Study of Singular Capillary Surfaces and Development of the Cluster Newton Method

by

Yasunori Aoki

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Applied Mathematics

Waterloo, Ontario, Canada, 2012

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In this thesis, we explore two important aspects of study of differential equations: analytical and computational aspects. We first consider a partial differential equation model for a static liquid surface (capillary surface). We prove through mathematical analyses that the solution of this mathematical model (the Laplace-Young equation) in a cusp domain can be bounded or unbounded depending on the boundary conditions. By utilizing the knowledge we have obtained about the singular behaviour of the solution through mathematical analysis, we then construct a numerical methodology to accurately approximate unbounded solutions of the Laplace-Young equation. Using this accurate numerical methodology, we explore some remaining open problems on singular solutions of the Laplace-Young equation. Lastly, we consider ordinary differential equation models used in the pharmaceutical industry and develop a numerical method for estimating model parameters from incomplete experimental data. With our numerical method, the parameter estimation can be done significantly faster and more robustly than with conventional methods.

Acknowledgements

I would like to thank the thesis examination committee for their time.

I would like to thank Professor David Siegel for supervising my PhD thesis and giving me an opportunity to further study mathematics.

I would like to thank Professor Hans De Sterck for supervising my PhD thesis. I am especially thankful that he has allowed me to follow my curiosity and intuition, and let me explore my "unusual" ideas.

I would like to thank Professor Ken Hayami and Professor Akihiko Konagaya for supervising my work presented in Chapter 5. I am thankful for Professor Hayami for the internship opportunity at the National Institute of Informatics in Tokyo, Japan, and Professor Konagaya for introducing me to the field of pharmacometrics.

I would like to thank the Department of Applied Mathematics and people in the department for always being very encouraging.

I would like to thank Ms. Alicia Wilkinson for proofreading the manuscripts, and for providing personal support for the last few years of my PhD.

I would like to thank my family for supporting me throughout my studies, and for understanding my passion for academic research.

for my mother and my sister

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In recent years, the use of mathematics has been increasingly spreading throughout many disciplines, and the role of an applied mathematician has been diversified greatly as society recognizes the power of mathematics. One of the most powerful modelling tools in practice is differential equations and their application ranges from physics, chemistry, and biology to the social sciences. Due to the complexity of differential equation based mathematical models that appear in diverse fields, there is a wide spectrum of techniques for studying differential equations. Analytical study of differential equations can give us important qualitative information about the mathematical model. Numerical study of differential equations can give us a way to quantitatively understand the solution behaviour. Experimental study can be conducted in order to calibrate a mathematical model through model parameter identification. In this thesis, we address these three aspects that are often used for differential equations in practice. We begin in Chapter 2 by mathematically proving analytic properties of the solution of the Laplace-Young equation in a cusp domain. In Chapters 3 and 4 we discuss numerical methods to accurately approximate singular solutions of elliptic partial differential equations. Lastly, in Chapter 5, we introduce a new computationally efficient algorithm to estimate the parameters of an ordinary differential equation system based on incomplete experimental data.

Since we regard the Laplace-Young equation which is, a mathematical model for a static liquid surface,

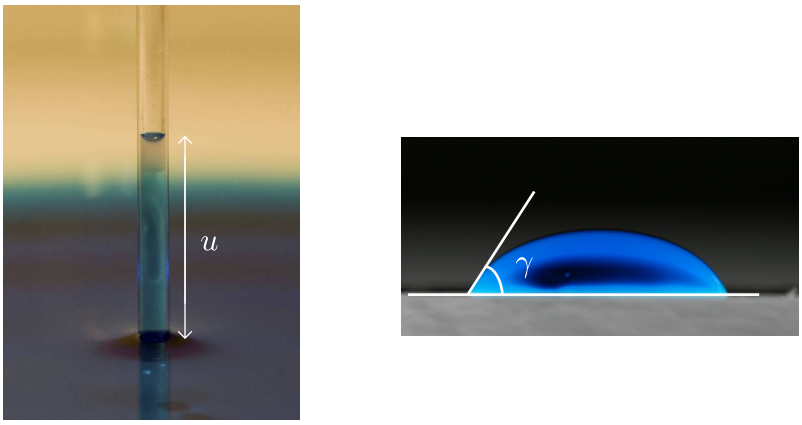


Figure 1.1: Two photos of capillarity experiments, indicating the capillary surface height u and the contact angle γ.

as one of the simplest equations of this kind with physical significance, the study of this equation can be a suitable starting point to deepen our understanding of nonlinear elliptic partial differential equations.

As may be familiar to the reader, the study of capillary surfaces originated in capillary tube experiment (cf. Figure 1.1), which dates back to the time of Aristoteles [15]. Even two millenniums later this is still a mysterious phenomenon to many, as the numerous "inventions" of the capillary tube perpetual motion machines indicate.[1] Ironically, as opposed to the perpetual "creation" of energy, capillary tube phenomena can be explained through the idea of energy minimization, as first discussed by Gauss (as presented in [13]). Given a domain Ω bounded by a vertical solid wall at the boundary, the wetting energy caused by the intermolecular forces can be written as in the following energy functional $E(u)$:

$$E(u) \quad = \quad \int_\Omega \left[g \frac{\rho_{\text{liquid}} - \rho_{\text{air}}}{\sigma_{\text{air-liquid}}} \frac{u^2}{2} + \sqrt{1 + |\nabla u|^2} \right] dA + \int_{\partial\Omega} \left[\frac{\sigma_{\text{liquid- solid}} - \sigma_{\text{air- solid}}}{\sigma_{\text{air-liquid}}} u \right] ds \,, \tag{1.1}$$

[1] One can easily find such "inventions" presented online still to date.

with

$$\begin{aligned}
u : \mathbb{R}^2 \to \mathbb{R} \quad &: \quad \text{the height of the liquid surface,} \\
\Omega \quad &: \quad \text{the domain ,} \\
\partial\Omega \quad &: \quad \text{the boundary of the domain,} \\
g \quad &: \quad \text{the acceleration due to gravity,} \\
\rho_{\text{liquid/air}} \quad &: \quad \text{the density of liquid/air,} \\
\sigma_{\text{air-liquid/air-solid/liquid-solid}} \quad &: \quad \text{the surface tension between air and liquid, etc.}
\end{aligned}$$

The Euler-Lagrange equation of (1.1) is called the Laplace-Young equation, or capillary surface equation:

$$\nabla \cdot \frac{\nabla u}{\sqrt{1+|\nabla u|^2}} = \kappa u \,, \tag{1.2}$$

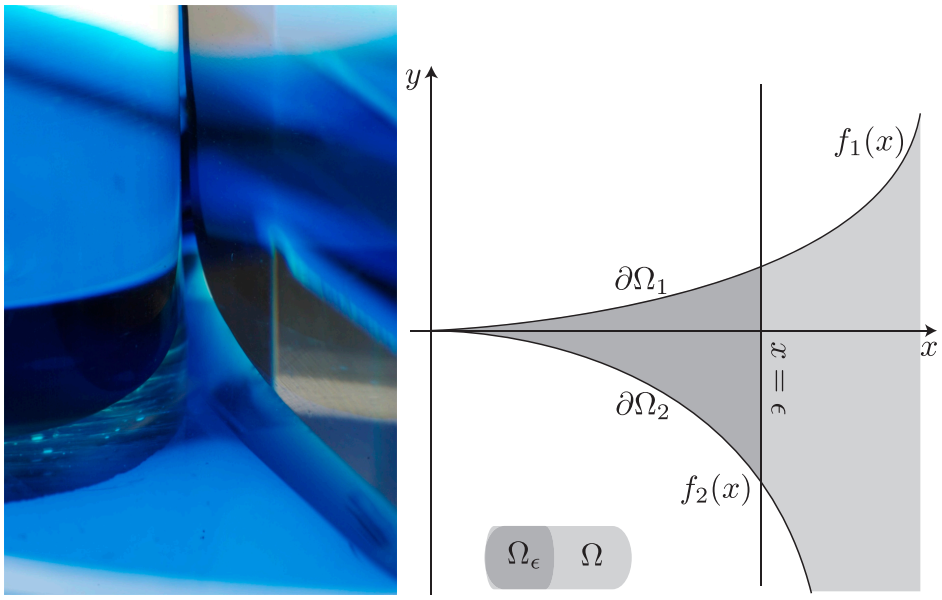
where $\kappa = g(\rho_{\text{liquid}} - \rho_{\text{air}})/\sigma_{\text{air-liquid}}$. Also, by minimizing the energy functional, we obtain a contact angle boundary condition:

$$\begin{aligned}
\nu \cdot \frac{\nabla u}{\sqrt{1+|\nabla u|^2}} &= -\frac{\sigma_{\text{liquid- solid}} - \sigma_{\text{air- solid}}}{\sigma_{\text{air-liquid}}} \,, &(1.3)\\
&= \cos\gamma \qquad \text{on } \partial\Omega \,, &(1.4)
\end{aligned}$$

where γ is the contact angle as depicted in Figure 1.1. We refer the reader to Section 1.4 of [13] for details, and to [1] for calculations.[2]

In everyday life, it is often safe to assume that the surface of still water is almost flat; however, a careful observation can tell us that the surface of water in a container can exhibit complicated geometry near the interface where the water meets the container. One of the most extreme examples of this geometry is when the container has a sharp boundary. As can be seen in Figure 1.2(a), in certain cases the liquid surface rises abruptly near the sharp boundary all the way to the top of the container. Such behaviour of the

[2]For the flow of our discussion, we have introduced the energy functional before the Laplace-Young equation; however, according to [13], the Laplace equation and Young equation were introduced in 1806 and 1805 based on physical observation, and then Gauss proposed the energy functional in 1830.



(a) A photograph of a capillary surface in a cusp domain (cusp formed by a circular cylinder and a straight wall).

(b) Domain Ω and boundaries.

Figure 1.2: Capillary surface in cusp domain.

static liquid surface (capillary surface) at a sharp boundary of a container can be studied by considering singular solutions[3] to the Laplace-Young capillary surface equation.

As noted in a celebrated book on capillarity by Robert Finn [13] the study of singular capillary surfaces can be traced back to Brook Taylor in 1712. Relatively recent contributions to the study of singular capillary surfaces by Concus and Finn [8, 9] and Miersemann [30] have spurred considerable interest in the field (e.g. King et al. [24], Scholz [39, 40], Norbury et al. [33], and Aoki [1]). In particular, work by Scholz on capillary surfaces in a domain containing a cusp where the boundaries can be approximated by power series (including fractional powers) has provided us with the asymptotic behaviour of the capillary surface. Scholz concluded that "[the capillary surface] rises with the same order [as] the order of contact of the two arcs, which form the cusp" [40]. As the above statement is very intuitive, our curiosity leads us to question

[3]We define a singularity to be a point where the smoothness of the function differs from the neighbouring points. In this thesis we especially focus on two types of singular solutions, where the solution becomes unbounded at a point or the derivative of the solution becomes unbounded at a point.

if this statement holds for the cases which Scholz did not consider in his paper [40]. In Chapter 2 we show that Scholz' statement holds for the cases he has left off and also prove that the solution is bounded if the contact angles of the two walls forming a cusp are supplementary angles under some conditions on the shapes of the boundary arcs forming the cusp.

With the work presented in Chapter 2 and other works on unbounded capillary surfaces ([9, 8, 30, 24, 33, 39, 40, 1]), almost all of the leading order asymptotic behaviours of unbounded capillary surfaces with a contact angle boundary condition are known. With all the asymptotic analysis results, one may think that there does not remain much room for exploration on unbounded solutions of the Laplace-Young equation; however, asymptotic analysis is a local analysis, and asymptotic approximations are only valid in a sufficiently small domain near the singularity. It is also not easy to determine the precise region of validity of the asymptotic analysis results. In applications, global approximations for singular capillary surfaces that are valid in the whole domain are desirable, and such approximations cannot be provided by asymptotic analysis. We wish to construct globally valid approximations of singular capillary surfaces which exhibit the proper asymptotic behaviour at the singular point, while also being valid away from the singularity.

Since the equation of interest, the Laplace-Young equation, is a nonlinear partial differential equation, we cannot expect to obtain a closed form solution or an analytic global approximation. Thus we take the path of numerical approximation using finite element approximations. As it is widely known ([17, 46]) that the lack of smoothness of the solution can spoil the accuracy of finite element approximations, it is easy to imagine that standard finite element approximations cannot accurately approximate the unbounded singularity. As a preliminary attempt to accurately approximate the singular solution of the partial differential equation, we have conducted numerical studies on the Finite Volume Element Method (FVEM) for derivative blowup singular solutions of the Poisson equation. Roughly speaking, the FVEM is a finite element like method formulated by integration over finite volume cells defined in a dual mesh. To treat derivative blowup singularity, augmentation of the trial function space by singular basis functions is often done [17, 46], to avoid the decrease in accuracy. To the author's best knowledge, augmentation of the trial function space for the FVEM has not been done before. In Chapter 3 we show that augmentation for

the FVEM applied to the Poisson equation can be done significantly simpler than that for the Galerkin Finite Element Method (FEM), and the FVEM can avoid the technical challenge of singular integrations, resulting in adequate high order convergence for singular solutions.

Following the preliminary work on singular solutions of the linear elliptic PDEs in Chapter 3, we aim at obtaining global numerical approximations of unbounded solutions of the Laplace-Young equation in Chapter 4. There are about a half dozen published papers on numerical solutions of the Laplace-Young equation [32, 22, 37, 36, 41], none of which consider unbounded singular solutions, except for the paper by Scott et al. [41]. Scott et al. use the FVEM to approximate solutions of the Laplace-Young equation, and one of their model problems is a corner problem with unbounded singularity. Although the work by Scott et al. [41] is important in that it is the first attempt to numerically approximate unbounded capillary surfaces, the accuracy of the approximation is not discussed in detail in their paper. We show in Section 4.3.1 that since Scott et al. have used a regular finite element expansion, their numerical approximation does not approximate the singular behaviour accurately.

Instead of directly approximating the solution with a standard finite element expansion, as has been done in [41], our idea is to incorporate knowledge obtained from asymptotic analysis into the finite element expansion in order to avoid inaccuracies introduced by the singularity. Roughly speaking, we first change the variable based on the known asymptotic order of the solution so that the new unknown function is bounded. Even though this new unknown function is bounded, it can still be discontinuous at the location of the original singularity. Inspired by knowledge of the leading order term of the asymptotic series solution, we change the coordinate system so that the unknown function is smooth with respect to the new coordinate variables. We then finally approximate the new smooth bounded unknown function with respect to the new coordinate variables using the FVEM or the Galerkin FEM.

Using the proposed accurate numerical methodology for computing singular solutions of the Laplace-Young equation, we investigate a few open problems of singular behaviour of the Laplace-Young equation in a cusp domain. While the work in Chapter 2 aims to address the solution behaviour for all possible cusp domains, attempting to generalize the results of Scholz [40], a few cases still remain open. Using our new highly accurate computational methodology, we numerically investigate the cases that we were not

able to address in Chapter 2 and make conjectures based on the numerical approximations.

In Chapters 2-4 we have studied elliptic PDEs analytically and numerically. Although both of the PDEs have physical interpretations, our study was motivated primarily by academic interest. We next consider ordinary differential equation-based mathematical models that are widely used in the industry.

Mathematically, ordinary differential equations (ODE) are well studied and ODEs now appear everywhere in industrial applications and computationally very efficient ODE solvers are available.

Recently, ODEs have made their way to the pharmaceutical industry as a way to model the behaviour of drugs in the human body. For example, pharmacokinetics models are ODE-based mathematical models for transportation, metabolization, and elimination of the drug in a body, used in Phase 1 and 2 clinical studies. Roughly speaking, a mathematical model for pharmacokinetics is constructed by linking compartments representing organs using diffusion processes and by expressing the metabolic reactions in the compartments with Michaelis-Menten kinetics. As drugs and metabolites have different diffusion rates for different organs and different metabolic reaction rates for different enzymes, pharmacokinetics models contain many model parameters. One often has reasonable ideas about these parameters from preclinical studies of the drug (i.e., from *in vitro* studies or animal studies); however, in Phase 1 clinical trials and sometimes in Phase 2 clinical trials one aims to estimate these parameters for a human body. Mathematically speaking, the parameter estimation can be thought of as solving an inverse problem where one does not know some of the coefficients of the system of ODEs, while measurements are available for some of the variables. As a reasonable initial guess is available for the parameters one wishes to estimate, one can easily imagine using a local optimization algorithm such as the Gauss-Newton method or the Levenberg-Marquard method to solve this inverse problem. However, as one can imagine, the information that can be obtained clinically from a live person (or a patient going through treatment, in case of Phase 2 study) is often much less extensive than the complexity of the internal activity in a patient's body. As a result, this parameter identification problem becomes a very underdetermined inverse problem. When confronted with an underdetermined inverse problem, one often adds mathematically convenient constraints and considers only one of many possible solutions, even though it may be beneficial for the application to consider multiple solutions. We propose an algorithm for simultaneously finding multiple solutions of an

underdetermined inverse problem that are suitably distributed, guided by *a priori* information on which part of the solution manifold is of interest. Through numerical experiments, we show that our algorithm is a fast, accurate, and robust solution method, especially applicable to this type of ODE coefficient identification problems. We give an example of applying this algorithm for the estimation of the parameters of a pharmacokinetics model of an anti-cancer drug (CPT-11) from mass-balance data collected in a Phase 2 study.

In this thesis, we have three major themes. These themes spans the field of applied mathematics, from applied analysis, to numerical computation, to mathematical biology. In Chapter 2 we focus on the mathematical analysis of the solution behaviour of a nonlinear elliptic PDE. In Chapter 3 and 4 we focus on numerical approximation methods for singular solutions of the elliptic PDE. In Chapter 5 we focus on parameter identification for a system of ODEs with pharmaceutical application. Each chapter is made to be as self-contained as possible.

Publications/Manuscripts Related to this Thesis

- Yasunori Aoki: *Analysis of Asymptotic Solutions for Cusp Problems in Capillarity*, Master's thesis, University of Waterloo, 2007. Available at http://hdl.handle.net/10012/3352
  - Chapter 2 of this thesis is an extension of the preliminary work presented in Chapters 2 and 4 of the author's Master's degree thesis.
- Yasunori Aoki and David Siegel: *Bounded and Unbounded Capillary Surfaces in a Cusp Domain*, Pacific Journal of Mathematics, Volume 257, Issue 1, 2012, Pages 143-165. DOI 10.2140/pjm.2012.257.143.
  - Chapter 2 of this thesis is based on this paper, Theorem 3.2 presented in this paper is not included in this thesis as it was the contribution of the coauthor.
- Yasunori Aoki and Hans De Sterck: *Augmented High Order Finite Volume Element Method for Elliptic PDEs in Non-smooth Domains*, Journal of Computational and Applied Mathematics, Volume 235, Issue 17, 2011, Pages 5177-5187.

– Chapter 3 of this thesis is based on this publication.

- Yasunori Aoki and Hans De Sterck: *Numerical Study of Unbounded Capillary Surfaces*, submitted to Pacific Journal of Mathematics, 2012

  – Chapter 4 of this thesis if based on this manuscript.

- Yasunori Aoki, Ken Hayami, Hans De Sterck and Akihiko Konagaya: *Cluster Newton Method for Sampling Multiple Solutions of an Underdetermined Inverse Problem: Parameter Identification for Pharmacokinetics*, submitted to Inverse Problems, 2011. Also appeared as a technical report at: http://www.nii.ac.jp/TechReports/11-002E.html.

  – Chapter 5 of this thesis is based on this manuscript.

Chapter 2

Bounded and Unbounded Capillary Surfaces in a Cusp Domain

In this chapter we extend the results of Scholz [40] in two directions, we first consider cusp domains which are not limited to the power-law cusp. Instead of approximating the boundary by power series, we directly use the distance between two arcs forming a cusp in the asymptotic expansion. Although one may argue that most of the shapes used in real life applications can be approximated by power series, our main focus was to justify the statement "[the capillary surface] rises with the same order [as] the order of contact of the two arcs, which form the cusp" [40] in a more direct and intuitive manner, by avoiding the extra approximation step. The second direction of extension is to include cases in which the contact angles of the boundary walls forming a cusp are supplementary angles. Although all the known results suggest that a capillary surface in a domain with a cusp is unbounded, we have shown that a capillary surface can be bounded if contact angles are supplementary angles.

Statement of the Problems

Here we state problems we are going to consider in this chapter. We first define a cusp domain. Without loss of generality, for the simplicity of writing, we consider the following domain (also depicted in Figure 1.2(b)):

$$\Omega \quad = \quad \{(x,y) : x > 0,\, f_2(x) < y < f_1(x)\}\,, \tag{2.1}$$

where

$$f_1(x), f_2(x) \quad \in \quad C^3(0,\infty)\,, \tag{2.2}$$

$$f_1(x) \quad > \quad f_2(x) \qquad \text{for } x > 0\,, \tag{2.3}$$

$$\lim_{x\to 0^+} f_1(x) \quad = \quad \lim_{x\to 0^+} f_2(x) \;=\; 0\,, \tag{2.4}$$

$$\lim_{x\to 0^+} f_1'(x) \quad = \quad \lim_{x\to 0^+} f_2'(x) \;=\; 0\,. \tag{2.5}$$

Also we denote the boundaries as follows:

$$\partial\Omega_1 \quad = \quad \{(x,y) : x > 0,\, y = f_1(x)\}\,, \tag{2.6}$$

$$\partial\Omega_2 \quad = \quad \{(x,y) : x > 0,\, y = f_2(x)\}\,. \tag{2.7}$$

Although we base our dicussion on this infinite domain, as all of the results presented in this chapter only depend locally on a domain sufficiently close to the cusp, the results hold for any domain which is equal to Ω in a neighbourhood of the origin.

We now state the partial differential equation of our interest, the Laplace-Young capillary surface equation. Let $u(x,y)$ be the height of a capillary surface in domain Ω, then $u(x,y)$ satisfies the following

boundary value problem (see [13] for derivation):

$$\nabla \cdot Tu = \kappa u \qquad \text{in } \Omega\,, \tag{2.8}$$

$$\vec{\nu}_1 \cdot Tu = \cos\gamma_1 \qquad \text{on } \partial\Omega_1\,, \tag{2.9}$$

$$\vec{\nu}_2 \cdot Tu = \cos\gamma_2 \qquad \text{on } \partial\Omega_2\,, \tag{2.10}$$

where

$$\kappa \;:\; \text{the capillarity constant,} \tag{2.11}$$

$$\vec{\nu}_1,\, \vec{\nu}_2 \;:\; \text{exterior unit normal vectors on the boundaries } \partial\Omega_1 \text{ and } \partial\Omega_2, \tag{2.12}$$

$$\gamma_1,\, \gamma_2 \;:\; \text{contact angles,} \tag{2.13}$$

$$Tu = \frac{\nabla u}{\sqrt{1+|\nabla u|^2}}\,. \tag{2.14}$$

The capillarity constant κ can be normalized by rescaling x, y, and u. In the following sections we let $\kappa = 1$. Here we introduce the big theta notation to replace the statement "is the same order as" to make this statement more mathematically precise. If $f(x) = \Theta(g(x))$ then there exist constants $k_1\,,\, k_2 > 0$ and $x_0 > 0$ such that

$$k_1|g(x)| < |f(x)| < k_2|g(x)| \qquad \text{for all } x < x_0\,. \tag{2.15}$$

We note that Θ is a more strict order relation than that of O, i.e., if $f(x) = \Theta(g(x))$ then $f(x) = O(g(x))$; however the converse is not true. Using the above notations, we can write our core research questions as follows:

- Does $u(x,y) = \Theta\left(\frac{1}{f_1(x)-f_2(x)}\right)$ hold for any $f_1(x)$ and $f_2(x)$ when $\gamma_1+\gamma_2 \neq \pi$?
- How does $u(x,y)$ behave asymptotically as $x \to 0^+$ when $\gamma_1+\gamma_2 = \pi$?

Structure of the chapter

As the title of this chapter suggests, there are two main parts of this chapter: unbounded and bounded cases.

In Section 2.1 we start with introducing the major tool for the analytical study of the capillary surfaces, the Concus-Finn comparison principle.

In Section 2.2, we consider unbounded capillary surfaces in cusp domains. We first prove in Section 2.2.1 that capillary surfaces are unbounded if $\gamma_1+\gamma_2 \neq \pi$. Then in section 2.2.2 the formal asymptotic expansion is presented. Using the formal asymptotic expansion, in Section 2.2.3 we prove the asymptotic behaviour of the solution. We list some examples of power-law cusps and non-power law cusps in Section 2.2.4 with an intention of comparing our findings with the previous results presented in the paper by Scholz [40].

In Section 2.3, we consider bounded capillary surfaces in cusp domains. We prove in Section 2.3.1 that the capillary surfaces are bounded if $\gamma_1 + \gamma_2 = \pi$ and the curvature of the boundaries are finite. We end this chapter with concluding remarks in Section 2.4 by summarizing our findings and suggesting some future extensions of our results.

2.1 The Concus-Finn Comparison Principle

In this chapter we utilize the Concus-Finn comparison principle extensively. We present it here in case the reader is not familiar with this principle (see pages 110-113 of [13] and [14] for detailed discussions and proofs). For the purpose of this chapter, the comparison principle can be rephrased as the following.

Theorem 2.1.1 (super-solution)**.** *Let $u(x, y)$ be the solution of the boundary value problem* (2.8)-(2.10)

and Ω_0 be a bounded subdomain of Ω. If a function $v^+(x,y)$ satisfies the following inequalities:

$$\begin{aligned}
\nabla \cdot Tv^+ - v^+ &\leq 0 \qquad \text{in } \Omega_0\,, && (2.16)\\
\vec{\nu}_1 \cdot Tv^+ - \cos\gamma_1 &\geq 0 \qquad \text{on } \partial\Omega_1 \cap \partial\Omega_0\,, && (2.17)\\
\vec{\nu}_2 \cdot Tv^+ - \cos\gamma_2 &\geq 0 \qquad \text{on } \partial\Omega_2 \cap \partial\Omega_0\,, && (2.18)\\
v^+(x,y) &\geq u(x,y) \qquad \text{on } \partial\Omega_0 \backslash (\partial\Omega_1 \cup \partial\Omega_2 \cup \{(0,0)\})\,, && (2.19)
\end{aligned}$$

where $\partial\Omega_0$ is the boundary of Ω_0 and $\{(0,0)\}$ is a measure zero set at the origin (cusp). Then $v^+(x,y)$ is a super-solution of the boundary value problem (2.8)-(2.10), i.e.,

$$v^+(x,y) \geq u(x,y) \qquad \text{in } \Omega_0\,. \tag{2.20}$$

Similarly sub-solution can be verified.

Also we make use of one of the corollaries of the comparison principle to construct an upper-bound of the solution (see [9] or pages 113-114 of [13]).

Corollary 2.1.1 (bound by hemispheres)**.** *Let $u(x,y)$ be the solution of the boundary value problem (2.8)-(2.10), and $B_{r_0}(x_0,y_0)$ be a disk of radius $r_0 > 0$ centred at (x_0,y_0). If $B_{r_0}(x_0,y_0) \subseteq \Omega$, then*

$$-\left(\frac{1}{r_0} + r_0\right) \leq u(x,y) \leq \frac{1}{r_0} + r_0 \qquad \text{in } B_{r_0}(x_0,y_0)\,. \tag{2.21}$$

Recalling the smoothness of the boundaries (i.e. equation (2.2)), it follows immediately from Corollary 2.1.1 that $u(x,y)$ can only be unbounded at the origin (cusp).

2.2 Unbounded Capillary Surfaces

In this section, we assume $\gamma_1 + \gamma_2 \neq \pi$ and aim to prove

$$u(x,y) = \Theta\left(\frac{1}{f_1(x) - f_2(x)}\right) \qquad \text{as } x \to 0^+\,, \tag{2.22}$$

with as few restrictions on $f_1(x)$ and $f_2(x)$ as possible.

2.2.1 Proof for the Unboundedness of the Capillary Surface when $\gamma_1 + \gamma_2 \neq \pi$

In this subsection we show that $u(x,y) \neq O(1)$. In fact, this is intuitively obvious from the remarkable result of Concus and Finn [8] as a cusp can be considered as a corner with zero opening angle.

Lemma 2.2.1 (Unboundedness of $u(x,y)$ when $\gamma_1 + \gamma_2 \neq \pi$)**.** *Let $u(x,y)$ be the solution of the boundary value problem (2.8)-(2.10).*

If $\cos\gamma_1 + \cos\gamma_2 > 0$, then $u(x,y)$ cannot be bounded from above.

If $\cos\gamma_1 + \cos\gamma_2 < 0$, then $u(x,y)$ cannot be bounded from below.

Proof. Similar to the proof given in Concus and Finn [8], we prove this by contradiction. First consider the case $\cos\gamma_1 + \cos\gamma_2 > 0$, and assume there exists a constant $M > 0$ such that $u(x,y) < M$ in Ω. Integrate the PDE (2.8) in a subdomain Ω_ϵ given as in the following:

$$\Omega_\epsilon \quad = \quad \{(x,y) : 0 < x < \epsilon,\ f_2(x) < y < f_1(x)\}\,. \tag{2.23}$$

By applying the divergence theorem and the boundary conditions (2.9)-(2.10), after some calculation we obtain the following equation:

$$\int_{x=0}^{\epsilon}\int_{y=f_2(x)}^{f_1(x)} u\,dy\,dx \quad = \quad \int_{x=0}^{\epsilon}\left(\cos\gamma_1\sqrt{1+f_1'^2} + \cos\gamma_2\sqrt{1+f_2'^2}\right)dx + \int_{y=f_2(\epsilon)}^{f_1(\epsilon)} \left.\frac{u_x}{\sqrt{1+u_x^2+u_y^2}}\right|_{x=\epsilon} dx\,. \tag{2.24}$$

The trick of this proof is to realize that the last term of equation (2.24) can be bounded from below, i.e.,

$$\frac{u_x}{\sqrt{1+u_x^2+u_y^2}} > -1 \tag{2.25}$$

$$\Rightarrow \quad \int_{y=f_2(\epsilon)}^{f_1(\epsilon)} \frac{u_x}{\sqrt{1+u_x^2+u_y^2}}\Bigg|_{x=\epsilon} dx > -(f_1(\epsilon)-f_2(\epsilon))\,. \tag{2.26}$$

We now apply the assumption $u(x,y) < M$ and inequality (2.26) to equation (2.24) and obtain the following inequality:

$$\epsilon M \max_{0<x\leq\epsilon}(f_1(x)-f_2(x)) + (f_1(\epsilon)-f_2(\epsilon)) > \int_{x=0}^{\epsilon}\left(\cos\gamma_1\sqrt{1+f_1'^2}+\cos\gamma_2\sqrt{1+f_2'^2}\right)dx\,. \tag{2.27}$$

Dividing both sides of inequality (2.27) by $\epsilon > 0$ and taking the limit as ϵ approaches to 0 gives,

$$\lim_{\epsilon\to 0^+} M \max_{0<x\leq\epsilon}(f_1(x)-f_2(x)) + \lim_{\epsilon\to 0^+}\frac{f_1(\epsilon)-f_2(\epsilon)}{\epsilon} \geq \lim_{\epsilon\to 0^+}\frac{\int_{x=0}^{\epsilon}\left(\cos\gamma_1\sqrt{1+f_1'^2}+\cos\gamma_2\sqrt{1+f_2'^2}\right)dx}{\epsilon}\,, \tag{2.28}$$

and then applying the definition of derivative and equations (2.2)-(2.5) gives the following inequalities:

$$f_1'(0)-f_2'(0) \geq \left(\cos\gamma_1\sqrt{1+f_1'(0)^2}+\cos\gamma_2\sqrt{1+f_2'(0)^2}\right), \tag{2.29}$$

$$\Rightarrow \quad 0 \geq \cos\gamma_1+\cos\gamma_2\,. \tag{2.30}$$

Hence we obtain a contradiction. Proof for the case where $\cos\gamma_1+\cos\gamma_2 < 0$ can be constructed similarly.

□

Note that Lemma 2.2.1 together with Corollary 2.1.1 gives that $u(x,y)$ is unbounded at the cusp and bounded away from the cusp.

2.2.2 Formal Asymptotic Expansion of BVP (2.8)-(2.10)

The main idea is to consider the following form of asymptotic expansion:

$$v(x,y) \;=\; \frac{A}{f_1(x)-f_2(x)} + g(x,y)\frac{f_1'(x)-f_2'(x)}{f_1(x)-f_2(x)} + h(x,y)\frac{(f_1'(x)-f_2'(x))^2}{f_1(x)-f_2(x)}\,, \tag{2.31}$$

where $g(x,y), h(x,y) \in O(1)$ as $x \to 0^+$. Recalling that $\lim_{x\to 0^+} f_1(x), \lim_{x\to 0^+} f_2(x) = 0$, we have the first term significantly larger than the second term near the cusp etc., also note that the leading order term is the same order as the reciprocal of the distance between two boundaries measured in $\vec{y}$ direction.

The aim of this subsection is to find $g(x,y)$ and $h(x,y)$ so that equation (2.31) satisfies PDE (2.8) and boundary conditions (2.9)-(2.10) asymptotically.

For the simplicity of computation, we introduce coordinate variables s and t defined as follows:

$$s \;:=\; x\,, \tag{2.32}$$

$$t \;:=\; \frac{2y-(f_1(x)+f_2(x))}{f_1(x)-f_2(x)}\,. \tag{2.33}$$

Note that we have chosen the coordinate variable t so that when $t=1$, $y=f_1(x)$ and when $t=-1$, $y=f_2(x)$.

Lemma 2.2.2 (First two terms of the formal asymptotic expansion)**.** *Let*

$$A \;=\; \cos\gamma_1 + \cos\gamma_2\,, \tag{2.34}$$

$$g(t) \;=\; -\sqrt{1-\left(\frac{\cos\gamma_1(t+1)+\cos\gamma_2(t-1)}{2}\right)^2} + C_1\,, \tag{2.35}$$

$$h(t) \;=\; 0\,, \tag{2.36}$$

where C_1 is an arbitrary constant. If $f_1(s)$ and $f_2(s)$ satisfy the following conditions

$$f_1(s) - f_2(s) = o\left(f_1'(s) - f_2'(s)\right), \tag{2.37}$$

$$\frac{f_1''(s) - f_2''(s)}{f_1(s) - f_2(s)} = o\left(\frac{f_1'(s) - f_2'(s)}{(f_1(s) - f_2(s))^2}\right), \tag{2.38}$$

$$\frac{f_1'''(s) - f_2'''(s)}{f_1'(s) - f_2'(s)} = o\left(\frac{1}{(f_1(s) - f_2(s))^2}\right), \tag{2.39}$$

as $s \to 0^+$, then

$$\vec{\nu}_1 \cdot Tv|_{t=1} = \cos\gamma_1 + o(1), \tag{2.40}$$

$$\vec{\nu}_2 \cdot Tv|_{t=-1} = \cos\gamma_2 + o(1), \tag{2.41}$$

$$\nabla \cdot Tv - v = o\left(\frac{1}{f_1(s) - f_2(s)}\right), \tag{2.42}$$

as $s \to 0^+$.

Rather tedious yet straightforward calculation will verify this lemma. Instead of showing this calculation, we briefly explain here how equations (2.34)-(2.35) were deduced. We first let

$$v(s,t) = \frac{A}{f_1(s) - f_2(s)} + g(t)\frac{f_1'(s) - f_2'(s)}{f_1(s) - f_2(s)}, \tag{2.43}$$

and after some lengthy calculation with assumptions (2.37)-(2.39) we obtain

$$\vec{\nu}_1 \cdot Tv|_{t=1} = \frac{2g'(1)}{\sqrt{A^2 + 4g'^2(1)}} + o(1), \tag{2.44}$$

$$\vec{\nu}_2 \cdot Tv|_{t=-1} = -\frac{2g'(-1)}{\sqrt{A^2 + 4g'^2(-1)}} + o(1), \tag{2.45}$$

$$\nabla \cdot Tv - v = \left(\frac{4g''(t)A^2}{(A^2 + 4g'^2(t))^{3/2}} - A\right)\frac{1}{f_1(s) - f_2(s)} + o\left(\frac{1}{f_1(s) - f_2(s)}\right). \tag{2.46}$$

We now impose equations (2.40)-(2.42) and obtain the following non-linear first order Ordinary Differential

Equation (ODE) of $g'(t)$ with boundary conditions:

$$\frac{4g''(t)A^2}{(A^2+4g'^2(t))^{3/2}} = A \qquad \text{for } -1<t<1\,, \tag{2.47}$$

$$\frac{2g'(1)}{\sqrt{A^2+4g'^2(1)}} = \cos\gamma_1\,, \tag{2.48}$$

$$-\frac{2g'(-1)}{\sqrt{A^2+4g'^2(-1)}} = \cos\gamma_2\,. \tag{2.49}$$

We observe that there are two boundary conditions for the first order ODE problem; however, we note that A is an unknown constant as well. Now both $g'(t)$ and A are determined by first integrating (2.47) and then imposing boundary conditions (2.48)-(2.49). One essential observation from this derivation is that a constant of the leading order term (i.e. A) was found with the second order term (i.e. $g(t)$). In fact this pattern continues and the constant on the second order term C_1 will be determined at the same time as the third order term of the formal asymptotic expansion is found.

Lemma 2.2.3 (First three terms of the formal asymptotic expansion)**.** *Let*

$$A = \cos\gamma_1+\cos\gamma_2\,, \tag{2.50}$$

$$g(t) = -\sqrt{1-\left(\frac{\cos\gamma_1(t+1)+\cos\gamma_2(t-1)}{2}\right)^2}+0\,, \tag{2.51}$$

$$h(t) = -\frac{A}{4}\left(\delta t+\frac{t^2}{2}\right)+\frac{1-\alpha}{2A}g(t)^2+C_2\,, \tag{2.52}$$

where C_2 is an arbitrary constant. If $f_1(s)$ and $f_2(s)$ satisfy the following conditions

$$
\begin{aligned}
f_1'(s) &> f_2'(s) \qquad \text{for } s > 0\,, && (2.53)\\
f_1(s) - f_2(s) &= o\left(f_1'(s) - f_2'(s)\right)\,, && (2.54)\\
\frac{f_1''(s) - f_2''(s)}{f_1(s) - f_2(s)} &= \alpha\frac{(f_1'(s) - f_2'(s))^2}{(f_1(s) - f_2(s))^2} + o\left(\frac{(f_1'(s) - f_2'(s))^2}{(f_1(s) - f_2(s))^2}\right)\,, && (2.55)\\
\frac{f_1'''(s) - f_2'''(s)}{f_1'(s) - f_2'(s)} &= O\left(\frac{(f_1'(s) - f_2'(s))^2}{(f_1(s) - f_2(s))^2}\right)\,, && (2.56)\\
f_1'(s) + f_2'(s) &= \delta(f_1'(s) - f_2'(s)) + o(f_1'(s) - f_2'(s))\,, && (2.57)\\
f_1''(s) + f_2''(s) &= O(f_1''(s) - f_2''(s))\,, && (2.58)
\end{aligned}
$$

as $s \to 0^+$, where α, $\delta \in \mathbb{R}$, then

$$
\begin{aligned}
\vec{\nu}_1 \cdot Tv|_{t=1} &= \cos\gamma_1 + o(f_1'(s) - f_2'(s))\,, && (2.59)\\
\vec{\nu}_2 \cdot Tv|_{t=-1} &= \cos\gamma_2 + o(f_1'(s) - f_2'(s))\,, && (2.60)\\
\nabla \cdot Tv - v &= o\left(\frac{f_1'(s) - f_2'(s)}{f_1(s) - f_2(s)}\right)\,, && (2.61)
\end{aligned}
$$

as $s \to 0^+$.

Again, a long tedious calculation will prove this lemma. We followed similar steps to determine $h(t)$, although solving the differential equation for $h(t)$ was not nearly as straight forward as for $g(t)$. Constant C_1 was determined to be 0 when $h(t)$ was determined and new unknown constant C_2 appeared in the third order term.

Comparing assumptions (2.37)-(2.39) with assumptions (2.54)-(2.58), we can see that the restrictions on f_1 and f_2 increase as the number of terms in the formal asymptotic expansion increases from two terms to three terms. Although these assumptions are not proven to be necessary conditions for these lemmas to hold, it is our suspicion that as the number of the terms in the asymptotic expansion increases, restrictions on f_1 and f_2 become more strict.

2.2.3 Asymptotic Behaviour of the Capillary Surface

The main result in Section 2.2 is stated and proven in this subsection. We first show that the asymptotic growth order of the solution is the same order as the reciprocal of the distance between two arcs forming a cusp.

Theorem 2.2.1 (Growth order behaviour of $u(x,y)$)**.** *Let $u(x,y)$ be the solution of the boundary value problem (2.8)-(2.10). If $f_1(s)$ and $f_2(s)$ satisfy the conditions (2.37)-(2.39) and $|\cos\gamma_1|, |\cos\gamma_2| \neq 1$, then there exist positive constants s_0, k_1 and k_2 such that*

$$k_2\left(\frac{1}{f_1(s)-f_2(s)}\right) < |u(s,t)| < k_1\left(\frac{1}{f_1(s)-f_2(s)}\right), \qquad \text{for } s < s_0\,. \tag{2.62}$$

Proof. The main idea of our proof is to construct super-solution and sub-solution by modifying the formal asymptotic expansion given in Lemma 2.2.2. We prove these modified equations are in fact super-solution and sub-solution by applying the Concus Finn comparison principle (Theorem 2.1.1). Let

$$v(s,t;K_1,K_2) \;=\; \frac{A(K_1)}{f_1(s)-f_2(s)} + g(t;K_1)\frac{f_1'(s)-f_2'(s)}{f_1(s)-f_2(s)} + K_2\,, \tag{2.63}$$

where

$$A(K_1) \;=\; \cos\gamma_1 + \cos\gamma_2 + K_1\,, \tag{2.64}$$

$$g(t;K_1) \;=\; -\frac{A}{-K_1/3+A}\sqrt{1-\left(\frac{\cos\gamma_1(t+1)+\cos\gamma_2(t-1)}{2}-\frac{K_1}{6}t\right)^2}\,, \tag{2.65}$$

where we choose K_1 and K_2 appropriately to construct super-solution and sub-solution. The trick of this proof is to realize that both first and second terms of the formal asymptotic expansion (i.e. both A and $g(t)$) need to be modified to obtain super-solution and sub-solution. We first impose the following

conditions for K_1 so that Equations (2.64)-(2.65) behave reasonably:

$$|K_1| < |\cos\gamma_1 + \cos\gamma_2|\,, \tag{2.66}$$

$$|K_1| < 6(1 - |\cos\gamma_1|)\,, \tag{2.67}$$

$$|K_1| < 6(1 - |\cos\gamma_2|)\,. \tag{2.68}$$

We put the restriction on the choice of K_1 so that the sign of $A(K_1)$ only depends on the sign of $\cos\gamma_1 + \cos\gamma_2$. Also, if K_1 is chosen to satisfy (2.66)-(2.68) then $g(t, K_1)$ is real and bounded. After some lengthy calculation assuming (2.37)-(2.39), we obtain the followings:

$$\vec{\nu}_1 \cdot Tv|_{t=1} = \cos\gamma_1 + \frac{K_1}{3} + o(1)\,, \tag{2.69}$$

$$\vec{\nu}_2 \cdot Tv|_{t=-1} = \cos\gamma_2 + \frac{K_1}{3} + o(1)\,, \tag{2.70}$$

$$\nabla \cdot Tv - v = -\frac{K_1}{3}\frac{1}{f_1(s) - f_2(s)} - K_2 + o\left(\frac{f_1'(s) - f_2'(s)}{f_1(s) - f_2(s)}\right)\,, \tag{2.71}$$

as $s \to 0^+$. The essential observation at this step of the proof is that equations (2.69) and(2.70) do not depend on K_2 including the "small o" terms. Similarly, Equation (2.71) has K_2 dependence only at the second term and not in the "small o" term.

