# NODAL METHODS: ANALYSIS, PERFORMANCE AND FAST ITERATIVE SOLVERS 

By<br>J. David Moulton<br>B. Eng. (Physics) McMaster University, 1988<br>M. Eng. (Physics) McMaster University, 1990<br>A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF<br>THE REQUIREMENTS FOR THE DEGREE OF<br>Doctor of Philosophy<br>in<br>THE FACULTY OF GRADUATE STUDIES<br>INSTITUTE OF APPLIED MATHEMATICS<br>DEPARTMENT OF MATHEMATICS<br>We accept this thesis as conforming to the required standard

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

## Mathematics

The University of British Columbia
2075 Wesbrook Place
Vancouver, Canada
V6T 1Z1

Date:



#### Abstract

Nodal Methods have long been one of the most popular discretization techniques employed within the reactor physics community, while remaining conspicuously absent from the mainstream numerical analysis literature. A fundamental reason for this anomaly is that the physical arguments which were used to develop and enhance these methods seemed at odds with more rigorous discretization techniques. To facilitate communication between these distinct communities, a detailed chronological study of the lowest-order nodal methods for linear second order elliptic problems is presented. The presentation highlights the underlying motivation of these methods and formalizes many of their renowned properties. In addition, various equivalence relations within this family of discretizations are demonstrated, and equivalences with specific non-conforming and mixed-hybrid finite element methods (FEMs) are established. Rigorous error bounds and stability properties follow immediately from these latter equivalence relations, corroborating the results of a more rudimentary truncation error analysis.

An inherent difficulty of reactor simulation is that the coefficients in the mathematical model exhibit severe variations on two significantly different length scales. As in many other applications this is treated by defining an appropriate homogenization procedure which yields a simplified model with piecewise constant coefficients on a coarse scale suitable for efficient computation. Significant enhancements in accuracy are possible if the processes of homogenization and discretization are unified. We review the popular techniques that are based on this premise and rely on certain properties of the nodal discretization. In addition, we address the factors that contribute to their success in reactor modelling and deter their generalization outside of the reactor physics community.


As an alternative to these highly specialized methods, we introduce a new multi-level homogenization technique which is readily applicable in a general setting, and is shown to have many important attributes.

Widespread acceptance of nodal methods has also been hindered by their use of nonstandard unknowns, as this results in stencils that appear awkward and incompatible with sophisticated iterative solution techniques. Specifically, equivalence with certain mixed-hybrid FEMs reveals that the nodal discretizations result in an indefinite system, which in two dimensions contains both cell-based and edge-based unknowns. Yet, inherent in this structure is a natural partitioning of the system which may be exploited to define a hierarchy of reduced systems (i.e. Schur complements) that are symmetric positive definite. Unfortunately, the reduced systems that involve unknowns of only one type, and hence seem most compatible with sophisticated iterative methods, suffer a loss of sparsity. However, the structure inherent in this hierarchy may be utilized in the construction of sparse approximate Schur complements for these systems. It is demonstrated that any one of these approximate operators, which are of either the standard 5 or 9-point family, may be utilized as an excellent preconditioner for conjugate gradient iterations. The efficiency of this approach is fully realized when the preconditioner is approximately inverted using only a single V- or W-cycle of a robust Black Box multigrid solver.

## Table of Contents

Abstract ..... ii
List of Tables ..... ix
List of Figures ..... xi
Acknowledgment ..... xiii
1 Introduction ..... 1
1.1 Background Sketch ..... 2
1.2 Outline ..... 5
2 Background Theory ..... 9
2.1 Neutron Transport Theory ..... 9
2.1.1 The Neutron Transport Equation ..... 9
2.1.2 Simplifying Assumptions ..... 13
2.1.3 Transport Boundary Conditions ..... 14
2.2 Diffusion Theory ..... 15
2.2.1 The $P_{1}$ Approximation ..... 15
2.2.2. Criticality: The Reactor Eigenvalue ..... 18
2.2.3 Approximate Boundary Conditions ..... 19
3 Nodal Discretizations ..... 21
3.1 Classical Approaches ..... 23
3.1.1 Finite Volume and Nodal Balance ..... 23
3.1.2 Standard Cell Centered Finite Differences ..... 25
3.2 Simulators ..... 27
3.2.1 Fine Mesh Correction - Cross Coupling ..... 29
3.2.2 FLARE ..... 29
3.3 Transverse Integration ..... 32
3.4 Polynomial Nodal Methods - NEM ..... 33
3.4.1 Components of the NEM Discretization ..... 34
A Polynomial Basis ..... 34
Transverse Leakage Approximations ..... 37
Weighted Residual Procedure ..... 39
3.4.2 The Lowest-Order NEM ..... 40
Interface Current Formulation ..... 41
Flux Formulation ..... 43
Edge Based Flux Formulation ..... 44
3.5 Analytic Nodal Methods - NIM ..... 45
3.5.1 The Nodal Integration Method ..... 46
3.5.2 The constant-constant NIM for Conservative Diffusion ..... 48
3.5.3 The constant-constant NIM for Nonconservative Diffusion ..... 50
The Flux Formulation ..... 53
Edge Based Flux Formulation ..... 54
3.6 Truncation Error Analysis ..... 55
3.6.1 Piecewise Constant Coefficients ..... 57
3.6.2 Piecewise Smooth Coefficients ..... 58
3.7 Numerical Experiments ..... 60
3.7.1 Example 1: A Simple Test ..... 60
3.7.2 Example 2: The Iron-Water Benchmark ..... 63
4 Homogenization ..... 69
4.1 Equivalence Techniques in Reactor Modelling ..... 70
4.1.1 Coarse-Scale Properties ..... 72
4.1.2 A One Dimensional Study ..... 73
4.1.3 Multidimensional Equivalence Techniques ..... 76
Flux Weighting ..... 77
Generalized Equivalence Theory ..... 78
4.2 Asymptotic Homogenization Techniques ..... 82
4.2.1 Multidimensional Periodic Analysis ..... 83
4.2.2 Multidimensional Nearly Periodic Analysis ..... 85
4.3 Multigrid Homogenization ..... 86
4.3.1 Obtaining the Diffusion Tensor ..... 87
4.3.2 A Numerical Homogenization Suite ..... 90
Shape Dependence ..... 91
Dependence on the Diffusivity ..... 91
A Nonsymmetric Inhomogeneity ..... 94
5 Finite Element Methods ..... 95
5.1 Primal Conforming FEM ..... 96
5.1.1 Variational Formulation ..... 96
5.1.2 Conforming Polynomial Basis ..... 98
5.2 Primal Nonconforming FEM ..... 102
5.2.1 The Edge and Cell Basis ..... 102
One-Dimensional Basis ..... 102
Two-Dimensional Basis ..... 105
5.2.2 Polynomial Basis Functions ..... 107
A One-Dimensional Study ..... 107
Equivalence in Two Dimensions ..... 114
5.2.3 Analytic Basis Functions ..... 116
A One-Dimensional Study ..... 116
Equivalences in Two Dimensions ..... 121
5.3 Mixed Hybrid Finite Element Methods ..... 123
5.3.1 Variational Formulation ..... 124
5.3.2 Polynomial Basis Functions ..... 126
5.3.3 Analytic Basis Functions ..... 128
6 Reduced Systems and Approximate Schur Complements ..... 129
6.1 Lumping The Edge-Based Schur Complement ..... 130
6.1.1 A Condition Bound ..... 132
6.1.2 Bounding the elements of $A_{u u}^{-1}$ ..... 136
6.2 A Two-Step Lumped Approximation ..... 137
6.3 Lumping the Cell-Based Schur Complement ..... 140
6.3.1 A Condition Bound ..... 141
6.4 Approximate Multi-level Inversion of the Preconditioner ..... 144
6.5 Numerical Experiments ..... 145
6.5.1 A Toy Problem ..... 146
6.5.2 Ground Water Flow ..... 148
6.5.3 A Diffusive Checkerboard ..... 151
7 Conclusions ..... 154
7.1 Research Summary ..... 154
7.2 Future Work ..... 159
Bibliography ..... 163

## List of Tables

3.1 Convergence results for Example 1: the transverse averaged scalar flux on a $10 \times 10$ uniform mesh. ..... 61
3.2 Convergence results for Example 1 in the $L_{\infty}$-norm. ..... 62
3.3 Convergence results for Example 1: the cell averages. ..... 62
3.4 The diffusivity $D$, and the absorption cross-section $\Sigma_{a}$, for Example 2 [25]. ..... 64
3.5 Errors for Example 2 on a $10 \times 10$ computational grid. ..... 67
4.1 A comparison of the shape dependence of the diffusivity relative to the square (i.e. the percent relative difference - \% RD) is presented for the results of Bourgat [12] and the black box homogenization. ..... 92
5.1 Canonical Definitions in 1D ..... 103
5.2 Canonical Definitions in 2D ..... 105
6.1 Iteration counts for the preconditioners, $\widetilde{\mathcal{S}_{\mu ; u}}$ and $\widetilde{\mathcal{S}_{\phi}}$ for the toy problem with constant mesh spacing on the domain $[0,1] \times[0,1]$. ..... 147
6.2 Iteration counts for the lumped preconditioner, $\widetilde{\mathcal{S}_{\mu ; u}}$ with constant mesh spacing on the domain $[0, a] \times[0, b] .\{$ Note: $\alpha=2 / a, \beta=2 / b\}$ ..... 147
6.3 Iteration counts for the two-step lumped preconditioner, $\widetilde{\mathcal{S}_{\mu ; u, v}}$ with con- stant mesh spacing on the domain $[0, a] \times[0, b]$. $\{$ Note: $\alpha=2 / a, \beta=2 / b\}$ ..... 147
6.4 Iteration counts for the cell-based preconditioner, $\widetilde{\mathcal{S}_{\phi}}$ with constant mesh spacing on the domain $[0, a] \times[0, b] .\{$ Note: $\alpha=2 / a, \beta=2 / b\}$ ..... 148
6.5 Iteration Counts with single $\mathrm{V}(1,1,1)$ cycles ..... 150
6.6 Iteration Timings, [s] for the flow problem with single $\mathrm{V}(1,1,1)$ cycles ..... 150
6.7 Iteration counts for $\gamma_{x}=\gamma_{y}=2 / 10$. ..... 153

## List of Figures

3.1 An $L \times M$ tensor product mesh imposed on the domain $\Omega$ ..... 22
3.2 The location of the unknowns ..... 25
3.3 Stencil weights of the flux formulation with $D=1$ and $\Sigma_{a}=0$. ..... 44
3.4 The edge-based flux formulation with $D=1$ and $\Sigma_{a}=0$. ..... 46
3.5 The transverse averaged scalar fluxes, $\left\{\phi_{i+\frac{1}{2}, j}, \phi_{i, j+\frac{1}{2}}\right\}$ obtained with the constant-constant NIM for this simple test case are shown in (a) while the corresponding node averages, $\phi_{i, j}$ are plotted in (b). ..... 62
3.6 Schematic of the iron-water benchmark problem [25] ..... 64
3.7 The constant-constant NIM solution of Example 2 computed on a $10 \times 10$ grid: (a) transverse averaged scalar fluxes (b) cell averages. ..... 67
3.8 Solution cross-sections for Example 2 are compared to a reference solution that was obtained using the finite volume discretization on a $640 \times 640$ uniform mesh: $y_{j}=10.5$ (a) full domain (b) zoom on layer. ..... 68
4.1 A compass based definition of an arbitrary 9-point stencil ..... 87
4.2 Three inhomogeneities with the same area of $1 / 4$, but different shape. ..... 92
4.3 A square inhomogeneity with diffusivity, $\lambda$ and an area of $1 / 9^{\text {th }}$. ..... 93
4.4 A comparison of homogenized diffusivities. ..... 93
4.5 The homogenization of an $L$ shaped inhomogeneity leads to a dense tensor. ..... 94
5.1 The quadratic canonical cell and edge basis functions. ..... 109
6.1 Upper bound of the condition number, $\kappa(6.6)$ and of the decay rate $\lambda_{\infty}(r) .137$
6.2 A schematic of Mosé et al.'s [56] first example. ..... 149
6.3 Solution $\phi(x, y)$ and streamlines of Mosé et al.'s [56] first example (above) ..... 150
6.4 A diffusive checkerboard configuration based on Dendy's [23] example. ..... 152
6.5 The solution, $\phi(x, y)$ of the diffusive checkerboard computed on a $24 \times 24$ uniform mesh ..... 153
6.6 2D Exponential Grid with the parameters $\gamma_{x}=\gamma_{y}=2 / 10$. ..... 153
7.1 Stream lines for the convection dominated flow of Smith and Hutton [67]. ..... 161
7.2 The convection dominated flow of [67] with $P e=[D]^{-1}=500$ computed using a constant-constant NIM: (a) cell averages (d) inflow and outflow profiles. ..... 161

## Acknowledgment

I would like to express my sincere appreciation to my advisor, Dr. Uri Ascher, for promoting and tolerating my diverse research interests. In addition I would like to thank him for his guidance, insight, patience and gregarious hospitality, all of which were vital factors in the completion of this thesis.

For many enthusiastic, and often impromptu, discussions I convey my sincere gratitude to Dr. Jim Varah. I would also like to thank the other members of my advisory committee, Drs. Brian Wetton and Michael Ward, both of whom have made positive contributions to this research.

I have had the good fortune of visiting Los Alamos several times over the past few years. I would like to express my gratitude to Dr. James Hyman for his key role in making my early visits possible and to acknowledge his incredibly contagious enthusiasm that continues to inspire my interest in homogenization well beyond its roots as a summer project. I also extend my thanks to Dr. Jim Morel for both my initial contact with nodal methods and the many encouraging and instructive discussions that we had during my latter visits. I am particularly grateful to Dr. Joel Dendy for both his insightful observations regarding my research and his generous hospitality.

To my parents and brother I convey my deepest thanks for their unconditional support and encouragement.

The long and arduous journey that has been this thesis, could not have been completed without the continued support and encouragement of Tracey Ward. I am especially indebted to her patience, sacrifice and understanding, which have been essential in the endless final stage of this work.

## Chapter 1

## Introduction

Nodal methods have long been one of the most popular discretization techniques employed within the reactor physics community to solve multi-group diffusion problems \{e.g. $[72,34,53,68]\}$. A survey of these methods can be found in [52]. Their success is a result of their exceptional accuracy which may be attributed to three distinct aspects of the nodal ideology. Specifically, akin to finite volume methods, nodal methods are physically motivated cell-based discretizations that, by construction, rigorously enforce cell balance. They utilize an intriguing choice of unknowns (consisting, in two dimensions for example, of cell and edge moments) which makes the nodal discretization naturally compatible with the various homogenization techniques that are crucial to reactor modelling. Moreover, the neutron current is obtained accurately and automatically from the discretization, thereby avoiding problematic finite difference approximations.

Despite these distinguishing characteristics and the appearance of diffusive phenomena in many disciplines, nodal methods have received only limited attention outside the reactor physics community. This anomaly is due, in part, to the limited analysis of these discretizations (see [39, 36, 25] ). In fact, the physical arguments which were used to develop and enhance these discretizations make it difficult to establish the strength of their mathematical foundation. In addition, their successful application in reactor modelling relies on homogenized diffusion and interaction coefficients that are typically approximated through idealized local auxiliary computations. Hence the corresponding discretization error is difficult to analyze and the application of nodal methods to general
diffusion problems seems nontrivial. Finally, the unusual choice of unknowns makes the resulting linear system difficult to solve as it appears incompatible with sophisticated iterative solution techniques.

The objective of this research is to address the key issues, namely analysis and solution techniques, that have thus far restricted the application of these methods. In particular, we conduct a detailed study of the lowest-order nodal discretizations. This investigation suggests that the homogenization techniques employed in reactor modelling are an integral part of the nodal methods dominance in this arena. The ultimate merit of nodal methods relative to other techniques is less clear in a general setting. However, for the related and important class of problems in which piecewise constant coefficients are assumed, nodal discretizations are ideal. To enhance their competitiveness in these applications we investigate a new general-purpose homogenization technique and we develop a new family of fast iterative solvers for the corresponding sparse linear systems.

### 1.1 Background Sketch

The neutron distribution in a reactor core is one of the fundamental quantities required in the simulation and design of nuclear reactors. An exact description of this distribution (i.e. the angular neutron density) is provided by the transport equation. Unfortunately, this integrodifferential equation is extremely difficult to solve and although its numerical treatment has improved significantly over the past 30 years, it is often replaced with approximate models [52]. The most popular of these is the $P_{N}$ approximation, which corresponds to an $N^{\text {th }}$ order angular expansion of the angular neutron density, and is asymptotic to the transport equation [51]. This expansion leads to a system of $N$ first order PDEs, and in particular the $P_{1}$ approximation leads to the first order form of the diffusion equation (i.e. a balance equation and Fick's law). In general, the energy
dependence of the angular neutron density is discretized into groups such that a system of diffusion equations is obtained. These multi-group diffusion equations have been the backbone of reactor modelling for some 35 years and continue to play a significant role today.

