FINITE ELEMENT METHODS FOR IMPLICIT SOLVENT MODELS

BY

JEHANZEB HAMEED CHAUDHRY

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science in the Graduate College of the University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Doctoral Committee:

Assistant Professor Luke N. Olson, Chair and Director of Research
Professor Michael T. Heath
Professor William D. Gropp
Assistant Professor Aleksei Aksimentiev
Professor Michael Holst
Abstract

In this thesis, we develop finite element methods (FEMs) for implicit solvent models. Implicit models treat the solvent in biomolecular systems as bulk continuum and are thus computationally efficient. The Poisson-Boltzmann equation (PBE) is one important example. We design a FEM for the linear PBE by directly exposing the flux through a first-order system. We propose a system that yields a tractable least-squares finite element formulation and establish theory to support this approach. Moreover, we explore the impact of weighting and the choice of elements on conditioning and adaptive refinement. In a series of numerical experiments, we compare the finite element methods when applied to the problem of computing the solvation free energy for realistic molecules of varying size.

The inclusion of steric effects is important in regions of high potential. Hence, we consider a modified PBE, in order to model these effects. We establish well-posedness of the weak problem along with convergence of an associated finite element formulation. We also examine several practical considerations such as conditioning of the linearized form of the nonlinear modified Poisson-Boltzmann equation, implications in numerical evaluation of the modified form, and utility of the modified equation in the context of the classical Poisson-Boltzmann equation.

The modified Poisson-Nernst-Planck equations model the dynamics of ions in solvated systems, while accounting for steric effects. We develop an efficient SUPG FEM method to prevent spurious values in regions there are high repulsive potentials. Our method takes into account conservation of ions and periodic boundary conditions common in molecular dynamics simulations. We apply our method to compute current due to the flow of ions through a nanopore.

Finally, we develop the theoretical and algorithmic framework for efficiently including quantities-of-interest (QOIs) into the least-squares framework. We show existence, uniqueness, and implications from a linear algebra point-of-view. Furthermore, we develop bounds on the error of the solution and the QOI. We outline an adaptive refinement algorithm based on our approach and support our approach with numerical result for several application areas, including solvation free energy for the PBE.
To Ami
Acknowledgments

First and foremost I want to thank my advisor, Luke Olson, for the major role he has played in my graduate studies. He provides good advice on specific technical issues as well as on broader research directions. However, he has qualities that make him an exceptional advisor. Chief among these qualities are his empathy and understanding of issues regarding all aspects of graduate life. I will forever be grateful for his involvement in fulfilling my research potential, and his encouragement that I engage with other researchers at conferences and present my research at other research institutions. His efforts have broadened my outlook on scientific computing, while simultaneously increasing my enthusiasm to further pursue research. He has always been available willing to provide me with guidance whenever I needed it. I am also grateful to him for helping improve my writing style tremendously over the course of my graduate studies.

I also want to thank Michael Heath for his efforts in promoting a collaborative research environment in scientific computing. His organization of the scientific computing seminar has played a significant role in my development as a graduate student. Not only did I learn about the research of other graduate students, but I also presented my research several times. His advice and critique on my talks have helped me hone my presentation skills. His knowledge of the history of the field has also been a source of inspiration to me.

I am also grateful to my committee members: Michael Heath, William Gropp, Michael Holst and Aleksei Aksimentiev. Aleksei Aksimentiev provided useful guidance in my work on the finite element method for the modified Poisson-Nernst-Planck equations. Michael Holst’s insights in my research have helped my improve my thesis. I am also thankful to Steve Bond for introducing me to the Poisson-Boltzmann equation. His willingness to talk about any aspect of numerical analysis has always been refreshing. I am also grateful to the Computation Science and Engineering, for the CSE fellowship which allowed me to focus my efforts on my research.

My fellow graduate students and friends have made a positive contribution to my graduate experience. Jacob Schroeder and Nana Arizumi have been supportive office mates. Eric Cyr, Bill Cochran, Jacob Schroeder, Russ Hewett, Mark Gates, Adam Reichert, Elena Caraba, Hormozd Gahvari, James Lai, Peng
Jiang, Michael Wolf and Dmytro Yershov have all helped in many forms like critiquing my talks, or distilling research ideas. Russ Hewett also provided me with the template used for this thesis. Jane Butterfield and Hemanta Maji have been a source of constant encouragement, and also of expert help in making Tikz figures. I also want to express gratitude to the CS office staff, especially Mary Beth Kelley, in providing guidance about the requirements of the PhD program.

My family have always been supportive of my efforts in obtaining a PhD. My parents always encouraged me to pursue the fields which I find interesting. My mother (whom I call Ami), instilled in me the desire for learning and curiosity to explore the unknown questions. My wife, Sana Mahmood has been a pillar of strength and support for me, and helped me numerous times when I have been overwhelmed by the graduate experience. My daughter, Zoya, while not understanding what I did, always gave me a smile when I sat down to work on my thesis, and probably wanted to help me type it, as she constantly tried to get to the keyboard of my laptop. I also want to point out the support of my friends Usman Mansoor, Salman Khan, Ahmad Javed, Waqas Arshad, Mansoor Sadiq and Ahmed Luqman. Finally, I want to thank the many great teachers, from my primary school to the graduate level, who inspired critical thinking and desire of knowledge in me. I especially want to point out the efforts of Tahir Mustafa, who was great in conveying complex concepts in mathematics, and encouraging me to apply for graduate studies.
## Table of Contents

Chapter 1  Introduction .................................................. 1
  1.1 Implicit Solvent Models ............................................. 2
  1.2 Finite Element Methods .......................................... 4
  1.3 Organization ....................................................... 5

Chapter 2  A First-Order Systems Least-Squares Finite Element Method for the Poisson-Boltzmann Equation .................................................. 10
  2.1 Introduction to Numerical Methods for the PBE ............... 10
  2.2 Poisson Boltzmann Equation ...................................... 12
  2.3 FOSLS Formulation of PBE ........................................ 16
    2.3.1 Ellipticity of FOSLS Functional ......................... 18
  2.4 Numerical Experiments ........................................... 23
  2.5 Conclusion ......................................................... 30

Chapter 3  Robust Finite Element Methods for the Poisson-Boltzmann Equation .................................................. 31
  3.1 Least-squares Finite Element Formulation ..................... 32
  3.2 Description of FEM Spaces ....................................... 32
  3.3 On the use of Piecewise Linear and Raviart–Thomas Elements ............... 33
  3.4 Comparison of Different Weightings of the LSFEM Functional ......................... 37
  3.5 Comparison of the LSFEM and Galerkin ......................... 39
  3.6 Interface Formulation ............................................ 40
  3.7 Conclusions ....................................................... 43

Chapter 4  Finite Element Approximation to a Finite-Size Modified Poisson-Boltzmann Equation .................................................. 45
  4.1 Motivation ......................................................... 45
  4.2 The Modified Poisson Boltzmann Equation ..................... 46
  4.3 Existence and Uniqueness ........................................ 50
  4.4 Finite Element Discretization ................................... 53
    4.4.1 Analysis of the Nonlinear Finite Element Approximation ......................... 54
    4.4.2 Nonlinear Solve .............................................. 57
  4.5 Solution to the Nonlinear PBE using the MPBE ............... 63
  4.6 Conclusion ....................................................... 66

Chapter 5  Finite Element Method for the Modified Poisson-Nernst-Planck Equations .................................................. 67
  5.1 Problem Description ............................................. 68
    5.1.1 Governing Equations ....................................... 69
    5.1.2 Description of the Model System ......................... 70
    5.1.3 Brownian Dynamics ........................................ 72
  5.2 Numerical Methods .............................................. 72
Chapter 1

Introduction

Computer simulations are vital for research in biomolecular systems. Among other things, they aid in discovering structure, dynamics, and function of proteins [16] and computing ionic flux through a pore in a membrane, surrounded by salt baths [93]. The complex interaction of forces within and between molecules pose a challenge for computer experiments. The forces can be classified as long range or short range forces. Short-range forces include van der Waals, bonding, angular, and torsional interaction forces. Long-range forces are determined by electrostatic interactions and are important at all length scales. Therefore, accurate and efficient evaluation of electrostatics is a significant challenge in computational molecular biology [67]. Models which aid in computation of such interactions greatly benefit computer simulations of biomolecular systems.

Computational methods to simulate electrostatic interactions are broadly classified as either explicit or implicit models. Explicit models simulate all atoms of a biomolecular system, including those belonging to the solvent. These methods require integration over a huge number of solvent degrees of freedoms. Hence in spite of recent improvements such as Ewald summations, Euler summations and reaction field theory explicit methods, they are extremely expensive. Implicit solvation models offer a computationally tractable alternative to explicit methods.

Implicit methods treat the solvent as bulk continuum and use a mean field approximation of the force exerted by solvent and counterions on the biomolecule. This results in a significant increase in computational efficiency, and hence implicit methods have enjoyed widespread use over recent years in solvated biomolecular systems. There are many other reasons to use implicit solvation models besides reduction in computational resources. Boundary conditions are easier to model in implicit models. Furthermore, explicit methods usually describe the system with the same resolution over the entire simulation domain. Thus, extensive computational effort is applied to simulate a nearly uniform solution where quantities of interest exhibit little variation. In contrast, continuum methods allow different regions of the same system to be described at varying levels of detail, and thus focus the computational effort on regions that require a more precise
description. However, implicit methods are not a panacea; there are challenges in both accurately solving implicit models, and improving models to better account for short range forces among solvent molecules which are averaged out in traditional implicit models.

This work explores finite element methods for simulation of implicit solvent models. There are two main approaches for solving implicit solvent models; the finite difference method (FDM) and the finite element method (FEM). The finite difference method is poor at treating curved and complex geometries due to its reliance on structured Cartesian grids. Moreover, they often require high regularity of the solution to prove convergence. On the other hand, finite element methods are easily implemented over unstructured simplicial meshes, which are excellent in modeling complicated structures. Quite often the theory of convergence of the finite element solution follows from the continuum theory. Finite element theory also handles problems with less regularity and easily models complex boundary conditions in a natural manner.

In this thesis we use a two fold strategy to accurately simulate biomolecular systems using finite elements. We explore new finite element methods for the common implicit solvent models, to expose more efficient ways of computing quantities of interest like the flux or free energy. To achieve this, we expose the first-order variables and quantities of interest directly in the problem formulation, making use of the least-squares finite element method. A second thrust is to develop finite element theory and methods for the more recent and accurate implicit solvent models which take ion size effects into account. These models retain the mean-field nature of the classical implicit models and hence are efficiently simulated. The methods we propose in this work are used to efficiently simulate older models in domains where they are valid, or to solve the newer models when the classical models break down due to ion-size effects. The remainder of this chapter gives a brief description of the implicit solvent models and different finite element methods used to solve them in this thesis.

1.1 Implicit Solvent Models

Implicit solvent models are classified into two broad categories: those that deal with systems in equilibrium and those which capture the dynamics of the system due to the motion of the solvent ions. The equilibrium case is described by the Poisson-Boltzmann equation and its modifications, while the dynamics are captured by the Poisson-Nernst-Planck equations and their modifications.

The Poisson-Boltzmann equation (PBE) is a second order nonlinear elliptic partial differential equation describing the electrostatic activity between molecules in an ionic solvent. The PBE arises from finding the critical point of a electrostatic free energy functional [62]. The PBE has been an important tool in numerous
biomedical and biophysical applications [12, 56]. It is used in electrostatic potential analysis, computation of solvation free energy, and $PK_a$ values [91, 69]. We describe the PBE in more detail in Chapter 2.

Although the Poisson-Boltzmann approach has been widely successful in describing a range of systems, it has its drawbacks. It neglects correlations and all specific (non-electrostatic) interactions between the ions including the ionic finite size. Thus, it predicts unrealistically high counterion concentration near a highly charged molecular surface, even if the bulk solution is very dilute [35]. At high ion densities achieved close to the surface, short-range ion-ion interactions become comparable to the Coulomb interaction and they can no longer be neglected. Borukhov et al [34] proposed a lattice gas formulation of solvent ions to derive the modified PBE (MPBE). In their model the solvent ions occupy cells in a lattice of size $a$. The modified PBE (MPBE) has also been called Poisson-Bikerman equation [87], as Bikerman proposed a model to account for steric effects due to finite size ions in 1942 [27]. The modified equation has the advantage of bounded concentrations of salt ions near the molecular surface. This is illustrated in Figure 1.1 which depicts the gathering of counterions near a charged surface. Due to the assumption of point ions in PBE, an unrealistically high number gather near the surface, leading to nonphysically large concentrations. In the modified implicit solvent models, due to the assumption of finite size of ions, only a bounded number gather round the surface, which keeps the concentration bounded. As the size of the salt ions go to zero, the modified equation converges to PBE. Since being proposed MPBE has recently been used in different applications where ordinary PBE gave unphysical results [88, 35, 87, 123], and we believe that this equation will see more widespread use instead of PBE. Further details on the MPBE are given in Chapter 4.

![Figure 1.1: Effect of ion size in implicit solvent models.](a) PNPE (b) MPNPE

While the Poisson-Boltzmann equation models the equilibrium electrostatic potential, dynamics of biomolecular systems are described by the Poisson-Nernst-Planck equations (PNPE). The PNPE are a coupled
system of partial differential equations which model the transport of ions under an electric field [18]. PNPE suffer from the same drawback as PBE; they assume point-like ions. The theory for MPBE is extended to obtain modified Poisson-Nernst-Planck equations (MPNPE) [88] which take into account for finite-size of the ions. The PNPE and the MPNPE are used to study the flow of ions through a channel in a membrane. In the presence of attractive potential, or a charged surface such as a deoxyribonucleic acid (DNA) molecule, PNPE give unrealistically high concentration. The ion-size parameter in MPNPE leads to more accurate modeling of such systems. Chapter 5 describes the MPNPE in greater depth.

1.2 Finite Element Methods

In this section, we outline the numerical methods that have been used to solve the implicit solvent models, as well as introducing our methods. A number of different directions for numerically solving the Poisson-Boltzmann equation have been pursued. Approaches such as finite difference and finite volume methods [129, 71, 55, 78, 15, 104, 17, 116, 139], finite elements methods [49, 110, 53, 79, 14, 119, 132, 131], boundary element methods [108, 137, 114, 133, 86, 128, 138, 26, 127, 97, 36, 99], and integral equations [92, 100] have been developed for this problem. Yet, as the complexity of applications increases so do the demands on the numerical approximation, and we are motivated to investigate additional computational tools that provide a medium for more robust and efficient simulation.

We investigate first-order systems least-squares (FOSLS) finite element method for numerically solving the PBE. There is a need to expose the gradient of the solution, which is used in calculating electrostatic force, e.g., in the Smoluchowski equation [121]. Traditional finite element methods have relied on second order Galerkin methods, or mixed methods. The second order Galerkin method does not expose the gradient directly. Gradients are incorporated directly by converting the (second order) system in a first-order form. There are two main options to solve the first-order systems, mixed methods, or FOSLS. Mixed methods constrain the choice of the finite element spaces which can be used, as they need to satisfy the inf-sup condition of Ladyzhenskaya-Babuška-Brezzi [38]. FOSLS is an exciting new alternative, which does not have this constraint, and offers a number of other advantages. The least-squares method begins by reforming the partial differential equation as a first-order system. A functional is then constructed based on the residual equations of the first-order system, and is minimized. A first-order system least-squares (FOSLS) approach to finite elements is shown to be effective for numerous problems [40, 41], including those with discontinuous coefficients [107, 25, 24]. FOSLS offers many improvements over existing methods. The functional is minimized using a variational principle, giving rise to a symmetric bilinear form. A discretization based on this form
leads to a symmetric positive-definite linear system, which is ideal for solvers such as preconditioned conjugate gradient. Also, the bilinear form is often elliptic with respect to a practical norm, and as a result the finite element spaces do not need to satisfy the discrete inf-sup condition, unlike mixed methods. A practical consequence is that basic finite element spaces, e.g. continuous piecewise linear polynomials, may be used for all variables. The FOSLS functional also defines a local error indicator, which is used for adaptive refinement. We detail the FOSLS methodology and its application to the PBE in Chapters 2 and 3. We often refer to the finite element method based on FOSLS as least-squares finite element method (LSFEM).

The modified Poisson-Boltzmann equation is a non-linear second order partial differential equation. The existence and uniqueness of the solution to the non-linear PBE were studied in [49]. We establish similar results for the MPBE. The finite-element method we investigate for the MPBE is the Galerkin method, which is useful for solving for the electrostatic potential accurately. The PNPE and the MPNPE present unique challenges in designing a stable FEM that were not present for the Poisson-Boltzmann models. One of these challenges arises from the accurate representation of an applied potential to the system. Typically, the PNPE solvers [83], ignore the sharp repulsive potential near the interface of the fluid and the membrane, and assume the potential to become infinity at the interface. We represent this repulsive potential explicitly to accurately capture the concentration profiles near the interface. However, traditional Galerkin method is unstable for such a system with a large advective term. To stabilize the system, and get meaningful physical results, we use the streamline upwind Petrov Galerkin (SUPG) method [80] to solve the PNPE and the MPNPE. The SUPG method adds a small amount of artificial diffusion along the streamlines to stabilize the system.

An important role of computer simulations is to evaluate quantities of interest (QOI) like the free energy, which are functions of the solution. The focus of traditional approaches to finite element methods is to minimize the error in the discrete solution globally. Goal-based finite element methods solve a dual problem to approximate the QOI, and guide local mesh refinement in computing the QOI efficiently. The goal-based FEM have been studied in the context of adjoint based error estimation [105, 60] and are based on classical FEMs like the Galerkin or mixed methods. An important contribution of this thesis is to explore goal-based methods for LSFEM and apply it to compute QOI in implicit solvent models.

1.3 Organization

The rest of the dissertation is organized as follows. We introduce the Poisson-Boltzmann equation and the least-squares finite element methodology in Chapter 2. The chapter begins with a description of the PBE, different transformations used to make it computationally amenable and its linearization. Then, the
least-squares finite element method, also referred to as first-order systems least squares (FOSLS) is introduced and its advantages explained. We point out that the flux typically used in least-squares methods does not work for the PBE. We design a new flux which takes into account the discontinuous nature of the coefficients, and ensures that it is in the Sobolev space $H(\text{div})$. We prove the ellipticity of the FOSLS functional and show we have a well-posed problem. We also give an adaptive refinement algorithm based on the FOSLS functional. Finally, we show the practical soundness of the method with numerical examples on a variety of molecules, using our adaptive refinement algorithm.

Chapter 3 further explores various issues related to LSFEM for the PBE. We examine $H(\text{div})$ conforming Raviart-Thomas elements for representing the flux variable, and argue that its leads to a robust method with better convergence properties. Then we look at different weighting of the LSFEM functional. We show that appropriately weighting the functional is vital for the accuracy of the solution and efficiency of the numerics. Finally, we compare the competitiveness of the LSFEM method to the standard Galerkin method, and show the advantages of the LSFEM approach when modeling first-order variables.

The modified Poisson-Boltzmann equation is addressed in Chapter 4. In this chapter, we establish the uniqueness and existence of the solution to this non-linear equation. This is followed by formulating a non-linear FEM to discrete solve this equation. We prove uniqueness and existence of the discrete solution, and show optimal convergence of the method. This first part of the chapter deals with the theoretical aspects of the equation and its discretization using FEM. In the latter part of the chapter, we show numerical studies validating our predictions, and investigate conditioning of the linear systems for the MPBE and the PBE. The chapter shows the physical and numerical advantages of the MPBE over PBE. Moreover, we show that if someone want to solve the PBE, solving MPBE as an initial step leads to a more efficient solver as compared to solving PBE directly.

Finite element methods for the PNPE and the MPNPE are discussed in Chapter 5. We discuss the equations, and note the non-linearity in the MPNPE which accounts for the finite size ion effects. We formulate a FEM to compute the steady state concentrations (or equivalently, ion probability distributions). The steady state solver is based on time evolution of the PNPE and the MPNPE equations. This is followed by a discussion of the non-linear FEM for the MPNPE. As we discussed earlier, traditional PNPE solvers ignore the repulsive potentials near the walls of the solvent domain. We capture this potential directly in our work. We show incorporating this potential leads to instabilities in the classical Galerkin method. Hence we propose the SUPG method for the PNPE and the MPNPE to accurately model the applied repulsive potential along the walls of the domain in a stable fashion. Finally the chapter looks at the difference between
the MPNPE and the PNPE for ions flowing through a channel in the presence of a DNA, and shows that MPNPE gives more physically meaningful results.

The last contribution of the thesis is to investigate QOI based LSFEM in Chapter 6. The chapter begins by outlining a modified LSFEM functional to account for the QOI directly. We then proceed to show some theoretical results on the convergence of the solution and the QOI based on this modified functional. Then we devise an algorithm to adaptively refine the mesh to resolve the QOI accurately and efficiently. Finally we compare the performance of the different functionals and estimators based on those functionals, and highlight the utility of this approach.

The contributions of this thesis are summarized in Figure 1.2. The figure lists both the broad and the specific contributions of the thesis.: least-squares for the PBE, FEM for the MPBE, FEM for the MPNPE and a novel framework for incorporating QOI in the least-squares framework. The outline of different chapters addressing these contributions is shown in Figure 1.3.
Figure 1.2: Contributions of the thesis
Poisson-Boltzmann Equation (PBE) 
LSFEM for PBE (Chapters 2,3) 
Linear PBE 
LSFEM for PBE (Chapters 2,3) 

High Potential: 
Crowding of ions 

Modified PBE (MPBE) 
Analysis and FEM for MPBE (Chapter 4) 
Quantity of Interest based Least-Squares (Chapter 6) 

Modified Poisson-Nernst-Planck Equations (MPNPE) 
Nonlinear FEM, SUPG for MPNPE (Chapter 5) 

Figure 1.3: Organization of the thesis
Chapter 2

A First-Order Systems Least-Squares Finite Element Method for the Poisson-Boltzmann Equation

The focus of this chapter is on numerical solutions to the Poisson-Boltzmann equation (PBE), which approximates the mean solvent forces by assuming the ions are distributed according to the Boltzmann distribution. This results in a unique electrostatic potential [49]. In this chapter, we seek a numerical solution of the linearization of the regularized PBE (RPBE). The use of a regularized formulation [49] is required because the original statement of the PBE yields singularities in the electrostatic potential. Regularization overcomes this issue by analytically subtracting the singularities from the electrostatic potential yielding a modified version of the original PDE. To further simplify the problem, and focus on the efficacy of our discretization, we linearize the RPBE. Consequently, the linearized version has many of the same challenges as the RPBE, however it features reduced computational cost [63] while remaining a physically accurate perturbation to the fully nonlinear problem [64].

The remainder of the chapter is organized as follows. Section 2.1 briefly introduces the PBE, discusses previous numerical approaches in solving the PBE, and lays out the foundation for the use of least-squares formulation. In Section 2.2, we summarize the PBE, its regularization and linearization, and the general problem domain. We outline the FOSLS terminology in Section 2.3 and introduce our formulation of the method. Furthermore, we establish theoretically the use our formulation and discuss implications and techniques for computational simulation. In Section 2.4, we provide numerical evidence of effectiveness of the FOSLS approach for a number of molecular systems. The method is shown to be effective for problems with known solutions (Born ion), for more complicated structures (Fasciculin 1 and methanol), and for a problem with low regularity (dipoles).

2.1 Introduction to Numerical Methods for the PBE

A number of different directions for numerically solving the Poisson-Boltzmann equation have been pursued. Approaches such as finite difference and finite volume methods [90, 129, 71, 55, 78, 15, 104, 17, 116, 139],
finite elements methods [49, 110, 53, 79, 14, 120, 132, 131], boundary element methods [108, 137, 114, 133, 86, 128, 138, 26, 127, 97, 36, 99], and integral equations [92, 100] have been developed for this problem. Yet, as the complexity of applications increases so do the demands on the numerical approximation, and we are motivated to investigate additional computational tools that provide a medium for more robust and efficient simulation.

There are several variational approaches to the Poisson-Boltzmann equation, and traditionally the Rayleigh-Ritz or Galerkin finite element method has been the method-of-choice with either the linearized equation [53, 131] or the non-linear equation [79, 14, 119]. The Galerkin method is a viable approach to Poisson-Boltzmann calculations due to its relative simplicity, strong theoretical base, and proven results in practice. Even so, there is an opportunity for more advanced approximations when considering other variational formulations of the problem. In this chapter, we focus on a variational setting for the PBE due to the underlying theoretical support for numerical methods and the established analysis of the equation. In particular, we propose a least-squares finite element formulation of the linear regularized Poisson-Boltzmann equation.

The least-squares method we follow begins by reforming the partial differential equation as a first-order system. A functional is then constructed based on the residual equations of the first-order system, and is minimized. A first-order system least-squares (FOSLS) approach to finite elements has shown to be effective for numerous problems. In particular, elliptic problems [40, 41] with discontinuous coefficients [107, 25, 24] are theoretically competitive and numerically plausible.

The existing FOSLS theory motivates our treatment of the PBE, yet the theoretical properties for the FOSLS form we pose in Section 2.3 are not fully developed. We establish these results and confirm the existence of a unique solution for our problem. Moreover, we propose a first-order system for the PBE in Section 2.3 that correctly addresses the jump discontinuity inherent in the problem. The PBE is described through a dielectric coefficient, $\epsilon(x)$, and Debye-Hückel parameter, $\kappa(x)$, that are discontinuous across an interface. Proper treatment of the flux term across this interface is critical to the variational formulation. To this end, we propose a unique form of the flux that both captures the underlying physics and yields a system amenable to a least-squares minimization.

The goal of this chapter is to outline a least-squares finite element method for use with existing computational tools, such as the Finite Element Toolkit (FETk) [76], which uses piecewise linear elements over tetrahedral tessellations of single domains. The result is a competitive and straightforward finite element method for the PBE using adaptive mesh refinement.
Adaptive refinement using finite elements has been studied for the Poisson-Boltzmann equation in a Galerkin formulation [79, 14]. These approaches focus on resolution of the singularities in the original PDE. Here, we use the functional provided by the least-squares formulation to guide refinement with similar success. Treatment of the interface condition is automatic in our formulation of the problem, naturally capturing the physics around the interface while still amenable to approximation by standard finite elements.

2.2 Poisson Boltzmann Equation

The Poisson-Boltzmann equation models the electrostatic activity between molecules in an ionic solvent. In this model, it is assumed that the ions in the solvent are distributed according to the Boltzmann distribution and that the potential of the mean force on a particle is simply the charge of the ion times the electrostatic potential. This yields the general Poisson-Boltzmann equation [64],

\[-\nabla \cdot (\epsilon(x) \nabla \phi(x)) = 4\pi \rho_f(x) + 4\pi \sum_{j=1}^{n_s} c_j^s Q_j^s \lambda_j(x) \exp \left( \frac{-1}{k_B T} Q_j^s \phi(x) \right), \quad \text{(2.1a)}\]

\[\lim_{\|x\| \to \infty} \phi(x) = 0. \quad \text{(2.1b)}\]

Here, \(\phi\) is the unknown electrostatic potential, \(\epsilon\) is the dielectric coefficient, \(\rho_f\) is the fixed charge distribution in the solute (biomolecule), \(k_B\) is the Boltzmann constant, and \(T\) is the temperature. It is assumed that the solvent is composed of \(n_s\) species of ions, each with charge \(Q_j^s\) and concentration \(c_j^s\). The accessibility of the \(j\)th ion-species to a point, \(x\), in space is described by \(\lambda_j(x)\).

For a solute in a 1:1 electrolyte solvent (e.g. NaCl), the charge of each ion species is \(\pm 1\) unit charge, and the general Poisson-Boltzmann equation simplifies [16] to

\[-\nabla \cdot (\epsilon(x) \nabla \phi(x)) + \hat{\kappa}^2(x) \left( \frac{k_B T}{\epsilon_c} \right) \sinh \left( \frac{\epsilon_c \phi(x)}{k_B T} \right) = 4\pi \sum_{i=1}^{m} Q_i \delta(x - x_i), \quad \text{(2.2a)}\]

\[\lim_{\|x\| \to \infty} \phi(x) = 0. \quad \text{(2.2b)}\]

Here, we have further assumed that solute contains a total of \(m\) fixed point charges, with the \(i\)th charge, \(Q_i\), centered at position \(x_i\). The resulting distribution, \(\rho_f\), is a linear combination of Dirac delta functions, \(\delta(x - x_i)\).

The domain for the problem, \(\mathbb{R}^3\), is subdivided into a molecular region, \(\Omega_m\), a solvent region, \(\Omega_s^\infty\), and an interface between the two, denoted by \(\Gamma\). The solute is surrounded by solvent, which is represented as
a continuum over the subdomain $\Omega_s^\infty = \mathbb{R}^3 \setminus \Omega_m$. In some Poisson-Boltzmann models, an additional ion exclusion or Stern layer is present between $\Omega_s$ and $\Omega_m$. The Stern layer provides separation between the solute and the ions of the solvent. As a result, the dielectric matches the dielectric in the solvent region and the ionic strength is zero ($\bar{\kappa} = 0$). In this chapter we focus on the more challenging issue of the jump in the dielectric, and neglect the Stern layer. The subdomains for a typical biomolecular solute are shown in Figure 2.1. The dielectric coefficient, $\epsilon(x)$, and modified Debye-Hückel parameter, $\bar{\kappa}(x)$, describe the accessibility of the solvent to the solute and are defined on $\Omega_m \cup \Omega_s^\infty$ by the piecewise constant functions

$$
\epsilon(x) = \begin{cases} 
\epsilon_m & x \in \Omega_m \\
\epsilon_s & x \in \Omega_s^\infty
\end{cases} \quad \text{and} \quad \bar{\kappa}^2(x) = \begin{cases} 
0 & x \in \Omega_m \\
\bar{\kappa}_s^2 = \epsilon_s \frac{8 \pi N_A e^2}{1000 k_B T} I_S & x \in \Omega_s^\infty
\end{cases}.
$$

(2.3)

Here, $\epsilon_m$ and $\epsilon_s$ are positive constants, $N_A$ is Avogadro’s number, and $e_c$ is the charge of a proton. The ionic strength, $I_S$, is a physical parameter which varies depending on the solvent.

For computational reasons, the unbounded solvent domain, $\Omega_s^\infty$, is typically truncated at a finite radius from the “center” of the molecule, which gives rise to a bounded solvent domain, $\Omega_s$. Dirichlet boundary conditions are imposed to capture the asymptotic behavior of the solution on an unbounded domain. Combining this with the change of variables, $\tilde{u}(x) = e_c \phi(x)/k_B T$, results in a dimensionless Poisson-Boltzmann equation on
the spherical domain $\Omega = \Omega_m \cup \Omega_s \cup \Gamma$:

$$-\nabla \cdot (\epsilon(x)\nabla \tilde{u}(x)) + \tilde{\kappa}^2(x) \sinh \tilde{u}(x) = \frac{4\pi e_c}{k_BT} \sum_{i=1}^{m} Q_i \delta(x - x_i), \quad x \in \Omega_m \cup \Omega_s,$$  \hspace{1cm} (2.4a)

$$\tilde{u}(x) = g(x), \quad x \in \partial \Omega_s,$$  \hspace{1cm} (2.4b)

$$\left[\epsilon(x) \frac{\partial \tilde{u}(x)}{\partial n}\right]_\Gamma = 0, \quad x \in \Gamma.$$  \hspace{1cm} (2.4c)

where the jump at the interface is defined as

$$\left[\epsilon(x) \frac{\partial \tilde{u}(x)}{\partial n}\right]_\Gamma = \lim_{\alpha \to 0^+} \epsilon(x + \alpha n) \frac{\partial \tilde{u}(x + \alpha n)}{\partial n} - \epsilon(x - \alpha n) \frac{\partial \tilde{u}(x - \alpha n)}{\partial n},$$

with $n$ as the unit normal to the interface $\Gamma$.

The boundary conditions are prescribed using a linear combination of Helmholtz Green's functions,

$$g = \frac{e_c}{k_BT} \sum_{i=1}^{m} \frac{Q_i}{\epsilon_s |x - x_i|} \exp\left(-\frac{\tilde{\kappa}_s |x - x_i|}{\sqrt{\epsilon_s}}\right).$$  \hspace{1cm} (2.5)

In contrast to (2.2a), the principal equation is defined over each subdomain and an interface condition is introduced on $\Gamma$. This restatement makes explicit the normal continuity implied by the strong form divergence of $\epsilon(x)\nabla \phi$ across the interface $\Gamma$ in the original PBE.

We denote the standard Sobolev spaces as $L^2(\Omega)$ and $H^k(\Omega)$, for $k \geq 0$. $H^k(\Omega)$ consists of functions over $\Omega$ having square integrable (weak) derivatives of order up to $k$. The norms on $L^2(\Omega)$ and $H^k(\Omega)$ are expressed as $\| \cdot \|_{0,\Omega}$ and $\| \cdot \|_{k,\Omega}$, with the $L^2(\Omega)$ inner product written $(\cdot, \cdot)_{0,\Omega}$. In addition, we define the Hilbert spaces

$$H(\text{div}; \Omega) := \{ q \in L^2(\Omega)^3 : \nabla \cdot q \in L^2(\Omega)\},$$

$$H_0(\text{div}; \Omega) := \{ q \in H(\text{div}; \Omega) : n \cdot q = 0 \text{ on } \partial \Omega\},$$

$$H_0^1(\Omega) := \{ u \in H^1(\Omega) : u = 0 \text{ on } \partial \Omega\},$$

with norms

$$\| q \|_{\text{div},\Omega}^2 = \| q \|_{0,\Omega}^2 + \| \nabla \cdot q \|_{0,\Omega}^2,$$  \hspace{1cm} (2.6)

$$\| u \|_{1,\Omega}^2 = \| u \|_{0,\Omega}^2 + \| \nabla u \|_{0,\Omega}^2.$$  \hspace{1cm} (2.7)

One difficulty with (2.2a) is regularity. The right-hand side $4\pi \sum_i Q_i \delta(x - x_i)$ is not in $H^{-1}(\Omega)$, i.e., the
dual space of $H^1_0(\Omega)$. Practically, the right-hand side induces singularities in $\tilde{u}$ at the solute atom centers $x_i$. These singularities are the familiar consequence of solute-solute electrostatic interactions satisfying Coulomb’s law. However, finite element and finite difference methods often require more smoothness in order to guarantee convergence. Following [49], we overcome this issue by decomposing $\tilde{u}$ into

$$\tilde{u} = u + u_c,$$

(2.8)

where $u$ is an unknown smooth function and $u_c$ is a known singular function. The Coulomb function, $u_c$, satisfies the Poisson equation

$$-\epsilon_m \nabla \cdot \nabla u_c(x) = \frac{4\pi e_c}{k_B T} \sum_{i=1}^m Q_i \delta(x - x_i),$$

(2.9)

and absorbs the singularities in $\tilde{u}$. Combining (2.8) with (2.4), we obtain the regularized PBE or RPBE

$$-\nabla \cdot \epsilon(x) \nabla u(x) + \tilde{k}^2(x) \sinh(u(x) + u_c(x)) = \nabla \cdot (\epsilon(x) - \epsilon_m) \nabla u_c(x), \quad x \in \Omega_s \cup \Omega_m,$n

$$u(x) = g(x) - u_c(x), \quad x \in \partial \Omega, \quad (2.10)$$

$$\left[ \left( \epsilon(x) \frac{\partial u(x)}{\partial n} \right) \right]_{\Gamma} = \left( \epsilon_m - \epsilon_s \right) \frac{\partial u_c(x)}{\partial n}, \quad x \in \Gamma.$$

Since $\tilde{k}(x)$ and $\epsilon(x) - \epsilon_m$ are zero inside $\Omega_m$, we avoid evaluating the Coulomb potential, $u_c$, near the singularities present at each point charge, $x_i \in \Omega_m$. This yields a right-hand side in (2.10) that is a well-defined distribution in $H^{-1}(\Omega)$ and, as a result, equation (2.10) is a well-defined nonlinear second-order elliptic equation with a unique weak solution $u$ in $H^1(\Omega)$ [49].