We now construct a super-solution v that satisfies Inequalities (2.16)-(2.19). We denote super-solution by v^+ and associated constants by K_1^+ and K_2^+, i.e. $v^+ = v(s, t; K_1^+, K_2^+)$. Firstly, K_1^+ are chosen to be a small enough positive real number so as to satisfy (2.66)-(2.68). Secondly, we choose a constant $s_0^+ > 0$ so that for all $s < s_0^+$ the following inequalities satisfy:

$$\vec{\nu}_1 \cdot Tv^+\big|_{t=1} - \cos\gamma_1 > 0\,, \tag{2.72}$$

$$\vec{\nu}_2 \cdot Tv^+\big|_{t=-1} - \cos\gamma_2 > 0\,, \tag{2.73}$$

$$\nabla \cdot Tv^+ - v^+ + K_2^+ < 0\,. \tag{2.74}$$

Based on our previous observation we note that the choice of s_0^+ is independent of K_2^+. Let Ω_0^+ be the

subdomain of Ω such that $s < s_0^+$. By adding a restriction on K_2^+ to be a positive real number, it follows from (2.74) that

$$\nabla \cdot Tv^+ - v^+ \quad < \quad 0 \qquad \text{in } \Omega_0^+ \,. \tag{2.75}$$

Note that v^+ now satisfies conditions (2.16)-(2.18) of the Concus Finn comparison principle. It remains to choose K_2^+ so as to satisfy condition (2.19). According to corollary 2.1.1, $u(s,t)$ is bounded at $s = s_0^+$. Hence there exists K_2^+ so that

$$v^+ \quad > \quad u \,, \qquad \text{on } s = s_0^+ \,. \tag{2.76}$$

Thus by the Concus Finn comparison principle (Theorem 2.1.1), we have shown that there exists Ω_0^+, K_1^+, K_2^+ such that

$$v^+(s,t;K_1^+,K_2^+) \quad > \quad u(s,t) \qquad \text{in } \Omega_0^+ \,. \tag{2.77}$$

Similarly we can construct a sub-solution $v^-(s,t;K_1^-,K_2^-)$ such that

$$v^-(s,t;K_1^-,k_2^-) \quad < \quad u(s,t) \qquad \text{in } \Omega_0^- \,. \tag{2.78}$$

Hence we have

$$v^- \; < \; u \; < \; v^+ \qquad \text{in } \Omega_0^+ \cap \Omega_0^- \,, \tag{2.79}$$

i.e.,

$$\frac{A(K_1^-)}{f_1(s)-f_2(s)} + g(t;K_1^-)\frac{f_1'(s)-f_2'(s)}{f_1(s)-f_2(s)} + K_2^- \; < \; u \; < \; \frac{A(K_1^+)}{f_1(s)-f_2(s)} + g(t;K_1^+)\frac{f_1'(s)-f_2'(s)}{f_1(s)-f_2(s)} + K_2^+ \tag{2.80}$$

in $\Omega_0^+ \cap \Omega_0^-$. As K_1^+ and K_1^- was chosen to satisfy (2.66), $A(K_1^+)$ and $A(K_1^-)$ has the same sign. Without

loss of generality assume $A(K_1^+) > 0$. Let

$$m_1(s) = A(K_1^+) + \left[\max_{-1<t<1} \left\{ g(t; K_1^+)(f_1'(s) - f_2'(s)) \right\} + K_2^+ (f_1(s) - f_2(s)) \right], \tag{2.81}$$

$$m_2(s) = A(K_1^-) + \left[\min_{-1<t<1} \left\{ g(t; K_1^-)(f_1'(s) - f_2'(s)) \right\} + K_2^- (f_1(s) - f_2(s)) \right]. \tag{2.82}$$

Since $(f_1'(s) - f_2'(s)), (f_1(s) - f_2(s)) \in o(1)$ and continuous, there exists $s_0 > 0$ so that $m_1(s), m_2(s) > 0$ for $s < s_0$. By choosing k_1 and k_2 as

$$k_1 = \max_{0<s<s_0} m_1(s), \tag{2.83}$$

$$k_2 = \min_{0<s<s_0} m_2(s), \tag{2.84}$$

we obtain (2.62). □

Note that the above proof holds for an arbitrarily small $|K_1^{\pm}|$. Hence it is natural to guess that $(\cos\gamma_1 + \cos\gamma_2)/(f_1(s) - f_2(s))$ is the correct leading order term of the asymptotic expansion. We now show that the leading order term of the formal asymptotic expansion is in fact the first order term of the asymptotic expansion of $u(s,t)$.

Theorem 2.2.2 (Leading order behaviour of $u(x,y)$)**.** *Let $u(x,y)$ be the solution of the boundary value problem (2.8)-(2.10). Assume that $f_1(s)$ and $f_2(s)$ satisfy the conditions (2.54)-(2.58). Then*

$$u(s,t) = \frac{\cos\gamma_1 + \cos\gamma_2}{f_1(s) - f_2(s)} + O\left(\frac{f_1'(s) - f_2'(s)}{f_1(s) - f_2(s)} \right) \qquad \text{as } s \to 0^+. \tag{2.85}$$

Proof. We let

$$v(s,t; K_3, K_4, K_5) = \frac{A}{f_1(s) - f_2(s)} + g(t, K_3) \frac{f_1'(s) - f_2'(s)}{f_1(s) - f_2(s)} + h(t; K_4) \frac{(f_1'(s) - f_2'(s))^2}{f_1(s) - f_2(s)} + K_5 \tag{2.86}$$

where

$$
\begin{aligned}
A &= \cos\gamma_1 + \cos\gamma_2\,, && (2.87)\\
g(t;K_3) &= -\sqrt{1-\left(\frac{\cos\gamma_1(t+1)+\cos\gamma_2(t-1)}{2}\right)^2} + K_3\,, && (2.88)\\
h(t;K_4) &= -\frac{A}{4}\left(\delta t+\frac{t^2}{2}\right) + \frac{1-\alpha}{2A}\left\{1-\left(\frac{\cos\gamma_1(t+1)+\cos\gamma_2(t-1)}{2}\right)^2\right\} + \frac{K_4}{2}t^2\,. && (2.89)
\end{aligned}
$$

We note that unlike for the proof of Theorem 2.2.1, we can choose K_3 and K_4 as any real numbers. After some calculation assuming (2.54)-(2.58), we obtain

$$
\begin{aligned}
\vec{\nu}_1\cdot Tv|_{t=1} &= \cos\gamma_1 + K_4\frac{(f_1'(s)-f_2'(s))}{(A^2+4(g'(t))^2)^{3/2}} + o(f_1'(s)-f_2'(s))\,, && (2.90)\\
\vec{\nu}_2\cdot Tv|_{t=-1} &= \cos\gamma_2 + K_4\frac{(f_1'(s)-f_2'(s))}{(A^2+4(g'(t))^2)^{3/2}} + o(f_1'(s)-f_2'(s))\,, && (2.91)\\
\nabla\cdot Tv - v &= \left\{\left(-12\frac{g'(t)t}{A^2+4(g'(t))^2} + \frac{4A^2}{(A^2+4(g'(t))^2)^{3/2}}\right)K_4 - K_3\right\}\frac{f_1'(s)-f_2'(s)}{f_1(s)-f_2(s)} - K_5 \\
&\quad + o\left(\frac{f_1'(s)-f_2'(s)}{f_1(s)-f_2(s)}\right)\,, && (2.92)
\end{aligned}
$$

as $s\to 0^+$.

We now construct a super-solution. Let v^+ denote the super-solution and K_3^+ K_4^+ and K_5^+ denote associate constants, i.e. $v^+ = v(s,t;K_3^+,K_4^+,K_5^+)$. We first choose a positive constant K_4^+. Then we choose a big enough constant K_3^+ so that

$$
\left\{\left(-12\frac{g'(t)t}{A^2+4(g'(t))^2} + \frac{4A^2}{(A^2+4(g'(t))^2)^{3/2}}\right)K_4^+ - K_3^+\right\} \;<\; 0 \qquad \text{for } -1<t<1\,. \tag{2.93}
$$

We now choose $0 < s_2^+$ so that

$$\vec{\nu}_1 \cdot Tv|_{t=1} - \cos\gamma_1 > 0 \tag{2.94}$$

$$\vec{\nu}_2 \cdot Tv|_{t=-1} - \cos\gamma_2 > 0 \tag{2.95}$$

$$\nabla \cdot Tv - v + K_5^+ < 0 \tag{2.96}$$

for $0 < s < s_2^+$. Let Ω_2^+ be the subdomain of Ω such that $s < s_2^+$. By Corollary 2.1.1, we know that $u(s_2^+, t)$ is bounded, hence there exists a large enough positive constant K_5^+ so that

$$v^+ > u \qquad \text{on } s = s_2^+ . \tag{2.97}$$

Thus by the Concus Finn comparison principle (Theorem 2.1.1) we have that

$$v^+ > u \qquad \text{in } \Omega_2^+ . \tag{2.98}$$

Similarly we can construct a sub-solution v^- by choosing suitable K_3^-, K_4^-, K_5^- and s_2^-. Thus we can bound the solution $u(s,t)$ by v^- and v^+, i.e.

$$v^- < u < v^+ \qquad \text{in } \Omega_2^+ \cap \Omega_2^- , \tag{2.99}$$

hence (2.85) holds. □

From this section, we conclude that the height of a capillary surface near a cusp is proportional to the reciprocal of the distance between two arcs forming the cusp, assuming these arcs satisfy (2.37)-(2.39).

2.2.4 Examples of Cusp Domains

In the previous subsection, we have shown the behaviour of the capillary surface near a cusp with certain assumptions for the shapes of boundaries (i.e. $f_1(x)$ and $f_2(x)$). Those assumptions expressed as (2.37)-

(2.39) or (2.54)-(2.58) are left in these forms in order to make the theorem as general as possible. On the other hand, it is hard to grasp what kind of cusps are allowed or not. In this subsection, we will show when the theorem is applicable and not applicable through examples.

It is easy to show that if $f_1(x)-f_2(x)$ can be written in the following form, they satisfy Equations (2.54)-(2.56).

$$f_1(x) - f_2(x) \quad = \quad c\, x^{a_0} \exp\left(\sum_{i=1}^{\infty} a_i x^{b_i}\right) , \tag{2.100}$$

where $c > 0$, $a_1 < 0$, $b_1 < 0$, $b_{i+1} > b_i$. An alternative way to write this is the following:

$$f_1(x) - f_2(x) \quad = \quad \exp\left(\int_c^x \frac{\sum_{i=0}^{\infty} \tilde{a}_i \zeta^{\tilde{b}_i}}{\sum_{i=0}^{\infty} a_i \zeta^{b_i}} d\zeta\right) , \tag{2.101}$$

where $c > 0$, $b_0 - \tilde{b}_0 \geq 1$, $b_{i+1} > b_i$, $a_0 > 0$ and $\tilde{a}_0 > 0$. As (2.54)-(2.56) are stricter requirements for $f_1(x)$ and $f_2(x)$ than (2.37)-(2.39), if $f_1(x) - f_2(x)$ can be written as (2.100) or(2.101), then they satisfy (2.37)-(2.39).

Note that equations (2.57)-(2.58) can be interpreted as saying some of the osculating cusps (a cusp with mutually tangential boundaries) are not allowed, and equation (2.53) can be interpreted as infinitely oscillating cusp boundaries are not allowed.

2.2.4.1 Fractional Power Cusp

We now consider a cusp that can be analyzed through the result of Scholz. Consider equation (2.101) and let $b_0 > 1$, $\tilde{a}_i = a_i\, b_i$ and $\tilde{b}_i = b_i - 1$, then we have

$$f_1(x) - f_2(x) \quad = \quad \tilde{c} \sum_{i=0}^{\infty} a_i x^{b_i} . \tag{2.102}$$

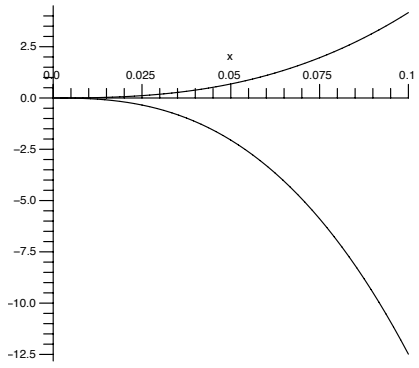


Figure 2.1: Fractional power cusp ($p = 1$, $q = -3$).

To be more specific, we consider the following cusp boundaries (see Figure 2.1)

$$f_1(x) = p\left(x^{5/2} + x^3\right), \tag{2.103}$$

$$f_2(x) = q\left(x^{5/2} + x^3\right), \tag{2.104}$$

where constants $p > q$. According to Theorem 2.2.2, we obtain the following asymptotic expansion:

$$u(x,y) = \frac{\cos\gamma_1 + \cos\gamma_2}{(p-q)\left(x^{5/2} + x^3\right)} + O(x^{-1}), \tag{2.105}$$

$$= \frac{\cos\gamma_1 + \cos\gamma_2}{p-q}\left(\frac{1}{x^{5/2}} - \frac{1}{x^2} + \frac{1}{x^{3/2}}\right) + O(x^{-1}). \tag{2.106}$$

as $x \to 0^+$. We note that this result is consistent with that of Scholz. It is noteworthy that by finding the first order term of our asymptotic expansion we find the first three terms of the asymptotic series solution in power series.

2.2.4.2 Exponential Cusp

We now consider cusps to which the results of Scholz do not apply. Equation (2.100) implies that $f_1(x)$ and $f_2(x)$ can contain exponential terms. We now consider a very sharp cusp, an "exponential cusp" where

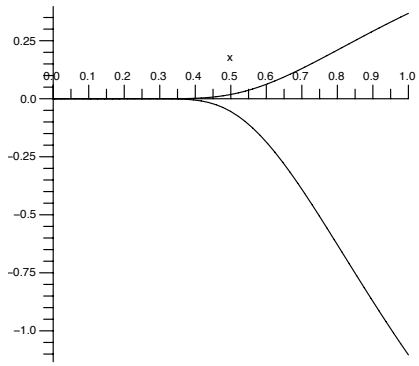


Figure 2.2: Exponential cusp ($p = 1$, $q = -3$).

$f_1(x)$ and $f_2(x)$ are chosen as follows (see Figure 2.2)

$$f_1(x) = p\, e^{-1/x^2}\,, \tag{2.107}$$

$$f_2(x) = q\, e^{-1/x^2}\,. \tag{2.108}$$

where constants $p > q$. According to Theorem 2.2.2, we obtain the following asymptotic expansion:

$$u(x,y) = \frac{\cos\gamma_1 + \cos\gamma_2}{p-q} e^{1/x^2} + O(x^{-3}) \qquad \text{as } x \to 0^+\,. \tag{2.109}$$

This example shows that our result has extended the result of Scholz on to the leading order behaviour of a capillary surface in a cusp domain.

2.2.4.3 Osculatory Cusp (Double Cusp)

We now consider the case where Theorem 2.2.2 cannot be applied. Consider the following cusp boundaries (see Figure 2.3)

$$f_1(x) = x^2 + px^3\,, \tag{2.110}$$

$$f_2(x) = x^2 + qx^3\,, \tag{2.111}$$

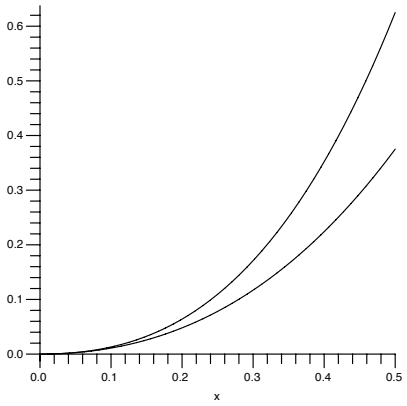


Figure 2.3: Osculatory cusp ($p = 3$, $q = 1$).

where constants $p > q$.

Above $f_1(x)$ and $f_2(x)$ do not satisfy (2.57)-(2.58), hence Theorem 2.2.2 does not apply. On the other hand, if $|\cos\gamma_1|, |\cos\gamma_2| \neq 1$, Theorem 2.2.1 applies as (2.110)-(2.111) satisfy (2.37)-(2.39). Hence even the case of the osculating cusp, we have shown that the height of the capillary surface rises as the same order as the reciprocal of the distance of two arcs forming a cusp, i.e.,

$$u(x,y) \quad = \quad \Theta\left(\frac{1}{x^3}\right). \tag{2.112}$$

We further investigate the solution behaviour of the Laplace-Young equation in the osculatory cusp in Chapter 4 by numerical approximation.

2.3 Bounded Capillary Surfaces

In this section we assume $\gamma_1 + \gamma_2 = \pi$ and aim to prove $u(x,y)$ is bounded.

2.3.1 Proof for the Boundedness of the Capillary Surface when $\gamma_1 + \gamma_2 = \pi$.

Theorem 2.3.1 (Boundedness of $u(x, y)$ when $\gamma_1 + \gamma_2 = \pi$)**.** *Let $u(x, y)$ be the solution of the boundary value problem (2.8)-(2.9) with $\gamma_1 = \gamma$ and $\gamma_2 = \pi - \gamma$. If the boundaries $\partial\Omega_1$ and $\partial\Omega_2$ have finite curvatures in the neighbourhood of the cusp, in other words, if there exists ϵ_o such that*

$$f_1(x),\ f_2(x) \quad \in \quad C^2([0, \epsilon_o])\,, \tag{2.113}$$

then $u(x, y)$ is bounded.

Proof. It follows immediately from Corollary 2.1.1 that $u(x, y)$ is bounded in the domain away from the origin. Hence our problem reduces to show that $u(x, y)$ is bounded in the neighbourhood of the origin.

First we show that $u(x, y)$ is bounded above at the origin by using the Concus Finn Comparison Principle (Theorem 2.1.1). In order to apply Theorem 2.1.1, we need to construct a surface that satisfies (2.16)-(2.19). The most difficult part of this proof is to construct a surface that satisfies both (2.17) and (2.18). Our unique idea is to construct a surface that satisfies (2.9) exactly hence (2.17) and also satisfies (2.18). Such surface can be constructed by a surface with contour lines parallel to the boundary $\partial\Omega_1$. In other words by letting the height of the surface only depends on the distance from the boundary $\partial\Omega_1$, we can easily construct a surface with exact constant contact angle γ on this boundary. We choose a surface so that the height and the mean curvature is bounded so that Inequalities (2.16) and (2.19) can easily be satisfied by shifting this surface upwards.

We now translate the above statement to the precise language of mathematics. Without loss of generality we assume $0 \leq \gamma \leq \pi/2$. First we define a coordinate system such that the one family of the coordinate curves is parallel curves of the boundary $\partial\Omega_1$ and another family of the coordinate curves is lines perpendicular to the boundary $\partial\Omega_1$. Let s and t be new coordinate variables defined implicitly as the following (note that s here has different meaning from s used in Section 2.2):

$$(x, y) \quad = \quad (s, f_1(s)) - t\,\vec{\nu}_1(s)\,, \tag{2.114}$$

where $\vec{\nu}_1(s)$ is the exterior unit normal vector of the boundary $\partial\Omega_1$ at $(s, f_1(s))$. More explicitly, the coordinate variables of Cartesian coordinate system x and y can be written using the new coordinate variables s and t as the followings:

$$x = s + t\frac{f_1'(s)}{\sqrt{1+(f_1'(s))^2}}, \tag{2.115}$$

$$y = f_1(s) - t\frac{1}{\sqrt{1+(f_1'(s))^2}}. \tag{2.116}$$

The coordinate variable t can be interpreted as the distance of the point from the boundary $\partial\Omega_1$. The coordinate curves are sketched in Figure 2.4(a).

The Jacobian of (2.115)-(2.116) is calculated to be

$$\frac{\partial(x,y)}{\partial(s,t)} = \frac{f_1'(s)^2-1}{\sqrt{1+(f_1'(s))^2}}\left(1 + t\frac{f_1''(s)}{(1+(f_1'(s))^2)^{3/2}}\right). \tag{2.117}$$

This gives that the point (x, y) in the Cartesian coordinate system can be specified uniquely by the new coordinate variables (s, t) defined by (2.115)-(2.116) if both of the following hold:

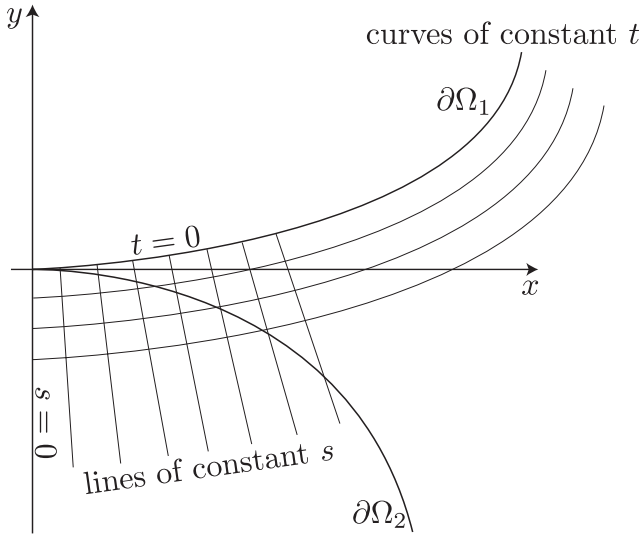
$$f_1'(s)^2 - 1 \neq 0, \tag{2.118}$$

$$1 + t\frac{f_1''(s)}{(1+(f_1'(s))^2)^{3/2}} \neq 0. \tag{2.119}$$

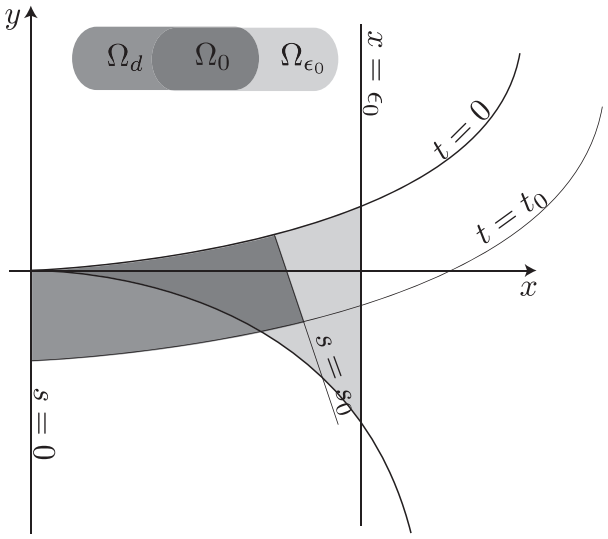
Since $f_1(s) \in C^2([0, \epsilon_o])$ and $\lim_{s\to 0^+} f_1(s) = 0$, there exists $0 < s_0 \leq \epsilon_0$ so that (2.118) is satisfied for all $s \in [0, s_0]$. Also due to the smoothness of $f_1(s)$, we can find $t_0 > 0$ such that (2.119) holds for all $t \in [0, t_0]$ in $s \in [0, s_0]$. That is to say the coordinate system defined in (2.115) and (2.116) is valid in the following domain Ω_d

$$\Omega_d := \{(s, f_1(s)) - t\,\vec{\nu}_1(s) \in \mathbb{R}^2 : 0 \leq s \leq s_0, 0 \leq t \leq t_0\}. \tag{2.120}$$

Then we choose the subdomain Ω_0 to be $\Omega_0 := \Omega_d \cap \Omega_{\epsilon_0}$ where $\Omega_{\epsilon_0} := \{(x, y) \in \mathbb{R}^2 : 0 < x < \epsilon_0, f_2(x) < y < f_1(x)\}$ as depicted in Figure 2.4(b). We notice that $\bar{\Omega}_0$ contains the cusp at the origin, hence finding



(a) coordinate lines of the s-t coordinate system



(b) domain Ω_0

Figure 2.4: s-t coordinate system.

an upper bound of the surface u in domain Ω_0 by using Theorem 2.1.1 would prove the capillary surface is bounded above at the cusp. We now construct a surface in Ω_0 using the parameters t and s as the following:

$$v^+(s,t) \;=\; \begin{cases} x(s,t) &= s + t\frac{f_1'(s)}{\sqrt{1+(f_1'(s))^2}}\,, \\ y(s,t) &= f_1(s) - t\frac{1}{\sqrt{1+(f_1'(s))^2}}\,, \\ z(s,t) &= g(t)\,. \end{cases} \tag{2.121}$$

Choice of the height function $g(t)$ depends on the contact angle γ. The simplest choice of the function $g(t)$ so that the surface v^+ satisfies (2.9) exactly and also satisfies (2.18), in our opinion, is the following:

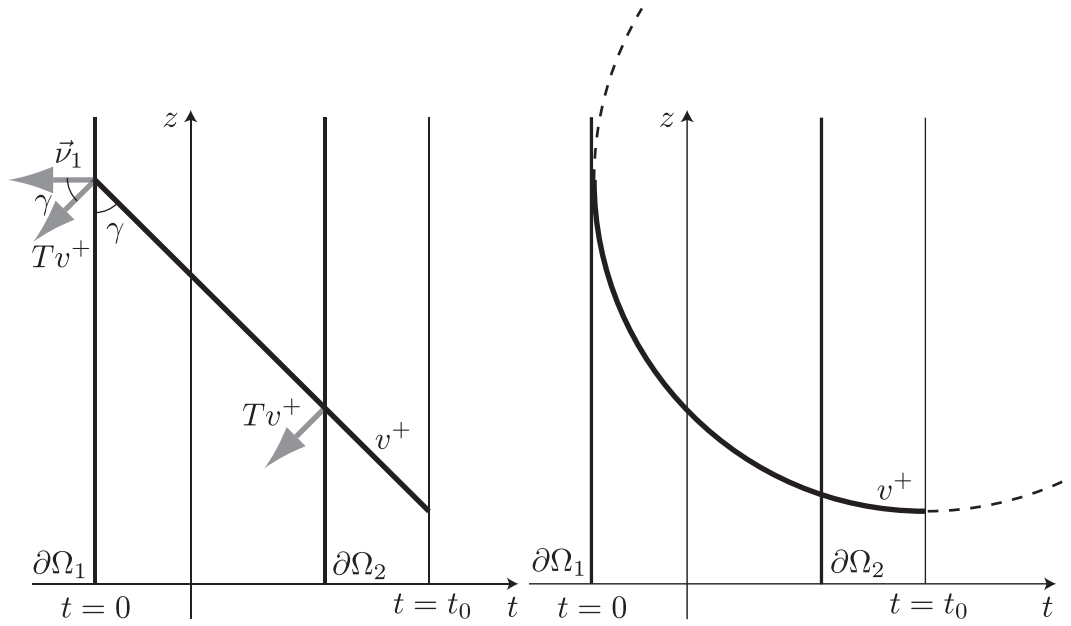
$$g(t) \;=\; \begin{cases} -\cot\gamma\, t + K & \text{for } \gamma \neq 0\,, \\ -\sqrt{t_0^2 - (t-t_0)^2} + K & \text{for } \gamma = 0\,, \end{cases} \tag{2.122}$$

where K is a constant that we will specify later. The cross section of this surface on a line of constant s is depicted in Figures 2.5(a) and 2.5(b). The surface $v^+(s,t)$ can be sketched as in Figures 2.6(a) and 2.6(b). For example, if the curve $\partial\Omega_1$ is a part of a circle, then the surface $v^+(s,t)$ for the case $\gamma \neq 0$ becomes a part of a cone, and for the case $\gamma = 0$ it becomes a part of a torus. We now verify that the surface $v^+(s,t)$ satisfies (2.9) exactly and also satisfies (2.18). We first consider the case $\gamma \neq 0$, as the vector Tv^+ can be interpreted as a unit downwards vector of the surface v^+, it follows immediately from Figure 2.5(a) that $Tv^+(s,t)$ can be written as

$$Tv^+ \;=\; \cos\gamma\,\vec{\nu}_1 - \sin\gamma\hat{z}\,, \tag{2.123}$$

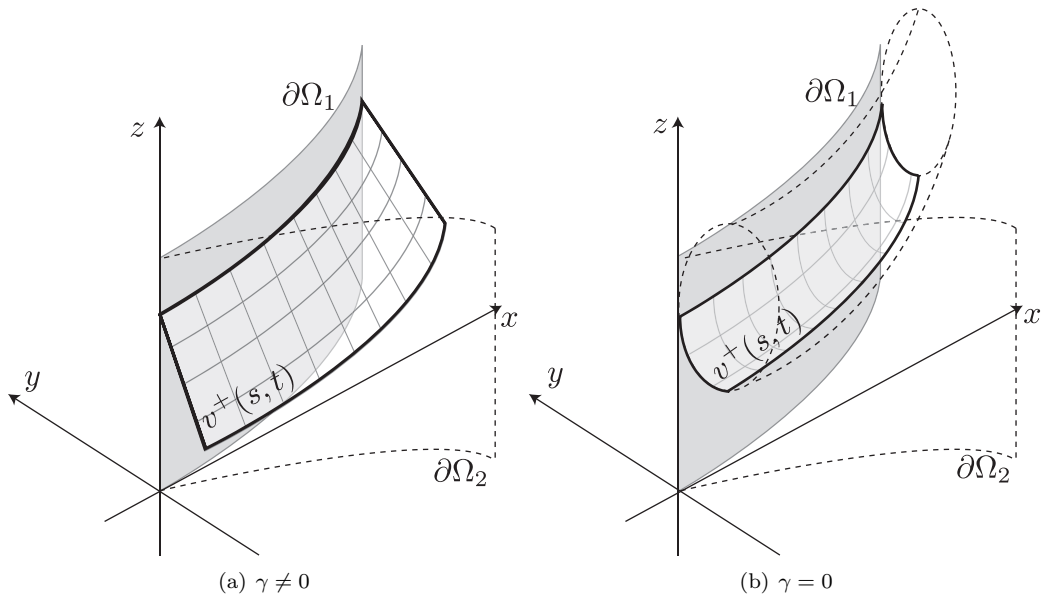
where $\hat{z}$ is a unit vector in z direction. Noting that the vector $\vec{\nu}_1$ is orthogonal to $\hat{z}$, we obtain that (2.9) is satisfied exactly by the surface $v^+(s,t)$, i.e.,

$$\vec{\nu}_1 \cdot Tv^+ \;=\; \cos\gamma \qquad \text{on } \partial\Omega_1 \cap \partial\Omega_0\,. \tag{2.124}$$



(a) Choice of function $g(t)$ for $\gamma \neq 0$ (b) Choice of function $g(t)$ for $\gamma = 0$

Figure 2.5: Cross section of a surface $v^+(s,t)$ on the line of constant s



(a) $\gamma \neq 0$ (b) $\gamma = 0$

Figure 2.6: Sketch of the surface $v^+(s,t)$

We now verify that the surface $v^+(s,t)$ satisfies Inequality (2.18). By noticing $\vec{\nu}_2$ and $\hat{z}$ are orthogonal and both $\vec{\nu}_1$ and $\vec{\nu}_2$ are unit vectors, we obtain the following inequality:

$$\begin{aligned} \vec{\nu}_2 \cdot Tv^+ &= \cos\gamma \vec{\nu}_1 \cdot \vec{\nu}_2 \,, && (2.125) \\ &> -\cos\gamma \,, && (2.126) \\ &= \cos(\pi - \gamma) \,. && (2.127) \end{aligned}$$

Although the case of $\gamma = 0$ may look complicated, it follows immediately from Figure 2.5(b) that the angle between the unit downward normal vector of v^+ and $\vec{\nu}_1$ are parallel on the boundary thus

$$\begin{aligned} \vec{\nu}_1 \cdot Tv^+ &= 1 \qquad \text{on } \partial\Omega_1 \cap \partial\Omega_0 \,, && (2.128) \\ &= \cos 0 \qquad \text{on } \partial\Omega_1 \cap \partial\Omega_0 \,. && (2.129) \end{aligned}$$

Also it follows immediately from the definition of the differential operator T that $|Tv^+| \leq 1$ (cf. Equation (2.14)). By noting that $\vec{\nu}_2$ is a unit vector, i.e., $|\vec{\nu}_2| = 1$, we have

$$\begin{aligned} \nu_2 \cdot Tv^+ &> -1 \,, && (2.130) \\ &= \cos(\pi - 0) \,. && (2.131) \end{aligned}$$

Hence the surface $v^+(s,t)$ defined by (2.121)-(2.122) satisfies Inequalities (2.17) and (2.18). We now show that the surface $v^+(s,t)$ satisfies Equations (2.16) by choosing large enough constant K.

Since $\nabla \cdot Tv^+$ is twice the mean curvature of the surface v^+, we can calculate $\nabla \cdot Tv^+$ as follows (cf. [31]):

$$\nabla \cdot Tv^+ = -2H(v^+) = -\frac{EN + GL - 2FM}{EG - F^2} \,, \qquad (2.132)$$

where

$$E = (x_s)^2 + (y_s)^2 + (z_s)^2 \,, \tag{2.133}$$

$$F = x_s x_t + y_s y_t + z_s z_t \,, \tag{2.134}$$

$$G = (x_t)^2 + (y_t)^2 + (z_t)^2 \,, \tag{2.135}$$

and

$$L = \frac{1}{\sqrt{EG - F^2}} \begin{vmatrix} x_{ss} & y_{ss} & z_{ss} \\ x_s & y_s & z_s \\ x_t & y_t & z_t \end{vmatrix} \,, \tag{2.136}$$

$$M = \frac{1}{\sqrt{EG - F^2}} \begin{vmatrix} x_{st} & y_{st} & z_{st} \\ x_s & y_s & z_s \\ x_t & y_t & z_t \end{vmatrix} \,, \tag{2.137}$$

$$N = \frac{1}{\sqrt{EG - F^2}} \begin{vmatrix} x_{tt} & y_{tt} & z_{tt} \\ x_s & y_s & z_s \\ x_t & y_t & z_t \end{vmatrix} \,. \tag{2.138}$$

After some calculation we obtain the following:

$$\nabla \cdot Tv^+ = \left(\frac{g_1''(t)}{(1 + (g'(t))^2)^{3/2}} + \frac{f_1''(s)}{(1 + (f_1'(s))^2)^{3/2} \left(1 + t \frac{f_1''(s)}{(1+(f_1'(s))^2)^{3/2}}\right)} \frac{g'(t)}{\sqrt{1 + (g'(t))^2}} \right) . \tag{2.139}$$

Recalling that we have chosen the domain Ω_0 so that Equation (2.119) satisfies in Ω_0 and $f_1''(s) \in C^2([0, \epsilon_o])$, in order to show $\nabla \cdot Tv^+$ is bounded, all we need is to show $\frac{g_1''(t)}{(1+(g'(t))^2)^{3/2}}$ is bounded, that is to say the curvature of the curve $g(t)$ is bounded. For the case of $\gamma \neq 0$, we have chosen $g(t)$ to be a linear function, thus $g''(t)$ is zero hence $\nabla \cdot Tv^+$ is bounded. For the case of $\gamma = 0$, we have chosen $g(t)$ to be the part of a circle with radius t_0 hence $\frac{g_1''(t)}{(1+(g'(t))^2)^{3/2}} = \frac{1}{t_0}$, this gives that $\nabla \cdot Tv^+$ is also bounded for this

case. We now consider the quantity $\nabla \cdot Tv^+ - v^+$, and it can be written as the following:

$$\nabla \cdot Tv^+ - v^+ \quad = \quad \nabla \cdot Tv^+ - (g(t) + K)\,. \tag{2.140}$$

It follows immediately from the choice of $g(t)$ that it is bounded in domain $\bar{\Omega}_0$ and also we have shown that twice the mean curvature $\nabla \cdot Tv^+$ is bounded and does not depend on K. Hence there exists a constant K_0 so that

$$\nabla \cdot Tv^+ - v^+ \quad = \quad \nabla \cdot Tv^+ - (g(t) + K) \leq 0 \qquad \text{for all } K \geq K_0. \tag{2.141}$$

Thus we have shown that the surface v^+ satisfies the Inequality (2.16) when K is chosen bigger than K_0.

We now put the last piece of the puzzle together by showing v^+ satisfies (2.19) for an appropriate choice of the constant K. Corrollary A.1 gives that the capillary surface u is bounded away from the cusp, hence it is bounded on $\partial\Omega_0 \backslash (\partial\Omega_1 \cup \partial\Omega_2 \cup \{(0,0)\})$. As $g(t)$ is bounded in the domain $\bar{\Omega}_0$, there exists a constant $K_1 \geq K_0$ such that $g(t) + K_1 > u$ on $\partial\Omega_0 \backslash (\partial\Omega_1 \cup \partial\Omega_2 \cup \{(0,0)\})$. Thus the surface v^+ satisfies (2.19) by choosing $K = K_1$.

As we have shown that the surface $v^+(s,t)$ defined in Equations (2.121)-(2.122) satisfies Inequalities (2.16)-(2.19), by the Concus-Finn comparison principle we have

$$v^+(s,t) \geq u(x,y) \qquad \text{in } \Omega_0\,. \tag{2.142}$$

Therefore the capillary surface at the cusp is bounded above when $\gamma_1 + \gamma_2 = \pi$ and each boundary $(\partial\Omega_1, \partial\Omega_2)$ has a finite curvature near cusp.

We can follow the similar steps for constructing the sub-surface to show that this capillary surface is bounded below. We first construct a coordinate system such that one of the families of the coordinate curves is parallel curves of the boundary $\partial\Omega_2$ and another is perpendicular lines of the boundary $\partial\Omega_2$. Then choose a surface v^- so that the heigh only depends on the distance from $\partial\Omega_2$ which satisfies the contact angle condition exactly on $\partial\Omega_2$ and also it satisfies $\vec{\nu}_1 \cdot Tv^- - \cos\gamma \leq 0$. By choosing v^- to have

the bounded height and the finite mean curvature, we can shift this surface downwards enough to satisfy $\nabla \cdot Tv^- - v^- \geq 0$ in Ω_0 and $v^- \leq u$ on $\partial\Omega_0 \backslash (\partial\Omega_1 \cup \partial\Omega_2 \cup \{(0,0)\})$. Then using the Concus Finn comparison principle, we can prove that $u(x,y)$ is bounded below.

Thus by showing that there exits bounded sub and super solutions of the Laplace-Young capillary surface equation, we have proven that the capillary surface is bounded if the contact angles of the boundaries are supplementary angles and boundaries have finite curvatures near the cusp. □

2.4 Concluding Remarks on Chapter 2

In this chapter we have shown that the statement "[the capillary surface] rises with the same order like the order of contact of the two arcs, which form the cusp" [40] is not only applicable to the power series cusp and also can be extended further. Our proof directly uses the functions $f_1(x)$ and $f_2(x)$ without approximating them by series. This idea has given us an advantage in the sense that our leading order term expression gives clearer intuitive understanding of the relationship between the shape of the domain and the shape of the singular capillary surface. Also as shown in an Example in Subsection 2.2.4.1, our leading order term gives first three terms of the power series asymptotic expansion, owing to the fact we have avoided approximating the boundary by the power series.

Even though we have extended the results beyond the power series cusps, our results still suffer from minor restrictions (2.54)-(2.58). Also a complete asymptotic series solution maybe desirable in order to claim the complete understanding of the asymptotic behaviour; however, it will require further assumptions to the boundary functions f_1 and f_2. The author would suspect the functions f_1 and f_2 of a form similar to the right hand side of Equation (2.100) can be a potential candidate for a type of cusp for which a complete asymptotic series can be determined.

Also we have shown the previously unknown phenomenon of the bounded capillary surface in a cusp domain is possible when the contact angles of the two walls are supplementary angles (i.e., $\gamma_1 + \gamma_2 = \pi$). Although our proof covers most of the cases when the boundaries are smooth except at the cusp, when the curvature of the boundary is not finite at the cusp, the behaviour of the capillary surface is unknown.

For example, it is unknown if the capillary surface is bounded or unbounded in a cusp domain bounded by $f_1 = x^{3/2}$ and $f_2 = -x^{3/2}$ when the contact angles of the two walls are supplementary angles.

This phenomena, the capillary surface can be bounded or unbounded in a cusp domain depending on the contact angle can be interesting physically as this result indicates that the gradual change of the contact angle (e.g. by changing the temperature of the liquid) can cause the dramatical change in the liquid surface from unbounded to bounded. However, as the bounded capillary surface in a cusp domain only appears when the contact angles are exactly supplementary, it is unknown to the author how easily this phenomena can be observed through an experiment.

Thus we end this chapter by remarking that the further exploration of the singular capillary surfaces through theoretical, experimental and possibly numerical analyses is desired.

Chapter 3

Augmented Finite Volume Element Method

The accuracy of a finite element numerical approximation of the solution of a partial differential equation can be spoiled significantly by singularities. This phenomenon is especially critical for high order methods. In this chapter, we show that the Finite Volume Element Method (FVEM) can be used to approximate singular solutions of linear elliptic PDEs with high-order accuracy by an augmentation procedure. If the PDE is linear and the singular basis functions are homogeneous solutions of the PDE, the augmentation of the trial function space can be done in a simple and elegant way. When the trial function space is augmented for the FVEM, all the entries in the matrix originating from the singular basis functions in the discrete form of the PDE are zero, and the singular basis functions only appear in the boundary conditions. That is to say, there is no need to integrate the singular basis functions over the elements and the sparsity of the matrix is preserved without special care. FVEM numerical convergence studies on two-dimensional triangular grids are presented using basis functions of arbitrary high order, confirming the same order of convergence for singular solutions as for smooth solutions.

Since there are many names and variations for the FVEM, the origin of this method is not entirely

clear; however, the paper by Bank and Rose [3] is usually cited as one of the first papers on this topic. About a decade after Bank and Rose's publication, this method was extended to a quadratic method by Liebau [29]. Later Li, Chen and Wu [28] and Xu and Zou [51] have carried out a convergence analysis for the quadratic method, see also [6]. Error estimates in the L_2 norm were given in [7]. Recently, Plexousakis and Zouraris [35] have proven a priori error estimates for the high order FVEM for one-dimensional problems (the ordinary differential equation case) with order higher than two. However, to our best knowledge, theoretical convergence results on the FVEM with arbitrary order piecewise polynomial basis functions for linear elliptic problems in two-dimensional domains are currently not yet available in the literature.

It is well-known that a solution of the Poisson equation with an analytic right hand side and analytic Dirichlet boundary data is not necessary analytic up to the boundary if the boundary is not smooth. Again, it is not easy to pin down the first discovery of such singular behaviour; we refer to the books by Grisvard [17, 18] for detailed discussion. The idea of augmentation of the trial function space for the finite element method (FEM) using results from regularity analysis can be found in [17] (Chapter 8.4.2) as well as in [46] (Chapter 8). Although it has been shown that the augmentation of the trial function space can recover the optimal convergence rate of the FEM, additional singular basis functions in the trial function space introduce complications. Integration of the singular basis function must be done analytically or using special quadrature rules, and special care is needed to preserve the sparsity of the stiffness matrix.

Error estimates for the first order FVEM for elliptic PDEs with a derivative blow-up singularity in a non-convex domain are presented by Chatzipantelidis and Lazarov [5]. They show that the rate of error convergence decreases when a singularity is present. Djadel et al. [12] have employed a grid refinement technique, similar to what is presented in Chapter 8.4.1 of [17], for the first order FVEM to improve the rate of convergence. On the other hand, to the best of our knowledge, augmentation of the trial function space has not been studied for the FVEM.

In this chapter, we demonstrate the augmentation of the trial function space for the arbitrary high order FVEM by presenting numerical convergence studies for the Poisson equation with derivative blow-up singularity at a reentrant corner. We first perform a numerical convergence study of the FVEM with arbitrary order piecewise polynomial basis functions on triangular grids for non-singular solutions using

the systematic way to construct control volumes that was proposed in [48], and that is a generalization of second order approaches in [29, 28, 51], see also [49]. Well-posedness of the FVEM on triangular grids for this particular way of constructing the control volumes was proved for second order methods in [51] under certain conditions on the angles of the triangular elements. For orders higher than two, it can be observed that H^1 convergence orders are the same as the optimal orders exhibited by the Galerkin FEM, but that L_2 convergence order is sub-optimal for polynomials of even order, similar to what has been observed before for the Discontinuous Galerkin method [34, 27, 19] and for the one-dimensional FVEM [35], see also [49]. We then show numerically that the presence of a derivative blow-up singularity at a reentrant corner can pollute the numerical solution and the rate of convergence. Finally we show numerically how the rate of convergence can be recovered by augmentation of the trial function space, leading to an elegant and efficient augmented high order FVEM.

3.1 Model Problems

In this section, we consider the Poisson equation with Dirichlet boundary condition

$$\nabla^2 u = f(x,y) \qquad \text{in } \Omega\,, \tag{3.1}$$

$$u = g(x,y) \qquad \text{on } \partial\Omega\,, \tag{3.2}$$

where $\Omega \subset \mathbb{R}^2$ is an open polygonal domain with a finite number of vertices, $\partial\Omega$ is the boundary of the polygonal domain, $f(x,y)$ is a function in $L^2(\Omega) \cap C^0(\Omega)$ and $g(x,y)$ is a function in $H^2(\partial\Omega)$. Note that, for simplicity, we only consider classical solutions of the PDE in this chapter ($f \in C^0(\Omega)$). The FVEM can also be used to approximate non-classical solutions, see [3].

3.1.1 Regularity of Solutions of the Poisson Problem

Since a polygonal domain is a Lipschitz domain, showing the existence and the interior smoothness of the solution of boundary value problem (BVP) (3.1)-(3.2) is straightforward. It can be proven by the

Lax-Milgram Theorem that there exists a unique solution $u \in H^1(\Omega)$ of BVP (3.1)-(3.2) (See Lemma 4.4.3.1 of [17]). Also, it is known that boundary derivative blow-up singularities can occur at the vertices of the domain. We now give a known regularity result for the solution at the vertices.

Let $(x_{s_i}, y_{s_i}) \in \partial\Omega$ be the vertices with singularities of the boundary of domain Ω, and let N_{vert} be the number of vertices with singularities. For each vertex (x_{s_i}, y_{s_i}), local polar coordinates r_i and θ_i are defined as depicted in Figure 3.1. Define the interior angle α_i for each vertex (x_{s_i}, y_{s_i}) so that $(r_i, \theta_i) \in \Omega$ if $0 < \theta_i < \alpha_i$ for sufficiently small $r_i > 0$. Using the above notation, the modified shift theorem can be stated as follows.