The diffusion model represents a significant simplification of the transport equation; however, the fine-scale variations in the coefficients in combination with the large physical size of the reactor core pose a complex computational problem. In particular, a naive fine-mesh three-dimensional treatment of the multi-group diffusion equations was inconceivable in the 60 's [40] and remains intractable today [52]. Yet the coarse-scale properties of the solution (e.g. cell averages), that appear on a practical computational mesh, are significantly influenced by the fine-scale structure of the coefficients. Problems containing two or more significantly different length scales arise frequently in physical applications and their treatment relies on various homogenization procedures. Specifically, the fine-scale structure of the coefficients is averaged in some sense to define an homogenized diffusion problem in which the coefficients are piecewise smooth functions, typically constants, on the computational mesh. Unfortunately, the earliest treatments of the multigroup diffusion equations, which applied standard discretization techniques to this homogenized problem, proved inadequate. Efforts to bridge this gap between computational resources and sophisticated modelling have produced a powerful class of discretization techniques referred to as nodal methods.

Common to all nodal methods is their rigorous enforcement of cell balance, a characteristic motivated by the underlying transport equation. In fact, the first nodal methods were based on a clever hybrid approach that combined transport physics and diffusion theory. However, the performance of these methods, now collectively known as nodal
simulators, relied on the tuning of several reactor dependent free parameters [34]. Moreover, these methods are not consistent with the multigroup diffusion equations, and as a result have largely been superseded by the modern, consistent family of nodal discretizations [52]. These particular methods are based on a transverse integration procedure that reduces the multidimensional diffusion equation to a coupled set of ODEs involving transverse averaged quantities. This not only facilitates the use of unknowns compatible with the popular homogenization techniques (cell and edge moments in two dimensions) but has contributed to their enhancement [65]. The treatment of these transverse integrated ODEs classifies the members of this family as either polynomial or analytic nodal methods.

However, the influence of the enhanced homogenization techniques that were developed specifically for reactor modelling is difficult to predict in a general setting. In addition, the utilization of transverse integration in conjunction with cell- and edge-based unknowns makes the discretizations themselves difficult to analyze. This uncertainty surrounding the mathematical foundation of nodal methods has significantly hindered their use in related applications outside of the reactor physics community. Recently, in the absence of homogenization, the equivalence of various nodal methods with certain nonconforming and mixed-hybrid finite element methods (FEMs) was established $[39,37,36]$. Hence in this restricted setting a solid theoretical foundation exists.

The widespread popularity of nodal discretizations has also been hindered by the awkward linear systems which result. This is complicated by the fact that each equivalent discretization (mixed, mixed-hybrid and non conformal FEMs) manifests this difficulty in a different way. For example, the mixed system is indefinite and although the elimination of certain unknowns leads to a positive definite system the relative sparsity is compromised. Various techniques have been employed to circumvent this limitation
\{e.g. [61, 44, 2]\}. Conversely, the nonconformal FEMs are sparse but typically contain only edge unknowns making the definition of a suitable grid hierarchy and the intergrid transfer operators difficult. Nevertheless, multilevel preconditioners have been developed for this system and include the multilevel substructuring preconditioners discussed in [29, 50]. However, the presence of strongly anisotropic diffusion may in general degrade the performance of these preconditioners. In this and most other cases, solvers for the nonconformal discretization utilize the mixed-hybrid method only as a tool to establish equivalence with the corresponding mixed method. We contend that the mixed-hybrid method is, in fact, central to the development of efficient solvers, as it is central to the equivalence of these discretizations.

### 1.2 Outline

To establish the setting in which nodal methods originate, a brief introduction to transport theory and the $P_{1}$ approximation is given in Chapter 2. The various terms of the transport equation are presented along with a discussion of the simplifying assumptions that are typically employed in reactor modelling. The one-group diffusion equation is derived in Section 2.2.1 and the diffusion approximation of the typical transport boundary conditions is addressed in Section 2.2.3.

An historical overview of the most popular nodal methods is presented in Chapter 3. We restrict our presentation to the one-group diffusion equation and begin by highlighting the key components that characterize nodal methods. Particularly, the classical cellbased finite volume discretization [Section 3.1] demonstrates the rigorous treatment of cell balance and serves to highlight the issues surrounding it. The motivation of the nodal simulators is discussed in Section 3.2 along with a detailed review of FLARE [21]. In addition to a rigorous treatment of cell balance, the modern, consistent, nodal
methods rely on transverse integration. We review this technique and resulting transverse integrated ODEs in Section 3.3. The most popular variants of the polynomial and analytic modern, consistent, nodal methods are presented chronologically. The nodal expansion method (NEM) is discussed in Section 3.4 and the nodal integration method (NIM) appears in Section 3.5. For the restricted case of conservative diffusion, a new equivalence is established between the lowest-order NEM and NIM discretizations. In Section 3.6 we begin to address the need for a strong mathematical foundation with the presentation of a new, rudimentary, analysis of these methods. The results of this analysis are verified in the numerical examples of Section 3.7. Specifically, a contrived problem is used to highlight the second order convergence of the flux, while the superior accuracy of these methods, particularly NIM, is demonstrated in a more practical setting.

The complex issues of homogenization are discussed in Chapter 4. Specifically, in Section 4.1 we review the equivalence-based techniques that are typically employed in reactor modelling and are motivated by the desire to preserve certain coarse-scale properties of the fine-scale solution. Unfortunately, closed-form expressions of the homogenized diffusion coefficient which satisfy exact equivalence relations are known in only a few cases, and thus, various approximations based on the numerical solution of local auxiliary problems have been developed. In Section 4.1.3 one of the most popular of these approximate techniques is discussed in detail and its crucial role is revealed. A more rigorous mathematical approach to homogenization utilizes a two-scale asymptotic analysis and is well suited to periodic and nearly periodic problems. We review the important results of this analysis in Section 4.2, and in particular, we identify the local auxiliary problems which define the homogenized diffusion tensor and deter its widespread application. In Section 4.3 we present a new multi-level technique [24] that does not rely on the solution of local auxiliary problems and demonstrates great promise for the periodic
case. Moreover, this new approach is readily extended to nonperiodic problems.
In Chapter 5 the equivalence of the lowest-order nodal discretizations with certain nonconforming and mixed-hybrid FEMs is studied. To provide an appropriate reference for comparison the standard bilinear conforming FEM is reviewed in Section 5.1 and several relevant properties are noted. In Section 5.2 we explore the equivalence of these nodal methods with certain nonconforming FEMs presented in [36, 37]. Specifically, we adopt a one-dimensional analysis that highlights the influence of the transverse integration procedure and provides new insight into the characteristics and relationships of the nonconformal methods. Of particular importance is the more rigorous equivalence with NIM that this new analysis establishes. Finally, the equivalence of mixed-hybrid FEMs with the lowest-order NEM, that was established by Hennart and del Valle [39], is presented in Section 5.3. In addition, we discuss the difficulty of extending this equivalence to NIM, but most importantly we demonstrate that the NIM system may be written in the indefinite form associated with the mixed-hybrid FEM. This latter property is crucial to the success of our cell-based solver.

In Chapter 6 we construct a new family of preconditioners that are based on the sparse approximation of various reduced systems. Specifically, we begin by eliminating the currents from the mixed-hybrid FEM system to obtain a reduced system for the scalar unknowns. Section 6.1 introduces the most common treatment of this reduced system, which eliminates the cell-based unknowns in favour of an exclusively edge-based system. This edge-based Schur complement is still sparse and, moreover, is naturally partitioned by edge type (i.e. horizontal or vertical). However, a further reduction through the elimination of either edge type degrades the sparsity, and hence, a lumping procedure is introduced to approximate the Schur complement of this reduced system. The resulting approximation has a 9 -point sparsity structure and is readily inverted with a standard
multigrid solver such as Black Box Multigrid [23]. We prove optimal convergence rates for this edge-based preconditioner. Unfortunately, these rates depend on the aspect ratio of the grid, and hence, in Section 6.2 we consider a two-step preconditioner which attempts to remove the directional bias of the lumping procedure. Ironically, this preconditioner is asymmetric, hence a more attractive possibility arises by returning to the reduced system for the scalar unknowns and proceeding to eliminate the edge variables [Section 6.3]. Again a loss of sparsity results and hence we apply a simple lumping to the original indefinite system that results in an approximate Schur complement which is the standard 5-point cell-centred discretization. Thus, efficient multigrid inversion of this preconditioner is also possible. However, performing an exact multigrid solve for each iteration of preconditioned conjugate gradient is undesirable, and hence in Section 6.4 we discuss the effectiveness of a single multigrid cycle as an approximate inversion of the preconditioner. In Section 6.5 we present a numerical study of our solvers that includes some non-trivial problems. In all cases, our preconditioners perform extremely well: only a small number of iterations, independent of the grid size, is required. The robustness of the two-step edge-based and the cell-based preconditioners is evident in the tests with high aspect ratio cells.

Finally, conclusions and considerations for future research are offered in Chapter 7.

## Chapter 2

## Background Theory

The foundation of reactor modelling is neutron transport theory. The corresponding equation, the neutron transport equation, is considered to be an essentially exact expression governing the neutron distribution in a reactor core. However, the generality and inherent complexity of this integrodifferential equation makes it exceedingly difficult to solve. Typically, simplifying approximations are made to reduce its complexity, and in cases such as diffusion theory, lead to a system of PDEs. Hence we present a concise introduction to transport theory which serves to highlight the natural hierarchy of the approximations that are inherent in any reactor model, and hence, any corresponding computation. An important implication of this hierarchy is that the terminology employed in virtually all approximate reactor models has its ancestry in transport theory. Thus, our overview of transport theory permits the introduction of this terminology in its natural environment.

### 2.1 Neutron Transport Theory

### 2.1.1 The Neutron Transport Equation

The fundamental quantity of interest in reactor modelling is the angular neutron density $n(\mathbf{r}, E, \widehat{\boldsymbol{\Omega}}, t)$. It is actually a density in the statistical sense that $n(\mathbf{r}, D, \widehat{\boldsymbol{\Omega}}, t) d \mathbf{r} d E d \widehat{\boldsymbol{\Omega}}$ is the expected number of neutrons located in $d \mathbf{r}$ about $\mathbf{r}$, within $d E$ of the kinetic energy $E$, travelling within the solid angle $d \widehat{\boldsymbol{\Omega}}$ of $\widehat{\boldsymbol{\Omega}}$, at a time in $d t$ about $t$. Although this
definition may seem excessively detailed, its generality is fully utilized in the derivation of its governing equation, the transport equation.

The mechanisms by which neutrons interact with the various components of the reactor naturally play a significant role in the development of this equation. Hence, these mechanisms are classified, with the two primary classes being scattering and absorption. In addition, these interactions must be quantified, thus a macroscopic cross-section $\Sigma_{\alpha}(\mathbf{r}, E)^{1}$ is defined for each class that gives the probability per unit length that a neutron will interact with a nucleus. Specifically, the probability per unit length of a neutron being absorbed at $(\mathbf{r}, E)$ is $\Sigma_{a}(\mathbf{r}, E)$. An analogous definition holds for the macroscopic scattering cross-section $\Sigma_{s}(\mathbf{r}, E)$. Interaction of any type combines these classes, and is described by the total macroscopic cross-section,

$$
\Sigma_{t}(\mathbf{r}, E)=\Sigma_{s}(\mathbf{r}, E)+\Sigma_{a}(\mathbf{r}, E)
$$

These general classes may be further subdivided. For example, absorption includes fission and radiative capture (among other mechanisms). Each of these mechanisms has a corresponding macroscopic cross-section, $\Sigma_{f}(\mathbf{r}, E)$ and $\Sigma_{\gamma}(\mathbf{r}, E)$ respectively.

Although a measure of the probability of neutron scatter has been defined, the angular dependence has not been considered. Moreover we have suppressed the dependence on the neutrons resultant energy $E^{\prime}$. These dependencies are described by the double differential scattering cross-section,

$$
\Sigma_{s}\left(\mathbf{r}, E \rightarrow E^{\prime}, \widehat{\Omega} \rightarrow \widehat{\Omega}^{\prime}\right)
$$

which gives the probability that a neutron initially in the state $(d E, d \widehat{\Omega})$ about $(E, \widehat{\Omega})$ is scattered into ( $d E^{\prime}, d \widehat{\Omega}^{\prime}$ ) about ( $E^{\prime}, \widehat{\Omega}^{\prime}$ ). Integrating over energy and solid angle we

[^0]obtain,
$$
\Sigma_{s}(\mathbf{r}, E)=\Sigma_{s}(\mathbf{r}, E, \widehat{\boldsymbol{\Omega}})=\int_{4 \pi} d \widehat{\boldsymbol{\Omega}}^{\prime} \int_{0}^{\infty} d E\left\{\Sigma_{s}\left(\mathbf{r}, E \rightarrow E^{\prime}, \widehat{\boldsymbol{\Omega}} \rightarrow \widehat{\boldsymbol{\Omega}}^{\prime}\right)\right\}
$$
clarifying the rather strange term: double differential cross-section. The dependence on incident direction is suppressed because the nuclei in any macroscopic sample are randomly oriented, thus averaging out any directional dependence.

Having defined the neutron density and the means of interaction we can now define the quantities that arise in the derivation and analysis of the transport equation. In particular, the speed of a neutron multiplied by its interaction probability gives an interaction frequency (i.e. $v \Sigma_{\alpha}$ ). Hence, we define a reaction rate density by,

$$
f(\mathbf{r}, E, \widehat{\Omega}, t)=v \Sigma_{\alpha}(\mathbf{r}, E) n(\mathbf{r}, E, \widehat{\Omega}, t)
$$

where the neutron speed is deduced from the kinetic energy, $v=\sqrt{2 m E}$. The product $v n(\mathbf{r}, E, \widehat{\boldsymbol{\Omega}}, t)$ is often referred to as the angular neutron flux

$$
\varphi(\mathbf{r}, E, \widehat{\boldsymbol{\Omega}}, t)=v n(\mathbf{r}, E, \widehat{\boldsymbol{\Omega}}, t)
$$

This unfortunate terminology is derived from the units, $[$ speed $] \times[$ density $]=[\mathrm{cm}]^{-2}[\mathrm{~s}]^{-1}$, which are consistent with those of the term flux in many other disciplines (e.g. electromagnetic theory). However, the term neutron flux is very misleading as it is a scalar quantity, not a vector, and hence characterizes the expected rate at which neutrons having a particular direction within $d \widehat{\boldsymbol{\Omega}}$ about $\widehat{\boldsymbol{\Omega}}$ are passing through a given elemental surface at time in $d t$ about $t$ regardless of the surface's orientation. ${ }^{2}$ In contrast we define the angular current density,

$$
\mathbf{j}(\mathbf{r}, E, \widehat{\Omega}, t)=\widehat{\Omega} \varphi(\mathbf{r}, E, \widehat{\Omega}, t)
$$

[^1]which assigns the direction $\widehat{\Omega}$ to the corresponding angular neutron flux and hence behaves much more like a traditional flux. Specifically, if we consider an infinitesimal area element $d \mathbf{A}$ located at $\mathbf{r}$ then $\mathbf{j}(\mathbf{r}, E, \widehat{\Omega}, t) d \mathbf{A} d E d \widehat{\boldsymbol{\Omega}}$ is the expected number of neutrons passing through the area $d \mathbf{A}$ per unit time with energy $E$ in $d E$, direction $\widehat{\boldsymbol{\Omega}}$ in $d \widehat{\boldsymbol{\Omega}}$, at a time in $d t$ about $t$.