A simplified version of (2.10) is the linear regularized Poisson-Boltzmann equation, which is obtained by linearizing the hyperbolic sine:

$$-\nabla \cdot \epsilon(x) \nabla u(x) + \tilde{k}^2(x) u(x) = \nabla \cdot (\epsilon(x) - \epsilon_m) \nabla u_c(x) - \tilde{k}^2(x) u_c(x), \quad x \in \Omega_s \cup \Omega_m,$$

(2.11a)

$$u(x) = g(x) - u_c(x), \quad x \in \partial \Omega, \quad (2.11b)$$

$$\left[ \left( \epsilon(x) \frac{\partial u(x)}{\partial n} \right) \right]_{\Gamma} = \left( \epsilon_m - \epsilon_s \right) \frac{\partial u_c(x)}{\partial n}, \quad x \in \Gamma. \quad (2.11c)$$

Physically, the linearization reduces the ionic response of the solvent to the solute. This approximation is acceptable unless the solute is highly charged [63]. In this case the magnitude of the electrostatic potential is large, and the approximation $\sinh(u) \approx u$ is not accurate [16].
2.3 FOSLS Formulation of PBE

The First-Order System Least Squares (FOSLS) finite element method is an alternative to standard and mixed Galerkin finite element methods [28]. FOSLS begins by converting the PDE to a first order system. Using the new set of equations, a functional is then defined whose minimizer solves the original PDE.

FOSLS offers a number of potential advantages over traditional methods. The functional is minimized using a variational principle, giving rise to a symmetric bilinear form. A discretization based on this form leads to a symmetric positive-definite linear system, which is ideal for solvers such as preconditioned conjugate gradient. Also, the bilinear form is often elliptic with respect to a practical norm, and as a result the finite element spaces do not need to satisfy the discrete inf-sup condition of Ladyzhenskaya-Babuška-Brezzi [38], unlike mixed methods. A practical consequence is that basic finite element spaces, e.g. continuous piecewise linear polynomials, may be used for all variables.

The FOSLS functional also provides a local a posteriori error estimate. Such estimates are complicated for other methods, but the FOSLS residual norm provides a straightforward and accurate estimate for our problem. This local error estimate is used for adaptively refining a mesh in our numerical experiments and we highlight the effectiveness of this tool.

Least-Squares finite element methods are not without limitation, however. The introduction of new variables to formulate the first-order system ultimately increases the degrees of freedom and complexity in computing the solution. This is not necessarily a disadvantage as the new variables are often physically meaningful and are often needed elsewhere in the simulation. For example, the FOSLS formulation of the PBE introduces a secondary “flux” variable, which is used to effectively compute potential of the mean force required in the solution of the Steady-State Smoluchowski Equation [121]. Another potential drawback is that FOSLS requires more regularity than might be present in the problem to ensure optimal error estimates. Optimal error estimates using a Least-Squares approach for PBE can be derived through a multi-domain approach for such problems [42]. In this chapter, we also use adaptive refinement to overcome these issues of computational complexity, yielding optimal convergence rates in our numerical experiments.

A typical approach to forming a first-order system of (2.11) over a single domain Ω is to introduce a flux,
\( \mathbf{\dot{q}} = \epsilon(x) \nabla u \) (e.g., see [107]). The resulting first-order system is

\[
\begin{align*}
\mathbf{\dot{q}} - \epsilon \nabla u &= 0 & \text{in } \Omega, \\
-\nabla \cdot \mathbf{\dot{q}} + \tilde{\kappa}^2 u &= \nabla \cdot (\epsilon(x) - \epsilon_m) \nabla u_c - \tilde{\kappa}^2 u_c & \text{in } \Omega, \\
u &= g - u_c & \text{on } \partial \Omega.
\end{align*}
\]

(2.12a)\hspace{2cm} (2.12b)\hspace{2cm} (2.12c)

An application of Green’s theorem on this system shows that across any surface in \( \Omega \) with normal \( \mathbf{n} \), \( \mathbf{n} \cdot \mathbf{\dot{q}} \) is continuous. In particular, solution to system (2.12) satisfies,

\[
\left[ \mathbf{\dot{q}} \cdot \mathbf{n} \right]_\Gamma = 0 \quad x \in \Gamma.
\]

However, since \( \mathbf{\dot{q}} = \epsilon \nabla u \), equation (2.11c) implies,

\[
\left[ \mathbf{\dot{q}} \cdot \mathbf{n} \right]_\Gamma = (\epsilon_m - \epsilon_s) \nabla u_c(x) \cdot \mathbf{n} \quad x \in \Gamma.
\]

This implies \( \mathbf{\dot{q}} \cdot \mathbf{n} \) is not continuous across the interface \( \Gamma \), and hence, a least squares approach based on system (2.12) is an incorrect formulation for solving the RPBE.

For a well-posed FOSLS formulation to system (2.11), we need to define a first-order variable \( \mathbf{q} \), whose normal component is not only continuous across the interface, but also satisfies the interface condition required by RPBE. To ensure these conditions, we define \( \mathbf{q} = \epsilon(x) \nabla u + (\epsilon(x) - \epsilon_m) \nabla u_c \), which results in,

\[
\begin{align*}
\frac{\mathbf{q}}{\epsilon(x)} - \nabla u &= \left( \left( \epsilon(x) - \epsilon_m \right) / \epsilon(x) \right) \nabla u_c & \text{in } \Omega, \\
-\nabla \cdot \mathbf{q} + \tilde{\kappa}^2 u &= -\tilde{\kappa}^2 u_c & \text{in } \Omega, \\
u &= g - u_c & \text{on } \partial \Omega, \\
\mathbf{n} \times \mathbf{q} &= \mathbf{n} \times \left( \epsilon_s \nabla g + (\epsilon(x) - \epsilon_m) \nabla u_c \right) & \text{on } \partial \Omega.
\end{align*}
\]

(2.13a)\hspace{2cm} (2.13b)\hspace{2cm} (2.13c)\hspace{2cm} (2.13d)

As a result, equations (2.11c) and (2.13) imply,

\[
\left[ \mathbf{q} \cdot \mathbf{n} \right]_\Gamma = 0 \quad x \in \Gamma.
\]

We now pose our problem in abstract form and establish a unique solution. To simplify the analysis we
consider homogeneous Dirichlet boundary conditions. Using a standard lifting argument, we obtain

$$q/\epsilon(x) - \nabla u = ((\epsilon(x) - \epsilon_m)/\epsilon(x))\nabla u_c \quad \text{in } \Omega,$$

$$- \nabla \cdot q + \bar{\kappa}^2 u = -\bar{\kappa}^2 u_c, \quad \text{in } \Omega,$$

$$u = 0 \quad \text{on } \partial\Omega,$$

$$n \times q = 0 \quad \text{on } \partial\Omega.$$ (2.14)

The least-squares functional based on (2.13) is as follows. For $q \in H_0(\text{div}; \Omega)$ and $u \in H^1_0(\Omega)$, we define

$$G(q, u; u_c) = \|q/\epsilon - \nabla u - ((\epsilon(x) - \epsilon_m)/\epsilon(x))\nabla u_c\|_{0,\Omega}^2 + \| - \nabla \cdot q + \bar{\kappa}^2 u + \bar{\kappa}^2 u_c\|_{0,\Omega}^2.$$ (2.15)

The solution of (2.13) solves the minimization problem

$$G(q, u; u_c) = \min_{(r, v) \in H_0(\text{div}; \Omega) \times H^1_0(\Omega)} G(r, v; u_c),$$ (2.16)

and leads to the variational problem

$$\mathcal{F}(q, u; r, v) = \ell(r, v),$$ (2.17)

where the bilinear form $\mathcal{F}$ and linear functional $\ell$ are

$$\mathcal{F}(q, u; r, v) = \langle q/\epsilon - \nabla u, r/\epsilon - \nabla v \rangle_{0,\Omega} + ( - \nabla \cdot q + \bar{\kappa}^2 u, - \nabla \cdot r + \bar{\kappa}^2 v \rangle_{0,\Omega},$$ (2.18)

$$\ell(r, v) = - (\bar{\kappa}^2 u_c, - \nabla \cdot r + \bar{\kappa}^2 v)_{0,\Omega} + ((\epsilon(x) - \epsilon_m)/\epsilon(x))\nabla u_c, \ n/\epsilon - \nabla v \rangle_{0,\Omega}.$$ (2.19)

### 2.3.1 Ellipticity of FOSLS Functional

To show the variational problem (2.17) is well-posed, it is sufficient to prove that $G(q, u; 0)\frac{1}{2}$ defines a norm equivalent to the $H(\text{div}) \times H^1$ norm (Theorem 1). This result also ensures that our finite element solution is the best approximation to the true solution under the norm defined by $G(q, u; 0)$. Before proving this norm equivalence, we start by stating and proving a lemma, which will be used in the proof of Theorem 1.

**Lemma 1.** Let $h(x)$ and $k(x)$ be two positive bounded functions on $\Omega$, i.e. $0 < c_1 < h(x) < c_2$ and $0 < c_1 < k(x) < c_2$ for all $x \in \Omega$, where $c_1$ and $c_2$ are constants. Then there exists positive constants $\alpha_1$ and $\alpha_2$ such that

$$\alpha_1 \tilde{\mathcal{F}}(q, w; q, u) \leq \mathcal{F}(q, w; q, u) \leq \alpha_2 \tilde{\mathcal{F}}(q, w; q, u),$$ (2.20)
where the bilinear form $\hat{F}$ is defined as

$$\hat{F}(q; u; r; v) = (\sqrt{h}(q/\epsilon - \nabla u), \sqrt{h}(r/\epsilon - \nabla v))_{0, \Omega} + (\sqrt{k}(-\nabla \cdot q + \kappa^2 u), \sqrt{k}(-\nabla \cdot r + \kappa^2 v))_{0, \Omega}. \quad (2.21)$$

**Proof of Lemma 1.** Taking $\alpha_1 = c_2^{-1}$ and $\alpha_2 = c_1^{-1}$ gives the desired result. \qed

**Theorem 1.** The bilinear form $F$ defines a norm equivalent to the $H(\text{div}) \times H^1$ norm. That is, there exists positive constants $\gamma_1$ and $\gamma_2$ such that

$$F(q; u; r; v) \leq \gamma_1 (\|q\|^2_{H(\text{div})} + \|u\|^2_{1, \Omega})^{1/2} (\|r\|^2_{H(\text{div})} + \|v\|^2_{1, \Omega})^{1/2} \quad (2.22)$$

and

$$F(q; u; q; u) \geq \gamma_2 (\|q\|^2_{H(\text{div})} + \|u\|^2_{1, \Omega}) \quad (2.23)$$

**Proof.** A proof for the general case is given in [40]. Here we offer a proof for our specific case, to obtain sharper constants of ellipticity; our proof is in the same spirit as a proof presented in [107].

First we prove boundedness of $F$ (equation (2.22)). An application of Cauchy-Bunyakovsky-Schwarz inequality to (2.18) leads to

$$F(q; u; r; v) \leq (F(q; u; q; u))^{1/2} (F(r; v; r; v))^{1/2} \quad (2.24)$$

Using the fact that $\epsilon$ is bounded away from zero in $\Omega$ yields

$$F(q; u; q; u) = \|q/\epsilon - \nabla u\|^2_{0, \Omega} + \| - \nabla \cdot q + \kappa^2 u\|^2_{0, \Omega}$$

$$\leq \gamma_3 (\|q\|^2_{0, \Omega} + \|\nabla u\|^2_{0, \Omega} + \|\nabla \cdot q\|^2_{0, \Omega} + \|u\|^2_{0, \Omega})$$

$$= \gamma_3 (\|q\|^2_{H(\text{div})} + \|u\|^2_{1, \Omega}), \quad (2.25)$$

where $\gamma_3 = \max(2, 2\kappa^4, 2\epsilon^{-2}) = \max(2, 2\kappa^4, 2\epsilon_m^{-2}, 2\epsilon_s^{-2})$. Combining equations (2.25) and (2.24) proves boundedness of $F$.

To prove coercivity, we consider a modified bilinear form, as defined by (2.21). We define $h(x)$ and $k(x)$...
as:
\[
h(x) = \begin{cases} 
\epsilon(x) & x \in \Omega_m \\
\tau \epsilon(x) & x \in \Omega_s 
\end{cases}
\]
and
\[
k(x) = \begin{cases} 
1 & x \in \Omega_m \\
\tau / \tilde{\kappa}^2 & x \in \Omega_s, 
\end{cases}
\]
where \( \tau \) is a constant such that \( 0 < \tau < 1 \).

We can decompose the integral over \( \Omega \) and evaluate \( F \) over \( \Omega_s \) and \( \Omega_m \),
\[
\hat{F}(q, u; q, u) = \hat{F}(q, u; q, u)|_{\Omega_s} + \hat{F}(q, u; q, u)|_{\Omega_m}
\]
(2.27)
where
\[
\hat{F}(q, u; q, u)|_{\Omega_s} = \| \sqrt{\tau} q / \sqrt{\epsilon} - \sqrt{\tau} \nabla u \|^2_{0, \Omega_s} + \| - \sqrt{\tau} / \tilde{\kappa} \cdot q + \sqrt{\tau} \tilde{\kappa} u \|^2_{0, \Omega_s}
\]
(2.28)
and
\[
\hat{F}(q, u; q, u)|_{\Omega_m} = \| q / \sqrt{\epsilon} - \sqrt{\tau} \nabla u \|^2_{0, \Omega_m} + \| \nabla \cdot q \|^2_{0, \Omega_m}.
\]
(2.29)
Integration by parts shows that
\[
\int_{\Omega_m} \nabla \cdot q u + \int_{\Omega_m} q \cdot \nabla u - \int_{\Gamma} u q \cdot n_m = 0,
\]
(2.30)
where \( n_m \) is the unit normal at \( \Gamma \), pointing from the solute region into the solvent region. Applying this result to (2.29), we obtain
\[
\hat{F}(q, u; q, u)|_{\Omega_m} = \| q / \sqrt{\epsilon} \|^2_{0, \Omega_m} + \| \sqrt{\tau} \nabla u \|^2_{0, \Omega_m} - 2 \int_{\Omega_m} q \cdot \nabla u + \| \nabla \cdot q \|^2_{0, \Omega_m} + 2 \tau \int_{\Omega_m} \nabla \cdot q u \\
+ 2 \tau \int_{\Omega_m} q \cdot \nabla u - 2 \tau \int_{\Gamma} u q \cdot n_m + \tau^2 \| u \|^2_{0, \Omega_m} - \tau^2 \| u \|^2_{0, \Omega_m}
\]
\[
= \| q / \sqrt{\epsilon} + (\tau - 1) \sqrt{\tau} \nabla u \|^2_{0, \Omega_m} + \| \nabla \cdot q + \tau u \|^2_{0, \Omega_m} - \tau^2 \| u \|^2_{0, \Omega_m}
\]
\[
+ (2 \tau - \tau^2) \| \sqrt{\tau} \nabla u \|^2_{0, \Omega_m} - 2 \tau \int_{\Gamma} u q \cdot n_m.
\]
(2.31)
Similarly, using integration by parts on equation (2.28) yields

\[
\mathcal{F}(q; u; q; u)|_{\Omega} = \|\sqrt{\tau}q/\sqrt{\tau}u\|^2_{0,\Omega} + \|\sqrt{\tau}\nabla u\|^2_{0,\Omega} - 2\tau \int_{\Omega_s} q \cdot \nabla u + \|\sqrt{\tau}\nabla \cdot q/\kappa_s\|^2_{0,\Omega} + \|\sqrt{\tau}\nabla u\|^2_{0,\Omega} - 2\tau \int_{\Omega_s} \nabla \cdot q u = \|\sqrt{(\tau/\kappa_s)}\nabla \cdot q\|^2_{0,\Omega} + \|\sqrt{\tau}q/\sqrt{\tau}u\|^2_{0,\Omega} + \|\sqrt{\tau}\nabla u\|^2_{0,\Omega} + \|\sqrt{\tau}\kappa_s u\|^2_{0,\Omega} - 2\tau \int_{\Gamma} u q \cdot n_s, \tag{2.32}
\]

where \(n_s = -n_m\) is the unit normal along \(\Gamma\), pointing from the solvent domain into the solute.

Using the Poincaré-Friedrichs inequality, we can assume

\[
\|u\|^2_{0,\Omega} \leq \lambda \|\nabla u\|^2_{0,\Omega}, \quad \text{with} \quad \lambda > 1. \tag{2.33}
\]

From equations (2.27), (2.31), (2.32), (2.33) and choosing \(\tau = \frac{1}{4\lambda} < 1\) we have

\[
\mathcal{F}(q; u; q; u) = \|\sqrt{\tau}q/\sqrt{\tau}u\|^2_{0,\Omega} + \|\sqrt{\tau}\nabla u\|^2_{0,\Omega} + \|\sqrt{\tau}\nabla \cdot q/\kappa_s\|^2_{0,\Omega} + \|\sqrt{\tau}\kappa_s u\|^2_{0,\Omega} + \|\sqrt{\tau}\nabla u\|^2_{0,\Omega} + \|\sqrt{\tau}\nabla \cdot q\|^2_{0,\Omega} + \|\sqrt{\tau}\kappa_s u\|^2_{0,\Omega}
\geq \|\sqrt{\tau}\nabla u\|^2_{0,\Omega} + (2\tau - \tau^2)\|\sqrt{\tau}\kappa_s u\|^2_{0,\Omega} + (2\tau - \tau^2)\|\sqrt{\tau}\nabla u\|^2_{0,\Omega}
\geq \tau \|\sqrt{\tau}\nabla u\|^2_{0,\Omega} - \tau^2\|\nabla u\|^2_{0,\Omega} \geq (\tau - \lambda \tau^2)\|\sqrt{\tau}\nabla u\|^2_{0,\Omega}
\geq \frac{1}{4\lambda} \|\sqrt{\tau}\nabla u\|^2_{0,\Omega} \geq 4\lambda \|\nabla u\|^2_{0,\Omega}, \tag{2.34}
\]

where \(\alpha_3 = \frac{1}{4\lambda} \min(\epsilon_m, \epsilon_s)\).

Now from equation (2.20), we get

\[
\mathcal{F}(q; u; q; u) \geq \alpha_4 \|\nabla u\|^2_{0,\Omega}, \tag{2.35}
\]

where \(\alpha_4 = \alpha_1 \alpha_3\). From the Poincaré-Friedrichs inequality (2.33), we find

\[
\mathcal{F}(q; u; q; u) \geq \alpha_5 \|u\|^2_{0,\Omega}. \tag{2.36}
\]

Moreover,

\[
\|q/\epsilon\|^2_{0,\Omega} \leq 2\|q/\epsilon - \nabla u\|^2_{0,\Omega} + \|\nabla u\|^2_{0,\Omega} \leq 2 \left(1 + \frac{1}{\alpha_4}\right) \mathcal{F}(q; u; q; u),
\]
and hence $F(q, u; q, u) \geq \alpha_6 \|q\|_{0, \Omega}^2$ for $\alpha_6 = \alpha_4 \left[2 (1 + \alpha_4) \max(\epsilon_s, \epsilon_m) \right]^{-1}$. Similarly,

$$\|\nabla \cdot q\|_{0, \Omega}^2 \leq 2 (\|\nabla \cdot q - \kappa^2 u\|_{0, \Omega}^2 + \|\nabla^2 u\|_{0, \Omega}^2) \leq 2 (1 + \kappa^4 \alpha_5) F(q, u; q, u),$$

(2.37)

and thus $F(q, u; q, u) \geq \alpha_7 \|\nabla \cdot q\|_{0, \Omega}^2$ for $\alpha_7 = \left[2 (1 + \kappa^4 \alpha_5) \right]^{-1}$.

Taking $\gamma_2 = \min(\alpha_4, \alpha_5, \alpha_6, \alpha_7)$ completes the proof.

The FOSLS functional (2.15) is $H(div) \times H^1$ equivalent. In some FOSLS formulations, a curl term of the form $\nabla \times (q/\epsilon) = 0$ is added to problem formulation (e.g., [107]), yielding a $H^1 \times H(div) \cap H(curl)$ equivalent FOSLS functional. The extra constraint is motivated by $\tilde{q} = \epsilon \nabla u$, which implies $\nabla \times \tilde{q}/\epsilon = 0$ (c.f., [72], Theorem 2.9). However, for our case, we cannot take the curl of $q/\epsilon$. This follows from our definition of $q = \epsilon \nabla u + (\epsilon - \epsilon_m) \nabla u_c$; the curl of $q/\epsilon$ is undefined at the interface. Hence we do not add the curl term to the formulation.

Traditionally, developing an effective error estimator for use in local adaptive refinement is challenging. Error estimators based on the Galerkin method are not immediately obvious from the problem formulation and local error bounds for the PBE can be complicated to derive [49]. In contrast, the FOSLS framework directly provides a natural error indicator through the functional. The local value of FOSLS functional is an \textit{a posteriori} lower error bound, and, under some restrictions on mesh refinement, the bound can be shown to be a sharp theoretical error estimate [23]. We exploit this fact and build an adaptive refinement scheme based on the value of the FOSLS functional.

Let $G_\tau(q, u; u_c)$ be the value of the FOSLS functional (2.15) restricted to element $\tau$. Note that if $S$ is the set of elements comprising the mesh, then

$$G(q, u; u_c) = \sum_{\tau \in S} G_\tau(q, u; u_c).$$

Let $\mu_\tau = \sqrt{G_\tau(q, u; u_c)}$ and $\mu_{\text{max}} = \max_{\tau \in S} \mu_\tau$. We mark simplex $\tau$ for refinement if $\mu_\tau \geq \gamma \mu_{\text{max}}$ where $\gamma \in (0, 1)$.

Our strategy is relatively straightforward, yet more advanced marking strategies based on the “solvation free-energy” [54] and FOSLS [23, 39] functionals have been proposed in the literature. However, in our numerical experiments, we did not find a significant difference in performance when the marking strategy is varied for our problem. When compared on the same mesh, FOSLS requires more memory and CPU time than the standard second-order Galerkin method. However, the meshes produced by the corresponding
adaptive refinement schemes are different, and the FOSLS approach is often able to achieve a more accurate solution with less refinement. As a result, the FOSLS approach is often more efficient than a standard second-order Galerkin method. The effectiveness of our scheme is highlighted in Section 2.4.

2.4 Numerical Experiments

We use a tetrahedral mesh of \( \Omega \) with globally continuous piecewise linear finite functions (P1 elements) and implement our finite element method and mesh refinement in FETK [76]. The meshes are generated using the Geometry-preserving Adaptive Mesher (GAMer), which is designed to produce simplicial meshes of molecular volumes and interfaces [136]. As a result, the solvent domain has a spherical outer boundary and the mesh is conforming at the interface of the solvent and molecule regions. For the first four numerical experiments, we choose \( \epsilon_m = 1, \epsilon_s = 78, \) and \( \bar{\kappa}_s = 0.918168 \), which corresponds to a typical ionic strength of 0.1M. In these experiments, we solve for the regularized potential and strongly impose boundary conditions. The experiments are performed on the Born ion, Fasciculin 1, methanol, and a simple dipole. Let \( q^h \) and \( u^h \) be our finite-element solution, and \( q \) and \( u \) the true solution. We verify convergence to the solution by monitoring the square-root of FOSLS functional, \( G(q^h, u^h; u_c)^{1/2} \), since the FOSLS functional measures the error in the norm induced by \( G \): \( G(q^h, u^h; u_c)^{1/2} = G(q^h - q, u^h - u; 0)^{1/2} \). Therefore, convergence of the FOSLS functional to zero implies convergence of our finite element solution to the true solution. We use uniform octal refinement and adaptive refinement to test the effectiveness of your method, with adaptive refinement being carried out by longest edge bisection. Since \( G(.,.,0)^{1/2} \) is equivalent to \( H(\text{div}) \times H^1 \) norm, a standard finite-element error estimate implies optimal convergence rate to be \( O(h) \) using uniform refinement with piecewise linear basis functions [40]. This optimal estimate assumes the problem to be \( H^2 \) regular. Consequently, convergence rate degrades as the solution becomes less smooth. We examine this scenario (dipole), and show that we still recover optimal convergence using adaptive refinement. In the following results, we refer to \( G(q^h, u^h; u_c)^{1/2} \) as the FOSLS norm and plot convergence rates normalized by the largest value. Finally, to validate the solutions generated by our implementation, we compute the solvation free energy of transcription factor PML (PDB code 1BOR). We compare the computed value with values found in the literature.
Born Ion

Due to the complex geometries associated with molecules there are few analytical solutions to the PBE or linearized PBE; however, it is possible to find an expression for the potential of a spherical ion in a solvent [77]. This system is referred to as the Born ion after its author Max Born [33]. The domain consists of a spherical solute of radius $R$ with a single point charge $Q_1$ at its center. The solute is surrounded by an unbounded solvent, $\Omega_s$, as depicted in Figure 2.2a.

Writing the linear regularized PBE in spherical coordinates yields

$$\frac{1}{r^2} \frac{d}{dr} \left( \epsilon(r) r^2 \frac{d}{dr} u(r) \right) + \bar{\kappa}^2(r) u(r) = -\bar{\kappa}^2(r) u_c(r), \quad r \neq R,$$

$$\left[ \epsilon(r) \frac{d}{dr} u(r) \right]_r = (\epsilon_m - \epsilon_s) \frac{d}{dr} u_c(r), \quad r = R,$$

$$u(\infty) = 0,$$

where $\omega = \bar{\kappa}_s / \sqrt{\epsilon_s}$. Following [77], we obtain the analytic solution

$$u(r) = \begin{cases} C_1 \exp \left[ -\omega (r - R) \right] / r - C_2 / r, & R \leq r, \\ (C_1 - C_2)/R, & 0 \leq r < R, \end{cases}$$

where

$$C_1 = \frac{e_c Q_1}{k_B T (1 + \omega R) \epsilon_s}, \quad \text{and} \quad C_2 = \frac{e_c Q_1}{k_B T \epsilon_m}.$$

\textbf{Figure 2.2:} Molecular description and convergence of solution for the Born Ion
Figure 2.3: Convergence Rates in the $L^2$ and FOSLS norm for the Born Ion

Figure 2.3a displays the convergence of the reaction potential $u$ in the $L^2$ norm, where the normalized $L^2$ error is plotted as a function of $N$, the number of points in the mesh. In three-dimensions, we observe a convergence rate of nearly $O(h^2)$ for uniform refinement, which corresponds to $O(N^{-2/3})$. On the other hand, for adaptive refinement, we observe a slightly better convergence rate. Figure 2.3b displays the FOSLS functional residual as the mesh is refined. In three-dimensions, a convergence rate of $O(h)$ corresponds to $O(N^{-1/3})$. We see that the FOSLS functional decreases nearly linear in $h$. During refinement, we ensure that new points on the solute/solvent interface lie on the analytically determined spherical boundary of the interface. As an example of convergence, in Figure 2.2b we display a slice of the true solution, a numerical solution on the initial mesh, and numerical solutions after two successive steps of uniform mesh refinement.

Fasciculin 1

The Born ion is a useful test case as the analytical solution is known; however, it is not a realistic simulation. To study the effectiveness of the FOSLS formulation on a realistic protein, we compute the regularized potential of Fasciculin 1 (1FAS in the Protein Data Bank) in an implicit solvent. 1FAS is a neurotoxin found in green mamba venom [94]. The dynamics and electrostatics of the Fasciculin 2 variant of this protein in its role as an acetylcholinesterase inhibitor have been studied in [113] and [13], where the electrostatics are argued to be important to its function. In our experiments we use the description of the molecule specified in the PDB file from the Protein Data Bank and strip off water molecules using VMD [82]. The molecule
region is not perfectly spherical, and we do not expect the solution to be symmetric as we did in the case of Born ion. It is assumed that the initial mesh defines the solute and solvent regions so that the solute/solvent interface in this case is polygonal and defined by the initial mesh. Consequently, refinement adds points to the polygonal interface. While the analytical solution for Fasciculin 1 is not known, we are able to monitor the convergence of FOSLS functional.

Figure 2.4 shows the normalized convergence rate of FOSLS functional. Both uniform and adaptive refinement perform well: the convergence rate is better than $O(h)$ for both cases. Figure 2.5 depicts adaptive refinement around the Fasciculin molecule. The adaptive scheme refines aggressively around the areas where the solution is changing sharply.

\[ \text{Figure 2.4: Convergence in FOSLS norm for Fasciculin 1} \]

**Methanol**

We examine our method in the more challenging setting of a methanol molecule, obtained from the APBS software package [15]. The model consists of three charged spheres representing charge groups: CH$_3$ and H with positive charges of 0.27 and 0.43 respectively, and an O atom with a negative charge of 0.7. The net charge on the molecule is zero. Figure 2.6a displays the methanol molecule.

We assume again that the initial mesh properly defines the solute and solvent regions. Figure 2.6b displays the FOSLS functional as a function of the number of vertices in the mesh. We see from the plot that the FOSLS functional does decrease, but the convergence is slightly less than $O(h)$. On the other hand, adaptive refinement is ideal for this problem since the solution varies sharply across the interface, indicating areas
where local refinement is useful. As Figure 2.6b shows, adaptive refinement yields slightly better than $O(h)$ convergence. The performance of adaptive refinement is shown in Figures 2.7 and 2.8, where the regularized electrostatic potential around the interface is displayed. Figure 2.7 shows the initial mesh and an adaptively refined mesh. Figure 2.8 displays a slice of the regularized solution, which highlights the areas in which the solution changes rapidly and also that the solution is not symmetric.

**Figure 2.5:** Adaptive refinement around the solute/solvent interface of Fasciculin 1

**Figure 2.6:** Molecular model and convergence for Methanol
Figure 2.7: Adaptive refinement around the solute/solvent interface of methanol

Figure 2.8: Methanol: Solution around the interface
Dipole

In this section we illustrate the performance of our scheme on a simple dipole, as depicted in Figure 2.9. The linearized PBE for ions inside a spherical molecular region has been studied in [89]. For our experiment the domain consists of a spherical molecular region of radius 2 units, with two equal, but opposite unit charges, $q^+$ and $q^-$, inside. The charges are placed on opposite sides of the x-axis, each at distance $d$ from the origin (see Figure 2.9). As $d$ is increased, the charges move closer to the interface, the solution becomes less well-behaved, developing a sharp gradient at the interface. Uniform refinement does not efficiently resolve the solution in this scenario. However, adaptive refinement is able to refine locally around the simplices at the interface, and gives a significantly better convergence rate than uniform refinement as shown in Figure 2.10. In particular, the rate of convergence for adaptive refinement is nearly insensitive to changes in the parameter $d$.

![Figure 2.9: Domain for a simple dipole.](image)

1BOR

Finally, we compute the electrostatic solvation free energy of transcription factor PML (PDB code 1BOR), and compare our value with the results in [68], where they choose $\epsilon_m = 1$, $\epsilon_s = 80$. The electrostatic free energy of solvation is defined by [68]

$$\Delta G_{sol} = \frac{1}{2} \sum_{j=1}^{n_s} Q_j (\phi(x_j) - \phi_{homo}(x_j)),$$

(2.38)

where $\phi_{homo}$ is the solution of (2.2a) in homogeneous environment, that is $\epsilon_m = \epsilon_s = 1$. In terms of the regularized potential $u$, the solvation free energy can be computed as,

$$\Delta G_{sol} = \frac{1}{2} k_B T \sum_{j=1}^{n_s} Q_j u(x_j).$$

(2.39)
On a mesh with 131086 vertices, we compute the free energy of solvation equal to -792.577 kcal/mol, which compares well with the value of -853.7 kcal/mol computed from the MIBPB-III method in [68]. The difference in the computed values is explained by the free energy of solvation is sensitive to the geometry of the protein surface. We use GAMer to define this interface geometry, and hence our result does not exactly match up with [68], who use MSMS [118] to generate their protein surface.

2.5 Conclusion

The interface jump condition in (2.11) presents a challenge to design a single-domain FOSLS approach. We overcome this difficulty with a choice of a vector parameter $\mathbf{q}$ that results in a consistent and well-posed first-order system. The approach is also useful for solving the non-linear equation using a Newton-FOSLS method [50], as each step of Newton’s method will effectively involve solution of a linearized Poisson-Boltzmann equation. In this chapter we show that the resulting FOSLS functional defines a norm equivalent to the norm on $H^1 \times H(\text{div})$, yet can be used in an existing finite element framework that uses more standard piecewise continuous elements in $H^1$.

We offer numerical evidence in support of this approach and test the methodology on several problems. We observe that adaptive refinement based on the FOSLS functional scheme yields a faster convergence rate than uniform refinement, and that this effect is more pronounced for solutions that are more sharply varying.
This chapter further explores issues related to the numerical efficacy of least-squares finite element methods as an alternative to the traditional Galerkin finite element approach. In particular, we highlight the utility of a first-order form, noting optimality, control of the flux variables, and flexibility in the formulation, including the choice of elements. We also explore the impact of weighting and the choice of elements on conditioning and adaptive refinement. In a series of numerical experiments, we compare the finite element methods when applied to the problem of computing the solvation free energy for realistic molecules of varying size.

A least-squares finite element method for the PBE in is proposed in Chapter 2, where it is shown that the least-squares functional is $H(\text{div}) \times H^1$ equivalent, yet can be used in an existing finite element scheme that uses (more standard) piecewise continuous elements in $H^1$. In numerical experiments, an adaptive refinement scheme based on the least-squares functional is shown to yield a faster convergence rate (compared to uniform refinement), especially for solutions with reduced regularity.