Theorem 3.1.1 (Modified Shift Theorem)**.** *Let $u(x, y)$ be the solution of the following boundary value problem:*

$$\nabla^2 u = f(x, y) \qquad \text{in } \Omega\,, \tag{3.3}$$

$$u = g(x, y) \qquad \text{on } \partial\Omega\,, \tag{3.4}$$

where $\Omega \subset \mathbb{R}^2$ is an open polygonal domain with a finite number of vertices, $\partial\Omega$ is the boundary of the polygonal domain, and $f(x, y), g(x, y)$ are C^∞ functions in $\bar{\Omega}$ and on $\partial\Omega$, respectively.
If $m\alpha_i/\pi \notin \mathbb{N}$ for all i, then there exist constants $k_{i,j} \in \mathbb{R}$ such that

$$u(x, y) - \sum_{0<\lambda_{i,j}<m+1} k_{i,j}\psi_{i,j} \in H^{m+1}(\Omega)\,, \tag{3.5}$$

where

$$\psi_{i,j} = \begin{cases} r_i^{\lambda_{i,j}} \sin(\lambda_{i,j}\theta_i) & \text{if } \lambda_{i,j} \notin \mathbb{Z}\,, \\ r_i^{\lambda_{i,j}} \left\{\ln r_i \sin(\lambda_{i,j}\theta_i) + \theta_i \cos(\lambda_{i,j}\theta_i)\right\} & \text{if } \lambda_{i,j} \in \mathbb{Z}\,, \end{cases} \tag{3.6}$$

with

$$\lambda_{i,j} = j\pi/\alpha_i\,. \tag{3.7}$$

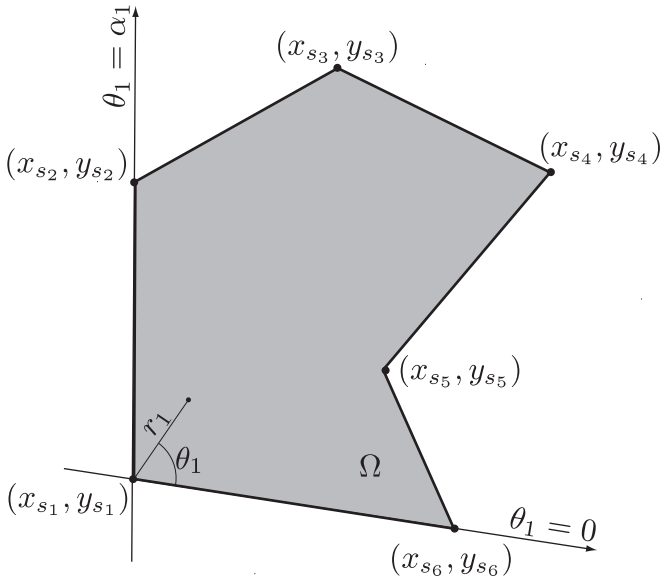


Figure 3.1: Example of a polygonal domain and the local polar coordinates for $i = 1$.

This theorem is a direct implication of Theorem 5.1.3.5 in Grisvard [17].

In order to illustrate the augmentation of the trial function space, we consider the following model problems.

3.1.2 Model Problem 1

Consider Poisson problem (3.1)-(3.2) with the following right hand side and boundary data:

$$f(x,y) = 20\,x^3y^4 + 12\,x^5y^2 \qquad \text{in } \Omega\,, \tag{3.8}$$

$$g(x,y) = x^5y^4 \qquad \text{on } \partial\Omega\,, \tag{3.9}$$

where domain Ω is a unit square domain. The exact solution is

$$u(x,y) = x^5y^4 \qquad \text{in } \Omega\,. \tag{3.10}$$

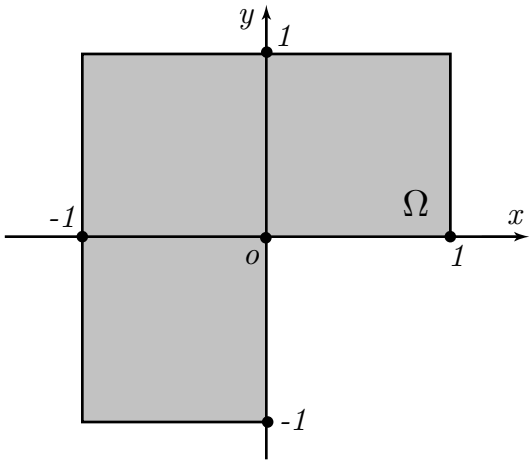


Figure 3.2: L-shaped domain.

3.1.3 Model Problem 2

Consider Poisson problem (3.1)-(3.2) with the following right hand side and boundary data:

$$f(x,y) = 20\,x^3y^4 + 12\,x^5y^2 \qquad \text{in } \Omega\,, \tag{3.11}$$

$$\begin{aligned} g(x,y) &= x^5y^4 + 2\,r^{\frac{2}{3}}\sin(\frac{2}{3}\theta) + 7\,r^{\frac{4}{3}}\sin(\frac{4}{3}\theta) + r^2\{\ln r\sin(2\theta) + \theta\cos(2\theta)\} \\ &\quad +8\,r^{\frac{8}{3}}\sin(\frac{8}{3}\theta) + 2\,r^{\frac{10}{3}}\sin(\frac{10}{3}\theta) + 8\,r^4\{\ln r\sin(4\theta) + \theta\cos(4\theta)\} \qquad \text{on } \partial\Omega\,, \end{aligned} \tag{3.12}$$

where domain Ω is as illustrated in Figure 3.2, and r and θ are polar coordinates centred at the origin. The exact solution is

$$\begin{aligned} u(x,y) &= x^5y^4 + 2\,r^{\frac{2}{3}}\sin(\frac{2}{3}\theta) + 7\,r^{\frac{4}{3}}\sin(\frac{4}{3}\theta) + r^2\{\ln r\sin(2\theta) + \theta\cos(2\theta)\} \\ &\quad +8\,r^{\frac{8}{3}}\sin(\frac{8}{3}\theta) + 2\,r^{\frac{10}{3}}\sin(\frac{10}{3}\theta) + 8\,r^4\{\ln r\sin(4\theta) + \theta\cos(4\theta)\} \qquad \text{in } \Omega\,. \end{aligned} \tag{3.13}$$

Note that the r-directional derivative of $u(x, y)$ blows up at the origin, but $g(x, y)$ is analytic on $\partial\Omega$ since $g(x, 0) = 0$ and $g(0, y) = (-3/2\, y^2 + 12\, y^4)\pi$.

3.2 Numerical Method

The FVEM approximates the solution of a BVP in a finite-dimensional trial function space by integrating the PDE over a finite number of control volumes. In this section, our construction of shape functions and control volumes is explained. Then, using the shape functions and the control volumes, the FVEM is built up using both a standard trial function space and an augmented trial function space.

3.2.1 Shape Functions

To approximate the solution of BVP (3.1)-(3.2), a finite-dimensional trial function space $S_p^h \subset H^1(\Omega)$ needs to be constructed. Similarly to the FEM, p^{th} order piecewise polynomial shape functions are used to form a basis of the trial function space. We construct the shape functions the same way as the FEM, as briefly described in the following. (See standard textbooks on the FEM, e.g., [4], for a more detailed discussion.)

Finite Element Triangulation of the Domain

First, a set of triangular open subdomains of Ω denoted by $\{T_1, T_2, ..., T_{N_e}\}$ is chosen in such a way that

$$T_i \cap T_j = \emptyset \quad \text{if } i \neq j\,, \tag{3.14}$$

$$\cup_{i=1,...,N_e} \bar{T}_i = \bar{\Omega}\,, \tag{3.15}$$

and no vertex of any triangle lies in the interior of an edge of another triangle [4]. We shall refer to these triangles as *element triangles*, with N_e denoting the number of element triangles. We let h be the maximum diameter of the element triangles and use it to quantitatively describe the resolution of the triangular mesh. The parameter h will be used in the convergence study of Section 3.3.

Node placement in the Finite Elements

In order to construct nodal basis functions, we now choose the location of nodes for each element triangle. Consider the element triangle T_i, and let $\{(x_1, y_1), (x_2, y_2), (x_3, y_3)\}$ be the vertices of this element triangle. The $(x, y) \in \bar{T}_i$ are defined by

$$x = (1 - \zeta - \eta)\, x_1 + \zeta\, x_2 + \eta\, x_3 \qquad \text{for } 0 \le \zeta + \eta \le 1 \text{ and } \eta, \zeta \ge 0\,, \tag{3.16}$$

$$y = (1 - \zeta - \eta)\, y_1 + \zeta\, y_2 + \eta\, y_3 \qquad \text{for } 0 \le \zeta + \eta \le 1 \text{ and } \eta, \zeta \ge 0\,. \tag{3.17}$$

We wish to construct a p^{th} order polynomial in $\bar{T}_i$ using a linear combination of nodal basis functions. Hence we require the same number of nodes in $\bar{T}_i$ as the degrees of freedom of a p^{th} order polynomial, i.e., $(p+1)(p+2)/2$. We define the coordinate of each node by choosing η and ζ as in the following (also depicted in Figure 3.3):

$$\eta = \frac{j}{p}\,, \qquad \zeta = \frac{l}{p}\,, \tag{3.18}$$

where

$$j, l = 0, 1, 2, ...p\,, \tag{3.19}$$

$$j + l \le p\,. \tag{3.20}$$

We let N_{node} be the total number of distinct nodes on the whole mesh and denote the location of these nodes by (x_i, y_i) for $i = 1, 2, ..., N_{node}$. Note that on the uniform grids we consider below $N_{node} = O(1/h^2)$. Let $\mathcal{N}$ be the set of node indices $\{1, 2, 3, ..., N_{node}\}$, and $\mathcal{N}_{int}$ and $\mathcal{N}_{bound}$ be disjoint subsets of $\mathcal{N}$ denoting the nodes in the interior and on the boundary of the domain, respectively, i.e.,

$$(x_i, y_i) \in \Omega^o \qquad \forall i \in \mathcal{N}_{int}\,, \tag{3.21}$$

$$(x_i, y_i) \in \partial\Omega \qquad \forall i \in \mathcal{N}_{bound}\,, \tag{3.22}$$

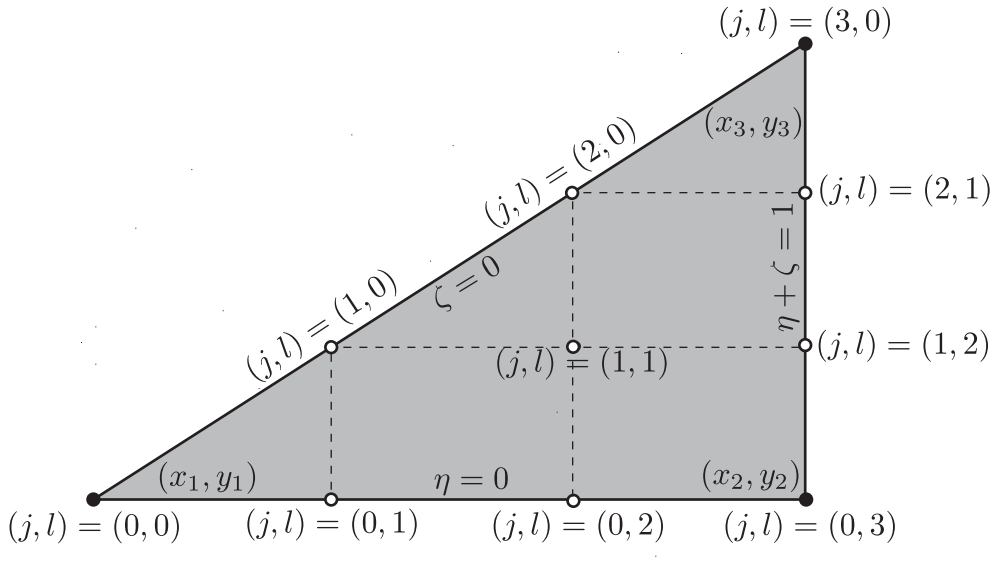


Figure 3.3: Placement of nodes in an element triangle ($p = 3$).

$$\mathcal{N}_{int} \cup \mathcal{N}_{bound} = \mathcal{N}\,. \tag{3.23}$$

We denote the number of nodes in $\mathcal{N}_{int}$ and $\mathcal{N}_{bound}$ by N_{int} and N_{bound}, respectively. In addition, we let $\mathcal{T}_i$ be the index set of all element triangles that contain node i in their closure.

Shape functions

We now construct a shape function $\phi_i(x, y)$ for each node with $\phi_i(x, y)$ p^{th} order piecewise polynomial in Ω and p^{th} order polynomial in T_j for all $j = 1, ..., N_e$, and with $\phi_i(x_j, y_j) = \delta_{i,j}$. It can be shown that $\phi_i(x, y) \in C^0(\Omega)$ and the set $\{\phi_1, \phi_2, ..., \phi_{N_{node}}\}$ is linearly independent. In addition we note that $\cup_{j \in \mathcal{T}_i} \bar{T}_j$ is the support of the shape function ϕ_i.

3.2.2 Control Volumes

We divide the domain Ω into a finite number of subdomains for finite volume integration. We follow the specific approach for constructing the control volumes for second order polynomials on triangles of Liebau [29] ([51] proves the inf-sup condition and convergence for, among others, this choice of control volumes, under certain conditions on the element angles), and we extend it to arbitrary high order in the same way as proposed in [48]. The subdomains are chosen so that the union of the closure of the subdomains is the closure of the domain, and the closure of each subdomain contains exactly one finite element node. We construct the subdomains by the following five steps (with each step depicted in Figure 3.4):

Step 1: Subdivide the element triangles in smaller triangles in a regular fashion, using the finite element nodes (see Figure 3.4).

Step 2: Determine the midpoint of each edge of the small triangles.

Step 3: Determine the centroids of the small triangles.

Step 4: Connect midpoints with centroids by line segments.

Step 5: Divide the domain Ω into N_{node} polygonal open subdomains enclosed by the line segments created in Step 4.

We shall refer to these polygonal open subdomains of Ω as *control volumes*. The control volume associated with the node at (x_i, y_i) is denoted as Ω_i.

3.2.3 FVEM with Standard Trial Function Space

We now formulate the discrete form of the boundary value problem. We first construct the standard trial function space using the shape functions and then discretize the PDE using the control volumes.

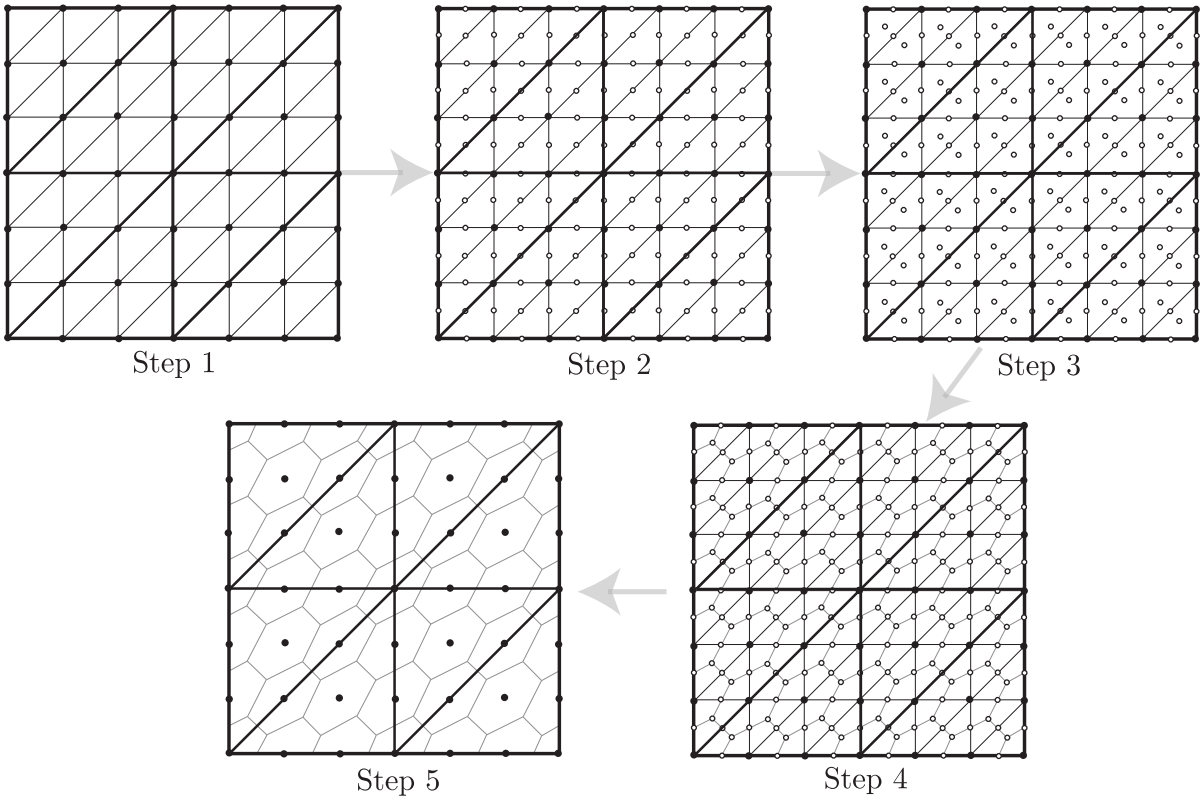


Figure 3.4: Construction of the control volumes for eight element triangles ($p = 3$).

Standard Trial Function Space

We construct a finite-dimensional function space S_p^h with the p^{th} order piecewise polynomial shape functions described in Section 3.2.1:

$$S_p^h \quad := \quad span\{\phi_1, \phi_2, ..., \phi_{N_{node}}\}\,. \tag{3.24}$$

We approximate the solution of the boundary value problem by a linear combination of the shape functions ϕ_i in S_p^h, the standard trial function space of our FVEM, i.e.,

$$u \approx u^h \quad := \quad \sum_{i=1}^{N_{node}} c_i\,\phi_i\,, \tag{3.25}$$

where $c_i \in \mathbb{R}$. We note that $S_p^h \subset W_1^2(\Omega) \subset H^1(\Omega)$.

Discrete Integral Form

The discrete integral form of PDE (3.1) can be motivated as follows. Since f is continuous in Ω, PDE (3.1) is equivalent to the following equation:

$$\int_{\Omega_\alpha} \nabla^2 u\, dA \quad = \quad \int_{\Omega_\alpha} f\, dA \qquad \text{for all open subdomains } \Omega_\alpha \subset \Omega. \tag{3.26}$$

Applying the divergence theorem we obtain

$$\int_{\partial\Omega_\alpha} \nu \cdot \nabla u\, ds \quad = \quad \int_{\Omega_\alpha} f\, dA \qquad \text{for all open subdomains } \Omega_\alpha \subset \Omega, \tag{3.27}$$

where ν is the unit outward normal vector on the boundary $\partial\Omega_\alpha$. We now approximate u by u^h and impose Equation (3.27) on the control volumes Ω_i associated with the interior nodes, and obtain

$$\sum_{j=1}^{N_{node}} c_j \int_{\partial\Omega_i} \nu \cdot \nabla\phi_j\, ds \quad = \quad \int_{\Omega_i} f\, dA \qquad \forall i \in \mathcal{N}_{int}\,, \tag{3.28}$$

where ν is the unit outward normal vector on the boundary of the control volume $\partial\Omega_i$. Note that $\nabla\phi_j$ in Equation (3.28) is well defined at all Gauss points for all $i \in \mathcal{N}_{int}$.

Boundary Conditions

Now consider boundary conditions (BCs) so that u^h satisfies BC (3.2) exactly at each boundary node, i.e.,

$$u^h(x_i, y_i) \;=\; \sum_{j=1}^{N_{node}} c_j \phi_j(x_i, y_i) \;=\; c_i \;=\; g(x_i, y_i) \qquad \forall i \in \mathcal{N}_{bound}\,. \tag{3.29}$$

This provides N_{bound} additional conditions which, together with the N_{int} equations of (3.28), fully specify the $N_{node} = N_{int} + N_{bound}$ unknown coefficients in linear combination (3.25).

FVEM with Standard Trial Function Space

The FVEM with the standard trial function space is thus given by:

Find $\{c_1, c_2, ..., c_{N_{node}}\}$ such that

$$\sum_{j=1}^{N_{node}} c_j \int_{\partial\Omega_i} \nu \cdot \nabla\phi_j \, ds \;=\; \int_{\Omega_i} f \, dA \qquad \forall i \in \mathcal{N}_{int}\,, \tag{3.30}$$

$$c_i \;=\; g(x_i, y_i) \qquad \forall i \in \mathcal{N}_{bound}\,. \tag{3.31}$$

We solve the system of Equations (3.30) and (3.31) by solving a matrix equation of the form $A \cdot \boldsymbol{x} = \boldsymbol{b}$ where A is a $N_{node} \times N_{node}$ matrix and $\boldsymbol{x}$ and $\boldsymbol{b}$ are vectors of length N_{node}. By rearranging the nodal indices so that

$$\mathcal{N}_{int} \;=\; \{1, 2, ..., N_{int}\}, \tag{3.32}$$

$$\mathcal{N}_{bound} \;=\; \{N_{int} + 1, N_{int} + 2, ..., N_{node}\}, \tag{3.33}$$

we obtain the matrix equation

$$\left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline 0 & I \end{array}\right] \cdot \left[\begin{array}{c} \boldsymbol{x}_1 \\ \hline \boldsymbol{x}_2 \end{array}\right] = \left[\begin{array}{c} \boldsymbol{b}_1 \\ \hline \boldsymbol{b}_2 \end{array}\right] \tag{3.34}$$

where

the i^{th} row, j^{th} column of the $N_{int} \times N_{node}$ matrix $[A_{11}\, A_{12}]$ is given by $\int_{\partial\Omega_i} \nu \cdot \nabla\phi_j \, ds$,

the i^{th} element of vector $[\boldsymbol{x}_1^T \boldsymbol{x}_2^T]^T$ is c_i,

the i^{th} element of vector $\boldsymbol{b}_1$ is $\int_{\Omega_i} f \, dA$,

and the i^{th} element of vector $\boldsymbol{b}_2$ is $g(x_{N_{int}+i}, y_{N_{int}+i})$.

Since the support of ϕ_j is $\cup_{k\in\mathcal{T}_j}\bar{T}_k$, the i^{th} row, j^{th} column of the matrix $[A_{11}\, A_{12}]$ is 0 if $\Omega_i \cap \big(\cup_{k\in\mathcal{T}_j}\bar{T}_k\big) = \emptyset$. Also, the number of non-zero elements of each row of the matrix $[A_{11} A_{12}]$ only depends on p. Hence the number of non-zeros of matrix A is $O(N_{node}) = O(1/h^2)$.

We shall refer to this approximation technique for the solution of the BVP with p^{th} order piecewise polynomials as the p^{th} *order FVEM*.

3.2.4 FVEM with Augmented Trial Function Space

The FVEM with an augmented trial function space can be formulated following similar steps to what was presented in Section 3.2.3. In this section, we describe the steps with special emphasis on how they avoid the integration of singular basis functions and preserve the sparsity of the matrix system. For simplicity, we assume there only exists one corner singularity (i.e., $N_{vert} = 1$). Extension to multiple points of singularity can be implemented easily. Due to our construction of the mesh and nodes, there is always a node at a point of singularity. We denote the location of the singularity by (x_s, y_s) (i.e., we let the index of the singular node be s).

Augmented Trial Function Space

In addition to the basis functions in the standard trial function space, we include the $\psi_{1,j}$ defined in Equation (3.6) in the basis of the trial function space. That is to say, the augmented trial function space $\hat{S}_p^h$ is defined as

$$\hat{S}_p^h \quad := \quad span\{\phi_1, \phi_2, ..., \phi_{N_{node}}, \psi_{1,1}, \psi_{1,2}, ..., \psi_{1,N_s}\}\,, \tag{3.35}$$

where N_s is the number of singular basis functions, which can be chosen according to Theorem 3.1.1 and the theory of polynomial interpolation. In our implementation, we choose $N_s = 2(p+1) - 1$ to guarantee that the solution is in H^{p+1} after subtracting a suitable linear combination of the singular basis functions, for any polygonal domain. It is beneficial to choose the $\psi_{1,j}$ as the singular basis functions, since they are harmonic functions (i.e., $\nabla^2 \psi_{1,j} = 0$) which leads to significant simplifications in the linear equations, see below. Similarly to Equation (3.25) we approximate u by $\hat{u}^h$ such that

$$u \approx \hat{u}^h \quad := \quad \sum_{i=1}^{N_{node}} c_i\, \phi_i + \sum_{i=1}^{N_s} k_i\, \psi_{1,i}\,, \tag{3.36}$$

where $c_i, k_i \in \mathbb{R}$. We note that $S_p^h \subsetneq \hat{S}_p^h$. The first few singular basis functions for Model Problem 2 are listed below:

$$\psi_{1,1} = r^{\frac{2}{3}} \sin(\frac{2}{3}\theta)\,, \tag{3.37}$$

$$\psi_{1,2} = r^{\frac{4}{3}} \sin(\frac{4}{3}\theta)\,, \tag{3.38}$$

$$\psi_{1,3} = r^2\{\ln r \sin(2\theta) + \theta\cos(2\theta)\}\,, \tag{3.39}$$

$$\vdots \quad .$$

Discrete Integral Form

Following Equation (3.27) the discrete integral form can be written as

$$\sum_{j=1}^{N_{node}} c_j \int_{\partial\Omega_i} \nu \cdot \nabla\phi_j \, ds + \sum_{j=1}^{N_s} k_j \int_{\partial\Omega_i} \nu \cdot \nabla\psi_{1,j} \, ds \quad = \quad \int_{\Omega_i} f \, dA \qquad \forall i \in \mathcal{N}_{int} \, . \tag{3.40}$$

By applying the divergence theorem, it can easily be seen that, for the singular basis functions $\psi_{1,j}$, it holds that

$$\int_{\Omega_i} \nabla^2\psi_{1,j} \, dA \quad = \int_{\partial\Omega_i} \nu \cdot \nabla\psi_{1,j} \, ds = 0 \qquad \forall i \in \mathcal{N}_{int} \, , \tag{3.41}$$

which means that Equation (3.40) can be simplified to

$$\sum_{j=1}^{N_{node}} c_j \int_{\partial\Omega_i} \nu \cdot \nabla\phi_j \, ds \quad = \quad \int_{\Omega_i} f \, dA \qquad \forall i \in \mathcal{N}_{int} \, , \tag{3.42}$$

which, surprisingly, is the same as Equation (3.28), i.e., the singular basis functions do not lead to extra terms in the equations for the interior nodes.

Boundary Conditions

Now consider the BCs so that $\hat{u}^h$ satisfies BC (3.2) exactly at each boundary node, i.e.,

$$\hat{u}^h(x_i, y_i) \; = \; \sum_{j=1}^{N_{node}} c_j \phi_j(x_i, y_i) + \sum_{j=1}^{N_s} k_j \psi_{1,j}(x_i, y_i) \; = \; g(x_i, y_i) \qquad \forall i \in \mathcal{N}_{bound} \, . \tag{3.43}$$

With $\phi_j(x_i, y_i) = \delta_{i,j}$, Equation (3.43) can be simplified as

$$c_i + \sum_{j=1}^{N_s} k_j \psi_{1,j}(x_i, y_i) \; = \; g(x_i, y_i) \qquad \forall i \in \mathcal{N}_{bound} \, . \tag{3.44}$$

Thus we see that the singular basis functions only appear in the boundary conditions. Also, to impose the boundary conditions there is no need for integrating the singular basis functions.

FVEM with Augmented Trial Function Space

We finally combine Equations (3.42) and (3.44) to formulate the FVEM. We have $N_{int}+N_{bound}$ equations for $N_{int}+N_{bound}+N_s$ unknowns, which is an underdetermined problem. Since we need N_s additional conditions, we make the natural choice of additionally imposing the integral form (3.42) for the control volumes Ω_i of the N_s boundary nodes that are closest to the point of singularity. Note that, in these integrals, the gradients on the control volume edges that are part of the domain boundary are to be evaluated in the limiting sense from the inside of the control volume.

Combining these equations, we seek $\{c_j\}_{j=1}^{N_{node}}$, $\{k_j\}_{j=1}^{s}$ satisfying

$$\sum_{j=1}^{N_{node}} c_j \int_{\partial\Omega_i} \nu\cdot\nabla\phi_j\, ds \;=\; \int_{\Omega_i} f\, dA \qquad \forall i\in\mathcal{N}_{int}\cup\mathcal{N}_{singular}\,, \tag{3.45}$$

$$c_i+\sum_{j=1}^{N_s} k_j\psi_{1,j}(x_i,y_i) \;=\; g(x_i,y_i) \qquad \forall i\in\mathcal{N}_{bound}\,, \tag{3.46}$$

where $\mathcal{N}_{singular}$ is the set of indices of the N_s boundary nodes closest to the point of singularity, chosen as depicted in Figure 3.5. The matrix equation $\hat{A}\hat{x}=\hat{b}$ corresponding to Equations (3.45)-(3.46) can be depicted as

$$\left[\begin{array}{cc|c} A_{11} & A_{12} & 0 \\ 0 & I & A_{23} \\ \hline A_{31} & A_{32} & 0 \end{array}\right]\cdot\left[\begin{array}{c} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \hline \boldsymbol{x}_3 \end{array}\right] \;=\; \left[\begin{array}{c} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \hline \boldsymbol{b}_3 \end{array}\right] \tag{3.47}$$

where matrices A_{11}, A_{12} and vectors $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are exactly as in (3.34), and

the j^{th} column of the $N_s\times N_{node}$ matrix $[A_{31}\,A_{32}]$ is $\int_{\partial\Omega_i}\nu\cdot\nabla\phi_j\,ds$ for $i\in\mathcal{N}_{singular}$,

the i^{th} row, j^{th} column of the $N_{bound}\times N_s$ matrix A_{23} is $\psi_{1,j}(x_{N_{int}+i},y_{N_{int}+i})$,

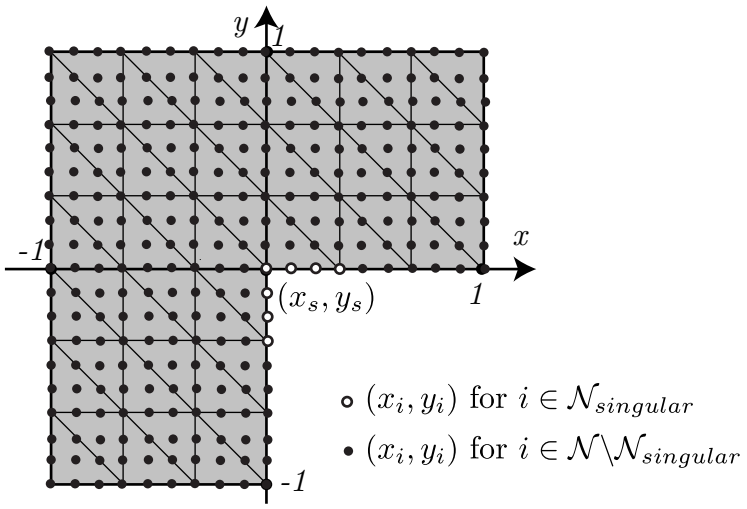


Figure 3.5: Placement of $\mathcal{N}_{singular}$ ($p = 3$, $N_e = 54$).

the i^{th} element of vector $\boldsymbol{x}_3$ is k_i ,

and the elements of vector $\boldsymbol{b}_3$ are $\int_{\Omega_i} f \, dA$ for $i \in \mathcal{N}_{singular}$.

Owing to the support of the basis functions $\{\phi_j\}_{j=1}^{N_{node}}$, matrix $[A_{31}A_{32}]$ has sparsity structure similar to matrix $[A_{11}A_{12}]$. Matrix A_{23} is usually a dense $N_{bound} \times N_s$ matrix, but the size of this matrix is much smaller than the size of $\hat{A}$. Since $N_s = 2(p+1)-1$, comparing with system (3.34), the size of system (3.47) has increased only by $2(p+1)-1$, and the number of non-zeros has increased only by $O(N_{bound}) = O(1/h)$. The total number of non-zeros in $\hat{A}$ remains $O(1/h^2)$. We shall refer to this approximation technique for the solution of the BVP by p^{th} *order Augmented FVEM.*

3.3 Numerical Experiments

3.3.1 Model Problem 1 : Regular Solution

We first consider Model Problem 1 described in Section 3.1.2. The solution to this BVP is a 9^{th} order polynomial. The H^1 and L_2 errors for our numerical experiments on a regular grid as in Figure 3.4 are

plotted in Figures 3.6 and 3.7, respectively. As can be seen, the error converges immediately to near machine accuracy for the 9^{th} and 10^{th} order FVEM (i.e., $p = 9, 10$). For other order FVEM, the error converges as follows:

$$||u - u^h||_{H^1} = O(h^p)\,, \tag{3.48}$$

$$||u - u^h||_{L_2} = \begin{cases} O(h^{p+1}) & \text{for } p = 1, 3, 5, 7, 8\ , \\ O(h^p) & \text{for } p = 2, 4, 6\ , \end{cases} \tag{3.49}$$

where u is the exact solution of BVP (3.1)-(3.2), u^h is the approximation obtained by the FVEM and h is the diameter of the element triangles. This convergence behaviour is similar to what is shown theoretically by Plexousakis and Zouraris [35] for the FVEM applied to one-dimensional problems. It is noteworthy that similar convergence behaviour for the L_2 norm with sub-optimal convergence rates for even polynomial order can also be found for the Discontinuous Galerkin Method [34]. To investigate whether this L_2 convergence behaviour may be the result of error cancellation on our regular mesh, we repeated the calculations on slightly perturbed grids with the vertices of element triangles in the interior of the domain randomly relocated within distances of h/10 from their original location. Figure 3.8 indicates that the odd-even dichotomy for the L_2 convergence is not due to error cancellation effects on regular grids. It is interesting that for Discontinuous Galerkin methods, the odd-even dichotomy only seems to occur on regular grids [34, 27, 19], while we and others also observe it on non-regular grids for the FVEM [49].

3.3.2 Model Problem 2 : Singular Solution

Second, we consider Model Problem 2 described in Section 3.1.3. The solution to this BVP has a boundary derivative blow-up singularity (i.e., the r-directional derivative blows up at the origin). We consider the FVEM with the standard trial function space (Section 3.2.3) first, and then the augmented trial function space (Section 3.2.4).

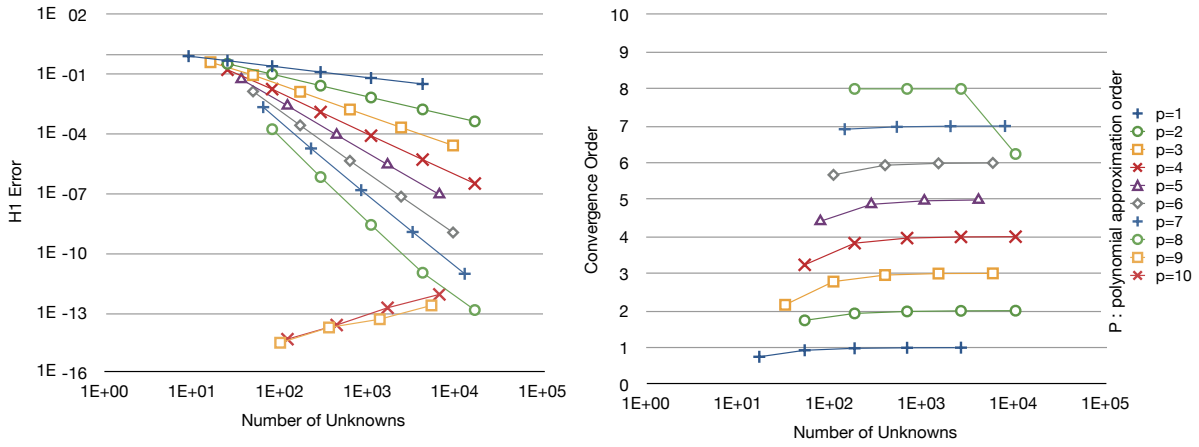


Figure 3.6: H^1 error convergence for the Poisson problem with 9^{th} order polynomial exact solution (Model Problem 1).

3.3.2.1 FVEM with Standard Trial Function Space

As can be seen in Figure 3.9 using the high order FVEM with standard trial function space, the rate of convergence does not increase as the order of the method increases. The 1^{st} and 2^{nd} order FVEM appear to converge initially with higher order than the 3^{rd} to 10^{th} order methods. We suspect that this is due to the fact that their errors are initially dominated by the regular part of the solution.

3.3.2.2 FVEM with Augmented Trial Function Space

As can be seen in Figure 3.10 using the augmented trial function space, we recover a rate of convergence similar to the smooth case (the BVP with 9^{th} order polynomial solution, Figure 3.6). That is to say, the error behaves like in Equation (3.48) and high order convergence rates are restored.

3.4 Concluding Remarks on Chapter 3

We have described how the idea of augmentation of the trial function space can be applied to the FVEM, and have presented numerical experiments that indicate that high order convergence rates can be recovered

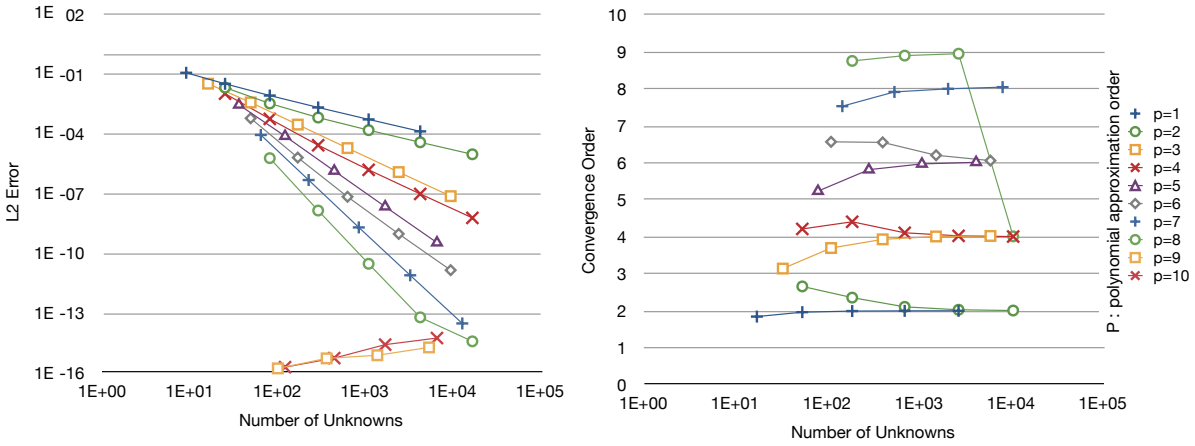


Figure 3.7: L_2 error convergence for the Poisson problem with 9^{th} order polynomial exact solution (Model Problem 1).

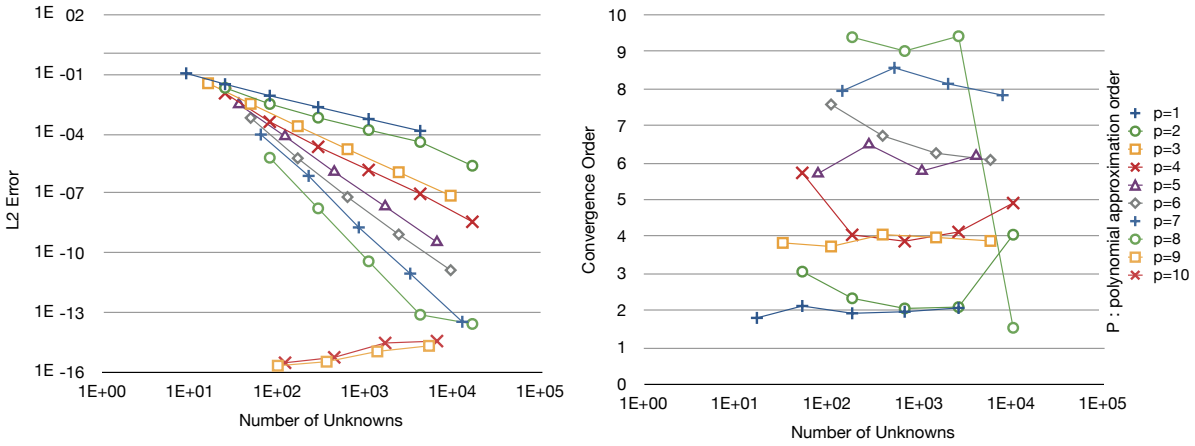


Figure 3.8: L_2 error convergence for the Poisson problem with 9^{th} order polynomial exact solution on a randomly perturbed grid (Model Problem 1).

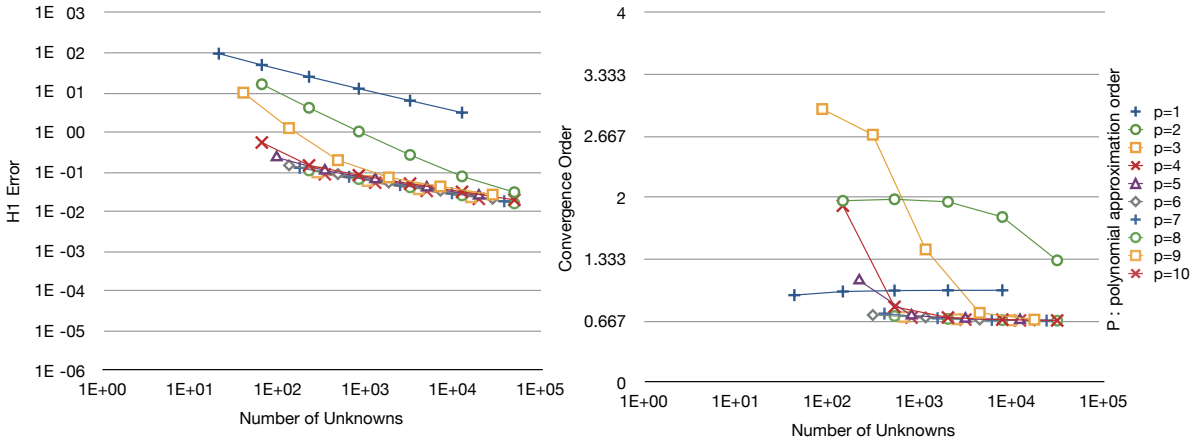


Figure 3.9: H^1 error convergence for derivative blow-up singular solution (Model Problem 2).

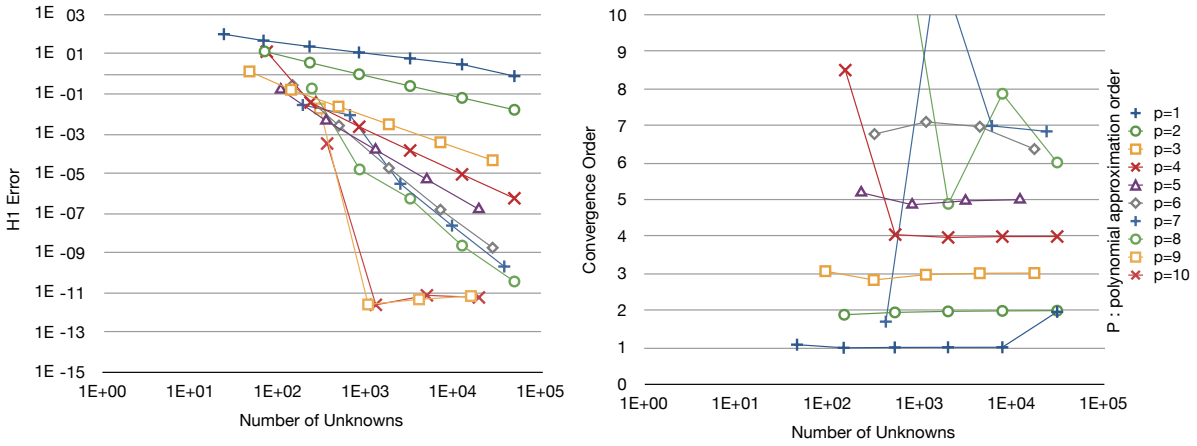


Figure 3.10: H^1 error convergence for derivative blow-up singular solution with augmented trial function space (Model Problem 2).

by augmenting the trial function space. Due to the nature of the discrete integral formulation of the FVEM, an augmented FVEM can be constructed in a much simpler way than the augmented FEM that was presented in Strang and Fix [46]. In particular, the singular basis functions only appear in the boundary conditions, which leads to the fact that there is no need for singular integration, and that the sparsity of the matrix is automatically maintained. Note that techniques exist for high-order accurate integration of singular functions, but our method avoids the need for using these techniques. Note that it may also be possible to augment the FEM similarly to what we have done for the FVEM, but we have not explored this.