Finally, we address the question of neutron sources within the reactor. The primary mechanisms of neutron production are fission and various forms of decay. To describe the fission source we assume that $\nu\left(E^{\prime}\right)$ is the average number of neutrons produced in a fission event that was induced by a neutron of energy $E^{\prime}$, such that the total rate at which fission neutrons are created is,

$$
\int_{4 \pi} d \hat{\Omega}^{\prime} \int_{0}^{\infty} d E^{\prime}\left\{\nu\left(E^{\prime}\right) \Sigma_{f}\left(E^{\prime}\right) \varphi(\mathrm{r}, E, \widehat{\Omega}, t)\right\}
$$

The fission neutrons will have an angular distribution that we assume to be isotropic, and a distribution in energy which is described by the fission spectrum $\chi(E)$. The final consideration in defining the fission source is the time required for a fission event to release the fission neutrons. Amazingly, most (i.e. about 99\%), of fission neutrons are emitted promptly (e.g. $\mathcal{O}\left(10^{-4}\right)$ [s] in a thermal reactor [26]) while less than one percent are delayed, being emitted with a temporal distribution that depends on the exact fission chain and isotopes which produce them (e.g. $\left.\mathcal{O}\left(10^{-1}\right)-\mathcal{O}\left(10^{2}\right)[\mathrm{s}]\right)$. If we approximate prompt as instantaneous then the prompt fission source may be written,

$$
s_{f}(\mathbf{r}, E, \widehat{\boldsymbol{\Omega}}, t)=\frac{\chi(E)}{4 \pi} \int_{4 \pi} d \widehat{\boldsymbol{\Omega}}^{\prime} \int_{0}^{\infty} d E^{\prime}\left\{\nu\left(E^{\prime}\right) \Sigma_{f}\left(\mathbf{r}, E^{\prime}\right) \varphi\left(\mathbf{r}, E, \widehat{\boldsymbol{\Omega}}^{\prime}, t\right)\right\}
$$

A detailed consideration of delayed neutrons, although fundamental in reactor control, is beyond the scope of this thesis. Thus, we collect all unspecified sources in the generic term, $q(\mathbf{r}, E, \widehat{\boldsymbol{\Omega}}, t)$ and write the total source,

$$
s(\mathbf{r}, E, \widehat{\boldsymbol{\Omega}}, t)=s_{f}(\mathbf{r}, E, \widehat{\boldsymbol{\Omega}}, t)+q(\mathbf{r}, E, \widehat{\boldsymbol{\Omega}}, t)
$$

The transport equation is commonly derived by considering the balance of neutrons within an arbitrary elemental volume. Many authors have presented discussions of this derivation (e.g. [20, 16, 26, 9] ) which results in the following integrodifferential equation,

$$
\begin{align*}
\frac{1}{v} \frac{\partial \varphi}{\partial t} & +\widehat{\boldsymbol{\Omega}} \cdot \nabla \varphi+\Sigma_{t}(\mathbf{r}, E) \varphi(\mathbf{r}, E, \widehat{\boldsymbol{\Omega}}, t) \\
& =\int_{4 \pi} d \widehat{\boldsymbol{\Omega}} \int_{0}^{\infty} d E^{\prime}\left\{\Sigma_{s}\left(\mathbf{r}, E \rightarrow E^{\prime}, \widehat{\boldsymbol{\Omega}} \rightarrow \widehat{\boldsymbol{\Omega}}^{\prime}\right) \varphi(\mathbf{r}, E, \widehat{\boldsymbol{\Omega}}, t)\right\}+s(\mathbf{r}, E, \widehat{\boldsymbol{\Omega}}, t) \tag{2.1}
\end{align*}
$$

where the first term represents the rate of change with time; the second, the net change due to neutron flow; the third term, the flux lost to scatter and absorption; the fourth, the gain in flux from scattered neutrons; and finally the source discussed above.

### 2.1.2 Simplifying Assumptions

Introducing the transport equation (2.1) in such a general setting serves to convey the fundamental issues of neutronics modelling. However, some additional assumptions may now be introduced which simplify the transport equation without jeopardizing the insight it provides. Specifically we first introduce the one-speed approximation. Indeed this is a useful approximation in its own right, provided that the energy dependence of the interaction coefficients is removed carefully [26]. A less practical analogy in neutron transport, although readily justified in other radiative settings such as monochromatic photon transport [57], is to assume that all scattering is elastic,

$$
\Sigma_{s}\left(\mathrm{r}, E \rightarrow E^{\prime}, \widehat{\boldsymbol{\Omega}} \rightarrow \hat{\boldsymbol{\Omega}}^{\prime}\right)=\Sigma_{s}\left(\mathrm{r}, E, \widehat{\boldsymbol{\Omega}} \rightarrow \widehat{\boldsymbol{\Omega}}^{\prime}\right) \delta\left(E^{\prime}-E\right)
$$

where $\delta\left(E^{\prime}-E\right)$ is the Dirac $\delta$-function. This assumption decouples the energy dependence of the scattering source leaving all terms evaluated at the same energy, hence the explicit energy dependence may be suppressed.

The second assumption is that the scattering is isotropic,

$$
\Sigma_{s}\left(\mathbf{r}, E \rightarrow E^{\prime}, \widehat{\boldsymbol{\Omega}} \rightarrow \widehat{\boldsymbol{\Omega}}^{\prime}\right)=\frac{1}{4 \pi} \Sigma_{s}\left(\mathbf{r}, E \rightarrow E^{\prime}\right)
$$

Technically this is a relatively poor assumption. However, a more rigorous treatment is not justified for our purposes.

Finally we are interested in steady state problems so we will assume that there is no time dependence. This requires that the fission source $s_{f}(\mathbf{r}, E, \widehat{\boldsymbol{\Omega}}, t)$, previously defined to be the source of prompt fission neutrons, be considered the source of all fission neutrons $s_{f}(\mathbf{r}, E, \widehat{\Omega})$. Moreover it was assumed that the fission source is isotropic, thus we will now assume the same of the generic source,

$$
q(\mathbf{r}, E, \widehat{\boldsymbol{\Omega}})=\frac{1}{4 \pi} Q(\mathbf{r}, E)
$$

making the total source isotropic,

$$
s(\mathbf{r}, E, \widehat{\Omega})=\frac{1}{4 \pi} S(\mathbf{r}, E)=\frac{1}{4 \pi}\left(S_{f}(\mathbf{r}, E)+Q(\mathbf{r}, E)\right)
$$

Combining all of these assumptions results in the steady-state, one-speed transport equation with isotropic scattering,

$$
\begin{equation*}
\widehat{\boldsymbol{\Omega}} \cdot \nabla \varphi+\Sigma_{t}(\mathbf{r}) \varphi(\mathbf{r}, \widehat{\boldsymbol{\Omega}})=\frac{\Sigma_{s}(r)}{4 \pi} \int_{4 \pi} \varphi(\mathbf{r}, \widehat{\boldsymbol{\Omega}}) d \widehat{\boldsymbol{\Omega}}^{\prime}+\frac{1}{4 \pi} S(\mathbf{r}) \tag{2.2}
\end{equation*}
$$

which, despite all of these simplifications, is still very difficult to solve.

### 2.1.3 Transport Boundary Conditions

To complete the mathematical description of the steady state neutron transport problem (2.2), we must specify a domain $V$ such that a boundary condition which prescribes the incident angular neutron flux may be imposed. The vacuum ${ }^{3}$ boundary condition is perhaps the most common in transport problems as it describes the interface between the medium of interest (i.e. the reactor core) and the surrounding vacuum. It simply

[^2]specifies that neutrons cannot return to the medium once they have left, and is expressed
\[

$$
\begin{equation*}
\varphi(\mathbf{r}, \widehat{\boldsymbol{\Omega}})=0 \quad \forall(\mathbf{r} \in \partial V, \widehat{\Omega} \cdot \mathbf{n}<0) \tag{2.3}
\end{equation*}
$$

\]

where $\mathbf{n}$ is the outward unit normal to $\partial V$.
Its complement is the reflecting boundary condition whose natural analogy as a perfect mirror suggests its most common application, the treatment of symmetry conditions. Mathematically the reflecting boundary condition is written

$$
\begin{equation*}
\varphi(\mathbf{r}, \widehat{\Omega})=\varphi\left(\mathbf{r}, \widehat{\Omega}_{r}\right) \quad \forall(\mathbf{r} \in \partial V, \widehat{\Omega} \cdot \mathbf{n}<0) \tag{2.4}
\end{equation*}
$$

where $\widehat{\boldsymbol{\Omega}}_{r}$ is the reflection defined such that $\widehat{\boldsymbol{\Omega}} \cdot \mathbf{n}=\widehat{\boldsymbol{\Omega}}_{r} \cdot \mathbf{n}$ and $\widehat{\boldsymbol{\Omega}} \times \widehat{\boldsymbol{\Omega}}_{r} \cdot \mathbf{n}=0$.
Finally both of these conditions are included in the more general albedo condition,

$$
\begin{equation*}
\varphi(\mathbf{r}, \widehat{\boldsymbol{\Omega}})=a \varphi\left(\mathbf{r}, \widehat{\boldsymbol{\Omega}}_{r}\right) \quad \forall(\mathbf{r} \in \partial V, \widehat{\boldsymbol{\Omega}} \cdot \mathbf{n}<0) \tag{2.5}
\end{equation*}
$$

where the albedo, $a=\Sigma_{s} / \Sigma_{t} \in[0,1]$. In some applications it is necessary to generalize this condition to include an angular dependent albedo [57].

### 2.2 Diffusion Theory

### 2.2.1 The $P_{1}$ Approximation

Many of the difficulties encountered in attempting to solve the transport equation arise from the scattering integral. A technique commonly employed to approximate the influence of this integral is the expansion of the angular flux in a truncated series of spherical harmonics. Specifically we assume that

$$
\begin{equation*}
\varphi(\mathbf{r}, \widehat{\boldsymbol{\Omega}})=\sum_{n=0}^{N} \sum_{m=-n}^{n}\left\{\frac{2 n+1}{4 \pi}\right\}^{\frac{1}{2}} \varphi_{n}^{m}(\mathbf{r}) Y_{n}^{m}(\widehat{\boldsymbol{\Omega}}) \tag{2.6}
\end{equation*}
$$

where the expansion coefficients are denoted $\varphi_{n}^{m}(\mathbf{r})$ and $Y_{n}^{m}(\widehat{\Omega})$ are the spherical harmonics [16]. The utilization of this approximation in the transport equation yields a
system of $N$ coupled first order partial differential equations that are referred to as the $P_{N}$ equations.

We will demonstrate that, under the simplifying assumptions of Section 2.1.2, the diffusion approximation arises naturally from the first order expansion (i.e. $N=1$ ) in spherical harmonics

$$
\varphi(\mathbf{r}, \widehat{\boldsymbol{\Omega}})=\frac{1}{\sqrt{4 \pi}} \varphi_{0}^{0}(\mathbf{r}) Y_{0}^{0}(\widehat{\boldsymbol{\Omega}})+\frac{3}{\sqrt{4 \pi}} \sum_{m=-\mathbf{1}}^{1} \varphi_{1}^{m}(\mathbf{r}) Y_{1}^{m}(\widehat{\boldsymbol{\Omega}})
$$

The validity of this truncated expansion has been studied through the asymptotic analysis of the transport equation. In fact, under certain scalings Larsen and Pomraning [51] have shown that, even with anisotropic scattering, $P_{N}$ theory is an asymptotic limit of transport theory. Under the assumption of isotropic scatter the required scaling for $P_{1}$ is such that the albedo $a=1-\mathcal{O}(\epsilon)$, thereby quantifying the more common stipulation that $\Sigma_{t} \gg \Sigma_{a}$.

Now expressing $\widehat{\Omega}$ in terms of spherical harmonics and utilizing their orthogonality this may be written in a more appealing form,

$$
\begin{equation*}
\varphi(\mathbf{r}, \widehat{\boldsymbol{\Omega}})=\frac{1}{4 \pi} \phi(\mathbf{r})+\frac{3}{4 \pi} \widehat{\boldsymbol{\Omega}} \cdot \mathbf{J}(\mathbf{r}) \tag{2.7}
\end{equation*}
$$

where the neutron $f l u x \phi(\mathbf{r})$ is defined by

$$
\phi(\mathbf{r})=\int_{4 \pi} \varphi(\mathbf{r}, \widehat{\Omega}) d \widehat{\Omega}
$$

and $\mathbf{J}(\mathbf{r})$ is the net current

$$
\mathbf{J}(\mathbf{r})=\int_{4 \pi} \varphi(\mathbf{r}, \widehat{\boldsymbol{\Omega}}) \widehat{\boldsymbol{\Omega}} d \widehat{\boldsymbol{\Omega}}=\int_{4 \pi} \mathbf{j}(\mathbf{r}, \widehat{\boldsymbol{\Omega}}) d \widehat{\boldsymbol{\Omega}}
$$

Inserting this $P_{1}$ approximation (2.7) into the transport equation (2.2) yields,

$$
\begin{align*}
\widehat{\boldsymbol{\Omega}} \cdot \nabla & {[\phi(\mathbf{r})+3 \widehat{\boldsymbol{\Omega}} \cdot \mathbf{J}(\mathbf{r})]+\Sigma_{t}(\mathbf{r})[\phi(\mathbf{r})+3 \widehat{\boldsymbol{\Omega}} \cdot \mathbf{J}(\mathbf{r})] } \\
& =\Sigma_{s}(\mathbf{r}) \int_{4 \pi}[\phi(\mathbf{r})+3 \widehat{\boldsymbol{\Omega}} \cdot \mathbf{J}(\mathbf{r})] d \widehat{\boldsymbol{\Omega}}^{\prime}+\chi \nu \Sigma_{f}[\phi(\mathbf{r})+3 \widehat{\boldsymbol{\Omega}} \cdot \mathbf{J}(\mathbf{r})]+Q(\mathbf{r}) \tag{2.8}
\end{align*}
$$

Next, integrating over $\widehat{\boldsymbol{\Omega}}$ one readily obtains

$$
\begin{equation*}
\nabla \cdot \mathbf{J}+\Sigma_{a}(\mathbf{r}) \phi(\mathbf{r})=\nu \chi \Sigma_{f}(\mathbf{r}) \phi(\mathbf{r})+Q(\mathbf{r}) \tag{2.9}
\end{equation*}
$$

which is commonly referred to as the neutron continuity equation [27] or the neutron balance equation [41]. In fact, this balance equation is equivalently the zeroth angular moment of the transport equation, and hence, it is considered to be an exact expression. To develop a second equation we multiply (2.8) by $\widehat{\boldsymbol{\Omega}}$ and integrate over all angles to obtain [16],

$$
\begin{equation*}
\frac{1}{3} \nabla \phi(\mathbf{r})+\Sigma_{t}(\mathbf{r}) \mathbf{J}(\mathbf{r})=0 \tag{2.10}
\end{equation*}
$$

which readily yields Fick's Law,

$$
\begin{equation*}
\mathbf{J}(\mathbf{r})=-D(\mathbf{r}) \nabla \phi(\mathbf{r}) \tag{2.11}
\end{equation*}
$$

with the diffusion coefficient given by $D(\mathbf{r})=\left[3 \Sigma_{t}(\mathbf{r})\right]^{-1}$. It is important to note that (2.11) does not correspond to the first angular moment of the transport equation. It is an approximate closure of the system made possible by the truncated angular expansion of $\varphi(\mathbf{r}, \Omega)$ [equation (2.7)]. In general an $N^{\text {th }}$ order expansion will lead to a system of $N$ PDEs which is comprised of the first $(N-1)$ angular moments of the transport equation and an approximate expression for the $N^{t h}$ angular moment that closes it.

We recognize that substitution of (2.11) into (2.9) gives the standard second order form of the diffusion equation

$$
-\nabla \cdot(D(\mathbf{r}) \nabla \phi(\mathbf{r}))+\Sigma_{a}(\mathbf{r}) \phi(\mathbf{r})=\chi \nu \Sigma_{f}(\mathbf{r}) \phi(\mathbf{r})+Q(\mathbf{r})
$$

However, in the following investigation of nodal discretizations it is important to preserve the distinction between the balance equation, which is an exact expression, and Fick's Law, which is an approximation. In addition, not all neutron transport problems arise
in a multiplying medium (i.e. $\Sigma_{f} \neq 0$ ), and perhaps more importantly, many significant phenomena outside of reactor physics are modelled as diffusive processes. Thus, we will primarily investigate the properties of nodal methods for general diffusion problems. To this end it is convenient to explicitly define two diffusion problems. The first, which we refer to as the one-group diffusion equation, is defined by the following first-order system,

$$
\begin{align*}
\nabla \cdot \mathbf{J}(\mathbf{r})+\Sigma_{a}(\mathbf{r}) \phi(\mathbf{r}) & =Q(\mathbf{r})  \tag{2.12a}\\
\mathbf{J}(\mathbf{r}) & =-D(\mathbf{r}) \nabla \phi(\mathbf{r}) \tag{2.12b}
\end{align*}
$$

and subject to a general boundary condition of the form,

$$
\begin{equation*}
c(\mathbf{r}) \phi+d(\mathbf{r}) \mathbf{J} \cdot \mathbf{n}=g(\mathbf{r}) \tag{2.13}
\end{equation*}
$$

The latter problem, which contains a fission source, will be defined in the following section.

