{C}: \ \alpha_j - \beta_{ij} \leq c_{ij} \quad c_{ij}\in[c_{min},c_{max}] \\
& \forall j\in\mathcal{F}: \sum_{j\in\mathcal{C}} \beta_{ij} \leq f_i \\
& \alpha_j, \beta_{ij} \geq 0
\end{aligned} \tag{4.4}$$

α_j can be interpreted as the price that client j would like to 'pay' for the facility opening cost, f_i, and the connection cost, β_{ij} [1].

4.2.2 Algorithm Description

Since we are using the same algorithm as in [1], here we briefly describe the algorithm and refer the reader to [1] for details.

The primal dual schema based algorithm has two phases. In Phase I, we first initialize all the dual variables to be 0, $\alpha_j = 0, j\in\mathcal{C}$ and $\beta_{ij} = 0, i\in\mathcal{F}, j\in\mathcal{C}$. We then increase all the $\alpha_j, j\in\mathcal{C}$, simultaneously until one α_j is equal to one connection cost c_{ij}. (This connection is the one that connects client j to facility i.) We say facility i and client j are *directly* connected, and then start increasing β_{ij} (starting from 0) simultaneously with the α_j. This guarantees the first set of constraints in DLP, $\alpha_j - \beta_{ij} \leq c_{ij}$, are always satisfied. At some point, when the sum of β_{ij} associated with facility i is equal to the opening cost f_i, i.e., $\sum_{j\in\mathcal{C}} \beta_{ij} = 1$, we *temporarily* open facility i, and freeze (stop increasing) all the α_i's that are connected to it. (Note that this will automatically freeze the nonzero β_{ij}'s.) We do this until all the clients are connected to at least one temporarily open facility.

In Phase II, we collect all the temporarily open facilities in the order of their openings and *connect* two of them if the two serve to one common client. If the connecting facility i of a client j is temporarily open and connects another temporarily open facility i', we define the nonexisting connection between j and i' to be an *indirect connection*. We then extract the maximal independent set from this temporarily open facility set based on their opening order. That is, we pick the first temporarily open facility, remove all the other ones connecting to it (sharing common connected clients with it), and pick the next facility until all are picked or removed. We *open* all the picked facilities, i.e., the corresponding $y_i = 1$, and assign them to the clients that are *directly* connected to it, i.e., the corresponding $x_{ij} = 1$. For those clients that are not directly connected to the open facilities, we assign the *indirectly* connected open facilities to them[3].

4.2.3 Performance Analysis

As mentioned, if we assume the connection costs comply with the triangle inequality, the above algorithm guarantees the approximation ratio of 3. Here we only have bounds on the connection costs. However, we are still able to show, in the following theorem, the performance ratio is bounded by $\frac{c_{max}}{c_{min}}$, the ratio between the upper and lower bounds of connection costs.

Theorem 1: Let x^*_{ij} and y^*_i denote the output of the algorithm described above, and OPT denote the optimal value for this problem. Then

a)

$$\sum_{i\in\mathcal{F},j\in\mathcal{C}} c_{ij}x^*_{ij} + \sum_{i\in\mathcal{F}} f_i y^*_i \leq \frac{c_{max}}{c_{min}} OPT.$$

b) The upper bound in a) is tight.

Following [1], we interpret dual variable α_j as the price that client j would like to pay to construct the feasible primal solution. The price includes the connection cost for the opened facility i assigned to client j and the opening cost of that facility. We use α^e_j and α^f_j to denote the two parts of α_j that contribute to the connection cost to and the opening cost of, the facility client j respectively. Thus, $\alpha^f_j = \alpha_j - c_{ij}$ and $\alpha^e_j = c_{ij}$. The crucial fact needed for the proof is given in the following lemma. Its proof is just a slight modification of the proof of Lemma 5 in [1].