In this chapter we investigate the use of $H(\text{div})$ conforming finite elements and propose a weighted least-squares formulation that leads to optimal convergence of the both the concentration and its gradient field. Also, in comparison to standard Galerkin, we see that our LSFEM approach is competitive. We find that our weighted least-squares method results in a balanced convergence of terms in the finite element approximation, leads to improved conditioning in the resulting linear systems, and yields a more accurate calculation of solvation energy in comparison to other finite element methods.

The remainder of the chapter is organized as follows. To support our finite element formulation we provide an overview of common finite element spaces in Section 3.2. In Section 3.3, we compare two least-squares formations: one using piecewise linear elements, and one which uses Raviart–Thomas elements. In Section 3.4, we review the least-squares formulation of the regularized PBE from the proceeding chapter, and investigate alternate weightings of the functional. We follow this discussion with an investigation of the relative advantages of this approach over a standard Galerkin formulation in Section 3.5. In Section 3.6, we consider the two-domain formulation of the problem, which includes an interface condition. Finally, we
present our conclusions in Section 3.7.

3.1 Least-squares Finite Element Formulation

A LSFEM for the PBE is described in detail in the preceding chapter. Here, we recall the first-order system,

\[
\begin{align*}
\frac{q}{\epsilon(x)} - \nabla u &= (\epsilon(x) - \epsilon_m)/\epsilon(x))\nabla u_c, & \text{in } \Omega, \\
-\nabla \cdot q + \tilde{\kappa}(x)^2 u &= -\tilde{\kappa}(x)^2 u_c, & \text{in } \Omega, \\
u &= g - u_c, & \text{on } \partial\Omega, \\
\mathbf{n} \times q &= \mathbf{n} \times (\epsilon_s \nabla g + (\epsilon(x) - \epsilon_m)\nabla u_c) & \text{on } \partial\Omega.
\end{align*}
\]

(3.1a, 3.1b, 3.1c, 3.1d)

We showed that the least-squares problems is well-posed in the Sobolev space \( H(\text{div}) \times H^1 \), since the least-squares functional arising from this first-order system is \( H(\text{div}) \times H^1 \) equivalent.

3.2 Description of FEM Spaces

Since the least-squares functional developed in the preceding chapter is \( H(\text{div}) \times H^1 \) equivalent, Cea’s Lemma ensures the best approximation in any finite-dimensional subspace of \( H(\text{div}) \times H^1 \). Here we consider spaces of piecewise polynomials on a tetrahedral mesh and denote the spaces corresponding to the discrete solutions \( q_h \) and \( u_h \) as \( m \)-dimensional \( W^h \) and \( n \)-dimensional \( V^h \) respectively, with \( W^h \subset H(\text{div}) \) and \( V^h \subset H^1 \).

Further, we denote basis functions of \( W^h \) by \( \psi(x) \) and those of \( V^h \) by \( \phi(x) \). Then any function in these subspaces is represented as

\[
q_h = \sum_{i=1}^{m} q_i \psi_i, \quad u_h = \sum_{i=1}^{n} u_i \phi_i.
\]

(3.2)

A common choice for \( H^1 \) subspaces are globally continuous piecewise linear functions — i.e. the restriction of such functions on any tetrahedron \( T \) belongs to \( \mathbb{P}_1(T) \), where \( \mathbb{P}_1(T) \) is the space of degree one polynomials. This is the choice we used in our numerical examples in the previous chapter. The relationship between spaces \( H(\text{div}), H(\text{curl}), \) and \( (H^1)^3 \) is illustrated in Figure 3.1a. Since \( (H^1)^3 \subset H(\text{div}) \), a subspace of \( (H^1)^3 \) is a valid representation of vector valued functions in \( H(\text{div}) \). Thus, representing each vector component as a continuous piecewise linear function (a 12-dimensional vector field over the tetrahedron) is valid mathematically. These elements are also attractive because degrees-of-freedom directly represent the quantities-of-interest and because they are widely available in many software packages.
While continuous $P_1^1$ elements are straightforward, one popular alternative are the vector-valued Raviart–Thomas elements [115, 109]. A vector-valued function belongs to $H(\text{div})$ only if its normal component is continuous across any surface in $\Omega$ and Raviart–Thomas elements are designed to satisfy this requirement by imposing normal continuity on the elements. That is,

$$RT_k := \{ \zeta \in L^2(\Omega)^3; \zeta|_T = c+kx, c \in P_k^3, k \in P_k, x \in T, \zeta \cdot n \text{ is continuous across neighboring faces} \}. \tag{3.3}$$

The set of lowest order Raviart–Thomas elements, $RT_0$, is a four dimensional subspace of $P_1^3(T)$ where each element is of the form $p = c + kx$ with $c \in \mathbb{R}^3, k \in \mathbb{R}$. The normal component $p \cdot n$ is constant on each tetrahedron face and identifies each degree-of-freedom. That is, zeroth order moments of $p \cdot n$, $\int_{\text{face}} p \cdot n \, d\gamma$, define the degrees-of-freedom. Raviart–Thomas elements are not tangentially continuous across the element boundary, and hence contain discontinuous functions; they impose only the continuity required by $H(\text{div})$. Figure 3.1b illustrates the degrees-of-freedom on a tetrahedron for both $P_1$ and $RT_0$ spaces.

### 3.3 On the use of Piecewise Linear and Raviart–Thomas Elements

In this section we consider the merits of $P_1$ and $RT_0$ elements for representing the flux $q$. We consider a tetrahedral mesh of $\Omega$ generated using the Geometry-preserving Adaptive Mesher (GAMer), which is designed to produce simplicial meshes of molecular volumes and interfaces [136].

We showed good convergence properties of the least-squares functional in the preceding chapter using $P_1$ elements. However the problem was shown to be $H(\text{div}) \times H^1$ equivalent, and $P_1$ elements cannot converge to the continuum solution if the flux $q$ is in $H(\text{div})$, as shown in [29]. Hence, for robustness of the numerical
method, we advocate the use of $H(\text{div})$ conforming elements for solving the PBE. Next we examine some experimental results for the $P_1$ and $RT_0$ elements.

In each experiment, we solve for the regularized potential $u$ and flux $q$, while strongly imposing the Dirichlet boundary conditions onto the finite element space. The first test case is that of a Born ion, since an analytical solution to the linearized form of the PBE is available, as we showed in the previous chapter. For a more realistic and non-symmetric geometry, we examine in a second test case of a methanol molecule [15]. The methanol domain, shown in Figure 3.2a, is constructed from a model with three ions of differing radii: two with positive charges, and one with a negative charge, for a total net charge of zero.

Finally, to compare the methods on more realistic and complex geometries, we perform a sequence of experiments with four molecules obtained from the Protein Data Bank [22]: 1bor, 1vii, 1sh1 and 1cbn. Figure 3.2b shows the structure of the 1sh1 molecule, surrounded by a slice of the molecular surface (the complete surface is shown in Figure 3.3a). Figure 3.3b is an enlarged view of the irregular and complicated mesh inside the molecular region (gray), and a small part of the surrounding solvent region (white). The other three remaining molecules have similarly complex geometries.

For the Born ion and methanol, we set parameters $\epsilon_m = 1$, $\epsilon_s = 78$, and $\bar{\kappa}_s = 0.918168$, which corresponds to an ionic strength of $0.1 \text{ M}$. These values are consistent with those used in [32]. For the other four molecules, we set $\epsilon_s = 80$.

![Figure 3.2: Schematic of problem domains for Methanol and 1sh1](image-url)
Convergence in the $L^2$ Norm

The convergence rates presented in Table 3.1 highlight the optimality of the LSFEM for the problem, which results in $O(h^2)$ convergence in the scalar potential $u$ for both $P_1$ and $RT_0$.

![Mesh along a 2-D slice of 1sh1 molecule](image)

**Figure 3.3:** Geometry of the 1sh1 molecule

<table>
<thead>
<tr>
<th>$h/h_0$</th>
<th>$P_1$ Value</th>
<th>$P_1$ Rate</th>
<th>$RT_0$ Value</th>
<th>$RT_0$ Rate</th>
<th>$P_1$ Value</th>
<th>$P_1$ Rate</th>
<th>$RT_0$ Value</th>
<th>$RT_0$ Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>284.4</td>
<td>303.5</td>
<td>383.3</td>
<td>378.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/2</td>
<td>81.6</td>
<td>1.80</td>
<td>84.0</td>
<td>1.85</td>
<td>189.5</td>
<td>1.02</td>
<td>189.0</td>
<td>1.00</td>
</tr>
<tr>
<td>1/4</td>
<td>21.5</td>
<td>1.92</td>
<td>21.8</td>
<td>1.94</td>
<td>94.4</td>
<td>1.01</td>
<td>94.3</td>
<td>1.00</td>
</tr>
<tr>
<td>1/8</td>
<td>5.4</td>
<td>1.99</td>
<td>5.5</td>
<td>1.99</td>
<td>46.9</td>
<td>1.01</td>
<td>46.9</td>
<td>1.01</td>
</tr>
</tbody>
</table>

**Table 3.1:** Convergence in $L^2$ and least-squares norms for the Born Ion

Convergence in the Least-Squares Norm

Since the least-squares functional is an error estimator, we also monitor convergence in the least-squares norm. Specifically, the least-squares functional measures the error in the norm $\|q^h, u^h\|_{LS} = G(q^h - q, u^h - u; 0)^{1/2}$.

Since we are using linear polynomials and the lowest-order $RT_0$ elements, the optimal convergence in an $H(\text{div}) \times H^1$ equivalent norm is expected to be $O(h)$. The convergence for the Born ion and methanol are summarized in Tables 3.1 and 3.2, which shows near optimal performance in both cases. In the tables, $h_0$ represents the mesh size parameter for the coarsest mesh. We observe that using $RT_0$ elements yields a
slightly more accurate solution for methanol.

\[
\begin{array}{cccc}
  \frac{h}{h_0} & \mathbb{P}_1 & \text{Value} & \text{Rate} & RT_0 & \text{Value} & \text{Rate} \\
 1 & 228.0 & 211.7 \\
 1/2 & 142.8 & 0.68 & 121.1 & 0.81 \\
 1/4 & 82.2 & 0.80 & 63.9 & 0.92 \\
 1/8 & 45.4 & 0.86 & 32.6 & 0.97 \\
\end{array}
\]

**Table 3.2:** Convergence in the least-squares norm for Methanol

For the other four molecules, the initial mesh generated by GAMer is fine enough that uniform octasection refinement is only feasible for a couple of mesh refinements. For this reason, we use uniform (longest edge) bisection refinement for these molecules. In computing convergence rates, we use the average value of the radius of the elemental circumspheres, \( h \). Table 3.3 shows the convergence rate for these four molecules. Refinement level indicates how many times the mesh has been refined, where level 0 is the initial mesh. Once again we see that \( RT_0 \) elements are more accurate for coarser meshes.

\[
\begin{array}{cccccccc}
  \text{Refinement Level} & 1\text{bor} & 1\text{vii} & 1\text{sh1} & 1\text{cbn} \\
 1 & 0.96 & 1.92 & 0.82 & 0.86 & 0.97 & 0.97 & 0.84 & 0.77 \\
 2 & 0.75 & 1.24 & 1.10 & 1.20 & 0.66 & 0.77 & 0.60 & 0.74 \\
 3 & 0.78 & 0.95 & 1.17 & 1.17 & 0.75 & 0.77 & 0.69 & 0.71 \\
\end{array}
\]

**Table 3.3:** Convergence Rates in the least-squares norm for 1bor, 1vii, 1sh1 and 1cbn

The slightly higher accuracy of \( RT_0 \) elements is not surprising since we are looking for a solution in \( H(\text{div}) \) and \( RT_0 \) elements approximate \( H(\text{div}) \) better than \( \mathbb{P}_1 \) by allowing inter-element jumps in the flux function \( \mathbf{q} \). \( RT_0 \) elements offer a more flexible space for minimization of the least-squares functional with fewer restrictions on continuity than \( \mathbb{P}_1 \).

**Cost of Solving the Discrete Problems**

We also need to take into account cost of the solving the discrete problems. The cost is indicated by the two properties, the number of non-zeros in the matrix, and its condition number. The number of non-zeros which is an indicator of the efficiency of a matrix-vector multiply in iterative methods, is nearly identical for both elements as shown in Tables 3.4 and 3.5.
One attractive quality of a least-squares finite element approximation is that the resulting matrix problems are symmetric and positive definite. Yet, if the conditioning of this linear system is poor, it leads to decreased overall performance. In Table 3.6, we report the condition number for the coarsest mesh for $P_1$ and $RT_0$ approximations; the condition number for a system arising from a LSFEM is known to be $O(h^{-2})$ [40]. The $P_1$ approximation yields systems with improved conditioning compared to $RT_0$, as mentioned in [29]. However, there exist preconditioners that are designed specifically for face elements [6, 75], hence this is not a significant factor in deciding which elements to use, as argued in [29]. From now on, unless we specify otherwise, we will use $RT_0$ elements for LSFEM.

### Table 3.4: Number of non-zeros for $RT_0$ versus $P_1$

<table>
<thead>
<tr>
<th>Refinement Level</th>
<th>1bor</th>
<th>1vii</th>
<th>1sh1</th>
<th>1cbn</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>1</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>2</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>3</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Table 3.5: Ratios of non-zeros of $RT_0$ and $P_1$ for 1bor, 1vii, 1sh1 and 1cbn

### Table 3.6: Conditioning of Linear Systems

<table>
<thead>
<tr>
<th></th>
<th>Born Ion</th>
<th>Methanol</th>
<th>1bor</th>
<th>1vii</th>
<th>1sh1</th>
<th>1cbn</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$</td>
<td>1.07E006</td>
<td>1.26E006</td>
<td>1.29E+008</td>
<td>1.30E+008</td>
<td>1.06E+008</td>
<td>8.79E+007</td>
</tr>
<tr>
<td>$RT_0$</td>
<td>8.07E009</td>
<td>8.66E009</td>
<td>5.05E+013</td>
<td>2.15E+015</td>
<td>5.22E+013</td>
<td>3.18E+013</td>
</tr>
</tbody>
</table>

### 3.4 Comparison of Different Weightings of the LSFEM Functional

In this section we investigate how different weights on the two terms of the functional affects the properties of the finite element solution. To motivate this section, we write (3.1) in a new form as,
\[
\alpha(x)(q - \epsilon(x)\nabla u) = \alpha(x)((\epsilon(x) - \epsilon_m))\nabla u_c, \quad \text{in } \Omega, \tag{3.4a}
\]
\[-\nabla \cdot q + \kappa(x)^2 u = -\kappa(x)^2 u_c, \quad \text{in } \Omega. \tag{3.4b}
\]

This leads to the LSFEM functional,

\[
G(q, u; u_c) = \|\alpha q - \alpha \epsilon \nabla u - \alpha((\epsilon - \epsilon_m))\nabla u_c\|_{0, \Omega}^2 + \| - \nabla \cdot q + \kappa^2 u + \kappa^2 u_c\|_{0, \Omega}^2. \tag{3.5}
\]

Here the first equation in (3.4) (or alternatively the first term of the LSFEM functional (6.121)), is weighted by a factor \(\alpha(x)\). If \(\alpha(x) = 1/\epsilon(x)\), we equation (3.1).

The primary motivation for choosing \(\alpha(x) = 1/\epsilon(x)\) in the preceding chapter was the observation that \(\epsilon(x) \geq 1\), which implies that \(\alpha(x) \leq 1\), and so the term in the LSFEM functional corresponding to (3.4a) is given a decreased weight. This will have the effect of minimizing (3.4b) more, and hence we expect to get a more accurate potential \(u\). However, we expect this to lead to a less accurate resolution of the flux \(q\).

Another weighting we investigate is to use \(\alpha(x) = 1/\sqrt{\epsilon}\). This weighs the flux term (3.4a) more, so we expect a more accurate \(q\) from this weighting, while \(u\) may not be as accurate. These observations are borne out for uniform refinement for the Born ion in Table 3.7.

<table>
<thead>
<tr>
<th>(h/h_0)</th>
<th>(L_2) error in (u^h) (\alpha = 1/\epsilon) (\alpha = 1/\sqrt{\epsilon})</th>
<th>(L_2) error in (q^h) (\alpha = 1/\epsilon) (\alpha = 1/\sqrt{\epsilon})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>303.5 362.5</td>
<td>881.5 233.6</td>
</tr>
<tr>
<td>1/2</td>
<td>84.0 98.8</td>
<td>225.0 63.2</td>
</tr>
<tr>
<td>1/4</td>
<td>21.8 25.3</td>
<td>57.0 17.6</td>
</tr>
<tr>
<td>1/8</td>
<td>5.4 6.3</td>
<td>14.7 5.7</td>
</tr>
</tbody>
</table>

Table 3.7: Comparison of different weighting schemes for the Born Ion

Next we examine the conditioning of the linear systems due to the different weightings. As we have noted earlier, the conditioning of the linear systems can greatly effect the performance of the linear solve. As before, we examine the condition number at the coarsest mesh level, for the Born ion and for methanol. The condition numbers are approximated by estimating the ratio of the largest and the smallest eigenvalue of linear system using Lanczos method. The results, shown in Table 3.8 show that the weight \(\alpha = 1/\sqrt{\epsilon}\) leads to better conditioned systems. This advantage of the weighting significantly increases the performance of
linear solvers.

<table>
<thead>
<tr>
<th></th>
<th>Born Ion</th>
<th>Methanol</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha = 1/\epsilon$</td>
<td>8.07E+09</td>
<td>8.66E+09</td>
</tr>
<tr>
<td>$\alpha = 1/\sqrt{\epsilon}$</td>
<td>1.05E+08</td>
<td>1.12E+08</td>
</tr>
</tbody>
</table>

**Table 3.8:** Conditioning of Linear Systems for different weighting

Finally we examine how the conditioning varies as the ratio $\epsilon_s/\epsilon_m$ is changed, for the weightings $\alpha = 1, \alpha = 1/\sqrt{\epsilon}$ and $\alpha = 1/\epsilon$. We vary $\epsilon_s$ and keep $\epsilon_m = 1$ constant, for the Born ion for the three weightings, in table 3.9. This shows, that for the range of $\epsilon_s$ considered in table 3.9, $1/\sqrt{\epsilon}$ is the optimal choice. From now on in this chapter, we will use this weighting $\alpha = 1/\sqrt{\epsilon}$, unless we specify otherwise.

<table>
<thead>
<tr>
<th>$\epsilon_s/\epsilon_m$</th>
<th>7.8</th>
<th>78</th>
<th>780</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha = 1$</td>
<td>1.57E+06</td>
<td>5.84E+07</td>
<td>6.11E+09</td>
</tr>
<tr>
<td>$\alpha = 1/\epsilon$</td>
<td>8.24E+07</td>
<td>8.07E+09</td>
<td>8.06E+11</td>
</tr>
<tr>
<td>$\alpha = 1/\sqrt{\epsilon}$</td>
<td>1.10E+07</td>
<td>1.05E+08</td>
<td>4.70E+09</td>
</tr>
</tbody>
</table>

**Table 3.9:** Conditioning of Linear Systems as $\epsilon_s$ is varied

### 3.5 Comparison of the LSFEM and Galerkin

The LSFEM has been shown to be an effective approximation method for the PBE [32]. In this section, we explore further the utility of a LSFEM discretization for this class of problems by detailing several advantages over a Galerkin formulation. We also discuss several quantitative implications such as the ability of the least-squares approximation to attain physical quantities—e.g. solvation energy—that are consistent with that of Galerkin.

The Galerkin approximation for the regularized PBE is defined by the weak problem: find $u$ so that

$$
\int_{\Omega} \epsilon \nabla u \cdot \nabla v + \kappa^2 u v \, dx = \int_{\Omega} - (\epsilon - \epsilon_m) \nabla u_c \cdot \nabla v - \kappa^2 \nabla u_c v \, dx, \quad \forall v.
$$

Notice that the flux variable is not explicitly expressed in the weak form—i.e. only the scalar potential $u$ is represented. As such, we examine the $L^2$ convergence of the error in the scalar potential and its gradient for both the Galerkin and the least-squares finite element approximations. Table 3.10 shows that both approximations achieve optimal $O(h^2)$ convergence for the scalar potential of the Born ion (recall the analytic solution is available in this case). The difference is in representing the flux. The Galerkin approximation is
only $O(h)$, while the least-squares approach yields $O(h^2)$ convergence, which is a notable advantage of the LSFEM.

<table>
<thead>
<tr>
<th>$h/h_0$</th>
<th>$u$</th>
<th>$\nabla u$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Value</td>
<td>Rate</td>
</tr>
<tr>
<td>1</td>
<td>346.7</td>
<td>362.5</td>
</tr>
<tr>
<td>1/2</td>
<td>94.3</td>
<td>1.88</td>
</tr>
<tr>
<td>1/4</td>
<td>24.2</td>
<td>1.96</td>
</tr>
<tr>
<td>1/8</td>
<td>6.0</td>
<td>2.01</td>
</tr>
</tbody>
</table>

**Table 3.10:** $L^2$ convergence of $u$ and $\nabla u$ for the Born Ion

**Comparison of Solvation Free Energy Calculations**

The electrostatic free energy of solvation measures the difference between electrostatic energy in solution and in vacuum [61]. For a set of discrete charges $Q_j$ and regularized potential $u$, the electrostatic free energy [67, 32] is defined by

$$\Delta G_{sol} = \frac{1}{2} k_B T \sum_{j=1}^{n_e} Q_j u(x_j).$$

Figure 3.4 shows the solvation energy results for the three schemes: Galerkin with uniform refinement and labeled “Galerkin”, LSFEM with uniform refinement and labeled “LSFEM” and LSFEM with adaptive refinement and labeled “Adaptive”. The adaptive refinement algorithm is the same one we used in the last chapter, and is based on the least-squares functional estimator. For adaptive refinement, the mesh size $h$ is approximated by $N^{-1/3}$, where $N$ is the number of nodes in the mesh. The results indicate that both Galerkin and LSFEM have similar rates of convergence, while we observe enhanced rate if we use adaptive refinement.

**3.6 Interface Formulation**

The discontinuity in coefficient functions $\bar{\kappa}$ and $\epsilon$ make the solution less smooth. While the LSFEM (3.1) results in the best approximation in a given finite-element space, convergence of the method cannot be
Figure 3.4: Comparison of the “Galerkin”, “LSFEM” and “Adaptive” methods for computing Solvation Free Energy. “Adaptive” is LSFEM using adaptive refinement.

guaranteed using piecewise linear elements. We recall from Chapter 2 that the linear regularized PBE has an interface equation,

$$
\left[ \langle \epsilon(x) \frac{\partial \tilde{u}(x)}{\partial n} \rangle \right]_\Gamma =, \quad x \in \Gamma.
$$

where the jump at the interface $\Gamma$ is defined as

$$
\left[ \langle \epsilon(x) \frac{\partial \tilde{u}(x)}{\partial n} \rangle \right]_\Gamma = \lim_{\alpha \to 0^+} \epsilon(x + \alpha \mathbf{n}) \frac{\partial \tilde{u}(x + \alpha \mathbf{n})}{\partial \mathbf{n}} \quad - \epsilon(x - \alpha \mathbf{n}) \frac{\partial \tilde{u}(x - \alpha \mathbf{n})}{\partial \mathbf{n}}.
$$

Piecewise linear elements are attractive because of their simplicity and one way to overcome the interface challenge is to explicitly represent the jump conditions in (3.7) [42]. In this section we briefly address this
approach and contrast it with using $P_1$ elements in a single domain formulation. For this section, we use the weighting $\alpha = 1/\epsilon$. While this weighting is inferior to $\alpha = 1/\sqrt{\epsilon}$, our primary purpose in this section is to compare interface formulation with a single domain formulation, and not the weighting.

Since the solution belongs to $H^2$ when restricted to $\Omega_m$ or $\Omega_s$, the interface approach to the LSFEM defines functionals on each of the two domains, while interface terms are added to couple the problems. Letting $(q_m, u_m)$ and $(q_s, u_s)$ be the restriction of the solution in $\Omega_m$ and $\Omega_s$, respectively, we define the least-squares functional $\tilde{G}$ as

$$
\tilde{G}(q, u; u_c) = \sum_{i=m,s} \left( \|q_i/\epsilon - \epsilon \nabla u_i/\epsilon - ((\epsilon - \epsilon_m)/\epsilon) \nabla u_c \|_{0,\Omega_i}^2 + \| - \nabla \cdot q_i + \kappa^2 u_i + \kappa^2 u_c \|_{0,\Omega_i}^2 \right) + \frac{1}{h^{1+c_0} \int_\Gamma [u]^2 d\Gamma} + \frac{1}{h^{c_1} \int_\Gamma [q.n]^2 d\Gamma},
$$

with $c_0, c_1 > 0$.

Since $u \in H^2$ in each sub-domain, we use $P_1$ for $q$ in each $\Omega_m$ and $\Omega_s$. At the interface $\Gamma$, there are two degrees-of-freedom at each vertex, one for each domain, as depicted in Figure 3.5a. These degrees-of-freedom are coupled through the interface term in the functional.

![Figure 3.5](image_url)

(a) degrees-of-freedom

(b) Comparison with single domain

**Figure 3.5:** Description and results for the interface formulation

In Figure 3.5b we compare the convergence of the least-squares functional for the single domain formulation in (2.15) with the interface formulation in (3.8), for methanol. Although similar convergence is achieved for both methods, the interface approach requires a more complex implementation due to the interface terms.
Effect of Mesh Refinement the LSFEM

For non-differentiable interfaces, \( \Gamma \), solutions exhibit singularities at the corners [107]. In this section, we examine the effect of refinement on the convergence of the least-squares functional, starting from an initial coarse representation of the solvent–molecular interface.

The solution of the Poisson-Boltzmann equation is known to be sensitive to the description of the solvent–molecular interface (or molecular surface) [16]. This interface is used to model the inability of solvent to penetrate the molecule, and is typically defined as a unions of spheres, splines, or a level set [16]. Here we denote this surface by \( \Gamma \), and label it the true interface.

We focus our numerical study on the case of the Born ion since the interface is given precisely as a sphere and eases analysis. Given a coarse representation of the mesh that conforms to the true interface, we monitor convergence of the least-squares functional, which is also an error estimator, and refine by three methods (see Figure 3.6a):

- **no pushing** New points are simply placed at midpoints of edges in the parent mesh. The concern here is that we are limited by the initial definition of the mesh.
- **pushing** New points that are introduced through refinement are pushed to the true interface, thereby ensuring a smooth surface in the limit. A disadvantage of this approach is that the meshes are not nested and that more management is needed during refinement. Moreover, in some situations simplices at the surface may exhibit flattening.
- **pushing with smoothing** To overcome flattening of elements this approach also smooths the pushed mesh.

This option is computationally the most expensive.

From the results shown in Figure 3.6b, we conclude that pushing to the interface does in fact yield a noticeable and sustained improvement in the convergence rate. Flattened elements are encountered after just a few refinements, thus requiring the use of smoothing.

3.7 Conclusions

In this chapter we examined the numerical properties of a least-squares finite element method for the linear Poisson-Boltzmann equation, and discussed it as an alternative to a standard Galerkin approach. The least-squares functional we considered is \( H(\text{div}) \times H^1 \) equivalent, which motivated our investigation of \( H(\text{div}) \) conforming Raviart–Thomas elements for the flux variable. The \( RT_0 \) elements are more robust and accurate choice as compared to \( P_1 \) elements. We then examined the effect of various weighting of the LSFEM functional,
and advocated one which had good accuracy and better conditioning than the one we proposed in the last chapter.

We followed this discussion by contrasting the LSFEM with the traditional Galerkin approach. Both methods have comparable accuracy for computing scalar variables and physical quantities such as solvation free energy. Yet, we observe improved control of the gradient of the solution in the LSFEM, and hence advocate this approach when evaluation of solution derivatives is desired. Through our discussion we concluded that LSFEMs are both accurate and computationally tractable, and offer a viable alternative to standard Galerkin approaches.

Finally, an interface approach \cite{42} was investigated, to address the jump condition in the problem while still using $P_1$ elements. The numerical results were found to be comparable to that of the single domain $RT_0$ LSFEM method. Given the success of this single domain approach, we concluded with refinement strategies for a curved interface.
Chapter 4

Finite Element Approximation to a Finite-Size Modified Poisson-Boltzmann Equation

The previous chapters focused on numerical methods for the PBE. The PBE has proved successful in approximating electrostatic interactions for moderate surface charge densities. The PBE adopts a continuum mean-field description of a solvent, assuming point-like ions in thermodynamic equilibrium, and neglecting statistical correlations [34]. However, the assumption of zero-size ions in this description leads to unphysical concentrations near the surface under high charge densities, even if the bulk solution is dilute [35].

This chapter discusses numerical methods for a modified form of the PBE. The modifications lead to bounded ionic concentration profiles and are consistent with the Poisson-Boltzmann equation in the limit of zero-size ions. Moreover, the modified equation fits well into existing finite element frameworks for the Poisson-Boltzmann equation. Section 4.1 introduces the modified PBE and lays out the motivation for the investigation of numerical methods for this equation. In Section 4.2 we describe the MPBE, its domain, and different transformations to make it amenable for numerical computations. In Section 4.3 we prove the existence of a unique weak solution to the regularized problem. We give a finite-element formulation for regularized MPBE in Section 4.4 and show optimality and convergence of the approximations. Finally, in Section 4.5, we discuss a relationship between the PBE and the MPBE as ion size is decreased in the MPBE.

4.1 Motivation

One approach to modeling the effects of finite-size ions is to consider a lattice gas model for the free ions [35], which leads to the modified PBE (MPBE). Also termed the Poisson-Bikerman equation [87, 27], the MPBE has the advantage of bounded concentrations of ions near the molecular surface. As the size of the ions tend toward zero, the modified equation converges to the PBE. The value of incorporating steric effects into the equation is highlighted through several recent applications that have adopted the MPBE [35, 87, 88, 123].

In addition to the MPBE, there have been other attempts to modify the Poisson-Boltzmann theory in an effort to account for steric effects. For example, through a series of modified equations [95], a new model
has been derived [111], which takes into account volume-exclusion effects of free ions. In another model based on density functional theory [124], in addition to the ion-exclusion effects, solvent exclusion effects are approximated by modeling solvent molecules as neutral, hard spheres. While effective, these alternate approaches involve non-trivial equations, and considerable computational effort is required to solve them. One advantage of the MPBE proposed by [35], and considered here, is in the simplicity and ease with which it can be incorporated with existing PBE numerical implementations.

The finite element method has been widely used for solving the PBE [49, 110, 53, 79, 14, 119, 132, 131, 32]. In this chapter we analyze the MPBE. We start by analytically subtracting the singularities in the potential through a common regularization procedure [49, 140] and follow by proving the existence and uniqueness of weak solutions to the regularized MPBE. We then examine the finite-element formulation for this equation and show that the corresponding approximation converges as the mesh is refined. As a practical consideration, we also examine the conditioning of linear systems which arise during discretization. Finally, we discuss a method for solving the unmodified PBE by first solving the modified equation to improve the initial guess. We argue that this approach is vital for fast convergence of Newton-like methods.

The unmodified PBE has been analyzed in Chen, Holst, and Xu [49]. Although our analysis of the modified PBE uses a similar approach, there are important differences. In particular, the nonlinearity in the modified equation is bounded, which simplifies the theory. When applicable, we use prior results for the unmodified PBE to contribute to our theory and discussion.

A slightly different form of the modified Poisson-Boltzmann equation has been analyzed by Li [96]. The focus is on deriving the equation from a free energy function, from properties of equilibrium concentrations, and by rigorously proving equivalence of different forms of the equation. In this chapter, we approach existence and uniqueness of solution differently, using results from convex optimization. Moreover, our results focus on discrete solutions to the problem.

### 4.2 The Modified Poisson Boltzmann Equation

Biomolecular systems are frequently modeled through classical molecular dynamics, where the dynamics of the atoms in the system are resolved, and thermodynamic properties are estimated by averaging in time [66]. Due to the large number of solvent molecules, and the small timesteps required in the simulation, explicit methods based on molecular dynamics are extremely expensive [135, 63]. A more tractable approach is to represent the solvent implicitly as a dielectric continuum, which results in the Poisson-Boltzmann model. Let $\phi(x)$ represent the unknown electrostatic potential, $\epsilon(x)$ the dielectric coefficient, and $\tilde{\epsilon}(x)$ the modified
Debye-Hückel parameter. Defining constant \( \beta = 1/(k_B T) \), \( k_B \) as the Boltzmann constant, \( e_c \) as the charge of an electron, and \( T \) as the temperature, the Poisson-Boltzmann equation for 1:1 solution (one positive and one negative ion per salt molecule, e.g. NaCl) is,

\[
-\nabla \cdot (\epsilon(x) \nabla \phi(x)) + \frac{1}{\beta e_c} \kappa^2(x) \sinh(\beta e_c \phi(x)) = 4\pi \sum_{i=1}^{N_m} Q_i \delta(x - x_i) \text{ in } \mathbb{R}^3,
\]

\[\phi(\infty) = 0.\]

Here, the solute contains \( N_m \) fixed points with charges \( Q_i \) at positions \( x_i \), and \( \delta \) represents the Dirac delta distribution.

The domain for the problem is \( \mathbb{R}^3 \) which is subdivided into a molecular region \( \Omega_m \), a solvent region \( \Omega_s^\infty \), and an interface between the two denoted by \( \Gamma \). The solute is surrounded by solvent, which is represented as a continuum over the subdomain \( \Omega_s^\infty = \mathbb{R}^3 \setminus \Omega_m \). The subdomains for a typical biomolecular solute are shown on the left in Figure 4.1.

**Figure 4.1:** Subdomains for the Poisson-Boltzmann equation

Let \( \epsilon_m \) and \( \epsilon_s \) be positive constants, and \( N_A \) represent Avogadro’s number. The dielectric coefficient, \( \epsilon(x) \), and modified Debye-Hückel parameter, \( \kappa^2(x) \), are piecewise constant functions given by

\[
\epsilon(x) = \begin{cases} 
\epsilon_m, & x \in \Omega_m, \\
\epsilon_s, & x \in \Omega_s^\infty,
\end{cases}
\quad \text{and} \quad
\kappa^2(x) = \begin{cases} 
0, & x \in \Omega_m, \\
\kappa^2_s = \frac{8\pi N_A e_c^2}{1000 k_B T c_b}, & x \in \Omega_s^\infty.
\end{cases}
\]

In the MPBE, the free ions occupy cells in a three-dimensional lattice of size \( a \), as shown on the right in
Let $a$ be the diameter of an ion and $c_b$ be the bulk concentration or ionic strength (which depends on the particular solvent being modeled). Defining $\nu = 2a^3c_b$, the modified PBE for a 1:1 solution, obtained from [35] (with a straightforward extension to account for the molecular region as is done for unmodified PBE in such cases) is,

$$-\nabla \cdot (\epsilon(x)\nabla \phi(x)) + \frac{1}{\beta e_c \kappa^2(x)} \frac{\sinh(\beta e_c \phi(x))}{1 - \nu + \nu \cosh(\beta e_c \phi(x))} = 4\pi \sum_{i=1}^{N_m} Q_i \delta(x - x_i), \quad \text{in } \mathbb{R}^3,$$

$$\phi(\infty) = 0.$$

For the remainder of this chapter, we assume that $0 < \nu < 1$, unless stated otherwise. Notice that as $a \to 0$ we recover the PBE.