Chapter 4

Numerical Study of Unbounded Capillary Surfaces

With the work presented in Chapter 2 and other works on the unbounded capillary surfaces ([9, 8, 30, 24, 33, 39, 40, 1]), almost all of the leading order asymptotic behaviours of the unbounded capillary surfaces are known, and it seems as if we have almost complete understanding of the unbounded capillary surfaces. However, the asymptotic series approximation always comes with a fine print warning "the approximation is only valid in a sufficiently small neighbourhood of the singularity", hence it is only a local approximation. We wish to obtain a global approximation of the solution through numerical methods; however, it is also known that the singularity of the solution spoils the accuracy of a standard finite element approximation and the approximation cannot reproduce the singularity accurately. In this chapter we propose a numerical methodology that constructs globally valid approximations of unbounded capillary surfaces, which exhibit the proper asymptotic behaviour at the singular point, while also being valid away from the singularity. This numerical methodology contains two simple but important ingredients, a change of variable and a change of coordinates, which are inspired by known asymptotic approximations for unbounded capillary surfaces. These ingredients are combined with the finite volume element or Galerkin finite element methods.

We verify the accuracy of this numerical methodology by comparing the numerical solution with asymptotic series approximations, and by conducting numerical convergence studies. We first show that the numerical solutions we obtain have the proper singular behaviour by comparing numerical solutions of the Laplace-Young equation with known asymptotic series approximations. Then we conduct numerical convergence studies to show that the numerical approximation is a globally valid approximation. In order to conduct numerical convergence studies, we need model problems with known closed-form solutions. However, there is no known unbounded closed-form solution of the Laplace-Young equation. On the other hand, a few closed-form solutions are known for the steep slope approximation of the Laplace-Young equation [24, 1] (we shall refer to this PDE as the Asymptotic Laplace-Young equation), and it is known that these solutions have the same asymptotic behaviour as the solution of the original problem, so we conduct the convergence study using the Asymptotic Laplace-Young equation.

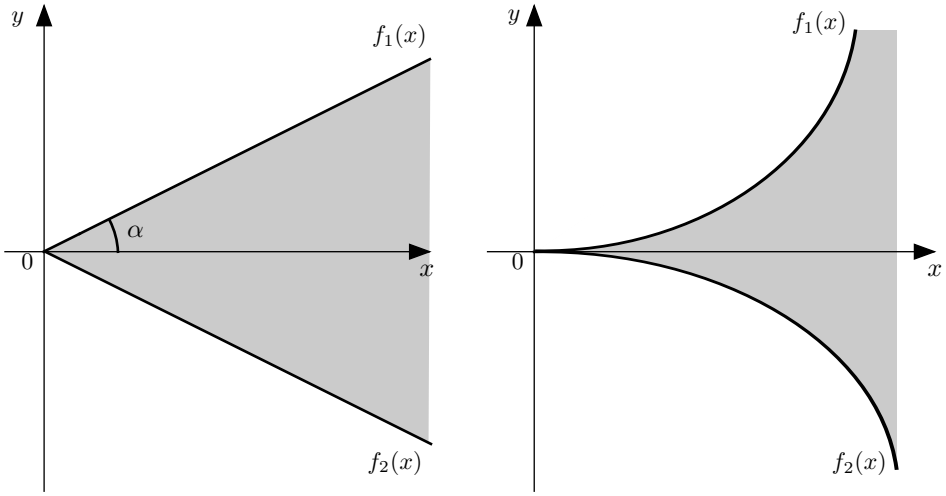
Using this accurate computational methodology, two open problems on the asymptotic behaviour of capillary surfaces in domains with a cusp are studied numerically, leading to two new conjectures that may guide future analytical work on these open problems.

4.1 The Boundary Value Problem

In this section we first formulate the Laplace-Young boundary value problem and summarize the asymptotic behaviour of its solutions in domains with a corner or a cusp and the function spaces these solutions belong to. We state some open problems on asymptotic behaviour for a domain with an osculatory cusp and a cusp with infinite curvature and define model problems that will be used in numerical tests. We then describe the asymptotic Laplace-Young equation and its known closed-form solutions on domains with a corner or a cusp, which are used to formulate additional numerical model problems.

4.1.1 Laplace-Young Boundary Value Problem

Let Ω be an unbounded open domain as in Figure 4.1 with boundaries $\partial\Omega_1$ and $\partial\Omega_2$ described by functions $f_1(x)$ and $f_2(x)$, and let $u \in C^2(\Omega)$ be the height of the capillary surface that satisfies the following



(a) a corner domain (b) a cusp domain

Figure 4.1: Unbounded domains with a corner and a cusp.

boundary value problem (the Laplace-Young boundary value problem) on this domain:

$$\nabla \cdot T(u) = \kappa u \quad \text{in } \Omega\,, \tag{4.1}$$

$$\vec{\nu}_1 \cdot T(u) = \cos\gamma_1 \quad \text{on } \partial\Omega_1\,, \tag{4.2}$$

$$\vec{\nu}_2 \cdot T(u) = \cos\gamma_2 \quad \text{on } \partial\Omega_2\,, \tag{4.3}$$

with

$$\kappa \;:\; \text{capillarity constant}\,, \tag{4.4}$$

$$\vec{\nu}_1,\, \vec{\nu}_2 \;:\; \text{exterior unit normal vectors on the boundaries } \partial\Omega_1 \text{ and } \partial\Omega_2, \tag{4.5}$$

$$\gamma_1,\, \gamma_2 \;:\; \text{contact angles (as indicated in Figure 1.1)}, \tag{4.6}$$

$$T(u) = \frac{\nabla u}{\sqrt{1+|\nabla u|^2}}. \tag{4.7}$$

Note that the capillarity constant κ can be normalized by rescaling x, y and u. In the following sections

we let $\kappa = 1$. The open domain Ω and boundaries $\partial\Omega_1$ and $\partial\Omega_2$ are defined more specifically as follows:

$$\Omega = \{(x,y) \in \mathbb{R}^2 : 0 < x, f_2(x) < y < f_1(x)\}, \tag{4.8}$$

$$\partial\Omega_1 = \{(x,y) \in \mathbb{R}^2 : 0 < x, y = f_1(x)\}, \tag{4.9}$$

$$\partial\Omega_2 = \{(x,y) \in \mathbb{R}^2 : 0 < x, y = f_2(x)\}, \tag{4.10}$$

with

$$f_1(x), f_2(x) \in C^3(0,\infty), \tag{4.11}$$

$$f_1(x) > f_2(x) \qquad \text{for } x > 0, \tag{4.12}$$

$$\lim_{x\to 0^+} f_1(x) = \lim_{x\to 0^+} f_2(x) = 0, \tag{4.13}$$

$$\lim_{x\to 0^+} f_1'(x) \neq \infty \neq \lim_{x\to 0^+} f_2'(x). \tag{4.14}$$

For simplicity of discussion we focus on the two specific types of domains depicted in Figure 4.1: a corner domain and a cusp domain, which are defined as

$$\text{corner domain} : f_1(x) = x\tan\alpha \quad \text{and} \quad f_2(x) = -x\tan\alpha, \qquad \text{where } 0 < \alpha < \pi/2, \tag{4.15}$$

$$\text{cusp domain} : \lim_{x\to 0^+} f_1'(x) = 0 \quad \text{and} \quad \lim_{x\to 0^+} f_2'(x) = 0. \tag{4.16}$$

4.1.1.1 Asymptotic Behaviour

It is known that the solution of the Laplace-Young boundary value problem in a corner domain is unbounded at $(0,0)$ if $\gamma_1 + \gamma_2 + 2\alpha < \pi$ (see [13]). Also, it can be shown that if $\gamma_1 + \gamma_2 \neq \pi$ the solution of the boundary value problem in a cusp domain is unbounded at $(0,0)$ (see Lemma 2.2.1). In addition, the following asymptotic behaviours are known:

Corner domain with $\gamma_1 + \gamma_2 + 2\alpha < \pi$ (see [30] for a proof)

If $\gamma_1 = \gamma_2 = \gamma$ and $\gamma + \alpha < \pi/2$ then the solution of the boundary value problem in the corner domain

has the following asymptotic behaviour:

$$u(r,\theta) \quad = \quad \frac{\cos\theta - \sqrt{k^2 - \sin^2\theta}}{kr} + O(r^3) \qquad \text{as } r \to 0\,, \tag{4.17}$$

where

$$(r,\,\theta) \quad : \quad \text{polar coordinate variables}\,, \tag{4.18}$$

$$k \quad = \quad \frac{\sin\alpha}{\cos\gamma}\,. \tag{4.19}$$

More formally, (4.17) can be written as:

$$\text{there exist constants } r_o \text{ and } M \text{ such that } \left| u - \frac{\cos\theta - \sqrt{k^2 - \sin^2\theta}}{kr} \right| < Mr^3 \quad \text{for } 0 < r < r_o\,. \tag{4.20}$$

The above inequality gives the following bounds for the solution u:

$$\frac{\cos\theta - \sqrt{k^2 - \sin^2\theta}}{kr} - Mr^3 < u < \frac{\cos\theta - \sqrt{k^2 - \sin^2\theta}}{kr} + Mr^3 \qquad \text{for } 0 < r < r_o\,. \tag{4.21}$$

The proof for the asymptotic relation (4.17) only provides the existence of these two constants and does not give any estimate of the size of these constants. Thus, even though (4.17) shows that the asymptotic approximation becomes more and more accurate as we get closer to the singularity, it does not give any quantitative description of the approximation error.

Also, it is easy to show from (4.21) that there exist positive constants M^+, M^-, and x_o such that

$$\frac{M^-}{f_1(x) - f_2(x)} < u < \frac{M^+}{f_1(x) - f_2(x)} \qquad \text{for } 0 < x < x_o. \tag{4.22}$$

Cusp domain with $\gamma_1 + \gamma_2 \neq \pi$ (see Theorem 2.2.2 for a proof)

An unbounded solution of the boundary value problem in a cusp domain has the following asymptotic

behaviour:

$$u(x,y) = \frac{\cos\gamma_1 + \cos\gamma_2}{f_1(x) - f_2(x)} + O\left(\frac{f_1'(x) - f_2'(x)}{f_1(x) - f_2(x)}\right) \qquad \text{as } x \to 0^+, \tag{4.23}$$

if $\gamma_1 + \gamma_2 \neq \pi$ and the boundary functions $f_1(x)$ and $f_2(x)$ satisfy the asymptotic relations

$$f_1(x) - f_2(x) = o\left(f_1'(x) - f_2'(x)\right), \tag{4.24}$$

$$\frac{f_1''(x) - f_2''(x)}{f_1(x) - f_2(x)} = \alpha\frac{(f_1'(x) - f_2'(x))^2}{(f_1(x) - f_2(x))^2} + o\left(\frac{(f_1'(x) - f_2'(x))^2}{(f_1(x) - f_2(x))^2}\right), \tag{4.25}$$

$$\frac{f_1'''(x) - f_2'''(x)}{f_1'(x) - f_2'(x)} = O\left(\frac{(f_1'(x) - f_2'(x))^2}{(f_1(x) - f_2(x))^2}\right), \tag{4.26}$$

$$f_1'(x) + f_2'(x) = \delta(f_1'(x) - f_2'(x)) + o(f_1'(x) - f_2'(x)), \tag{4.27}$$

$$f_1''(x) + f_2''(x) = O(f_1''(x) - f_2''(x)), \tag{4.28}$$

as $x \to 0$, where α, $\delta \in \mathbb{R}$.

Note that most boundary functions forming cusp domains satisfy the asymptotic conditions (4.24)-(4.28). One known exception is when the boundary functions forming a cusp are osculatory at the cusp. Curves are said to be osculatory if they intersect and share the tangent line and the osculating circle at the intersection point. Again it follows from (4.23) that there exist positive constants M^+, M^-, and x_o such that

$$\frac{M^-}{f_1(x) - f_2(x)} < u < \frac{M^+}{f_1(x) - f_2(x)} \qquad \text{for } 0 < x < x_o. \tag{4.29}$$

4.1.1.2 Open Problems

To the author's knowledge there are two major open problems in the solution behaviour of the Laplace-Young equation in a domain with a cusp. We here summarize these open problems.

Open Problem 1: osculatory cusp with non-supplementary contact angles ($\gamma_1 + \gamma_2 \neq \pi$)

An osculatory cusp is a cusp formed by two osculating curves, and the asymptotic expansion from the

previous section does not apply. For example, the following two boundary functions $f_1(x)$ and $f_2(x)$ form an osculatory cusp at the origin:

$$f_1(x) = x^2 + x^3\,, \tag{4.30}$$

$$f_2(x) = x^2 - x^3\,. \tag{4.31}$$

We note that the asymptotic orders of the sum and the difference of these boundary functions $f_1(x)$ and $f_2(x)$ are different, (i.e., $f_1(x) - f_2(x) = O(x^3)$ while $f_1(x) + f_2(x) = O(x^2)$ as $x \to 0$). Hence these choices of $f_1(x)$ and $f_2(x)$ do not satisfy the asymptotic relations (4.27)-(4.28) so that the leading order asymptotic behaviour of the solution at this cusp is unknown. The main reason why the proof for the leading order asymptotic behaviour could not be constructed for the osculatory cusp case is because the second order term of the formal asymptotic series could not be found (see Chapter 2 for details).

Open Problem 2: Infinite curvature cusp with supplementary contact angles ($\gamma_1 + \gamma_2 = \pi$)

As was noted before, the solution of the Laplace-Young equation in a cusp domain is unbounded if $\gamma_1 + \gamma_2 \neq \pi$, but it is not necessarily true that the solution is bounded if $\gamma_1 + \gamma_2 = \pi$. In Chapter 2 we have shown that the solution is bounded if $\gamma_1 + \gamma_2 = \pi$ and the curvatures of the boundary functions are finite, (i.e., $\lim_{x\to 0} f_1''(x) \neq \infty$ and $\lim_{x\to 0} f_2''(x) \neq \infty$). However, the nature of the solution for the case where the curvatures of one or both boundary functions are infinite is not known (e.g., $f_1(x) = x^{3/2}$ and $f_2(x) = -2x^{3/2}$).

4.1.1.3 Model Problems 1 and 2

For the numerical experiments to be reported on below we consider the following Model Problems.

Consider bounded open domains Ω as depicted in Figure 4.2. Let $u \in C^2(\Omega)$ be the height of the capillary

surface that satisfies the following boundary value problem:

$$\nabla \cdot T(u) = u \quad \text{in } \Omega\,, \tag{4.32}$$

$$\vec{\nu}_1 \cdot T(u) = \cos\gamma_1 \quad \text{on } \partial\Omega_1\,, \tag{4.33}$$

$$\vec{\nu}_2 \cdot T(u) = \cos\gamma_2 \quad \text{on } \partial\Omega_2\,, \tag{4.34}$$

$$\vec{\nu}_3 \cdot T(u) = 0 \quad \text{on } \partial\Omega_3\,, \tag{4.35}$$

with

$$\vec{\nu}_1,\, \vec{\nu}_2,\, \vec{\nu}_3 \;:\; \text{exterior unit normal vectors on the boundaries } \partial\Omega_1,\, \partial\Omega_2 \text{ and } \partial\Omega_3, \tag{4.36}$$

$$\gamma_1,\, \gamma_2 \;:\; \text{contact angles.} \tag{4.37}$$

The bounded open domain Ω and boundaries $\partial\Omega_{1,2,3}$ are defined more specifically as follows:

$$\Omega = \{(x,y) \in \mathbb{R}^2 : 0 < x < 1, f_2(x) < y < f_1(x)\}\,, \tag{4.38}$$

$$\partial\Omega_1 = \{(x,y) \in \mathbb{R}^2 : 0 < x < 1, y = f_1(x)\}\,, \tag{4.39}$$

$$\partial\Omega_2 = \{(x,y) \in \mathbb{R}^2 : 0 < x < 1, y = f_2(x)\}\,, \tag{4.40}$$

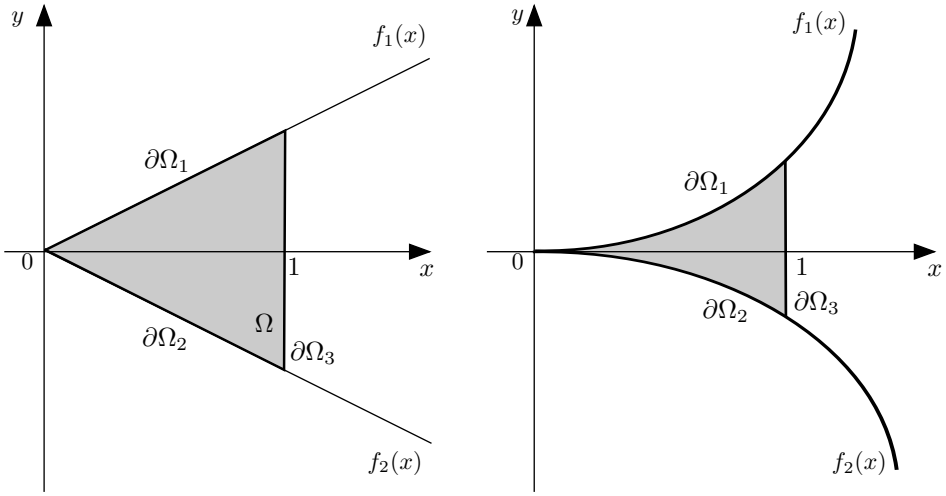
$$\partial\Omega_3 = \{(x,y) \in \mathbb{R}^2 : x = 1, f_2(1) < y < f_1(1)\}\,. \tag{4.41}$$

The boundary functions and the contact angles are chosen for each model problem as tabulated in Tables 4.1 and 4.2.

Table 4.1: Model Problem 1: Laplace-Young equation in a domain with a corner. All three model problems have $\alpha = \pi/7$ and $\gamma_1 + \gamma_2 + 2\alpha < \pi$, resulting in solutions that are unbounded at (0,0).

Problem	$f_1(x)$	$f_2(x)$	γ_1	γ_2	
1-1	$x\tan(\pi/7)$	$-x\tan(\pi/7)$	$\pi/6$	$\pi/6$	Corner (unbounded)
1-2	$x\tan(\pi/7)$	$-x\tan(\pi/7)$	$\pi/4$	$\pi/4$	Corner (unbounded)
1-3	$x\tan(\pi/7)$	$-x\tan(\pi/7)$	$\pi/3$	$\pi/3$	Corner (unbounded)

These model problems are chosen carefully so that the singularity may only occur at the corner or cusp



(a) Model Problems 1 and 3

(b) Model Problems 2 and 4

Figure 4.2: Computational domains for model problems with a corner and a cusp at (0,0).

at the origin, although there are three nonsmooth points on the boundary of the domain Ω. Following immediately from the regularity result of Simon [43], this implies that the solutions u of these model problems are differentiable up to the boundary except at the origin, i.e., $u \in C^1(\bar{\Omega}\backslash\{(0,0)\})$. Also, the asymptotic behaviour of the solution at the origin is known to be as stated in (4.17) for Model Problem 1 ($\gamma_1 + \gamma_2 + 2\alpha < \pi$, unbounded) and as in (4.23) for Model Problems 2a ($\gamma_1 + \gamma_2 \neq \pi$, unbounded) and 2c-1 and 2c-3 ($\gamma_1 + \gamma_2 \neq \pi$, unbounded). The asymptotic behaviour for Model Problems 2b (osculatory cusp with non-supplementary contact angles) and 2c-2 (infinite curvature cusp with supplementary contact angles) are open problems.

4.1.1.4 Solution Function Spaces

It is interesting to discuss the function spaces where the solutions of Model Problem 1 and Model Problems 2a, 2c-1 and 2c-3 reside.

Proposition 4.1.1. *For any fixed p with $1 \leq p < \infty$, the solutions of Model Problem 1, Model Problems 2a and Model Problems 2c-1 and 2c-3 are in the $L_p(\Omega)$ function space if and only if the following integral is*

Table 4.2: Model Problem 2: Laplace-Young equation in a domain with a cusp. Model Problems 2a and 2c-1 and 2c-3 have unbounded solutions at (0,0), and the asymptotic behaviour for Model Problems 2b and 2c-2 at (0,0) are open problems.

Problem	$f_1(x)$	$f_2(x)$	γ_1	γ_2	
2a-1	$x^2/6$	$-x^3/8$	$\pi/6$	$\pi/3$	Regular cusp (unbounded)
2a-2	$x^3/6$	$-x^3/8$	$\pi/3$	$\pi/4$	Regular cusp (unbounded)
2a-3	$x^5/6$	$-x^4/8$	$\pi/3$	$\pi/4$	Regular cusp (unbounded)
2b-1	$(x^2+x^3)/6$	$(x^2-3/4x^3)/6$	$\pi/3$	$\pi/4$	Osculatory cusp (open problem)
2b-2	$(3x^2+x^3)/6$	$(3x^2-3/4x^3)/6$	$\pi/3$	$\pi/4$	Osculatory cusp (open problem)
2b-3	$(x^{3/2}+x^3)/6$	$(x^{3/2}-3/4x^3)/6$	$\pi/3$	$\pi/4$	Osculatory cusp (open problem)
2c-1	$(x^{3/2})/6$	$(x^{3/2})/8$	$5\pi/6-\pi/180$	$\pi/6$	∞ curvature cusp (unbounded)
2c-2	$(x^{3/2})/6$	$(x^{3/2})/8$	$5\pi/6$	$\pi/6$	∞ curvature cusp (open problem)
2c-3	$(x^{3/2})/6$	$(x^{3/2})/8$	$5\pi/6+\pi/180$	$\pi/6$	∞ curvature cusp (unbounded)

finite for any ϵ in the interval $(0,1]$:

$$\int_0^{\epsilon} \frac{1}{(f_1(x)-f_2(x))^{p-1}}\,dx\,. \tag{4.42}$$

Proof. We first note that, for the case of Model Problem 1 and Model Problem 2a, the comparison principle (see [13]) gives that $u>0$, and also recall that there exist positive constants M^+, M^-, and x_o such that

$$\frac{M^-}{f_1(x)-f_2(x)} < u < \frac{M^+}{f_1(x)-f_2(x)} \qquad \text{for } 0<x<x_o. \tag{4.43}$$

We now bound the integral $\int_\Omega |u|^p \mathrm{d}A$ from above:

$$\begin{aligned}
\int_\Omega |u|^p \,\mathrm{d}A &= \int_\Omega u^p \,\mathrm{d}A \qquad (\text{since } u > 0), && (4.44)\\
&= \int_{x=0}^{1} \int_{f_2(x)}^{f_1(x)} u^p \mathrm{d}y\mathrm{d}x\,, && (4.45)\\
&= \int_{x=0}^{x_o} \int_{f_2(x)}^{f_1(x)} u^p \,\mathrm{d}y\mathrm{d}x + \int_{x=x_o}^{1} \int_{f_2(x)}^{f_1(x)} u^p \,\mathrm{d}y\mathrm{d}x\,, && (4.46)\\
&\leq \int_{x=0}^{x_o} \int_{f_2(x)}^{f_1(x)} u^p \,\mathrm{d}y\mathrm{d}x + \int_{x=x_o}^{1} \int_{f_2(x)}^{f_1(x)} \max_{x_o<x<1}(u)^p \,\mathrm{d}y\mathrm{d}x\,, && (4.47)\\
&< \int_{x=0}^{x_o} \int_{f_2(x)}^{f_1(x)} \frac{(M^+)^p}{(f_1(x)-f_2(x))^p} \,\mathrm{d}y\mathrm{d}x + \max_{x_o<x<1}(u)^p \int_{x=x_o}^{1} \int_{f_2(x)}^{f_1(x)} 1 \,\mathrm{d}y\mathrm{d}x\,, && (4.48)\\
&= (M^+)^p \int_0^{x_o} \frac{1}{(f_1(x)-f_2(x))^{p-1}} \,\mathrm{d}x + \max_{x_o<x<1}(u)^p \int_{x=x_o}^{1} (f_1(x)-f_2(x)) \,\mathrm{d}x\,, && (4.49)
\end{aligned}$$

If p is chosen so that integral (4.42) is finite for any $\epsilon \in (0,1]$, then the first term of (4.49) is finite. Also, noting that u is bounded away from the origin ($u \in C^1(\bar{\Omega}\backslash\{0\})$) and that the domains Ω for the Model Problems are bounded domains, the second term of (4.49) is also finite. Thus if p is chosen so that integral (4.42) is finite then the solution of Model Problems 1 and 2a are in the $L_p(\Omega)$ function space.

We now bound the integral $\int_\Omega |u|^p \mathrm{d}A$ from below:

$$\begin{aligned}
\int_\Omega |u|^p \,\mathrm{d}A &= \int_{x=0}^{x_o} \int_{f_2(x)}^{f_1(x)} u^p \,\mathrm{d}y\mathrm{d}x + \int_{x=x_o}^{1} \int_{f_2(x)}^{f_1(x)} u^p \,\mathrm{d}y\mathrm{d}x\,, && (4.50)\\
&> \int_{x=0}^{x_o} \int_{f_2(x)}^{f_1(x)} u^p \,\mathrm{d}y\mathrm{d}x && (4.51)\\
&> \int_{x=0}^{x_o} \int_{f_2(x)}^{f_1(x)} \frac{(M^-)^p}{(f_1(x)-f_2(x))^p} \,\mathrm{d}y\mathrm{d}x && (4.52)\\
&= (M^-)^p \int_{x=0}^{x_o} \frac{1}{(f_1(x)-f_2(x))^{p-1}} \,\mathrm{d}x && (4.53)
\end{aligned}$$

This gives that if p is chosen so that integral (4.42) is not finite, then the solutions of Model Problems 1 and 2a are not in the L_p function space.

The proof for Model Problems 2c-1 and 2c-3 is slightly more complicated because $u > 0$ does not hold.

A sketch of the proof for these cases is as follows. Since $u \in C^1(\bar{\Omega}\backslash\{0\})$, there is a neighbourhood Ω_s of the singularity where the solution is either positive or negative. Using the approach above, it can be shown that $u \in L_p(\Omega_s)$ if and only if integral (4.42) is finite, which is equivalent to $u \in L_p(\Omega)$ since u is bounded away from the singularity. □

Corollary 4.1.1.

(A) The solution of Model Problem 1 is in the $L_{2-\delta}$ function space for any $\delta > 0$.

(B) The solution is in the $L_{1+1/2-\delta}$ function space for of Model Problem 2a-1, is in the $L_{1+1/3-\delta}$ function space for Model Problem 2a-2 , and is in the $L_{1+1/4-\delta}$ function space for Model Problem 2a-3, for any $\delta > 0$.

(C) The solution is in the $L_{1+2/3-\delta}$ function space for of Model Problems 2c-1 and 2c-3, for any $\delta > 0$.

Note finally that all solutions of the Laplace-Young equation in a bounded domain Ω are in L_1, which is consistent with the physical interpretation that the volume of the fluid under the capillary surface is finite.

4.1.2 Asymptotic Laplace-Young Equation

There are no closed-form solutions for the Laplace-Young equation in domains with a corner or a cusp, but closed-form solutions exist for the following simplification of the Laplace-Young PDE. These closed-form solutions will be used in Section 4.3.2 for convergence studies of the numerical methodology we propose in Section 4.2.

Assuming the slope of the solution of the Laplace-Young boundary value problem ((4.1)-(4.3)) is steep, i.e., $|\nabla u| >> 1$, we can approximate the PDE and the boundary condition, by ignoring the 1 in the denominator of the differential operator $T(\cdot)$, and obtain the following boundary value problem:

$$\nabla \cdot \tilde{T}(u) = u \qquad \text{in } \Omega\,, \tag{4.54}$$

$$\vec{\nu}_1 \cdot \tilde{T}(u) = \cos\gamma \qquad \text{on } \partial\Omega_1\,, \tag{4.55}$$

$$\vec{\nu}_2 \cdot \tilde{T}(u) = \cos\gamma \qquad \text{on } \partial\Omega_2\,, \tag{4.56}$$

where

$$\tilde{T}(u) \quad = \quad \frac{\nabla u}{|\nabla u|}\,. \tag{4.57}$$

This approximation is called the "steep slope approximation" [24] of the Laplace-Young boundary value problem and unbounded closed-form solutions of this boundary value problem are known for two types of domains: the unbounded corner domain of Figure 4.1(a) [24] and the circular cusp domains of Figure 4.3 [1]. Also, it has been shown that the exact solutions of this boundary value problem are good asymptotic approximations of the solutions of the original Laplace-Young equation on the same domains [30, 1]. We shall refer to this boundary value problem as the *Asymptotic Laplace-Young boundary value problem*. Note that this boundary value problem is a rare case of a nonlinear PDE with nonlinear boundary conditions for which one can find closed-form solutions in some nontrivial domains.

4.1.2.1 Closed-form Solutions

Corner domain (Figure 4.1(a), $\gamma + \alpha < \pi/2$)

Let $u \in C^2(\Omega)$ be a solution of the boundary value problem (4.54)-(4.56) on the unbounded corner domain defined as in (4.8)-(4.10) with the boundary functions

$$f_1(x) = x \tan\alpha\,, \tag{4.58}$$

$$f_2(x) = -x \tan\alpha\,. \tag{4.59}$$

If $\gamma + \alpha < \pi/2$, u is given as the following closed-form expression in terms of the polar coordinate variables r and θ:

$$u(r,\theta) = \frac{\cos\theta - \sqrt{k^2 - \sin^2\theta}}{kr}\,, \tag{4.60}$$

where $k = \frac{\sin\alpha}{\cos\gamma}$. This closed-form solution of the asymptotic Laplace-Young equation first explicitly appears in [24].

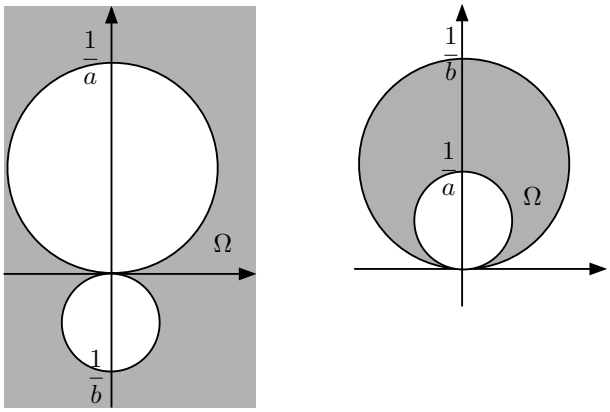


Figure 4.3: Circular cusp domains.

Circular cusp domain (Figure 4.3, $\gamma \neq \pi/2$)

Let $u \in C^2(\Omega)$ be a solution of the boundary value problem (4.54)-(4.56) with $\gamma \neq \pi/2$ and with the domain defined as

$$\Omega \quad := \quad \begin{cases} \left\{(x,y) \in \mathbb{R}^2 \backslash \left(\bar{B}_{\frac{1}{2a}}(0,\frac{1}{2a}) \cup \bar{B}_{-\frac{1}{2b}}(0,\frac{1}{2b})\right)\right\} & \text{for } b<0, \\ \left\{(x,y) \in \left(B_{\frac{1}{2b}}(0,\frac{1}{2b}) \backslash \bar{B}_{\frac{1}{2a}}(0,\frac{1}{2a})\right)\right\} & \text{for } b>0, \end{cases} \tag{4.61}$$

where $B_r(x_o, y_o)$ is the open disc of radius r centred at (x_o, y_o), i.e.,

$$B_r(x_o,y_o) \quad = \quad \{(x,y) \in \mathbb{R}^2 : (x-x_o)^2 + (y-y_o)^2 < r^2\}. \tag{4.62}$$

A closed-form expression for u is given by

$$u(p,q) \quad = \quad Ap^2 - 2\sqrt{1 - A^2(q-q_0)^2}p - A(q-q_0)^2 + Aq_0^2 \tag{4.63}$$

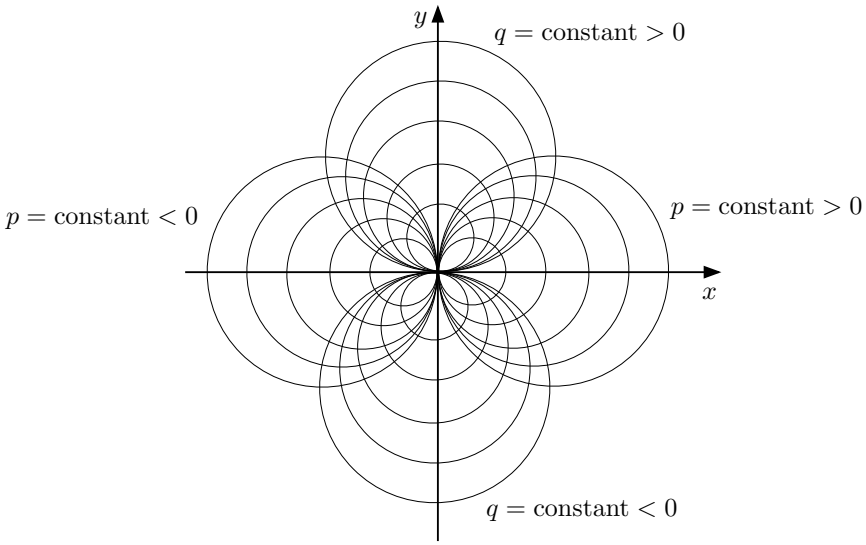


Figure 4.4: Tangent cylindrical coordinate system.

where

$$A = \frac{2\cos\gamma}{a-b}, \tag{4.64}$$

$$q_0 = \frac{a+b}{2}, \tag{4.65}$$

and p, q are the coordinate variables of the tangent cylindrical coordinate system introduced in [31], which is depicted in Figure 4.4 and is defined as

$$p = \frac{x}{x^2+y^2}, \tag{4.66}$$

$$q = \frac{y}{x^2+y^2}. \tag{4.67}$$

This closed-form solution of the Asymptotic Laplace-Young equation first appears in [1]. Note that $\lim_{(x,y)\to(0,0)} p = \infty$ and the solution (4.63) behaves like $1/x^2$ as $x \to 0$, hence it exhibits a more severe singularity than the singularity of the Asymptotic Laplace-Young PDE in a corner domain, which features a $1/r$ singularity.

Table 4.3: Model Problems 3 and 4: Asymptotic Laplace-Young equation in domains with a corner and a cusp. Model Problem 3 has $\alpha = \pi/7$ and $\gamma + \alpha < \pi/2$, resulting in a solution that is unbounded at (0,0). Model Problem 4 has $\gamma \neq \pi/2$, and its solution is also unbounded at (0,0).

Name	$f_1(x)$	$f_2(x)$	γ_1	γ_2	
Model Problem 3	$x\tan(\pi/7)$	$-x\tan(\pi/7)$	$\pi/6$	$\pi/6$	Corner (unbounded)
Model Problem 4	$-\sqrt{5^2 - x^2} + 5$	$\sqrt{10^2 - x^2} - 10$	$\pi/6$	$\pi/6$	Circular cusp (unbounded)

4.1.2.2 Model Problems 3 and 4

For the numerical experiments on the Asymptotic Laplace-Young equation we consider the following Model Problems on the corner and cusp domains of Figure 4.2.

Let $u \in C^2(\Omega)$ be the function that satisfies the boundary value problem with mixed boundary conditions

$$\nabla \cdot \tilde{T}(u) = u \quad \text{in } \Omega\,, \tag{4.68}$$

$$\vec{\nu}_1 \cdot \tilde{T}(u) = \cos\gamma_1 \quad \text{on } \partial\Omega_1\,, \tag{4.69}$$

$$\vec{\nu}_2 \cdot \tilde{T}(u) = \cos\gamma_2 \quad \text{on } \partial\Omega_2\,, \tag{4.70}$$

$$u = u_{\text{exact}} \quad \text{on } \partial\Omega_3\,, \tag{4.71}$$

with

$$\vec{\nu}_1,\, \vec{\nu}_2,\, \vec{\nu}_3 \;:\; \text{exterior unit normal vectors on the boundaries } \partial\Omega_1, \partial\Omega_2, \text{ and } \partial\Omega_3 \tag{4.72}$$

$$\gamma_1,\, \gamma_2 \;:\; \text{contact angles,} \tag{4.73}$$

$$u_{\text{exact}} \;:\; \text{the closed-form solutions given in (4.60) or (4.63).} \tag{4.74}$$

The bounded open domain Ω and boundaries $\partial\Omega_{1,2,3}$ are defined as in (4.38)-(4.41), see Figure 4.2. The boundary functions and the contact angles are chosen for each model problem as tabulated in Tables 4.3, resulting in model problems with unbounded solution. The closed-form solutions of these two model problems are given by (4.60) and (4.63), respectively.

4.2 Numerical Method

In this Section, we propose a numerical methodology to accurately find global numerical approximations of singular solutions of the Laplace-Young equation in domains with a corner or a cusp. The starting point of our approach is the finite volume element method (FVEM) [3] or the Galerkin finite element method (FEM) [46, 4], and two simple but crucial additional steps are made to arrive at a method that can accurately capture the singular behaviour. The first step is to consider a change of variable, with the new solution variable being smoother than the capillary height variable u and more amenable to accurate numerical approximation. The second step is to solve the PDE numerically in a new coordinate system, which allows us to accurately represent the discontinuous behaviour of the new solution variable at the singular point. We describe these two crucial ingredients of our methodology along with the FEM and FVEM discretizations, and show in the numerical results of Section 4.3 that this approach leads to a global approximation method for singular solutions of the Laplace-Young equation that recovers the proper asymptotic behaviour and is more accurate and has better convergence properties than numerical methods that were considered previously.

4.2.1 Change of Variable

From the asymptotic analysis results (4.22)-(4.29) we observe that the solutions we wish to approximate have the following asymptotic behaviour:

$$u(x,y) = O\left(\frac{1}{f_1(x)-f_2(x)}\right) \quad \text{as } x \to 0\,, \tag{4.75}$$

$$= \frac{O(1)}{f_1(x)-f_2(x)} \quad \text{as } x \to 0\,. \tag{4.76}$$

This implies that, if we transform the unknown function $u(x,y)$ as follows, the new unknown function $v(x,y)$ is a bounded function:

$$u(x,y) = \frac{v(x,y)}{f_1(x)-f_2(x)}\,. \tag{4.77}$$

We aim to approximate the solution of the boundary value problem, $u(x, y)$, by numerically approximating the new unknown function $v(x, y)$. Since $v(x, y)$ is bounded while $u(x, y)$ is unbounded, we expect a better quality of numerical approximation.

4.2.2 Change of Coordinates

An appropriate choice of coordinate system is essential for the asymptotic analysis of unbounded solutions of the Laplace-Young equation, as shown in Chapter 2 (also in [30, 40, 1]). We have also observed that an appropriate choice of coordinate system is also beneficial for the numerical approximation of unbounded solutions.

For Model Problem 1, we can observe as follows that the new unknown function v is discontinuous at the origin. From (4.21), we know that the solution u of Model Problem 1 behaves like $\frac{\cos\theta - \sqrt{k^2 - \sin^2\theta}}{kr}$ near the origin $r = 0$. This gives that the new unknown function v behaves like $\frac{\cos\theta - \sqrt{k^2 - \sin^2\theta}}{k}$ near the origin. Hence, as $r \to 0$, v approaches different values depending on the angle θ, so the new unknown function v has a jump discontinuity at the origin. Our idea is to expand the point of singularity on the boundary into a boundary line segment through a coordinate transformation in order to accurately approximate the discontinuous behaviour of v.

For Model Problem 2, since the boundaries for the cusp domain are curved boundaries, we would need special boundary elements (e.g., isoparametric elements) to accurately represent the cusp domain when approximating the unknown function through finite element approximation in the standard (x, y) coordinate system. However, the change to (s, t) coordinates introduced in Chapter 2 and illustrated in Figure 4.5 transforms a cusp domain into a rectangular domain, and hence no special treatment is needed for curved boundaries.

We use this (s, t) coordinate system for numerical simulation on domains with a corner or a cusp at

$(0,0)$. The (s,t) coordinate transformation as depicted in Figure 4.5 is given by

$$t = \frac{2y-(f_1+f_2)}{f_1-f_2}\,, \tag{4.78}$$

$$s = x\,. \tag{4.79}$$

The Cartesian coordinates can be expressed using the above coordinate system as

$$x = s\,, \tag{4.80}$$

$$y = \frac{t(f_1(s)-f_2(s))+(f_1(s)+f_2(s))}{2}\,, \tag{4.81}$$

$$= \frac{1+t}{2}f_1(s)+\frac{1-t}{2}f_2(s)\,. \tag{4.82}$$

When $t=1$, $y=f_1(x)$, and when $t=-1$, $y=f_2(x)$, so the domain of interest in the curvilinear (s,t) coordinate system can be written as (see Figure 4.5)

$$\Omega = \{(s,t)\in\mathbb{R}^2 : 0<s<1,\, -1<t<1\}\,. \tag{4.83}$$

With some calculation, the left-hand side of the Laplace-Young PDE can be rewritten in the curvilinear coordinate system as

$$\begin{aligned}\nabla\cdot T(u) = {} & \frac{\partial}{\partial s}\frac{u_x}{\sqrt{1+u_x^2+u_y^2}}+\frac{f_1'-f_2'}{f_1-f_2}\frac{u_x}{\sqrt{1+u_x^2+u_y^2}} \\ & +\frac{\partial}{\partial t}\left(\left(\frac{2}{f_1-f_2}\right)\frac{u_y}{\sqrt{1+u_x^2+u_y^2}}+\left(-\frac{f_1'+f_2'}{f_1-f_2}-t\frac{f_1'-f_2'}{f_1-f_2}\right)\frac{u_x}{\sqrt{1+u_x^2+u_y^2}}\right)\,,\end{aligned} \tag{4.84}$$

where

$$u_x = \frac{v_s}{f_1-f_2}-\frac{v(f_1'-f_2')}{(f_1-f_2)^2}-v_t\frac{(f_1'+f_2')+t(f_1'-f_2')}{(f_1-f_2)^2}\,, \tag{4.85}$$

$$u_y = \frac{2v_t}{(f_1-f_2)^2}\,. \tag{4.86}$$

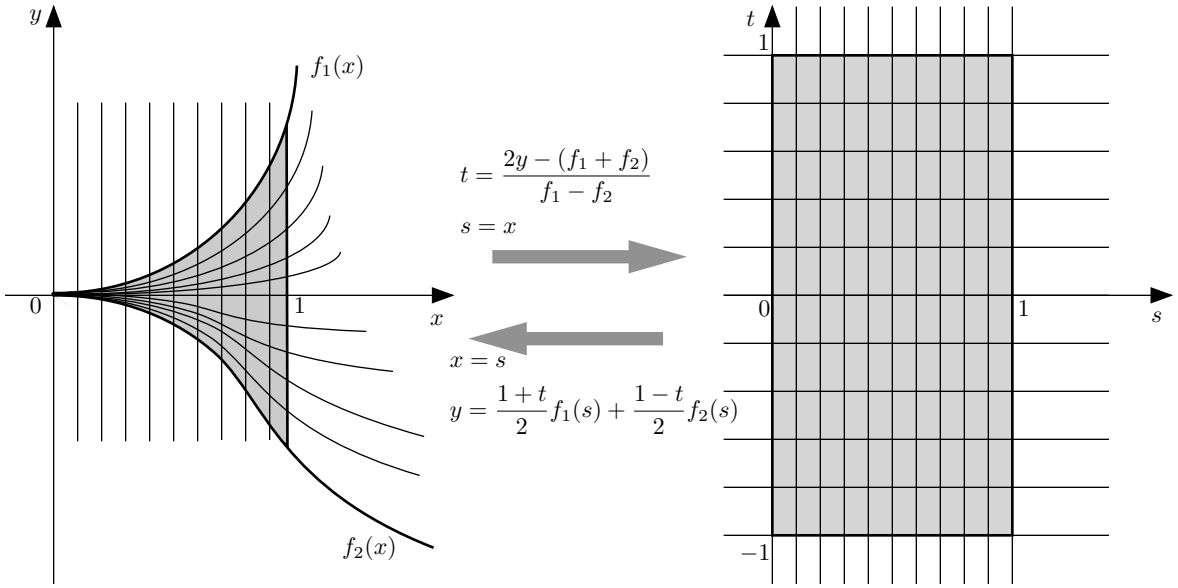
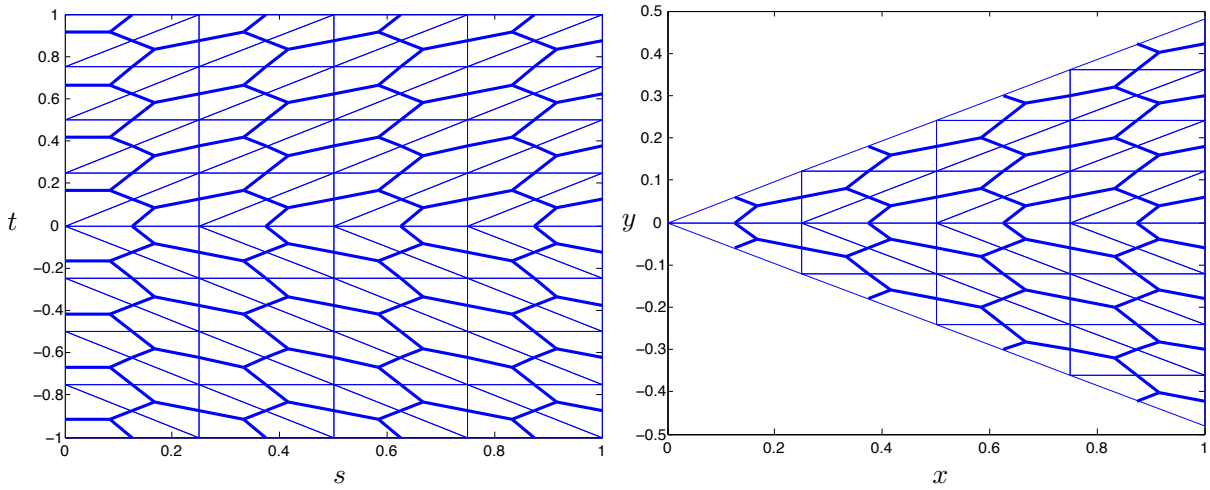


Figure 4.5: Coordinate transformation.