Lemma: If client j is indirectly connected to facility i then

$$c_{ij} \leq \frac{c_{max}}{c_{min}}\alpha^e_j.$$

Proof From the proof of Lemma 5 in [1], there exists a facility i' and a client j' such that

$$3\alpha_j \geq c_{i'j} + c_{ij'} + c_{i'j'} \geq 3c_{min}.$$

Thus,

$$\alpha_j \geq c_{min} = c_{max}\frac{c_{min}}{c_{max}}$$

and hence

$$c_{ij} \leq c_{max} \leq \frac{c_{max}}{c_{min}}\alpha_j = \frac{c_{max}}{c_{min}}\alpha^e_j$$

Proof of Theorem 1

a) For a directly connected client j we have

$$c_{ij} = \alpha^e_j \leq \frac{c_{max}}{c_{min}}\alpha^e_j.$$

Combining this inequality with the result from the above lemma we have

[3] Note that each client must have either a direct or an indirect connection to an open facility, because otherwise the first temporarily opened facility that directly connects it must be opened.

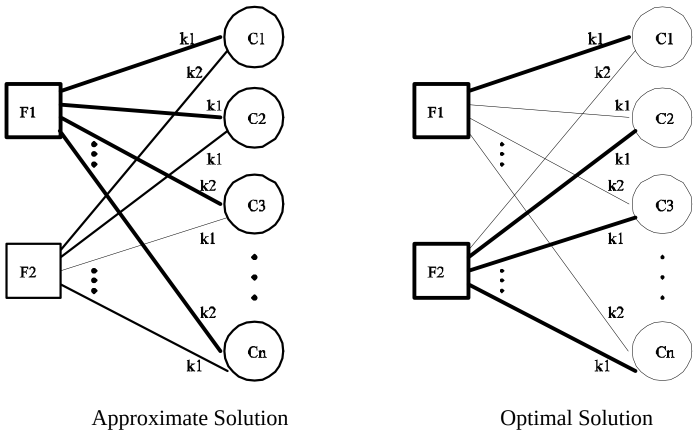
$$\sum_{i\in\mathcal{F},j\in\mathcal{C}} c_{ij}x^*_{ij} \le \frac{c_{max}}{c_{min}} \sum_{j\in\mathcal{F}} \alpha^e_j.$$

Since the opening cost of each open facility are ‘paid’ by those clients that *directly connects* to it, we have $\sum_i f_i \le \sum_j \alpha^f_j$. (This simple fact is called Corollary 4 in [1].)

So we have,

$$\begin{aligned}\sum_{i,j} c_{ij}x^*_{i,j} + \sum_i f_i y^*_i &\le \frac{c_{max}}{c_{min}} \sum_j (\alpha^e_j + \alpha^f_j) \\ &= \frac{c_{max}}{c_{min}} \sum_j \alpha_j \\ &\le \frac{c_{max}}{c_{min}} OPT\end{aligned}$$

b) Here we construct a family of examples that establish that the ratio c_{max}/c_{min} is tight.



Approximate Solution

Optimal Solution

Figure 4.1: A tight example that the performance of the primal dual algorithm approaches $\frac{k_2}{k_1}$.

The construction is as follows, shown in Fig. 4.1. For n clients ($n > 3$), $c_1, c_2, \ldots, c_n$ and 2 facilities F_1, F_2, the connecting costs satisfy the following: $c_{ij} \in \{k_1, k_2\}$, where $k_1 < k_2$; $c_{11} = c_{12} = c_{2l_1} = k_1$ for $2 \le l_1 \le n$, and $c_{1l_2} = c_{21} = k_2$ for $3 \le l_2 \le n$. Furthermore, the two facilities have equal opening costs $f_1 = f_2 = f > 0$ and $\frac{f}{n-1} < k_2 - k_1$ and $\frac{f(n-2)}{n-1} < k_2 - k_1$.

The optimal solution for this setup is to open both facilities and connect every client with the one facility for which the connecting cost is k_1. The corresponding optimal cost is $2f + nk_1$. On the other hand, the solutions given by our algorithm yield the end result of opening only facility 1 and connecting all clients to it. And the corresponding overall cost is $f + 2k_1 + (n-2)k_2$.

This implies the performance ratio is $\frac{f+2k_1+(n-2)k_2}{2f+nk_1}$. As n approaches infinity, this ratio approaches k_2/k_1 in the limit. This implies the approximation ratio, $\frac{c_{max}}{c_{min}}$, stated in the above theorem, is tight.