The assumption of zero-size ions in the unmodified PBE leads to unphysical concentrations near the surface. This is illustrated in Figure 4.2 for the Born ion (described in Section 4.4.2) with a charge of $5e_c$, which shows counter-ion concentrations calculated using our finite-element formulation with a free ion size of 5 Å and bulk concentration of 0.1M. The PBE results in high concentrations of counter-ions near the surface of the molecule, whereas the MPBE yields a realistically bounded profile.

![Figure 4.2: Concentration profile near the Born ion interface.](image-url)
Truncation

For computation, the unbounded solvent domain, $\Omega^\infty_s$, is truncated to a bounded domain, $\Omega_s$, and the resulting problem domain, $\Omega = \Omega_s \cup \Gamma \cup \Omega_m$, has a convex and Lipschitz-continuous boundary $\partial \Omega$. Dirichlet boundary conditions are imposed to capture the asymptotic behavior of the solution on an unbounded domain. Together with a change of variables, $\tilde{u}(x) = \beta c \phi(x)$, this results in a dimensionless modified Poisson-Boltzmann equation on $\Omega$:

$$-\nabla \cdot (\epsilon(x) \nabla \tilde{u}(x)) + \bar{\kappa}^2(x) \frac{\sinh(\tilde{u}(x))}{1 - \nu + \nu \cosh(\tilde{u}(x))} = 4\pi \beta c \sum_{i=1}^{N_m} Q_i \delta(x - x_i) \quad \text{in } \Omega, \quad (4.2)$$

$$\tilde{u}(x) = g(x) \quad \text{on } \partial \Omega. \quad (4.3)$$

The boundary conditions are prescribed using a linear combination of Helmholtz Green’s functions,

$$g(x) = \beta c \sum_{i=1}^{m} Q_i \epsilon_s |x - x_i| \exp \left( -\bar{\kappa}_s |x - x_i| \right), \quad (4.4)$$

which captures the asymptotic behavior of the solution. This boundary condition is the same as boundary conditions commonly used for the PBE [49] and can be derived from the asymptotic properties of (4.2).

Similar to the unmodified case, the solution to the modified PBE must satisfy two conditions at the interface $\Gamma$,

$$\left[ \tilde{u}(x) \right]_{\Gamma} = 0 \quad \text{and} \quad \left[ \epsilon(x) \frac{\partial \tilde{u}(x)}{\partial n} \right]_{\Gamma} = 0, \quad x \in \Gamma, \quad (4.5)$$

where $n$ is the unit normal at the interface and the jump at the interface is defined as

$$\left[ v(x) \right]_{\Gamma} = \lim_{\alpha \to 0^+} v(x + \alpha n) - v(x - \alpha n).$$

Regularization

The right hand side in (4.2) contains delta distributions which are not in $H^{-1}(\Omega)$ — i.e., the dual space of $H^1_0(\Omega)$ (see [49]). This precludes seeking a solution in $H^1(\Omega)$, the typical Sobolev space for second order equations. The lack of smoothness of the solution $\tilde{u}$ also hinders the design of a convergent finite element method for the MPBE. Following [140], we overcome this issue by decomposing $\tilde{u}$ into

$$\tilde{u} = u + u_c, \quad (4.6)$$

49
where $u$ is an unknown smooth function and $u_c$ is a known singular function which absorbs the singularities in $\tilde{u}$. The Coulomb function, $u_c$, satisfies the Poisson equation

$$-\epsilon_m \nabla \cdot \nabla u_c(x) = 4\pi e_c \beta \sum_{i=1}^{m} Q_i \delta(x - x_i). \quad (4.7)$$

Combining (4.6) with (4.2), we obtain the regularized MPBE or RMPBE:

$$-
abla \cdot (\epsilon(x) \nabla u(x)) + \bar{\kappa}^2(x) \frac{\sinh(u(x) + u_c(x))}{1 - \nu + \nu \cosh(u(x) + u_c(x))} = \nabla \cdot ((\epsilon(x) - \epsilon_m) \nabla u_c(x)) \quad \text{in } \Omega,$nabla \cdot ((\epsilon - \epsilon_m) \nabla u_c(x)) \quad \text{in } \Omega,$$

$$u(x) = g(x) - u_c(x) \quad \text{on } \partial \Omega. \quad (4.9)$$

### 4.3 Existence and Uniqueness

In this section we establish the existence and uniqueness of the weak solution to the RMPBE. Although our approach is similar to the approach used by Chen, Holst, and Xu [49] to analyze the unmodified RPBE, the assumptions differ and we are able to make several simplifications to make the theory more accessible. In particular, we do not need an $L^\infty$ bound since the nonlinearity in the modified equation is bounded.

We denote the $L^2(\Omega)$ inner product by $(\cdot, \cdot)$ and the duality pairing between functions from $H^1_0(\Omega)$ and $H^{-1}(\Omega)$ by $\langle \cdot, \cdot \rangle$. Define

$$M := \{ v \in H^1(\Omega) | v = g - u_c \text{ on } \partial \Omega \} \quad (4.10)$$

and $V = H^1_0(\Omega)$. The weak problem for (4.8) becomes: Find $u \in M$ such that

$$A(u, v) + (N(u), v) + \langle f_c, v \rangle = 0, \quad \forall v \in V, \quad (4.11)$$

where

$$A(u, v) = (\epsilon \nabla u, \nabla v),$$

$$(N(u), v) = (\bar{\kappa}^2 \frac{\sinh(u + u_c)}{1 - \nu + \nu \cosh(u + u_c)}, v), \text{ and}$$

$$\langle f_c, v \rangle = \int_{\Omega} (\epsilon - \epsilon_m) \nabla u_c \cdot \nabla v \, d\mathbf{x}.$$

For equation (4.11) to be well defined, we need $N(u) \in L^2(\Omega)$. This is true since $N(u) \in L^\infty(\Omega)$, which
follows from the bound
\[
\frac{\sinh(u + u_c)}{1 - \nu + \nu \cosh(u + u_c)} \leq \frac{1}{\nu} \quad \text{in } \Omega.
\] (4.12)

We seek an energy functional on M with a minimum that satisfies (4.11). Define \( E \) as
\[
E(w) = \int_\Omega \frac{\epsilon}{2} \nabla w \cdot \nabla w + \frac{\kappa^2}{\nu} \ln(1 - \nu + \nu \cosh(w + u_c)) \, dx + \langle f_c, w \rangle.
\]
The motivation for choosing this energy is given by the following lemma, which shows that the solution to the RMPBE is the minimizer of this energy.

**Lemma 2.** If \( u \) is the solution of the optimization problem, i.e.
\[
E(u) = \inf_{w \in M} E(w),
\] (4.13)
then \( u \) is the solution of (4.11).

Similar to the proof of Lemma (4.1) in [49], we consider the function \( F : \mathbb{R} \to \mathbb{R} \) given by
\[
F(\tau) = E(u + \tau v),
\]
for any \( v \in H^1_0(\Omega) \). By assumption \( \tau = 0 \) is a critical point of this function. Thus \( F'(\tau)|_{\tau=0} = 0 \). This gives
\[
F'(\tau)|_{\tau=0} = \frac{d}{d\tau} \left[ \int_\Omega \frac{\epsilon}{2} (\nabla u + \tau \nabla v) \cdot (\nabla u + \tau \nabla v) + \frac{\kappa^2}{\nu} \ln(1 - \nu + \nu \cosh(u + \tau v + u_c)) \, dx + \langle f_c, u + \tau v \rangle \right] |_{\tau=0}
= (\epsilon \nabla u, \nabla v) + (\kappa^2 \frac{\sinh(u + u_c)}{1 - \nu + \nu \cosh(u + u_c)}, v) + \langle f_c, v \rangle.
\]

With a suitable energy function defined, the following theorem guarantees the existence and uniqueness of the solution to the minimization problem (4.13).

**Theorem 2 (Existence and Uniqueness).** There exists a unique \( u \in M \subset H^1(\Omega) \) such that
\[
E(u) = \inf_{w \in M} E(w).
\]
To prove this theorem we make use of results from variational analysis. The requisite definitions and theorems that we use are now given. Let $f : V \to \mathbb{R}$ be a functional on a separable reflexive Banach space $V$ with norm $\| \cdot \|$. We denote convergence in norm by $\to$ and weak convergence by $\rightharpoonup$. Additionally, recall that a functional $f$ is lower semi-continuous at $v \in V$ if $f(v) \leq \liminf_{n \to \infty} f(v_n)$ for any sequence $\{v_n\}$ such that $v_n \to v$ and is weakly lower semi-continuous at $v \in V$ if $f(v) \leq \liminf_{n \to \infty} f(v_n)$ for any sequence $\{v_n\}$ such that $v_n \rightharpoonup v$. We also make use of coercive and proper functions. A function $f : V \to \mathbb{R} \cup \{\infty\}$ is said to be coercive if $\lim_{\|x\| \to \infty} f(x) = \infty$, and is proper if $f \neq \infty$.

Now we give three results that we use in our proof of Theorem 2.

**Theorem 3** (cf. Theorem 3.3.4 in [8]). Let $(V, \| \cdot \|)$ be a reflexive Banach space, $D$ a closed convex subset of $V$ and $f : D \to \mathbb{R} \cup \{\infty\}$ a convex, lower semi-continuous and coercive function. Then there exists $u \in D$ which minimizes $f$ on $D$,

$$f(u) = \inf_{v \in D} f(v),$$

and the minimizer $u$ is unique if $f$ is strictly convex.

**Lemma 3** (cf. Theorem 7.2.3 in [5]). If $f$ is a convex functional on a closed convex set $D \subset V$ and $f$ is Gâteaux differentiable, then $f$ is weakly lower semi-continuous on $D$.

**Lemma 4** (cf. Theorem 3.3.3 in [8]). If a proper, convex functional on $V$ is weakly lower semi-continuous, then it is also lower semi-continuous.

With these standard definitions and theorems, we are ready to prove Theorem 2.

**Proof of Theorem 2.** To prove existence of minimizer, we need to verify that $M$ is a convex set, and $E$ is a convex, coercive and semi lower-continuous functional on $M$. To show convexity of $M$, we consider $w = tu + (1 - t)v$ for any $u, v \in M$ and a scalar $t$ such that $0 < t < 1$. Since $M \subset H^1(\Omega)$, we have $w \in H^1(\Omega)$ as $H^1(\Omega)$ is a linear space. On the boundary, $\partial \Omega$, we have

$$w = tu + (1 - t)v = t(g - u_c) + (1 - t)(g - u_c) = g - u_c$$

on $\partial \Omega$,

and hence we conclude that $w \in M$. This shows that the set $M$ is convex.

From the proof of Lemma 2, it follows that $E(w)$ is Gâteaux differentiable with

$$DE(u)[v] = A(u, v) + \langle N(u), v \rangle + \langle f_c, v \rangle.$$
Thus from Lemma 3, $E$ is weakly lower semi-continuous. Since $E$ is a proper function, i.e. $E(0) \neq \infty$, then Lemma 4 immediately implies that it is lower semi-continuous.

The convexity of $E$ follows from the convexity of $x^2$ and the convexity of $r(t) = \ln(1 - \nu + \nu \cosh(t))$ as a function from $\mathbb{R}$ to $\mathbb{R}$. The convexity of latter function is seen from

$$r''(t) = \frac{\nu + (1 - \nu) \cosh(t)}{(1 - \nu + \nu \cosh(t))^2} > 0,$$

since $0 < \nu < 1$. Thus $E$ is strictly convex.

To show coercivity of $E$, we have by a generalization of the arithmetic-geometric mean inequality, for any $\delta > 0$

$$\langle f_c, v \rangle \leq \epsilon_s \|
abla u_c\|_{\Omega_s} \|
abla v\|_{\Omega_s} \leq \frac{1}{\delta} \|
abla u_c\|_{\Omega_s}^2 + \delta \epsilon_s^2 \|
abla v\|_{\Omega_s}^2. \tag{4.14}$$

Since $\cosh(t) \geq 1$, it follows that $\ln(1 - \nu + \nu \cosh(v + u_c)) \geq 0$ and hence $E(v) \geq C(\epsilon, \delta) \|
abla v\| - \frac{1}{\delta} \|
abla u_c\|_{\Omega_s}^2.$

By choosing $\delta$ sufficiently small, $C(\epsilon, \delta) > 0$ is satisfied. Then the Poincaré-Friedrich’s inequality gives us

$$E(v) \geq C(\epsilon, \delta) \|
abla v\|_1^2 + C(u_c, g) \tag{4.15}$$

which shows the coercivity of $E$. Finally, Theorem 3.3.4 in [8] shows the existence of the minimizer. Since $E$ is strictly convex, the minimizer is unique.

### 4.4 Finite Element Discretization

With the theoretical framework of weak solutions presented in the previous section, we next consider the implications on associated finite element discretizations. In particular, we show ellipticity of the bilinear form, $A(\cdot, \cdot)$, and establish several properties on the functional $N(u)$. From this we are able to verify optimality through an a priori error estimate and demonstrate convergence in the $H^1$-norm.

This section illustrates the numerical robustness of the finite element method for the MPBE relative to the unmodified equation. Convergence proofs for solutions of the unmodified PBE include restrictive mesh conditions [49] which are generally not satisfied by standard meshing software designed for the PBE, e.g. GAMer [136]. These conditions arise from the need for the discrete solution to satisfy $L^\infty$ estimates. In this section, we show that the MPBE does not have the same restrictions, as the nonlinearity is bounded.

As a result, this yields a provably convergent method for the MPBE using meshes generated by existing meshing software, without the need to satisfy the discrete $L^\infty$ bounds. We also note that the linear systems
arising from discretization of the MPBE are better conditioned (compared to the PBE), which leads to a more efficient numerical method.

4.4.1 Analysis of the Nonlinear Finite Element Approximation

Consider \( u_D \) that satisfies the Dirichlet boundary conditions in (4.8). We then solve for \( u_0 \in H^1_0(\Omega) \) with \( u_0 = u - u_D \). Correspondingly, we define \( V^h \) to be the space of globally continuous piecewise linear elements that satisfy Dirichlet boundary conditions. That is, \( V^h = \{ v \in H^1(\Omega), v|_{\tau} \in P_1(\tau), v = g - u_c \text{ on } \partial \Omega, \forall \tau \in T_h \} \), where \( T_h \) is the triangulation of \( \Omega \), and \( V^h_0 = V^h \cap H^1_0(\Omega) \). From this, we consider the finite element approximation based on (4.11) to be: Find \( u_h \in V^h \) such that

\[
A(u_h, v_h) + (N(u_h), v_h) + \langle f_c, v_h \rangle = 0, \quad \forall v_h \in V^h_0.
\] (4.16)

Several results hold in this discrete setting as a direct consequence of the theory provided in the previous sections. From Lemma 2, for example, we have that \( u_h \) is the minimizer of \( E \) in \( V^h \). Also, since \( V^h \) is a convex and closed subset of \( H^1(\Omega) \), based on Theorem 2, existence and uniqueness of the discrete solution follows as well.

For finite element approximations it is important to verify not only the existence of a unique solution, but also optimality of the solution. We present this result below in Theorem 5 and follow in Theorem 6 by establishing convergence of the discrete solution as the mesh is refined. Before proceeding, we recall several results in functional analysis that are instrumental in proving our proposed theory.

First, the following lemmas generalize the mean value theorem for functions on Banach spaces and relate convexity to Gâteaux differentiability. They will be used in proving properties of \( A(\ , \ , ) \) and \( N(\ , ) \) in Lemma 6.

**Lemma 5** (cf. Proposition 5.3.11 in [7]). Let \( V \) be a Banach space and \( F : V \rightarrow V \) be a Gâteaux differentiable function such that the derivative \( DF(u) \) is a continuous function of \( u \) on \( V \) to \( \mathcal{L}(V,V) \). Then for \( u, v \in V \)

\[
\|F(u) - F(w)\|_V \leq \sup_{0 \leq \theta \leq 1} \|DF((1-\theta)u + \theta w)\||u - w\|_V.
\]

**Theorem 4** (cf. Theorem 5.3.17 in [7]). Let \( V \) be a Banach space and \( M \subset V \) be a non-empty convex subset. Assume \( f : M \rightarrow \mathbb{R} \) is Gâteaux differentiable. Then \( f \) is convex implies

\[
(Df(u) - Df(v))[u - v] \geq 0 \quad \forall u, v \in M.
\]
The following lemma establishes important properties of bilinear form $A(\cdot, \cdot)$ and functional $N(\cdot)$, which are then used to show optimality of the discrete solution in Theorem 5. We choose to present the statement of the lemma and theorem to coincide with the unmodified Poisson-Boltzmann equation theory as in Lemma 6.1 and Theorem 6.2 in [49]. The relationship between the results of the PBE and the MPBE in this context is important, but it is notable that our hypothesis and assumptions differ from the existing PBE theory. As a result, the mechanics of our proofs are different and the simplicity of our theory further highlights the accessibility and ease-of-use of a modified form of the Poisson-Boltzmann equation.

**Lemma 6 (Properties of $A(\cdot, \cdot)$ and $N(\cdot)$).**

(a) The bilinear form $A(u, v)$ is coercive and bounded. That is

$$c \|u\|^2 \leq A(u, u), \quad \text{and} \quad A(u, v) \leq C \|u\|_1 \|v\|_1.$$

(b) The operator $N$ satisfies the property

$$(N(u) - N(v), u - v) \geq 0.$$

(c) The operator $N$ is bounded in the sense that for $u, v \in H^1(\Omega), w \in L^2(\Omega),$

$$(N(u) - N(v), w) \leq K\|u - v\||w||.$$

**Proof.** The proof of (a) is standard.

To prove (b), consider the functional $R(u) = \int_{\Omega} \bar{\kappa} \ln(1 - \nu + \nu \cosh(u))$. From the proof of Theorem 2, we have that $R$ is a convex and Gâteaux differentiable with $DR(u)[v] = \int_{\Omega} N(u)v = (N(u), v)$. Using the convexity and differentiability of $R$ Theorem 4 then implies,

$$(DR(u) - DR(v))[u - v] = (N(u) - N(v), u - v) \geq 0.$$

To show (c), we use the Cauchy-Schwarz inequality,

$$(N(u) - N(v), w) \leq \|N(u) - N(v)||w||.$$
Now consider \( n(t) = \frac{\kappa_2 \sinh(t)}{1 - \nu + \nu \cosh(t)} \) as a function from \( \mathbb{R} \) to \( \mathbb{R} \). Then,

\[
n'(t) = \frac{\kappa_2 \nu + (1 - \nu) \cosh(t)}{(1 - \nu + \nu \cosh(t))^2} = \frac{\nu \operatorname{sech}(t) + 1 - \nu}{(\operatorname{sech}(t) - \nu \operatorname{sech}(t) + \nu)^2} \leq \frac{\kappa_2^2}{\nu^2} \leq K.
\]

Then for any \( v \in L^2(\Omega) \) we have

\[
\|DN(v)\| = \sup_{w \in L^2(\Omega)} \frac{\|DN(v)[w]\|_{L^2(\Omega)}}{\|w\|_{L^2(\Omega)}} \leq \sup_{w \in L^2(\Omega)} \frac{\sqrt{\int_{\Omega} (n'(v)w)^2}}{\|w\|_{L^2(\Omega)}} \leq \sup_{w \in L^2(\Omega)} K \frac{\|w\|_{L^2(\Omega)}}{\|w\|_{L^2(\Omega)}} = K.
\]

Finally, from the generalized Mean-Value Theorem for Gâteaux differentiable functions Lemma 5, we have

\[
\|N(u) - N(v)\| \leq K\|u - v\|,
\]

which shows (c).

With these properties on bilinear form \( A(\cdot, \cdot) \) and functional \( N(\cdot) \), one can show that the discrete solution is quasi-optimal in the \( H^1 \)-norm. This result resembles Céa’s lemma, which shows quasi-optimality of discrete solutions for bilinear forms.

**Theorem 5** (Quasi-optimal a priori error estimate). Let \( u \) and \( u_h \) be the solution of RMPBE and its finite element approximation, respectively. Then the finite element solution \( u_h \) is quasi optimal, i.e.

\[
\|u - u_h\|_1 \leq C \inf_{v_h \in V_h} \|u - v_h\|_1.
\]

**Proof.** The proof directly follows the proof of [49, Theorem 6.2].

From this \textit{a priori} estimate we are now able to relate the error to a given mesh in the following convergence result.

**Theorem 6** (Convergence in norm of the discrete solution). Let \( u \in H^1(\Omega) \) be the weak solution of RMPBE and \( u_h \) be the finite element approximation. Further we assume the following,
• The interface $\Gamma$ is polygonal and is exactly represented by the faces of simplices in $\mathcal{T}_h$.

• The maximum diameter of the elements in the mesh, $h$, goes to 0.

• The family of triangulations $\mathcal{T}_h$ is shape regular.

Then there exists a constant $C$ independent of $h$ such that

$$\|u - u_h\|_{1, \Omega} \leq Ch \left( \sum_{\tau \in \mathcal{T}_h} |u|_{2, \tau} \right)^{1/2},$$

and hence

$$\lim_{h \to 0} \|u - u_h\|_1 = 0.$$

Proof. Because of our assumption on the interface $\Gamma$, $\epsilon$ and $\bar{\kappa}$ are piecewise constants on each element $\tau \in \mathcal{T}_h$. If $\tau \subset \Omega_m$, then for $v \in H^1_0(\tau)$ we extend it by 0 outside $\tau$ to $\tilde{v}$ so that $\tilde{v} \in H^1_0(\Omega)$. By definition of the weak solution we have,

$$\langle \epsilon \nabla u, \nabla \tilde{v} \rangle_{\Omega} = 0,$$

$$\epsilon_m \langle \nabla u, \nabla v \rangle_{\tau} = 0 \quad \forall v \in H^1_0(\tau).$$

A regularity result of elliptic partial differential equations [70] says that $u \in H^2(\tau)$ for any $\tau \subset \Omega_m$. Similarly, since $N(u) \in L^2(\Omega)$ we have $u \in H^2(\tau)$ for any $\tau \subset \Omega_s$. Letting $I_h u \in V^h$ be the finite element interpolant of $u$, our quasi-optimal result (4.17) and standard approximation theory [37] results in

$$\|u - u_h\|_{1, 1} \leq C \|u - I_h u\|_{1, 1} = C \sum_{\tau \in \mathcal{T}_h} \|u - I_h u\|_{1, \tau} \leq Ch^2 \sum_{\tau \in \mathcal{T}_h} |u|_{2, \tau},$$

which completes the proof. \qed

4.4.2 Nonlinear Solve

The previous sections establish important results regarding the weak form of the nonlinear problem. In this section we examine the linearized modified Poisson-Boltzmann equation. We begin by defining the functional $F(\cdot)$ by

$$\langle F(u), v \rangle = \langle \epsilon \nabla u, \nabla v \rangle + \langle \bar{\kappa}^2 \frac{\sinh(u + u_c)}{1 - \nu + \nu \cosh(u + u_c)} \cdot \nabla v \rangle + \langle (\epsilon - \epsilon_m) \nabla u_c \cdot \nabla v \rangle.$$  \hspace{1cm} (4.18)
Using $F$, the nonlinear weak problem given in (4.11) can be restated as: Find $u \in H^1(\Omega)$ such that

$$
\langle F(u), v \rangle = 0 \quad \forall v \in H^1_0(\Omega).
$$

(4.19)

Newton Linearization

To solve the nonlinear equation in (4.19), we apply Newton’s method, which uses the linearization of the nonlinear form $\langle F(u), v \rangle$. Thus, for the MPBE, the bilinear form $\langle DF(u)[w], v \rangle$ is given by

$$
\langle DF(u)[w], v \rangle = (\epsilon \nabla w, \nabla v) + (\bar{\kappa}^2 \frac{\nu + (1 - \nu) \cosh(u + u_c)}{(\nu \cosh(u + u_c) - \nu + 1)^2} w, v).
$$

(4.20)

Newton’s method with damping ($\gamma \in [0, 1]$) solves for the unknown $u$, which can be summarized by

<table>
<thead>
<tr>
<th>Input: $u \in H^1(\Omega)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>while not converged</td>
</tr>
<tr>
<td>Find $w \in H^1_0(\Omega)$ so that $\langle DF(u)[w], v \rangle = -(F(u), v)$, $\forall v \in H^1_0(\Omega)$.</td>
</tr>
<tr>
<td>$u \leftarrow u + \gamma w$</td>
</tr>
<tr>
<td>end while</td>
</tr>
</tbody>
</table>

For the discrete problem, the spaces $H^1(\Omega)$ and $H^1_0(\Omega)$ are replaced by $V^h$ and $V^h_0$ respectively. Convergence criteria for a damped Newton’s method may be $|\langle F(u_h), v_h \rangle| < TOL$ for all $v_h$ or $\|u^{i+1}_h - u^i_h\| < TOL$, where $TOL$ is some user specified tolerance and $u^i_h$ and $u^{i+1}_h$ are approximations to the solution from successive iterations of the algorithm. In our experiments we found that convergence of the damped Newton iterates was insensitive to the method of selecting $\gamma$.

The Newton iteration above requires an initial approximation to the solution and convergence is dependent on this selection. Moreover, an accurate initial approximation leads to improved performance. For our approach, we use the solution of a linearized PBE, where the linearized PBE arises from linearizing the nonlinear term in either the PBE or MPBE. For example, near $\tilde{u} = 0$, the linearized PBE or MPBE is

$$
-\nabla \cdot (\epsilon(x) \nabla \tilde{u}(x)) + \kappa^2(x)\tilde{u}(x) = 4\pi \beta e_c \sum_{i=1}^{N_m} Q_i \delta(x - x_i).
$$

This equation is then regularized as described for the nonlinear the MPBE and solved using the finite element method.
Convergence Rate of the Discrete Solution

Theorem 6 states that the expected rate of convergence for discrete solution is $O(h)$ for the $H^1$ norm. For second order elliptic problems with continuous coefficients, one can often prove that the convergence rate for the $L^2$ norm is $O(h^2)$. We are unable to show this rate for the $L^2$ norm for the MPBE because the coefficients $\epsilon$ and $\kappa$ are discontinuous. However we show $O(h^2)$ convergence in $L^2$ norm experimentally, and also verify the claim of $O(h)$ convergence for the $H^1$ norm by examining the convergence of our finite-element scheme for the Born ion.

The domain for the Born ion consists of a spherical solute of radius $R$ with a single point charge $Q_1$ at its center [33]. The solute is surrounded by a solvent, with the solvent domain truncated at a finite spherical radius, as depicted in Figure 4.3a.

Analytical solutions for the MPBE do not exist. To obtain a “true” solution, we exploit the spherical symmetry in the Born ion, and recast (4.8) in spherical coordinates,

\[-\frac{1}{r^2} \frac{d}{dr} \left( r^2 \epsilon \frac{du}{dr} \right) + \frac{\kappa^2}{1 - \nu + \nu \cosh(u + u_c)} \frac{\sinh(u + u_c)}{\nu + \nu \cosh(u + u_c)} = \frac{1}{r^2} \frac{d}{dr} \left( r^2 (\epsilon - \epsilon_m) \frac{du}{dr} \right).\]

We solve this one-dimensional equation using a finite volume method on a fine mesh, and consider the resulting solution to be the “true” solution.

One advantage of the modified form of the Poisson-Boltzmann equation that we pursue is the natural fit into existing PBE finite element software. We employ the finite element package FETK [76] which is

![Figure 4.3: Models used in the numerical experiments](image-url)
frequently used for PBE computation. The finite element approximation benefits from an accurate meshing of the subdomains and we use a three-dimensional tetrahedral mesh generated using Geometry-preserving Adaptive Mesher (GAMer) [136]. In our convergence results, we normalize the mesh parameter, \( h \), by its value for the coarsest mesh, \( h_{\text{max}} \). Note that the \( L^2 \) and \( H^1 \) errors are absolute errors, and depend on the size of the domain and choice of units.

As an example, we consider \( a = 1, \epsilon_m = 1, \epsilon_s = 78, \) ionic-strength of 0.1\( M \), and temperature of 300\( K \). Table 4.1 summarizes the convergence rate of the FEM solution for the MPBE and the results indicate that convergence is approaching optimality.

<table>
<thead>
<tr>
<th>( h/h_{\text{max}} )</th>
<th>( L^2 ) Error</th>
<th>( L^2 ) Convergence rate</th>
<th>( H^1 ) Error</th>
<th>( H^1 ) Convergence rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>247.8</td>
<td>—</td>
<td>442.9</td>
<td>—</td>
</tr>
<tr>
<td>1/2</td>
<td>75.6</td>
<td>1.7</td>
<td>201.4</td>
<td>1.1</td>
</tr>
<tr>
<td>1/4</td>
<td>22.0</td>
<td>1.8</td>
<td>96.5</td>
<td>1.1</td>
</tr>
<tr>
<td>1/8</td>
<td>5.7</td>
<td>2.0</td>
<td>47.1</td>
<td>1.0</td>
</tr>
</tbody>
</table>

**Table 4.1:** \( L^2 \) convergence of \( u \)

### Conditioning of Linear Systems in Newton’s Iteration

The dominant cost in each iteration of Newton’s method is the solution of the linear problem, which in this case results in a large, sparse linear system. Consequently we examine bounds on condition numbers from systems arising from unmodified PBE and MPBE to quantify the computational effort needed at each Newton iteration.

For simplicity, initially we consider homogeneous physical parameters \( \epsilon = \bar{\kappa}_s = 1. \) Since \( \epsilon \) and \( \bar{\kappa} \) are common to each equation, the dependence on them in the following discussion of condition numbers is the same.

The condition number of a linear system at a particular Newton iteration depends on the coercivity and continuity constants of the bilinear forms \( \langle DF(u)[w], v \rangle \) and \( \langle D\tilde{F}(u)[w], v \rangle \), which correspond to the MPBE and PBE respectively. Its a simple exercise to verify that the linear systems arising from forms \( \langle DF(u)[w], v \rangle \) and \( \langle D\tilde{F}(u)[w], v \rangle \) are symmetric and positive-definite.

We will assume that the family of meshes \( \mathcal{T}_h \) satisfies inverse assumption, i.e. there exists a positive constant \( \theta \) such

\[
h \leq \theta \text{ diam}(\tau) \quad \forall \tau \in \mathcal{T}_h.
\]
Let \( \phi_1, \ldots, \phi_N \) be the basis for \( V^h \). Any \( v \in V^h \) can be represented as

\[
v = \sum_{i=1}^{N} \eta_i \phi_i
\]

Let \( \eta \) denote the vector \([\eta_1, \ldots, \eta_N]\). Then from [40] its easy to see that the \( l^2 \) norm of \( \eta \), is related to the \( H^1 \) norm of \( v \) by

\[
\alpha_1(h) |\eta| \leq ||v||_{1, \Omega} \leq \alpha_2(h) |\eta|
\]

where \( \alpha_1 \) and \( \alpha_2 \) are constants depending on \( h \). We are not concerned with dependence on \( h \), but rather on how the differences in PBE and MPBE affect the condition number. Under the above assumptions, the coercivity constant for the bilinear forms gives an upper bound on the largest eigenvalue of the resulting linear system, and boundedness constant gives a lower bound on the smallest eigenvalue. This can be seen by the fact that the largest and smallest eigenvalues of symmetric matrices are given by

\[
\lambda_{\text{max}} = \max_{x \in \mathbb{R}^3} \frac{x^T A x}{|x|^2}, \quad \lambda_{\text{min}} = \min_{x \in \mathbb{R}^3} \frac{x^T A x}{|x|^2}
\]

and the fact that for a bilinear form \( a(u, v) \) and the corresponding matrix \( A \) obtained from FEM we have

\[
a(v, v) = \eta^T A \eta
\]

And hence from equation (4.21) we get,

\[
c_1 \frac{\eta^T A \eta}{|\eta|^2} \leq \frac{a(v, v)}{||v||_{1, \Omega}^2} \leq c_2 \frac{\eta^T A \eta}{|\eta|^2}
\]

The condition number for symmetric positive definite systems in then simply the ratio of the largest and smallest eigenvalues. From the Poincaré-Friedrich’s inequality and the fact that hyperbolic cosine is a positive function, the coercivity bounds for PBE are

\[
\langle D\tilde{F}(u)[w], w \rangle = (\nabla w, \nabla w) + (\tilde{\kappa}^2 \cosh(u + u_c)w, w)
\]

\[
\geq (\nabla w, \nabla w) \geq C ||w||_{1}^2.
\]
Furthermore, for boundedness we use the Cauchy-Schwarz inequality,

$$\langle \tilde{D}F(u)[w], w \rangle \leq \max_{x \in \Omega} \cosh(u(x) + u_x)[(\nabla w, \nabla v) + (w, v)]$$

$$\leq \max_{x \in \Omega} \cosh(u(x) + u_x) \|w\|_1 \|v\|_1.$$ 

For the MPBE, the coercivity bound follows similarly to the PBE case since

$$\nu + (1 - \nu) \cosh(x) \left( \nu \cosh(x) + 1 - \nu \right)^2 \geq 0.$$ 

For boundedness, we note that

$$\frac{\nu + (1 - \nu) \cosh(x)}{(\nu \cosh(x) + 1 - \nu)^2} \leq 1 < \frac{1}{\nu^2}$$

to get,

$$\langle DF(u)[w], w \rangle \leq \frac{1}{\nu^2} \|w\|_1^2.$$ 

From these bounds, we argue that the condition number for the PBE could be as large as $O(\cosh(u))$, while for the MPBE it is $O(1/\nu^2)$. Since $\cosh(u)$ depends exponentially on $u$, the conditioning for the PBE is likely to be considerably worse than for the MPBE if $u$ is large. Table 4.2 gives the condition numbers (estimated to high accuracy using Lanczos iteration) of the linear systems for the first three iterations of Newton’s method for the Born ion, which provides numerical evidence for the above hypothesis. For the MPBE, $\nu$ corresponding to ionic radius of $a = 1$ was used.