Also, the boundary conditions on $\partial\Omega_1$ and $\partial\Omega_2$ can be written as

$$
\begin{aligned}
& \vec{\nu}_{1,2} \cdot T(u) \\
= \;& \vec{\nu}_{1,2} \cdot \hat{s} \left(\frac{f_1 - f_2}{2} \frac{u_x}{\sqrt{1 + u_x^2 + u_y^2}} \right) + \vec{\nu}_{1,2} \cdot \hat{t} \left(\frac{u_y}{\sqrt{1 + u_x^2 + u_y^2}} + \frac{-(f_1' + f_2') - t(f_1' - f_2')}{2} \frac{u_x}{\sqrt{1 + u_x^2 + u_y^2}} \right) \\
= \;& \sqrt{1 + f_{1,2}'(s)^2} \cos \gamma_{1,2,} \qquad \text{on } \partial\Omega_{1,2} \,. \qquad (4.87)
\end{aligned}
$$

The boundary conditions for the boundary at $s = 0$ for all the model problems and $\partial\Omega_3$ of Model Problems 1 and 2 are as in (4.87) but with zero right-hand side. The left-hand side and the boundary conditions of the Asymptotic Laplace-Young PDE in the s, t coordinate system can be obtained by just neglecting the 1 in the denominator in the expressions above. Note that the point $(x, y) = (0, 0)$ corresponds to the line segment $(s = 0, t \in [-1, 1])$ in the (s, t) coordinate system.



(a) (s,t) space

(b) a corner domain in (x,y) space (without change of coordinates)

Figure 4.6: Finite elements and control volumes for the numerical methods. The thin lines give the finite element triangulation, which is used in both the FEM and the FVEM. The thick lines give the control volumes that are used in the FVEM. The grid in panel (a) can be used for corner domains or for cusp domains (depending on the boundary functions f_1 and f_2 that enter into the coordinate transformation formulas), and the grid in panel (b) is used for comparison simulations for corner domains (without coordinate transformation).

4.2.3 Discretized Boundary Value Problem

We approximate the new unknown function $v(s,t)$ in the new coordinate variables s and t numerically on the Cartesian grid in (s,t)-space of Figure 4.6(a) in the numerical results of Section 4.3, and for comparison we also perform some calculations on the corner domain of Figure 4.6(b) without a change of coordinates. We now describe the Galerkin Finite Element Method (FEM) and the Finite Volume Element Method (FVEM) discretizations.

4.2.3.1 Galerkin Finite Element method discretization

We follow the construction of the Finite Element space presented in Chapter 3 of Brenner and Scott [4]. Let N_{node} be the number of nodes created by finite element triangulation of the domain and $\mathcal{N}$ be the set of indices of the nodes, i.e., $\mathcal{N} = \{1, 2, ..., N_{\text{node}}\}$. The triangulation of the domain is as depicted

in Figure 4.6(a) (or Figure 4.6(b) for the corner problem without a change of coordinates). Also, we let $\mathcal{N}_{\text{Dirichlet}}$ be the indices of the nodes on the boundary with Dirichlet boundary condition. That is to say, for Model Problems 3 and 4,

$$(s_i, t_i) \in \overline{\partial\Omega_3} \qquad \Rightarrow \qquad i \in \mathcal{N}_{\text{Dirichlet}}, \tag{4.88}$$

where (s_i, t_i) is the location of the ith node, and for Model Problems 1 and 2 $\mathcal{N}_{\text{Dirichlet}} = \emptyset$ since there is no Dirichlet boundary. Let $\phi_i(s,t)$ be the standard continuous piecewise linear nodal basis function (tent function) that corresponds to node i in the finite element triangulation on domain Ω. We have that

$$\phi_i(s_j, t_j) = \delta_{i,j} \tag{4.89}$$

where $\delta_{i,j}$ is the Kronecker delta function. We approximate the unknown function v with a linear combination of these basis functions, i.e.,

$$v \quad \approx \quad v^h := \sum_{i=1}^{N_{\text{node}}} c_i \phi_i \,. \tag{4.90}$$

The $\{c_1, c_2, ..., c_{N_{\text{node}}}\}$ are the unknowns of the discretized boundary value problem. The Galerkin Finite Element discretization of Model Problems 1 and 2 can then be written as follows (the discretization of Model Problems 3 and 4 can be derived similarly):

$$\int_\Omega \left(\nabla \cdot T\left(\frac{\sum_{i=1}^{N_{\text{node}}} c_i \phi_i}{f_1(s) - f_2(s)} \right) \right) \phi_j \mathrm{d}A \quad = \quad \int_\Omega \frac{\sum_{i=1}^{N_{\text{node}}} c_i \phi_i}{f_1(s) - f_2(s)} \phi_j \mathrm{d}A \qquad \text{for } j \in \mathcal{N} \backslash \mathcal{N}_{\text{Dirichlet}} \,, \tag{4.91}$$

$$\frac{c_i}{f_1(s_i) - f_2(s_i)} \quad = \quad u_{\text{exact}}(s_i, t_i) \qquad \text{for } i \in \mathcal{N}_{\text{Dirichlet}} \,. \tag{4.92}$$

By the divergence theorem we can rewrite (4.91) as

$$\int_{\partial\Omega} \left(\nu \cdot T\left(\frac{\sum_{i=1}^{N_{\text{node}}} c_i \phi_i}{f_1(s) - f_2(s)} \right) \right) \phi_j \mathrm{d}l - \int_\Omega \left(T\left(\frac{\sum_{i=1}^{N_{\text{node}}} c_i \phi_i}{f_1(s) - f_2(s)} \right) \right) \cdot \nabla \phi_j \mathrm{d}A \quad = \quad \int_\Omega \sum_{i=1}^{N_{\text{node}}} c_i \frac{\phi_i \phi_j}{f_1(s) - f_2(s)} \mathrm{d}A$$
$$\text{for } j \in \mathcal{N} \backslash \mathcal{N}_{\text{Dirichlet}} \,. \tag{4.93}$$

By imposing the boundary conditions (4.87), we obtain the following equations:

$$\begin{aligned}
&\int_\Omega \left(T\left(\frac{\sum_{i=1}^{N_{\text{node}}} c_i \phi_i}{f_1(s) - f_2(s)} \right) \right) \cdot \nabla \phi_j \mathrm{d}A - \sum_{i=1}^{N_{\text{node}}} c_i \int_\Omega \frac{\phi_i \phi_j}{f_1(s) - f_2(s)} \mathrm{d}A \\
= \quad & \int_{\partial\Omega_1} \sqrt{1 + f_1'(s)^2} \cos\gamma_1 \phi_j \mathrm{d}l + \int_{\partial\Omega_2} \sqrt{1 + f_2'(s)^2} \cos\gamma_2 \phi_j \mathrm{d}l \qquad \text{for } j \in \mathcal{N} \backslash \mathcal{N}_{\text{Dirichlet}} . \quad (4.94)
\end{aligned}$$

After some calculation we can rewrite (4.94) together with (4.92) as the following system of nonlinear equations:

$$\begin{aligned}
&\int_{t=-1}^{1} \int_{s=0}^{1} (\phi_j)_s \left(\frac{f_1 - f_2}{2} \frac{u_x^h}{\sqrt{1 + (u_x^h)^2 + (u_y^h)^2}} \right) \\
&+ (\phi_j)_t \left(\frac{u_y^h}{\sqrt{1 + (u_x^h)^2 + (u_y^h)^2}} + \frac{-(f_1' + f_2') - t(f_1' - f_2')}{2} \frac{u_x^h}{\sqrt{1 + (u_x^h)^2 + (u_y^h)^2}} \right) \mathrm{d}s\mathrm{d}t \\
&- \sum_{i=1}^{N_{\text{node}}} c_i \int_{t=-1}^{1} \int_{s=0}^{1} \phi_i \phi_j \mathrm{d}s\mathrm{d}t \\
= \quad & \int_{\partial\Omega_1} \sqrt{1 + f_1'(s)^2} \cos\gamma_1 \phi_j \mathrm{d}l + \int_{\partial\Omega_2} \sqrt{1 + f_2'(s)^2} \cos\gamma_2 \phi_j \mathrm{d}l \qquad \text{for } j \in \mathcal{N} \backslash \mathcal{N}_{\text{Dirichlet}} , \quad (4.95) \\
c_i \quad = \quad & u_{\text{exact}}(x_i, y_i) \qquad \text{for } i \in \mathcal{N}_{\text{Dirichlet}} , \quad (4.96)
\end{aligned}$$

where

$$\begin{aligned}
u_x^h \quad &= \quad \sum_{i=1}^{N_{\text{node}}} c_i \left(\frac{(\phi_i)_s}{f_1 - f_2} - \frac{(\phi_i)(f_1' - f_2')}{(f_1 - f_2)^2} - (\phi_i)_t \frac{(f_1' + f_2') + t(f_1' - f_2')}{(f_1 - f_2)^2} \right) , \quad (4.97) \\
u_y^h \quad &= \quad \sum_{i=1}^{N_{\text{node}}} c_i \frac{2(\phi_i)_t}{(f_1 - f_2)^2} \quad (4.98)
\end{aligned}$$

and $(\phi_i)_s$ and $(\phi_i)_t$ are the partial derivatives of ϕ_i with respect to s and t. We can construct a system of nonlinear equations by integrating each of the terms in (4.95) numerically. Note that, although we are integrating the unbounded functions $v^h \phi_j / (f_1(s) - f_2(s))$, due to the change of coordinates the area element $\mathrm{d}A$ becomes $(f_1(s) - f_2(s))/2 \mathrm{d}s\mathrm{d}t$, hence the integrand becomes $2v^h \phi_j$ which is a piecewise quadratic polynomial, hence we avoid singular integration. We solve this system of nonlinear equations

with the Levenberg-Marquardt method to obtain the unknowns $\{c_1, c_2, ..., c_{N_{\text{node}}}\}$.[1] This gives a numerical approximation for v, and hence a numerical approximation of the solution of the boundary value problem u.

4.2.3.2 Finite Volume Element method discretization

The Finite Volume Element Method (FVEM) is a type of Petrov-Galerkin method that uses piecewise constant functions as test functions in the weak form, instead of using the finite element basis functions as in the Galerkin FEM. The test functions for the FVEM are chosen as follows:

$$\psi_j(s,t) \quad = \quad \begin{cases} 1 & \text{if } (s,t) \in \Omega_j \,, \\ 0 & \text{otherwise,} \end{cases} \tag{4.99}$$

where Ω_j are the control volumes constructed as in [3] (note that in [3] the control volumes are called "boxes"). As depicted in Figure 4.6(a) (and Figure 4.6(b)), the control volumes $\{\Omega_j\}_{j=1}^{N_{\text{node}}}$ are constructed by first computing the centroids of the finite element triangles, and then connecting those element centroids with the midpoints of the finite element triangle edges. This construction divides each finite element triangle into three quadrilaterals. The control volume Ω_j for finite element node j is then constructed as the union of the quadrilaterals adjacent to node j.

By substituting the test functions ϕ_j by ψ_j in the Galerkin Finite Element discretization (4.91) and

[1] In order to accelerate the speed of convergence of the Levenberg-Marquardt method, we first compute the numerical solution on a coarser grid and then interpolate it to the original grid and use it as the initial iterate of the Levenberg-Marquardt method.

after some calculation, we obtain the following system of nonlinear equations for the FVEM:

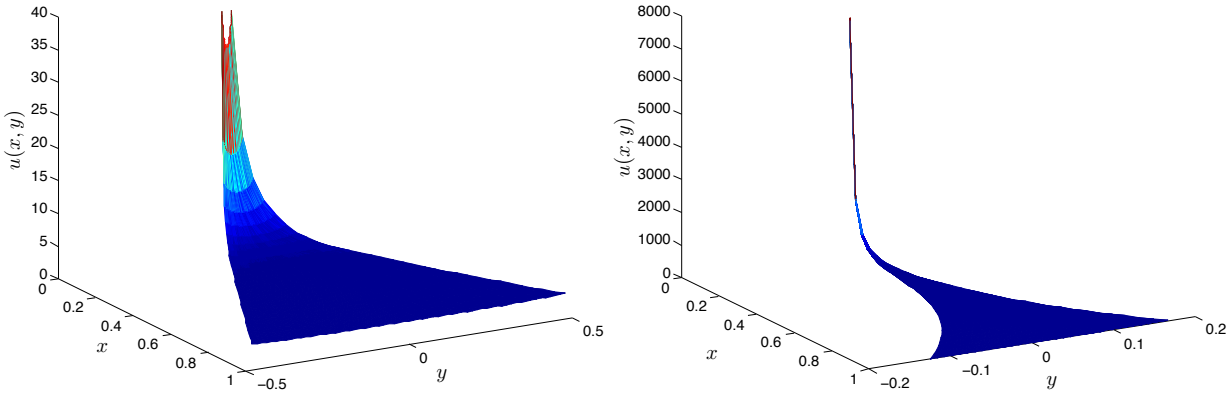
$$\begin{aligned}
&\int_{\partial\Omega_j} \vec{\nu}\cdot\hat{s}\left(\frac{f_1-f_2}{2}\frac{u_x^h}{\sqrt{1+(u_x^h)^2+(u_y^h)^2}}\right)\\
&+\vec{\nu}\cdot\hat{t}\left(\frac{u_y^h}{\sqrt{1+(u_x^h)^2+(u_y^h)^2}}+\frac{-(f_1'+f_2')-t(f_1'-f_2')}{2}\frac{u_x^h}{\sqrt{1+(u_x^h)^2+(u_y^h)^2}}\right)\mathrm{d}l\\
&-\sum_{i=1}^{N_{\text{node}}} c_i\int\int_{\Omega_j}\phi_i\mathrm{d}s\mathrm{d}t\\
=\;&\int_{\partial\Omega_1\cap\partial\Omega_j}\sqrt{1+f_1'(s)^2}\cos\gamma_1\mathrm{d}l+\int_{\partial\Omega_2\cap\partial\Omega_j}\sqrt{1+f_2'(s)^2}\cos\gamma_2\mathrm{d}l \qquad \text{for } j\in\mathcal{N}\backslash\mathcal{N}_{\text{Dirichlet}}\,,
\end{aligned} \tag{4.100}$$

where u_x and u_y are defined as in (4.97) and (4.98).

Again, we avoid singular integration by the change of coordinates, hence the integration can be done numerically without any special treatment for singular integration. We solve the resulting system of nonlinear equations with the Levenberg-Marquardt method.

Note that we choose the triangulations of Figures 4.6(a) and 4.6(b) symmetric with respect to the $t = 0$ and $y = 0$ axes, respectively. While this is not a requirement, we made this choice because some of our model problems are symmetric with respect to the $t = 0$ and $y = 0$ axes, and this choice of grid leads to numerical solutions that closely retain this symmetry.

The FEM is known to achieve optimality in the energy norm for linear elliptic PDEs, but it does not have a local conservation property. The FVEM has a local conservation property like the Finite Volume method; however, it does not necessarily produce an optimal approximation. We have conducted numerical experiments using both methods, and the results we obtained were very similar. For brevity, we mainly present the numerical experiment results obtained by the FVEM, except in a few places where we compare with the Galerkin FEM.



(a) Model Problem 1-1: Corner Problem

(b) Model Problem 2a-1: Cusp Problem

Figure 4.7: Model Problems 1-1 and 2a-1. FVEM solution on the (s,t)-type grid of Figure 4.6(a) with 33 $\times$ 65 nodes. Surface plots of the unbounded capillary surfaces in the corner and cusp domains.

4.3 Numerical Results

We now show that the numerical approximations we obtain with the computational methodology proposed in Section 4.2 for singular solutions of the Laplace-Young equation in domains with a corner or a cusp are accurate global approximations. As an initial illustration, surface plots for two numerical approximations of singular solutions of the Laplace-Young equation in domains with a corner and with a cusp are shown in Figure 4.7. In what follows, we first show how our numerical methods obtain accurate global solutions for unbounded solutions of the Laplace-Young equation in domains with a corner or a cusp, by carefully comparing with known asymptotic expansions and formal asymptotic series. We then numerically investigate the convergence behaviour of the methods we propose using known closed-form unbounded solutions for the Asymptotic Laplace-Young equation. The numerical results confirm that the computational methods we propose are accurate and have good convergence properties, and that they can be used with confidence to numerically investigate open problems on asymptotic solutions of the Laplace-Young equation in Section 4.4.

4.3.1 Laplace-Young Equation: Asymptotic Behaviour

We now investigate how well our numerical solutions can approximate the singular behaviour by comparing the numerical solutions to known asymptotic solutions for the Laplace-Young equation.

Model Problem 1: Corner Problem

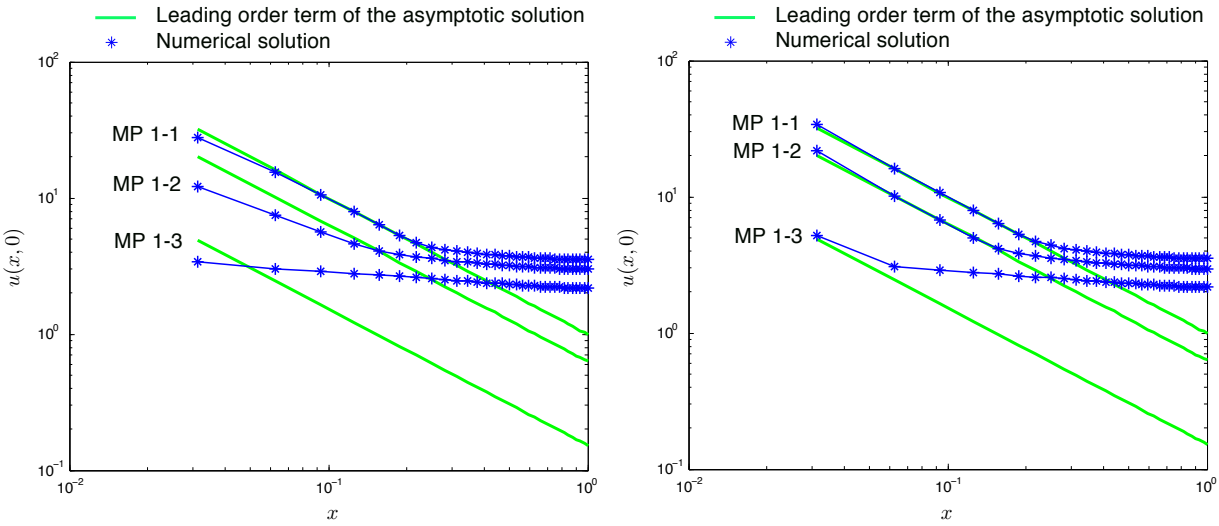
As given in (4.17), the leading order term of the asymptotic series solution of the Laplace-Young equation at a sharp corner is known. In Figure 4.8, we plot a horizontal cross-section (a cross-section along the x-axis or s-axis, see Figure 4.5) of the numerical approximation and the asymptotic approximation in log-log scale. In Figure 4.9, we plot a vertical cross-section (a cross-section along the line $x = 1/2^5$ or $s = 1/2^5$, see Figure 4.5) of the numerical approximation and the asymptotic approximation.

In order to illustrate the crucial benefits of the change of variable and change of coordinates that are the essential building blocks of the numerical methodology we proposed in Section 4.2, we compare four different choices for obtaining the numerical approximation using the FVEM: with or without change of variable, and with or without change of coordinates. The only published work on numerical approximation of singular capillary surfaces [41] also uses the FVEM, but it does not use a change of variable nor a change of coordinates, and thus corresponds to Figures 4.8(a) and 4.9(a).

As can be seen in Figures 4.8(d) and 4.9(d), the change of variable and the change of coordinates proposed in Sections 4.2.1 and 4.2.2 are very beneficial for the accuracy of the numerical approximations for the singular behaviours of the solutions of the test problems on a domain with a sharp corner. Note that we cannot conduct a numerical convergence study for these unbounded solutions of the Laplace-Young equation, as there is no known closed-form solution.

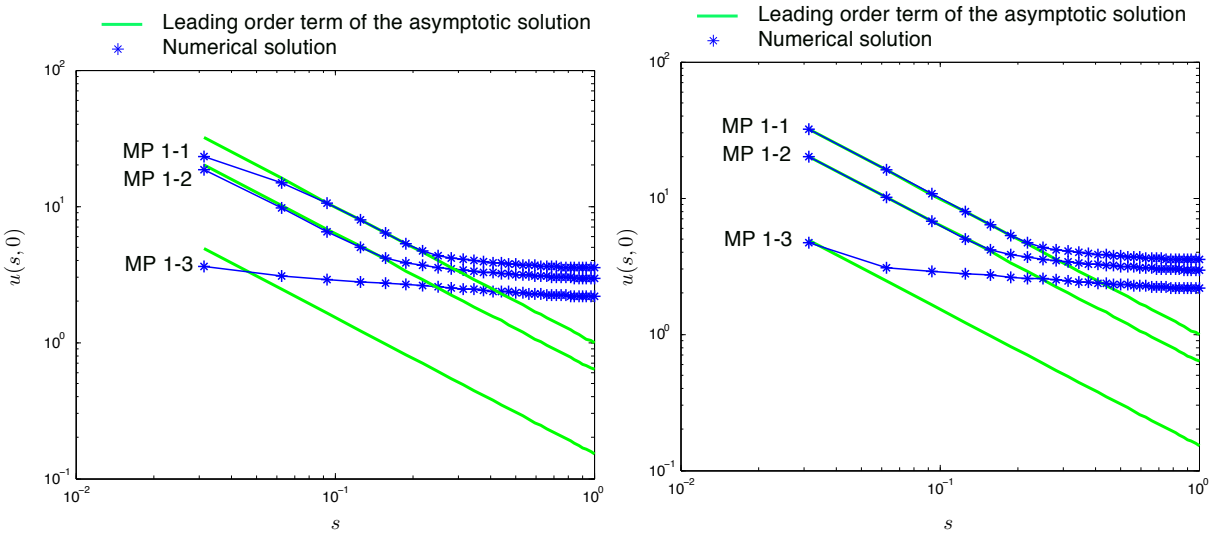
Model Problem 2: Cusp Problem

We now consider the Laplace-Young equation in a domain with a cusp. Unbounded cusp solutions are known to have a more severe singularity than the sharp corner problem. The leading order term of the



(a) without change of coordinates and without change of variable

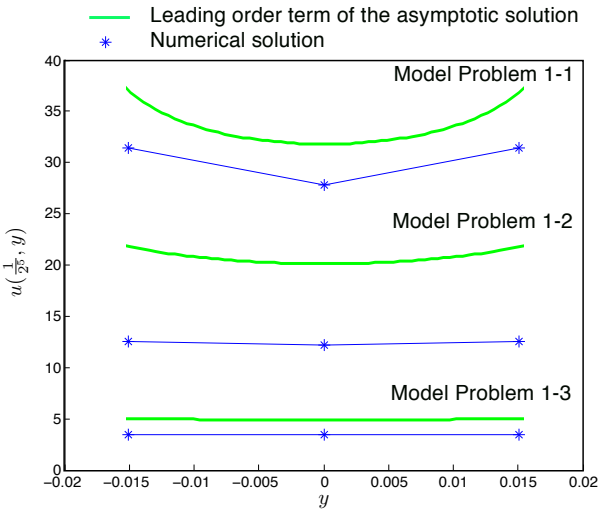
(b) without change of coordinates and with change of variable



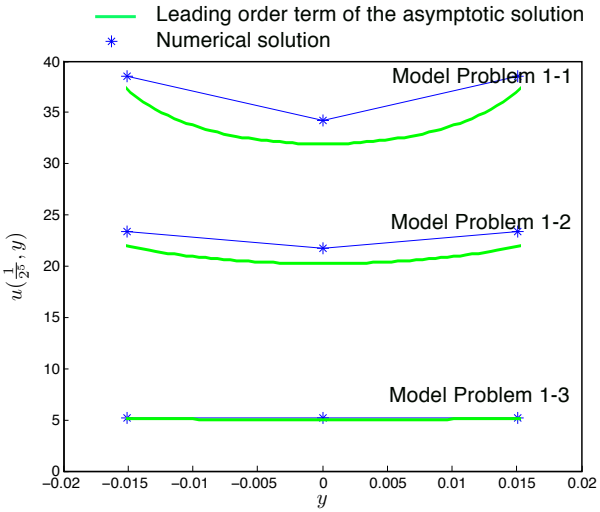
(c) with change of coordinates and without change of variable

(d) with change of coordinates and with change of variable

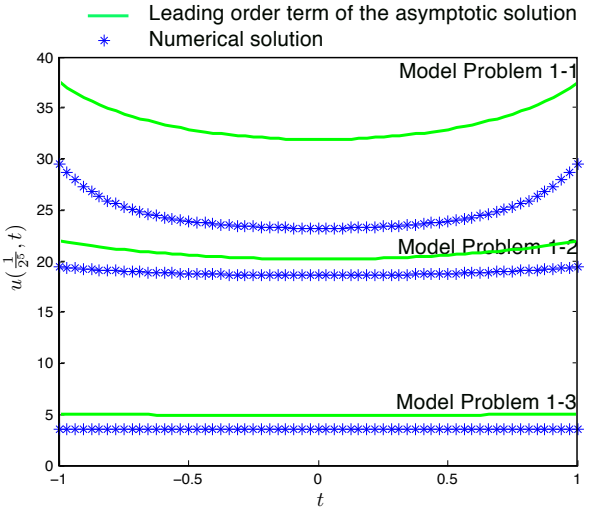
Figure 4.8: Model Problems 1-1, 1-2 and 1-3 (unbounded corner solutions). Panels (a) and (b) show FVEM solutions on the (x, y)-type grid of Figure 4.6(b) with 1089 nodes (no change of coordinates). Panels (c) and (d) show FVEM solutions on the (s, t)-type grid of Figure 4.6(a) with 33×65 nodes (with change of coordinates). Panels (a) and (c) are for computation of the original variable u, and panels (b) and (d) are for computation of the transformed variable v. The log-log plots show a comparison of the numerical solutions and the first-order asymptotic approximations in a horizontal cross section at $y = 0$ or $t = 0$. Panel (d) clearly gives the most accurate numerical solutions.



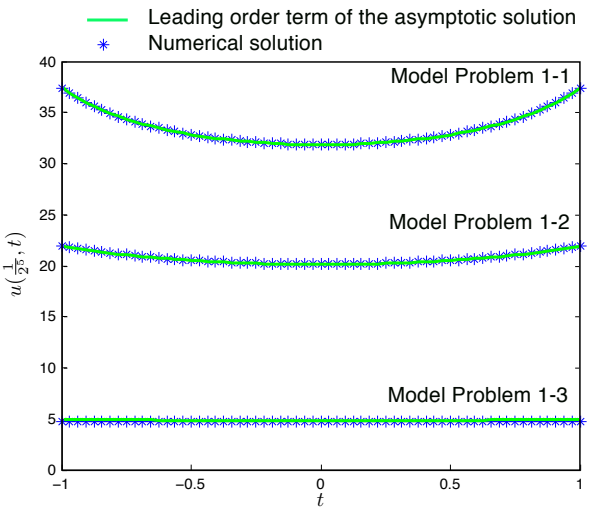
(a) without change of coordinates and without change of variable



(b) without change of coordinates and with change of variable



(c) with change of coordinates and without change of variable



(d) with change of coordinates and with change of variable

Figure 4.9: Model Problems 1-1, 1-2 and 1-3 (unbounded corner solutions). Panels (a) and (b) show FVEM solutions on the (x, y)-type grid of Figure 4.6(b) with 1089 nodes (no change of coordinates). Panels (c) and (d) show FVEM solutions on the (s, t)-type grid of Figure 4.6(a) with 33×65 nodes (with change of coordinates). Panels (a) and (c) are for computation of the original variable u, and panels (b) and (d) are for computation of the transformed variable v. The plots show a comparison of the numerical solutions and the first-order asymptotic approximations in a vertical cross section at $x = \frac{1}{2^5}$ or $s = \frac{1}{2^5}$ (the grid points closest to the singular point). Panel (d) clearly gives the most accurate numerical solutions.

asymptotic series solution is known, see (4.23). Also, as shown in Lemma 2.2.3, the first two terms of the formal asymptotic series $\tilde{u}$ are known:

$$\tilde{u} = \frac{\cos\gamma_1 + \cos\gamma_2}{f_1(s) - f_2(s)} - \sqrt{1 - \left(\frac{\cos\gamma_1(t+1) + \cos\gamma_2(t-1)}{2}\right)^2} \frac{f_1'(s) - f_1'(s)}{f_1(s) - f_2(s)}\,. \tag{4.101}$$

The formal asymptotic series of a boundary value problem is a series that satisfies the PDE and the boundary condition asymptotically, but, as opposed to the case of an asymptotic expansion, a bound on the error has not been proven. (There is no $O(\cdot)$ term in (4.101), but there is one in the asymptotic expansion (4.23).)

As can be seen in Figure 4.10, the numerical solution we obtain for Model Problem 2 with the change of variable and the change of coordinates proposed in Sections 4.2.1 and 4.2.2 accurately approximates the singular behaviour.

Although it is not know if the second order term of the formal asymptotic series of this problem is in fact the second order term of the asymptotic series solution, it can be seen in Figure 4.10(b) that the numerical solution appears to match better with the second order formal asymptotic series than with the first order asymptotic solution. It is particularly interesting that the domain where the asymptotic approximation is a good approximation seems to expand by adding a second term to the asymptotic series.

4.3.2 Asymptotic Laplace-Young Equation: Convergence Study

In the previous section, we have shown that the numerical approximations with a change of variable and a change of coordinates as proposed in Sections 4.2.1 and 4.2.2 exhibit the correct singular behaviour for singular solutions of the Laplace-Young equation. Since our interest is to obtain global approximations which are accurate both at the singularity and away from the singularity, we now show that the numerical solution in fact converges to the exact solution everywhere. It would be desirable to conduct a numerical convergence study for the Laplace-Young equation, but there is no known closed-form singular solution, and hence we cannot conduct a numerical convergence study. As we have discussed in Section 4.1.2, there are known exact solutions of the Asymptotic Laplace-Young equation, and it is known that they have

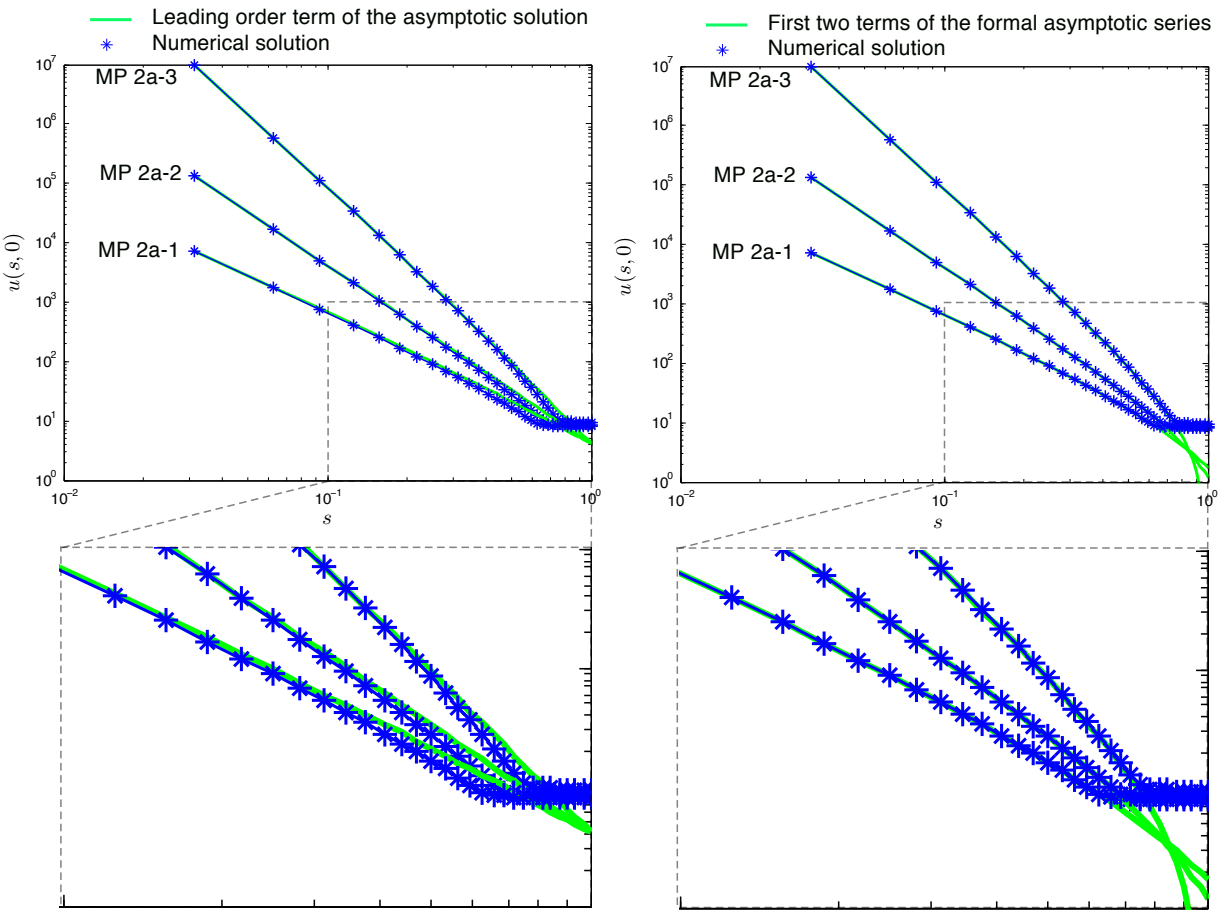


Figure 4.10: Model Problems 2a-1, 2a-2 and 2a-3 (unbounded cusp solutions). FVEM solutions on the (s,t)-type grid of Figure 4.6(a) with 33×65 nodes (with change of coordinates and with change of variable). The log-log plots in the left panels show a comparison of the numerical solutions with the first-order asymptotic solution in a horizontal cross section at $t = 0$. The log-log plots in the right panels show a comparison of the numerical solutions with the first two terms of the formal asymptotic series in a horizontal cross section at $t = 0$. It is clear that accurate numerical solutions are obtained.

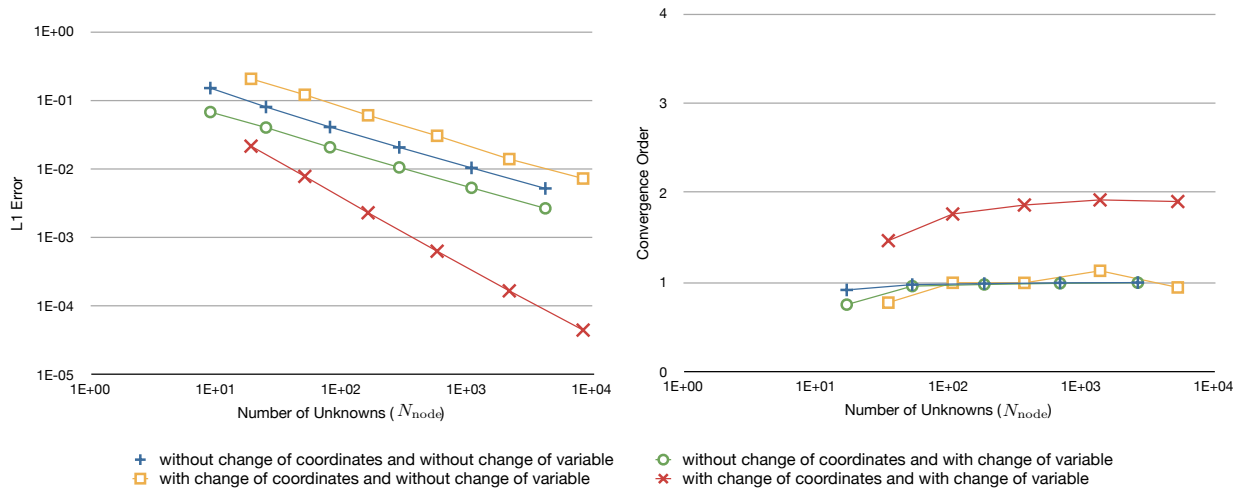


Figure 4.11: Model Problem 3 (unbounded corner solution for Asymptotic Laplace-Young). FVEM solutions on (s,t)-type grids (Figure 4.6(a)) and on (x,y)-type grids (Figure 4.6(b)), with and without change of variable. The plots show L_1 convergence of the numerical solutions obtained by the FVEM to the closed-form solution. The plots indicate that all four approaches converge, but it is clear that the method with change of variable and with change of coordinates converges significantly faster (with nearly second order accuracy) than the other approaches.

the same singular behaviour as the corresponding solutions of the Laplace-Young equation. We therefore conduct a numerical convergence study for the Asymptotic Laplace-Young equation in corner and cusp domains. Since the exact solution is in the L_1 function space but not in L_2, we conduct the convergence study in the L_1 norm.

Model Problem 3: Corner Problem

As can be seen in Figure 4.11, the FVEM numerical approximation with change of variable and change of coordinates as proposed in Sections 4.2.1 and 4.2.2 converges to the closed-form solution nearly quadratically, whereas the other approaches (no change of variable or no change of coordinates) only converge linearly.

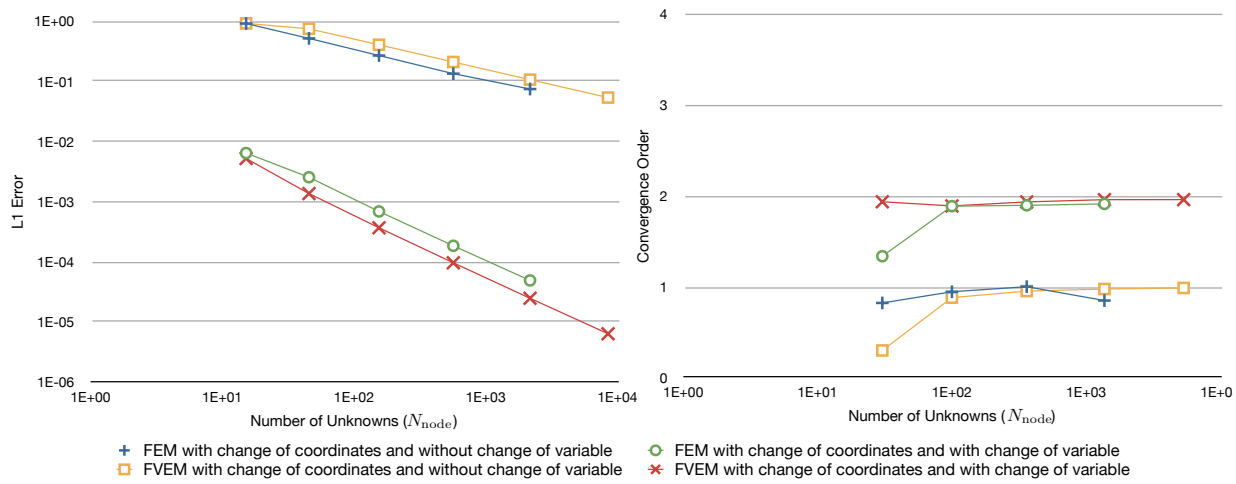


Figure 4.12: Model Problem 4 (unbounded cusp solution for Asymptotic Laplace-Young). FVEM and FEM solutions on (s,t)-type grids (Figure 4.6(a), with change of coordinates), with and without change of variable. The plots show L_1 convergence of the numerical solutions obtained by the FVEM and FEM to the closed-form solution. The plots indicate that all four approaches converge, but it is clear that the methods with change of variable converge significantly faster (with nearly second order accuracy).

Model Problem 4: Cusp Problem

We have also conducted a numerical convergence study for the circular cusp problem, where the solution has a more severe singularity than for the corner problem. For this problem, we have used both the Galerkin Finite Element Method (FEM) and the Finite Volume Element Method (FVEM) to show that both numerical schemes work well with the change of variable and the change of coordinates proposed in Sections 4.2.1 and 4.2.2. As can be seen in Figure 4.12, both the FEM and the FVEM achieve near-quadratic convergence with the change of variable and change of coordinates, while only linear convergence can be achieved without change of variable.

4.4 Conjectures on Open Problems

As shown in the previous section, we can obtain a globally accurate approximation of unbounded solutions of the Laplace-Young equation using the numerical methodology proposed in Section 4.2. We now

numerically approximate the solutions of two problems where the singular behaviour is not known yet analytically. Our numerical results will allow us to formulate conjectures on asymptotic behaviour for these open problems, which may guide further analytical study of these open problems.

4.4.1 Open Problem 1: Osculatory Cusp with non-Supplementary Contact Angles

As stated in Section 4.1.1.2, the leading order asymptotic behaviour of the unbounded solution of the Laplace-Young equation at an osculatory cusp is not known: in summary, a proof for the leading order asymptotic behaviour could not be obtained in Chapter 2 for the osculatory cusp because the author was not able to determine the formal asymptotic series. As shown in Lemma 2.2.3, the first two terms of the formal asymptotic series are known for the osculatory cusp case up to an additive constant in the coefficient of the second order term, i.e.,

$$\tilde{u} \quad = \quad \frac{\cos\gamma_1 + \cos\gamma_2}{f_1(s) - f_2(s)} + \left(-\sqrt{1 - \left(\frac{\cos\gamma_1 (t+1) + \cos\gamma_2 (t-1)}{2} \right)^2} + C_1 \right) \frac{f_1'(s) - f_1'(s)}{f_1(s) - f_2(s)}, \quad (4.102)$$

where $\tilde{u}$ asymptotically satisfies the boundary value problem, $C_1 = 0$ if the cusp is not an osculatory cusp, and C_1 is unknown if it is an osculatory cusp. One can see from the proofs in Chapter 2, that the unknown additive constant C_1 is the elusive key to the proof of the leading order behaviour of the osculatory cusp problem. The coefficient C_1 is unknown and may depend on the specific functional form of the boundary functions $f_1(s)$ and $f_2(s)$.

Physical intuition suggests that the singular behaviour of the unbounded capillary surface near a sharp corner or a cusp may be governed only by the distance between the two boundaries forming the sharp corner or cusp. In other words, one may think that the asymptotic behaviour should only depend on $f_1(s) - f_2(s)$ and its derivatives and not on $f_1(s)$ and $f_2(s)$ separately. This would imply that the formal asymptotic series would be the same for the four Model Problems 2a-2 and 2b, since $f_1(s) - f_2(s) = 7/24\, x^3$ for all these cases. If so, then $C_1 = 0$ is required also for the osculatory cusps of Model Problems 2b, since $C_1 = 0$ for the regular cusp of Model Problem 2a-2. But it is also possible that C_1 depends on the precise

functional form of $f_1(s)$ and $f_2(s)$.

In order to investigate this, we now numerically approximate the second order term of the formal asymptotic series by the following change of variable for the unknown function u:

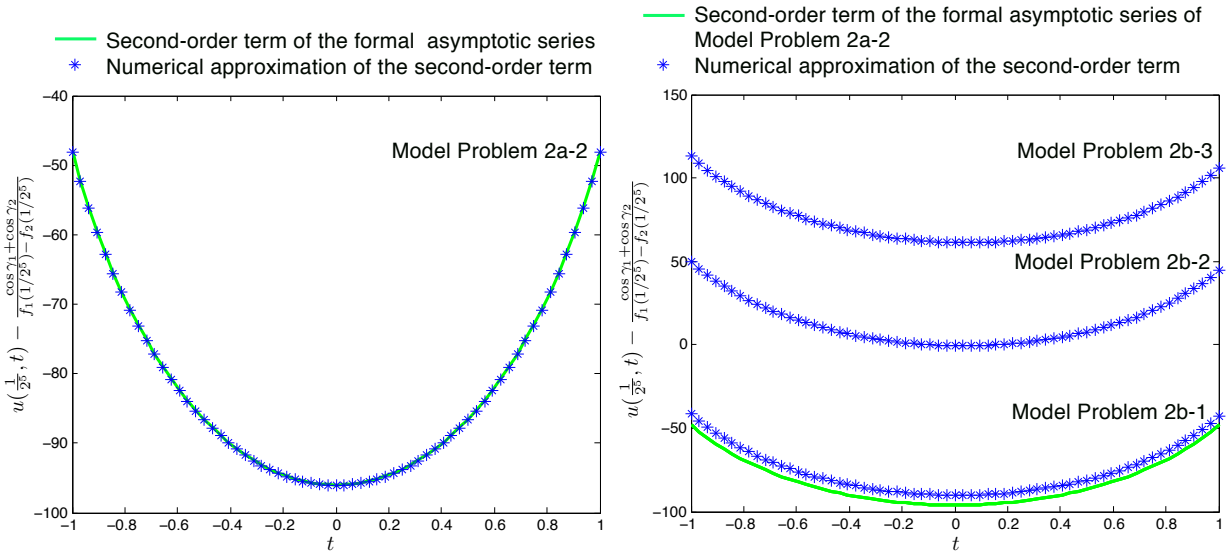
$$u(s,t) \quad = \quad \frac{\cos\gamma_1 + \cos\gamma_2}{f_1(s) - f_2(t)} + w(s,t)\frac{f_1'(s) - f_2'(s)}{f_1(s) - f_2(s)}\,. \tag{4.103}$$

We numerically approximate the new unknown function $w(s,t)$ in (s,t) coordinates, and we plot the second order term $w(s,t)\frac{f_1'(s)-f_2'(s)}{f_1(s)-f_2(s)}$ (or equivalently $u(s,t) - \frac{\cos\gamma_1+\cos\gamma_2}{f_1(s)-f_2(t)}$) obtained from the numerical approximation in Figure 4.13.