### 2.2.2 Criticality: The Reactor Eigenvalue

In a steady-state, one-group, source-free calculation we have the diffusion equation,

$$
-\nabla \cdot(D(\mathbf{r}) \nabla \phi(\mathbf{r}))+\Sigma_{a}(\mathbf{r}) \phi(\mathbf{r})=\chi \nu \Sigma_{f}(\mathbf{r}) \phi(\mathbf{r})
$$

which, under homogeneous boundary conditions, may not have a solution. Yet the existence of a solution is precisely the objective of a modelling or control computation as it corresponds to a configuration which supports a self-sustained chain reaction. Thus, for practical computations we introduce the multiplication constant $k$, and write

$$
-\nabla \cdot(D(\mathbf{r}) \nabla \phi(\mathbf{r}))+\Sigma_{a}(\mathbf{r}) \phi(\mathbf{r})=\frac{\chi \nu}{k} \Sigma_{f}(\mathbf{r}) \phi(\mathbf{r})
$$

where $k^{-1}$ performs the role of an eigenvalue. In fact we are only interested in the largest eigenvalue, as this corresponds to the only nonnegative eigenfunction [26].This
transformation to an eigenvalue problem guarantees the existence of a solution, and moreover, allows one to characterize the criticality of the core:

$$
\begin{array}{ll}
k>1 & \text { supercritical } \\
k=1 & \text { critical } \\
k<1 & \text { subcritical }
\end{array}
$$

suggesting in rough terms the control which is needed to maintain a critical reactor core. Thus, we define the one-group, fission-source, diffusion problem,

$$
\begin{align*}
\nabla \cdot \mathbf{J}(\mathbf{r})+\Sigma_{a}(\mathbf{r}) \phi(\mathbf{r}) & =\frac{\chi \nu}{k} \Sigma_{f}(\mathbf{r}) \phi(\mathbf{r})  \tag{2.14a}\\
\mathbf{J} & =-D(\mathbf{r}) \nabla \phi(\mathbf{r}) \tag{2.14b}
\end{align*}
$$

subject to the general (homogeneous) boundary condition (2.13), as the second problem.

### 2.2.3 Approximate Boundary Conditions

In practical computations the physical boundary conditions of a problem must be approximated within the context of the mathematical model. This is particularly evident with diffusion theory as the exact transport boundary conditions of Section 2.1.3 depend on $\widehat{\boldsymbol{\Omega}}$, while the flux $\phi(\mathbf{r})$ and the net current $\mathbf{J}(\mathbf{r})$ are integrated quantities. Hence, it is not possible for diffusion theory to satisfy the transport boundary conditions exactly, and we must consider enforcing them in only the integral or weak sense. To this end we decompose the net current into partial currents,

$$
\begin{equation*}
\mathbf{J} \cdot \mathbf{n}=j_{\mathbf{n}}^{+}-j_{\mathbf{n}}^{-} \tag{2.15}
\end{equation*}
$$

which are obtained by integrating over the respective half space

$$
\begin{equation*}
j_{\mathbf{n}}^{ \pm}(\mathbf{r})=\int_{\widehat{\Omega} \cdot \mathbf{n} \gtrless 0}\{\mathbf{j}(\mathbf{r}, \widehat{\Omega}) \cdot \mathbf{n}\} d \widehat{\boldsymbol{\Omega}} \tag{2.16}
\end{equation*}
$$

The partial currents give the expected number of neutrons passing through an elemental surface area, with directions characterized by either $\widehat{\boldsymbol{\Omega}} \cdot \mathbf{n}<0$ or $\widehat{\boldsymbol{\Omega}} \cdot \mathbf{n}>0$. To apply
these definitions within diffusion theory we must also employ the $P_{1}$ approximation of the angular flux (2.7),

$$
\begin{equation*}
j_{\mathbf{n}}^{ \pm}(\mathbf{r}) \approx \frac{1}{4 \pi} \int_{\widehat{\boldsymbol{\Omega}} \cdot \mathbf{n} \gtrless 0}\{[\phi(\mathbf{r})+3 \widehat{\boldsymbol{\Omega}} \cdot \mathbf{J}(\mathbf{r})][\widehat{\boldsymbol{\Omega}} \cdot \mathbf{n}]\} d \widehat{\boldsymbol{\Omega}}=\frac{1}{4} \phi(\mathbf{r}) \pm \frac{1}{2} \mathbf{J}(\mathbf{r}) \cdot \mathbf{n} \tag{2.17}
\end{equation*}
$$

With this expression in hand, the approximation of the transport boundary conditions becomes transparent. In particular, the vacuum boundary condition is satisfied in an integral sense by considering

$$
\int_{\widehat{\Omega} \cdot \mathbf{n}<0}\left\{\varphi\left(\mathbf{r}_{\partial V}, \widehat{\boldsymbol{\Omega}}\right)[\widehat{\boldsymbol{\Omega}} \cdot \mathbf{n}]\right\} d \widehat{\boldsymbol{\Omega}}=0
$$

which is precisely the requirement that $j_{\mathbf{n}}^{-}(\mathbf{r})=0$, and hence, from (2.17) we have the diffusion theory vacuum boundary condition,

$$
\begin{equation*}
j_{\mathbf{n}}^{-}(\mathbf{r})=\frac{1}{4} \phi(\mathbf{r})-\frac{1}{2} \mathbf{J}(\mathbf{r}) \cdot \mathbf{n}=0 \tag{2.18}
\end{equation*}
$$

Similarly one can show that a reflecting boundary is defined by

$$
\begin{equation*}
j_{\mathbf{n}}^{-}(\mathbf{r})=j_{\mathbf{n}}^{+}(\mathbf{r}) \tag{2.19a}
\end{equation*}
$$

and hence, corresponds to the following zero net current condition,

$$
\begin{equation*}
\mathbf{J}(\mathbf{r}) \cdot \mathbf{n}=0 \tag{2.19b}
\end{equation*}
$$

In the general case of the albedo boundary condition, we obtain

$$
\begin{equation*}
j_{\mathrm{n}}^{-}(\mathbf{r})=a j_{\mathrm{n}}^{+}(\mathbf{r}) \tag{2.20a}
\end{equation*}
$$

which may be written,

$$
\begin{equation*}
\frac{1}{4}(1-a) \phi(\mathbf{r})-\frac{1}{2}(1+a) \mathbf{J}(\mathbf{r}) \cdot \mathbf{n}=0 \tag{2.20b}
\end{equation*}
$$

We note that the limiting behaviour of (2.20b) in the albedo is consistent with the transport conditions (i.e. $a=0 \Rightarrow j_{\mathbf{n}}^{-}(\mathbf{r})=0$ and $a=1 \Rightarrow \mathbf{J}(\mathbf{r}) \cdot \mathbf{n}=0$ ).

## Chapter 3

## Nodal Discretizations

The combination of severe fine-scale variations in the interaction coefficients and the large physical size of the reactor core made a naive, fine-mesh, three-dimensional computation inconceivable in the 60 's [40] and continues to make it intractable today [52]. Nevertheless, the safe operation and continued development of nuclear reactors depend on sophisticated neutronics modelling. Research efforts to bridge the gap between computational resources and modelling demands has spanned some 35 years and is ongoing today. The most significant contribution of this research is a broad class of discretization techniques referred to as nodal methods. The distinct view of nodal ideology is that the diffusion equation, when expressed as a first order system of PDEs (2.12), is composed of an exact expression governing the neutron balance over an arbitrary volume, the balance equation (2.12a), and an approximate relationship between the net current and the neutron flux, given by Fick's law (2.12b). This perspective differs significantly from that found in the numerical analysis community, which considers the entire system as an exact model, and it strongly influences the choices that are made throughout the development of the various nodal discretizations.

To investigate these techniques we consider various nodal discretizations of the onegroup diffusion equation (2.12) in two dimensions (i.e. $\mathbf{r}=(x, y)$ and $d \mathbf{r}=d x d y$ ) on a rectangular domain $\Omega$. The discretizations will employ an $L \times M$ tensor product mesh having nodes $\Omega_{i, j}=\left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right] \times\left[y_{j+\frac{1}{2}}, y_{j-\frac{1}{2}}\right]$. The cell centres are simply denoted $\left(x_{i}, y_{j}\right)$, and it is convenient to define the mesh spacing $\Delta x_{i}=x_{i+\frac{1}{2}}-x_{i-\frac{1}{2}}$ and $\Delta y_{j}=y_{j+\frac{1}{2}}-y_{j-\frac{1}{2}}$.


Figure 3.1: An $L \times M$ tensor product mesh imposed on the domain $\Omega$
Inherited naturally from the nodal perspective, and common to all nodal discretizations, is the choice of cell and edge-based unknowns. Generally these are taken to be moments up to some specified order, although we focus our discussion on the lowestorder case and hence employ simple averages. Specifically, the cell-based unknowns are averages defined by

$$
\begin{equation*}
\phi_{i, j}=\frac{1}{\Delta x_{i} \Delta y_{j}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \phi(x, y) d x d y \tag{3.1}
\end{equation*}
$$

while the edge-based scalar unknowns, namely edge averages of the flux, are given by

$$
\begin{array}{ll}
\phi_{i+\frac{1}{2}, j}=\phi_{j}\left(x_{i+\frac{1}{2}}\right), & \phi_{j}(x)=\frac{1}{\Delta y_{j}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \phi(x, y) d y \\
\phi_{i, j+\frac{1}{2}}=\phi_{i}\left(y_{j+\frac{1}{2}}\right), & \phi_{i}(y)=\frac{1}{\Delta x_{i}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \phi(x, y) d x \tag{3.2b}
\end{array}
$$

Similarly, the edge averaged currents are written

$$
\begin{array}{ll}
J_{i+\frac{1}{2}, j}^{ \pm}=\lim _{x \rightarrow x_{i+\frac{1}{2}}^{ \pm}} J_{j}(x), & J_{j}(x)=\frac{1}{\Delta y_{j}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}}\left[\mathbf{J}(x, y) \cdot \mathbf{n}_{x}\right] d y \\
J_{i, j+\frac{1}{2}}^{ \pm}=\lim _{y \rightarrow y_{j+\frac{1}{2}}^{ \pm}} J_{i}(y), & J_{i}(y)=\frac{1}{\Delta x_{i}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}\left[\mathbf{J}(x, y) \cdot \mathbf{n}_{y}\right] d x \tag{3.3b}
\end{array}
$$

where $\mathbf{n}_{x}=[1,0]^{T}$ and $\mathbf{n}_{y}=[0,1]^{T}$.

### 3.1 Classical Approaches

### 3.1.1 Finite Volume and Nodal Balance

The nodal view of the first order system (2.12) suggests that a natural starting point is to integrate the exact balance equation (2.12a) over an arbitrary cell $\Omega_{i, j}$. Thus, in the spirit of finite volume techniques we integrate equation (2.12a) over $\Omega_{i, j}$ and divide through by the cell volume $\Delta \mathbf{r}_{i, j}$ to obtain

$$
\begin{equation*}
\frac{1}{\Delta \mathbf{r}_{i, j}} \iint_{\Omega_{i, j}} \nabla \cdot \mathbf{J} d \mathbf{r}+\frac{1}{\Delta \mathbf{r}_{i, j}} \iint_{\Omega_{i, j}} \Sigma_{a}(\mathbf{r}) \phi(\mathbf{r}) d \mathbf{r}=\frac{1}{\Delta \mathbf{r}_{i, j}} \iint_{\Omega_{i, j}} Q(\mathbf{r}) d \mathbf{r} \tag{3.4}
\end{equation*}
$$

The first term of equation (3.4) is naturally handled by the divergence theorem,

$$
\frac{1}{\Delta \mathbf{r}_{i, j}} \iint_{\Omega_{i, j}} \nabla \cdot \mathbf{J} d \mathbf{r}=\frac{1}{\Delta \mathbf{r}_{i, j}} \int_{\partial \Omega_{i, j}} \mathbf{J} \cdot \mathbf{n} d s
$$

where the line integral motivates the definition of the transverse averaged normal currents (3.3) and is thus discretized exactly by,

$$
\begin{equation*}
\frac{1}{\Delta \mathbf{r}_{i, j}} \int_{\partial \Omega_{i, j}} \mathbf{J} \cdot \mathbf{n} d s=\frac{1}{\Delta x_{i}}\left[J_{i+\frac{1}{2}, j}^{-}-J_{i-\frac{1}{2}, j}^{+}\right]+\frac{1}{\Delta y_{j}}\left[J_{i, j+\frac{1}{2}}^{-}-J_{i, j-\frac{1}{2}}^{+}\right] \tag{3.5}
\end{equation*}
$$

We are not as fortunate with the second term, which is approximated by

$$
\begin{equation*}
\frac{1}{\Delta \mathbf{r}_{i, j}} \iint_{\Omega_{i, j}} \Sigma_{a}(\mathbf{r}) \phi(\mathbf{r}) d \mathbf{r}=\left(\Sigma_{a}\right)_{i, j} \phi_{i, j} \tag{3.6}
\end{equation*}
$$

where $\phi_{i, j}$ is the node average defined in (3.1) and the cell-based absorption cross-section has been derived a priori with an homogenization procedure,

$$
\left(\Sigma_{a}\right)_{i, j} \equiv \mathcal{H}_{\Sigma_{a}}\left(\Sigma_{a}(x, y)\right) \quad \forall \mathbf{r} \in \Omega_{i, j}
$$

A discussion of the complex issues that surround this process will be presented in Chapter 4. However, we note that if we continued to pursue an exact discretization of the balance equation, the homogenized absorption cross-section must in fact be the mean value suggested by (3.6),

$$
\begin{equation*}
\left(\Sigma_{a}\right)_{i, j}=\frac{\iint_{\Omega_{i, j}} \Sigma_{a}(\mathbf{r}) \phi(\mathbf{r}) d \mathbf{r}}{\iint_{\Omega_{i, j}} \phi(\mathbf{r}) d \mathbf{r}} \tag{3.7}
\end{equation*}
$$

which involves the unknown solution. Thus, the homogenized treatment of the second term is exact only if an iterative procedure is implemented, or in the unlikely event that $\Sigma_{a}(\mathbf{r})$ is truly constant over $\Omega_{i, j}$. For many, this approach may seem esoteric as it could be argued that (3.6) is a midpoint approximation of the integral, implying that $\left(\Sigma_{a}\right)_{i, j}$ and $\phi_{i, j}$ are equivalently cell-centred point values. But the distinction is critical to nodal discretizations and typifies the desire to preserve the exact balance equation.

Finally, we accommodate the last term by defining the node averaged source,

$$
Q_{i, j}=\frac{1}{\Delta \mathbf{r}_{i, j}} \iint_{\Omega_{i, j}} Q(\mathbf{r}) d \mathbf{r}
$$

such that combining this with (3.5) and (3.6) in equation (3.4) yields,

$$
\begin{equation*}
\frac{1}{\Delta x_{i}}\left[J_{i+\frac{1}{2}, j}^{-}-J_{i-\frac{1}{2}, j}^{+}\right]+\frac{1}{\Delta y_{j}}\left[J_{i, j+\frac{1}{2}}^{-}-J_{i, j-\frac{1}{2}}^{+}\right]+\left(\Sigma_{a}\right)_{i, j} \phi_{i, j}=Q_{i, j} \tag{3.8}
\end{equation*}
$$

This is a discrete balance equation over $\Omega_{i, j}$, and despite the aforementioned difficulty associated with defining the homogenized absorption cross-section, it is still considered to be of greater physical significance than Fick's Law [cf. Section 2.2.1] Hence, the various
nodal methods have a common foundation in the balance equation, and are characterized by the techniques they employ to derive the additional equations necessary to close the discrete system.

We note that this finite volume treatment of the one-group fission-source balance equation (2.14a) leads to the analogous discrete balance relation,

$$
\begin{equation*}
\frac{1}{\Delta x_{i}}\left[J_{i+\frac{1}{2}, j}^{-}-J_{i-\frac{1}{2}, j}^{+}\right]+\frac{1}{\Delta y_{j}}\left[J_{i, j+\frac{1}{2}}^{-}-J_{i, j-\frac{1}{2}}^{+}\right]+\left(\Sigma_{a}\right)_{i, j} \phi_{i, j}=\frac{\nu \chi}{k}\left(\Sigma_{f}\right)_{i, j} \phi_{i, j} \tag{3.9}
\end{equation*}
$$

where the homogenized fission cross-section $\left(\Sigma_{f}\right)_{i, j}$ is defined in analogy with $\left(\Sigma_{a}\right)_{i, j}$.