<table>
<thead>
<tr>
<th>Newton Iteration</th>
<th>PBE</th>
<th>MPBE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.21e+08</td>
<td>1.76e+04</td>
</tr>
<tr>
<td>2</td>
<td>1.18e+08</td>
<td>1.75e+04</td>
</tr>
<tr>
<td>3</td>
<td>4.34e+07</td>
<td>1.75e+04</td>
</tr>
</tbody>
</table>

Table 4.2: Comparison of condition numbers for the Born ion

Next we examine the $\nu$-dependence of the condition number of the linear systems arising in the MPBE. Table 4.3 shows the variation in condition number as the parameter $\nu$ is changed in the MPBE for the Born ion. Specifically, the table shows the condition numbers of matrices for the first iteration of Newton’s method. The results indicate that as $\nu$ goes to 0, the conditioning of the linear systems in the Newton iteration deteriorates. Since the $O(1/\nu^2)$ bound on the condition number was derived from coercivity and continuity bounds and since the bounds are not tight, there is not a sharp agreement with the condition numbers seen in practice.

The condition number for $\nu = 0$ corresponds to the unmodified PBE, for which we have already seen that the condition number is bounded by $O(\cosh(u))$. This explains why the condition number does not increase as $O(1/\nu^2)$, but stays bounded as $\nu$ is decreased.

Finally we examine the dependence of the condition number as the parameters $\epsilon$ and $\bar{\kappa}_s$ are varied.
Table 4.3: Condition numbers for varying $\nu$

<table>
<thead>
<tr>
<th>$\nu$</th>
<th>Condition Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2e-01</td>
<td>1.79e+04</td>
</tr>
<tr>
<td>1.2e-04</td>
<td>1.76e+04</td>
</tr>
<tr>
<td>1.2e-07</td>
<td>1.35e+06</td>
</tr>
<tr>
<td>1.2e-10</td>
<td>4.34e+08</td>
</tr>
<tr>
<td>0</td>
<td>3.21e+08</td>
</tr>
</tbody>
</table>

Table 4.4: Condition numbers for varying $\epsilon_s/\epsilon_m$

<table>
<thead>
<tr>
<th>$\epsilon_s/\epsilon_m$</th>
<th>Condition Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19.2e+03</td>
</tr>
<tr>
<td>10</td>
<td>9.3e+03</td>
</tr>
<tr>
<td>100</td>
<td>22.4e+03</td>
</tr>
<tr>
<td>1000</td>
<td>21.9e+04</td>
</tr>
<tr>
<td>10000</td>
<td>21.8e+05</td>
</tr>
</tbody>
</table>

Table 4.4 and 4.5 show the variation in condition numbers for the first iteration of Newton’s method. We observe only a small change in the condition number as $\bar{\kappa}_s$ is increased by four orders of magnitude, which is expected since $\bar{\kappa}$ effects the mass matrix part of the full assembled matrix. On the other hand, condition number is more sensitive to the jump $\epsilon_s/\epsilon_m$ and varies linearly as expected.

4.5 Solution to the Nonlinear PBE using the MPBE

The modified form of the PBE that we consider in this chapter is motivated from physics and results in a tractable model from a theoretical point-of-view. In addition to the advantageous numerical properties, such as optimal accuracy and improved conditioning, the MPBE is also a useful tool when considering the unmodified PBE. Indeed, the finite element approach to the MPBE fits well into an existing finite element computational framework for the PBE. Even more, as we motivate in this section, it is also a valuable approach for generating an initial guess for the unmodified PBE.

Table 4.5: Condition numbers for varying $\bar{\kappa}_s$

<table>
<thead>
<tr>
<th>$\bar{\kappa}_s$</th>
<th>Condition Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.78e+04</td>
</tr>
<tr>
<td>0.1</td>
<td>1.73e+04</td>
</tr>
<tr>
<td>1</td>
<td>1.77e+04</td>
</tr>
<tr>
<td>10</td>
<td>2.01e+04</td>
</tr>
<tr>
<td>100</td>
<td>7.56e+04</td>
</tr>
</tbody>
</table>
As the ion size goes to 0 in the MPBE, we recover the PBE. Yet, for a finite-size with \( \nu > 0 \), the nonlinearity in the MPBE is numerically evaluated more effectively. To see this, consider the forms \( \langle \tilde{F}(u), (v) \rangle \) and its linearization \( \langle D\tilde{F}(u)[w], v \rangle \) for the PBE

\[
\langle \tilde{F}(u), (v) \rangle = (\epsilon \nabla u, \nabla v) + ((\epsilon - \epsilon_m) \nabla u_c \cdot \nabla v),
\]

\[
\langle D\tilde{F}(u)[w], v \rangle = (\epsilon \nabla w, \nabla v) + ((\epsilon - \epsilon_m) \nabla u_c \cdot \nabla v).
\]

In particular, each step of a Newton iteration requires evaluation of \( \cosh(u + u_c) \) and \( \sinh(u + u_c) \). If the initial approximation is not close to the solution, then these hyperbolic quantities may be large, leading to overflow. One method of avoiding this issue is by capping the input [15]. However, even with truncation the resulting solution process may be complex. We propose using the MPBE as an initial guess in Newton’s method to solve the unmodified PBE and we investigate this in the remainder of the chapter.

To highlight the differences with the MPBE, we consider the nature of the nonlinearities. The nonlinearities in \( \langle F(u), v \rangle \) and \( \langle DF(u)[w], w \rangle \) for the MPBE (see (4.18) and (4.20)) are \( N(u + u_c) \) and \( N'(u + u_c) \), which are bounded by \( \bar{\kappa}^2_s / \nu \) and \( \bar{\kappa}^2_s / \nu^2 \) respectively. Moreover if the input to hyperbolic functions are large, the nonlinearities in MPBE go to 0.

Thus, if a solution to the unmodified PBE is desired — without taking into account steric effects — using the MPBE solution provides a physically accurate and computationally tractable initial approximation. For example, choosing a small ion size so that \( 1 / \nu^2 \) is modest in size may provide a reasonable initial approximation for the PBE. This hypothesis is experimentally verified for the Born ion and for methanol (see Figure 4.3b). The parameters and subdomains for methanol were obtained from the APBS software package and meshed using GAMer [136].

For a comparison, we consider solving the nonlinear PBE with either a linear approximation of the PBE (labeled PBE in Table 4.6) or with the nonlinear MPBE (labeled MPBE in Table 4.6). In the case where the MPBE is used to form an initial approximation to the PBE, we report the sum of the Newton iterations from solving both the PBE and MPBE. An ion size of \( 10^{-1} \) is used in the MPBE, but other sizes are also effective. In both ion configurations, using the MPBE accelerates convergence of the nonlinear PBE. The improvement is more pronounced for methanol, which is particularly difficult to solve without using the MPBE to generate the initial guess. In the case of the Born ion, the linear approximation provides an adequate initial approximation, which unlike methanol, has a more uniform and basic geometry. Here we see that using the MPBE may provide a significant improvement in the convergence of the Newton iteration.
Remark 7. As noted earlier, convergence of Newton’s method is highly dependent on choosing a suitable initial approximation. One way to generate this is to use homotopy continuation methods [4]. A homotopy $H(x,t)$ between two functions $f(x)$ and $g(x)$ from a space topological space $X$ to another space $Y$ is a continuous function from $X \times [0,1]$ such $H(x,0) = f(x)$ and $H(x,1) = g(x)$ for all $x \in X$, see Figure 4.4. If we want to find zeros of $g$ using Newton’s method, then in homotopy continuation methods the function $f$ is chosen so that it is easy to solve for $t = 0$, and as the parameter $t$ varies from 0 to 1, the initial guess for $H(\cdot, t_n)$ is generated from the solution of $H(\cdot, t_{n-1})$. Homotopy methods aim to accelerate the solution of the target problem ($t = 1$) by solving a sequence of modified problems, as $t$ increases from 0 to 1.

For solving the PBE, we designate $g(\cdot)$ to be the nonlinear function corresponding to the FEM formulation of PBE in (4.22). The function $f(\cdot)$ is defined as the nonlinear function corresponding to the MPBE in (4.18). Varying $t$ from 0 to 1 then corresponds to varying $\nu$ in MPBE from some small initial value to 0. This is equivalent to solving MPBE with decreasing ion size in successive nonlinear solves until the ion size is 0, at which point we solve the PBE. In our numerical experiments we use only one partition of the homotopy map, but for more difficult problems, more homotopy resolution may be necessary.
4.6 Conclusion

In this chapter we have considered finite element approximations to a modified form of the Poisson-Boltzmann equation. The modified equation is important as it accounts for the steric or finite-size effects of ion interactions in a molecular simulation. Previously, modeling and computation efforts using this equation have been considered, but in this chapter we established several important theoretical results regarding the modified Poisson-Boltzmann equation and related the modified form to the standard form. In particular, we established existence and uniqueness of a weak solution to the problem thereby enabling a finite element formulation. To this end, we also showed convergence and optimal mesh dependence of a finite element approximation to the problem. As a practical consideration we examined the relationship between the conditioning and the standard Poisson-Boltzmann equation, showing that the conditioning is improved for the modified form and that the relationship is dependent on the ion size in a consistent manner. We also discussed the numerical evaluation of the modified equation, showing that numerical artifacts of evaluation in the standard form are not present in the modified form. Finally, we discussed the use of the modified equation to assist in the iterative solution of the nonlinear Poisson-Boltzmann equation, ultimately reducing the total number of linear solves and total time to solution.
Chapter 5

Finite Element Method for the Modified Poisson-Nernst-Planck Equations

In previous chapters we develop finite element methods for implicit solvent models for systems in equilibrium. This chapter focuses on systems in which ions are in motion. Transport properties of ionic solutions can be simulated using explicit ion methods such as all-atom molecular dynamics or Brownian dynamics [3, 2], or by using continuum models such as the Poisson-Boltzmann and Poisson-Nernst-Planck equations [56, 125]. While the explicit ion methods provide the most accurate description of the system’s behavior, both in spatial and temporal domains, they are stochastic in nature and thus require long, computationally expensive simulations to obtain average properties. Furthermore, the application of an explicit ion method usually requires the system to be described with the same resolution over the entire simulation domain. Often, this leads to a situation where a majority of the computational effort is applied to simulate a nearly uniform solution where quantities of interest exhibit little variation. In contrast, continuum methods allow different regions of the same system to be described at varying levels of detail, and thus focus the computational effort on regions that require a more precise description. In addition to being more computationally efficient, continuum models more easily incorporate certain types of boundary conditions that arise in physical systems, such as boundaries of fixed concentration or electrostatic potential.

The traditional continuum approach to modeling ionic transport is based on the Poisson-Nernst-Planck equations (PNPE). Although the PNPE have been applied successfully to model the electro-diffusion phenomena [93, 101], the equations are not without drawbacks. Within the PNPE approach, ions are modeled as mathematical points of negligible physical dimension, thereby allowing for accumulation of ions at unrealistically high concentrations in certain regions of the system. Here, we consider a modified formulation of the PNPE, called the modified Poisson-Nernst-Planck equations (MPNPE) [88], that explicitly takes the physical dimensions of ions into consideration, which limits the maximum concentration that can be attained in the system. The advantage of using MPNPE over PNPE becomes apparent in the systems that contain regions subject to strong attractive potentials, for example, near charged surfaces.

Here, we explore the MPNPE approach for modeling equilibrium and transport properties of ionic
solutions in realistic three-dimensional geometries subject to realistic applied potentials. The finite difference method has been widely used to solve the Nernst-Planck equations in one or three dimensions [31, 43, 51, 93]. Although the finite difference method is straightforward to implement, applying this method to systems that have curved boundaries and complicated geometries is challenging. In this respect, using a finite element method is more appropriate as it naturally handles complex geometries, such as the molecular surfaces of DNA molecules and ion channels.

A finite element method for solving the three-dimensional PNPE has been described [102, 103]. Numerical studies of the MPNPE have been limited to one-dimensional systems [88] and have not been applied to simulate ion flow through a solid-state nanopore, which is the main process considered in this work. In contrast to the previous studies, here we present a finite element method to solve the MPNPE that conserves ion concentration, takes into account the sharp repulsive potentials present near walls of an ion channel [93, 84], and is able to reproduce the results of explicit ions simulations.

Furthermore, the present study also fully describes a three dimensional MPNPE solver for the simulation of ionic current through nanopores, which takes into account complex geometry of the channel and realistic applied potentials. The presence of a sharp, repulsive potential at the interface of fluid and solid-state domains in the nanopore systems necessitates formulation of a new stable numerical method for finding the solution of the MPNPE. Thus, when applied to such systems, we find that the standard finite element method becomes unstable, producing spurious results such as negative concentrations. Below, we describe a numerical procedure that stabilizes the finite element method in the presence of sharp repulsive potentials, which is one of the main results of this work.

The remainder of the chapter is organized as follows. Section 5.1 introduces the systems and the governing equations. Our nonlinear finite element method for solving the MPNPE is described in Section 5.2. Also in this section, we provide the Galerkin formulation for the equations that do not have large advective terms, and a streamline-upwind-Petrov-Galerkin (SUPG) method for the equations in which such terms are present. In Section 5.3 we describe the results of several computational experiments that highlight the utility of the MPNPE and the necessity of having a stabilized algorithm. The chapter concludes with final remarks.

5.1 Problem Description

In this section we give a brief overview of the problem and review the relevant equations.
5.1.1 Governing Equations

We consider the Poisson-Nernst-Planck Equations (PNPE) for a 1:1 electrolyte solution (referred to as solvent) described over a computational domain, denoted by $\Omega = \Omega_s \cup \Omega_m$, which includes both the solvent region, represented as $\Omega_s$, as well as a molecular or membrane region, $\Omega_m$, which is void of solvent. The time dependent PNPE are given as

$$
\frac{\partial c_\pm}{\partial t} = D_\pm \nabla \cdot \left[ \nabla c_\pm + \frac{1}{k_B T} \left[ \pm e (c_\pm \nabla \phi) + (c_\pm \nabla U) \right] \right] \quad \text{in } \Omega_s,
$$

$$
- \nabla \cdot \epsilon \nabla \phi = e (c_+ - c_-) \quad \text{in } \Omega,
$$

where $\phi$ is the electrostatic potential and $U$ is the potential due to non-electrostatic interactions, which is assumed to be the same for both ionic species. In the Nernst-Planck equation, (5.1), the concentration of positive and negative ions are $c_+$ and $c_-$, respectively, $k_B$ is the Boltzmann’s constant, $T$ is temperature, $e$ is the charge on an electron and $D_\pm$ are the diffusivities of the positive and negative ions, respectively. In the Poisson equation, (5.2), we assume a piecewise constant dielectric coefficient $\epsilon$ that is defined in the two sub-domains, $\Omega_s$ and $\Omega_m$. For simplicity, we write the total potential energy experienced by an ion as

$$
V_\pm = \pm e \phi + U.
$$

The modified form of the PNPE (MPNPE) adds a nonlinear term to each of the two Nernst-Planck equations in (5.1) to model the steric repulsion. The Poisson equation remains unchanged, however the modified Nernst Planck equations are

$$
\frac{\partial c_\pm}{\partial t} = D_\pm \nabla \cdot \left[ \nabla c_\pm + \frac{1}{k_B T} c_\pm \nabla V_\pm + a^3 \left( \frac{c_\pm (c_+ + c_-)}{1 - c_+ a^3 - c_- a^3} \right) \right] \quad \text{in } \Omega_s,
$$

$$
- \nabla \cdot \epsilon \nabla \phi = e (c_+ - c_-) \quad \text{in } \Omega.
$$

Here, $a$ is the size of the ion (assumed to be the same for both species). As a result, in this model the maximum permitted concentration is bounded by $1/a^3$, which we refer to as the steric limit. To simplify the presentation of the material that follows, we write the PNPE and the MPNPE as

$$
\frac{\partial c_\pm}{\partial t} = D_\pm \nabla \cdot \left[ \nabla c_\pm + \frac{1}{k_B T} c_\pm \nabla V_\pm + N_a (c_\pm) \right] \quad \text{in } \Omega_s,
$$

69
where

\[ N_\alpha(c_\pm) = \alpha \left( a^3 \frac{c_\pm \nabla(c_+ + c_-)}{1 - c_+ a^3 - c_- a^3} \right), \]  

and

\[ \alpha = \begin{cases} 
0 & \text{for PNPE} \\
1 & \text{for MPNPE}. 
\end{cases} \]

5.1.2 Description of the Model System

A primary focus of this chapter is the application of MPNPE solver to nanopores, wherein we compute the ionic current through a pore in a solid-state membrane. The domain we consider is depicted in Figure 5.1a. Here, solution reservoirs above and below the membrane are connected through a nanopore, allowing positive and negative ions to pass from one side of the membrane to the other. We also consider a system where a DNA molecule is present inside the pore. Thus, the membrane (and the DNA, if present) comprise the domain \( \Omega_m \), and the ionic solution, which consists of the solution reservoirs above and below the membranes and the nanopore, comprise the domain \( \Omega_s \).

With the concentration profiles of (5.6), one important quantity is the ionic current \( J \) through a surface \( \Gamma \) with normal \( n \) (see Figure 5.1b), which is defined as

\[ J = \sum_{\pm} \pm \int_{\Gamma} e D_\pm n \cdot \left[ \nabla c_\pm + \frac{1}{k_B T} [\pm e (c_\pm \nabla \phi) + c_\pm \nabla V] \right]. \]  

For example, in the 2D cross-section of the problem domain shown in Figure 5.1b, we measure the ionic current through the plane in the middle of the pore, denoted by a dotted line.

The Poisson portion of the PNPE in (5.2) is solved with Dirichlet boundary conditions specified by \( \phi_t \) and \( \phi_b \) at the top and the bottom of the domain, and periodic boundary conditions along the other four sides. Further, the (unmodified and modified) Nernst-Planck equations in (5.1) and (5.4) use blocking boundary conditions on the interface of the membrane and the ionic solution, which is denoted \( \partial \Omega_{s,n} \) and is displayed with a dotted line in Figure 5.1b, while periodic boundary conditions are set at the remaining boundaries. Specifically, we consider blocking boundary conditions of the form

\[ \left[ \nabla c_\pm + \frac{1}{k_B T} c_\pm \nabla V_\pm + N_\alpha(c_\pm) \right] \cdot n = 0 \quad \text{on} \ \partial \Omega_{s,n}, \]

where \( n \) is the unit normal on the surface \( \partial \Omega_{s,n} \). One consequence of the blocking boundary conditions is
All-atom model of a nanopore system. The nanopore is shown as a smooth semi-transparent surface, the explicit ions are shown as spheres and the entire solvent domain is shown as a cut-away semitransparent molecular surface.

**Figure 5.1**: Description of the problem domain. The 3D domain is approximately 7.2 nm in length and 4 nm in width. The pore has a radius of 0.9 nm and a height of 4 nm. Dashed lines in the figure on the right represent a blocking boundary, while solid lines represent a periodic boundary for Nernst-Planck equations, and Dirichlet boundary at the top and bottom for the Poisson equation.

that the integral of the concentration remains constant. That is, the total number of ions of each ion species is conserved:

\[
\frac{\partial}{\partial t} \int_{\Omega_s} c_{\pm} \, dx = \int_{\Omega_s} D_{\pm} \nabla \cdot \left[ \nabla c_{\pm} + \frac{1}{k_B T} c_{\pm} \nabla V_{\pm} + N_\alpha (c_{\pm}) \right] \, dx = \int_{\partial \Omega_s,n} \nabla c_{\pm} \cdot \left[ \nabla c_{\pm} + \frac{1}{k_B T} c_{\pm} \nabla V_{\pm} + N_\alpha (c_{\pm}) \right] \cdot n = 0.
\] (5.11)

(5.12)

Additionally, the number of ions in the domain is also conserved in the case of a partially periodic boundary.
5.1.3 Brownian Dynamics

Several methods attempt to simulate the motion of ions by enforcing these principles of conservation, for example molecular dynamics. In this chapter, we refer to a Brownian Dynamics (BD) simulation [3] as a reference point, and we briefly describe it here.

The interaction between each pair of ions consists of a short-range portion, which is computed from all-atom molecular dynamics simulations, and a long-range Coulomb portion. The Coulomb portion is calculated using a uniform (relative) dielectric constant of 92 for the solvent, which is near the bulk value for the TIP3P water model used in the molecular dynamics simulations [134].

The short range portion of the interaction is calculated by the weighted histogram analysis method [117] using the results of many umbrella sampling molecular dynamics simulations. The umbrella sampling molecular dynamics simulations are performed using NAMD [112] and the protocols that have been described previously [52], including a 1 fs timestep, particle-mesh Ewald electrostatics, and a Langevin thermostat with a damping constant of 0.2 ps$^{-1}$. Interactions between the atoms of the systems (TIP3P water and ions) are calculated using the CHARMM force field [106], which includes the ions parameters for K$^+$ and Cl$^-$ described in [21]. The simulation systems consist of a periodic box of water that measures 5.8$\times$5.8$\times$5.9 nm$^3$ after equilibration at 1 atm of pressure.

From the position distributions of the ions in these simulations, the weighted histogram analysis method [117] yields radial potentials for K$^+$–K$^+$, K$^+$–Cl$^-$, and Cl$^+$–Cl$^+$ having 0.01 nm resolutions. These potentials include water-mediated effects and have a form similar to those used in [84]. The potentials are shifted to match the Coulomb energies at an ion separation of 1.4 nm. Beyond 1.4 nm, the Coulomb energies are used.

In the Brownian dynamics simulations, the stochastic equation of motion is integrated using a 10 fs timestep [84]. The interaction between all pairs of ions is computed by cubic interpolation of the potentials constructed above, and the diffusivities are the same as those used in the continuum models. The potential energy due to the pore $U$ is imparted by cubic interpolation from a uniform grid having a 0.03 nm resolution [130]. The exact shape of the potential used to model the pore wall is shown in Figure 5.12a.

5.2 Numerical Methods

In this section we describe a numerical approach to solving the MPNPE. For the modified Nernst-Planck equations, we use the finite element method for spatial discretization, and backward Euler for time discretiza-
tion due to the stiff nature of the solutions. Similarly, we also use a finite element discretization for solving the Poisson’s equation. The goal of our simulation is the steady state, and we evolve the solution until the measured temporal change reaches a desired tolerance.

Given some initial concentrations $c_\pm$ and electrostatic potential $\phi$, we solve (5.4) using the finite element method. The updated values of the concentrations are then used to solve (5.5), and the updated value of the electric potential is then used as input to (5.4) for the next time step. This scheme is similar to previous attempts at solving the PNPE [101].

We now proceed to describe our method in more detail (the approach is implemented in the finite element package Dolfin [98]). In the following we devise two finite element methods: one for the case in which there is only an electric field, and one in which a non-electric potential is also applied. We seek to highlight the benefits of incorporating the steric effects into the modified Nernst-Planck equations. Hence we derive our finite element method for both PNPE and MPNPE, highlighting several important differences. The methods we propose are used further when we devise a method for solving the MPNPE in the presence of an applied non-electrostatic potential.

5.2.1 Non-Linear Finite Element Method for Modified Nernst-Planck Equations

For the (modified) Nernst-Planck equations, we seek a solution in $H^1(\Omega_s) \times H^1(\Omega_s)$ for $c_\pm$, where $H^1$ is a Sobolev space with standard notation [37]. The weak form of (5.6) (which represents both (5.1) and (5.4)) is found by integrating against a test function $v_\pm = (v_+, v_-) \in [H^1(\Omega_s)]^2$: find $c_\pm \in [H^1(\Omega_s)]^2$ such that

$$
\int_{\Omega_s} \left( \frac{\partial c_\pm}{\partial t} - D_\pm \nabla \cdot \left[ \nabla c_\pm + \frac{1}{k_B T} c_\pm \nabla \phi_\pm + N_\alpha(c_\pm) \right] \right) \cdot v_\pm \, dx = 0,
$$

(5.14)

$$
\Rightarrow \int_{\Omega_s} \left( \frac{\partial c_\pm}{\partial t} \cdot v_\pm + D_\pm \left[ \nabla c_\pm + \frac{1}{k_B T} c_\pm \nabla \phi_\pm + N_\alpha(c_\pm) \right] \cdot \nabla v_- \right) \, dx = 0
$$

(5.15)

for all $v_\pm \in [H^1(\Omega_s)]^2$. From this, we apply the backward Euler method in time to (5.15) and arrive at

$$
\int_{\Omega_s} \left( \frac{c_\pm^{n+1} - c_\pm^n}{\Delta t} \cdot v_\pm + D_\pm \left[ \nabla c_\pm^{n+1} + \frac{1}{k_B T} c_\pm^{n+1} \nabla \phi_\pm + N_\alpha(c_\pm) \right] \cdot \nabla v_- \right) \, dx = 0,
$$

(5.16)
for all \( v_\pm \in [H^1(\Omega_s)]^2 \). Here, \( c^{k}_\pm \) are the concentrations at time step \( k \), while \( \Delta t \) is the length of the time step. Multiplying by \( \Delta t \) results in

\[
\int_{\Omega_s} \left( (c_{\pm}^{n+1} - c_{\pm}^n) \cdot v_\pm + \Delta t D_\pm \left[ \nabla c_{\pm}^{n+1} + \frac{1}{k_B T} c_{\pm} \nabla V_\pm + N_\alpha(c_{\pm}) \right] \cdot \nabla v_\pm \right) \, dx = 0 \tag{5.17}
\]

for all \( v_\pm \in [H^1(\Omega_s)]^2 \).

For the PNPE, (5.17) are a pair of uncoupled equations, linear in the unknown variables \( c_{\pm} \). For the MPNPE, (5.17) are a pair of coupled non-linear equations, and hence result in a more complicated form than PNPE. To address the nonlinearity, we use a straightforward application of Newton’s method to find the concentrations \( c_{\pm} \) for the MPNPE. Specifically, we define the form in (5.17) as \( \langle F_\alpha(c_{\pm}), v_\pm \rangle \):

\[
\langle F_\alpha(c_{\pm}), v_\pm \rangle = \int_{\Omega_s} \left( (c_{\pm}^{n+1} - c_{\pm}^n) \cdot v_\pm + \Delta t D_\pm \left[ \nabla c_{\pm}^{n+1} + \frac{1}{k_B T} c_{\pm} \nabla V_\pm + N_\alpha(c_{\pm}) \right] \cdot \nabla v_\pm \right) \, dx \tag{5.18}
\]

Here we have again used the \( \alpha \) notation to write the weak form for both the NP and MNP equations.

The MNPE in (5.18) are nonlinear in \( c_{\pm} \) and linear in \( v_{\pm} \). The first step in a Newton’s method is to linearize this form with respect to \( c_{\pm} \), to arrive at a bilinear form in \( w_\pm \) and \( v_{\pm} \). This is accomplished by taking a variational derivative of \( \langle F_\alpha(c_{\pm}), v_\pm \rangle \) with respect to \( c_{\pm} \),

\[
\langle DF_\alpha(c_{\pm}^{n+1}) w_{\pm}, v_\pm \rangle = \frac{d}{d\tau} \langle F_\alpha(c_{\pm}^{n+1} + \tau w_{\pm}), v_\pm \rangle |_{\tau=0}, \tag{5.19}
\]

\[
= \int_{\Omega_s} \left( w_\pm \cdot v_\pm + \Delta t D_\pm \left[ \nabla w_\pm + \frac{1}{k_B T} c_{\pm} w_\pm \nabla V_\pm \right] \cdot \nabla v_\pm + DN_\alpha(c_{\pm}) \right) \, dx, \tag{5.20}
\]

where

\[
DN_\alpha(c_{\pm}) = \alpha \frac{a^3}{(1 - c_{\pm}^{n+1} a^3 - c_{\pm}^{n+1} + a^3)^2} \left[ a^3 (w_+ + w_-) c_{\pm}^{n+1} \nabla(c_{\pm}^{n+1} + c_{\pm}^{n+1}) \right. \\
\left. - (a^3 (c_{\pm}^{n+1} + c_{\pm}^{n+1}) - 1) (\nabla(c_{\pm}^{n+1} + c_{\pm}^{n+1}) w_{\pm}) \right. \\
\left. + \nabla(w_+ + w_-) c_{\pm}^{n+1} \right] \cdot \nabla v_\pm. \tag{5.21}
\]

As a consequence, with \( \alpha = 0 \) for PNPE, the derivative \( DN_0 \) is absent from (5.20).

With the derivative defined, we arrive at Newton’s method, which first computes \( w_\pm \) by solving

\[
\langle DF_\alpha(c_{\pm}^{n+1}) w_{\pm}, v_\pm \rangle = -(F_\alpha(c_{\pm}), v_\pm) \tag{5.22}
\]
followed by the estimate of $c_n^{n+1}$ with

$$c_n^{n+1} = c_n^{n+1} + w_\pm. \quad (5.23)$$

Solving the weak problem in (5.22) reduces to a linear system of the form $Aw = f$, where stiffness matrix $A$ is formed by $\langle DF_\alpha(c_n^{n+1})w, v_\pm \rangle$, the force vector $f$ by $\langle F_\alpha(c_\pm), v_\pm \rangle$, and where $w$ represents the unknown coefficients for the function $w_\pm$.

For the NP equations, $\alpha = 0$, and so the form $DF_\alpha$ in (5.22) does not depend on $c_n^{n+1}$, and the Newton’s method converges in one iteration (as we expect for linear problems). For simplicity, we drop the arguments to the forms $DF_\alpha$ and $F_\alpha$. Then $DF_0$ and $DF_1$ represent the forms in (5.20). Similarly $F_0$ and $F_1$ are the forms for NP and MMP in (5.18). The relationship between these forms is expressed as,

$$DF_1 = DF_0 + DN_1, \quad (5.24)$$
$$F_1 = F_0 + N_1 \cdot \nabla v_\pm, \quad (5.25)$$

which allows us to view the Newton step for MNP as,

$$DF_0 + DN_1 = -(F_0 + N_1 \cdot \nabla v_\pm). \quad (5.26)$$

From this equation, we notice that the Newton step for the MPNPE is exactly the Newton step for the PNPE with the addition of stabilizing terms $DN_1(c_\pm)$ and $N_1(c_\pm) \cdot \nabla v_\pm$.

### 5.2.2 A Stabilized Finite Element Method for the Modified Nernst Planck Equations

The discretization scheme we have developed so far works well if there is no applied non-electrostatic potential — i.e. $U = 0$. However, in the presence of a non-electrostatic applied potential, the Nernst-Planck equations have a strong advective or drift term, which is a challenge to standard Galerkin methods. Using the standard Galerkin approach results in a solution with spurious values [30] — e.g. the concentration becomes negative in portions of the domain. One remedy is to augment the Galerkin weak form by adding artificial dissipative terms to stabilize the method. To this end, we use a variant of streamline upwind Petrov-Galerkin method (SUPG) for stabilizing our scheme in presence of steep gradients in $U$ [80, 81].

We develop stabilized schemes for the PNPE and MPNPE. Specifically, we develop two SUPG schemes for the MPNPE, one that arises from the standard application of SUPG to the MPNPE, and one by adding
the nonlinear terms of the MPNPE to the SUPG scheme for the PNPE. The latter scheme, which we call “Fast SUPG” improves on the former, called the “Full SUPG” scheme, by increasing its efficiency, as we explain later. We use the “Fast SUPG” method in this chapter, unless we specify otherwise. The relationship between different SUPG schemes is shown in Figure 5.2. To simplify the presentation of our SUPG scheme, we introduce some notation. The differential operator is given as

$$\mathcal{L}_\alpha(c_\pm) = D_\pm \nabla \cdot \left[ \nabla c_\pm + \frac{1}{k_B T} c_\pm \nabla V_\pm + N_\alpha(c_\pm) \right],$$  \hspace{1cm} (5.27)

which consists of a flow field governed by the applied potential $V_\pm$ and denoted by

$$b_\pm = -\frac{D_\pm}{kT} \nabla V_\pm.$$  \hspace{1cm} (5.28)

We also isolate the Péclet number $Pe$ with a stability parameter of the form

$$\sigma_\pm = \frac{h_\tau}{2\|b_\pm\|_2} \psi(Pe_\tau),$$  \hspace{1cm} (5.29)
where $h_\tau$ denotes the diameter of the element $\tau$, and with

$$\text{Pe}_\tau = \left( \frac{0.33 \, \|b_\pm\|_2 h_\tau}{2D_\pm} \right) \text{ and } \psi(q) = \begin{cases} 1 & \text{if } q > 1 \\ q & \text{otherwise} \end{cases}.$$  \hfill (5.30)

The Péclet number of element $\tau$, denoted as $\text{Pe}_\tau$ is an indication of the strength of advection. Specifically, a Péclet number greater than 1 indicates that advection is dominating the flow and that stabilization may be necessary, and we use the values developed in [65, 30].

Using the above notation, we write (5.14) as,

$$\int_{\Omega_s} \left( \frac{\partial c_\pm}{\partial t} - L_\alpha(c_\pm) \right) \cdot v_\pm \, dx = 0.$$  \hfill (5.31)

From this, we define the SUPG weak form using integration-by-parts similar to the derivation in (5.14), to arrive

$$\langle F_{\alpha,\text{supg}}(c_\pm), v_\pm \rangle = \int_{\Omega_s} \left( \frac{\partial c_\pm}{\partial t} - L_\alpha(c_\pm) \right) \cdot v_\pm \, dx + \sum_{\tau \in T} \int_{\tau} \left( \frac{\partial c_\pm}{\partial t} - L_\alpha(c_\pm) \right) \cdot v_{\pm,\text{supg}} \, dx = 0,$$  \hfill (5.32)

where we write the test functions in stabilized form are

$$v_{\pm,\text{supg}} = \sigma_\pm b_\pm \cdot \nabla v_\pm.$$  \hfill (5.33)

For the PNPE, (5.32) is solved as a pair of uncoupled equations. However, for the MPNPE, we use a nonlinear solver, as discussed in the previous section. The nonlinear SUPG scheme is similar to the approach taken in [57] for use with nonlinear Navier-Stokes equations.

**A Faster SUPG Scheme for MPNPE**

The SUPG method arising from (5.32) stabilizes the MPNPE, however it is costly to implement. This cost arises because of the presence of the strong form of the operator $L_1$ in the “stabilization” part of (5.32) — i.e., we do not use integration by parts. Thus, one of the terms in the weak form $\langle F_{1,\text{supg}}(c_\pm), v_\pm \rangle$ is $\nabla \cdot N_1(c_\pm)$, which is lengthy, and is costly to evaluate. Moreover, in Newton’s method, the variational derivative of this term is needed, which makes the implementation even costlier. By considering a Newton’s method for the SUPG method for NPE we are able to derive a more efficient scheme for the modified form of the equations.
To further motivate this, consider the Newton step for the Galerkin method for MPNPE, as given in (5.26). We observe that the Newton step for MPNPE in (5.26) is the Newton step for PNPE along with additional stabilization terms. Thus, we take a similar view in designing the faster SUPG method for MPNPE: first we form the Newton step for the SUPG method for the PNPE followed by the addition of stabilization terms.