In conclusion, for this variant of FL in which the connection costs are within a range, we have shown the application of the primal dual algorithm achieves a tight performance ratio, the ratio between upper and lower bounds of the connection costs. Further along this line, as a generalization, we also considered an LP version where the constraints $x_{ij}, y_i \in \{0, 1\}$ are replaced by $x_{ij}, y_i \ge 0$. The optimal solution in this case will necessarily have $x_{ij}, y_i \le 1$. Therefore the x_{ij} and y_i variables represent “partially connecting” and “partially opening”. We would also like to investigate the approximation ratio of a generalized version of the primal dual algorithm for this problem.

4.3 Line Facility and Point/Rectangular Client in 2D

The second problem we considered was motivated by industrial applications. Consider an airplane hangar with randomly dispersed workshops of different sizes, Fig. 4.2. These specific workshops need to be serviced with tools and parts by cranes which are fixed on the ceiling of the workshop. To reduce the construction costs of these cranes, our goal is to reduce the number of cranes needed to service all the workshops.

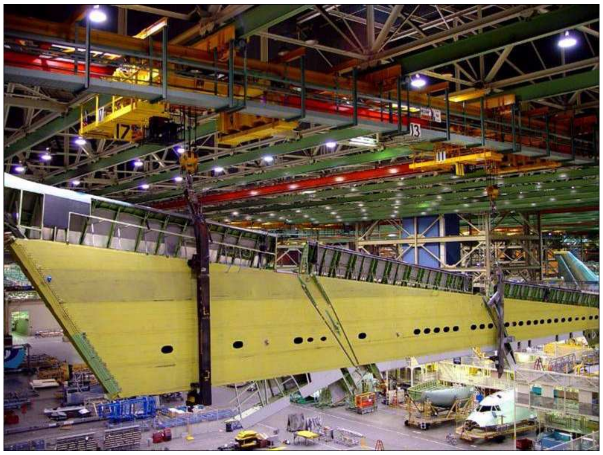


Figure 4.2: An Example.

To further motivate the problem formulation, let us see an abstract version of the problem as shown in Fig. 4.3. This is a 2D "warehouse" setting with scattered workshops, the rectangles, and possible line facility positions, horizontal and vertical lines. We would like to choose (construct) the minimum number of lines, a subset of all the possible lines, and at the same time, still be able to service all the clients, i.e., each client has to have at least one chosen line servicing (going through) it.

In the following, we are going to formulate the program using IP formulations, apply the primal dual algorithm as before, and analyze its performance ratio. First, we consider the case where the clients can be modelled as points. Next we consider the case where we have rectangular clients.

4.3.1 Point Client Case

Problem Definition

For line facilities (restricted to be either horizontal or vertical) specified by either the x coordinate (vertical lines) or the y coordinate (horizontal lines), we denote the line set by $\mathcal{L} = \{l_1, l_2, l_3, \ldots, l_N\}$. For the point workshop i, or point client, specified by $(x(i), y(i))$ ($x(i)$ and $y(i)$ give the $x-$ and $y-$ coordinates of point client i), we denote it by the line labels in $\mathcal{L}$ corresponding to its coordinates respectively, i.e., $(x(i), y(i)) \leftrightarrow (l_{x(i)}, l_{y(i)})$. We denote the point client set by $\mathcal{C} = \{c_1, c_2, \ldots, c_M\}$, where $c_i = (l_{x(i)}, l_{y(i)})$. Note that in the point case, each point has only two corresponding lines.

The problem is to select a subset of $\mathcal{L}$ such that each point workshop is either serviced by the horizontal line or the vertical line going through it.