Figure 3.2: The location of the unknowns

### 3.1.2 Standard Cell Centered Finite Differences

One of the first techniques employed in reactor modelling was the utilization of finite difference approximations of Fick's law (2.12b) to eliminate the transverse averaged normal currents from the discrete balance equation (3.8) [40]. Specifically, to approximate the transverse averaged normal currents we substitute Fick's law (2.12b) into their definition (3.3), to obtain,

$$
J_{i+\frac{1}{2}, j}^{-}=\lim _{x \rightarrow x_{i+\frac{1}{2}}^{-}}=\frac{1}{\Delta y_{j}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}^{\prime}}\left[-D(x, y) \frac{\partial}{\partial x}\{\phi(x, y)\}\right] d y
$$

along with similar expressions for the other edges. We further assume that an homogenized diffusion coefficient, whose evaluation will be discussed in Chapter 4, has been
defined on each cell [i.e. $D_{i, j}=\mathcal{H}_{D}(D(x, y))$ ] such that,

$$
J_{i+\frac{1}{2}, j}^{-}=\lim _{x \rightarrow x_{i+\frac{1}{2}}^{-}}\left[D_{i, j} \frac{\partial}{\partial x}\left\{\phi_{j}(x)\right\}\right]
$$

for which we use a one-sided difference approximation,

$$
J_{i+\frac{1}{2}, j}^{-}=-\frac{2 D_{i, j}}{\Delta x_{i}}\left(\phi_{i+\frac{1}{2}, j}-\phi_{i, j}\right)
$$

Similar arguments also yield,

$$
J_{i+\frac{1}{2}, j}^{+}=-\frac{2 D_{i+1, j}}{\Delta x_{i+1}}\left(\phi_{i+1, j}-\phi_{i+\frac{1}{2}, j}\right)
$$

Thus, enforcing continuity of the normal current in the weak sense of the transverse average, we can eliminate $\phi_{i+\frac{1}{2}, j}$ to obtain,

$$
\begin{equation*}
J_{i+\frac{1}{2}, j}=-\frac{D_{i+\frac{1}{2}, j}}{\Delta x_{i+\frac{1}{2}}}\left(\phi_{i+1, j}-\phi_{i, j}\right) \tag{3.10}
\end{equation*}
$$

where $\Delta x_{i+\frac{1}{2}}=\left(x_{i+1}-x_{i}\right)$ and the interface diffusion coefficient $D_{i+\frac{1}{2}, j}$ is given by the harmonic mean,

$$
D_{i+\frac{1}{2}, j}=\left[\frac{1}{\Delta x_{i+\frac{1}{2}}}\left\{\frac{\Delta x_{i}}{2 D_{i, j}}+\frac{\Delta x_{i+1}}{2 D_{i+1, j}}\right\}\right]^{-1}=\frac{2 D_{i, j} D_{i+1, j} \Delta x_{i+\frac{1}{2}}}{D_{i, j} \Delta x_{i+1}+D_{i+1, j} \Delta x_{i}}
$$

Developing an analogous expression for $J_{i, j+\frac{1}{2}}$, followed by substitution into the discrete balance equation, we obtain a discrete system governing the node average,

$$
\begin{aligned}
& -\frac{\Delta y_{j}}{\Delta x_{i+\frac{1}{2}}} D_{i+\frac{1}{2}, j} \phi_{i+1, j}-\frac{\Delta y_{j}}{\Delta x_{i-\frac{1}{2}}} D_{i-\frac{1}{2}, j} \phi_{i-1, j}-\frac{\Delta x_{i}}{\Delta y_{j+\frac{1}{2}}} D_{i, j+\frac{1}{2}} \phi_{i, j+1}-\frac{\Delta x_{i}}{\Delta y_{j-\frac{1}{2}}} D_{i, j-\frac{1}{2}} \phi_{i, j-1} \\
& {\left[\left\{\frac{D_{i+\frac{1}{2}, j}}{\Delta x_{i+\frac{1}{2}}}+\frac{D_{i-\frac{1}{2}, j}}{\Delta x_{i-\frac{1}{2}}}\right\} \Delta y_{j}+\left\{\frac{D_{i, j+\frac{1}{2}}}{\Delta y_{j+\frac{1}{2}}}+\frac{D_{i, j-\frac{1}{2}}}{\Delta y_{j-\frac{1}{2}}}\right\} \Delta x_{i}+\Delta x_{i} \Delta y_{j}\left(\Sigma_{a}\right)_{i, j}\right] \phi_{i, j}=\Delta x_{i} \Delta y_{j} Q_{i, j}}
\end{aligned}
$$

The convergence properties of this well known 5-point scheme have been investigated by a number of authors including Weiser and Wheeler [74] who have shown it is second-order under suitable assumptions. But the objective of nodal methods is to attain sufficient
accuracy on a very coarse mesh, approximately one cell per assembly. In this case the errors introduced by the approximations of $J_{i+\frac{1}{2}, j}$ and $J_{i, j+\frac{1}{2}}$ can be severe [40]. Conversely, the fine mesh required to reduce these errors, particularly in three dimensions, is too costly in practice [52].

### 3.2 Simulators

The main objective of a neutronics simulator is the accurate and reliable prediction of incore neutron distributions that may, in turn, be coupled to a global reactor model. Work on such simulators began in the early 60 's and despite the severe memory restrictions of that era (FLARE [21] was developed for a system with 32Kb of RAM) 3D simulations were still targeted. The necessary compromise was the use of an extremely coarse mesh. However, a corresponding loss of accuracy could not be tolerated. This paradox was resolved with the development of a family of nodal methods, simulators, which are based on a clever hybrid of transport physics and diffusion theory. Simulators achieve impressive accuracy ${ }^{1}$ once the associated free parameters are tuned to a particular reactor [34]. They have been used successfully in a wide range of reactors including PWRs, LWRs, BWRs and even fast breeder reactors (see the review by Gupta [34] and the references therein). However, from the perspective of truncation error analysis these methods are inconsistent with diffusion theory. Thus, it is primarily for completeness that we provide a brief historical development of these methods.

In simulator development, the nodal view of the first order system (2.14) is adopted. Since the discrete balance equation (3.9) is viewed as an exact relationship governing the average quantities it contains, it is the discretization of the net currents (3.10) which

[^3]is considered the primary source of error and hence must be improved. In fact, it was generally accepted that the underlying problem was the need to approximate the net current at an interface where it naturally embodied information from the neighbouring nodes. Thus, in the following development due to Henry [40] and discussed by Gupta [34] we consider rewriting the discrete balance equation (3.9) in terms of partial currents (2.16). In particular we approximate the partial currents through the introduction of the partial current coupling coefficients,
\[

$$
\begin{align*}
& j_{i+\frac{1}{2}, j}^{+}=\frac{1}{\Delta y_{j}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}}\left\{j_{\mathbf{n}_{\mathbf{x}}}^{+}\left(x_{i+\frac{1}{2}}, y\right)\right\} d y=\alpha_{(i, j)}^{(i+1, j)} \Delta x_{i} \phi_{i, j}  \tag{3.11a}\\
& j_{i+\frac{1}{2}, j}^{-}=\frac{1}{\Delta y_{j}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}}\left\{j_{\mathbf{n}_{\mathbf{x}}}^{-}\left(x_{i+\frac{1}{2}}, y\right)\right\} d y=\alpha_{(i+1, j)}^{(i, j)} \Delta x_{i+1} \phi_{i+\mathbf{1}, j} \tag{3.11b}
\end{align*}
$$
\]

with analogous definitions for the $y$-oriented coupling coefficients $\alpha_{(i, j)}^{(i, j+1)}$ and $\alpha_{(i, j+1)}^{(i, j)}$. The implication of (3.11) is an approximation of the net current that may be written,

$$
\begin{equation*}
J_{i+\frac{1}{2}, j}=\left(j_{i+\frac{1}{2}, j}^{-}-j_{i+\frac{1}{2}, j}^{+}\right)=-\left[\alpha_{(i+1, j)}^{(i, j)} \Delta x_{i+1}^{\prime} \phi_{i+1, j}-\alpha_{(i, j)}^{(i+1, j)} \Delta x_{i} \phi_{i, j}\right] \tag{3.12}
\end{equation*}
$$

with the $J_{i, j+\frac{1}{2}}$ defined similarly. Thus the partial current coupling coefficients represent a significant increase in the number of degrees of freedom. Finally, utilizing (3.12) the balance equation (3.9) becomes,

$$
\begin{aligned}
& -\alpha_{(i+1, j)}^{(i, j)} \frac{\Delta x_{i+1}}{\Delta x_{i}} \phi_{i+1, j}-\alpha_{(i-1, j)}^{(i, j)} \frac{\Delta x_{i-1}}{\Delta x_{i}} \phi_{i-1, j}-\alpha_{(i, j+1)}^{(i, j)} \frac{\Delta y_{j+1}}{\Delta y_{j}} \phi_{i, j+1}-\alpha_{(i, j-1)}^{(i, j)} \frac{\Delta y_{j-1}}{\Delta y_{j}} \phi_{i, j-1} \\
& \quad+\left[\left(\alpha_{(i, j)}^{(i+1, j)}+\alpha_{(i, j)}^{(i-1, j)}+\alpha_{(i, j)}^{(i, j+1)}+\alpha_{(i, j)}^{(i, j-1)}\right)+\left(\Sigma_{a}\right)_{i, j}-\frac{\chi \nu}{k}\left(\Sigma_{f}\right)_{i, j}\right] \phi_{i, j}=0
\end{aligned}
$$

which, provided we are able to evaluate the $\alpha$ 's, is still of the 5 -point cell-centred variety. Unfortunately, this stencil will be asymmetric in general, making its solution significantly more difficult than its symmetric counterpart.

### 3.2.1 Fine Mesh Correction - Cross Coupling

The Cross Coupling method of Wachspress [71] was perhaps the first neutronics simulator and is based on the partial current coupling approach. It utilizes the diffusion theory expression for the partial currents [equation (2.17)] in conjunction with fine mesh calculations to compute the coupling coefficients. Specifically, a series of partial core fine mesh calculations are performed to obtain a standard solution $\phi_{i, j}^{(f m)}, \phi_{i+\frac{1}{2}, j}^{(f m)}$, and $\phi_{i, j+\frac{1}{2}}^{(f m)}$. Based on these fine mesh values the coupling coefficient is computed by first evaluating the partial currents, e.g.

$$
j_{i+\frac{1}{2}, j}^{+}=\frac{1}{4} \phi_{i+\frac{1}{2}, j}^{(f m)}-\frac{1}{2} \frac{2 D_{i+\frac{1}{2}, j}^{(f f)}}{\Delta x_{i}^{(f m)}+\Delta x_{i+1}^{(f m)}}\left[\phi_{i+1, j}^{(f m)}-\phi_{i, j}^{(f m)}\right]
$$

where we acknowledge that the approximation of $\mathbf{J} \cdot \mathbf{n}$ employed here, although deemed inadequate on very coarse meshes, was considered reasonable for this fine mesh calculation. With the partial currents in hand, the coupling coefficients are readily obtained from equation (3.11).

Unfortunately, the Cross Coupling method has the obvious weakness that if the problem one desires to solve is not sufficiently close to the problem for which the fine mesh computations were performed, the results can be disastrous. It was for this reason, in conjunction with the additional cost of the fine mesh calculations, that other methods were sought. Nevertheless, attempts to utilize partial core or reduced dimension fine mesh computations to enhance the accuracy of the computed coarse mesh solution appears in many of the early methods, and will appear again in the more formal setting of Equivalence Theory [Section 4.1.3].

### 3.2.2 FLARE

The first method that successfully approximated the coupling coefficients independently of diffusion theory and standard discretization techniques was Delp et al.'s FLARE [21].

Their approach first acknowledges that the primary quantity of interest is the power produced in each node, which is denoted $s_{i, j}$ and defined by

$$
s_{i, j}=\Delta x_{i} \Delta y_{j}\left(\Sigma_{f}\right)_{i, j} \phi_{i, j}
$$

Hence they introduce the modified coupling coefficients,

$$
\begin{aligned}
w_{(i, j)}^{(i+1, j)} & =\frac{k}{\nu\left(\Sigma_{f}\right)_{i, j}} \alpha_{(i, j)}^{(i+1, j)}=\frac{k \cdot j_{i+\frac{1}{2}, j}^{+}}{\nu\left(\Sigma_{f}\right)_{i, j} \phi_{i, j} \Delta x_{i} \Delta y_{j}} \\
w_{(i+1, j)}^{(i, j)} & =\frac{k}{\nu\left(\Sigma_{f}\right)_{i+1, j}} \alpha_{(i+1, j)}^{(i, j)}=\frac{k \cdot j_{i+\frac{1}{2}, j}^{-}}{\nu\left(\Sigma_{f}\right)_{i+1, j} \phi_{i+1, j} \Delta x_{i+1} \Delta y_{j}}
\end{aligned}
$$

where the substitution of $\alpha_{(i, j)}^{(k, l)}$ from (3.11) has revealed that, $w_{(i, j)}^{(k, l)}$ is the ratio of, the total number of neutrons leaving $\Omega_{i, j}$ and entering $\Omega_{l, k}$ through their common edge, to the number of neutrons born in $\Omega_{i, j}$. In the FLARE model the assumption is made that all neutrons born in $\Omega_{i, j}$ are absorbed either there, or in one of its 4 nearest neighbours. This fundamental assumption implies that $w_{(i, j)}^{(k, l)}$ is equivalently the probability that a neutron born in $\Omega_{i, j}$ is absorbed in its neighbour $\Omega_{k, l}$ and facilitates the use of transport probabilities to estimate the coupling. Thus, the form of the balance equation employed in most simulators is,

$$
\left[\frac{k}{k_{\infty}^{(i, j)}}-w_{(i, j)}^{(i, j)}\right] s_{i, j}-w_{(i+1, j)}^{(i, j)} s_{i+1, j}-w_{(i-1, j)}^{(i, j)} s_{i-1, j}-w_{(i, j+1)}^{(i, j)} s_{i, j+1}-w_{(i, j-1)}^{(i, j)} s_{i, j-1}=0
$$

where the infinite medium neutron multiplication constant is given by,

$$
k_{\infty}^{(i, j)}=\frac{\nu\left(\Sigma_{f}\right)_{i, j}}{\left(\Sigma_{a}\right)_{i, j}}
$$

and $w_{(i, j)}^{(i, j)}=\left[1-w_{(i, j)}^{(i+1, j)}-w_{(i, j)}^{(i-1, j)}-w_{(i, j)}^{(i, j+1)}-w_{(i, j)}^{(i, j-1)}\right]$, which is obtained directly from the balance equation, is also consistent with the definition of $w_{(i, j)}^{(k, l)}$.

This form of the balance equation is valid for all interior nodes with fuel bearing neighbours. Boundaries are treated with an albedo condition. For example at $x=0$ we have,

$$
\left[\frac{k}{k_{\infty}^{(1, j)}}-w_{(1, j)}^{(1, j)}\right] s_{1, j}-w_{(2, j)}^{(1, j)} s_{2, j}-w_{(1, j+1)}^{(1, j)} s_{1, j+1}-w_{(1, j-1)}^{(1, j)} s_{1, j-1}=0
$$

with $w_{(1, j)}^{(1, j)}=1-w_{(1, j)}^{(2, j)}-\left(1-a_{1, j)} w_{(1, j)}^{(0, j)}-w_{(1, j)}^{(1, j+1)}-w_{(1, j)}^{(1, j-1)}\right.$ for $j=2, \ldots, M-1$ where $a_{1, j}$ is the albedo of the fictitious boundary cell and is considered a free parameter. A peculiarity of choosing $s_{i, j}$ as the unknowns is that non-fuel bearing cells must also be treated with albedo type boundary conditions because in the limit of $\nu\left(\Sigma_{f}\right)_{i, j} \rightarrow 0$, $s_{i, j} \rightarrow 0$ but $w_{(i, j)}^{(k, l)} \rightarrow \infty$.

The absorption probability, $w_{(i, j)}^{(k, l)}$ implemented in FLARE is given by,

$$
w_{(i, j)}^{(k, l)}=(1-g) \frac{M_{i, j}}{2 \Delta}+g \frac{M_{i, j}^{2}}{\Delta^{2}}
$$

where $g$ is a free mixing parameter, $M_{i, j}^{2}$ is the migration area ${ }^{2}$ and $\Delta$ is the mesh spacing (assumed constant) in the appropriate coordinate direction. Both terms arise from onedimensional calculations, with the first based on transport theory and the second arising from a diffusion theory analysis. Although these approximations may seem crude, they are deceptively clever, chosen carefully to perform best on extremely coarse meshes. Naturally as simulators have evolved these kernels have improved, multiple energy groups have been considered, and the strict nearest neighbour interaction assumption has been relaxed [34]. But despite these improvements and their many successes, simulators have not escaped the criticism of being inconsistent methods which are laden with free parameters that must be tuned.