As before, to form the Newton step for NPE, we take the variational derivative of the form $F_{0,\text{supg}}$ to arrive at the form $\langle DF_{0,\text{supg}}(c_{\pm}), w_{\pm}, v_{\pm} \rangle$. The Newton iteration for the SUPG method for the NPE is then to find functions $w_{\pm} \in [H^1(\Omega_s)]^2$ such that

$$\langle DF_{0,\text{supg}}(c_{\pm}), w_{\pm}, v_{\pm} \rangle = -\langle F_{0,\text{supg}}(c_{\pm}), v_{\pm} \rangle. \quad (5.34)$$

From $w_{\pm}$, we then update using (5.23): $c_{\pm}^{n+1} = c_{\pm}^{n+1} + w_{\pm}$.

As in the previous case, for the unmodified form of the NPE, Newton’s method reduces to one iteration. For the MNPE however, we add additional terms present in Newton’s method for the Galerkin method in (5.26), namely $DN_1(c_{\pm})$ and $N_1(c_{\pm}) \cdot \nabla v_{\pm}$, to (5.34). As a result, the SUPG method for the MNPE is defined by the following Newton’s step,

$$\langle DF_{0,\text{supg}}(c_{\pm}), w_{\pm}, v_{\pm} \rangle + DN_1(c_{\pm}) = -\langle F_{0,\text{supg}}(c_{\pm}), v_{\pm} \rangle - N_1(c_{\pm}) \cdot \nabla v_{\pm}. \quad (5.35)$$

We use the SUPG form in (5.35) in our numerical test in Section 5.3

### 5.2.3 Galerkin Method for the Poisson Equation

A stabilized scheme is unnecessary for the Poisson portion of the PNPE in (5.5). Thus, we use a standard Galerkin finite element method which is stated as the following: find $\phi \in \phi_d + H^1_0(\Omega)$ such that,

$$\int_{\Omega} \nabla \phi \cdot \nabla v \, dx = \int_{\Omega} e(c_+ - c_-) v \, dx, \quad (5.36)$$

for all $v \in H^1_0(\Omega)$ and where $\phi_d$ is a function satisfying Dirichlet boundary conditions.

### 5.3 Numerical Experiments

In this section we illustrate the use of our finite element method through four numerical experiments. The first experiment examines the steric effects in an electrolyte solution confined between two charged plates.
Then, we apply our method to compute the ionic current through a nanopore in a solid-state membrane in the absence of a non-electrostatic potential and a DNA molecule. We follow with a third experiment in which we introduce a non-electrostatic potential into the ionic current calculations, mimicking the effect of a realistic nanopore surface. This system highlights the significant difference between MPNPE and PNPE calculations. Furthermore, we relate the results of PNPE and MPNPE calculations to the results of Brownian dynamics simulation. Finally we examine the steric effects in the nanopore system when a DNA molecule (modeled as a cylinder) is present in the nanopore.

For the steady state experiments, we evolve an initial uniform concentration until the temporal change in the concentration is less than a fixed tolerance. More specifically, the temporal change is measured as

\[
\frac{1}{N} \sum_{i=1}^{N} \left| \frac{c_{\pm}^{n+1} - c_{\pm}^n}{\Delta t} \right| \leq \theta, \tag{5.37}
\]

where \( N \) is the number of nodes in the mesh, and \( n \) refers to the temporal time step. For example, in the case of the nanopore in Experiments 2 and 3, a tolerance of \( \theta = 10^{-2} \) is used, while we enforce a tolerance of \( 10^{-3} \) in the case of DNA in Experiment 4. With a tolerance of \( 10^{-2} \), the computations remain generally static and are representative of the final steady state. Higher concentrations demand increased resolution and thus require more time steps, as shown in Table 5.1.

<table>
<thead>
<tr>
<th>Experiment:</th>
<th>Concentration ((c_+):)</th>
<th>#2</th>
<th>#3</th>
<th>#4</th>
</tr>
</thead>
<tbody>
<tr>
<td>PNPE</td>
<td>0.1082M</td>
<td>17</td>
<td>74</td>
<td>1672</td>
</tr>
<tr>
<td>MPNPE</td>
<td>0.1082M</td>
<td>17</td>
<td>74</td>
<td>1672</td>
</tr>
<tr>
<td></td>
<td>1.623M</td>
<td>50</td>
<td>149</td>
<td>1672</td>
</tr>
<tr>
<td></td>
<td>2.705M</td>
<td>125</td>
<td>155</td>
<td>305</td>
</tr>
<tr>
<td></td>
<td>1.623M</td>
<td>120</td>
<td>155</td>
<td>305</td>
</tr>
<tr>
<td></td>
<td>2.705M</td>
<td>116</td>
<td>155</td>
<td>305</td>
</tr>
<tr>
<td></td>
<td>1.559M</td>
<td>116</td>
<td>155</td>
<td>305</td>
</tr>
</tbody>
</table>

Table 5.1: Number of temporal iterations to reach steady state.

One of the goals in the numerical experiments is to highlight the utility of using the modified form of the equations. To this end, we present the sum of concentrations — i.e. \( c = c_+ + c_- \) — to illustrate the behavior of the solutions with respect to the steric limit of the problem.

**Choice of the Effective Ion Size Parameter \( a \)**

In order to compute concentrations from the MPNPE an effective ion size \( a \) in (5.4) needs to be determined. The effective ion diameter captures the size of the ion in its solvated state, and may be larger than its bare ion size \([19]\). Moreover, the effective ion size in the case of same-sign ion crowding differs from the effective ion size in the presence of oppositely charged crowding ions, as argued in \([19]\). In our experiments, we encounter
both situations. In Experiment 1 with charged metal plates, and in Experiment 4 with a charged cylinder, there is a crowding of ions of the same charge. Hence, we consider the effective ion size for these situations by the effective solvated diameter of 0.66 nm from Table 2 in [19] for $K^+$ and $Cl^-$. In the other cases, the crowding of ions is due to the non-electrostatic potential, which is same for both ion species. As a result, the mixture remains essentially neutral, and we thus set the effective ion size in Experiment 3 to the value that reproduces the results of our BD simulations. We find this ion size to be 3 nm, and we justify it below.

Validating the Fast SUPG Scheme and the Ion Size

We verify that our fast SUPG scheme in (5.35) results in approximations similar to that of the full SUPG scheme in (5.32), and that using an ion size of 3 nm is the correct selection. To support this, we perform computational experiments on a box domain of dimensions 1 nm $\times$ 1 nm $\times$ 2 nm with an applied, non-electrostatic potential $U$. The cross-section of the box domain is shown in Figure (5.3a). The profile of this potential along the $z$-axis is shown in Figure 5.3b. Since there is no applied electrostatic potential, the system converges rapidly to a steady state. Hence, we use a tolerance of $10^{-5}$ in the temporal change of concentration for the experiments with the box; in this case, experiments generally required fewer than 10 iterations to converge to the steady state.

![Cross-section of the box domain.](image1)

![Profile of the non-electrostatic potential $U$ $(kT$ units) along the $z$-axis.](image2)

**Figure 5.3:** Cross-section (left) and profile of non-electrostatic potential (right) for the box domain.

We compare our results for the box domain with Brownian Dynamics (BD) simulations for these systems,
which capture the effects of explicit ion-ion interaction. The steady state concentrations obtained from BD, PNPE, and MPNPE are illustrated in Figure 5.4, which shows the results for systems that contain 3 $K^+$ ions and 3 $Cl^-$ ions. Consequently, when using an ion size of 0.3 nm, the concentration profile accurately matches the BD solution. As a comparison, we also plot the profile for the MPNPE with an ion size of 0.36 nm and a profile using the PNPE. Using an ion size of 0.36 nm lowers the peak, while solving the PNPE, which ignores steric effects, produces a much higher peak, as expected.

![Concentration profile](image)

**Figure 5.4:** Profile of concentration $c_+$ for BD, PNPE, and MPNPE for the box domain.

To further justify the 0.3 nm size, we repeat the same experiment using 4 and 5 ions of each K and Cl. Similarly, the results shown in Figure 5.5 demonstrate that an ion size of 0.3 nm captures the profile given by BD simulations.

Next, we verify that our fast SUPG scheme (5.35) yields results similar to the full SUPG scheme (5.32). We show the similarity for the above box domain, in an experiment with 3 ions of each species, in Figure 5.6a, where we show the profiles using standard Galerkin, full SUPG, and our fast SUPG schemes. For this experiment, we use a fine enough mesh to ensure that the standard Galerkin method results in accurate approximations. However, for more realistic domains this is a significant computational limitation, since highly refined meshes decrease the efficiency of the continuum models, which is one of their main advantages. We also show the similarity of the concentration profiles for the nanopore experiment in Figure 5.6b, in a system with 60 ions of each species (see Experiment 3 for details). The close agreement of the profiles in the figures indicates that our fast SUPG scheme accurately computes the concentrations in comparison to the other expensive methods, and we proceed with its use through the remainder of the chapter.

Finally we demonstrate that the fast SUPG scheme results in significant savings in computational effort.
Figure 5.5: Profile of concentration $c_+$ for BD and MPNPE with an ion size of 0.3 nm for the box domain.

Figure 5.6: Profile of concentration $c_+$ for different numerical schemes for MPNPE.

Table 5.2 shows the speedup of the fast scheme as compared to the full scheme for the box and pore domains. The table illustrates the utility of the fast scheme by showing that the fast scheme speeds up the solution process by a factor of more than 2 for the nanopore, and more than 4 for the box domain.

<table>
<thead>
<tr>
<th>System</th>
<th>Box - 3 ions</th>
<th>Box - 4 ions</th>
<th>Nanopore - 60 ions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speedup</td>
<td>4.47</td>
<td>4.40</td>
<td>2.26</td>
</tr>
</tbody>
</table>

Table 5.2: Speedup of the fast SUPG scheme versus the full SUPG scheme.
Experiment 1: Ionic Solution Between Two Charged Plates

In this experiment, we consider a 1 nm×1 nm×4 nm parallelepiped domain placed between two charged plates, as shown in Figure 5.7. At the beginning of the numerical experiments, the concentration of ions is uniform in the solvent domain. We solve the PNPE and the MPNPE for the concentration of ions in the presence of an electrostatic potential difference between the plates. Specifically, the upper plate is set at a potential of -800mV and the lower plate at +800mv, leading to a total potential difference of -1600mV — see Figure 5.7. In these calculations, we use blocking boundary conditions for the ionic concentrations throughout the domain.

![Figure 5.7: Cross-section of the parallel plates domain considered in Experiment 1.](image)

The parameters for this numerical experiment are given in Table 5.3. Specifically, an ion size of 0.66 nm in diameter corresponds to an effective size of K or Cl ion, which is larger than the ion size in the bulk solution, an approximation that is known to work well at high charge densities [19]. Consequently, the large ion size captures the crowding effect of ions due to a large voltage, resulting in a scenario where we expect to see the effects of the steric limit.

From Figures 5.8a and 5.8b, which show 2-D cross-sections of the concentration after 149 times steps of the simulation, we see that the steric limit is violated in the case of PNPE (colored red in the figure), but that the concentration remains appropriately bounded in the case of MPNPE (Figure 5.8c), even after 700 iterations. Furthermore, Figure 5.9 shows that the local sum of the concentrations along the z-axis of the domain is unrealistically high in the case of the PNPE, whereas the sum for the MPNPE remains bounded and within steric limit. Computationally, we also observe that consecutive approximations to the PNPE are increasingly difficult to compute as the simulation progresses due to the sharp gradients, whereas
<table>
<thead>
<tr>
<th>Name</th>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time step</td>
<td>$\Delta t$</td>
<td>$10^{-13}$ s</td>
</tr>
<tr>
<td>Diffusivity for positive ions</td>
<td>$D_+$</td>
<td>$2.27 \times 10^{-9}$ m$^2$/s</td>
</tr>
<tr>
<td>Diffusivity for negative ions</td>
<td>$D_-$</td>
<td>$2.27 \times 10^{-9}$ m$^2$/s</td>
</tr>
<tr>
<td>Relative Permittivity</td>
<td>$\epsilon$</td>
<td>92</td>
</tr>
<tr>
<td>Ion diameter</td>
<td>$a$</td>
<td>0.66 nm</td>
</tr>
<tr>
<td>Initial concentration</td>
<td>$c_{init}$</td>
<td>1 M</td>
</tr>
<tr>
<td>Potential Drop</td>
<td>$u_{top} - u_{bottom}$</td>
<td>-1600 mV</td>
</tr>
<tr>
<td>Steric Limit for 0.66A</td>
<td>$\nu$</td>
<td>5.776 M</td>
</tr>
</tbody>
</table>

**Table 5.3:** Summary of parameters in Experiment 1.

approximations to solutions of the MPNPE maintain a more subtle profile.

**Figure 5.8:** The sum of concentrations $c = c_+ + c_-$ for the PNPE and the MPNPE.
Figure 5.9: Comparison of the solutions of PNPE and MPNPE for the parallel plates domain system. The local sum of positive and negative ion concentrations is plotted along the $z$-axis. A red hue indicates a violation of the steric limit.

Experiment 2: Ionic Current through a Nanopore in the Membrane

In this experiment, our aim is to accurately compute the current through a nanopore in a membrane. The system setup is shown in Figure 5.1b. A three dimensional view of the solvent domain $\Omega_s$ is given in Figure 5.10, where the dimensions of the full domain, $\Omega_s \cup \Omega_m$ are $4 \text{ nm} \times 4 \text{ nm} \times 7.2 \text{ nm}$, while the length of the pore (modeled as a cylinder) is $4 \text{ nm}$ with a radius of $0.9 \text{ nm}$. We summarize additional parameters for this experiment in Table 5.4.

<table>
<thead>
<tr>
<th>Name</th>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diffusivity for positive ions</td>
<td>$D_+$</td>
<td>$2.27 \times 10^{-9} \text{ m}^2/\text{s}$</td>
</tr>
<tr>
<td>Diffusivity for negative ions</td>
<td>$D_-$</td>
<td>$2.41 \times 10^{-9} \text{ m}^2/\text{s}$</td>
</tr>
<tr>
<td>Relative Permittivity of membrane</td>
<td>$\epsilon_m$</td>
<td>92</td>
</tr>
<tr>
<td>Relative Permittivity of solvent</td>
<td>$\epsilon_s$</td>
<td>92</td>
</tr>
<tr>
<td>Ion diameter</td>
<td>$a$</td>
<td>0.3 nm</td>
</tr>
<tr>
<td>Potential Drop</td>
<td>$u_{\text{top}} - u_{\text{bottom}}$</td>
<td>-180 mV</td>
</tr>
<tr>
<td>Temperature</td>
<td>$T$</td>
<td>295 K</td>
</tr>
</tbody>
</table>

Table 5.4: Summary of parameters used in Experiment 2.

Periodic boundary conditions are applied in all directions for the concentrations, except for the boundary of the pore, $\partial \Omega_{s,n}$ (see Figure 5.1b), where blocking conditions are applied. For the potential, we enforce periodic conditions in the $x$ and $y$ direction, along with Dirichlet conditions at the top and bottom of the domain — i.e., the top of the domain is set at a potential of -90mV and the bottom at +90mV.
The ion size is taken to be the approximate size of the K or Cl ions in the bulk [19]. In this setting, the PNPE do not exhibit steric effects. Consequently, we expect to see similar results for PNPE and MPNPE, which we examine by computing steady state concentration profiles and by computing the ionic currents. The difference between the steady state concentrations for the positively charged ions for a uniform initial concentration of 1.623 M computed by PNPE and MPNPE is shown in Figure 5.11. The figure shows that both MPNPE and PNPE result in similar concentration profiles, with the majority of the differences appearing near the channel edges where the gradient is the largest.

We consider the circular slice through the center of the channel which has a unit-normal vector in the $z$-coordinate direction, for the calculation of the electric current in (5.9). The values are shown in Table 5.5 where we see that the concentrations vary below the steric limit and that MPNPE matches PNPE closely in terms of computed current, as expected.

<table>
<thead>
<tr>
<th>Concentration (M)</th>
<th>PNPE Current (pA)</th>
<th>MPNPE Current (pA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1082</td>
<td>198.10</td>
<td>198.00</td>
</tr>
<tr>
<td>1.623</td>
<td>2592.78</td>
<td>2590.12</td>
</tr>
<tr>
<td>2.705</td>
<td>4275.77</td>
<td>4272.34</td>
</tr>
</tbody>
</table>

Table 5.5: Currents (pA) for PNPE and MPNPE.
Figure 5.11: Difference in approximating $c_+^{}$ for PNPE and MPNPE.

**Experiment 3: Ionic Current through a Nanopore in the Presence of a Non-Electrostatic Potential**

Next, we consider calculation the ionic current through the nanopore domain described in Figures 5.1b and 5.10. In this experiment we apply a non-electric potential at the boundary of the membrane and the solvent domains denoted $\partial \Omega_{s,n}$. The dimensions of the domain and parameters for the experiment are the same as specified in Table 5.4 for Experiment 2, with the exception of the diffusivities $D_+$ and $D_-$, which are used for the computation of the ionic currents. The diffusivities are obtained from the BD simulations with the same setup, and are listed in Table 5.6.

<table>
<thead>
<tr>
<th>Atoms</th>
<th>$D_+$ (m$^2$/s)</th>
<th>$D_-$ (m$^2$/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>1.90584</td>
<td>1.96144</td>
</tr>
<tr>
<td>100</td>
<td>1.71593</td>
<td>1.72708</td>
</tr>
</tbody>
</table>

*Table 5.6:* Diffusivities used for the ionic current computations in Table (5.7).

To illustrate the applied non-electric potential for this experiment, we display a plot that is parallel to the $z$-axis at $(x,y) = (1.5 \text{ nm}, 1.5 \text{ nm})$ in Figure 5.12a. The upper wall of the pore is at $z = 2 \text{ nm}$, where the applied potential is large, thus resulting in a highly advective concentration. Consequently, the traditional
Galerkin method is not stable, and we apply our stabilized SUPG method to arrive at the solution. To illustrate the instability in the Galerkin method, the element-wise Péclet number (defined in Section 5.2.2) along the $z$-axis at $(x, y) = (1.5 \text{ nm}, 1.5 \text{ nm})$ is shown in Figure 5.12b. As the Péclet number surpasses a value of 1.0 near the pore, instabilities are more likely in the Galerkin method (see Section 5.2.2), thus requiring use of our SUPG scheme.

![Figure 5.12](image)

**Figure 5.12:** Profiles parallel to the $z$-axis at $x = y = 1.5 \text{ nm}$, demonstrating a high non-electrostatic potential near the membrane wall at $z = 2 \text{ nm}$, which separates the solvent region from the membrane, for the nanopore domain defined in Figure 5.1b. This particular profile is shown since it highlights the sharpest repulsive gradients present in the solution, which cause spurious modes in the Galerkin approximation.

For the computation of the ionic current through the nanopore we use 60 and 100 ions, which correspond to uniform initial concentrations of 1623.26 mM and 2705.43 mM for each ion species. The profiles of the concentration of the positive ions, $c_+$, in the steady state solution, for the PNPE and MPNPE along the $x$-axis are shown in the nanopore system Figure 5.13. The figure indicates that MPNPE captures the steric effects at high concentrations, whereas, PNPE shows high concentrations in regions of large potential gradients.

Furthermore, the calculated electric currents for the PNPE and MPNPE, as shown in Table 5.7, are lower in the case of MPNPE. It is important to note that current calculations are insensitive to small variations in the concentration since the form of the current calculation is an integral. Thus, while the difference in the PNPE and MPNPE current calculations are only around 2%, this represents a significant deviation in the concentration as supported by Figure 5.13. The ionic currents predicted by both PNPE and MPNPE are in reasonable agreement with the results of the BD simulations.
Figure 5.13: Concentration of positively charged ions \((c_+\)) along the \(x\)-axis (at \(y, z = 0\) nm) in the nanopore system (Figure 5.1b) with an applied, non-electric potential. The overall ion concentration is equivalent to the presence of 100 ions in the solvent domain.

<table>
<thead>
<tr>
<th>Atoms</th>
<th>Concentration (M)</th>
<th>BD Current (pA)</th>
<th>PNPE Current (pA)</th>
<th>MPNPE Current (pA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>1.623 ± 9</td>
<td>2170</td>
<td>2421.54</td>
<td>2396.30</td>
</tr>
<tr>
<td>100</td>
<td>2.705 ± 18</td>
<td>2960</td>
<td>3575.12</td>
<td>3514.29</td>
</tr>
</tbody>
</table>

Table 5.7: Currents (pA) for BD, PNPE and MPNPE.

Experiment 4: Ionic Current Through a Nanopore Blocked by a Charged Cylinder

Finally, we consider the presence of a DNA strand in the nanopore. Here, we approximate a DNA molecule by a cylinder with a uniformly distributed constant surface charge density. While this is a simplistic representation, it captures the essential features of the system, including the narrow channels in the solvent domain surrounding the DNA that create sharp gradients and large values in the concentration. As a result, this experiment underscores the need for a successful numerical methods for the modified form of the NPE.

In these computational experiments, the dimensions of the full domain, \(\Omega_s \cup \Omega_m\), are \(8\) nm \(\times\) \(8\) nm \(\times\) \(15\) nm, while the length of the pore (modeled as a cylinder) is \(5\) nm with a radius of \(1.81\) nm. An electrostatic potential difference of \(-120\) mV is applied from the top to bottom. Furthermore, due to the presence of the charged cylinder, we also impose Neumann boundary condition on the surface of the cylinder, \(\Gamma_{DNA}\) [34], of the form

\[ \epsilon_s \nabla u \cdot n = -\sigma. \]  

(5.38)

The remaining parameters of this DNA cylinder are summarized in Table 5.8, and a cross-section of the
domain in the $yz$ plane is shown in Figure 5.14. We see that the presence of the DNA cylinder (represented by a red region) creates a narrow channel in the pore, producing spikes in the concentration profiles that necessitates careful treatment of steric effects.

\[ \Omega_s \neg \Omega_m \neg \partial \Omega_{s,n} \neg \phi_t \neg \phi_b \neg \Gamma_{DNA} \]

\[ z \ (\text{nm}) \]

\[ x \ (\text{nm}) \]

$\Omega_s = \Omega_m = 2.5, x = -1.81, 0, 1.81$

**Figure 5.14:** Cross section of the nanopore system with a DNA cylinder.

In this experiment, we use different initial uniform concentrations for the positive and negative ions to maintain overall electric neutrality of the system and we set $U = 0$, so that the non-electrostatic potential does not influence the results. The initial concentrations and the computed electric currents are summarized in Table 5.9.

<table>
<thead>
<tr>
<th>Name</th>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radius of DNA cylinder</td>
<td>$a$</td>
<td>0.66 nm</td>
</tr>
<tr>
<td>Surface charge density of cylinder</td>
<td>$\sigma$</td>
<td>-0.83876 e/nm$^2$</td>
</tr>
<tr>
<td>Ion diameter</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 5.8:** Parameters for experiment 4.

From these numerical experiments, we observe PNPE currents that are unrealistically high and that are in violation of the steric limit, which is in stark contrast to the currents obtained through MPNPE. Indeed, Figure 5.15 shows that the sum of concentrations for the PNPE and MPNPE differs greatly in the channel, as expected. Furthermore, from the cross-sectional view of the domain, shown in Figure 5.16, it is evident
that the concentration produced by solving the PNPE exhibits large deviations from the steric limit near the cylinder.

<table>
<thead>
<tr>
<th>$c_+ \text{ (M)}$</th>
<th>$c_- \text{ (M)}$</th>
<th>PNPE Current (pA)</th>
<th>MPNPE Current (pA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.559</td>
<td>1.319</td>
<td>2966.7</td>
<td>2815.0</td>
</tr>
</tbody>
</table>

**Table 5.9:** Nanopore ionic currents (pA) computed by PNPE and MPNPE.

**Figure 5.15:** 2D density plot of sum of concentrations (M) $c = c_+ + c_-$ in the nanopore/DNA system.

### 5.4 Conclusions

The modified Poisson-Nernst-Planck equations account for steric effects and lead to physically realistic results. In this chapter we develop a finite element method for the MPNPE, with a focus on computing steady state concentrations. We achieve this by evolving the system forward in time, thus ensuring that the number of ions in the system remains conserved. The presence of a high potential gradient near the wall of the membrane causes instabilities in the Galerkin finite element method, which is indicated by high Péclet number. In response, we stabilize the solution by employing a SUPG-type finite element method. Initially our SUPG
method augments the Galerkin weak form with a strong form residual. However, this strong form residual is expensive to evaluate for the MPNPE. Thus, we design a new SUPG method for the MPNPE which is derived from the relationship between the weak Galerkin forms of the PNPE and the MPNPE. We highlight our finite element solver for a variety of experiments, first showing the utility of MPNPE in the case of flow between two oppositely charged plates. We then consider the flow of ions through a nanopore, where we determine the ion size using BD simulations. Finally, we explore the flow of ions around a DNA molecule, modeled as a charged cylinder. Our results indicate that the MPNPE account for steric effects, and yield more physically meaningful solutions.
Chapter 6

A Goal-Oriented Approach to Least-Squares Finite Element Method

Previous chapters have focused on efficient finite element methods for implicit solvent models, with particular emphasis on reducing global errors. In this chapter, we introduce a mechanism for targeting particular quantities in least-squares finite element approximations, e.g., the solvation free energy. In many simulations, local calculations and quantities are often a primary objective. To this end, we develop an approach for least-squares finite element methods to incorporate local quantities-of-interest.

Consider a first-order system of partial differential equations

\[ L u = f \quad \text{in } \Omega, \quad (6.1) \]

where \( L \) is a quasi-linear operator. We seek to minimize the least-squares functional

\[ F(u; f) = \| L u - f \|^2, \quad (6.2) \]

over all \( u \in V \), where \( \| \cdot \| \) denotes the \( L^2 \)-norm. We denote the unique minimizer of (6.2) by \( u^* \), which satisfies the weak equation

\[ \langle L u - f, L v \rangle = 0 \quad \forall v \in V. \quad (6.3) \]

Least-squares formulations offer many attractive properties. The discrete problems follow directly from (6.3) and the resulting systems are symmetric and positive-definite. Moreover, the least-squares functional itself, (6.2), is an effective error estimator, since the error \( \| u - u^* \|_V \) is equivalent to

\[ F(u, f) = F(u - u^*; 0). \quad (6.4) \]

This estimator is shown to be sharp in certain settings [23] and is often successful in practice [1].

Many applications are primarily interested in a particular aspect of a solution, such as the lift on a wing, conservation of mass or energy, positivity of the solution, a calculated expression over a region, etc. These
examples are expressed as integral quantities of the solution and are not directly considered by a standard least-squares formulation. The focus of this chapter is to develop a methodology that maintains an accurate solution to the minimum of (6.2), while targeting additional accuracy in a particular quantity-of-interest (QOI). We denote the QOI or goal as $Q$, and assume that the exact solution to (6.1) satisfies

$$Q(u^*) = \gamma.$$  \hfill (6.5)

From (6.5), there are two cases to consider: when $\gamma$ is known a priori and when the value of $\gamma$ is unknown. For example, $\gamma$ is known in the case of conserved quantities, such as the flow through a cross-section when the inflow is specified. In contrast, in applications such as the lift on an airfoil, the value of $\gamma$ is unknown prior to computation (and the quantity itself is an important result). In this chapter we present a framework that handles the more general scenario of an unknown $\gamma$, however we do consider the case of known quantities of interest when the resulting analysis is simplified and enlightening.

To account for the QOI in (6.5), we add a term to functional (6.2) and consider the following

$$F_Q(u) = \|Lu - f\|^2 + \beta^2 |Q(u) - \gamma|^2.$$  \hfill (6.6)

Minimizers to (6.2) still minimize (6.6). However, because of the inclusion of the goal, the new functional assesses the quality of the global approximation as well as the fitness of the approximation to the quantity of interest. As a result, with this new functional, we are able to guide the approximation process to the original least-squares solution, but now along a particular direction that yields an improved prediction of the QOI. As we show in later sections, we are able to maintain the global convergence properties of the original formulation, but we inherit a constraint by the inclusion of the QOI. The result is an improved QOI without sacrificing global approximation properties.

The form of (6.6) is directly fit for computation if the value of $\gamma$ is known. However, for unknown $\gamma$, the process is not clear. Consequently, in Section 6.2, we develop a duality argument that yields a practical method. The concept of duality is commonly used in finite element analysis, such as the Aubin-Nitsche Lemma, which relies on duality to define an a priori bound on the $L^2$ error [37]. Furthermore, the use of duality arguments to develop algorithms for goal-oriented adaptivity and error estimation has been developed in other contexts (e.g. [9, 10, 11]). These techniques apply the Riesz Representation Theorem to relate the goal-functional to an adjoint function. This approach relates an adjoint function to a domain-of-influence [20, 58] of the functional. Based on this approach, methods have been derived resulting in error estimates and indicators.
suitable for goal-oriented adaptive mesh refinement, including the popular dual-weighted residual method [20]. Goal-oriented refinement has also been applied to a large number of applications, including elasticity, fluid-flow, fluid structure interaction, and molecular biophysics [122, 73, 44, 126].

In this chapter, we develop the theoretical and algorithmic framework for efficiently including QOIs into the least-squares minimization process. The outline of the chapter is as follows. In Section 6.1, we outline several immediate observations for the goal-based functional (6.6), including existence, uniqueness, and implications from a linear algebra point-of-view. In Section 6.2 we develop bounds on the error for (6.6) and, from the theory, introduce a modified functional. A schematic of this approach is depicted in Figure 6.1 where an approximated form of (6.6) is introduced to account for the unknown quantity $\gamma$. Specifically, a functional of the form

$$F_{Q^*} = \| Lu - f \|^2 + \beta^2 |Q(u) - \gamma_h|^2$$

(6.7)

is introduced where $\gamma_h$ is approximated using a subspace $W^h \subset V$, whereas the approximation $u^h \in V^h \subset V$. An additional modification results in

$$F_{Q^h} = \| Lu - f \|^2 + \beta^2 |Q^h(u) - \gamma_h|^2$$

(6.8)

by approximating the QOI with $Q^h$ with the same space $W^h$. We derive these functionals in detail in Section 6.2. In Section 6.3 we follow with improved error bounds for these functionals and in Section 6.4 we bound the error in the QOI. Finally, in Section 6.5 we outline an adaptive refinement algorithm based on our approach and in Section 6.6 we support our approach with numerical result for several application areas.

**Figure 6.1:** Relationship between functionals.
6.1 Assumptions, Immediate Theory and Discrete Motivation

To further elucidate our proposed methodology, in this section we consider the relationship between the standard least-squares functional \( F \) in (6.2) and the goal-based least-squares functional \( F_\ast \) in (6.6), as well as their associated discrete forms.

6.1.1 Assumptions and Notation

We first outline several properties of the spaces and functionals that are assumed throughout the manuscript. We denote by \( \| \cdot \| \) or \( \| \cdot \|_0 \) the standard \( L^2 \)-norm over \( \Omega \) and use standard notation for Sobolev spaces. We also make frequent use of generic constants that are independent of the meshing, and when the specific constant is irrelevant we denote it by \( C \).

The standard least-squares functional is required to maintain coercivity and continuity over \( V \), which is summarized by the following assumption.

**Assumption 1 (Continuity and Coercivity).** Given the least-squares functional \( F(u; f) : V \to \mathbb{R} \) defined in (6.2), assume there exists constants \( c_0 \) and \( c_1 \) such that

\[
F(u; 0) = \langle L u, L u \rangle \geq c_0 \| u \|^2_V \quad \forall u \in V, \tag{6.9}
\]

and

\[
\langle L u, L v \rangle \leq c_1 \| u \|_V \| v \|_V \quad \forall u, v \in V. \tag{6.10}
\]

With the QOI in (6.5) and in view of functional (6.6), we restrict our attention to bounded quantities as summarized in the following.

**Assumption 2 (Bounded QOI).** Given the linear QOI functional \( Q : V \to \mathbb{R} \) in (6.5), assume there exists constant \( c_Q \) such that

\[
Q(v) \leq c_Q \| v \|_V \quad \forall v \in V. \tag{6.11}
\]

In our theoretical development it is occasionally useful to consider more specific bounds, such as \( L^2 \) on the QOI; we modify this assumption when necessary.

We also assume general approximation properties on the discrete subspaces of \( V \) as follows.

**Assumption 3 (Approximation Property).** Given a shape regular mesh \( M \) of \( \Omega \), and a finite dimensional
subspace $S^h$ of $V$, we have for any $v \in V$,

$$\inf_{v^h \in S^h} \|v - v^h\|_V \leq C h^t \|u\|_{t+1},$$

(6.12)

where $C$ is independent of $h$ and $t > 0$ is in integer. This property holds for $V$ being a Sobolev space like $H^1$, $H(\text{div})$ or their product space, and for $S^h$ being the space of piecewise polynomial of degree $t$. We will assume this approximation property for the finite dimensional subspaces we use in our theory.

Note that assumptions 1 and 3 also imply the following convergence rate holds for the error $e^h$ for the least-squares problem $\langle Lu^*, Lv \rangle = \langle f, Lv \rangle$,

$$\|Le^h\| \leq Ch^r \|u^*\|_{r+1},$$

(6.13)

for $u^* \in H^{r+1}(\Omega)$.

We use these observations in several of the proofs in the following sections.

### 6.1.2 Theoretical Observations

An immediate consequence of ellipticity of $F$ is that it extends to $F_{Q^*}$. This is summarized in the following Lemma, and we note that multiple (additive) quantities-of-interest are also straightforward to accommodate:

**Lemma 7.** Suppose that $F$ is coercive and continuous over $V$, with constants $c_0$ and $c_1$, respectively. If $Q$ is a bounded linear functional over $V \subset L^2$ with constant $c_Q$, then $F_{Q^*}$ is coercive and continuous over $V$, with constants $c_0$ and $\hat{c}_1 = \sqrt{c_1^2 + \beta^2 c_Q^2}$, respectively.