In the Integer Program (IP) setting, by using $z_k \in \{0, 1\}$ to denote whether line k is open (1) or not (0), we will have

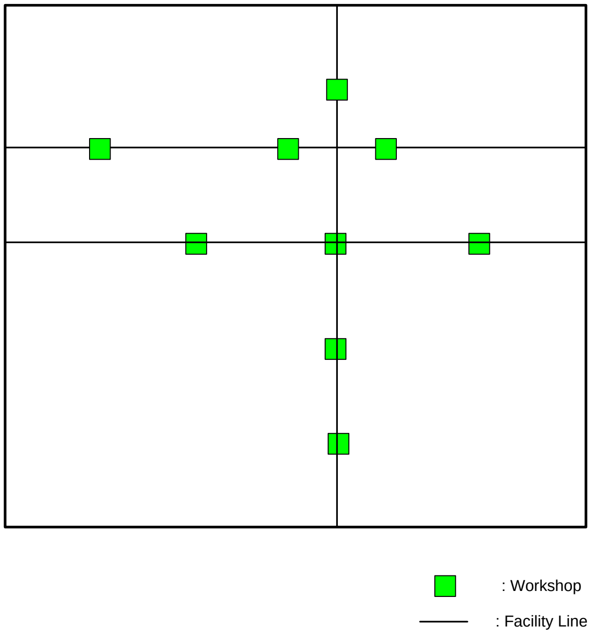


Figure 4.3: A Warehouse Example.

$$\begin{aligned} \textbf{IP:} \qquad & \min \sum_{j=1}^{N} z_j \\ \text{s. t. } \forall \text{ client } i,\ & z_{l_{x(i)}} + z_{l_{y(i)}} \geq 1 \\ & z_k \in \{0,1\} \end{aligned} \tag{4.5}$$

By relaxing the integers solutions to positive reals, the IP above can be written as a Linear Program (LP) problem as follows,

$$\begin{aligned} \textbf{LP:} \qquad & \min \sum_{j=1}^{N} z_j \\ \text{s. t. } \forall \text{ client } i,\ & z_{l_{x(i)}} + z_{l_{y(i)}} \geq 1 \\ & z_k \geq 0 \end{aligned} \tag{4.6}$$

By setting the above LP as our primal problem, the corresponding dual linear problem (DLP) of the above can be formulated as:

$$\begin{aligned} \textbf{DLP:} \qquad & \max \sum_{i=1}^{M} \alpha_i \\ \text{s. t. } \forall \text{ line } j,\ & \sum_{\text{client } k \text{ is on line } j} \alpha_k \leq 1 \\ & \alpha_i \geq 0 \end{aligned} \tag{4.7}$$

Primal-dual Algorithm and Performance Analysis

Before the description of the primal dual algorithm, we would like to motivate its usage by first analyzing some simple algorithms, such as the greedy algorithm described below, and get some insights into this problem.

The greedy algorithm is the most natural way of attacking this particular problem. It works in the following fashion. The set of *unsatisfied clients* is initialized to include all the clients. Then the facility which intersects the most number of unsatisfied clients is picked. Once these clients are removed from the unsatisfied clients set the next facility which intersects the most unsatisfied clients is picked. These steps are repeated until the set of unsatisfied clients is empty.

However, the performance of this greedy algorithm can be quite bad as shown by Hassin and Megiddo. Suppose the number of horizontal lines is N and we deploy the points in the way shown in Fig. 4.4. The number of output lines given by the greedy algorithm is $(1 + 1/2 + 1/3 + \cdots + 1/N)N$. It is apparent that the optimal solution to this case is N horizontal lines. Therefore, the approximation ratio is $(1 + 1/2 + 1/3 + \cdots + 1/N)$, which can be in the order of N.

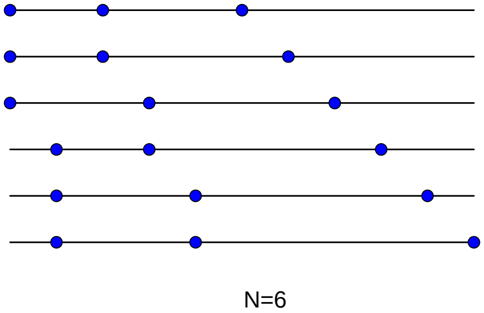


Figure 4.4: An example that shows the performance ratio of the greedy algorithm is $\log n$.