[^4]
### 3.3 Transverse Integration

Maintaining Henry's view [40] (i.e. Section 3.2) that the net current approximation is the primary source of error in the discretization process, the use of partial currents is considered a physically appealing proposition. However, the partial current coupling coefficients that are employed in nodal simulators lack any theoretical foundation. Thus researchers sought consistent discretization procedures which utilize the partial currents directly. Finnemann et al. [31] developed one of the first and most popular discretizations of this type, by introducing what has become a cornerstone of modern nodal methods, the transverse integration procedure.

The transverse integration procedure first assumes that a homogenized diffusion coefficient and absorption cross-section are defined on each cell. Equivalently, one may consider this an assumption that the problem itself has piecewise constant coefficients. In two dimensions transverse integration implies operating on an equation with either of the following,

$$
\frac{1}{\Delta y_{j}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}}(\cdot) d y, \quad \frac{1}{\Delta x_{i}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}(\cdot) d x
$$

Thus, transverse integrating the balance equation (2.12a) we obtain,

$$
\begin{align*}
-\frac{\partial}{\partial x}\left\{J_{j}(x)\right\}+\left(\Sigma_{a}\right)_{i, j} \phi_{j}(x) & =-\frac{1}{\Delta y_{j}}\left\{J_{j+\frac{1}{2}}^{-}(x)-J_{j-\frac{1}{2}}^{+}(x)\right\}+Q_{j}(x)  \tag{3.13a}\\
-\frac{\partial}{\partial y}\left\{J_{i}(y)\right\}+\left(\Sigma_{a}\right)_{i, j} \phi_{i}(y) & =-\frac{1}{\Delta x_{i}}\left\{J_{i+\frac{1}{2}}^{-}(y)-J_{i-\frac{1}{2}}^{+}(y)\right\}+Q_{i}(y) \tag{3.13b}
\end{align*}
$$

where $x \in\left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right]$ and $y \in\left[y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}\right]$. The pseudo-source terms $J_{j \pm \frac{1}{2}}^{\mp}(x)$ and $J_{i \pm \frac{1}{2}}^{\mp}(y)$ are one-sided limits of the point-wise normal currents along the respective edges of the cell, while the transverse averaged sources $Q_{j}(x)$ and $Q_{i}(y)$ are defined in analogy with the transverse averaged flux. The term pseudo-source is adopted because it will prove useful to treat these transverse edge currents as sources despite the obvious fact
that they are fundamental unknowns. To emphasize this role the transverse leakage is commonly defined [52],

$$
\begin{align*}
\mathcal{L}_{i, j}^{x}(x) & =\frac{1}{\Delta y_{j}}\left\{J_{j+\frac{1}{2}}^{-}(x)-J_{j-\frac{1}{2}}^{+}(x)\right\}  \tag{3.14a}\\
\mathcal{L}_{i, j}^{y}(y) & =\frac{1}{\Delta x_{i}}\left\{J_{i+\frac{1}{2}}^{-}(y)-J_{i-\frac{1}{2}}^{+}(y)\right\} \tag{3.14b}
\end{align*}
$$

Similarly the transverse average of Fick's Law, which relies on the assumption of an homogenized diffusion coefficient, is simply,

$$
\begin{array}{ll}
J_{j}(x)=-D_{i, j} \frac{\partial}{\partial x}\left\{\phi_{j}(x)\right\} & x \in\left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right] \\
J_{i}(y)=-D_{i, j} \frac{\partial}{\partial y}\left\{\phi_{i}(y)\right\} & y \in\left[y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}\right] \tag{3.15b}
\end{array}
$$

Inserting (3.14) and (3.15) into (3.13) we obtain a system of second order ODEs

$$
\begin{align*}
-D_{i, j} \frac{\partial^{2}}{\partial x^{2}}\left\{\phi_{j}(x)\right\}+\left(\Sigma_{a}\right)_{i, j} \phi_{j}(x) & =\mathcal{L}_{i, j}^{x}(x)+Q_{j}(x)  \tag{3.16a}\\
-D_{i, j} \frac{\partial^{2}}{\partial y^{2}}\left\{\phi_{i}(y)\right\}+\left(\Sigma_{a}\right)_{i, j} \phi_{i}(y) & =\mathcal{L}_{i, j}^{y}(y)+Q_{i}(y) \tag{3.16b}
\end{align*}
$$

There are two distinct applications of the transverse integrated equations. The first is the weighted residuals procedure of the nodal expansion method (NEM) which we outline below [Section 3.4.1] and the second is the Nodal Integration Method (NIM) which we discuss in Section 3.5.1.

### 3.4 Polynomial Nodal Methods - NEM

The development of modern consistent nodal discretizations began in the mid 70 's. These methods were based on local polynomial expansions. The first popular polynomial method was the nodal expansion method (NEM) of Finnemann et al. [31]. In fact, although some variations and improvements have been considered [52], the NEM
ideology still dominates the polynomial class of nodal methods. In its lowest-order form, NEM considers a quadratic expansion of the transverse averaged flux [i.e. $\phi_{j}(x)$ and $\left.\phi_{i}(y)\right]$ on each cell. The expansion coefficients are determined by applying Fick's Law in combination with the discrete balance equation (3.8) and continuity of the normal current. Considerable effort has been made to utilize higher order polynomial expansions within NEM. The difficulty this creates is centred around the evaluation of the higher order expansion coefficients. In particular, the weighted residual procedure that is typically used relies on transverse integration of the balance equation (2.12a) and as a result an approximation of the transverse normal currents (i.e. the transverse leakage) is also required. In the following discussion we highlight the issues surrounding the weighted residual procedure and the approximation of the transverse leakage, although we focus on the lowest-order case.

### 3.4.1 Components of the NEM Discretization

## A Polynomial Basis

The NEM treatment of the transverse integrated ODEs (3.16) is based on a low order polynomial expansion of the transverse integrated flux. Specifically, we write,

$$
\begin{align*}
\phi_{j}(x) & =\sum_{l=0}^{N_{x}} a_{i, j}^{(x, l)} \psi_{l}(x)  \tag{3.17a}\\
\phi_{i}(y) & =\sum_{k=1}^{N_{y}} a_{i, j}^{(y, k)} \psi_{k}(y) \tag{3.17b}
\end{align*}
$$

where $\mathcal{P}_{N} \in \operatorname{span}\left\{\psi_{0}, \ldots, \psi_{N}\right\}$ is a polynomial of degree $N$ [i.e. either $N_{x}$ or $N_{y}$ ]. The basis functions employed in [31] are chosen to satisfy certain properties at the expense of universal orthogonality.

Definition 3.4.1. The first five NEM basis functions are given by,

$$
\begin{aligned}
& \psi_{0}(\xi)=1 \\
& \psi_{1}(\xi)=\xi \\
& \psi_{2}(\xi)=3 \xi^{2}-\frac{1}{4} \\
& \psi_{3}(\xi)=\xi\left(\xi-\frac{1}{2}\right)\left(\xi+\frac{1}{2}\right) \\
& \psi_{4}(\xi)=\left(\xi^{2}-\frac{1}{20}\right)\left(\xi-\frac{1}{2}\right)\left(\xi+\frac{1}{2}\right)
\end{aligned}
$$

where $\xi=\left(x-x_{i}\right) / \Delta x_{i}$ in the expansion of $\phi_{j}(x)$ while $\xi=\left(y-y_{j}\right) / \Delta y_{j}$ for $\phi_{i}(y)$.

Constructing the NEM basis. To construct the basis we first specify that $\psi_{l}(x) \forall l \geq 1$ be orthogonal to $\psi_{0}(x)$. Hence multiplying (3.17a) by $\psi_{0}(x)$ and integrating gives,

$$
a_{i, j}^{(x, 0)} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}\left[\psi_{0}(x)\right]^{2} d x=\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \phi_{j}(x) \psi_{0}(x) d x
$$

It is natural to choose $\psi_{0}(x)=1$ obtaining $a_{i, j}^{(x, 0)}=\phi_{i, j}$. As a consequence $a_{i, j}^{(y, 0)}=\phi_{i, j}$ and thus these expansions are, by definition, consistent with the node average,

$$
\phi_{i, j}=\frac{1}{\Delta x_{i}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \phi_{j}(x) d x=\frac{1}{\Delta y_{j}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \phi_{i}(y) d y
$$

Now requiring that the linear and quadratic basis functions be orthogonal to $\psi_{0}(x)$ and $\left\{\psi_{0}(x), \psi_{1}(x)\right\}$ respectively, we obtain

$$
\begin{align*}
\psi_{\mathbf{1}}(\xi) & =\alpha_{\mathbf{1}} \xi  \tag{3.18a}\\
\psi_{2}(\xi) & =\alpha_{2}\left[\xi^{2}-\frac{1}{12}\right] \tag{3.18b}
\end{align*}
$$

where the normalization constants $\alpha_{1}, \alpha_{2}$ are defined as follows. Recall that the definition of the transverse averaged flux implies that $\phi_{i \pm \frac{1}{2}, j}=\phi_{j}\left(x_{i \pm \frac{1}{2}}\right)$, providing two constraints with which to define the normalization constants. Taking $N_{x}=2$, substitution of $x_{i \pm \frac{1}{2}}$
in (3.18) results in the following linear system

$$
\left[\begin{array}{cc}
-\frac{1}{2} \alpha_{1} & \frac{1}{6} \alpha_{2}  \tag{3.19}\\
\frac{1}{2} \alpha_{1} & \frac{1}{6} \alpha_{2}
\end{array}\right]\left[\begin{array}{c}
-a_{i, j}^{(x, 1)} \\
-a_{i, j}^{(x, 2)}
\end{array}\right]=\left[\begin{array}{c}
\phi_{i-\frac{1}{2}, j}-\phi_{i, j} \\
\phi_{i+\frac{1}{2}, j}-\phi_{i, j}
\end{array}\right]
$$

from which the coefficients are easily obtained,

$$
\begin{aligned}
& a_{i, j}^{(x, 1)}=\frac{1}{\alpha_{1}}\left(\phi_{i+\frac{1}{2}, j}-\phi_{i-\frac{1}{2}, j}\right) \\
& a_{i, j}^{(x, 2)}=\frac{3}{\alpha_{2}}\left(\phi_{i+\frac{1}{2}, j}-2 \phi_{i, j}+\phi_{i-\frac{1}{2}, j}\right)
\end{aligned}
$$

We take $\alpha_{1}=1$ and $\alpha_{2}=3$ to obtain the basis functions $\psi_{1}(\xi)$ and $\psi_{2}(\xi)$.
This unusual construction of the quadratic expansion ensures the global continuity of the transverse averaged flux. However, we must prevent higher order terms from altering the edge averaged flux and thus higher order basis functions must satisfy,

$$
\psi_{l}\left(x_{i \pm \frac{1}{2}}\right)=0 \quad l=3, \ldots, N
$$

For example, for the cubic basis function, this constraint implies the form,

$$
\psi_{3}(\xi)=\left[c_{0}+c_{1} \xi\right]\left(\xi-\frac{1}{2}\right)\left(\xi+\frac{1}{2}\right)
$$

so orthogonality with $\psi_{0}(x)$ implies that $c_{0}=0$. The normalization remains, and appears to be chosen arbitrarily, as setting $c_{1}=1$ we obtain $\psi_{3}(\xi)$. Similarly for the quartic basis functions we have,

$$
\psi_{4}(\xi)=\left[c_{0}+c_{1} \xi+c_{2} \xi^{2}\right]\left(\xi-\frac{1}{2}\right)\left(\xi+\frac{1}{2}\right)
$$

and orthogonality gives $c_{1}=0$ and $c_{2}=-20 c_{0}$. Again the normalization seems arbitrary, and taking $c_{2}=1$ immediately yields $\psi_{4}(\xi)$.

The consequence of this approach is that expansions of degree less than two are not considered while the expansion of degree two is completely determined by the edge and cell based fluxes. These quadratic expansions are given in the following proposition.

Proposition 3.1. The coefficients of the lowest-order NEM expansion of the transverse averaged flux [i.e. $N_{x}=N_{y}=2$ in equation (3.17)] are given by

$$
\begin{array}{ll}
a_{i, j}^{(x, 0)}=\phi_{i, j} & a_{i, j}^{(y, 0)}=\phi_{i, j} \\
a_{i, j}^{(x, 1)}=\phi_{i+\frac{1}{2}, j}-\phi_{i-\frac{1}{2}, j} & a_{i, j}^{(y, 1)}=\phi_{i, j+\frac{1}{2}}-\phi_{i, j-\frac{1}{2}} \\
a_{i, j}^{(x, 2)}=\phi_{i+\frac{1}{2}, j}-2 \phi_{i, j}+\phi_{i-\frac{1}{2}, j} & a_{i, j}^{(y, 2)}=\phi_{i, j+\frac{1}{2}}-2 \phi_{i, j}+\phi_{i, j-\frac{1}{2}}
\end{array}
$$

## Transverse Leakage Approximations

In addition to the transverse averaged flux, the transverse integrated ODEs depend on the transverse leakages (3.14). Although these terms are treated as sources they are in fact unknown, coupling the ODEs through Fick's law (3.15). Thus, in NEM the transverse leakages are also approximated with a polynomial expansion,

$$
\begin{align*}
\mathcal{L}_{i, j}^{(x)}(x) & =\sum_{l=0}^{M_{x}} b_{i, j}^{(x, l)} \varphi_{l}(x) \tag{3.20a}
\end{align*} \quad x \in\left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right], ~ x \in\left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right]
$$

where $\mathcal{P}_{M} \in \operatorname{span}\left\{\varphi_{0}, \ldots, \varphi_{M}\right\}$ is a polynomial of degree $M$ [i.e. either $M_{x}$ or $M_{y}$ ]. In the following discussion the expansion of $\mathcal{L}_{i, j}^{(x)}$ is considered in detail; analogous definitions and relations are assumed for $\mathcal{L}_{i, j}^{(y)}$.

The quadratic expansion suggested in [31] utilizes the NEM basis [i.e. Definition 3.4.1], Hence, the expansion coefficients follow immediately from Definition 3.1,

$$
\begin{aligned}
& b_{i, j}^{(x, 0)}=\mathcal{L}_{i, j}^{(x, 0)} \\
& b_{i, j}^{(x, 1)}=\mathcal{L}_{i+\frac{1}{2}, j}^{-}-\mathcal{L}_{i-\frac{1}{2}, j}^{+} \\
& b_{i, j}^{(x, 2)}=\mathcal{L}_{i+\frac{1}{2}, j}^{-}-2 \mathcal{L}_{i, j}^{(x, 0)}+\mathcal{L}_{i-\frac{1}{2}, j}^{+}
\end{aligned}
$$

where we have adopted the notation

$$
\begin{aligned}
\mathcal{L}_{i, j}^{(x, 0)} & =\frac{1}{\Delta x_{i}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathcal{L}_{i, j}^{(x)}(x) d x \\
\mathcal{L}_{i+\frac{1}{2}, j}^{-} & =\lim _{x \rightarrow x_{i+\frac{1}{2}}^{-}} \mathcal{L}_{i, j}^{(x)}(x) \\
\mathcal{L}_{i-\frac{1}{2}, j}^{+} & =\lim _{x \rightarrow x_{i-\frac{1}{2}}^{+}} \mathcal{L}_{i, j}^{(x)}(x)
\end{aligned}
$$

Unfortunately, the edge values of the transverse leakage are unknown and moreover are not considered within the NEM discretization. Hence additional constraints are obtained by enforcing the continuity of the transverse leakage and its first derivative at cell edges. However, rather than incorporating these constraints into the discretization they are viewed in [31] as auxiliary equations, forming a tridiagonal system that relates the edge values $\mathcal{L}_{i+\frac{1}{2}, j}$ to the cell averages $\mathcal{L}_{i, j}^{(x, 0)}$. This implies that $(L+M)$ such systems must be inverted for each iteration of the global discretization.