**Proof.** Since

$$c_0^2 \|u\|_V^2 \leq \|Lu\|^2 \leq c_1^2 \|u\|_V^2 \quad \text{and} \quad |Q(u)|^2 \leq c_Q^2 \|u\|_V^2$$

(6.14)

hold for all $u \in V$,

$$c_0^2 \|u\|_V^2 \leq \|Lu\|^2 + \beta^2 |Q(u)|^2 \leq (c_1^2 + \beta^2 c_Q^2) \|u\|_V^2.$$  \hspace{1cm} (6.15)

immediately follows. \hfill \Box

### 6.1.3 Discrete Relationships

Additionally, if we assume that the value of the QOI, $\gamma$, is known, then the discrete problem adds intuition to our approach. Consider the weak problem associated with the goal-based functional $F_{Q^*}$: find $u \in V$ such
that

\[ \langle Lu - f, Lv \rangle + \beta^2 (Q(u) - \gamma) Q(v) = 0 \quad \forall v \in \mathcal{V}, \tag{6.16} \]

and the associated discrete problem with \( \mathcal{V}^h \subset \mathcal{V} \): find \( u^h \in \mathcal{V}^h \subset \mathcal{V} \) such that

\[ \langle Lu^h, Lv^h \rangle + \beta^2 Q(u^h) Q(v^h) = \langle f, Lv^h \rangle + \gamma \beta^2 Q(v^h) \quad \forall v^h \in \mathcal{V}^h. \tag{6.17} \]

Let \( \{ \phi_j^h \}_{j=1}^n \) be a basis for \( \mathcal{V}^h \) and write \( u^h = \sum_i u_i \phi_i^h \). From this we define matrix \( A \) with entries

\[ A_{ij} = \langle L\phi_i^h, L\phi_j^h \rangle, \]

and vectors \( \mathbf{q} \) and \( \mathbf{f} \) with entries \( q_i = Q(\phi_i^h) \) and \( f_i = \langle f, L\phi_i^h \rangle \), respectively. This results in the matrix problem

\[ (A + \beta^2 \mathbf{q}\mathbf{q}^T) \mathbf{u} = \mathbf{f} + \gamma \beta^2 \mathbf{q}. \tag{6.18} \]

We notice that the goal-based functional simply adds a rank-one update to our sparse matrix problem. Furthermore, in the case of \( k \) quantities-of-interest, the matrix problem results in \( k \) rank-one updates, which are handled effectively using the method of Sherman-Morrison-Woodbury [74], for example.

### 6.2 Error Bounds and a Modified Functional

In this section, we establish initial error bounds on the goal-based functional. Additionally, as a result of dissecting the method, we see that two modified forms of the goal-based least-squares functional emerge as more convenient forms. We first consider the weak form of \( F_{Q^*} \) in (6.16) and notice that the QOI, \( Q \) is not computable in the general case. Thus, we make use of the Reisz Representation Theorem to express \( Q \): there exists a unique \( p \in \mathcal{V} \) such that

\[ Q(v) = \langle Lv, Lp \rangle \quad \forall v \in \mathcal{V}. \tag{6.19} \]

With this, (6.16) becomes

\[
\langle Lu, Lv \rangle + \beta^2 Q(u) Q(v) = \langle f, Lv \rangle + \beta^2 Q(u^*) Q(v) \\
= \langle f, Lv \rangle + \beta^2 \langle Lu^*, Lp \rangle Q(v) \\
= \langle f, Lv \rangle + \beta^2 \langle f, Lp \rangle Q(v) \quad \forall v \in \mathcal{V}. \tag{6.20}
\]
Here, $p$ is unknown, so we introduce a subspace $W^h \subset V$, and approximate $p$ by solving the weak problem: find $p^h \in W^h$ such that
\[ \langle Lw^h, Lp^h \rangle = Q(w^h) \quad \forall w^h \in W^h. \] (6.23)

**First Modified Functional**

The above approximation leads us to our first modified functional, $F_Q$, in Figure 6.1. Substituting $p^h$ into (6.22) results in the weak problem: find $u \in V$ such that
\[ \langle Lu, Lv \rangle + \beta^2 Q(u)Q(v) = \langle f, Lv \rangle + \beta^2 \langle f, Lp^h \rangle Q(v) \quad \forall v \in V. \] (6.24)

Here, (6.24) is the weak form of the modified least-squares functional
\[ F_Q(u; f) = \|Lu - f\|^2 + \beta^2|Q(u) - (f, Lp^h)|^2. \] (6.25)

**Second Modified Functional**

The second modified functional, $F_{Q^h}$, in Figure 6.1, is obtained by considering a modified QOI $Q^h(u)$ as
\[ Q^h(u) = \langle Lu, Lp^h \rangle \forall u \in V. \] (6.26)

Thus, for exact solution $u^*$, we have $Q^h(u^*) = \langle f, Lp^h \rangle$, which becomes
\[ F_{Q^h}(u; f) = \|Lu - f\|^2 + \beta^2(Q^h(u) - Q^h(u^*))^2. \] (6.27)

Likewise, using the definition of $Q^h$, this yields
\[ F_{Q^h}(u; f) = \|Lu - f\|^2 + \beta^2(Lu - f, Lp^h)^2. \] (6.28)

We discuss the relationships and convergence properties of the two modified functionals shortly. However, first we examine the error due to the approximation of $p^h$ in $F_Q$, which is summarized by the following:

**Lemma 8** (Dual estimate: effect of $p^h \in W^h$). The error due to approximation of $p$ by $p^h$ is,
\[ \|e\|_V \leq C_{\tilde{h}^r} \|f\|_r \|p\|_{r+1}. \] (6.29)
Proof. Since $F_{Q^*}$ is coercive and continuous in $V$, $F_{O}$ is also coercive and continuous in $V$, and (6.24) has a unique solution, $\tilde{u} \in V$. From this, we consider the error in (6.24)

\[ e = \tilde{u} - u^*, \]  

(6.30)

by subtracting (6.24) from (6.22), to arrive at

\[ \langle L e, L v \rangle + \beta^2 Q(e)Q(v) = \beta^2 \langle f, L(p^h - p) \rangle Q(v) \quad \forall v \in V. \]  

(6.31)

Together with ellipticity of $F_{Q^*}$ in Lemma 7, this yields

\[ \|e\|_V^2 \leq \frac{\beta^2}{c_0} |\langle f, L(p^h - p) \rangle Q(e)\| \]  

(6.32)

\[ \leq \frac{\beta^2 c_1 c_0}{c_0} \|f\| \|p^h - p\| V \|e\|_V, \]  

(6.33)

or

\[ \|e\|_V \leq \frac{\beta^2 c_1 c_0}{c_0} \|f\| \|p^h - p\| V, \]  

(6.34)

which characterizes the error in terms of the approximation to $p$. By the assumption on approximation property of $W^h$ (Assumption 3), we have

\[ \|e\|_V \leq C\hat{h}^r \|f\|_r \|p\|_{r+1}. \]  

(6.35)

\[ \Box \]

Now we consider a discrete solution of (6.24) and specify the relationship between functionals $F_{Q}$ and $F_{Q^h}$. First we examine the discrete analogue of (6.24): find $u^h \in V^h \subset V$ such that

\[ \langle L u^h, L v^h \rangle + \beta^2 Q(u^h)Q(v^h) = \langle f, L v^h \rangle + \beta^2 \langle f, L(p^h - p) \rangle Q(v^h) \quad \forall v^h \in V^h. \]  

(6.36)

If we assume that $V^h \subset W^h$, then applying (6.23) for any $v^h \in V^h$ we have,

\[ Q^h(v^h) = \langle L v^h, L p^h \rangle = Q(v^h). \]  

(6.37)

Thus when $V^h \subset W^h$, (6.36) becomes the discrete weak form of the functional $F_{Q^h}$ (6.28). In this case we
have,
\[ F_Q(u^h; f) = F_{Q^h}(u^h; f) = \| Lu^h - f \|^2 + \beta^2 (Lu^h - f, Lp^h)^2. \] (6.38)

The next lemma provides an error bound for discrete problem.

**Lemma 9** (Primal estimate: effect of \( u^h \in \mathcal{V}^h \)). The finite element approximation \( u^h \in \mathcal{V}^h \) satisfies
\[ \| u^h - u^* \|_{\mathcal{V}} \leq Ch^s \| \tilde{u} \|_{s+1} + C\hat{h}^r \| f \|_{r+1}. \] (6.39)

**Proof.** By the approximation property of \( \mathcal{V}^h \) (Assumption 3), ellipticity of \( F_Q \) and Lemma 8 we arrive at
\[ \| u^h - u^* \|_{\mathcal{V}} \leq \| u^h - \tilde{u} \|_{\mathcal{V}} + \| \tilde{u} - u^* \|_{\mathcal{V}} \leq Ch^s \| \tilde{u} \|_{s+1} + C\hat{h}^r \| f \|_{r+1}. \] (6.40)

As a result, we see that the convergence of \( u^h \) suffers with the inclusion of the QOI in the functional, however as shown in (6.39), the degradation is directly associated with the ability to approximate \( p \) by \( p^h \), as expected, and by the weight and continuity coefficients of the QOI, namely \( \beta \) and \( c_Q \). Moreover, if \( \hat{h} = h \), \( \mathcal{V}^h \subset \mathcal{W}^h \), and \( r = s \), then (6.39) reduces to
\[ \| u^h - u^* \|_{\mathcal{V}} \leq Ch^r (\| \tilde{u} \|_{r+1} + \| p \|_{r+1}), \] (6.41)

which is used in the following sections.

We have seen the weak form for \( F_Q \) in (6.24) and its discretization in (6.36). Now we define the weak forms of the functionals \( F \) and \( F_{Q^h} \), which we use in our analysis below. Consider the standard least-squares functional \( F \) in (6.2), the (unique) weak solution is the function \( u^* \in \mathcal{V} \) such that
\[ \langle Lu^*, Lz \rangle = \langle f, Lz \rangle \quad \forall z \in \mathcal{V}. \] (6.42)

Similarly, the discrete weak solution \( u^h \in \mathcal{V}^h \subset \mathcal{V} \) satisfies
\[ \langle Lu^h, Lv^h \rangle = \langle f, Lv^h \rangle = \langle Lu^*, Lv^h \rangle, \quad \forall v^h \in \mathcal{V}^h. \] (6.43)
As a result, the discrete weak problem for the functional $F_{Q^h}$ is: find $u^h \in \mathcal{V}^h$, such that

$$
(\mathcal{L}u^h, \mathcal{L}v^h) + \beta^2 Q^h(u^h)Q^h(v^h) = (f, \mathcal{L}v^h) + \beta^2 Q^h(u^*)Q^h(v^h),
$$

(6.44)

Remark 8. The space $W^h$ can be viewed as an “adjoint” space in the sense that we are establishing a representation of the QOI, $\gamma$, through the Reisz Representation Theorem. In the discussion above, we seek this representation in a finer space than that of the primal approximation $u^h$. In contrast, it is natural to consider the case when $W^h \subseteq \mathcal{V}^h$. The following shows that there is no advantage to using $F_{Q^h}$ in place of the standard functional $F$, as expected.

Lemma 10. Suppose $W^h \subseteq \mathcal{V}^h$. If $\tilde{u}^h$ satisfies the discrete weak form associated with $F_{Q^h}$, (6.44), then it also satisfies the discrete weak form associated with $F$, (6.43).

Proof. From this we have the orthogonality condition on the solution in (6.43) we have,

$$
(\mathcal{L}(u^* - u^h), \mathcal{L}v^h) = 0, \quad \forall v^h \in \mathcal{V}^h.
$$

(6.45)

Using our assumption, $W^h \subseteq \mathcal{V}^h$ and (6.45),

$$
(\mathcal{L}u^h - f, \mathcal{L}p^h) = Q^h(u^h) - Q^h(u^*) = 0.
$$

Thus, $\tilde{u}^h$ satisfies (6.44), the weak form associated with $F_{Q^h}$. By uniqueness of the solution of the weak form, $u^h = \tilde{u}^h$ and the proof is complete.

6.3 Improved Error Bounds

In this section we improve upon the error analysis introduced in in Section 6.2. We consider two cases: a known value for the QOI, in which the goal-based functional $F_{Q^*}$ is used, and also an unknown value for the QOI, in which we establish estimates based on $F_Q$, which was introduced in Section 6.2. To establish improved error bounds, we assume $H^2$ regularity of the functionals, which we define below.

Assumption 4 ($H^2$ regularity). A least-squares problem arising from first order system $Lu^* = f$ is $H^2$ regular if there exists a constant $C$ such that for any $f \in H^r(\Omega)$, there exists a solution $u^* \in H^{r+2}(\Omega)$ such that,

$$
\|u^*\|_{r+2} \leq C\|f\|_r.
$$

(6.46)
6.3.1 Goal-based Least-Squares with a Known QOI

Consider the case when the QOI in (6.5) has a known value \( \gamma \) — e.g., in the case of conservation. This assumption eases the analysis (see the next section), and allows us to focus directly on function \( F_{Q^*} \) since an approximation to \( \gamma \) is not needed.

**Lemma 11.** If the functional \( F_{Q^*} \) is \( H^2 \) regular, then the error \( e^h \) satisfies,

\[
\|e^h\| \leq Ch^{p+1}\|u^*\|_{p+1}. \tag{6.47}
\]

**Proof.** Consider the solution \( w \in V \) to the auxiliary weak problem

\[
\langle Lw, Lv \rangle + \beta^2 Q(w)Q(v) = \langle e^h, v \rangle, \quad \forall v \in V. \tag{6.48}
\]

Setting \( v = e^h \) we arrive at a relationship with auxiliary variable \( w \) and the error:

\[
\langle Lw, Le^h \rangle + \beta^2 Q(w)Q(e^h) = \langle e^h, e^h \rangle. \tag{6.49}
\]

Then, from the orthogonality induced by the weak form (6.17), we have

\[
\langle e^h, e^h \rangle = \langle L(w - v^h), Le^h \rangle + \beta^2 Q(w - v^h)Q(e^h), \quad \forall v^h \in V^h, \tag{6.50}
\]

which becomes

\[
\|e^h\|^2 \leq \|L(w - v^h)\|\|Le^h\| + \beta^2 |Q(w - v^h)||Q(e^h)| \tag{6.51}
\]

\[
\leq (\|L(w - v^h)\|^2 + \beta^2 Q(w - v^h)^2)^{1/2} (\|Le^h\|^2 + \beta^2 Q(e^h)^2)^{1/2}. \tag{6.52}
\]

Finally, we use the approximation property of \( V^h \) and \( H^2 \) regularity of \( F_{Q^*} \) to arrive at

\[
\inf_{v^h \in V^h} (\|L(w - v^h)\|^2 + \beta^2 Q(w - v^h)^2)^{1/2} \leq Ch\|w\|_2 \leq Ch\|e^h\|, \tag{6.53}
\]

and hence,

\[
\|e^h\| \leq Ch (\|Le^h\|^2 + \beta^2 Q(e^h)^2)^{1/2} = ChF_{Q^*}^{1/2}(e^h; 0) \leq Ch^{p+1}\|u^*\|_{p+1}. \tag{6.54}
\]
6.3.2 Goal-based Least-Squares with an Unknown QOI

Now, we consider the more general case of an unknown value for the QOI in (6.5). Recall from Section 6.2 the modified functional in (6.27),

\[
\mathcal{F}_{Q^h}(u; f) = \| Lu - f \|^2 + \beta^2 (\hat{Q}^h(u) - \hat{Q}^h(u^*))^2. \tag{6.55}
\]

From this form of the functional, we arrived at a discrete weak problem in (6.44): find \( u^h \in \mathcal{V}^h \), such that

\[
\langle Lu^h, Lv^h \rangle + \beta^2 Q^h(u^h)Q^h(v^h) = \langle f, Lv^h \rangle + \beta^2 \hat{Q}^h(u^*) \hat{Q}^h(v^h), \tag{6.56}
\]

for every \( v^h \in \mathcal{V}^h \). With this we arrive at the following estimate of the error for \( \mathcal{F}_{Q^h} \). Recall that if we assume \( \mathcal{V}^h \subset \mathcal{W}^h \), then from (6.38) we see that the following result also holds for \( \mathcal{F}_Q \).

**Lemma 12.** If the functional \( \mathcal{F}_{Q^h} \) is \( H^2 \) regular, then the error \( e^h \) satisfies,

\[
\| e^h \| \leq Ch^{p+1} \| u^* \|_{p+1}. \tag{6.57}
\]

**Proof.** In comparing this modified weak form to that of the functional \( \mathcal{F}_Q \) in (6.17), we see that the error estimate follows that of the previous section. As a result, we have that

\[
\| e^h \| \leq Ch^{1/2} \langle u^* - u^h, 0 \rangle \leq Ch^{p+1} \| u^* \|_{p+1}. \tag{6.58}
\]

\hfill \Box

6.4 Bounding the Quantity-of-Interest

In the previous section, we showed that convergence of the global solution is maintained (in \( L^2 \)) when a goal is added to the functional. In this section, we derive bounds on the error in the QOI for both \( Q \) and \( Q^h \) in (6.26), which affirms that the approximations do indeed conform better to the goal than in standard least-squares.

We assume that \( Q \) is bounded in \( L^2 \) over \( \mathcal{V} \), that is,

**Assumption 5 (Bounded QOI in \( L^2 \)).** Given the linear QOI functional \( Q : \mathcal{V} \to \mathbb{R} \) in (6.5), assume there
exists constant \( c_L \) such that
\[
Q(v) \leq c_L\|v\| \quad \forall v \in \mathcal{V}.
\]

This results in a bounded \( Q^h \) in \( L^2 \) over \( \mathcal{V}^h \subset \mathcal{W}^h \). Similar to the representation in (6.19), under assumption 5, we define \( q \in L^2 \) to be the function such that
\[
Q(v) = \langle v, q \rangle \quad \forall v \in \mathcal{V}.
\]

The next two lemmas show enhanced convergence of \( Q(u - \tilde{u}^h) \) and \( Q^h(u - \tilde{u}^h) \) under the above assumption.

**Lemma 13.** Given Assumptions 4 and 5 and approximations \( u^h \) based on either \( F_Q^* \) or \( F_{Q^h} \), the error in the QOI \( Q \) satisfies,
\[
\frac{|Q(u^* - u^h)|}{\|Q\|_{L^2}} \leq \|u^* - u^h\| \leq Ch^{p+1}\|u^*\|_{p+1}.
\]

**Proof.** From (6.60), \( q = L^* Lp \), weakly, so that
\[
\|Q\| = \sup_{v \in L^2 \atop \|v\| = 1} |Q(v)| = \sup_{v \in L^2 \atop \|v\| = 1} |\langle v, q \rangle| = \|q\| = \|L^* Lp\|.
\]

With this, letting \( e^h = u^* - u^h \) results in
\[
|Q(e^h)| \leq \|Q\|\|e^h\|.
\]

Thus, for approximations \( u^h \) from \( F_Q^* \) (using (6.47)) or \( F_{Q^h} \) (using (6.57)), we have that
\[
\frac{|Q(u^* - u^h)|}{\|Q\|_{L^2}} \leq \|u^* - u^h\| \leq Ch^{p+1}\|u^*\|_{p+1}.
\]

\[\Box\]

Similarly, we also show that the error in the approximated QOI, \( Q^h \), also exhibits improved convergence with the modified least-squares functional \( F_{Q^h} \).

**Lemma 14.** Given the finite element solution \( u^h \) and Assumptions 4 and 5, the error in the modified QOI \( Q^h \) in \( F_{Q^h} \) satisfies,
\[
\frac{|Q^h(u^* - u^h)|}{\|Q^h\|_{L^2}} \leq Ch^{p+1}\|u^*\|_{p+1}.
\]

105
Proof. We have \( p^h \in W^h \) such that (in (6.23))

\[
\langle Lw^h, Lp^h \rangle = Q(u^h) \quad \forall w^h \in W^h.
\]  

(6.66)

Consequently, while \( \langle Lv, Lp^h \rangle \) may not be bounded over \( V \), we do have for \( v^h \in W^h \)

\[
Q^h(v^h) = \langle Lv^h, Lp^h \rangle = Q(v^h) = \langle v^h, q \rangle \leq \|v^h\|\|q\|. 
\]

(6.67)

That is, \( Q^h \) is bounded over the finer subspace. Thus,

\[
\|Q^h\| := \|q\|. 
\]

(6.68)

Now, let \( u^h \) be a weak solution using either \( F_{Q^*} \) or \( F_{Q^h} \). Then, we bound \( Q^h \) by observing that

\[
|Q^h(e^h)| = |\langle Le^h, Lp^h \rangle| 
\]

\[
\leq |\langle Le^h, Lp \rangle| + |\langle Le^h, L(p^h - p) \rangle| 
\]

\[
\leq \|q\|\|e^h\| + C\|h\|\|Le^h\||p\|_2 
\]

\[
\leq \|q\|\|e^h\| + C\|h\|\|Le^h\||q\|, 
\]

which follows from the interpolation assumptions on \( W^h \). Then, based on (6.57) we have that

\[
\frac{|Q^h(u^* - u^h)|}{\|Q^h\|_{L^2}} \leq \|u^* - u^h\| + C\|h\|\|L(u^* - u^h)\| \leq Ch^{p+1}\|u^*\|_{p+1}, 
\]

(6.73)

\[
\square
\]

In conclusion, we have shown that if \( Q \) is bounded in \( L^2 \), then enhanced convergence of \( Q(u - \tilde{u}^h) \) and \( Q^h(u - \tilde{u}^h) \) achieved, regardless of the choice of \( W^h \). However, as we detail in the following sections, while the asymptotic rate is not changed, bounds on the error in the goal for \( F_{Q^*} \) and \( F_Q \) yield a smaller constant.

### 6.4.1 Sharp Bounds for a Known QOI

Next, we establish sharp bounds on the error in the QOI. The approach that we take is similar to that of [105] in the case of a known QOI, which we assume here. That is, assume that the value of the QOI is a known quantity, namely \( Q(u^*) = \gamma \).
Lemma 15 (Dependence of error in $Q$ from $F_{Q^*}$ on $\beta^2$). Let $u^h$ be the solution to the discrete weak form, (6.17) associated with $F_{Q^*}$. Then for $e^h = u^* - u^h$ we have,

$$ Q(e^h) \leq \frac{Ch_{p+1}}{(1 + \beta^2(Lw^h, Lw^h))} \|u^*\|_{p+1} \|Q\|. \quad (6.74) $$

Proof. We have, for all $y^h \in V^h$, from orthogonality that

$$ Q(e^h) = \langle L e^h, Lp \rangle \quad (6.75) $$

$$ = \langle L e^h, Lp \rangle - \langle (L e^h, Ly^h) + \beta^2 Q(e^h) Q(y^h) \rangle \quad (6.76) $$

$$ = \langle L e^h, L(p - y^h) \rangle - \beta^2 Q(e^h) Q(y^h) \quad (6.77) $$

This leads to

$$ Q(e^h) = \frac{\langle L e^h, L(p - y^h) \rangle}{(1 + \beta^2 Q(y^h))}, \quad \forall y^h \in V^h. \quad (6.78) $$

From here, we choose a particular $y^h = w^h \in V^h$ so that

$$ \langle L(p - w^h), Lv^h \rangle = 0, \quad \forall v^h \in V^h. \quad (6.79) $$

By full $H^2$ regularity of the problem (Assumption 4) and from the approximation properties of $V^h$ (Assumption 3), we have

$$ \|L(p - w^h)\| = \inf_{v^h \in V^h} \|L(p - v^h)\| \leq ch\|p\|_2 \leq ch\|q\|, \quad (6.80) $$

where $q$ is derived in (6.60). Then, this leads to

$$ Q(e^h) = \frac{\langle L e^h, L(p - w^h) \rangle}{(1 + \beta^2(Lw^h, Lw^h))} \quad (6.81) $$

$$ \leq \frac{1}{(1 + \beta^2(Lw^h, Lw^h))} \|L e^h\| \|L(p - w^h)\| \quad (6.82) $$

$$ \leq \frac{Ch_{p+1}}{(1 + \beta^2(Lw^h, Lw^h))} \|u^*\|_{p+1} \|Q\|. \quad (6.83) $$

From Lemma 15 we observe that the constant becomes smaller as the weight $\beta^2$ on the QOI constraint in (6.6) increases. That is, as we strengthen the QOI term in the least-squares functional, the approximate solution satisfies the QOI with higher accuracy, as expected.
6.4.2 Sharp Bounds for an Unknown QOI

Next we consider the more general scenario of an unknown value for the QOI $Q(u^*)$. The next lemma yields a bound similar to that of (6.74), but for the modified QOI.

**Lemma 16.** Let $u^h$ be the weak solution to the minimization of the modified least-squares functional $F_{Q^h}$.

Then we have,

$$Q^h(e^h) \leq \frac{C_{\tilde{h}}p+1}{(1 + \beta^2 \langle Lw^h, Lw^h \rangle)} \|u^*\|_{p+1} \|Q\|,$$

(6.84)

where $\tilde{h} = \text{max}(h, \hat{h})$.

**Proof.** From (6.26) and from orthogonality, this $u^h$ satisfies for all $y^h \in V^h$,

$$Q^h(e^h) = \langle Lu^h, Lp^h \rangle = \langle Lc^h, L(p^h - y^h) \rangle - \beta^2 Q^h(e^h)Q^h(y^h),$$

(6.85)

where $e^h = u^* - u^h$. Thus, we have

$$Q^h(e^h) = \frac{\langle Le^h, L(p^h - y^h) \rangle}{(1 + \beta^2 Q^h(y^h))} \forall y^h \in V^h.$$  

(6.86)

From here, we choose a particular $y^h = w^h \in V^h$ so that

$$\langle Lw^h, Lz^h \rangle = Q^h(z^h) = \langle Lp^h, Lz^h \rangle \forall z^h \in V^h.$$  

(6.87)

As a result, by $H^2$ regularity of the problem (Assumption 4) and from the approximation properties of $V^h$ and $W^h$ (Assumption 3), we bound the error in the modified QOI by

$$Q^h(e^h) \leq \frac{\|Le^h\| \|L(p^h - w^h)\|}{1 + \beta^2 \langle Lw^h, Lw^h \rangle}$$

(6.88)

$$\leq \frac{\|Le^h\| \|L(p^h - p + p - w^h)\|}{1 + \beta^2 \langle Lw^h, Lw^h \rangle}$$

(6.89)

$$\leq \frac{\|Le^h\| \|L(p^h - p)\| + \|Le^h\| \|L(p - w^h)\|}{1 + \beta^2 \langle Lw^h, Lw^h \rangle}$$

(6.90)

$$\leq \frac{C_{\tilde{h}}p+1}{(1 + \beta^2 \langle Lw^h, Lw^h \rangle)} \|u^*\|_{p+1} \|Q\|.$$  

(6.91)
In order to bound \( Q(e^h) \) we observe that

\[
Q(e^h) = \langle L e^h, L p \rangle = \langle L e^h, L p^\hat{h} \rangle + \langle L e^h, L (p - p^\hat{h}) \rangle = Q^\hat{h}(e^h) + \langle L e^h, L (p - p^\hat{h}) \rangle.
\]

(6.92) (6.93) (6.94)

Therefore, if \( V^h \subseteq W^\hat{h} \), then

\[
Q(e^h) = Q^\hat{h}(e^h) + \langle L(e^h - y^\hat{h}), L(p - p^\hat{h}) \rangle,
\]

(6.95)

for all \( y^\hat{h} \in W^\hat{h} \). Moreover, the second term is bounded in terms of \( \hat{h} \), which yields

\[
|\langle L(e^h - y^\hat{h}), L(p - p^\hat{h}) \rangle| \leq C\hat{h}^{p+1}\|u^*\|_{p+1}\|Q\|.
\]

(6.96)

As a result we have,

\[
Q(e^h) \leq \frac{C\hat{h}^{p+1}}{(1 + \beta^2(L_{w^h}, L_{w^h}))}\|u^*\|_{p+1}\|Q\| + C\hat{h}^{p+1}\|u^*\|_{p+1}\|Q\|.
\]

(6.97)

In contrast, suppose that \( W^\hat{h} \subseteq V^h \), then from Lemma 10, we have \( Q^\hat{h}(u^* - \tilde{u}^h) = 0 \) and

\[
Q(e^h) = \langle L e^h, L(p - z^h) \rangle \quad \forall z^h \in V^h.
\]

(6.98)

In either of these two cases, the result is that

\[
\frac{|Q(e^h) - Q^\hat{h}(e^h)|}{\|Q\|} \leq C\hat{h}^{p+1}\|u^*\|_{p+1}.
\]

(6.99)

For the unknown QOI, \( \beta \) affects only one term of the error. In this case, the asymptotic rate is similar to the one derived in Lemma 12. Consequently, this shows that while the modified QOI in (6.26) is based on an approximation of \( p \) in the representation of the QOI, the difference is reduced as the mesh is refined, as expected.
6.5 Adaptive Refinement

In this section we detail an adaptive refinement algorithm that effectively targets the QOI, while maintaining convergence of the approximation to the global PDE. The algorithm is outlined in Algorithm 1 and uses the modified functional, \( F_Q \) in (6.25).

Consider the goal-based functional, \( F_{Q^*} \) in (6.6):

\[
F_{Q^*}(u; (f, \gamma)) = \|Lu - f\|^2 + \beta^2|Q(u) - \gamma|^2.
\]  

(6.100)

An effective refinement strategy relies on the balance of the terms in the functional. However, as motivation by the modified form of the functional in Section 6.2, the QOI term is not directly computable. Instead, we bound the error in this term by

\[
|Q(u_h^*) - \gamma| = |Q(u^*) - Q(u_h^*)| = |Q(u^* - u_h^*)|
\]  

(6.101)

\[
\leq C|\langle Lu^* - u^h, L\hat{p}\rangle| \quad \text{by (6.23)}
\]  

(6.102)

\[
= |\langle f - Lu_h^*, L\hat{p}\rangle|
\]  

(6.103)

\[
\leq \sum_{\tau \in \text{mesh}} |\langle f - Lu_h^*, L\hat{p}\rangle|_{\tau}.
\]  

(6.104)

That is, we use the approximation to the QOI in the modified form of the functional. As a result, we consider an error estimator, per element \( \tau \in \Omega^h \) of the form

\[
\eta_\tau(\alpha, \beta_h) = \alpha\|Lu_h^* - f\|_{\tau} + \beta_h|\langle f - Lu_h^*, L\hat{p}\rangle|_{\tau}
\]  

(6.105)

In our numerical tests, we restrict \( \alpha \in \{0, 1\} \) —ie., use either a pure goal estimator \( (\alpha = 0) \) or a combined estimator \( (\alpha = 1) \), with the weighting determined by \( \beta_h \). The presence of the QOI in both the functional (see (6.6)) and the error estimator (see (6.105)) is important and we underscore this in the numerical experiments below. Specifically, the fitness of the approximation \( u_h^* \) to the QOI is guided by the control of the QOI in both the weak problem and in the refinement pattern.

In addition to the modified functional \( F_Q \), we also consider the standard least-squares function \( F \) in (6.2) and with a goal-based error estimator such as \( \eta_\tau(0, \beta_h) \). We utilize an additional algebraic step based on
Algorithm 1: Goal-based least-squares adaptive refinement algorithm

**Input:** \( \Omega^h \): initial mesh
\( \beta \): QOI functional weight
\( \alpha, \beta_h \): least-squares and QOI error estimate weight
\( \theta_{ref} \): refinement threshold

**Return:** \( \Omega^h \): refined mesh

\[ u^h \leftarrow \text{initialize} \]

while \( \mathcal{F}_\mathcal{Q}(u^h; f) > \varepsilon_{tol} \)

1. construct \( \mathcal{V}^h \) and \( \mathcal{W}^h \) on \( \Omega^h \); \{h or p refinement\}

2. \( p^h \leftarrow \langle \mathcal{L}p^h, \mathcal{L}u^h \rangle \quad \forall u^h \in \mathcal{W}^h \); \{approximate QOI\}

3. \( u^h \leftarrow \min_{\mathcal{V}^h} \mathcal{F}_\mathcal{Q}(u; f, p^h, \beta) \quad \parallel \mathcal{L}u - f \parallel_2^2 + \beta^2 \langle \mathcal{L}u - f, \mathcal{L}p^h \rangle^2 \); \{goal-based least-squares\}

4. for each \( \tau \in \Omega^h \)

5. \( e^F_\tau \leftarrow \| \mathcal{L}u^h - f \|_\tau \); \{least-squares error\}

6. \( e^Q_\tau \leftarrow |\langle f - \mathcal{L}u^h, \mathcal{L}p^h \rangle_\tau | \); \{QOI error\}

7. \( E^F = \{ e^F_\tau | \tau \in \Omega^h \} \)

8. \( E^Q = \{ e^Q_\tau | \tau \in \Omega^h \} \)

9. \( \beta_h, \beta \leftarrow \text{compute weight}(E^F, E^Q); \{ \text{determine weights} \}

10. \( T_{ref} = \{ \tau \in \Omega^h | \alpha e^F_\tau + \beta_h e^Q_\tau \geq \theta_{ref} \max_\tau e_\tau \} \)

11. for each \( \tau \in T_{ref} \)

12. \( \Omega^h \leftarrow \text{refine } \tau \); \{refine in h\}

Galerkin orthogonality, common in literature for adjoint-based error estimators, namely

\[
|\langle f - \mathcal{L}u^h, \mathcal{L}p^h \rangle| = |\langle f - \mathcal{L}u^h, \mathcal{L}(p^h - p^h) \rangle| \quad \forall p^h \in \mathcal{V}^h.
\]

(6.106)

As a result, our QOI based estimator, \( \eta_\tau(0,1) \), for functional (6.2) is \( |\langle f - \mathcal{L}u^h, \mathcal{L}(p^h - p^h) \rangle|_\tau \). Furthermore, \( p^h \) can be chosen in multiple ways, e.g., nodal projection or \( L_2 \) projection of \( p^h \) from \( \mathcal{W}^h \) to \( \mathcal{V}^h \). In our experiments, we use a nodal projection, which is relatively straightforward.

Algorithm 1 requires an initial coarse representation of the mesh, \( \Omega^h \), weights for the functional and error estimator — i.e., \( \beta, \alpha, \) and \( \beta_h \) — and a threshold for selecting elements of the mesh to refine, labeled \( \theta_{ref} \) in Line 7. In Line 1 a space \( \mathcal{V}^h \) is used for mesh \( \Omega^h \) and a refinement space \( \mathcal{W}^h \), either in \( h \) or \( p \), for mesh \( \Omega^h \). With these spaces, an approximation to the QOI and to the modified weak problem are determined in Lines 2 and 3, respectively. Based on the approximations \( u^h \) and \( p^h \), the error in the solution and in the QOI are approximated per element in Lines 4 and 5. Then, based on this element-wise view of the error, the weights are recomputed in Line 6 (see Section 6.5.1) — e.g., one approach in \( \text{compute weight} \) is to let \( \beta \) and \( \beta_h \) remain the same. Finally, a certain portion of the error is selected for refinement in Line 7, using \( \theta_{ref} \) as a
threshold.

### 6.5.1 The Influence of $\beta$

As the refinement algorithm proceeds, the least-squares and QOI terms may not remain in balance. To simplify our motivation, consider the case of $h = h$ in spaces $W^h$ and $V^h$. Since the QOI integral differs in scope from that of the $L^2$-norm in the least-squares functional — i.e., it may be local and lower in dimension — we expect that $\beta$ to retain some dependence on $h$ in order to maintain balance of the terms in functional $F_Q$. If we consider the exponent $r$ in (6.29), we see that $\beta(h) < Ch^{-r/2}$ is necessary for the solution to converge. Moreover, if $\beta(h) = Ch^{-s/2}$, with $s < r$, then the convergence of $\|e\|_V$ is reduced to $O(h^{r-s})$, thus impacting the convergence inequalities of (6.39), (6.47), (6.57), and (6.61). This indicates that the rate of convergence of both the approximation and the QOI may suffer. In contrast, it is notable that a larger $\beta$ has a positive effect on the two sharper inequalities concerning the convergence in the QOI, namely (6.74) and (6.88). More precisely, as the constraint on the QOI becomes larger, the QOI becomes more accurate, as expected.