Again, we resort to the primal dual schema based algorithm described above, and show that it achieves 2-approximation ratio. Here we recapitulate the algorithm in this different setting (where connection costs are either 0, if the corresponding line is on the point, or ∞, if the line is not on the point). In this algorithm, a notion of time is defined, which is associated with each event as it happens. The algorithm starts at time 0. The dual variables α_i associated with point client i are all initialized to be 0. Then the algorithm raises the dual variables at a uniform rate. When the sum of α_i of the points on a line is equal to 1, the line is *open*. All the points on the line are serviced by the line and we freeze their dual variables. This process iterates till all the points are serviced.

Theorem 2: The Primal-dual Algorithm has an approximation ratio of 2.

Proof: Because each point has two connections to the lines at most, the dual variable of each point contributes to two lines at most. Therefore, in the worst case, the number of opened lines equals to twice that of the optimal. Factoring the relationship of the primal and dual programs, i.e., the primal solution is always greater than the dual solution, we can easily show the performance ratio is at most 2.

Remark: This analysis is similar to the analysis of the primal-dual approximation for vertex cover.

However, even a simple greedy algorithm can achieve the same approximation ratio. It works as follows. Starting from any un-serviced point client, open both the horizontal and vertical lines going through it, and use them to service all the clients on it. Repeat this until all the clients are serviced. This shows the sophisticated primal dual algorithm is not superb for this simple case.

4.3.2 Rectangular Client Case

Now we consider the case where the clients are rectangles with varying sizes. This case is more general and the point case described above can be treated as a special case of it.

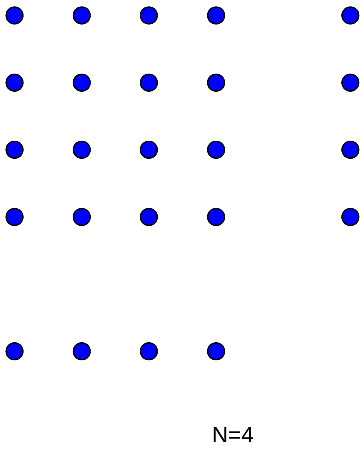


Figure 4.5: A Worst Case.

Remark: This problem is also known as the rectangle stabbing problem and has been studied by various authors in different contexts.

Problem Formulation

The formulation is slightly more complicated. Now, the facility lines (either horizontal or vertical) are specified by the edge coordinates of the rectangular clients[4]. More importantly, each square client has more than two lines going through it.

Again, we denote the line set by $\mathcal{L} = \{l_1, l_2, l_3, \ldots, l_N\}$. For each rectangular client, we specify them by the labels of lines going through it. The modified IP is given by:

$$\begin{aligned} \textbf{IP:} \quad & \min \sum_{j=1}^{N} z_j \\ \text{s. t. } \forall \text{ client } i, \quad & \sum_{\text{line } k \text{ is on client i}} z_k \geq 1 \\ & z_k \in \{0,1\} \end{aligned} \tag{4.8}$$

Again, by relaxing integer solutions to be positive reals, we have the relaxed LP problem:

$$\begin{aligned} \textbf{LP:} \quad & \min \sum_{j=1}^{N} z_j \\ \text{s. t. } \forall \text{ client } i, \quad & \sum_{\text{line } k \text{ is on client i}} z_k \geq 1 \\ & z_k \geq 0 \end{aligned} \tag{4.9}$$

The corresponding DLP is given by:

[4] If the lines are on the edges, we could always move it to the boundary without breaking the covering relations.

$$\textbf{DLP:} \qquad \max \sum_{i=1}^{M} \alpha_i$$

$$\text{s. t. } \forall \text{ line } j, \quad \sum_{\text{client } k \text{ is on line } j} \alpha_k \leq 1$$

$$\alpha_i \geq 0 \tag{4.10}$$

Primal-dual Algorithm and Performance Analysis

One crucial observation is that the number of serving lines for a rectangular client is not bounded. By reducing this problem to a special set cover problem, the cardinality of each set can not be bounded. This is intuitively why, as we show later, that even sophisticated algorithms, e.g., the Primal-dual schema based algorithm mentioned in the introduction, can only achieve $O(\log n)$ performance ratio.

We still use a primal-dual method to solve the problem in the dual space first. Unlike the point case, where the number of edges connecting each point to lines is fixed to two, in the rectangle case, the number of edges incident to each rectangle varies.