As can be seen in Figure 4.13(a), the known second order term of the formal asymptotic series for the regular cusp (Model Problem 2a-2) is approximated correctly using the change of variable (4.103). Also, Figure 4.13(b) shows that the second order term of the formal asymptotic series of the osculatory cusp case differs from the regular cusp case and is shifted up by constants, consistent with (4.102). These numerical results guide us in conjecturing that the additive constant C_1 of the coefficient of the second order formal asymptotic series changes depending on the leading order term of the boundary functions $f_1(s)$ and $f_2(s)$, and is strictly greater than 0. The numerical evidence from Figure 4.13(b) indeed indicates that C_1 is not zero for osculatory cusps and that the asymptotic behaviour depends on $f_1(s)$ and $f_2(s)$, and not just on the difference $f_1(s) - f_2(s)$. This conjecture on the unknown constant C_1 in (4.102), obtained from careful numerical investigation, can guide future analytical study of this case.

4.4.2 Open Problem 2: Infinite-curvature Cusp with Supplementary Contact Angles

Another open problem on the singular behaviour of the Laplace-Young equation in a cusp domain is the infinite-curvature boundary cusp (i.e., $\lim_{x\to 0} f_1''(x) = \infty$ or $\lim_{x\to 0} f_2''(x) = \infty$) with supplementary contact angles (i.e., $\gamma_1 + \gamma_2 = \pi$). It was proven in Chapter 2 that the cusp solution is bounded if the contact angles are supplementary angles and the boundaries forming the cusp have finite curvatures (but it is unbounded if the contact angles are not supplementary).



(a) Regular Cusp: Model Problem 2a-2

(b) Osculatory Cusp: Model Problem 2b

Figure 4.13: Model Problems 2a-2 (unbounded cusp solution) and 2b (unbounded osculatory cusp solution, open problem). FVEM solutions on an (s,t)-type grid (Figure 4.6(a)) with 33×65 nodes, with change of variable. The plots show vertical cross sections at $s = \frac{1}{2^5}$. The left panel shows how the numerical solution tracks the second-order term of the formal asymptotic series. The right panel supports the conjecture that $C_1 > 0$ in (4.102).

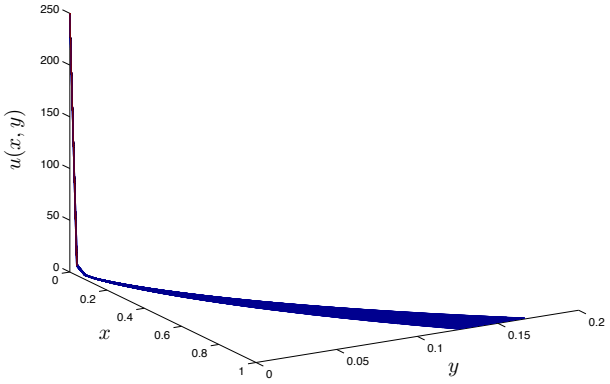
We conduct numerical experiments for Model Problem 2c (infinite curvature cusp) without change of variable. Lemma 2.2.1 gives that the solutions of Model Problems 2c-1 and 2c-3 are unbounded, and Model Problem 2c-2 is the supplementary contact angle case with unknown behaviour.

As can be seen in Figure 4.14, the numerical solution surface is bounded if the contact angles are supplementary for this case where the boundaries forming a cusp have infinite curvature at the cusp. We have conducted various other numerical experiments, however, we were not able to find any evidence of unbounded solutions if the contact angles are supplementary angles. Guided by these careful numerical results we conjecture that the solution of the Laplace-Young equation in a domain with a cusp is always bounded if the contact angles of the boundaries forming the cusp are supplementary angles. We also note that, as an additional check on the validity of our numerical approach, we have conducted further numerical experiments with cusps with finite curvature boundaries and with the same contact angles as Model Problems 2c, and we have confirmed numerically the theoretical prediction that the solution is bounded for supplementary contact angles, and unbounded otherwise.

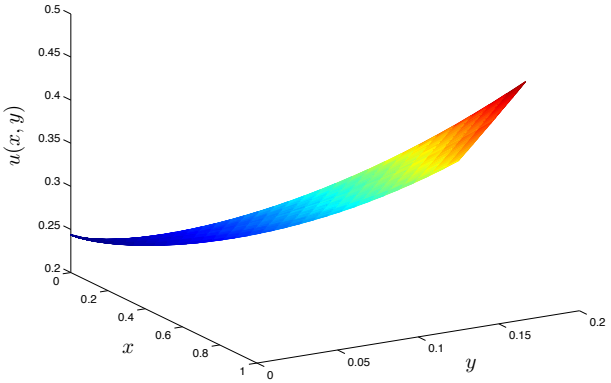
To conclude, we conjecture that the capillary surface in a cusp domain is bounded if the contact angles of the boundaries forming the cusp are supplementary angles, even if the curvatures of the boundaries are infinite. This conjecture on the open problem of the asymptotic behaviour of capillary surfaces in domains with a cusp and supplementary contact angles, obtained from careful numerical investigation, can guide further analytical study of this case.

4.5 Concluding Remarks on Chapter 4

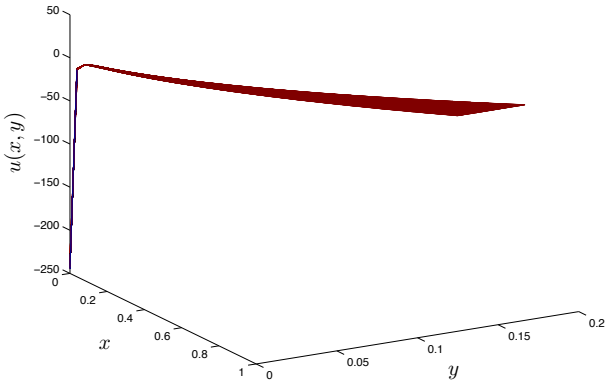
In this chapter, we have proposed a new methodology for the numerical study of unbounded capillary surfaces in domains with a sharp corner or a cusp. The methodology was developed by incorporating knowledge obtained from asymptotic analysis into a finite-element based approximation method. It contains two crucial ingredients that are combined with the finite volume element method (FVEM) [3] or the Galerkin finite element method (FEM) [46, 4]. The first ingredient is to consider a change of variable, with the new solution variable being smoother than the capillary height variable and more amenable to accurate



(a) Model Problem 2c-1: $\gamma_1 + \gamma_2 = \pi - \frac{1}{180}\pi$



(b) Model Problem 2c-2: $\gamma_1 + \gamma_2 = \pi$ (supplementary angles)



(c) Model Problem 2c-3: $\gamma_1 + \gamma_2 = \pi + \frac{1}{180}\pi$

Figure 4.14: Model Problems 2c-1, 2c-2 and 2c-3 (infinite curvature cusp). It is known that the solutions for 2c-1 and 2c-3 are unbounded, but the behaviour for 2c-2 is an open problem. FVEM solutions on an (s,t)-type grid (Figure 4.6(a)) with 33×65 nodes, with change of variable. Surface plots of the capillary surfaces are shown. The numerical result for Model Problem 2c-2 supports the conjecture that the solution is bounded in this case.

numerical approximation. The second ingredient is to solve the PDE numerically in a new coordinate system that is inspired by asymptotic analysis work, which allows us to accurately represent the discontinuous behaviour of the new solution variable at the singular point. We have shown in extensive numerical tests in domains with a sharp corner or a cusp that this approach leads to a global approximation method for singular solutions of the Laplace-Young equation that recovers the proper asymptotic behaviour and is more accurate and has better convergence properties than numerical methods that were considered for singular capillary surfaces before [41]. Although we have only considered the Laplace-Young equation and its steep slope approximation, it is likely that the methodology we have proposed can also be useful for other nonlinear elliptic PDEs with singularities. One important limitation of our approach is that in its present form it only works for problems with one singular point. Extension to problems with multiple singular points is a subject for further research.

Lastly, we were able to formulate conjectures for two open problems on the asymptotic behaviour of capillary surfaces in domains with a cusp. These conjectures are derived from very careful numerical investigation of these open problems using the powerful and highly accurate numerical methodology we propose, and they may guide future analytical work on these open problems.

Chapter 5

Cluster Newton Method for an Underdetermined Inverse Problem

I would like to dedicate this chapter of my PhD thesis to Mrs. Evie Geminiano, my Canadian guardian, who has bravely fought against a cancer and passed away in June 2011.

Our interest in the underdetermined inverse problem was initiated by the parameter identification problem of a pharmacokinetics model for the anti-cancer drug CPT-11 (also known as Irinotecan [2]). The essential difference of this problem from other parameter identification approaches is that instead of finding a single set of parameters that is suitable for the pharmacokinetics model to reproduce the clinically observed data, its aim is to find multiple sets of such parameters (see [25, 26]). The reason for finding these multiple sets of parameters is to take into account multiple relevant possibilities of the drug kinetics in the patient's body.

For underdetermined inverse problems, it is customary to add extra constraints to make the solution unique (e.g., the solution closest to some initial guess). If only one of many solutions is considered, it is often hard to know whether the characteristics of that solution are representative of all solutions, or whether they are a consequence of the choice of the extra constraints. Hence, we wish to sample many

solutions from the solution manifold of the underdetermined inverse problem. However, for a problem as complicated as a pharmacokinetics model aiming to model whole body drug kinetics, even to find one set of model parameters that fits a clinical observation can be time consuming. Thus, trying to find multiple sets of model parameters, one set by one set, can take computational time that is unrealistic for practical use, if the traditional method like the Levenberg-Marquardt method is used.

Motivated by this problem, we have constructed an algorithm to simultaneously find multiple solutions of an underdetermined inverse problem, in a new way that is significantly more robust and efficient than solving many separate inverse problems with different initial iterates. Our iterative scheme starts with a set (cluster) of initial points and by computing the forward problem at each point and fitting a hyperplane to the solution values in the sense of least squares, we obtain a linear approximation of the function that corresponds to the forward problem. This linear approximation aims to approximate the function in the broad domain covered by the cluster of initial points. Then by using this linear approximation, we estimate the solution of the inverse problem and move the cluster of points accordingly. By repeating this iteratively, the cluster of points becomes stationary and the points are close to being solutions of the inverse problem. In a next step, if the desired accuracy has not been met, we use Broyden's method to improve the accuracy by moving each point individually, using different approximated Jacobians, until the desired accuracy is achieved. Throughout this chapter, we shall refer to this method we have constructed as the Cluster Newton method (CN method).

Through numerical experiments, we have found that the Cluster Newton method requires far less function evaluations than applying the Levenberg-Marquardt method multiple times for various initial iterates. The Cluster Newton method is similar to Newton's method in the sense that it iteratively improves the approximation by approximating the forward problem by a linear function and inverting the linear approximation. Aside from moving a cluster of points instead of a single point, the Cluster Newton method differs from the traditional Newton's method in the sense that instead of approximating the Jacobian locally, we estimate it more globally in the domain covered by the cluster of points. The global approximation of the Jacobian acts as a regularization and we have observed that the Cluster Newton algorithm is robust against local optima caused by small-scale "roughness" of the function (e.g., caused by

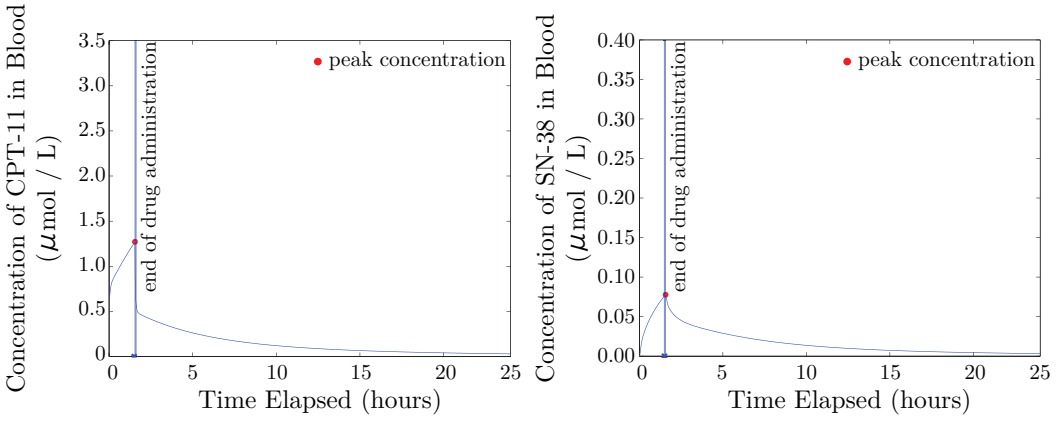
the inaccuracy of solving the forward problem), compared to a method like Levenberg-Marquardt. Such roughness appears in the coefficient identification problem of a system of ODEs when the ODE is solved numerically.

5.1 Motivation for Finding Multiple Solutions of the Underdetermined Inverse Problem

Our motivation for finding multiple solutions instead of a single solution of the underdetermined inverse problem came from the parameter identification problem of Arikuma et al.'s pharmacokinetics model for the anti-cancer drug CPT-11 [2]. Figure 5.1 shows the concentrations of CPT-11 and SN-38 (a metabolite of CPT-11) in blood simulated by the pharmacokinetics model using a set of model parameters found by the Levenberg-Marquardt method based on clinically measured data. The Levenberg-Marquardt method iteratively finds a solution of the underdetermined inverse problem near the initial iterate. We have chosen the initial iterate as the "typical" values of the model parameters listed in Arikuma et al. [2]. From Figure 5.1, we observe that the peak concentration occurs at time $t = 1.5$ (we denote this time as T_{max}) for both CPT-11 and SN-38. Also, observe that the peak concentration is around 1.3 $\mu mol/L$ for CPT-11 and 0.08 $\mu mol/L$ for SN-38. (We denote peak concentrations by C_{max})

In order to investigate further whether these obtained values are specific to the choice of the initial iterate or similar for most of the solutions of this underdetermined inverse problem, we have computed multiple solutions (multiple sets of model parameters) using the Levenberg-Marquardt method with different initial iterates close to the "typical value" listed in [2]. Figure 5.2 shows the concentrations of CPT-11 and SN-38 in blood simulated by the pharmacokinetics model using 1,000 sets of model parameters found by the Levenberg-Marquardt method with 1,000 different initial iterates.

We observe from Figure 5.2 that only the observation that $T_{max} = 1.5$ for CPT-11 seems independent of the choice of the initial iterate and may be a common feature among the solutions on the solution manifold of this underdetermined inverse problem. Other values (e.g., C_{max} for both CPT-11 and SN-38



(a) Concentration of the anti-cancer drug CPT-11.

(b) Concentration of the metabolite SN-38.

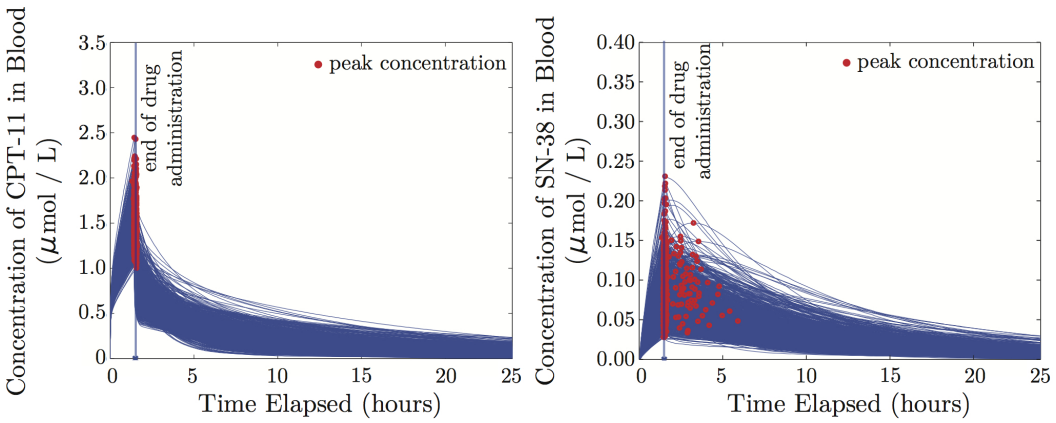
Figure 5.1: Drug and Metabolite concentration simulation based on a single set of model parameters found by the Levenberg-Marquardt Method.

and T_{max} for SN-38) are heavily dependent on the choice of the initial iterate. That is to say, these values cannot be determined precisely due to the underdetermined nature of the inverse problem. Although these values cannot be determined precisely, information on the range of C_{max} and T_{max} as obtained from the multiple solutions shown in Figure 5.2 is still of significant value in the context of the application problem.

The values of C_{max} and T_{max} can be measured, even though this is only realistic in a clinical experiment setting and not necessarily reliable (e.g., 16 blood samples are required from a patient in a day). Slatter et al. have obtained for 8 patients that C_{max} for CPT-11 is on average 2.26 $\mu mol/L$ (with standard deviation of 0.21), C_{max} of SN-38 is on average 0.04 $\mu mol/L$ (with standard deviation of 0.017), T_{max} of CPT-11 is 1.5 hours (with zero standard deviation) and T_{max} of SN-38 is on average 2.3 hours (with standard deviation of 1.0). All of the measured values except T_{max} of CPT-11 are different from what can be predicted from the single solution of Figure 5.1. However, Table 5.1 shows that these measured values for a small set of patients are within the range of the values of C_{max} and T_{max} obtained by solving for multiple solutions of the underdetermined inverse problem as shown in Figure 5.2. While the numbers in Table 5.1 indicate that the pharmacokinetics model is not perfect yet, this example does show that obtaining multiple solutions of the underdetermined inverse problem is useful for determining the general characteristics of the solutions on the solution manifold of the underdetermined inverse problem.

Table 5.1: Summary of C_{max} and T_{max} predicted from Figures 5.1 and 5.2, and clinically measured values.

	Predicted value from Figure 5.1	Range from Figure 5.2	Measured value in [44] (avg ± sd)
C_{max} of CPT-11 ($\mu mol/L$)	1.3	1.0∼2.5	2.26 ±0.21
C_{max} of SN-38 ($\mu mol/L$)	0.08	0.02∼0.23	0.04 ± 0.017
T_{max} of CPT-11 (hours)	1.5	1.5	1.5 ± 0
T_{max} of SN-38 (hours)	1.5	1.5∼6	2.3 ± 1.0



(a) Concentration of the anti-cancer drug CPT-11.

(b) Concentration of the metabolite SN-38.

Figure 5.2: 1,000 model parameter sets found by multiple application of the Levenberg-Marquardt Method.

The MATLAB implementation of the Levenberg-Marquardt method we used took 3.3 minutes to compute the single set of model parameters that was used to simulate the concentration of CPT-11 in blood plotted in Figure 5.1, using one core of an Intel Xeon X7350 3GHz processor. It took about 7 hours with a server machine with two quad-core Intel Xeon X7350 3GHz processors to find the 1,000 model parameters used to produce Figure 5.2. Our goal in this chapter is to develop an algorithm that can find such sets of parameters with significantly less computational cost.

5.2 Problem Statement

In this chapter, we consider the following underdetermined inverse problem (see 5.3 for an explanation of the matrix and vector notation used in this chapter):
Find parameter vector $\boldsymbol{x}$ such that

$$\boldsymbol{f}(\boldsymbol{x}) \quad = \quad \boldsymbol{y}^* \,, \tag{5.1}$$

where $\boldsymbol{y}^*$ is a given constant vector in $\mathbb{R}^n$, $\boldsymbol{f}$ is a C^1 vector function from $\mathcal{X} \subset \mathbb{R}^m$ to $\mathbb{R}^n$ with $m > n$, and the solution of (5.1) is not unique. We assume this inverse problem has the following properties:

- The evaluation of $\boldsymbol{f}$ (solving the forward problem) is computationally expensive. Thus, we would like to minimize the number of function evaluations.
- The Jacobian of $\boldsymbol{f}$ is not explicitly known.

We now denote a subset $\mathcal{X}_\epsilon^*$ of $\mathcal{X}$ to be the set containing all the values of $\mathcal{X}$ which approximately satisfy (5.1) with maximum norm relative residual less than ϵ, i.e.,

$$\mathcal{X}_\epsilon^* \quad := \quad \left\{ \boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^m \,:\, \max_{i=1,\cdots,n} \left| \left(f_i(\boldsymbol{x}) - y_i^* \right) / y_i^* \right| < \epsilon \right\}. \tag{5.2}$$

We note that in most cases $\mathcal{X}_\epsilon^*$ is an infinite set and often it is an unbounded set. We are only interested in a part of this set $\mathcal{X}_\epsilon^*$, namely the part that is relevant in the context of the problem, and that corresponds to a range of reasonable physiological parameters. For real physiological parameter identification applications, experiments described in the literature often provide a typical value for each of the parameters in the parameter vector $\boldsymbol{x}$, and a range in which each parameter can be expected to vary. The goal is to find multiple sets of parameter values in $\mathcal{X}_\epsilon^*$ that are close to these physiologically relevant values in the following sense. We use these typical values and range in an algorithm to define a box in $\mathcal{X}$ in which the initial cluster of points for our method is chosen in a uniformly random manner. This cluster is used to initialize the algorithm, and the location and size of this initial box will also influence the location and size

of the final solution cluster obtained by the algorithm. As explained below, the algorithm moves the initial cluster points towards the solution manifold by solving least squares problems in a way that minimizes the distance between successive iterates for each cluster point. The algorithm thus targets finding cluster points on the solution manifold that remain close to the centre of the initial box, and have a range that is similar to the range of the initial box. In this way, the selection of cluster points on the solution manifold is guided by the physiologically relevant parameter values that define the initial box. So we assume that we know the following regarding the physiologically relevant values of $\boldsymbol{x} \in \mathcal{X}$:

- a typical value of $\boldsymbol{x}$ is known (we denote this value as $\hat{\boldsymbol{x}}$, and it is used as the centre of the initial cluster of points for our algorithm)
- the typical relative ranges of the parameter values in $\boldsymbol{x}$ are known and we denote the typical relative range of the ith parameter in the parameter vector as v_i (see Equation 5.10 below; the ranges v_i are used to define the size of the initial cluster centred about $\hat{\boldsymbol{x}}$).

5.3 Matrix Notation

In general, we use a capital letter for a matrix, a bold lowercase letter for a vector and a lower case letter for a scalar quantity. Also, we introduce the following matrix related notations:

$$\boldsymbol{y} \quad : \quad \text{column vector,} \tag{5.3}$$

$$y_i \quad : \quad i\text{th component of column vector } \boldsymbol{y}, \tag{5.4}$$

$$x_{i\,j} \quad : \quad \text{element of matrix } X \text{ in column } i \text{ and row } j, \tag{5.5}$$

$$\boldsymbol{x}_{\cdot j} \quad : \quad j\text{th column of matrix } X \text{ as a column vector,} \tag{5.6}$$

$$\boldsymbol{x}_{i\cdot} \quad : \quad i\text{th row of matrix } X \text{ as a row vector,} \tag{5.7}$$

$$diag(\boldsymbol{y}) \quad : \quad \text{diagonal matrix whose } i\text{th column, } i\text{th row entry is } y_i, \tag{5.8}$$

$$||X||_F \quad : \quad \text{Frobenius norm of a matrix } X. \tag{5.9}$$

5.4 Algorithm: Cluster Newton Method

The Cluster Newton method for finding a cluster of parameter vectors $\boldsymbol{x}$ in the desired solution set $\mathcal{X}_\epsilon^*$ proceeds in two stages.

Starting from an initial cluster of points chosen uniformly randomly in the initial box centred about the typical value $\hat{\boldsymbol{x}}$ for the parameter vector, the first stage of the algorithm iteratively moves the initial cluster of points towards the desired solution set $\mathcal{X}_\epsilon^*$. In each iteration, a linear approximation of $\boldsymbol{f}(\boldsymbol{x})$ is constructed in the neighbourhood of the current cluster using the solutions of the forward problem at all the points in the cluster, and then this linear approximation is used to move the cluster closer to the desired solution set $\mathcal{X}_\epsilon^*$. The linear approximation is created using least squares fitting of a hyperplane to the solutions of the forward problem. A visual illustration of this iterative process is provided for a simple example problem in Section 5.5 (see Figure 5.5), and detailed pseudocode for all the steps of the algorithm is provided below.

Since Stage 1 of our algorithm moves all the points in the cluster collectively using one linear approximation, the improvement in the accuracy of the approximation eventually stalls after a number of iterations have been conducted. If the accuracy achieved in the first stage of the algorithm is not sufficient, we proceed to Stage 2 of the algorithm.

In the second stage of the algorithm, we further move each point towards the desired solution set $\mathcal{X}_\epsilon^*$ but now individually, using Broyden's method. For each point, we take the collective linear approximation of $\boldsymbol{f}(\boldsymbol{x})$ from Stage 1 to obtain an initial approximation for the Jacobian in Broyden's method. This avoids the large number of function evaluations that would be required for approximating the initial Jacobian for Broyden's method for each solution, for example by a finite difference scheme.

5.4.1 Algorithm Pseudocode (see also Figure 5.5 for graphical illustration of the algorithm)

Stage 1

1: Set up the initial cluster points and the target values.

1-1: Uniformly randomly generate an initial cluster of l points $\{\boldsymbol{x}_{\cdot j}^{(0)}\}_{j=1}^{l}$ in $\mathbb{R}^m$ in a box defined by the following inequalities:

$$\left|\frac{x_{ij}^{(0)} - \hat{x}_i}{\hat{x}_i}\right| < v_i \qquad \text{for } i = 1, 2, \cdots, m,\ j = 1, 2, \cdots, l\,, \tag{5.10}$$

where $\hat{x}_i$ is the typical value of the ith parameter and v_i is the typical relative range of the ith parameter. We require the number of points in the cluster to satisfy $l \geq m + 1$ so that we can construct a linear approximation to $\boldsymbol{f}(\boldsymbol{x})$ using the values of $\boldsymbol{f}$ at all the points in the cluster. We typically choose l to be much larger than $m + 1$ in order to make the algorithm more robust against small scale roughness in the function values and because we are interested in obtaining many more solutions than $m + 1$. The vectors $\boldsymbol{x}_{\cdot j}^{(0)}$ are stored in the columns of matrix $X^{(0)} = [\boldsymbol{x}_{\cdot 1}^{(0)}, \boldsymbol{x}_{\cdot 2}^{(0)}, \cdots, \boldsymbol{x}_{\cdot l}^{(0)}] \in \mathbb{R}^{m \times l}$.

1-2: Generate randomly perturbed target vectors $\{\boldsymbol{y}_{\cdot j}^{*}\}_{j=1}^{l}$ near $\boldsymbol{y}^{*}$. We choose each vector $\boldsymbol{y}_{\cdot j}^{*}$ so that

$$\max_{i=1,2,\cdots,n} \left|\frac{y_{ij}^{*} - y_i^{*}}{y_i^{*}}\right| < \eta\,, \tag{5.11}$$

with $\eta = 0.1$. The random perturbation is necessary to keep the algorithm well-posed, as explained in Section 5.4.2.3. We put these vectors in the columns of matrix Y^{*}.

2: For $k = 0, 1, 2, \cdots, K_1$

2-1: Solve the forward problem for each point $\boldsymbol{x}_{\cdot j}^{(k)}$, i.e., compute

$$\boldsymbol{y}_{\cdot j}^{(k)} \quad = \quad \boldsymbol{f}(\boldsymbol{x}_{\cdot j}^{(k)}) \qquad \text{for } j = 1, 2, \cdots, l\,, \tag{5.12}$$

with $Y^{(k)} = [\boldsymbol{y}_{\cdot 1}^{(k)}, \boldsymbol{y}_{\cdot 2}^{(k)}, \cdots, \boldsymbol{y}_{\cdot l}^{(k)}]$.

2-2: Construct a linear approximation of $\boldsymbol{f}$, i.e.,

$$\boldsymbol{f}(\boldsymbol{x}) \approx A^{(k)}\boldsymbol{x} + \boldsymbol{y}_o^{(k)}\,, \tag{5.13}$$

by fitting a hyperplane to $\{(\boldsymbol{x}_{\cdot j}^{(k)}, \boldsymbol{y}_{\cdot j}^{(k)})\}_{j=1}^{l}$. Recalling that we have chosen the number of the points in the cluster to be l, where $l \geq m+1$, the slope matrix $A^{(k)} \in \mathbb{R}^{n\times m}$ and the shift constant $\boldsymbol{y}_o^{(k)} \in \mathbb{R}^n$ can be found as the least squares solution of an overdetermined system of linear equations:

$$\min_{A^{(k)}\in\mathbb{R}^{n\times m}\,,\,\boldsymbol{y}_o^{(k)}\in\mathbb{R}^n} ||Y^{(k)} - (A^{(k)}X^{(k)} + Y_o^{(k)})||_F\,, \tag{5.14}$$

where $Y_o^{(k)}$ is a $n \times l$ matrix whose columns are all $\boldsymbol{y}_o^{(k)}$.

2-3: Find an update vectors $\boldsymbol{s}_{\cdot j}$ for each $\boldsymbol{x}_{\cdot j}^{(k)}$ using the linear approximation, i.e., find $\boldsymbol{s}_{\cdot j}$ s.t.

$$\boldsymbol{y}_{\cdot j}^{*} \quad = \quad A^{(k)}(\boldsymbol{x}_{\cdot j}^{(k)} + \boldsymbol{s}_{\cdot j}^{(k)}) + \boldsymbol{y}_o{}^{(k)} \qquad \text{for } j = 1, 2, \cdots, l\,. \tag{5.15}$$

As can be seen from the fact that matrix $A^{(k)} \in \mathbb{R}^{n\times m}$ is a rectangular matrix with more columns than rows, this is an underdetermined system of linear equations. Hence, we cannot uniquely determine $\boldsymbol{s}_{\cdot j}^{(k)}$ that satisfies Equation (5.15). Instead, we choose the vector $\boldsymbol{s}_{\cdot j}^{(k)}$ with the shortest scaled length, among all the solutions of Equation (5.15), as follows:

$$\min_{\boldsymbol{s}_{\cdot j}^{(k)}\in\mathbb{R}^m} ||(diag(\hat{\boldsymbol{x}}))^{-1}\boldsymbol{s}_{\cdot j}^{(k)}||_2 \tag{5.16}$$

$$\text{s.t.} \qquad \boldsymbol{y}_{\cdot j}^{*} = A^{(k)}(\boldsymbol{x}_{\cdot j}^{(k)} + \boldsymbol{s}_{\cdot j}^{(k)}) + \boldsymbol{y}_o^{(k)}\,, \tag{5.17}$$

for $j = 1, 2, \cdots, l$, where $\hat{\boldsymbol{x}} = [\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_m]^{\mathrm{T}}$.

We note that we have scaled Expression (5.16) with the diagonal matrix $diag(\hat{\boldsymbol{x}})^{-1}$ since the order of magnitude of the parameter values in vector $\hat{\boldsymbol{x}}$ can vary significantly, and finding the vector with shortest scaled length leads to a more robust method.

2-4: Find new points approximating the solution manifold $\mathcal{X}^*$ by updating $X^{(k)}$. If necessary, we first shrink the length of the vector $\boldsymbol{s}_{\cdot j}^{(k)}$ until the point $\boldsymbol{x}_{\cdot j}^{(k)} + \boldsymbol{s}_{\cdot j}^{(k)}$ is in the domain of the function $\boldsymbol{f}$ by the following simple procedure:

For $j = 1, 2, \cdots, l$

While $(\boldsymbol{x}_{\cdot j}^{(k)} + \boldsymbol{s}_{\cdot j}^{(k)}) \notin \mathcal{X}$

$$\boldsymbol{s}_{\cdot j}^{(k)} = \frac{1}{2}\boldsymbol{s}_{\cdot j}^{(k)} \tag{5.18}$$

End while

End for

Then

$$\boldsymbol{x}_{\cdot j}^{(k+1)} \quad = \quad \boldsymbol{x}_{\cdot j}^{(k)} + \boldsymbol{s}_{\cdot j}^{(k)} \qquad \text{for } j = 1, 2, \cdots, l\,, \tag{5.19}$$

with $X^{(k+1)} = [\boldsymbol{x}_{\cdot 1}^{(k+1)}, \boldsymbol{x}_{\cdot 2}^{(k+1)}, \cdots, \boldsymbol{x}_{\cdot l}^{(k+1)}]$.

End for.

Stage 2: Broyden's method

3: Set up the initial Jacobian approximation for each point $\boldsymbol{x}_{\cdot j}^{(k)}$:

$$J_{(j)}^{(K_1+1)} \quad = \quad A^{(K_1)} \qquad \text{for } j = 1, 2, \cdots, l\,. \tag{5.20}$$

4: For $k = K_1 + 1, \cdots, K_2$

4-1: Solve the forward problem for each point $\boldsymbol{x}_{\cdot j}^{(k)}$, i.e.,

$$\boldsymbol{y}_{\cdot j}^{(k)} \quad = \quad \boldsymbol{f}(\boldsymbol{x}_{\cdot j}^{(k)}) \qquad \text{for } j = 1, 2, \cdots, l\,. \tag{5.21}$$

4-2: If $k \neq K_1 + 1$ then update the Jacobian for each point using Broyden's method (see [23] or [11]) as follows:

$$J_{(j)}^{(k)} = J_{(j)}^{(k-1)} + (\boldsymbol{y}_{\cdot j}^{(k)} - \boldsymbol{y}^*) \frac{(\boldsymbol{s}_{\cdot j}^{(k-1)})^{\mathrm{T}}}{||\boldsymbol{s}_{\cdot j}^{(k-1)}||^2} \qquad \text{for } j = 1, 2, \cdots, l. \tag{5.22}$$

4-3: Find the search direction vector $\boldsymbol{s}_{\cdot j}^{(k)}$ for each $\boldsymbol{x}_{\cdot j}^{(k)}$ using the approximate Jacobian, i.e., $\boldsymbol{s}_{\cdot j}^{(k)}$ is given by the minimum norm solution of an underdetermined linear system:

$$\min_{\boldsymbol{s}_{\cdot j}^{(k)} \in \mathbb{R}^m} ||(diag(\hat{\boldsymbol{x}}))^{-1} \boldsymbol{s}_{\cdot j}^{(k)}||_2 \tag{5.23}$$

$$\text{s.t.} \qquad \boldsymbol{y}^* - \boldsymbol{y}_{\cdot j}^{(k)} = J_{(j)}^{(k)} \boldsymbol{s}_{\cdot j}^{(k)} \tag{5.24}$$

for $j = 1, 2, \cdots, l$.

4-4: Find new points approximating the solution manifold $\mathcal{X}^*$ by updating $X^{(k)}$, i.e.,

For $j = 1, 2, \cdots, l$

While $(\boldsymbol{x}_{\cdot j}^{(k)} + \boldsymbol{s}_{\cdot j}^{(k)}) \notin \mathcal{X}$

$$\boldsymbol{s}_{\cdot j}^{(k)} = \frac{1}{2} \boldsymbol{s}_{\cdot j}^{(k)} \tag{5.25}$$

End while

End for

Then

$$\boldsymbol{x}_{\cdot j}^{(k+1)} = \boldsymbol{x}_{\cdot j}^{(k)} + \boldsymbol{s}_{\cdot j}^{(k)} \qquad \text{for } j = 1, 2, \cdots, l\,. \tag{5.26}$$

End for.

5.4.2 Further Details on Some Key Steps in the Algorithm

5.4.2.1 Finding the linear approximation (line 2-2)

When constructing the linear approximation in Stage 1, we solve for the least squares solution of the overdetermined system of linear equations (5.14). To show how this computation can be done, we first rewrite (5.14) in standard matrix-vector multiplication form and then solve a set of least squares problems. We first rewrite the matrix in expression (5.14), by considering its n rows:

$$
\begin{aligned}
& Y^{(k)} - (A^{(k)}X^{(k)} + Y_o^{(k)}) \\
&= \begin{bmatrix} \boldsymbol{y}_{1\cdot}^{(k)} \\ \boldsymbol{y}_{2\cdot}^{(k)} \\ \vdots \\ \boldsymbol{y}_{n\cdot}^{(k)} \end{bmatrix} - \left(\begin{bmatrix} \boldsymbol{a}_{1\cdot}^{(k)} \\ \boldsymbol{a}_{2\cdot}^{(k)} \\ \vdots \\ \boldsymbol{a}_{n\cdot}^{(k)} \end{bmatrix} X^{(k)} + \begin{bmatrix} y_{o1}^{(k)} \cdots y_{o1}^{(k)} \\ y_{o2}^{(k)} \cdots y_{o2}^{(k)} \\ \vdots \\ y_{on}^{(k)} \cdots y_{on}^{(k)} \end{bmatrix} \right),
\end{aligned} \tag{5.27}
$$

where $\boldsymbol{y}_{i\cdot}^{(k)} \in \mathbb{R}^{1\times l}$ and $\boldsymbol{a}_{i\cdot}^{(k)} \in \mathbb{R}^{1\times m}$. By taking the transpose of both sides we obtain the following expression:

$$
\begin{aligned}
& (Y^{(k)} - (A^{(k)}X^{(k)} + Y_o^{(k)}))^{\mathrm{T}} \\
&= \begin{bmatrix} \boldsymbol{y}_{1\cdot}^{(k)\mathrm{T}} \cdots \boldsymbol{y}_{n\cdot}^{(k)\mathrm{T}} \end{bmatrix} - \left(X^{(k)\mathrm{T}} \begin{bmatrix} \boldsymbol{a}_{1\cdot}^{(k)\mathrm{T}} \cdots \boldsymbol{a}_{n\cdot}^{(k)\mathrm{T}} \end{bmatrix} + \begin{bmatrix} y_{o1}^{(k)} \cdots y_{on}^{(k)} \\ y_{o1}^{(k)} \cdots y_{on}^{(k)} \\ \vdots \\ y_{o1}^{(k)} \cdots y_{on}^{(k)} \end{bmatrix} \right),
\end{aligned} \tag{5.28}
$$

where $\boldsymbol{y}_{i\cdot}^{(k)\mathrm{T}} \in \mathbb{R}^{l\times 1}$, $X^{(k)\mathrm{T}} \in \mathbb{R}^{l\times m}$ and $\boldsymbol{a}_{i\cdot}^{(k)\mathrm{T}} \in \mathbb{R}^{m\times 1}$. We now observe that expression (5.14) is equivalent to n independent least squares problems:

$$\min_{\boldsymbol{a}_{i\cdot}^{(k)}\in\mathbb{R}^m,\, y_{oi}^{(k)}\in\mathbb{R}} \left\| \boldsymbol{y}_{i\cdot}^{(k)\mathrm{T}} - \left(X^{(k)\mathrm{T}}\boldsymbol{a}_{i\cdot}^{(k)\mathrm{T}} + \begin{bmatrix} y_{oi}^{(k)} \\ y_{oi}^{(k)} \\ \vdots \\ y_{oi}^{(k)} \end{bmatrix} \right) \right\|_2 \qquad \text{for } i = 1, 2, \cdots, n. \tag{5.29}$$

This can be concisely written as follows

$$\min_{\tilde{\boldsymbol{a}}_{i\cdot}^{(k)}\in\mathbb{R}^{m+1}} ||\boldsymbol{y}_{i\cdot}^{(k)\mathrm{T}} - \tilde{X}^{(k)\mathrm{T}}\tilde{\boldsymbol{a}}_{i\cdot}^{(k)\mathrm{T}}||_2 \qquad \text{for } i = 1, 2, \cdots, n, \tag{5.30}$$

where

$$\tilde{x}_{ij}^{(k)} = \begin{cases} x_{ij}^{(k)} & \text{for } i \le m \\ 1 & \text{for } i = m+1 \end{cases} \qquad \tilde{a}_{ij}^{(k)} = \begin{cases} a_{ij}^{(k)} & \text{for } j \le m \\ y_{oi}^{(k)} & \text{for } j = m+1\,. \end{cases} \tag{5.31}$$

Noting that $\tilde{X}^{(k)\mathrm{T}}$ is an $l \times (m+1)$ matrix, the solutions of the least squares problems (5.30) are the least squares solutions of overdetermined systems of linear equations if $l > m+1$. Assuming $\mathrm{rank}(\tilde{X}^{(k)\mathrm{T}}) = m+1$ (see Section 5.4.2.3), the normal equations of the first kind (see [38]) provide the least squares solution:

$$\tilde{\boldsymbol{a}}_{i\cdot}^{(k)\mathrm{T}} \;=\; (\tilde{X}^{(k)}\tilde{X}^{(k)\mathrm{T}})^{-1}(\tilde{X}^{(k)}\boldsymbol{y}_{i\cdot}^{(k)\mathrm{T}}) \qquad \text{for } i = 1, 2, \cdots, n. \tag{5.32}$$

The actual computation is done using QR decomposition for numerical stability reasons. As a result we obtain the matrix $A^{(k)}$ and the vector $\boldsymbol{y}_o^{(k)}$ such that $\boldsymbol{y}_{\cdot j}^{(k)} \approx A^{(k)}\boldsymbol{x}_{\cdot j}^{(k)} + \boldsymbol{y}_o^{(k)}$ for $j = 1, 2, \cdots, l$.

5.4.2.2 Solving the underdetermined system of linear equations (line 2-3)

The minimum norm solution of the underdetermined systems of linear equations (5.16)-(5.17) can be computed as the solution of the normal equation of the second kind (see [38]). Equation (5.17) can be rewritten as follows:

$$\boldsymbol{y}^*_{\cdot j} - A^{(k)}\boldsymbol{x}^{(k)}_{\cdot j} - \boldsymbol{y}^{(k)}_o = A^{(k)}(diag(\hat{\boldsymbol{x}}))(diag(\hat{\boldsymbol{x}}))^{-1}\boldsymbol{s}^{(k)}_{\cdot j} \tag{5.33}$$

for $j = 1, 2, \cdots, l$. Thus by using the normal equations of the second kind, $\boldsymbol{s}^{(k)}_{\cdot j}$ can be expressed as follows:

$$\boldsymbol{s}^{(k)}_{\cdot j} = (diag(\hat{\boldsymbol{x}}))^2 A^{(k)\mathrm{T}} \left(A^{(k)}(diag(\hat{\boldsymbol{x}}))^2 A^{(k)\mathrm{T}}\right)^{-1} (\boldsymbol{y}^*_{\cdot j} - A^{(k)}\boldsymbol{x}^{(k)}_{\cdot j} - \boldsymbol{y}^{(k)}_o)\,, \tag{5.34}$$

for $j = 1, 2, \cdots, l$. Again, the actual computation is done using QR decomposition.

5.4.2.3 Randomly perturbed target values (line 1-2)

In line 1-2 of the algorithm, we generate randomly perturbed target values $\{\boldsymbol{y}^*_{\cdot j}\}_{j=1}^{l}$. This step is necessary in order to iteratively repeat Stage 1 of the algorithm, since it makes the matrix $X^{(k)}$ in line 2-2 full rank, which is required to make the overdetermined system (5.14) uniquely solvable.