Unwilling to accept the additional expense of these auxiliary calculations, most implementations of NEM have employed the local technique suggested in [10]. Specifically, a quadratic expansion in $\left\{1, \xi, \xi^{2}\right\}$ with $\xi=\left(x-x_{i}\right) / \Delta x_{i}$ is adopted, and the expansion coefficients are determined such that the average transverse leakage over the central node $\Omega_{i, j}$ and its nearest neigbours $\Omega_{i-1, j}, \Omega_{i+1, j}$ are reproduced. Thus the expansion coefficients of this local quadratic are characterized by the following system,

$$
\left[\begin{array}{ccc}
\beta_{i-1, i} & -\frac{1}{2} \gamma_{i-1, i}^{(1)} & \frac{1}{12} \gamma_{i-1, i}^{(2)}  \tag{3.21}\\
1 & 0 & \frac{1}{12} \\
\beta_{i+1, i} & \frac{1}{2} \gamma_{i+1, i}^{(1)} & \frac{1}{12} \gamma_{i+1, i}^{(2)}
\end{array}\right]\left[\begin{array}{c}
b_{i, j}^{(x, 0)} \\
b_{i, j}^{(x, 1)} \\
b_{i, j}^{(x, 2)}
\end{array}\right]=\left[\begin{array}{c}
\mathcal{L}_{i-1, j}^{(x, 0)} \\
\mathcal{L}_{i, j}^{(x, 0)} \\
\mathcal{L}_{i+1, j}^{(x, 0)}
\end{array}\right]
$$

where $\beta_{l, k}=\Delta x_{l} / \Delta x_{k}$ and $\gamma_{l, k}^{(1)}=\beta_{l, k}\left(\beta_{l, k}+1\right), \gamma_{l, k}^{(2)}=\beta_{l, k}\left(4 \beta_{l, k}^{2}+6 \beta_{l, k}+3\right)$.
An alternative local scheme based on a piecewise linear expansion was also suggested in [31]. However, we are specifically interested in the lowest-order NEM for which a
simple constant approximation,

$$
\begin{equation*}
\mathcal{L}_{i, j}^{(x)}(x)=\mathcal{L}_{i, j}^{(x, 0)}=\frac{1}{\Delta x_{i}}\left[J_{i+\frac{1}{2}, j}-J_{i-\frac{1}{2}, j}\right] \tag{3.22}
\end{equation*}
$$

is employed. We note that this zeroth moment approximation is consistent with the balance equation.

## Weighted Residual Procedure

The development of the transverse integrated ODEs was motivated by the desire to utilize higher order polynomial expansions of the transverse averaged flux and transverse leakage. The transverse integrated ODEs facilitate this through a weighted residual procedure developed in [31]. Specifically, we consider a set of weight functions $w_{n}(\xi)$, $n=0, \ldots, N_{w}$, such that operating on the transverse integrated ODEs (3.16) with

$$
\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} w_{n}(x)(\cdot) d x, \text { and } \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} w_{m}(y)(\cdot) d y
$$

yields,

$$
\begin{aligned}
& \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} w_{n}(x)\left[-D_{i, j} \frac{\partial^{2}}{\partial x^{2}}\left\{\phi_{j}(x)\right\}\right] d x+\left(\Sigma_{a}\right)_{i, j} \phi_{i, j}^{(n, 0)}=Q_{i, j}^{(n, 0)}+\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} w_{n}(x) \mathcal{L}_{i, j}^{(x)}(x) d x \\
& \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} w_{m}(y)\left[-D_{i, j} \frac{\partial^{2}}{\partial y^{2}}\left\{\phi_{i}(y)\right\}\right] d y+\left(\Sigma_{a}\right)_{i, j} \phi_{i, j}^{(0, m)}=Q_{i, j}^{(0, m)}+\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} w_{m}(y) \mathcal{L}_{i, j}^{(y)}(y) d y
\end{aligned}
$$

where $n=0, \ldots, N_{w}^{(x)}, m=0, \ldots, N_{w}^{(y)}$ and $\phi_{i, j}^{(l, k)}$ defines the weighted integral of the scalar flux,

$$
\phi_{i, j}^{(l, k)}=\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} w_{l}(x) w_{k}(y) \phi(x, y) d x d y
$$

It is assumed that $w_{0}(\xi) \equiv 1$, with $x=x_{i}+\Delta x_{i} \xi$ and $y=y_{j}+\Delta y_{j} \xi$. We note that the most popular choice of weight functions is the NEM basis given in Definition 3.4.1. The resulting procedure is typically referred to as moment weighting [52].

Finally, to determine $N_{w}^{(x)}$ and $N_{w}^{(y)}$ observe that with a quadratic expansion of the transverse average flux and a constant approximation of the transverse leakage no additional equations are required. Moreover, with $w_{0}(\xi) \equiv 1$ both weighted transverse integrated ODEs yield the balance equation (3.8), implying that the quadratic/constant expansion is consistent with this zeroth order weighted residual procedure. Thus the required number of weighted residual equations in each spatial direction is $N_{w}^{(x)}=$ $N_{x}+M_{x}-2$ and $N_{w}^{(y)}=N_{y}+M_{y}-2$.

### 3.4.2 The Lowest-Order NEM

Definition 3.4.2. The lowest-order NEM discretization of the one-group diffusion equation (2.12) is comprised of the cell balance equation (3.8) and the following cell-based normal currents,

$$
\begin{align*}
& J_{i+\frac{1}{2}, j}^{-}=-2 \frac{D_{i, j}}{\Delta x_{i}}\left\{\phi_{i-\frac{1}{2}, j}-3 \phi_{i, j}+2 \phi_{i+\frac{1}{2}, j}\right\}  \tag{3.23a}\\
& J_{i-\frac{1}{2}, j}^{+}=-2 \frac{D_{i, j}}{\Delta x_{i}}\left\{-2 \phi_{i-\frac{1}{2}, j}+3 \phi_{i, j}-\phi_{i+\frac{1}{2}, j}\right\} \tag{3.23b}
\end{align*}
$$

with analogous expressions defining $J_{i, j \pm \frac{1}{2}}^{ \pm}$. The continuity of the normal current is imposed across all interior edges, in the weak sense of the transverse average,

$$
\begin{equation*}
J_{i+\frac{1}{2}, j}^{-}=J_{i+\frac{1}{2}, j}^{+}, \quad J_{i, j+\frac{1}{2}}^{-}=J_{i, j+\frac{1}{2}}^{+} \tag{3.24}
\end{equation*}
$$

in conjunction with the discrete boundary equations to complete the discretization.
Derivation of the Lowest-Order NEM. The lowest-order NEM assumes a constant leakage approximation [i.e. $M_{x}=M_{y}=0$ ] and hence the zeroth weighted residual of either transverse integrated ODE (3.16) is the balance equation (3.8). A quadratic flux expansion [i.e. $N_{x}=N_{y}=2$ ] is also characteristic of the lowest-order NEM. Substitution of this expansion (3.17a) into the transverse averaged Fick's Law (3.15a) immediately yields (3.23). The analogous treatment follows for the $y$ oriented net currents.

## Interface Current Formulation

Recall the motivation was to develop a discretization that employed the partial currents directly. This is now possible if we utilize the diffusion theory approximation of the partial currents (2.17) to write

$$
\left.\begin{array}{rl}
\phi_{i+\frac{1}{2}, j} & =2\left(j_{i+\frac{1}{2}, j}^{+}+j_{i+\frac{1}{2}, j}^{-}\right.
\end{array}\right)
$$

We note that enforcing the continuity of the normal current (3.24), in conjunction with the continuity of the scalar flux ensures continuity of the partial currents.

Definition 3.4.3. The interface current formulation of the lowest-order NEM [Definition 3.4.2] is defined by the following expressions for the outgoing partial currents,

$$
\begin{align*}
& \tau_{i, j}^{(O, x)} j_{i+\frac{1}{2}, j}^{+}=-4 \frac{D_{i, j}}{\Delta x_{i}} j_{i-\frac{1}{2}, j}^{+}+\tau_{i, j}^{(I, x)} j_{i+\frac{1}{2}, j}^{-}+6 \frac{D_{i, j}}{\Delta x_{i}}\left\{1+4 \frac{D_{i, j}}{\Delta x_{i}}\right\} \phi_{i, j}  \tag{3.26a}\\
& \tau_{i, j}^{(O, x)} j_{i-\frac{1}{2}, j}^{-}=\tau_{i, j}^{(I, x)} j_{i-\frac{1}{2}, j}^{+}-4 \frac{D_{i, j}}{\Delta x_{i}} j_{i+\frac{1}{2}, j}^{-}+6 \frac{D_{i, j}}{\Delta x_{i}}\left\{1+4 \frac{D_{i, j}}{\Delta x_{i}}\right\} \phi_{i, j} \tag{3.26b}
\end{align*}
$$

where

$$
\begin{aligned}
\tau_{i, j}^{(O, x)} & =\left(1+8 \frac{D_{i, j}}{\Delta x_{i}}\right)^{2}-\left(4 \frac{D_{i, j}}{\Delta x_{i}}\right)^{2} \\
\tau_{i, j}^{(I, x)} & =\left(1-8 \frac{D_{i, j}}{\Delta x_{i}}\right)^{2}+\left(4 \frac{D_{i, j}}{\Delta x_{i}}\right)^{2}
\end{aligned}
$$

with analogous definitions for the outgoing partial currents in $y$ (i.e. $j_{i, j+\frac{1}{2}}^{+}$and $j_{i, j-\frac{1}{2}}^{-}$) and the coefficients $\tau_{i, j}^{(O, y)}$ and $\tau_{i, j}^{(I, y)}$. The partial current balance equation is written in the form,

$$
\begin{array}{r}
\frac{8 D_{i, j}}{\tau_{i, j}^{(O, x)} \Delta x_{i}^{2}}\left\{\frac{D_{i, j}}{\Delta x_{i}}-1\right\}\left(j_{i+\frac{1}{2}, j}^{-}+j_{i-\frac{1}{2}, j}^{+}\right)+\frac{8 D_{i, j}}{\tau_{i, j}^{(O, y)} \Delta y_{j}^{2}}\left\{\frac{D_{i, j}}{\Delta y_{j}}-1\right\}\left(j_{i, j+\frac{1}{2}}^{-}+j_{i, j-\frac{1}{2}}^{+}\right)  \tag{3.27}\\
\\
+\left[12 \frac{D_{i, j}}{\Delta x_{i}^{2}}\left\{1+4 \frac{D_{i, j}}{\Delta x_{i}}\right\}+12 \frac{D_{i, j}}{\Delta y_{j}^{2}}\left\{1+4 \frac{D_{i, j}}{\Delta y_{j}}\right\}+\left(\Sigma_{a}\right)_{i, j}\right] \phi_{i, j}=Q_{i, j}
\end{array}
$$

The discrete boundary conditions complete the discretization.

Derivation of the Interface Current Formulation. The objective is to eliminate the net currents and the edge averaged scalar flux from the NEM in favour of the partial currents. Substitution of (3.25) in (3.23) gives,

$$
\begin{align*}
& \left\{1+8 \frac{D_{i, j}}{\Delta x_{i}}\right\} j_{i+\frac{1}{2}, j}^{+}=\left\{1-8 \frac{D_{i, j}}{\Delta x_{i}}\right\} j_{i+\frac{1}{2}, j}^{-}-4 \frac{D_{i, j}}{\Delta x_{i}}\left(j_{i-\frac{1}{2}, j}^{+}+j_{i-\frac{1}{2}, j}^{-}\right)+6 \frac{D_{i, j}}{\Delta x_{i}} \phi_{i, j}  \tag{3.28a}\\
& \left\{1-8 \frac{D_{i, j}}{\Delta x_{i}}\right\} j_{i-\frac{1}{2}, j}^{+}=\left\{1+8 \frac{D_{i, j}}{\Delta x_{i}}\right\} j_{i-\frac{1}{2}, j}^{-}+4 \frac{D_{i, j}}{\Delta x_{i}}\left(j_{i+\frac{1}{2}, j}^{+}+j_{i+\frac{1}{2}, j}^{-}\right)+6 \frac{D_{i, j}}{\Delta x_{i}} \phi_{i, j} \tag{3.28b}
\end{align*}
$$

The interface current formulation now expresses the outgoing partial currents (i.e. $j_{i-\frac{1}{2}, j}^{-}$ and $j_{i+\frac{1}{2}, j}^{+}$) in terms of the incoming partial currents (i.e. $j_{i-\frac{1}{2}, j}^{+}$and $j_{i+\frac{1}{2}, j}^{-}$) and the node average. Specifically, substituting the outgoing current $j_{i-\frac{1}{2}, j}^{-}$defined by (3.28b) into (3.28a), collecting terms and simplifying yields (3.26a). Similarly we obtain (3.26b) and moreover, the rotated analogues hold in $y$.

To express the balance equation in terms of incoming partial currents we employ (3.25b) to rewrite the average transverse leakage,

$$
\frac{1}{\Delta x_{i}}\left[J_{i+\frac{1}{2}, j}-J_{i-\frac{1}{2}, j}\right]=\frac{1}{\Delta x_{i}}\left[\left(j_{i+\frac{1}{2}, j}^{+}-j_{i+\frac{1}{2}, j}^{-}\right)-\left(j_{i-\frac{1}{2}, j}^{+}-j_{i-\frac{1}{2}, j}^{-}\right)\right]
$$

followed by substitution of the outgoing partial currents from (3.26), such that upon collecting terms and simplifying the coefficients we obtain,

$$
\frac{1}{\Delta x_{i}}\left[J_{i+\frac{1}{2}, j}-J_{i-\frac{1}{2}, j}\right]=\frac{8 D_{i, j}}{\tau_{i, j}^{(O, x)} \Delta x_{i}}\left\{\frac{D_{i, j}}{\Delta x_{i}}-1\right\}\left(j_{i+\frac{1}{2}, j}^{-}+j_{i-\frac{1}{2}, j}^{+}\right)+12 \frac{D_{i, j}}{\Delta x_{i}}\left\{1+4 \frac{D_{i, j}}{\Delta x_{i}}\right\} \phi_{i, j}
$$

Thus equation (3.27) is obtained immediately with the substitution of this expression for the average transverse leakage in $x$, along with its $y$-oriented analogue into the balance equation (3.8).

A distinct advantage of the interface current formulation is that higher order approximations of the transverse leakage may be employed without significantly altering the sparsity structure of the system. This is not the case for the flux formulations which follow. Moreover, numerical studies have demonstrated that higher-order approximations
of the transverse leakage decrease the magnitude of the discretization error on a given mesh (e.g. [31, 52, 45]). Although, to the best of our knowledge their influence on the order of the truncation error has not been studied.

## Flux Formulation

Definition 3.4.4. The flux formulation of the lowest-order NEM [Definition 3.4.2] is defined by the following flux based balance equation,

$$
\begin{equation*}
-6 \frac{D_{i, j}}{\Delta x_{i}^{2}}\left(\phi_{i-\frac{1}{2}, j}+\phi_{i+\frac{1}{2}, j}\right)+\frac{12 D_{i, j}}{\Delta x_{i} \Delta y_{j}} \frac{\mu_{i, j}^{(x, y)}}{\varrho_{i, j}^{(x, y)}} \phi_{i, j}-6 \frac{D_{i, j}}{\Delta y_{j}^{2}}\left(\phi_{i, j-\frac{1}{2}}+\phi_{i, j+\frac{1}{2}}\right)=Q_{i, j} \tag{3.29}
\end{equation*}
$$

where

$$
\varrho_{i, j}^{(x, y)}=\frac{\Delta x_{i} \Delta y_{j}}{\Delta x_{i}^{2}+\Delta y_{j}^{2}}, \quad \mu_{i, j}^{(x, y)}=1+\left[\frac{\Delta x_{i} \Delta y_{j}}{12 D_{i, j}}\right]\left(\Sigma_{a}\right)_{i, j} \varrho_{i, j}^{(x, y)}
$$

and one-dimensional stencils that govern the transverse averaged flux

$$
\begin{equation*}
2 \frac{D_{i, j}}{\Delta x_{i}} \phi_{i-\frac{1}{2}, j}-6 \frac{D_{i, j}}{\Delta x_{i}} \phi_{i, j}+4\left[\frac{D_{i, j}}{\Delta x_{i}}+\frac{D_{i+1, j}}{\Delta x_{i+1}}\right] \phi_{i+\frac{1}{2}, j}-6 \frac{D_{i+1, j}}{\Delta x_{i+1}} \phi_{i+1, j}+2 \frac{D_{i+1, j}}{\Delta x_{i+1}} \phi_{i+\frac{3}{2}, j}=0 \tag{3.30}
\end{equation*}
$$

with the rotated analogue defining $\phi_{i, j+\frac{1}{2}}$ [Figure 3.3].
Derivation of the Flux Formulation. Utilizing the expressions for the one-sided currents given in (3.23) we obtain,

$$
\frac{1}{\Delta x_{i}}\left[J_{i+\frac{1}{2}, j}^{-}-J_{i-\frac{1}{2}, j}^{+}\right]=6 \frac{D_{i, j}}{\Delta x_{i}^{2}}\left\{-\phi_{i-\frac{1}{2}, j}+2 \phi_{i, j}-\phi_{i+\frac{1}{2}, j}\right\} .
$$

along with an analogous expression for the $y$-oriented leakage. Substitution into the balance equation (3.8) immediately gives equation (3.29). The one-dimensional stencils that govern the transverse averaged flux [i.e. equation (3.30)] are obtained directly from imposing the continuity of normal currents (3.24).