We design a greedy algorithm to solve the dual problem. Again, we use the notion of time. All the dual variables α_i are initialized to 0. Then the algorithm raises all the dual variables at a uniform rate. When the sum of α_i of the clients on a line reaches 1, these lines are open and we freeze α_i of these clients. These clients are serviced by this open line. We then raise the unfrozen α_i until we get the next set of tight edges. This process iterates till all the clients are serviced.

Theorem 3: The Primal-dual Algorithm has an approximation ratio of $\log(n)$, where n is the number of clients and the bound is tight.

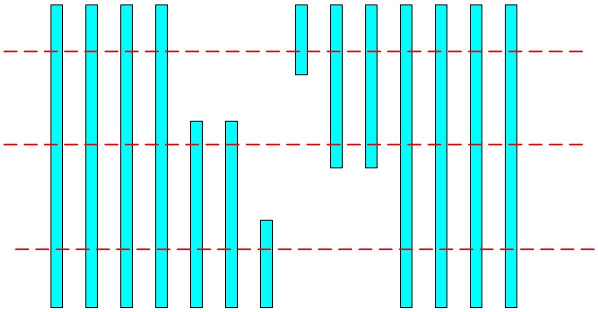


Figure 4.6: A Tight Example.

Proof: The facility location problem with zero connection cost is a special case of the set cover problem. It is well known that set cover approximated within a factor of $\log n$ using the primal-dual schema. In the example shown in Fig. 4.6, we observe that the approximation ratio of our primal-dual algorithm is actually $\log(n)$. (The primal dual algorithm will open the middle line first since the number of clients is the largest for this line. After freezing the corresponding dual variables, only the dual variables on the upper and lower halves are unfrozen. Then the algorithm opens the middle line of each half, and so on and so forth. This is why the output, the number of lines picked is $\log(n)$.

In conclusion, for line facility serving clients problems, we looked into two cases. In the first step, we applied the Primal-dual algorithm for a point client case. The primal dual algorithm was shown to have a tight approximation ratio of 2, which can also be easily achieved by a simple greedy algorithm. An intuitive explanation is that a point client can be serviced by at most two line facilities. In the dual setting, this means the dual variable assigned to one point client can contribute to at most twice the opening costs. So the primal solution is bounded by twice the dual solution, which gives 2 approximation result.

We then studied the rectangular client case using the primal dual algorithm. Due to the fact that the number of serving lines is not bounded, the approximation ratio is not bounded.

4.4 Comparison of Primal-dual Algorithm with Local Search Strategy

We implemented a simple local search algorithm and compared its performance with the primal dual algorithm described above. We show that, although both algorithms are proven to be 3-approximation for the metric facility location problem, the simple local search algorithm outperforms the primal dual algorithm in every problem instance (without triangle inequality) that we tested.

4.4.1 Motivation

Facility location optimization has been the subject of extensive research. The objective is to open a subset of facilities such that the total cost of opening the facilities plus the cost of serving the clients is minimized. Jain and Vazirani [1] proposed a 3-factor approximation algorithm for the metric facility location problem using the primal-dual schema. The triangle inequality constraint is essential for their algorithm to guarantee 3-factor approximation. As we have shown in the previous sections, without this constraint, the primal dual algorithm has either the approximation ratio of c_{max}/c_{min}, where c_{max} and c_{min} are the maximum and minimum connection costs respectively, or the approximation ratio of $\log(n)$. This implies the primal dual algorithm is not a constant-factor approximation algorithm in general. Most recently, Arya et. al. [2] showed that a simple local search algorithm when applied to MFL is 3-approximation, the same factor as the primal dual algorithm. This motivates us to compare the performance of the two algorithms for general FL problems. Moreover, by running the local search algorithm on the worst-case scenarios for the primal dual algorithm, one might get some insights into why the primal dual algorithm performs so poorly in those cases. In the following, we will describe the local search algorithm, and compare the performances of the primal dual algorithm with the local search algorithm on some FL problem instances.