Since the exact dependence on $h$ is difficult to determine in practice, we propose a heuristic based on the observations above. A large value of $\beta$ emphasizes the goal factor $\langle f - Lu^h, Lp^h \rangle$ in $F_Q$. Moreover, if the approximation to $p^h$ is accurate, then a large $\beta$ yields a solution that resolves the goal more accurately than standard least-squares. However, it is also necessary to retain convergence of the global least-squares error, which is represented by the term $\|Lu^h - f\|$. As a result, we seek to balance these two terms in the functional by modifying $\beta$ and $\beta_h$ for each level of refinement in Algorithm 1. More precisely, we write the least-squares error and the approximate goal error in each element $\tau$ as

\begin{align}
    e^F_{\tau} &= \|Lu^h - f\|_{\tau}, \\
    e^Q_{\tau} &\approx |\langle f - Lu^h, Lp^h \rangle|_{\tau},
\end{align}

respectively. Then, in Line 6 of Algorithm 1, we compute a new weight as

\begin{equation}
    \beta_h = \frac{\max_{\tau} e^F_{\tau}}{\max_{\tau} e^Q_{\tau}},
\end{equation}

which is summarized in Line 1 of Algorithm 2. In order to reduce the initial unrealistic influence of the QOI, a small value of $\beta$ is chosen for the first iteration. Similarly, a finite bound on $\beta$ is also imposed to avoid pathological cases of overweighting and to guarantee convergence.
**Algorithm 2**: Adaptive $\beta$ algorithm

<table>
<thead>
<tr>
<th>Input: $\mathcal{E}^F, \mathcal{E}^Q$: list of least-squares and QOI errors per element</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e_{F_{\text{max}}} \leftarrow \max \mathcal{E}^F$; {least-squares error}</td>
</tr>
<tr>
<td>$e_{Q_{\text{max}}} \leftarrow \max \mathcal{E}^Q$; {QOI error}</td>
</tr>
<tr>
<td>$1, \beta_h \leftarrow \frac{e_{F_{\text{max}}}}{e_{Q_{\text{max}}}}$; {ratio of least-squares and QOI errors}</td>
</tr>
<tr>
<td>$\beta = \beta_h$</td>
</tr>
</tbody>
</table>

6.6 Numerical Experiments

In this section we detail several numerical experiments comparing the traditional least-squares finite element methods with the goal-oriented approach developed in this chapter. We explore the effect of the weighting parameters $\alpha$ and $\beta$ on the efficiency of the methods, which underscores the robustness and broad applicability of the goal-oriented approach. Several, representative PDEs and goals are considered. In particular, an illustrative Poisson problem, a convection-diffusion problem, and a Stokes flow with both known and unknown goals.

In the following experiments, $\mathcal{F}$ denotes a standard least-squares functional (6.2) and $\mathcal{F}_Q(\beta)$ represents the modified functional in (6.6) with a particular $\beta$. The error estimator, used for adaptivity in Algorithm 1, is represented by $\eta(\alpha, \beta)$, where $\alpha$ controls the least-squares term and $\beta$ controls the QOI as in (6.105). For each of the experiments, a single value of $\beta$ is used for both the error indicator and the modified functional $\mathcal{F}_Q$. An adaptive beta is indicated by $\hat{\beta}$ and is defined in Algorithm 2. For example,

- $\mathcal{F}$ $\eta(1,0)$ standard least-squares with standard estimator
- $\mathcal{F}$ $\eta(1,10^3)$ standard least-squares with a goal-based estimator with $\alpha = 1$ and $\beta = 10^3$
- $\mathcal{F}_Q(10^3)$ $\eta(0,10^3)$ goal-based least-squares with $\beta = 10^3$ in (6.6) and an estimator only based on the QOI, thus $\alpha = 0$ and $\beta_h = 10^3$ in Algorithm 1
- $\mathcal{F}_Q(\hat{\beta})$ $\eta(1,\hat{\beta})$ goal-based least-squares and estimator with an adaptive weight using Algorithm 2

The efficiency of these methods are compared by considering the computational work required for each method. Our metric for work or cost, defined to be the number of unknowns in the dual problem if we solve it, otherwise it is the number of unknowns in the primal problem.
6.6.1 Diffusion

We first consider a Poisson problem considered in [105]. The PDE is posed on the unit square with homogenous Dirichlet boundary conditions. More formally, $u$ satisfies

$$\begin{align*}
-\nabla^2 u &= f(x) \quad x \in \Omega \\
u(x) &= 0 \quad x \in \partial \Omega.
\end{align*}$$

(6.110)

The source term, $f$, is defined so that the solution to (6.110) is

$$u(x, y) = 5x^2(1 - x)^2(e^{10x^2} - 1)y^2(1 - y)^2(e^{10y^2} - 1),$$

(6.111)

where as the QOI interest for the the problem is

$$Q(u) = \int_{\Omega_Q} u,$$

(6.112)

where $\Omega_Q = [0.2, 0.4]^2$. Figure 6.2 shows the contours of $u$ and the region defining the goal, $\Omega_Q$. We see that the exact solution (6.111) yields relatively steep gradients at the northeast corner, while the region corresponding to the goal functional is well separated from this area. Rewriting (6.110) as a first-order system
\[
\mathcal{L} u = \begin{bmatrix}
-\nabla u + q \\
-\nabla \cdot q \\
\nabla \times \end{bmatrix} = \begin{bmatrix}
0 \\
f \\
0
\end{bmatrix} = f
\] (6.113)

defines the functional \( F \) and, with the addition of the goal \( Q \), defines the functional \( F_Q \) for this problem.

In our experiments, we use linear elements to discretize the primal problem — \( \mathcal{V}^h = (\mathbb{P}^2_1, \mathbb{P}_1) \) — ie., and quadratic elements for the dual — ie., \( \mathcal{W}^\hat{h} = (\mathbb{P}^2_2, \mathbb{P}_2) \).

Figure 6.3 shows the relative error in the QOI as a function of the cost. These plots compare the effect of different parameterizations and choices for the functional and error indicators. In the first image, Figure 6.3a, the convergence of the QOI is compared for methods using the standard least-squares functional \( F \) with several parameterizations of the error indicator (possibly including the goal). Also included in the plot, is the adaptively parameterized goal-oriented FOSLS algorithm \( "F_Q(\hat{\beta}); \eta(1, \hat{\beta})" \). The plot indicates that using the fully adaptive goal functional \( F_Q \) and error estimator based on it, labeled \( "F_Q(\hat{\beta}); \eta(1, \hat{\beta})" \), outperforms the other formulations which only solve the standard least-squares functional. In particular, the adaptive goal-based method outperforms the standard least-squares error estimator, labeled \( "F; \eta(1, 0)" \) and the standard least-squares functional with a pure goal-oriented error indicator labeled \( "F; \eta(0, 1)" \). Additionally, the goal-based least squares estimator, \( "F; \eta(1, 10^3)" \) results in improved approximations in comparison to the standard last-squares estimator. This suggests the importance of the \( \beta \) parameter in the performance of the error indicator.

**Figure 6.3:** [Diffusion] Convergence of the relative error in the QOI, \( Q(e^h)/Q(u^*) \), versus total cost of the spaces — ie. total number of nodes in \( \mathcal{V}^h \) and \( \mathcal{W}^\hat{h} \).
In Figure 6.3b we compare the performance of using the $\mathcal{F}_Q$ functional with different indicators applied to drive adaptive refinement. We compare using the least-squares estimator “$\mathcal{F}_Q(10^3); \eta(1, 0)$”, the goal-based estimator “$\mathcal{F}_Q(10^3); \eta(0, 10^3)$”, an estimator based on both the least-squares and QOI “$\mathcal{F}_Q(10^3); \eta(1, 10^3)$”, but with a fixed weight of $\beta = 10^3$. The value for $\beta$ is chosen empirically to maximize the convergence of these formulations to the true value of the goal functional. The last formulation considered in this figure is based on our adaptive weight “$\mathcal{F}_Q(\hat{\beta}); \eta(1, \hat{\beta})$”. The plots show that the use of the optimal weight ($\beta = 10^3$) leads to several formulations that have comparable performance. This suggests the additional benefit of the use of the modified functional on convergence to the goal. It is also important to note that the adaptively weighted formulation, $\mathcal{F}_Q(\hat{\beta}); \eta(1, \hat{\beta})$, performs in some cases better than the formulations with a fixed weight. This indicates the quality of the weighting constructed for this problem.

The previous discussion demonstrated that the modified formulations and error indicators can effectively reduce the error in the QOI. To see that they also satisfy the PDE consider Figure 6.4 which shows the convergence of the formulations measured in the least-squares error. We have included the traditional least-squares functional “$\mathcal{F}; \eta(1, 0)$” and, for reference, a shifted version “$\mathcal{F}; \eta(1, 0)(\text{adj})$” to account for the increased size of the dual problem in this figure. We see that the convergence rate of the least-squares functional is roughly the same for all the methods. As a result, the goal-based least squares method targets the goal, while retaining convergence to the solution of the PDE.

Figures 6.5a and 6.5b show the mesh after 12 levels of refinement using the least-squares estimator “$\mathcal{F}; \eta(1, 0)$” and estimators based on the modified functional, “$\mathcal{F}_Q(\hat{\beta}); \eta(1, \hat{\beta})$”. The figures for the meshes show that standard least-squares refines heavily in the top right corner, where the gradients in the solution are large. In contrast, the adaptively weighted goal-based formulation refines in the top right less aggressively while refining more heavily in the region $\Omega_Q$ corresponding to the QOI. The more targeted refinement pattern of the goal-based least-squares formulation attempts to balance the resolution required to satisfy the PDE with the resolution required to compute the goal functional.

### 6.6.2 Convection-Diffusion

In this section we consider convection-diffusion in the 2D domain $\Omega = [0, 10] \times [0, 2]$. This problem has been previously considered in [59]. More precisely we seek a solution to

\[
-\nabla \cdot (a(x)\nabla u(x)) + b \cdot \nabla u = 1 \quad x \in \Omega
\]

\[
u(x) = 0 \quad x \in \partial \Omega,
\]

(6.114)
Figure 6.4: [Diffusion] Convergence of the least-squares solution, $\|Lu - f\|$, with and without effects of the QOI-term.

Figure 6.5: [Diffusion] Mesh after 12 levels of adaptive refinement.
where $a(x) > 0$ is a smoothly varying function, defined as

$$a(x) = 0.05 + \tanh(10(x - 5)^2 + 10(y - 1)^2),$$

and the vector field $b = [-1, 0]^T$. Rewriting (6.114) as a first-order system results in

$$\mathbf{q} - \sqrt{a} \nabla u = 0,$$
$$-\nabla \cdot (\sqrt{a} \mathbf{q}) + b \cdot \nabla u = 1,$$
$$\nabla \times (\mathbf{q}/\sqrt{a}) = 0.$$

The goal functional of interest is to evaluate the integral of $u$ over the small patch $\Omega_Q = [1, 1.25] \times [0.25, 0.5]$. Formally, in terms of the variables of the first-order system, the QOI is stated as

$$\mathcal{Q}(\mathbf{q}, u) = \int_{\Omega_Q} u.$$

Both the domain and the goal region are depicted in Figure 6.6. To solve the primal problem piecewise linear basis functions are used, while the solution to the dual is approximated by quadratics — ie., we use $\mathcal{V}^h = (P^2_1, P_1)$ and $\mathcal{W}^h = (P^2_2, P_2)$.

![Figure 6.6: Convection-Diffusion Computation domain with goal region denoted by $\Omega_Q$.](image)

We look at the effect of the functional and the error estimator on the error in the QOI in Figure 6.7. In Figure 6.7a we compare different error estimators for the standard least-squares functional $\mathcal{F}$, with an estimator for the functional $\mathcal{F}_Q$. The figure indicates that using the functional $\mathcal{F}_Q$ and an error estimator based on it, labeled “$\mathcal{F}_Q(\hat{\beta}); \eta(1, \hat{\beta})$“, outperforms all other error estimators which only solve the standard least-squares functional.

Figure 6.7b compares different estimators based on the modified functional $\mathcal{F}_Q$. We compare using the
least-squares estimator “$F_{Q}(10^{3.5}); \eta(1,0)$”, the goal-based estimator “$F_{Q}(10^{3.5}); \eta(0,10^{3.5})$”, an estimator based on both the least-squares and QOI “$F_{Q}(10^{3.5}); \eta(1,10^{3.5})$”, but with fixed $\beta = 10^{3.5}$, and finally the estimator based on our adaptive weight “$F_{\hat{\beta}}(\hat{\beta}); \eta(1,\hat{\beta})$”. This figure indicates that again, choosing an estimator based on our adaptive weight $\hat{\beta}$ outperforms other weightings.

Finally, we examine the convergence of the least-squares error, which is a global measure, for different estimators in Figure 6.8. We see that estimators that focus on the QOI, “$F_{Q}(10^{3.5}); \eta(0,10^{3.5})$” and “$F; \eta(0,1)$” perform relatively poorly as expected. The least-squares estimator is adequate for the global error, yet estimators based on the modified functional $F_{Q}$ and having both the QOI and least-squares part, “$F_{Q}(10^{3.5}); \eta(1,10^{3.5})$” outperform those based only on the QOI.

Figures 6.7a, 6.7b, and 6.8 show that we need the least-squares part and the QOI part in both the functional and the error estimator. Moreover, such estimators are successful at reducing the QOI error, while maintaining the convergence rate for the global least-squares error.

Next we examine the effect of the value of $\beta$ on the performance of the experiment in Figure 6.9. From these results, $\beta = 10^{3.5}$ is the optimal for a fixed value. However, as illustrated in the figure, our adaptively chosen value, $\hat{\beta}$ for the functional and the estimator performs similarly. Figure 6.10 shows the mesh after 12 levels of refinement using the least-squares estimator “$F; \eta(1,0)$” and estimators based on the modified functional, “$F_{\hat{\beta}}(\hat{\beta}); \eta(1,\hat{\beta})$”. The figures for the meshes indicate that the QOI-based estimator refines more in the “downwind” direction as compared to the standard least-squares estimator.
6.6.3 Stokes Equations

In this section we consider a Stokes flow with a known goal. The Stokes equations on a domain $\Omega$ are,

$$ -\Delta u + \nabla p = f \quad \text{in } \Omega $$

$$ \nabla \cdot u = 0 \quad \text{in } \Omega. $$
Here \( u, p \) and \( f \) denote the velocity, pressure, and body force per unit mass, respectively. We specify the boundary condition after describing the domain \( \Omega \) in more detail below.

We use the velocity-vorticity-pressure \((u, \omega, p)\) formulation,

\[
\begin{align*}
\nabla \times \omega + \nabla p &= f \quad \text{in } \Omega \\
\nabla \times u - \omega &= 0 \quad \text{in } \Omega \\
\n\nabla \cdot u &= 0 \quad \text{in } \Omega.
\end{align*}
\]

This leads to the \( H^{-1} \) norm based least-squares functional,

\[
\mathcal{F}(u, \omega, p; f) = \| \nabla \times \omega + \nabla p - f \|_{-1}^2 + \| \nabla \times u - \omega \|_{0}^2 + \| \nabla \cdot u \|_{0}^2. \tag{6.115}
\]

In the discrete setting, the norm \( \| \phi^h \|_{-1} \) is implemented as \( h \| \phi^h \|_0 \) for some discrete function \( \phi \). The domain is a backward step, shown in Figure (6.11). The boundary conditions we use (with \( u = (u_0, u_1) \)) are as
follows:

\[ u_0 = \begin{cases} 
  y(1-y) & \text{at } x = 0 \\
  0 & \text{for } x \neq 0 \text{ and } x \neq 10 
\end{cases} \]  

(6.116)

\[ u_1 = 0 \text{ on } \partial \Omega \]  

(6.117)

\[ p = 0 \text{ at } x = 10. \]  

(6.118)

Figure 6.11: [Stokes] Backward step domain with a QOI line integral at \( x = 8 \).

**Known QOI**

Since we know that

\[
\text{inflow} \left( \int_{x=0} u \cdot n \right) = 1/6,
\]

(6.119)

the QOI shown by the dashed line in Figure 6.11, is known. Specifically,

\[
Q(u, \omega, p) = \int_{x=8} u \cdot n,
\]

(6.120)

where \( n = (1, 0) \). For the primal problem, we use \( \mathcal{V}^h = (P_2^2, P_1, P_1) \), while for the dual problem, we use \( \mathcal{W}^h = (P_3^2, P_1, P_1) \).

In Figure 6.12 we compare the performance of an estimator based on the least-squares functional “\( F; \eta(1, 0) \)” with QOI-based functionals. We see that the relative error in the QOI is handled more effectively using the adaptive, modified functional in comparison to standard least-squares.

**Multiple QOIs**

We next consider multiple regions where a QOI is desired, in order to highlight the robustness of our approach. We compute the flow at locations \( x = 1, x = 2, x = 3, x = 6 \) and \( x = 8 \), as shown in Figure 6.13.
Cost
|Q(eh)|
|Q(\hat{u})|
F; η(1, 0)
F_Q(20); η(1, 20)
F_Q(\hat{β}); η(1, \hat{β})

Figure 6.12: [Stokes] Convergence of the relative error in the QOI, Q(e_h)/Q(u*), versus total cost of the spaces — ie. total number of nodes in \(V_h\) and \(W_h\).

We compare the conservation of flow at these five locations for three different functionals and estimators:

"\(F; η(1, 0)\)": the standard least-squares estimator,

"\(F_Q(20); η(1, 20)\)": the estimator based on \(F_Q\) with a single QOI at \(x = 8\) and with \(Q = Q_8\), and

"\(F_Q(20); η(1, 20), Q_8\)": the estimator based on \(F_Q\) which has five QOI with \(Q = \sum_i Q_i\).

This last estimator, which takes into account all five QOI locations, outperforms the other two, less-targeted estimators. It is important to note that this estimator minimizes all five QOIs, resulting the smallest, total relative error across all five regions.

Figure 6.13: [Stokes] Regions of multiple QOI.
Figure 6.14: [Stokes] Convergence of the relative error in multiple QOI versus total cost of the spaces.
6.6.4 The Poisson-Boltzmann Equation

As a final application we consider the Poisson-Boltzmann equation, which is described in detail in Chapter 2. We use the LSFEM functional for the PBE formed in Chapter 3, which uses $\alpha = \frac{1}{\sqrt{\epsilon}}$ weighting. That is, we use the following LSFEM functional,

$$G(q, u; u_c) = \|\frac{1}{\sqrt{\epsilon}} q - \sqrt{\epsilon} \nabla u - \frac{1}{\sqrt{\epsilon}} ((\epsilon - \epsilon_m)) \nabla u_c\|_{0,\Omega}^2 + \|\nabla \cdot q + \kappa^2 u + \kappa^2 u_c\|_{0,\Omega}^2. \quad (6.121)$$

The QOI is to compute the solvation free energy. We recall from Chapter 3 that the solvation free energy for a set of discrete charges $Q_j$ at positions $x_j$ and regularized potential $u$ is defined by

$$\Delta G_{sol} = \frac{1}{2} k_B T e_c \sum_{j=1}^{n \epsilon} Q_j u(x_j). \quad (6.122)$$

Solvation free energy as defined in (6.122) is a sum of evaluation of the solution at points $x_j$. However, point evaluation of a function is not in the space $H^{-1}(\Omega)$, i.e. the dual space of $H^1(\Omega)$, since point evaluation is linear, but not bounded. To see this, consider the function $f(x) = 1/\|x\|^\theta$ on $B(1, \mathbb{R}^3)$, for $0 < \theta < 0.5$, where $B(1, \mathbb{R}^3)$ is the ball of radius 1 in $\mathbb{R}^3$; indeed this function is in $H^1(B(1, \mathbb{R}^3))$ [7], and hence $\|f\|_1$ is a finite value. However, the function has an infinite value at $x = 0$, and hence there cannot exist a (finite) constant $C$ such that $f(0) < C \|f\|_1$. That is, point evaluation is not bounded for this function. Hence, we cannot use the definition of solvation free energy in (6.122) as our QOI. To define a bounded linear functional, we may approximate point evaluation by a local average, as is done for the PBE in [54], or instead use mollification [105].

To define a QOI which is a linear functional on $H^1(\Omega)$, we mollify our regularized potential. This involves convolution of the regularized potential with a “mollifier” such that evaluation of the mollified function gives an accurate approximation to the regularized potential. Now we proceed to describe this process. First, we define the “mollifier” $\rho(x)$, which is a $C^\infty$ function, as,

$$\rho(x) = \begin{cases} c_\rho \exp \left( \frac{1}{\|x\|^{2-\theta}} \right) & \|x\| < 1, \\ 0 & \text{otherwise.} \end{cases} \quad (6.123)$$

where $c_\rho$ is chosen so that $\int_{\mathbb{R}^3} \rho(x) dx = 1$. Note that $\rho(x)$ obtains its maximum value at $x = 0$. Now we
define the mollification of $u$, which we call the mollified potential, $u_\delta$,

$$u_\delta(x) = \delta^{-3} \int_\Omega \rho \left( \frac{x - \xi}{\delta} \right) u(\xi) d\xi,$$

(6.124)

for $0 < \delta < \text{dist}(x, \partial \Omega)$, and approximate $u(x_j)$ by $u_\delta(x_j)$. Note that the mollifier $\rho \left( \frac{x - \xi}{\delta} \right)$ is centered at $\xi$ and has support in a ball of radius $\delta$. Now we show that $u_\delta(x_j)$ is a linear functional on $H^1(\Omega)$ and that it converges to $u(x_j)$ as $\delta \to 0$.

**Lemma 17.** The mollified potential, $u_\delta(x_j)$, as defined in (6.124), belongs to the dual space of $H^1(\Omega)$, that is $u_\delta \in H^{-1}(\Omega)$.

**Proof.** First we verify boundedness. Let $h(\xi) = \rho \left( \frac{x - \xi}{\delta} \right)$. Then using the Cauchy-Bunyakovsky-Schwarz inequality we obtain,

$$u_\delta(x_j) = \delta^{-3} \int_\Omega \rho \left( \frac{x_j - \xi}{\delta} \right) u(\xi) d\xi,$$

(6.125)

$$= \delta^{-3} \int_\Omega h(\xi) u(\xi) d\xi,$$

(6.126)

$$\leq c \|h\|_{0, \Omega} \|u\|_{0, \Omega},$$

(6.127)

$$\leq c \|h\|_{0, \Omega} \|u\|_{1, \Omega}$$

(6.128)

where $c$ is a generic constant. Furthermore using the fact that $\Omega$ is a bounded set in $\mathbb{R}^3$ we obtain,

$$\|h\|^2_{0, \Omega} = \int_\Omega \rho^2 \left( \frac{x_j - \xi}{\delta} \right) d\xi,$$

(6.129)

$$\leq \rho^2(0) \text{Vol}(\Omega),$$

(6.130)

$$\leq M,$$

(6.131)

for some $M$ which is sufficiently large. From equations (6.128) and (6.131), we see that $u_\epsilon(x_j)$ is a bounded functional on $H^1(\Omega)$. 
Next we verify linearity. Let $u_1$ and $u_2$ be two functions in $H^1(\Omega)$. Then,

\[
(u_1 + u_2)_\delta = \delta^{-3} \int_\Omega \rho \left( \frac{x_j - \xi}{\delta} \right) (u_1 + u_2)(\xi) d\xi
\]

\[
= \delta^{-3} \int_\Omega \rho \left( \frac{x_j - \xi}{\delta} \right) (u_1 + u_2)(\xi) d\xi
\]

\[
= \delta^{-3} \int_\Omega \rho \left( \frac{x_j - \xi}{\delta} \right) (u_1(\xi) + u_2(\xi)) d\xi
\]

\[
= u_1_\delta + u_2_\delta.
\]

Thus, we have that $u_\delta \in H^{-1}(\Omega)$. Now we examine the convergence of $u_\delta$.

**Lemma 18.** As $\delta$ approaches 0, the mollified potential $u_\delta$ approaches the potential $u$ point-wise. That is, as $\delta \to 0$, $u_\delta(x_j) \to u(x_j)$.

**Proof.** See [85] for details.

**Definition of QOI for the PBE**

Based on the above discussion, we define the QOI as the approximation of the solvation free energy by the mollified regularized potential $u_\delta$ as,

\[
Q(q, u) = \frac{1}{2} \frac{k_B T}{e_c} \sum_{j=1}^{n_s} Q_j u_\delta(x_j).
\]

**Lemma 19.** The mollified solvation free energy, as defined by $Q(q, u)$ in (6.136) belongs to $V^*$, where $V^*$ is the dual space to $V = H(\text{div}) \times H^1$.

**Proof.** Each $u_\delta(x_j)$ is a bounded linear functional on $V$ by Lemma 17, and hence belongs to the vector space $V^*$. $Q$ is a linear combination of functionals, and hence by the vector space property, also belongs to $V^*$.

**Results**

We compare the performance of an estimator based on the least-squares functional “$F; \eta(1,0)$” with the one based on the modified functional for the Born ion and methanol and use $\delta = 0.1$ in our experiments. For the primal problem, we use $V^h = (P_1^3, P_1)$, while for the dual problem, we use $W^h = (P_2^3, P_2)$. We write the expression for the true value of the solvation free energy for the Born ion using the mollified potential $u_\delta$, $x_j = 0$ and $Q_j = 1$ as,

\[
Q(q, u) = \frac{1}{2} \frac{k_B T}{e_c} u_\delta(0).
\]
Hence the value of \( u(0) \) is,

\[
\begin{align*}
u(0) &= \delta^{-3} \int_{\|\xi\| < \delta} \rho\left(\frac{\xi}{\delta}\right) u(\xi) d\xi, \\
&= \delta^{-3} \int_{\|\xi\| < \delta} \rho\left(\frac{\xi}{\delta}\right) u(\xi) d\xi. 
\end{align*}
\]

(6.138)

(6.139)

Recall from Chapter 2 that \( u(\xi) \) has a constant value, denoted \( c_u \) for \( \|\xi\| < a \), where \( a \) is the radius of the Born ion. Since \( a = 2 \), and \( \delta = 0.1 < a \), we transform the above integral into spherical coordinates with \( r = \|x\| \) as,

\[
u(0) = 4\pi \delta^{-3} \int_{r < \delta} \rho\left(\frac{r}{\delta}\right) c_u r^2 dr.
\]

(6.140)

We compute this integral numerically, to obtain a value of the mollified solvation free energy (6.137) as -82.134874762320194. For methanol, we approximate the solvation free energy by solving the PBE using the Galerkin method on a mesh with a large number of vertices. This value, while not being the actual value of solvation free energy, serves as a reference to compare the two estimators. The results for the two estimators are shown in Figure 6.15. The figure indicates “\( F_{mod}(\hat{\beta}; \eta(1, \hat{\beta})) \)” performs significantly better than the simple least-squares based method. We obtain very good approximation to the solvation free energy, even for coarse meshes. We conclude that QOI-based least-squares methods provide significant computational advantages in computing the solvation free energy.

**Figure 6.15:** Solvation free energy versus total cost of the spaces — ie. total number of nodes in \( V^h \) and \( W^h \).
Chapter 7

Summary and Future Directions

This thesis develops finite element methods for partial differential equations arising from implicit solvent models. We developed finite element methods based on least-squares methodology for the linear Poisson-Boltzmann equation, and analyzed methods for modern implicit solvent models which take ion-size into account. These models, which are the modified Poisson-Boltzmann equation for systems in equilibrium and the modified Poisson-Nernst-Planck equations for systems in dynamics, requires new analysis and more complex numerical methods for efficient simulation. In the remainder of this chapter, we detail the primary contributions of this thesis, and outline directions for future research.

7.1 Contributions

**Least-Squares finite element method for the PBE** [32]: We developed a first order systems least squares method for the linearized PBE, wherein we designed a novel flux variable that takes into account the discontinuous nature of the gradient of the regularized potential at the solvent/solute interface. From this, we show that the traditional flux variables, which are based only on the gradient (or a multiple of the gradient), are not adequate for the regularized PBE. Furthermore, we showed ellipticity of the least-squares functional, which results in a well-posed problem. Finally we designed an efficient numerical method based on adaptive refinement using the least-squares functional, and demonstrated its effectiveness on a variety of examples.

**Analysis of Least-Squares finite element method for the PBE** [46]: We analyzed $H(\text{div})$ conforming Raviart-Thomas elements for the PBE to illustrate the robustness of our approach. Moreover, we examined different weighting of the least-squares functional based on the discontinuous coefficient $\epsilon$, and showed how a weighting based on $1/\sqrt{\epsilon}$ leads to better condition systems with good convergence properties. Consequently the least-squares method, in comparison to the scalar Galerkin method, showed least-squares computes the flux variables significantly more accurately.

**Analysis of weak solution to the nonlinear modified Poisson-Boltzmann equation** [45]: The
linearized PBE is a good approximation to the PBE when the electrostatic potential is low. On the other hand, at large potentials, this approximation breaks down. However, at large potentials, the PBE itself gives unphysical results, as it assumes point ions. In contrast, the MPBE improves on the PBE by assuming finite sized ions. Hence we defined a weak solution to the modified Poisson-Boltzmann equation, and showed it exists and is unique. To show this, we defined an energy functional, the critical point of which is the MPBE. Then using functional analysis, we showed this energy functional has a unique minimizer.

**Nonlinear finite element method for the modified Poisson-Boltzmann equation [45]:** We developed a convergent nonlinear finite element method for the MPBE and showed that a unique discrete solution to the MPBE exists. Further, the discrete solution converges to the continuum solution in the limit of the mesh refinement. We also showed optimal convergence for our finite element method. One consequence is that the finite element method for the MPBE places less restrictions on the mesh than the one for the PBE. This is due the modified nonlinearity in the MPBE, which makes the analysis simpler, whereas PBE requires special mesh conditions to satisfy $L^\infty$ bounds. We showed that the MPBE leads to better conditioned linear systems than the PBE. Moreover, MPBE also provides an efficient way to solve the PBE by providing a good initial guess for Newton’s method.

**Finite Element method for the modified Poisson-Nernst-Planck equations [47]:** The modified Poisson-Nernst-Planck equations model the dynamics of ions in a fluid while taking into account finite ion size. We formed a nonlinear finite element method for computing the steady state concentration for the MPNPE. Our scheme uses a staggered scheme for solving the modified Nernst-Planck and Poisson equations respectively. The solution process is more complex than that of the PNPE because of the presence of nonlinearity in the modified Nernst-Planck equations to model ion size effects. Our numerical scheme conserves the number of ions in the system to mimic the behavior of molecular dynamics simulations. We demonstrate how the presence of large potential leads to instabilities in the Galerkin method. In response, we develop a stabilized finite element method based on the streamline-upwind Petrov/Galerkin (SUPG) scheme. This method adds artificial diffusion along the streamlines to prevent spurious negative concentrations in the simulation. We also develop a fast SUPG method to accelerate the solution process for the MPNPE and demonstrate our method to compute flow of ions through a nanopore and around a charged cylinder.

**Design and analysis of QOI based least-squares finite element method [48]:** Traditional finite element methods are efficient at reducing global errors, but not at resolving a quantity of interest. Adjoint based methods resolve this deficiency somewhat, but this comes at the cost of loss of accuracy in global norms. We augment the least-squares functional with a term based on the QOI to design a numerical scheme
which accurately computes the QOI, while still retaining convergence in global norms. We develop theory for our method, demonstrating its convergence properties and also develop theory for the Aubin-Nitsche trick for the modified functional to obtain faster convergence rates for the case of $L^2$ QOIs. Next we extend our analysis to show enhanced convergence for the QOI, and develop sharp bounds for the error in the QOI. Finally we develop an error estimator based on the new functional, and numerically show the advantages of incorporating QOI in the functional and error estimator. We apply for QOI-based least-squares method to compute solvation free energy for the PBE. Our results indicate that QOI-based least-squares methods perform significantly better than simple least-squares based method.

### 7.2 Future Directions

The research in this thesis has opened up a number of interesting directions to pursue further research. Here we outline these challenging directions.

**Least-squares for stability and ionic current computation in MPNPE** This thesis explored SUPG finite element for the MPNPE in the presence of sharp gradients. Another viable alternative if the first-order systems least squares method. As has been mentioned earlier, LSFEM exposes flux variables, the accurate computation of which is required for ionic current calculation. One complication in applying least-squares is that because of the periodic boundary conditions, symmetry of the system is violated. An implementation of periodic boundary conditions which take into account the symmetry is essential for good performance of the least-squares. Further, QOI-based least squares is promising for current calculation, by making the ionic current as the QOI. However, this means solving either the steady state problem, with the integral of the concentration as a further QOI, or solving a time dependent problem. In the former, conditioning of the linear systems is a challenge, while in the latter, QOI-based methods for least-squares which take time integration into account have to be developed. This application has quite a few subtleties and complications, however, it also has immense promise.

**QOI-based least squares for nonlinear problems and nonlinear QOI** The theory developed for least-squares based QOI in this thesis is motivated by accurate calculation of solvation free energy for the linearized PBE, which is a linear functional for a linear problem. Many interesting problems are not linear, for example the MPBE, the MPNPE and the Navier-Stokes equations. Moreover, the QOI may not be a linear functional. Such nonlinearities present challenges in designing QOI based least-squares methods for these problems. For instance, we apply Riesz-Representation theorem to define the QOI. If the QOI is nonlinear, then we cannot apply the Riesz-Representation theorem directly. Such challenges make this an exciting area.
of further research.

**FEM for more sophisticated modified models** The MPNPE is one model for incorporating ion size effects for dynamical systems. The MPNPE, as studied in this thesis, assume uniform ion size for both ion species throughout the domain. More sophisticated models have been studied [19] which relax some of these assumptions. Designing efficient finite element methods for these models in an interesting area of research.

**Parallelization of assembly for MPNPE** The SUPG method for the MPNPE has many terms in its weak form, and hence leads to a computationally intensive assembly of the linear system and the load vector. However, this process is also amenable to parallelization. Assembling the system in parallel may lead to significant performance advantages. The challenges in addressing parallelization for the MPNPE is of great practical importance.
References


Author Biography

Jehanzeb Hameed Chaudhry was born on October 7, 1980, in Rawalpindi, Pakistan. He attended primary school at the Beaconhouse Public School in Rawalpindi. He went to high school at the Fauji Foundation College, Rawalpindi. He went to college (senior high school in the US) at the F.G. Sir Syed College, The Mall, Rawalpindi. He got his Bachelor of Science in Computer Science and Engineering degree from the Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi, Pakistan in 2003. He enrolled in the PhD program in Computer Science in the University of Illinois in the fall of 2005.