We first observe that each of the overdetermined systems (5.30) has a unique least square solution if and only if rank $\tilde{X}^{(k)} = m + 1$. Further, rank $\tilde{X}^{(k)} = m + 1$ requires rank $X^{(k)} = m$. On the other hand, when solving the underdetermined system of linear equations (5.16)-(5.17), if we had not randomly perturbed the target (i.e., if we take $\boldsymbol{y}^*_{\cdot j} = \boldsymbol{y}^*$ for all j), then we would have the following relationships:

$$A^{(k)}\left(\boldsymbol{x}^{(k)}_{\cdot i} + \boldsymbol{s}^{(k)}_{\cdot i}\right) = A^{(k)}\left(\boldsymbol{x}^{(k)}_{\cdot j} + \boldsymbol{s}^{(k)}_{\cdot j}\right) \equiv \boldsymbol{b} \qquad \text{for any } i, j = 1, 2, \cdots l. \tag{5.35}$$

In other words,

$$[A^{(k)}, -\boldsymbol{b}] \begin{pmatrix} \boldsymbol{x}_{\cdot j}^{(k)} + \boldsymbol{s}_{\cdot j}^{(k)} \\ 1 \end{pmatrix} = \mathbf{0} \qquad \text{for } j = 1, 2, \cdots, l\,. \tag{5.36}$$

Hence, if we define $\tilde{A}^{(k)} = [A^{(k)}, -\boldsymbol{b}] \in \mathbb{R}^{n \times (m+1)}$, then

$$\begin{pmatrix} \boldsymbol{x}_{\cdot j}^{(k)} + \boldsymbol{s}_{\cdot j}^{(k)} \\ 1 \end{pmatrix} \in \mathcal{N}(\tilde{A}^{(k)}) \qquad \text{for } j = 1, 2, \cdots l, \tag{5.37}$$

where $\mathcal{N}(\tilde{A}^{(k)})$ denotes the null space of $\tilde{A}^{(k)}$. Also note that dim $\mathcal{N}(\tilde{A}^{(k)}) = m + 1 -$ rank $\tilde{A}^{(k)} < m + 1$ unless $\tilde{A}^{(k)} = 0$. (Note that rank $\tilde{A}^{(k)} =$ rank $A^{(k)}$.)

Hence, if $\boldsymbol{x}_{\cdot j}^{(k+1)}$ is chosen to be $\boldsymbol{x}_{\cdot j}^{(k)} + \boldsymbol{s}_{\cdot j}^{(k)}$, rank $\tilde{X}^{(k+1)} \leq$ dim $\mathcal{N}(\tilde{A}^{(k)}) < m + 1$. This would imply that each least squares problem (5.30) in the the $k + 1$th iteration is rank deficient and does not have a unique solution.

On the other hand, if we seek the minimum norm solution of the underdetermined system of linear equations (5.16)-(5.17) for randomly perturbed targets $\{\boldsymbol{y}_{\cdot j}^*\}_{j=1}^l$, then generically rank $\tilde{X}^{(k+1)} = m + 1$ and each least squares problem (5.30) in the $k + 1$th iteration will have a unique solution.

5.4.2.4 Some further implementation details

Note that the function evaluations at each point in the cluster in lines 2-1 and 4-1 are independent of each other. Hence, these lines can be implemented in an embarrassingly parallel way. Since most of the computational cost is spent on computing the function values in these two lines, the computation time required by the Cluster Newton method is almost inversely proportional to the number of CPU cores that can be utilized. Also, for simplicity of presentation, we have fixed the number of iterations for each of the stages (via parameters K_1 and K_2). However, one can easily modify the implementation so that the iteration stops once a desired accuracy has been achieved.

5.4.3 Relation to Other Methods

As illustrated in the pseudo code, the Cluster Newton method only requires one function evaluation per iteration per solution point, so it is computationally efficient. We now compare the method with existing algorithms.

5.4.3.1 Gradient based algorithms

In this Chapter, we are interested in problems where the Jacobian of the forward problem is not explicitly given. Hence, the Jacobian needs to be approximated when using a gradient based algorithm. Finite differences are the standard way of approximating the Jacobian; however, it takes at least $m+1$ function evaluations to estimate the Jacobian through a finite difference approximation. Thus, finding multiple solutions one by one with a gradient based algorithm for ODE coefficient identification will require a large number of function evaluations.

On the other hand, if the Jacobian is explicitly given or the number of unknowns is small, finding each parameter vector individually through a conventional gradient based algorithm can be faster than the Cluster Newton method.

Also, it is worth mentioning that the linear approximation in Stage 1 of the Cluster Newton method becomes arbitrarily close to the tangent hyper-plane of the surface $\boldsymbol{f}$ as the size of the cluster becomes increasingly small. Thus Stage 1 of the Cluster Newton method reduces to Newton's method (with numerical determination of the Jacobian) if one takes the initial size of the cluster and the size of the random perturbation in line 1-2 (η) increasingly small.

We numerically compare the Cluster Newton method with the Levenberg-Marquardt method in the following sections. Also, we briefly compare the Cluster Newton method with Broyden's method in Section 5.6.6.

5.4.3.2 Global optimization algorithms

Gradient based algorithms are local optimization algorithms, as they find local optima rather than global

optima. Global optimization algorithms (like Genetic Algorithms) are necessary in the case where a reasonable initial guess is not available for the problem. For our problem, we are given a reasonable initial guess, and global optimization algorithms are not necessary. Also, among the many solutions of the underdetermined inverse problem, we are only interested in the ones near the initial guess, in the sense explained before. Thus, global algorithms that do not necessarily find solutions near the initial guess can be disadvantageous.

We will briefly discuss numerical comparison of the Cluster Newton method to a Genetic Algorithm in Section 5.6.6.

5.4.3.3 Interval analysis

The only other method that aims to solve an underdetermined ODE coefficient identification problems by finding more than one possible solutions, to the author's best knowledge, is an algorithm that uses interval analysis (see [47]). This algorithm rigorously determines the "infeasible" intervals and determines the intervals where the solutions may reside. If the inverse problem is underdetermined, it finds the intervals that covers all the possible solutions. This algorithm is mathematically rigorous. However, it is computationally very expensive, especially when dealing with a solution space of large dimension. Thus, for our pharmacokinetics problem, which requires identification of 60 parameters, it is not practical to use the interval analysis algorithm.

5.5 Simple Model Problem (Example 1)

Before we attempt to solve the parameter identification problem of the pharmacokinetics model, we will illustrate our algorithm using a simple example that is easy to visualize.

5.5.1 Model Problem Description

Our model problem is as follows: find a set of l points in $\mathbb{R}^2$ near a box $\mathcal{X}^0$, such that

$$f(\boldsymbol{x}) = y^* \,, \tag{5.38}$$

where

$$\begin{aligned} f(\boldsymbol{x}) &= (x_1^2 + x_2^2) + \frac{1}{100}\sin(10000x_1)\cdot\sin(10000x_2)\,, & (5.39)\\ y^* &= 100\,, & (5.40)\\ \mathcal{X}^0 &= \{\boldsymbol{x}\in\mathbb{R}^2 \,:\, \max_{i=1,2} |\,(x_i - 2.5)\,/2.5| < 1\}\,. & (5.41) \end{aligned}$$

The function is a paraboloid perturbed by a oscillatory perturbation with a small amplitude. As depicted in Figure 5.3, the solution manifold of this inverse problem $\mathcal{X}^*$ is approximately a circle of radius 10 centred at the origin in the x_1-x_2 plane. Thus, we aim to find the points on this curve $\mathcal{X}^*$ near the box $\mathcal{X}^0$. The perturbation mimics 'roughness' or 'noise' that can be found in many realistic high-dimensional applications. For example, as we will illustrate in the following section, when the forward problem involves numerical solution of a system of ODEs, a similar kind of 'roughness' can be observed for the function evaluation, caused by numerical error. The initial box $\mathcal{X}^0$ may signify some *a priori* knowledge about where physically relevant solutions are expected. We choose $l = 100$ in the numerical examples below.

5.5.2 Levenberg-Marquardt Method

We first discuss how the well-known Levenberg-Marquardt method (see e.g. [38]) performs when applied to this problem. We create l random points in $\mathcal{X}^0$ and then apply the Levenberg-Marquardt method using each random point as an initial point. We have used the Levenberg-Marquardt implementation in the MATLAB optimization toolbox (version 2010b) for our numerical experiment. We observe that the algorithm terminates with the error "Algorithm appears to be converging to a point that is not a root" for all initial points we tried. As can be seen in Figure 5.4, we fail to find points close to the solution manifold

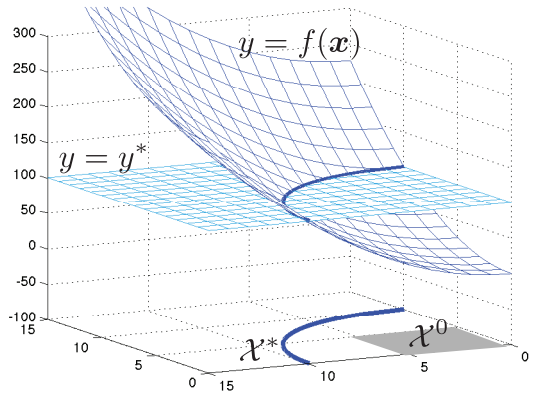


Figure 5.3: Example 1: the function f and the solution manifold $\mathcal{X}^*$ (in the first quadrant).

$\mathcal{X}^*$.

5.5.3 Cluster Newton (CN) Method

We now use the Cluster Newton method to find multiple points on the level curve. Noting that this example is a special case of (5.1) with $m = 2$ and $n = 1$, we directly apply the algorithm presented in Section 5.4 with the following user defined parameters:

$$\text{typical value of } \boldsymbol{x} \qquad \hat{\boldsymbol{x}} = [2.5, 2.5]^{\mathrm{T}}, \tag{5.42}$$

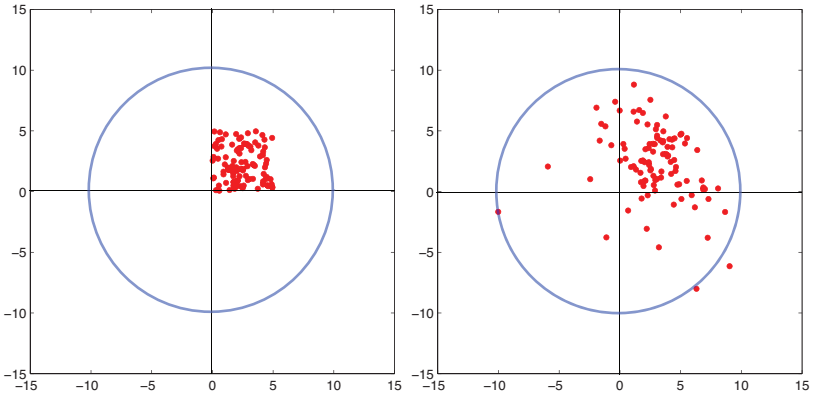
$$\text{relative range of } \boldsymbol{x} \qquad \boldsymbol{v} = [1, 1]^{\mathrm{T}}, \tag{5.43}$$

$$\text{domain of } \boldsymbol{f} \qquad \mathcal{X} = \mathbb{R}^2, \tag{5.44}$$

$$\text{number of Stage 1 iterations} \qquad K_1 + 1 = 6 \quad (\text{so } K_1 = 5), \tag{5.45}$$

$$\text{number of total iterations} \qquad K_2 + 1 = 24 \quad (\text{so } K_2 = 23). \tag{5.46}$$

In order to illustrate the fundamental idea of the algorithm, Stage 1 of the Cluster Newton method is graphically explained in Figure 5.5 using this example. As can be seen in Figure 5.5(b), the Cluster Newton method constructs a linear approximation from points in the cluster and computes the gradient



(a) Initial points in $\mathcal{X}^0$.

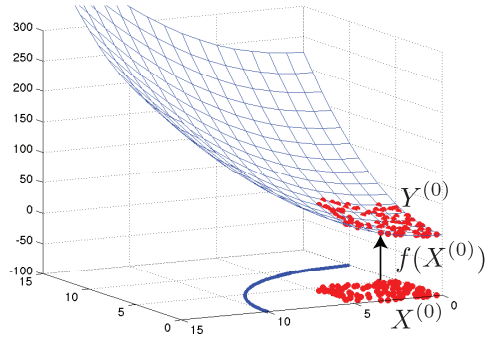
(b) Final Points obtained by the Levenberg-Marquardt method.

Figure 5.4: Example 1: an attempt with the Levenberg-Marquardt method.

using all points in the cluster. Also, it constructs only one linear approximation in each iteration, so only one function evaluation per iteration per point in the cluster is required.

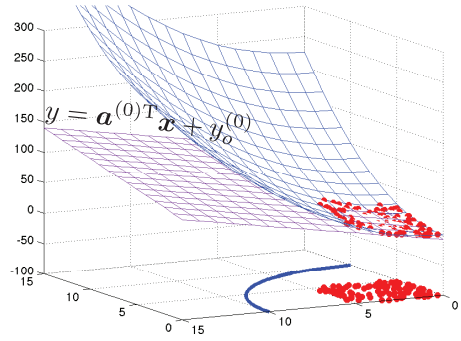
5.5.3.1 Graphical description of the solutions found by the Cluster Newton method

As depicted in Figure 5.6(i), the multiple points found by our algorithm trace part of the solution curve accurately. As can be seen in Figure 5.6(a), we start with an initial cluster of points uniformly distributed in the domain bounded by $x_1 = 0$, $x_1 = 5$ and $x_2 = 0$, $x_2 = 5$. Figures 5.6(a)-5.6(f) show that Stage 1 of the Cluster Newton method moves the points close to the solution manifold $\mathcal{X}^*$, without getting stuck in the local minima created by the oscillatory perturbation with small amplitude. Once the algorithm moves on to Stage 2 when $k = 6$, we see that the points quickly line up with the solution curve as they were already near the solution owing to Stage 1 of our scheme. Recall from Figure 5.4 that the Levenberg-Marquardt method was not able to find solutions. Also note that the size of the final cluster in the direction parallel to the solution manifold is similar to the size of the initial cluster as specified by the typical relative range $\boldsymbol{v}$.

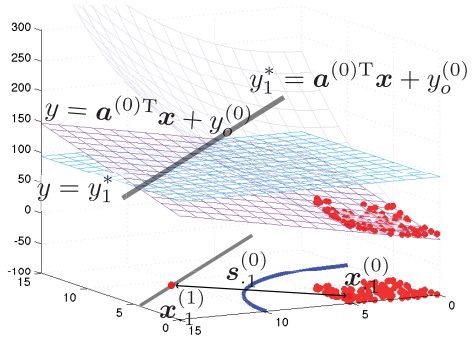


[scale=0.43]

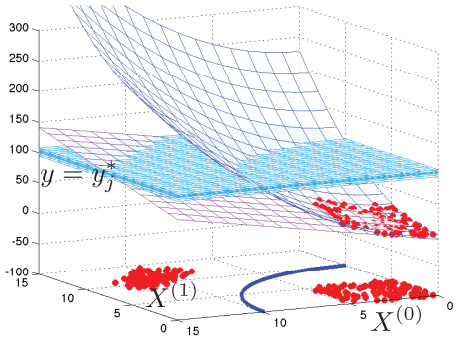
(a) Line 2-1: Evaluate f at each point in $X^{(0)}$ to generate $Y^{(0)}$.



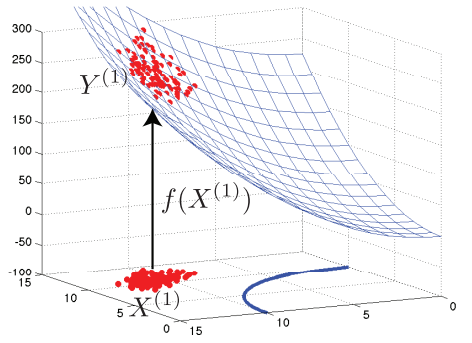
(b) Line 2-2: Construct a linear approximation $y = \boldsymbol{a}^{(0)\mathrm{T}}\boldsymbol{x} + y_o^{(0)}$.



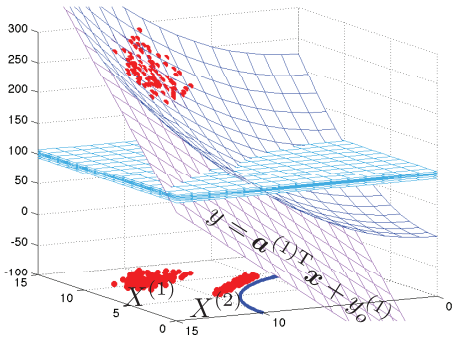
(c) Line 2-3: Find the update vectors $\boldsymbol{s}_{\cdot j}$.



(d) Line 2-4: Find a new cluster of points $X^{(1)}$.

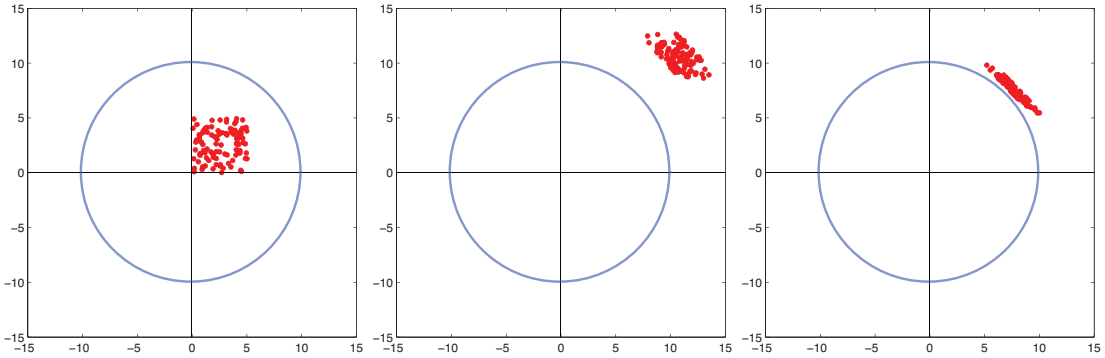


(e) Line 2-1: Evaluate f at each point in $X^{(1)}$ to generate $Y^{(1)}$.

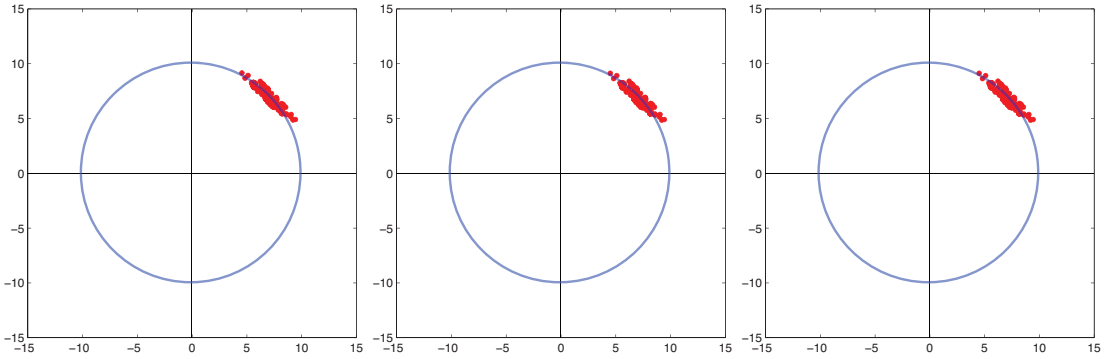


(f) Line 2-2,...,2-4: Obtain $X^{(2)}$

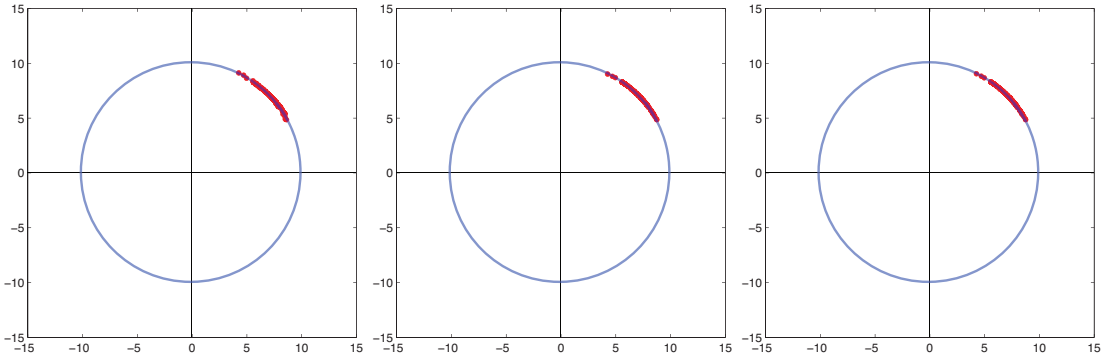
Figure 5.5: Example 1: Graphical description of Stage 1 of the Cluster Newton method.



(a) $X^{(0)}$ (b) $X^{(1)}$ (c) $X^{(2)}$

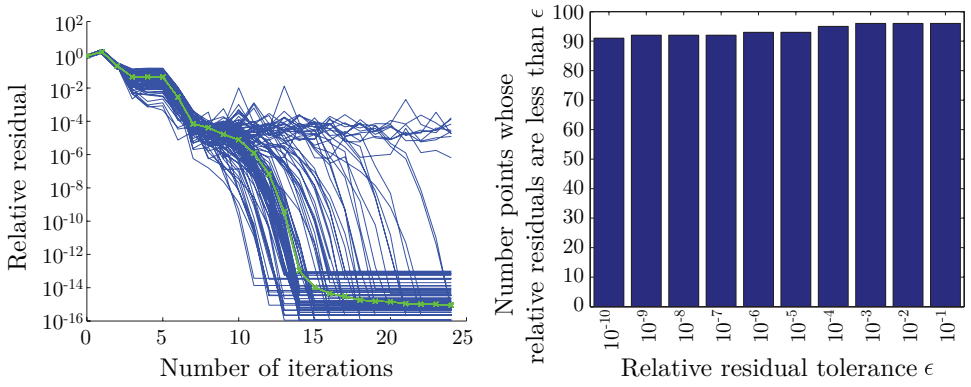


(d) $X^{(3)}$ (e) $X^{(4)}$ (f) $X^{(5)}$



(g) $X^{(6)}$ (h) $X^{(7)}$ (i) $X^{(8)}$

Figure 5.6: Example 1: Plots of the points in $X^{(k)}$. $X^{(0)}$ to $X^{(5)}$ correspond to Stage 1, and $X^{(6)}$ to $X^{(8)}$ to Stage 2.



(a) Speed of relative residual reduction.

(b) Accuracy of the final set $X^{(24)}$ after 2500 total function evaluations.

Figure 5.7: Example 1: Speed and accuracy of the Cluster Newton method.

5.5.3.2 Speed and accuracy of the Cluster Newton method

In Figure 5.7(a), the relative residuals are plotted against the number of iterations. For Example 1, we define the relative residual to be the following:

$$r_j^{(k)} = \frac{|y_j^{(k)} - y^*|}{y^*} = \frac{|y_j^{(k)} - 100|}{100}. \tag{5.47}$$

As can be seen in Figure 5.7(a), the residual is reduced rapidly.

In Figure 5.7(b), the number of points in the final set $X^{(24)}$ whose relative residuals are less than a relative residual tolerance ϵ is plotted. The graph shows that more than 90% of the points in the final set have relative residuals less than 10^{-10}.

Thus, we have shown that the Cluster Newton method can find multiple accurate solutions of an underdetermined inverse problem. The method is highly efficient because of the collective operator fitting in Stage 1, and the collective fitting also acts as a regularization against the small 'roughness' of the function. Hence, the method is robust.

5.6 Pharmacokinetics ODE Coefficient Identification Problem (Example 2)

We now introduce the original problem that led us to construct the Cluster Newton algorithm for simultaneously finding multiple solutions of an underdetermined inverse problem. This inverse problem can be categorized as a coefficient identification problem of a system of ordinary differential equations. We use the system of ODEs modelling the metabolic and transportation processes of the anti-cancer drug CPT-11 and its metabolites developed by Arikuma et al. [2]. By solving this system of ODEs with fixed parameters, we simulate the amount of CPT-11 and its metabolites excreted in urine and bile. These quantities are determined by integrating specific parts of the ODE system solution over a long time (see Section 5.6.1 for a full description of the model). The amount of these chemical compounds in urine and bile can be measured clinically for individual patients (cf. Slatter et al. [44] and de Jong et al. [10]). The set of model parameters (we denote them as $x_1, x_2, \cdots, x_{60}$) represents a biological state of a patient (e.g., amounts of enzymes, blood flow rates) that cannot be measured directly.

By solving this inverse problem, we aim to find multiple possible biological states of a patient that are consistent with the clinical observations.

5.6.1 Forward Problem: Physiologically Based Pharmacokinetics Model

We here briefly describe the physiologically based pharmacokinetics (PBPK) model of the intravenous (i.v.) drip infusion of CPT-11 (also known as Irinotecan). We use the PBPK model developed by Arikuma et al. [2] to model the concentration $u_1(t), \cdots, u_{25}(t)$ of CPT-11 and its metabolites (SN-38, SN-38G, NPC and APC) in each compartment of the body (Blood, Adipose, Gastrointestinal tract (denote as GI), Liver, and everything-else (denoted as NET)). All chemical compounds in the compartments are connected by pathways, $l_1, \cdots, l_{55}$, representing the inflow and outflow of chemical compounds. By noting that a change in concentration is due to the flow of the chemical compounds, we can construct a system of first order ordinary differential equations (ODEs) for the concentrations $u_i(t)$ as a function of time t.

Table 5.2: Kinetic parameters related to blood flow transportation of the drugs, where PBR is the protein binding ratio (typical values).

Organ		Tissue-Blood Distribution Ratio Adipose	GI	Liver	NET	PBR
Compound	Label	$i=1$	$i=6$	$i=11$	$i=16$	$i=21$
CPT-11	$\hat{x}_i$	10.00	1.00	1.00	3.00	0.37
SN-38	$\hat{x}_{i+1}$	2.00	1.00	1.00	0.70	0.05
SN-38G	$\hat{x}_{i+2}$	2.80	1.00	1.00	0.08	1.00
NPC	$\hat{x}_{i+3}$	6.00	1.00	1.00	2.00	0.37
APC	$\hat{x}_{i+4}$	1.50	1.00	1.00	0.06	0.37

As the information presented in Arikuma et al. [2] is not sufficient to construct the ODE model, for the sake of reproducibility of our results, we here present enough details to construct a system of ODEs that models the drug kinetics of CPT-11. We shall refer the reader to [2] for biological justification and verification of this PBPK model.

5.6.1.1 Mathematical model of the pathways

Arikuma et al. have modelled the pathways (labeled as $l_1, \cdots, l_{55}$ in Figure 5.8) of the PBPK model as listed below. There are four kinds of pathways: intravenous (i.v.) drip pathways, blood flow pathways, metabolic pathways, and excretion pathways. Each quantitatively describes the flow rate of the drug, with units of $[nmol/min]$. There are 60 parameters associated with these pathways and these parameters are labelled $x_1, \cdots, x_{60}$ and are listed in Tables 5.2-5.6 with typical values. The goal of solving this inverse problem will be to determine better estimates of these parameters than the typical values listed in the tables, based on clinically observed data.

The i.v. drip feed into blood is expressed by the following step function:

$$l_0 = \begin{cases} x_{59}/x_{60} & \text{for } 0 < t < x_{60} \\ 0 & \text{otherwise} \end{cases}. \tag{5.48}$$

The drug transportation by the blood flow pathways is modelled using the fact that the amount of drug

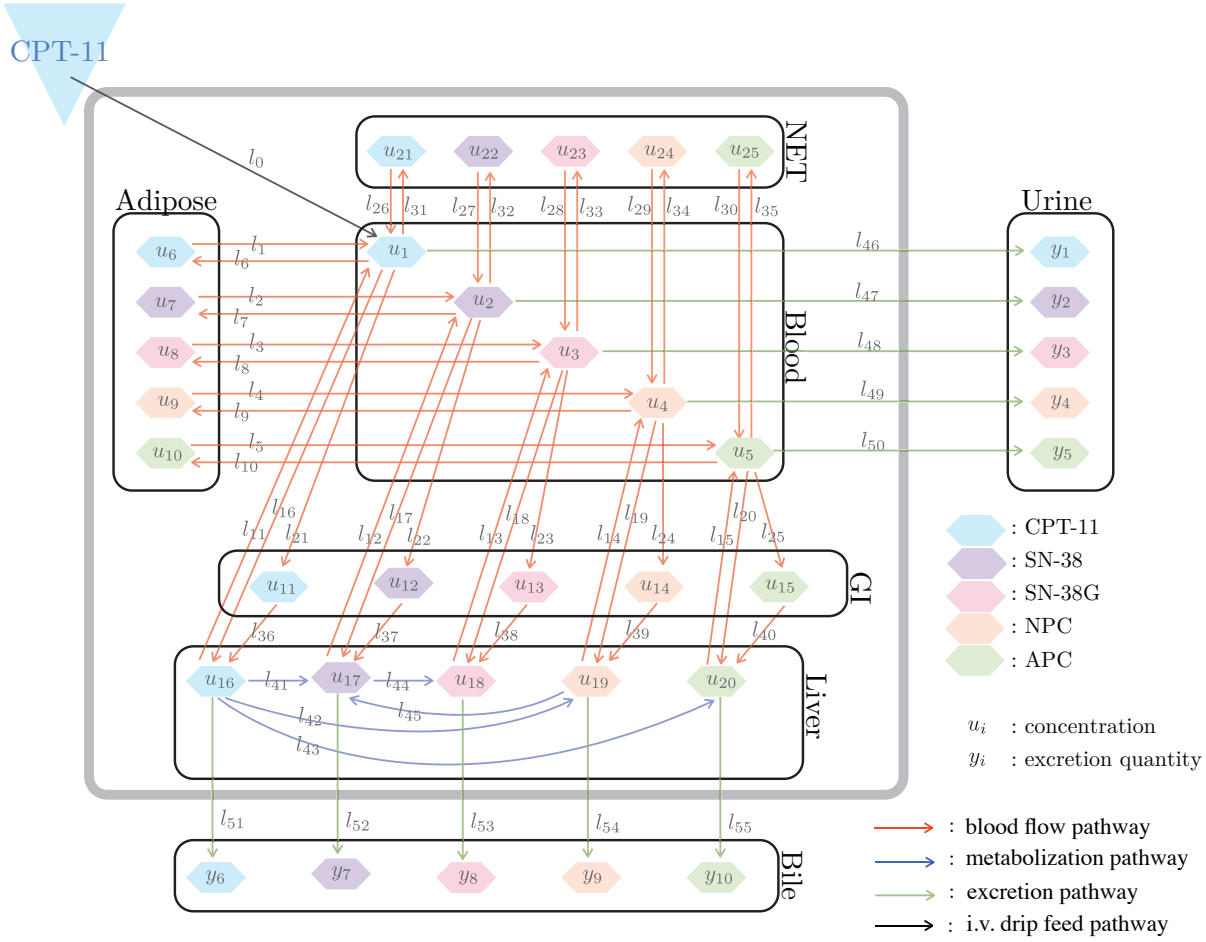


Figure 5.8: Schematic diagram of the PBPK model.

Table 5.3: Kinetic parameters related to the clearances (typical values).

Unit		Urinary Clearance $[mL/min/kg]$	Biliary Clearance $[mL/min/kg]$
Compound	Label	$i = 26$	$i = 31$
CPT-11	$\hat{x}_i$	6.15	10.6
SN-38	$\hat{x}_{i+1}$	9.91	103
SN-38G	$\hat{x}_{i+2}$	1.44	2.03
NPC	$\hat{x}_{i+3}$	1.49	14.5
APC	$\hat{x}_{i+4}$	1.47	5.45

Table 5.4: Kinetic parameters related to the Michaelis-Menten kinetics equation, where the units of K_m, V_{max} and α are $[nmol/mL]$, $[nmol/min/mg_{protein}]$ and $[mg_{protein}/g_{tissue}]$, respectively (typical values).

				K_m	V_{max}	α
Enzyme	Substrate	Product	Label	$i = 36$	$i = 41$	$i = 46$
Carboxylesterase	CPT-11	SN-38	$\hat{x}_i$	2.30	0.00211	128
Carboxylesterase	NPC	SN-38	$\hat{x}_{i+1}$	2.30	0.00211	128
CYP3A4	CPT-11	APC	$\hat{x}_{i+2}$	18.4	0.0260	73.3
CYP3A4	CPT-11	NPC	$\hat{x}_{i+3}$	48.2	0.0741	11.7
UGT1A1	SN-38	SN-38G	$\hat{x}_{i+4}$	3.80	0.0508	750

Table 5.5: Physiological parameters obtained from Willmann et al. [50].

Unit	Blood flow rate $[mL/min/kg]$		Volume $[mL/kg]$	
Organ	Label	Typical Value	Label	Typical Value
Blood	-	-	$\hat{x}_{55}$	51.0
Adipose	$\hat{x}_{51}$	4.45		$1000 - (x_{55} + x_{56} + x_{57} + x_{58})$
GI	$\hat{x}_{52}$	13.4	$\hat{x}_{56}$	32.1
Liver	$\hat{x}_{53}$	5.79	$\hat{x}_{57}$	32.3
NET	$\hat{x}_{54}$	37.4	$\hat{x}_{58}$	681

Table 5.6: Drug administration parameters.

Parameters	Unit	Label	Typical values
Dosing amount	$[nmol/kg]$	$\hat{x}_{59}$	4860
Drip feed duration	$[min]$	$\hat{x}_{60}$	90

flowing out of a compartment is proportional to the concentration of the drug in the compartment, with the parameters x_i used to model the constants of proportionality. This can be written as follows:

$$l_i = \begin{cases} x_{51}/x_i \cdot u_{i+5}(t) & \text{for } i = 1, \cdots, 5\,, \\ x_{51} \cdot u_{i-5}(t) & \text{for } i = 6, \cdots, 10\,, \\ (x_{52} + x_{53})/x_i \cdot u_{i+5}(t) & \text{for } i = 11, \cdots, 15\,, \\ x_{53} \cdot u_{i-15}(t) & \text{for } i = 16, \cdots, 20\,, \\ x_{52} \cdot u_{i-20}(t) & \text{for } i = 21, \cdots, 25\,, \\ x_{54}/x_{i-10} \cdot u_{i-5}(t) & \text{for } i = 26, \cdots, 30\,, \\ x_{54} \cdot u_{i-30}(t) & \text{for } i = 31, \cdots, 35\,, \\ x_{52}/x_{i-30} \cdot u_{i-25}(t) & \text{for } i = 36, \cdots, 40\,, \end{cases} \tag{5.49}$$

where $u_1(t), \cdots, u_{25}(t)$ are the concentrations of CPT-11 and its metabolites in the compartments, as labelled in Figure 5.8. The drug transformations by the metabolic pathways are modelled using the Michaelis-Menten kinetics equation as follows:

$$l_i = \begin{cases} (x_{41} \cdot x_{46} \cdot x_{57})/(\frac{x_{36} \cdot x_{11}}{x_{21} \cdot u_{16}(t)} + 1) & \text{for } i = 41\,, \\ (x_{44} \cdot x_{49} \cdot x_{57})/(\frac{x_{39} \cdot x_{11}}{x_{21} \cdot u_{16}(t)} + 1) & \text{for } i = 42\,, \\ (x_{43} \cdot x_{48} \cdot x_{57})/(\frac{x_{38} \cdot x_{11}}{x_{21} \cdot u_{16}(t)} + 1) & \text{for } i = 43\,, \\ (x_{45} \cdot x_{50} \cdot x_{57})/(\frac{x_{40} \cdot x_{12}}{x_{22} \cdot u_{17}(t)} + 1) & \text{for } i = 44\,, \\ (x_{42} \cdot x_{47} \cdot x_{57})/(\frac{x_{37} \cdot x_{14}}{x_{24} \cdot u_{19}(t)} + 1) & \text{for } i = 45\,. \end{cases} \tag{5.50}$$

The drug elimination by the excretion pathway is modelled using the fact that the amount of drug flowing out of a compartment is proportional to the concentration of the drug in the compartment. This can be written as

$$l_i = \begin{cases} x_{i-20} \cdot x_{i-25} \cdot u_{i-45}(t) & \text{for } i = 46, \cdots, 50\,, \\ (x_{i-20} \cdot x_{i-30})/x_{i-40} \cdot u_{i-35}(t) & \text{for } i = 51, \cdots, 55\,. \end{cases} \tag{5.51}$$

These pathways are assembled together to form the system of ODEs, as we discuss next.

5.6.1.2 Mathematical model of the concentrations

Let $u_1(t), \cdots, u_{25}(t)$ be the concentrations of CPT-11 and its metabolites in the compartments, as labelled in Figure 5.8, with units of $[\mu mol/L]$. As a change of concentration, $\frac{\mathrm{d}u_i}{\mathrm{d}t}$, is due to the inflow/outflow of the drug via pathways, we can construct the following system of ODEs to model the concentrations $u_i(t)$:

$$\begin{aligned} \frac{d}{dt}u_{5(j-1)+k} &= \left(\sum_{i \in \mathcal{N}_{5(j-1)+k}} l_i(x_1, \cdots, x_{60}, u_1, \cdots, u_{25}; t) \right) / V_j \\ &\quad - \left(\sum_{i \in \mathcal{M}_{5(j-1)+k}} l_i(x_{1,2,\cdots,60}, u_{1,2,\cdots,25}; t) \right) / V_j \,, \end{aligned} \tag{5.52}$$

with the initial conditions:

$$u_{5(j-1)+k}(t=0) = 0\,, \tag{5.53}$$

where

$$\begin{aligned} j = 1, \cdots, 5 &: \text{compartments} \\ &\quad \text{(Blood, Adipose, GI, Liver, NET)}, \\ k = 1, \cdots, 5 &: \text{drug and its metabolites} \\ &\quad \text{(CPT-11, SN-38, SN-38G, NPC, APC)}, \end{aligned}$$

and

t : time in minutes,

$\mathcal{N}_i$: index set of the inflow pathways of u_i as listed in Equations (5.59),

$\mathcal{M}_i$: index set of the outflow pathways of u_i as listed in Equations (5.60),

l_i : flow rate of the drug in each pathway as listed in Equations (5.48)-(5.51),

x_i : model parameter,

V_j : volume of the compartment j per weight of the patient $[mL/kg]$.

The compartment volumes V_j can be written as

$$V_1 = x_{55}\,, \tag{5.54}$$

$$V_2 = 1000 - (x_{55} + x_{56} + x_{57} + x_{58})\,, \tag{5.55}$$

$$V_3 = x_{56}\,, \tag{5.56}$$

$$V_4 = x_{57}\,, \tag{5.57}$$

$$V_5 = x_{58}\,. \tag{5.58}$$

Equation (5.55) is derived based on the formulation used in Willmann et al. [50] and assuming that the volume mass ratio of the human body is $1000mL/kg$. The following index sets indicate the inflow pathways

of u_i as in Figure 5.8:

$$\mathcal{N}_i = \begin{cases} \{0,\, 1,\, 11,\, 26\} & \text{for } i = 1\,, \\ \{i,\, i+10,\, i+25\} & \text{for } i = 2, \cdots, 5\,, \\ \{i\} & \text{for } i = 6, \cdots, 10\,, \\ \{i+10\} & \text{for } i = 11, \cdots, 15\,, \\ \{16,\, 36\} & \text{for } i = 16\,, \\ \{17,\, 37,\, 41,\, 45\} & \text{for } i = 17\,, \\ \{18,\, 38,\, 44\} & \text{for } i = 18\,, \\ \{19,\, 39,\, 42\} & \text{for } i = 19\,, \\ \{20,\, 40,\, 43\} & \text{for } i = 20\,, \\ \{i+10\} & \text{for } i = 21, \cdots, 25\,. \end{cases} \tag{5.59}$$

The following index sets indicate the outflow pathways of u_i as in Figure 5.8:

$$\mathcal{M}_i = \begin{cases} \{i+5,\, i+15,\, i+20,\, i+30,\, i+45\} & \text{for } i = 1, \cdots, 5\,, \\ \{i-5\} & \text{for } i = 6, \cdots, 10\,, \\ \{i+25\} & \text{for } i = 11, \cdots, 15\,, \\ \{11,\, 41,\, 42,\, 43,\, 51\} & \text{for } i = 16\,, \\ \{12,\, 44,\, 52\} & \text{for } i = 17\,, \\ \{13,\, 53\} & \text{for } i = 18\,, \\ \{14,\, 45,\, 54\} & \text{for } i = 19\,, \\ \{15,\, 55\} & \text{for } i = 20\,, \\ \{i+5\} & \text{for } i = 21, \cdots, 25\,. \end{cases} \tag{5.60}$$

For example, the ODE model for the concentration of SN-38G ($k = 3$) in the Liver ($j = 4$) can be

Table 5.7: Computational costs of numerically solving the forward problem for Example 2. The computational time was measured on a server machine with Intel Xeon X7350 3GHz processor.

Absolute/Relative tolerance	Computational time (sec)
10^{-2}	0.097
10^{-3}	0.12
10^{-4}	0.19
10^{-5}	0.25
10^{-6}	0.37
10^{-7}	0.42
10^{-8}	0.56
10^{-9}	0.76
10^{-10}	1.13

written as follows:

$$\begin{aligned}
\frac{d}{dt}u_{18} &= \left(\sum_{i\in\{18,38,44\}} l_i - \sum_{i\in\{13,53\}} l_i\right)/V_4 \\
&= \left(x_{53}\cdot u_3(t) + \frac{x_{52}}{x_8}\cdot u_{13}(t) + \frac{x_{45}\cdot x_{50}\cdot x_{57}}{\frac{x_{40}\cdot x_{12}}{x_{22}\cdot u_{17}(t)}+1}\right)/x_{57} \\
&\quad - \left(\frac{x_{52}+x_{53}}{x_{13}}\cdot u_{18}(t) + \frac{x_{33}\cdot x_{23}}{x_{13}}\cdot u_{18}(t)\right)/x_{57}\,. \qquad (5.61)
\end{aligned}$$

By solving this system of ODEs, we obtain a vector valued function $\boldsymbol{u}(x_1,\cdots,x_{60};t)$ depending on the variable t and parameters $x_1,\cdots,x_{60}$. Using the typical parameters $\hat{\boldsymbol{x}}$, the concentration of, for example, SN-38G in the Liver can be simulated by the above system of ODEs as depicted in Figure 5.9(a).

Note:

- We solve this system of ODEs using the MATLAB 2010b stiff ODE solver ODE15s [42]. The computational costs of numerically solving this system of ODEs with various integration tolerances are tabulated in Table 5.7.
- It can be shown easily that this ODE model conserves the amount of drug. Assuming all the parameters x_i and V_2 are positive real numbers, we further can show that $u_i(t) \geq 0$ for all i which leads to the proof of the existence of $u_i(t)$ for all $0 < t < \infty$.

- We will only consider the physiologically relevant cases where all parameters x_i and the compartment volume V_2 are positive.

- This system of ODEs cannot easily be solved numerically. We first note that one of the terms on the right hand side of the ODEs corresponding to the i.v. drip feed is discontinuous at $t = x_{60}$ (cf. Equation (5.48)). This discontinuous term causes abrupt changes of the solution u_i at time $t = x_{60}$. Hence, this system of ODEs is considered a stiff system numerically, and it requires a stiff ODE solver. Also, it is noteworthy that the ODE system is well-posed only if $u_i(t) \geq 0$ for all i and $t \geq 0$, while $\lim_{t\to\infty} u_i(t) = 0$. Thus we have to make sure that the ODE solver chooses small enough time steps especially when $u_i(t)$ is small, so that the numerical solution does not go below 0. This stiffness and positivity requirement causes the high computational cost of solving the forward problem.

5.6.1.3 Numerical simulation of the excretion profile using the PBPK model

We now describe how to compute the excretion profile using the PBPK model. Let $y_1, \cdots, y_{10}$ be the total excretion amount of CPT-11 and its metabolites through urine and bile in units of $[nmol/kg]$ (see Figure 5.8). More specifically, the variables $y_1, \cdots, y_5$ are the total excretion amount of CPT-11, SN-38, SN-38G, NPC and APC in urine, respectively, and the variables $y_6, \cdots, y_{10}$ are the total excretion amount of CPT-11, SN-38, SN-38G, NPC and APC in bile, respectively. The set of total excretion amounts of CPT-11 and its four metabolites through urine and bile (a set of ten numbers) is an *excretion profile*. The total excretion amount of each chemical compound can be computed as follows:

$$y_i = \begin{cases} \int_0^\infty x_{i+25} \cdot x_{i+20} \cdot u_i(t)\,\mathrm{d}t & \text{for } i = 1, \cdots, 5 \\ \int_0^\infty (x_{i+25} \cdot x_{i+15})/x_{i+5} \cdot u_{i+10}(t)\,\mathrm{d}t & \text{for } i = 6, \cdots, 10 \end{cases}, \tag{5.62}$$

$$= \int_0^\infty l_{i+45}\,\mathrm{d}t \qquad \text{for } i = 1, \cdots, 10\,, \tag{5.63}$$

$$= \lim_{t\to\infty} \int_0^t l_{i+45}\,\mathrm{d}\tau \qquad \text{for } i = 1, \cdots, 10\,, \tag{5.64}$$

$$\approx \int_0^{\mathrm{T}} l_{i+45}\,\mathrm{d}t \qquad \text{for } i = 1, \cdots, 10\,, \tag{5.65}$$

where $u_i(t)$ is the concentration of CPT-11 and its metabolites obtained by solving the system of ODEs (5.52), and T is a sufficiently large constant. In our implementation, we have chosen $T = 10^5$ to ensure that $u_i(T) \approx 0$ for all $i = 1, \cdots, 5, 16, \cdots, 20$. It is a direct consequence of the conservation property of the total amount of the drug in the system and $u_i > 0$ that $\int_0^t l_{i+45}\, \mathrm{d}\tau$ is bounded and monotonically increasing. Thus, by the monotone convergence theorem, there exist y_i which satisfy Equations (5.62). Also by the conservation property we have $x_{59} = \sum_{i=1}^{10} y_i$.