4.4.2 Local Search Algorithm

Local search is a very simple and yet powerful technique proposed by Kuehn and Hamburger [3] four decades ago, and has been shown to exhibit good practical performance [4, 5]. The algorithm starts with any feasible solution and then iteratively improves the solution by repeatedly moving to the best "neighbouring" feasible solution. For the facility location problems, a "neighbouring" solution is defined as a solution obtained by either adding, deleting or swapping facilities.

Although the basic idea of the local search algorithm is very simple, it is very hard to analyze its performance theoretically. Most recently, Arya et. al. [2] provided a locality gap of 3 for the local search algorithm on metric FL problems. This algorithm provides the same approximation ratio as the primal dual algorithm used for the metric FL problems.

Algorithm

1. $S \leftarrow$ an arbitrary feasible solution in S.
2. While $\exists S' \in \beta(S)$ such that cost$(S') <$ cost(S), do $S \leftarrow S'$.
3. return S.

In the above, $\beta(S)$ is the "neighbourhood" of S that can be reached from S by one operation, namely adding, dropping or swapping a facility.

4.4.3 Empirical Results

We implemented a simple local search algorithm which randomly chooses an initial feasible solution as well as operations. The stopping criteria is when some number of operations has been reached and results in no improvement in the cost function. We compared the performance of this simple local search algorithm with the primal dual algorithm on some general FL problem instances. For small problems, we also computed the optimal solution by exhaustive search. Table 4.1 shows the results for some worst-case scenarios of the primal dual algorithm. The first three are as in Fig. 4.1 and the last one is the same as in Fig. 4.6. The local search algorithm finds optimal solutions in all these cases. Table 4.2 shows the results for some randomly generated data sets. We only computed the optimal solutions for the cases with only 5 facilities, where exhaustive searching is possible. For all the cases, the local search algorithm outperforms the primal dual algorithm, especially for large problems. It is also interesting to note that for all the test cases, the local search algorithm was able to find the known optimal solution. This shows the power of random search schema over the greedy approaches. The price is the overhead in the running time: while the primal-dual algorithm guarantees to run in polynomial time, the local search algorithm might take exponential time to converge.

Since the local search algorithm is implemented in Matlab while the primal dual algorithm is implemented in C++, we are not reporting running times here. However, the local search algorithm runs much slower than the primal dual algorithm. To meet the deadline, we weren't able to obtain results for the last 2 cases with the local search algorithm.

Problem Size	Optimal Solution	Primal-dual	Local Search
2*5	25.01	160.005	25.01
2*7	105.02	405.01	105.02
2*8	24.002	606.001	24.002
5*62	2	1.5e+11	2

Table 4.1: Test Results on the worst-case scenarios of the primal dual algorithm.

4.5 Conclusion and Future Work

In this modelling camp, we have studied several variants of the facility location (FL) problem. Motivated by the primal dual schema based algorithm applied to the metric version of FL that achieves constant (3) ratio performance, we applied the same primal dual algorithm to these variants while relaxing the metric assumption, i.e., triangle inequalities. The first problem is the case where the connection costs are within a certain range. We show that the performance ratio of primal dual algorithm is bounded by the ratio of the upper and lower bounds of the connection costs. The second problem is the line facility rectangular client case where the connection costs are either 0 or ∞. We show that when the clients can be further simplified as points, the primal dual algorithm will have approximation ratio of 2, while in the general case, rectangular clients, the ratio is $\log(n)$. Then we empirically compare a popular simple algorithm in practice, local search strategy, with the primal dual algorithm. For all the tested examples, some difficult problem instances such as the worst case examples used in the analysis of the two variants of FL, the local search algorithm outperforms the primal dual algorithm. This validates the usages of local search in the common practice.

In our future work, we would like to investigate further along the following lines. For the bounded connection cost case, we would like to study the performance of a generalized version of the primal dual algorithm. For the line facility rectangular client case, we would like to investigate the possibility of designing a more sophisticated primal dual algorithm for achieving constant approximation ratio.

Problem Size	Optimal Solution	Primal-dual	Local Search
5*20	610.1	1255.2	610.1
	478.5	1069.4	478.5
	749.4	1062.7	749.4
	699.9	1056.7	699.9
	538.1	811.9	538.1
20*30	/	1060	432.8
	/	1111.5	393.5
	/	1173.4	500.9
	/	1384.9	460.1
	/	1298	479.5
50*100	/	3602.9	740.2
	/	3411.6	682.9
	/	3812.3	801.1
	/	4401	776.6
	/	4521.9	864.7
100*200	/	6865.7	1134.3
	/	7666.5	1060.8
	/	6954.1	1157.6
	/	6749.1	1110.4
	/	7287.2	1137
300*400	/	15808.8	1137.4
	/	15186.8	1243.8
	/	15922.4	1154.1
	/	16088.7	/
	/	17977.1	/

Table 4.2: Test results on randomly generated data sets.

Bibliography

[1] K. Jain and V. Vazirani. ‘Approximation Algorithms for Metric Facility Location and k-Median Problems Using the Primal-dual Schema and Lagrangian Relaxation’. *Journal of ACM*, 48(2): 27–296, 2001.

[2] V. Arya, N. Garg, R. Khandekar, A. Meyerson, K. Munagala and V. Pandit. ‘Local Search Heuristics For k-Median and Facility Location Problems’. *SIAM Journal on Computing*, Vo. 33, No. 3, pp 544–562, 2004.

[3] A. Kuehn and M. Hamburger. ‘A Heuristic Program For Locating Warehouses’. *Management Science*. 9: 643–666, 1963.

[4] G. Diehr. ‘An Algorithm for the p-Median Problem’. Technical Report No. 191, Western Management Science Institute, UCLA, 1972.

[5] M. Teitz and P. Bart. ‘Heuristic Methods For Estimating the Generalized Vertex Median of a Weighted Graph’. *Operations Research*, 16: 955–961, 1968.

List of Participants

Organizing Committee

Elena Braverman	University of Calgary
Hadi Kharaghani	University of Lethbridge

Mentors

C. Sean Bohun	Penn State University
Chris Bose	University of Victoria
Lou Fishman	MDF International
Daya Gaur	University of Lethbridge

Students

Enkeleida Lushi	Simon Fraser University
Pengpeng Wang	Simon Fraser University
Robin Clysdale	University of Calgary
Matthew Emmett	University of Calgary
Chad Hogan	University of Calgary
Mahyar Mohajer	University of Calgary
Hui Huang	University of British Columbia
Jihong Ren	University of British Columbia
Chong Liu	University of Victoria
Maryam Mizani	University of Victoria
Liang Xu	University of Washington
Zachary Friggstad	University of Lethbridge
Massih Khorvash	University of Lethbridge
Dallas Thomas	University of Lethbridge
Mahmoud Akelbek	University of Regina
Sandra Fital	University of Regina
Xiaoping Liu	University of Regina
Ojenie Artoun	Concordia University
Peter Smith	Memorial University of Newfoundland
John Gonzalez	Northeastern University
Diana David-Rus	Rutgers University
Vesselin Marinov	Rutgers University
Sarah Williams	University of California, Davis

Jisun Lim	University of Colorado
Hatesh Radia	University of Massachusetts, Lowell
Yaling Yin	University of Saskatchewan
Zhidong Zhang	University of Saskatchewan
Amirhossein Amiraslani	University of Western Ontario
Parisa Jamali	University Western Ontario
Mohammadrahim Nouri	University of Western Ontario
James Odegaard	University of Western Ontario
Nargol Rezvani	University of Western Ontario
Oulu Xu	York University
Naveen Vaidya	York University

PIMS Contact Information

email: pims@pims.math.ca
http://www.pims.math.ca

- **Director: I. Ekeland**
  Phone: 604-822-3922
  Fax: 604-822-0883
  email: director@pims.math.ca
- **Deputy Director: A. Adem**
  email: deputy@pims.math.ca
- **SFU-Site Director: R. Choksi**
  email: sfu@pims.math.ca
- **UAlberta-Site Director: G. Cliff**
  email: ua@pims.math.ca
- **UCalgary-Site Director: G. Chen**
  email: uc@pims.math.ca
- **UVictoria-Site Director: C. Bose**
  email: uvic@pims.math.ca
- **UWashington-Site Director: G. Uhlmann**
  email: uw@pims.math.ca