It is possible to compute the y_i at the same time as solving the PBPK model. That is to say, this numerical integration can be included in the system of ODEs by adding 10 unknown functions and re-writing Equations (5.65) in the following form:

$$\frac{d}{dt} u_{i+25} \quad = \quad l_{i+45} \qquad \text{for } i = 1, \cdots, 10\,, \tag{5.66}$$

which leads to

$$y_i \quad \approx \quad u_{i+25}(T) \qquad \text{for } i = 1, \cdots, 10\,. \tag{5.67}$$

As an example, the graphs of $u_{26}(t)$ and y_1 are depicted in Figure 5.9(b) using the typical values of the parameters $\hat{\boldsymbol{x}}$.

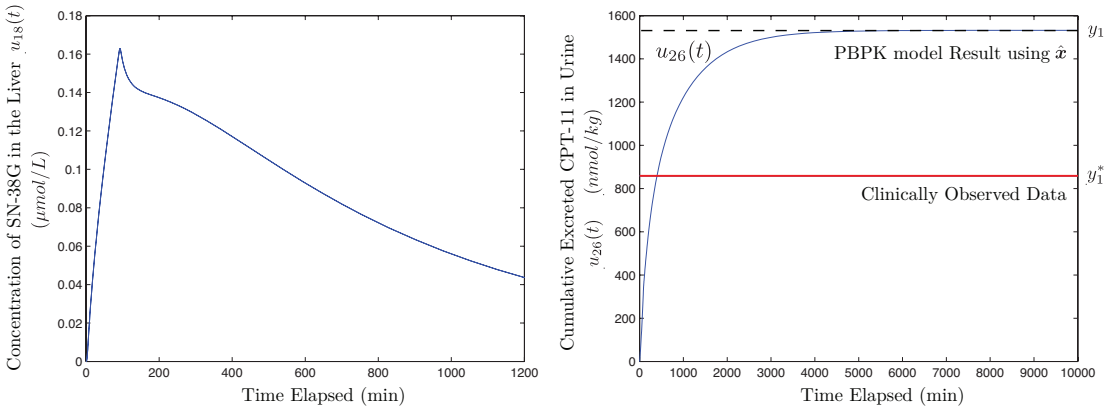
Recalling that $u_1, \cdots, u_{25}$ depend on the parameters $x_1, \cdots, x_{60}$, now we have obtained a function that maps the parameters to the excretion profile. For simplicity we denote this vector function by $\boldsymbol{f}$, i.e.,

$$[y_1, y_2, \cdots, y_{10}]^{\mathrm{T}} \quad = \quad [u_{26}(x_1, \cdots, x_{60}; T), \cdots, u_{35}(x_1, \cdots, x_{60}; T)]^{\mathrm{T}}\,, \tag{5.68}$$

$$\quad = \quad [f_1(x_1, \cdots, x_{60}), \cdots, f_{10}(x_1, \cdots, x_{60})]^{\mathrm{T}}\,, \tag{5.69}$$

or

$$\boldsymbol{y} \quad = \quad \boldsymbol{f}(\boldsymbol{x})\,, \tag{5.70}$$



(a) Concentration of SN-38G in the Liver

(b) Excretion amount of CPT-11 in Urine

Figure 5.9: Example solutions from a PBPK model simulation using the typical value $\hat{\boldsymbol{x}}$

where

$$\boldsymbol{f} \quad : \quad \mathbb{R}^{60} \to \mathbb{R}^{10}\,, \tag{5.71}$$

$$\boldsymbol{x} \quad = \quad [x_1, \cdots, x_{60}]^{\mathrm{T}} \ : \text{a vector that represents the parameters}\,, \tag{5.72}$$

$$\boldsymbol{y} \quad = \quad [y_1, \cdots, y_{10}]^{\mathrm{T}} \ : \text{a vector that represents the excretion profile}\,. \tag{5.73}$$

5.6.2 Inverse Problem: Parameter Identification of a Pharmacokinetics Model

Using Arikuma et al.'s physiologically based pharmacokinetics (PBPK) model in Section 5.6.1, we have a function that maps the parameters of the PBPK model to the quantities of CPT-11 and its metabolites excreted in urine and bile (faeces). We shall refer to a set of quantities of excreted CPT-11 and its metabolites as *an excretion profile*. In order to create the mathematical model, we wish to identify these model parameters using clinically measured excretion profiles. As there are 60 model parameters and 10 clinically measurable excretion profile quantities, we have an underdetermined inverse problem.

We note that the model parameters appear as coefficients of the system of ODEs. Also note that solving the forward problem involves solving a system of ODEs numerically and is computationally expensive.

Table 5.8: The amount of drug and its metabolite in excreta in units [nmol/kg] calculated from published data of Slatter et al. [44].

		Patient 1
CPT-11 in Urine	y_1^*	859.0
SN-38 in Urine	y_2^*	35.5
SN-38G in Urine	y_3^*	473.9
NPC in Urine	y_4^*	3.55
APC in Urine	y_5^*	305.0
CPT-11 in Bile + Faeces	y_6^*	975.4
SN-38 in Bile + Faeces	y_7^*	127.1
SN-38G in Bile + Faeces	y_8^*	105.4
NPC in Bile + Faeces	y_9^*	24.5
APC in Bile + Faeces	y_{10}^*	219.4
total dosage	$\sum_{i=1}^{10} y_i^*$	3946

5.6.2.1 Clinically measured excretion profile by Slatter et al. [44]

We use the clinical data published by Slatter et al. [44]. Based on their data, we calculate the excretion profile of two patients as shown in Table 5.8. The published data by de Jong et al. [10] lacks the information of the patient's weight so that we could not calculate the excretion profile for this data. Without loss of generality, for our numerical experiments, we use the clinically measured data of Patient 1 as the target for the output of the pharmacokinetics model: the goal is to determine multiple sets of parameter values that are consistent with the excretion profile of Patient 1.

5.6.2.2 Typical value and relative range of the model parameters $x_1, x_2, \cdots, x_{60}$

We use the typical values of the model parameters derived through literature search and educated estimates by Arikuma et al. [2] and denote them as $\hat{x}_1, \cdots, \hat{x}_{60}$. These values are re-tabulated in Tables 5.2-5.6 in Section 5.6.1.

We choose the variability to be $\pm 50\%$ for the kinetic parameters (Tables 5.2-5.4), $\pm 30\%$ for the physiological parameters (Table 5.5) and $\pm 5\%$ for the drug administration parameters (Table 5.6), i.e., the

relative ranges are

$$v_i \;=\; \begin{cases} 0.5 & \text{for } i = 1, 2, \cdots, 50 \\ 0.3 & \text{for } i = 51, 52, \cdots, 58 \\ 0.05 & \text{for } i = 59, 60. \end{cases} \tag{5.74}$$

The variability of the kinetic parameters was chosen guided by the fact that the inter-subject variability of these values is usually less than $\pm 50\%$ as shown in [16, 21, 20, 45]. The variability of the physiological parameters was motivated by [50]. The variability of the drug administration parameters is chosen to be small since it is only influenced by the experimental precision of the drug administration procedure.

5.6.2.3 Statement of the inverse problem

We now state the PBPK model parameter identification problem as follows: find a set of l points in $\mathcal{X} \subset \mathbb{R}^{60}$ near a box $\mathcal{X}^0$, such that

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{y}^* \,, \tag{5.75}$$

where

$$\begin{aligned} \boldsymbol{f} \;&:\; \mathcal{X} \subset \mathbb{R}^{60} \to \mathbb{R}^{10} \qquad \text{a function that maps the model parameters} \\ &\quad\; \text{to the excretion profile, as defined in Section 5.6.1 ,} && (5.76) \\ \boldsymbol{y}^* \;&:\; \text{clinically measured data from patient 1 as in Table 5.8 ,} && (5.77) \\ \mathcal{X} \;&=\; \{\boldsymbol{x} \in \mathbb{R}^{60} : x_i > 0 \text{ and } \sum_{i=55}^{58} x_i < 1000\} \qquad \text{(see Section 5.6.1) ,} && (5.78) \\ \mathcal{X}^0 \;&=\; \{\boldsymbol{x} \in \mathbb{R}^{60} : \max_{i=1,2,\cdots,60} |\left(x_i - \hat{x}_i\right)/(\hat{x}_i v_i)| < 1\} \,. && (5.79) \end{aligned}$$

So $m = 60$ and $n = 10$ for this problem. We note that evaluating the function value $\boldsymbol{f}(\boldsymbol{x})$ involves numerically solving the system of ODEs. Hence, the function cannot be evaluated exactly. The computational cost of computing this forward problem with various discretization accuracy is tabulated in Table 5.7.

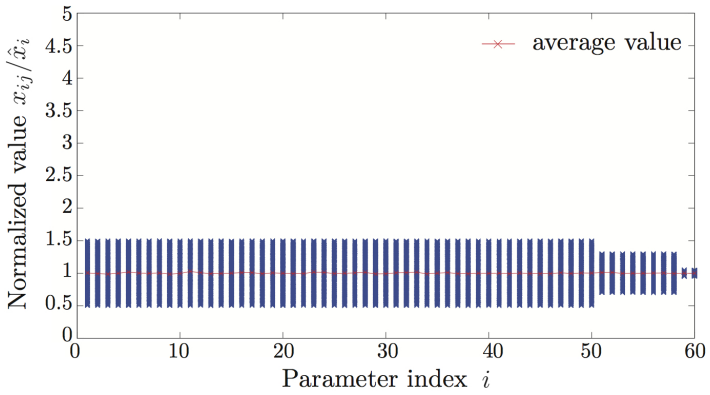


Figure 5.10: Example 2: Initial set of points in $\mathcal{X}^0$.

5.6.3 Levenberg-Marquardt Method

We first create random points in $\mathcal{X}^0$ and then apply the Levenberg-Marquardt method using each point as an initial guess. We do this computation in parallel as each run of the Levenberg-Marquardt method is independent of the others. Due to a limitation of the MATLAB Parallel Computing Toolbox, we utilize at most 8 cores. A visual representation of the 1,000 randomly chosen points in $\mathcal{X}^0$ is given in Figure 5.10. Each red $\times$ indicates the average of the normalized parameter over all points. We have used the Levenberg-Marquardt implementation in the MATLAB optimization toolbox (version 2010b) to find the root of the following function in our numerical experiment:

$$\tilde{\boldsymbol{f}}(\tilde{\boldsymbol{x}}) \quad = \quad \begin{cases} (diag(\boldsymbol{y}^*))^{-1}\boldsymbol{f}(diag(\hat{\boldsymbol{x}})\tilde{\boldsymbol{x}}) - 1 & \text{if } (diag(\hat{\boldsymbol{x}})\tilde{\boldsymbol{x}}) \in \mathcal{X} \\ 10^5\,[1,1,1,1,1,1,1,1,1,1]^{\mathrm{T}} & \text{otherwise} \end{cases}, \tag{5.80}$$

where $\tilde{\boldsymbol{x}}$ is a normalized model parameter vector. Note that we have used the normalization by $diag(\boldsymbol{y}^*)$ because it improved convergence. The way adopted in (5.80) to force solutions to lie in the domain of the function $\boldsymbol{f}$ turns out to work satisfactorily.

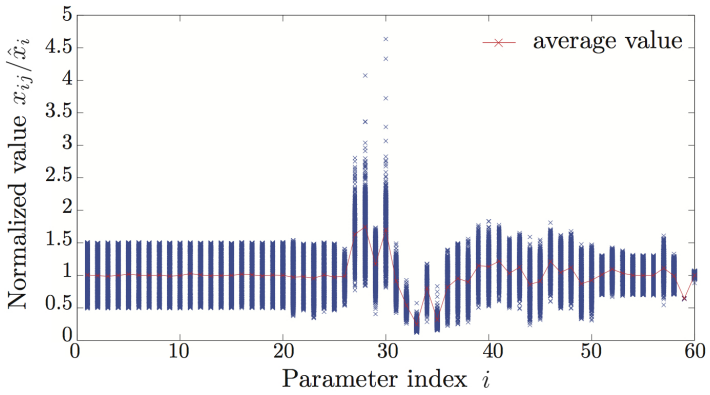


Figure 5.11: Final set of points found by the Levenberg-Marquardt method.

5.6.3.1 Visual representation of the solution found by the Levenberg-Marquardt method

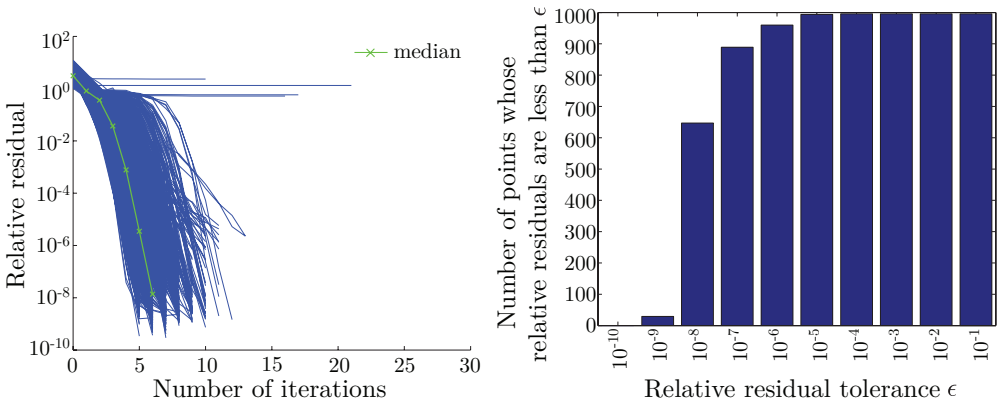
Figure 5.11 visually represents the final set of points found by the Levenberg-Marquardt method after 469,439 function evaluations using an error tolerance of 10^{-9} when solving the system of ODEs (i.e., we choose $\delta_{ODE} = 10^{-9}$ and set the error tolerance to be RelTol= δ_{ODE}, AbsTol= δ_{ODE} for MATLAB ODE solver ODE15s). Note that this small tolerance is required for the Levenberg-Marquardt method to converge, while the Cluster Newton method converges with a significantly less restrictive tolerance (see below).

5.6.3.2 Speed and accuracy obtained with the Levenberg-Marquardt method

In Figure 5.12(a), the relative residual was plotted against the number of iterations. For Example 2, we define the relative residual of parameter vector $\boldsymbol{x}$ to be:

$$r_j^{(k)}(\boldsymbol{x}) \quad = \quad \max_{i=1,2,\cdots,10} \left| \frac{y_{ij}^{(k)} - y_i^*}{y_i^*} \right| , \tag{5.81}$$

where $\boldsymbol{y}_{\cdot j}^{(k)} = \boldsymbol{f}(\boldsymbol{x}_{\cdot j}^{(k)})$ with $\delta_{ODE} = 10^{-11}$. We use $\delta_{ODE} = 10^{-11}$ to make sure we obtain a residual that is close to the true residual that would be determined when the function $\boldsymbol{f}$ is evaluated exactly. As can be seen in Figure 5.12(a), it takes on average seven iterations to find solutions accurate up to



(a) Relative residual reduction.

(b) Accuracy of the final set of points after 469,439 function evaluations with the error tolerance of $\delta_{ODE} = 10^{-9}$.

Figure 5.12: Speed and accuracy of the Levenberg-Marquardt method applied to Example 2.

the accuracy of the function evaluation (δ_{ODE}). We note that since the Jacobian of the function is not explicitly given, the MATLAB implementation of the Levenberg-Marquardt method estimates the Jacobian by finite differences. Hence, at each iteration, the function is evaluated at least 61 times. In each function evaluation, we solve the system of ODEs to high accuracy. Thus, this method can be computationally very expensive (e.g., to obtain the solution presented in Figure 5.12(b), about 8 hours of computation is used on a server machine with two quad-core Intel Xeon X7350 3GHz processors).

In Figure 5.12(b), the number of points in the final set obtained by the Levenberg-Marquardt method (after 469,439 function evaluations) whose relative residual is less than the relative residual tolerance ϵ is plotted. As can be seen in Figure 5.12(b), about 95% of the points achieve a relative residual less than 10^{-6} and about 65% of the points achieve a relative residual less than 10^{-8}.

As already stated in Section 5.1, we can obtain 1,000 different model parameter vectors of interest through multiple application of the Levenberg-Marquardt method. However, this requires accurate function evaluation with tolerance $\delta_{ODE} = 10^{-9}$ and a large number of function evaluations per iteration. Thus, multiple application of the Levenberg-Marquardt method to obtain multiple model parameter vectors is computationally very expensive.

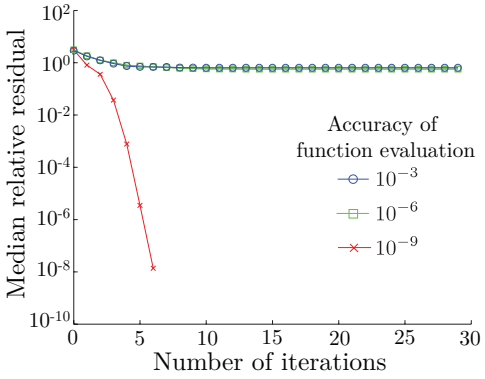


Figure 5.13: Influence of the accuracy of the function evaluation (δ_{ODE}) on the Levenberg-Marquardt method.

5.6.3.3 Influence of the accuracy of the function evaluation on the convergence of the Levenberg-Marquardt method

In Figure 5.13, we have plotted the median relative residual against the number of iterations with different accuracy of the function evaluation (δ_{ODE}). As can be seen from Figure 5.13, the function needs to be evaluated accurately in order for the Levenberg-Marquardt method to find solutions with small relative residual. These numerical experiments show that $\delta_{ODE} = 10^{-9}$ or smaller is required for the MATLAB implementation of the Levenberg-Marquardt method to stably find the solution.

5.6.4 Cluster Newton Method

We now show that our Cluster Newton method is a much more computationally efficient way to find multiple solutions of the underdetermined inverse problem than multiple applications of the Levenberg-Marquardt method. The computational efficiency of the Cluster Newton method is due to the following characteristics: robustness against 'roughness' (numerical error) in the function evaluation, and significantly smaller number of required function evaluations per iteration. These characteristics follow from the collective way in which the points are updated and linear approximations are computed in Stage 1 of the Cluster Newton method. We directly apply the algorithm presented in Section 5.4 with $m = 60$ and

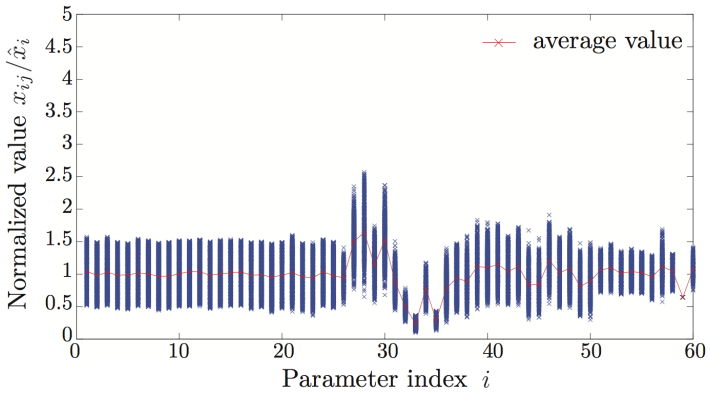


Figure 5.14: Example 2: Set of points $X^{(30)}$ found by the Cluster Newton method.

$n = 10$ and with the following user defined parameters:

$$\text{number of points in the cluster} \qquad l \;=\; 1{,}000\,, \tag{5.82}$$

$$\text{number of Stage 1 iterations} \qquad K_1 + 1 \;=\; 11 \qquad (\text{so } K_1 = 10)\,, \tag{5.83}$$

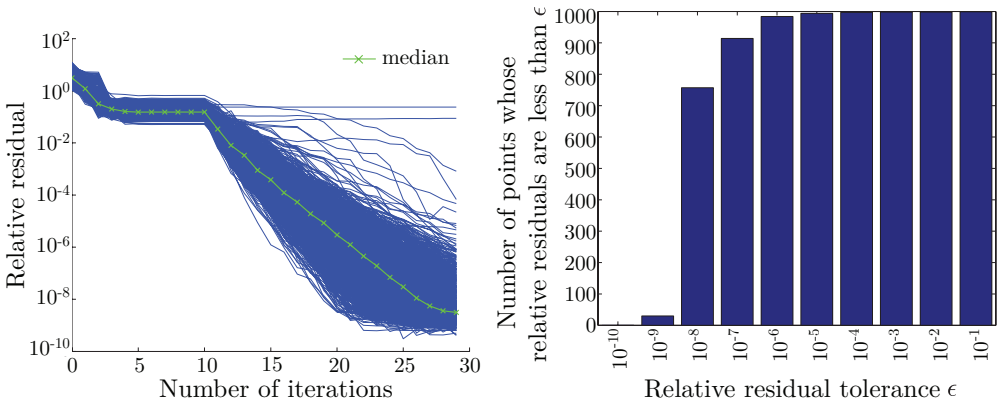
$$\text{number of total iterations} \qquad K_2 + 1 \;=\; 30 \qquad (\text{so } K_2 = 29)\,. \tag{5.84}$$

5.6.4.1 Visual representation of the solution found by the Cluster Newton method

Figure 5.14 visually represents the final set of points found by the Cluster Newton method using a total of 30,000 function evaluations with $\delta_{ODE} = 10^{-9}$.

5.6.4.2 Speed and accuracy obtained with the Cluster Newton method

In Figure 5.15(a), the relative residual calculated as in Equation (5.81) is plotted against the number of iterations. It can be seen that 30 iterations are sufficient to find the solution accurate up to the accuracy of the function evaluation. We note that the Cluster Newton method only requires one function evaluation per point in the cluster per iteration. Thus, to get the final solution, only 30,000 function evaluations are



(a) Relative residual reduction.

(b) Accuracy of the final set of points after 30,000 function evaluations with the error tolerance of $\delta_{ODE} = 10^{-9}$.

Figure 5.15: Example 2: Speed and accuracy of the Cluster Newton method.

necessary (recall that the Levenberg-Marquardt method required 469,439 function evaluations).

In Figure 5.15(b), the number of points in the final set obtained by the Cluster Newton method whose relative residuals are less than the relative residual tolerance ϵ is plotted. As can be seen in Figure 5.15(b), almost all the points achieve relative residual less than 10^{-6} and about 75% of the points achieve relative residual less than 10^{-8}.

As the Cluster Newton method requires only one function evaluation per point in the cluster per iteration, we can obtain the solution presented in Figure 5.15(b) in about 30 minutes on a server machine with two quad-core of Intel Xeon X7350 3GHz processors, which is a factor of 16 faster than the Levenberg-Marquardt method.

5.6.4.3 Influence of the accuracy of the function evaluations on the convergence of the Cluster Newton method

In Figure 5.16, we have plotted the median of the relative residuals against the number of iterations with different accuracy of function evaluation (δ_{ODE}). As can be seen from Figure 5.16, the Cluster Newton method finds solutions with accuracy close to the accuracy of the function evaluation for all values of δ_{ODE}.

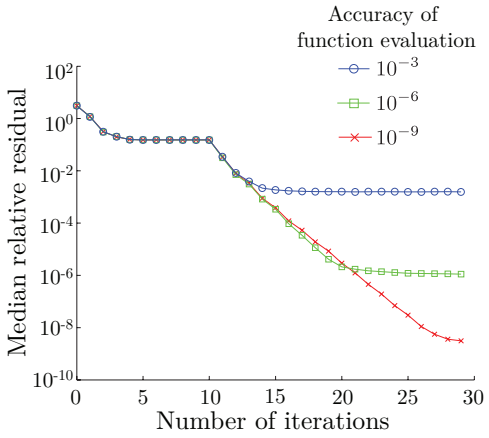


Figure 5.16: Influence of the accuracy of the function evaluation (δ_{ODE}) on the Cluster Newton method.

Thus, we observe that the Cluster Newton method is robust against small errors in the function evaluation caused by the discretization of the system of ODEs. This characteristic is especially advantageous if the desired accuracy for the solution of the inverse problem is not very high, so that we can reduce the accuracy of the numerical solution of the ODEs to save computational cost. In this way, the Cluster Newton method can, for example, obtain 1,000 solutions with relative accuracy of 10^{-3} within 4.7 minutes which is 100 times faster than the Levenberg-Marquardt method.

5.6.5 Comparison Between the Levenberg-Marquardt Method and the Cluster Newton Method

We now further compare the Levenberg-Marquardt method and the Cluster Newton method. We first show that the Cluster Newton method finds solutions similar to what the Levenberg-Marquardt method finds. Then we compare the time it takes to compute each set of solutions.

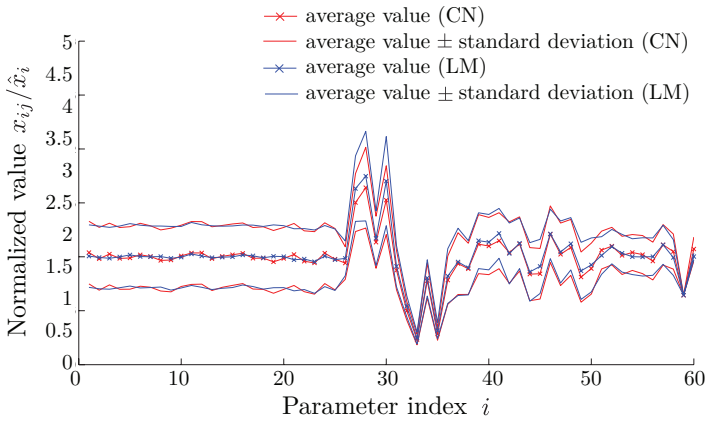


Figure 5.17: Average and standard deviation of the set of points found by the Levenberg-Marquardt method and the Cluster Newton method.

5.6.5.1 Solutions found by the Levenberg-Marquardt method and the Cluster Newton method

In Figure 5.17, we have plotted the average and the standard deviation of the points found by the Levenberg-Marquardt method and the Cluster Newton method. The average values correspond to the location of the centre of mass of the set of these points. The standard deviation is a measure for the size of the point set. As can be seen in Figure 5.17, the difference between the set of points found by the Levenberg-Marquardt method and the Cluster Newton method is marginal.

5.6.5.2 Computational cost of the Levenberg-Marquardt method and the Cluster Newton method

In Figure 5.18, we have plotted the relative residual versus the computational time. We have conducted this numerical experiment on a server machine with two quad-core Intel Xeon X7350 3GHz processors with fully-parallelized code using the MATLAB parallel computing toolbox for both methods.

In order for the Cluster Newton method to take advantage of the robustness against the numerical error of the function evaluation, we have implemented the Cluster Newton method so that the error tolerance of the function evaluation is initially set to 10^{-3} and then decreases as the number of iterations increases.

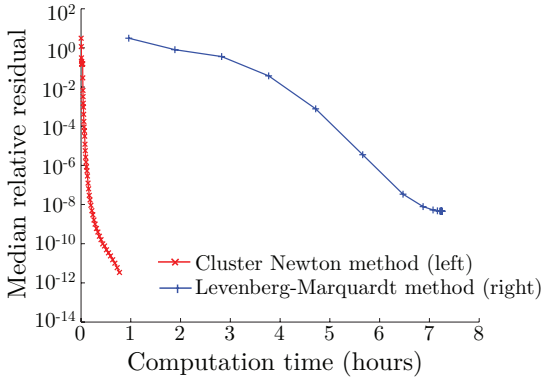


Figure 5.18: Comparison of computational cost between the Levenberg-Marquardt method and the Cluster Newton method.

The Levenberg-Marquardt method requires 61 times more function evaluations per iteration in order to estimate the Jacobian by finite differences, and the function evaluation tolerance has to be less than 10^{-9} in order for the method to converge. Thus the computational time required by the Levenberg-Marquardt method is significantly more than the time required for the Cluster Newton method when seeking multiple solutions of the underdetermined inverse problem.

This difference in computational time becomes prominent if the desired accuracy of the solution is not very high. For example, if one only requires to find solutions whose relative residual is around 10^{-3}, then the Cluster Newton method takes only about 5 minutes to find 1,000 sets of solutions. However, the MATLAB implementation of the Levenberg-Marquardt method requires over 7 hours in order to find a similar set of solutions.

5.6.6 A Brief Numerical Comparison with Other Algorithms

In addition to the Levenberg-Marquardt method, we have conducted numerical experiments using two other algorithms: Broyden's method and a Genetic Algorithm.

When the Cluster Newton method is run without Stage 1, it is essentially Broyden's method. We first use a finite difference scheme to approximate the Jacobian at each initial point. Using this as the initial Jacobian approximation, we have conducted a numerical experiment using Stage 2 of the Cluster Newton

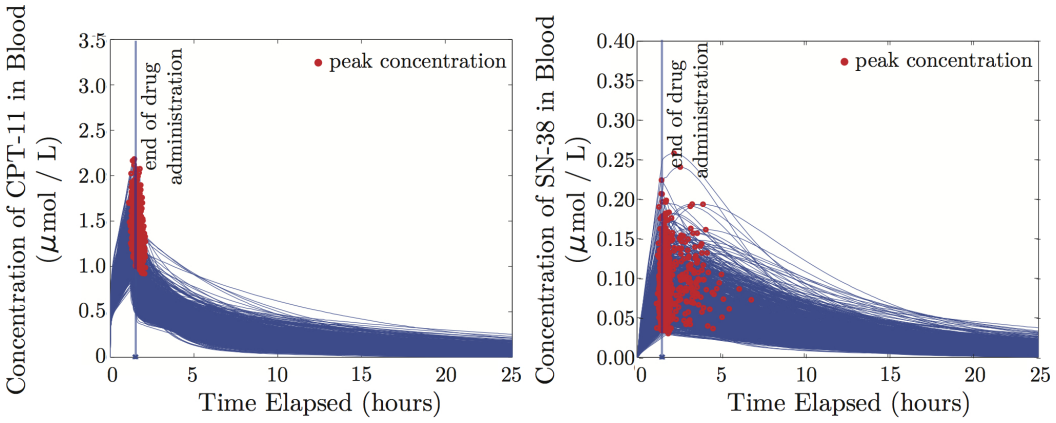
method (Broyden's method). From the numerical experiments, we have observed that, even after as many as 50 iterations, Broyden's method finds only about 350 points (out of 1,000 points) in $\mathcal{X}^*_{10^{-8}}$ compared to about 750 points for the Cluster Newton method. Also, as this approach requires approximating the initial Jacobian by a finite difference scheme, it takes 110,000 function evaluations in total (compared to 30,000 for the fully converged Cluster Newton result). Thus, we observe that Stage 1 of the Cluster Newton method is essential for its accuracy and computational efficiency.

Also, we have attempted to solve this inverse problem using the Genetic Algorithm implemented in the MATLAB optimization toolbox. Even after much trial and error with different parameters for the algorithm, we were only able to obtain solutions with relative residual larger than 10^{-1} after 72,000 function evaluations, and the method was not able to find any solution in $\mathcal{X}^*_{10^{-1}}$. Thus the Genetic Algorithm was not able to obtain accurate solutions.

5.7 Concluding Remarks on Chapter 5

We have introduced a new idea of sampling multiple points from the solution manifold of an underdetermined inverse problem for problems for which multiple solutions are of interest. We have also proposed a new computationally efficient, easy to parallelize, robust algorithm for simultaneously finding these multiple solutions of an underdetermined inverse problem. Our algorithm was applied to a coefficient identification problem of a system of nonlinear ODEs modelling the drug kinetics of an anti-cancer drug, and we demonstrated that it efficiently traces the part of the solution manifold of our interest. The efficiency and robustness of the algorithm follow from the collective way in which a linear approximation to the forward function is computed from all the points in the cluster in Stage 1. Multiple solutions are of interest in this application because they give representative samples of the possible biological states of a patient which according to the model, reproduce the observed data. This information about the patient can potentially be used, for example, to assess or design treatment plans.

Using our algorithm, 1,000 sets of model parameters can be estimated with relative accuracy 10^{-3} from clinically observed data in half an hour using a four year old laptop computer (MacBook Pro with



(a) Concentration of the anti-cancer drug CPT-11 (b) Concentration of the metabolite SN-38

Figure 5.19: 1,000 model parameter sets found by the Cluster Newton method.

2.33 GHz Intel Core2Duo processor with 4 GB of RAM). Figure 5.19 shows the predicted concentration of CPT-11 and SN-38 in blood calculated using the model parameters found by the Cluster Newton method. Detailed comparison verifies that this solution and the solution obtained through multiple applications of the Levenberg-Marquardt method (cf. Figure 5.2), after 8 hours of computation using a server machine, are very similar.

We recognize that there are many ways to further accelerate our algorithm. For example, we can choose the number of Stage 1 iterations based on the residual, or can use more sophisticated selection algorithms for the step size (i.e., the size of the update vectors $\boldsymbol{s}_{\cdot j}^{(k)}$ in lines 2-4 and 4-4).

Similarly to what has been done for the inverse problem of the drug kinetics model, we expect that applying our algorithm to other underdetermined inverse problems will efficiently provide useful information, and will also lead to new insights about the applicability of our algorithm.

Chapter 6

Conclusion

Several studies related to differential equations were presented in this thesis.

In Chapter 2, we have analytically proven that the statement "[the capillary surface] rises with the same order like the order of contact of the two arcs, which form the cusp" [40] is not only applicable to power series cusps as shown by Scholz [40], but can be extended further. In addition, with a simple geometrical argument, we have analytically proven that the capillary surface is bounded if the contact angles of the boundary arc forming the cusp are supplementary angles and the curvatures of the boundaries forming the cusp are finite. The existence of a bounded capillary surface in cusp domains was suspected; however, it was not analytically proven until our work.

In Chapter 3 we have shown that augmentation of the FVEM for linear elliptic PDEs with singular solutions can be done significantly simpler than for the FEM. This simplification resulted from the fact that the singular functions used to construct singular basis functions are harmonic functions. Our discovery indicates that the FVEM is advantageous compared to the FEM, not only for its local conservation property, but also for its efficiency in approximating singular solutions of linear elliptic PDEs.

In Chapter 4, we have introduced a new numerical methodology for constructing an accurate global approximation for unbounded capillary surfaces. The methodology was motivated by the asymptotic be-

haviour of singular solutions obtained in Chapter 2. Since approximations obtained by this new numerical methodology exhibit correct singular behaviour and are valid in the entire domain, it can be used to increase our understanding of unbounded capillary surfaces. We have built conjectures for some open problems on the singular behaviour of capillary surfaces in a cusp domain using this numerical methodology. In addition, using this methodology, we may be able to further investigate how the boundary conditions distant from the singularity may influence the singular behaviour of capillary surfaces.

In Chapter 5, we have introduced a new numerical method that simultaneously finds multiple solutions of an underdetermined inverse problem. This method efficiently finds multiple solutions by solving two different least squares problems. We have shown its robustness and efficiency through estimating the model parameters of a pharmacokinetics model from an incomplete clinical data set. Although the work presented in this thesis is still at a proof of concept stage, further improvement of this algorithm and collaboration with industry may lead to the development of useful software for parameter estimation in the pharmaceutical industry.

Through the numerical and analytical studies presented in this thesis, we have learnt that small modifications to existing numerical methods, motivated by analytical properties of the problem, can make the numerical methods significantly more accurate and efficient. Thus we would like to conclude this thesis with the following statement: if analytical and numerical studies of differential equations are combined, we can more efficiently use the power of mathematics in many fields of study that utilize differential equations.

References

[1] Y. Aoki. Analysis of asymptotic solutions for cusp problems in capillarity. Master's thesis, University of Waterloo, 2007.

[2] T. Arikuma, S. Yoshikawa, R. Azuma, K. Watanabe, K. Matsumura, and A. Konagaya. Drug interaction prediction using ontology-driven hypothetical assertion framework for pathway generation followed by numerical simulation. *BMC Bioinformatics*, 9(suppl 6), 2008.

[3] R. E. Bank and D. J. Rose. Some error estimates for the box method. *Numer. Anal.*, 24:777–787, 1987.

[4] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, 1994.

[5] P. Chatzipantelidis and R.D. Lazarov. Error estimates for a finite volume element method for elliptic PDEs in nonconvex polygonal domains. *SIAM journal on numerical analysis*, 42(5):1932, 2005.

[6] L. Chen. A new class of high order finite volume methods for second order elliptic equations. *SIAM journal on numerical analysis*, 47(6):4021–4042, 2010.

[7] Z. Chen, R. Li, and A. Zhou. A Note on the Optimal L^2-Estimate of the Finite Volume Element Method. *Advances in Computational Mathematics*, 16(4):291–303, 2002.

[8] P. Concus and R. Finn. On the Behavior of a Capillary Surface in a Wedge. *Proceedings of the National Academy of Sciences of the United States of America*, 63(2):292–299, 1969.

[9] P. Concus and R. Finn. On capillary free surfaces in a gravitational field. *Acta Mathematica*, 132(1):207–223, July 1974.

[10] F.A. de Jong, J.J. Kitzen, P. de Bruijn, J. Verweij, and W.J. Loos. Hepatic transport, metabolism and biliary excretion of irinotecan in a cancer patient with an external bile drain. *Cancer Biology and Therapy*, 5(9):1105–10, 2006.

[11] P. Deuflhard. *Newton methods for nonlinear problems: affine invariance and adaptive algorithms*. Springer-Verlag, 2004.

[12] K. Djadel, S. Nicaise, and J. Tabka. Some refined finite volume methods for elliptic problems with corner singularities. *Journal of Numerical Mathematics*, 12(4):255, 2004.

[13] R. Finn. *Equilibrium Capillary Surfaces*. Springer-Verlag, 1986.

[14] R. Finn and J.F. Hwang. On the comparison principle for capillary surfaces. *Journal of The Faculty of Science, The University of Tokyo, Section IA, Mathematics*, 36(1):131–134, 1989.

[15] Robert Finn. The contact angle in capillarity. *Physics of Fluids*, 18(4):047102, 2006.

[16] J. F. Gagne, V. Montminy, P. Belanger, K. Journault, G. Gaucher, and C. Guillemette. Common human ugt1a polymorphisms and the altered metabolism of irinotecan active metabolite 7-ethyl-10-hydroxycamptothecin (sn-38). *Molecular pharmacology*, 62(3):608–617, 2002.

[17] P. Grisvard. *Elliptic problems in nonsmooth domains*. Monographs and studies in Mathematics. Pitman, 1985.

[18] P. Grisvard. *Singularities in boundary value problems*. Springer-Verlag, 1992.

[19] J. Guzman and B. Riviere. Sub-optimal Convergence of Non-symmetric Discontinuous Galerkin Methods for Odd Polynomial Approximations. *Journal of Scientific Computing*, 40:273–280, 2009.

[20] M. C. Haaz, C. Riche, L. P. Rivory, and J. Robert. Biosynthesis of an aminopiperidino metabolite of irinotecan [7-ethyl-10-[4-(1-piperidino)-1-piperidino]carbonyloxycamptothecine] by human hepatic microsomes. *Drug Metabolism and Disposition*, 26(8):769–774, 1998.

[21] M.C. Haaz, L. Rivory, C. Rich, L. Vernillet, and J. Robert. Metabolism of irinotecan (cpt-11) by human hepatic microsomes: Participation of cytochrome p-450 3a and drug interactions. *Cancer Research*, 58(3):468–472, 1998.

[22] U. Hornung and H.D. Mittelmann. A finite element method for capillary surfaces with volume constraints. *Journal of Computational Physics*, 87(1):126–136, 1990.

[23] C.T. Kelley. *Iterative Methods for Optimization*. SIAM, 1999.

[24] J. King, J. Ockendon, and H. Ockendon. The Laplace-Young Equation Near a Corner. *The Quarterly Journal of Mechanics and Applied Mathematics*, 52(1):73–97, 1999.

[25] A. Konagaya. *Bioinformatics (in Japanese)*. The Institute of Electronics, Information and Communication Engineering, Tokyo, 2009.

[26] A. Konagaya. *Towards an In Silico Approach to Personalized Pharmacokinetics*, pages 263–28. Molecular Interactions. InTech, 2012.

[27] M. G. Larson and A. J. Niklasson. Analysis of a family of discontinuous Galerkin methods for elliptic problems: the one dimensional case. *Numerische Mathematik*, 99:113–130, September 2004.

[28] R. Li, Z. Chen, and W. Wu. *Generalized Difference Methods for Differential Equations*. Marcel Dekker Inc., 2000.

[29] F. Liebau. The finite volume element method with quadratic basis functions. *Computing*, 57:281–299, 1996.

[30] E. Miersemann. Asymptotic Expansion at a Corner for the Capillary Problem: the Singular case. *Pacific Journal of Mathematics*, 157(1):95–107, 1993.

[31] P. Moon and D. E. Spencer. *Field Theory Handbook*. Springer-Verlag, 1970.

[32] N. J. Nigro, B. P. Zellmer, D. Shangguan, and P. S. Lee. A modified finite element method for determining equilibrium capillary surfaces of liquids with specified volumes. *International Journal for Numerical Methods in Fluids*, 33(6):833–846, July 2000.

[33] J. Norbury, G. Sander, and C. Scott. Corner solutions of the Laplace-Young equation. *The Quarterly Journal of Mechanics and Applied Mathematics*, 58(1):55–71, 2005.

[34] J.T. Oden, I. Babuska, and C.E. Baumann. A discontinuous *hp* finite element method for diffusion problems: 1-D analysis. *Journal of Computational Physics*, 146(2):491–519(29), 1998.

[35] M. Plexousakis and G.E. Zouraris. On the construction and analysis of high order locally conservative finite volume type methods for one dimensional elliptic problems. *SIAM journal on numerical analysis*, 42(3):1226, 2004.

[36] V. K. Polevikov. Methods for numerical modeling of equilibrium capillary surfaces. *Differ. Equations*, 35(7):985–991, 1999.

[37] V. K. Polevikov. Methods for numerical modeling of two-dimensional capillary surfces. *Computational Methods in Applied Mathematics,*, 4(1):66–93, 2004.

[38] Å. Björck. *Numerical methods for least squares problems*. Society for Industrial and Applied Mathematics, Philadelphia, 1996.

[39] M. Scholz. *Über das Verhalten von Kapillarflächen in Spitzen*. PhD thesis, Universität Leipzig, 2001.

[40] M. Scholz. On the asymptotic behaviour of capillary surfaces in cusps. *Zeitschrift für Angewandte Mathematik und Physik*, 55:216–234, 2003.

[41] C. Scott, G. Sander, and J. Norbury. Computation of capillary surfaces for the Laplace-Young equation. *The Quarterly Journal of Mechanics and Applied Mathematics*, 58(2):201–212, 2005.

[42] L. F. Shampine and M. W. Reichelt. The matlab ode suite. *SIAM journal on scientific computing*, 18(1):1, 1997.

[43] L. Simon. Regularity of capillary surfaces over domains with corners. *Pacific J. Math.*, 88(2):363–377, 1980.

[44] J. G. Slatter, L. J. Schaaf, J. P. Sams, K. L. Feenstra, M. G. Johnson, P. A. Bombardt, K. S. Cathcart, M. T. Verburg, L. K. Pearson, L. D. Compton, L. L. Miller, D. S. Baker, C. V. Pesheck, and R. S. Lord III. Pharmacokinetics, metabolism, and excretion of irinotecan (cpt-11) following i.v. infusion of [14c]cpt-11 in cancer patients. *Drug Metabolism and Disposition*, 28(4):423–433, 2000.

[45] J. G. Slatter, P. Su, J. P. Sams, L. J. Schaaf, and L. C. Wienkers. Bioactivation of the anticancer agent cpt-11 to sn-38 by human hepatic microsomal carboxylesterases and the in vitro assessment of potential drug interactions. *Drug Metab Dispos*, 25(10):1157–1164, 1997.

[46] G. Strang and G. J. Fix. *An Analysis of the Finite Element Method*. Prentice-Hall, 1973.

[47] W. Tucker, Z. Kutalik, and V. Moulton. Estimating parameters for generalized mass action models using constraint propagation. *Mathematical Biosciences*, 208(2):607–620, 2007.

[48] A. Vogel. Higher order finite volume method for elliptic partial differential equations. In *European Multigrid Conference 2008 (oral presentation)*, 2008.

[49] A. Vogel, J. Xu, and G. Wittum. A generalization of the vertex-centered finite volume scheme to arbitrary high order. *Computing and visualization in science*, 13(5):221–228, 2010.

[50] S. Willmann, K. Hohn, A. Edginton, M. Sevestre, J. Solodenko, W. Weiss, J. Lippert, and W. Schmitt. Development of a physiology-based whole-body population model for assessing the influence of individual variability on the pharmacokinetics of drugs. *Journal of pharmacokinetics and pharmacodynamics*, 34(3):401, 2007.

[51] J. Xu and Q. Zou. Analysis of linear and quadratic simplicial finite volume methods for elliptic equations. *Numerische Mathematik*, 111(3):469–492, 2009.