Figure 3.3: Stencil weights of the flux formulation with $D=1$ and $\Sigma_{a}=0$.

## Edge Based Flux Formulation

Definition 3.4.5. The edge-based flux formulation of the lowest-order NEM [Definition 3.4.2] is defined by the following 7 point nearest neighbour flux based stencil,

$$
\begin{align*}
& \vartheta_{i, j}^{(x, y)} \phi_{i-\frac{1}{2}, j}+\left[\zeta_{i, j}^{(x, y)}+\zeta_{i+1, j}^{(x, y)}\right] \phi_{i+\frac{1}{2}, j}+\vartheta_{i+1, j}^{(x, y)} \phi_{i+\frac{3}{2}, j} \\
&+3 D_{i, j} \frac{\varrho_{i, j}^{(x, y)}}{\mu_{i, j}^{(x, y)}}\left\{\left(\phi_{i+\frac{1}{2}, j}+\phi_{i-\frac{1}{2}, j}\right)-\left(\phi_{i, j+\frac{1}{2}}+\phi_{i, j-\frac{1}{2}}\right)\right\} \\
&+3 D_{i+1, j} \frac{\varrho_{i+1, j}^{(x, y)}}{\mu_{i+1, j}^{(x, y)}}\left\{\left(\phi_{i+\frac{3}{2}, j}+\phi_{i+\frac{1}{2}, j}\right)-\left(\phi_{i+1, j+\frac{1}{2}}+\phi_{i+1, j-\frac{1}{2}}\right)\right\}  \tag{3.31}\\
&=\frac{\Delta y_{j}^{2}}{2}\left\{\frac{\varrho_{i, j}^{(x, y)}}{\mu_{i, j}^{(x, y)}} Q_{i, j}+\frac{\varrho_{i+1, j}^{(x, y)}}{\mu_{i, j}^{(x, y)}} Q_{i+1, j}\right\}
\end{align*}
$$

where

$$
\begin{align*}
\vartheta_{i, j}^{(x, y)} & =D_{i, j} \frac{\Delta y_{j}}{\Delta x_{i}}\left[3\left(1-\left[\mu_{i, j}^{(x, y)}\right]^{-1}\right)-1\right]  \tag{3.32a}\\
\zeta_{i, j}^{(x, y)} & =D_{i, j} \frac{\Delta y_{j}}{\Delta x_{i}}\left[3\left(1-\left[\mu_{i, j}^{(x, y)}\right]^{-1}\right)+1\right] \tag{3.32~b}
\end{align*}
$$

and $\mu_{i, j}^{(x, y)}, \varrho_{i, j}^{(x, y)}$ are given in Definition 3.4.4. The $y$-oriented analogue governs $\phi_{i, j+\frac{1}{2}}$.

Derivation of the Edge Based Flux Formulation. Solving (3.29) for the node average,

$$
\phi_{i, j}=\frac{1}{2} \frac{\varrho_{i, j}^{(x, y)}}{\mu_{i, j}^{(x, y)}}\left\{\frac{\Delta y_{j}}{\Delta x_{i}}\left(\phi_{i+\frac{1}{2}, j}+\phi_{i-\frac{1}{2}, j}\right)+\frac{\Delta x_{i}}{\Delta y_{j}}\left(\phi_{i, j+\frac{1}{2}}+\phi_{i, j-\frac{1}{2}}\right)+\frac{\Delta x_{i} \Delta y_{j}}{6 D_{i, j}} Q_{i, j}\right\}
$$

Hence substituting into (3.23a) and multiplying through by $\Delta y_{j}$ we obtain,

$$
\begin{align*}
\Delta y_{j} J_{i+\frac{1}{2}, j}^{-}= & -D_{i, j} \frac{\Delta y_{j}}{\Delta x_{i}}\left\{\left[2-3 \frac{\Delta y_{j}}{\Delta x_{i}} \frac{\varrho_{i, j}^{(x, y)}}{\mu_{i, j}^{(x, y)}}\right] \phi_{i-\frac{1}{2}, j}+\left[4-3 \frac{\Delta y_{j}}{\Delta x_{i}} \frac{\varrho_{i, j}^{(x, y)}}{\mu_{i, j}^{(x, y)}}\right] \phi_{i+\frac{1}{2}, j}\right\}  \tag{3.33}\\
& +3 D_{i, j} \frac{\varrho_{i, j}^{(x, y)}}{\mu_{i, j}^{(x, y)}}\left\{\phi_{i, j+\frac{1}{2}}+\phi_{i, j-\frac{1}{2}}\right\}-\frac{\Delta y_{j}^{2}}{2} \frac{\varrho_{i, j}^{(x, y)}}{\mu_{i, j}^{(x, y)}} Q_{i, j}
\end{align*}
$$

Similarly substitution into (3.23b) yields,

$$
\begin{align*}
\Delta y_{j} J_{i-\frac{1}{2}, j}^{+}= & D_{i, j} \frac{\Delta y_{j}}{\Delta x_{i}}\left\{\left[4-3 \frac{\Delta y_{j}}{\Delta x_{i}} \frac{\varrho_{i, j}^{(x, y)}}{\mu_{i, j}^{(x, y)}}\right] \phi_{i-\frac{1}{2}, j}\left[2-3 \frac{\Delta y_{j}}{\Delta x_{i}} \frac{\varrho_{i, j}^{(x, y)}}{\mu_{i, j}^{(x, y)}}\right] \phi_{i+\frac{1}{2}, j}\right\}  \tag{3.34}\\
& -3 D_{i, j} \frac{\varrho_{i, j}^{(x, y)}}{\mu_{i, j}^{(x, y)}}\left\{\phi_{i, j+\frac{1}{2}}+\phi_{i, j-\frac{1}{2}}\right\}+\frac{\Delta y_{j}^{2}}{2} \frac{\varrho_{i, j}^{(x, y)}}{\mu_{i, j}^{(x, y)}} Q_{i, j}
\end{align*}
$$

Thus we consider manipulating the coefficients to deduce,

$$
\begin{equation*}
D_{i, j} \frac{\Delta y_{j}}{\Delta x_{i}}\left[2-3 \frac{\Delta y_{j}}{\Delta x_{i}} \frac{\varrho_{i, j}^{(x, y)}}{\mu_{i, j}^{(x, y)}}\right]=D_{i, j} \frac{\Delta y_{j}}{\Delta x_{i}}\left[2-3\left[\mu_{i, j}^{(x, y)}\right]^{-1}\right]+3 D_{i, j} \frac{\varrho_{i, j}^{(x, y)}}{\mu_{i, j}^{(x, y)}} \tag{3.35}
\end{equation*}
$$

but $2-3\left[\mu_{i, j}^{(x, y)}\right]^{-1}=3\left(1-\left[\mu_{i, j}^{(x, y)}\right]^{-1}\right)-1$ and hence enforcing continuity of the normal current and utilizing the definitions of $\zeta_{i, j}^{(x, y)}$ and $\vartheta_{i, j}^{(x, y)}$ given in (3.32), gives (3.60).

### 3.5 Analytic Nodal Methods - NIM

The second distinct class of nodal discretizations utilize analytic information that is obtained from the solution of the cell-based homogenized diffusion equation. Although the analytic solution of the two-dimensional cell-based homogenized diffusion equation has been considered [58], typically analytic nodal methods are based on a particular treatment


Figure 3.4: The edge-based flux formulation with $D=1$ and $\Sigma_{a}=0$.
of the transverse integrated ODEs (3.16). Methods of this type include the Nodal Integration Method (NIM) [32, 6] and the Nodal Green's Function Method (NGFM) [54]. In the lowest-order case these discretizations are in fact equivalent [6], thus we focus on the more popular form, NIM. In contrast to the polynomial methods, the nodal integration method defines the transverse averaged flux [i.e. $\phi_{j}(x)$ and $\phi_{i}(y)$ ] as the analytic solution of the corresponding transverse averaged ODE (3.16). This is facilitated by specifying cell-based boundary conditions in conjunction with an approximation of the transverse edge currents [i.e. transverse leakage (3.14)] and the transverse averaged source. In the following we consider the assumption that these currents are approximately constant along their respective edges, and thus derive the constant-constant NIM discretization of the one-group diffusion equation (2.12).

### 3.5.1 The Nodal Integration Method

Posing the cell-based transverse integrated ODEs (3.16) which govern the transverse average flux is the first step in the NIM discretization. To this end we must specify the cell-based boundary conditions as well as the functional form of both the transverse edge currents and the transverse integrated source.

The definition of the edge average (3.2a) naturally yields Dirichlet boundary conditions for each cell [6],

$$
\begin{array}{ll}
\phi_{j}\left(x_{i-\frac{1}{2}}\right)=\phi_{i-\frac{1}{2}, j} & \phi_{j}\left(x_{i+\frac{1}{2}}\right)=\phi_{i+\frac{1}{2}, j} \\
\phi_{i}\left(y_{j-\frac{1}{2}}\right)=\phi_{i, j-\frac{1}{2}} & \phi_{i}\left(y_{j+\frac{1}{2}}\right)=\phi_{i, j+\frac{1}{2}} \tag{3.36b}
\end{array}
$$

although other boundary conditions have been considered. In particular, Neumann boundary conditions were considered by Greenman et al. [33], but in the conservative case suffer the obvious problem of nonuniqueness. Fischer and Finnemann [32] employed partial current boundary conditions as these naturally yield an interface current formulation of the discrete equations. In fact, this discretization is equivalent to the flux formulation that arises naturally with the Dirichlet boundary conditions (3.36). Moreover, we are specifically interested in the flux formulation, and thus restrict our discussion to the Dirichlet case.

We assume that the edge-based normal currents are approximately constant,

$$
J_{j \pm \frac{1}{2}}^{\mp}(x) \approx J_{i, j \pm \frac{1}{2}}^{\mp}, \quad J_{i \pm \frac{1}{2}}^{\mp}(y) \approx J_{i \pm \frac{1}{2}, j}^{\mp}
$$

or equivalently we assume a constant transverse leakage. Various higher order approximations are possible and were discussed in Section 3.4.1

Finally, the exact functional form of the source term $Q(x, y)$ may be unknown, so we employ a piecewise constant mean value definition,

$$
Q_{i, j}=\frac{1}{\Delta x_{i}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} Q_{j}(x) d x=\frac{1}{\Delta y_{j}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} Q_{i}(y) d y
$$

Hence, the transverse integrated ODEs (3.16) become,

$$
\begin{align*}
-\frac{\partial}{\partial x}\left\{D_{i, j} \frac{\partial}{\partial x} \phi_{j}(x)\right\}+\left(\Sigma_{a}\right)_{i, j} \phi_{j}(x) & =-\frac{1}{\Delta y_{j}}\left\{J_{i, j+\frac{1}{2}}^{-}-J_{i, j-\frac{1}{2}}^{+}\right\}+Q_{i, j}  \tag{3.37a}\\
-\frac{\partial}{\partial y}\left\{D_{i, j} \frac{\partial}{\partial y} \phi_{i}(y)\right\}+\left(\Sigma_{a}\right)_{i, j} \phi_{i}(y) & =-\frac{1}{\Delta x_{i}}\left\{J_{i+\frac{1}{2}, j}^{-}-J_{i-\frac{1}{2}, j}^{+}\right\}+Q_{i, j} \tag{3.37b}
\end{align*}
$$

where $x \in\left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right]$ and $y \in\left[y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}\right]$ respectively, subject to the Dirichlet BC 's given in (3.36). These ODEs are readily solved and thus, with expressions for $\phi_{j}(x)$ and $\phi_{i}(y)$ in hand, we outline the NIM discretization procedure.

First, we note that two independent definitions of the cell average are possible,

$$
\begin{equation*}
\phi_{i, j}=\frac{1}{\Delta x_{i}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \phi_{j}(x) d x=\frac{1}{\Delta y_{j}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \phi_{i}(y) d y \tag{3.38}
\end{equation*}
$$

yielding 2 equations per node. Furthermore, utilizing the transverse averaged Fick's Law (3.15) we obtain expressions for the edge averaged normal currents,

$$
\begin{equation*}
J_{i \pm \frac{1}{2}, j}^{\mp}=\lim _{x \rightarrow x_{i \pm \frac{1}{2}}^{\mp}} J_{j}(x) \quad J_{i, j \pm \frac{1}{2}}^{\mp}=\lim _{y \rightarrow y_{j \pm \frac{1}{2}}^{\mp}} J_{i}(y) \tag{3.39}
\end{equation*}
$$

These 6 equations constitute only 5 linearly independent equations per node as the cell balance equation arises from two linear combinations. Imposing the continuity of the normal current in the weak sense of the transverse average yields an equation for each interior edge [i.e. $(L-1) M+L(M-1)$ equations] while the boundary conditions give rise to $2 L+2 M$ discrete boundary equations. Thus we have $7 L M+L+M$ equations in as many unknowns.

### 3.5.2 The constant-constant NIM for Conservative Diffusion

Theorem 3.1. The constant-constant NIM discretization of the conservative one-group diffusion equation (2.12) on an $L \times M$ tensor product mesh is equivalent to the lowestorder NEM discretization given in Definition 3.4.2.

Proof. We begin by solving the transverse integrated ODEs in a conservative medium. The solution of (3.37a) with $\left(\Sigma_{a}\right)_{i, j}=0$ may be expressed as,

$$
\begin{align*}
\phi_{j}(x)=\frac{1}{\Delta x_{i}} & \left(x_{i+\frac{1}{2}}-x\right) \phi_{i-\frac{1}{2}, j}+\frac{1}{\Delta x_{i}}\left(x-x_{i-\frac{1}{2}}\right) \phi_{i+\frac{1}{2}, j} \\
& +\frac{1}{2 D_{i, j}}\left\{Q_{i, j}-\frac{1}{\Delta y_{j}}\left[J_{i, j+\frac{1}{2}}-J_{i, j-\frac{1}{2}}\right]\right\}\left(x_{i+\frac{1}{2}}-x\right)\left(x-x_{i-\frac{1}{2}}\right) \tag{3.40}
\end{align*}
$$

hence substitution of $\phi_{j}(x)$ into Fick's Law (3.15a) gives

$$
\begin{equation*}
J_{i \pm \frac{1}{2}, j}^{\mp}=-\frac{D_{i, j}}{\Delta x_{i}}\left(\phi_{i+\frac{1}{2}, j}-\phi_{i-\frac{1}{2}, j}\right) \pm \frac{\Delta x_{i}}{2}\left\{Q_{i, j}-\frac{1}{\Delta y_{j}}\left[J_{i, j+\frac{1}{2}}^{-}-J_{i, j-\frac{1}{2}}^{+}\right]\right\} \tag{3.41a}
\end{equation*}
$$

Similarly, the analogous expression for $\phi_{i}(y)$ gives,

$$
\begin{equation*}
J_{i, j \pm \frac{1}{2}}^{\mp}=-\frac{D_{i, j}}{\Delta y_{j}}\left(\phi_{i, j+\frac{1}{2}}-\phi_{i, j-\frac{1}{2}}\right) \pm \frac{\Delta y_{j}}{2}\left\{Q_{i, j}-\frac{1}{\Delta x_{i}}\left[J_{i+\frac{1}{2}, j}^{-}-J_{i-\frac{1}{2}, j}^{+}\right]\right\} \tag{3.41b}
\end{equation*}
$$

Evaluating the consistency expressions of the node average,

$$
\begin{align*}
\phi_{i, j} & =\frac{1}{\Delta x_{i}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \phi_{j}(x) d x \\
& =\frac{1}{2}\left(\phi_{i+\frac{1}{2}, j}+\phi_{i-\frac{1}{2}, j}\right)+\frac{\Delta x_{i}^{2}}{12 D_{i, j}}\left\{Q_{i, j}-\frac{1}{\Delta y_{j}}\left[J_{i, j+\frac{1}{2}}-J_{i, j-\frac{1}{2}}\right]\right\} \tag{3.41c}
\end{align*}
$$

and,

$$
\begin{align*}
\phi_{i, j} & =\frac{1}{\Delta y_{j}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \phi_{i}(y) d y \\
& =\frac{1}{2}\left(\phi_{i, j+\frac{1}{2}}+\phi_{i, j-\frac{1}{2}}\right)+\frac{\Delta y_{j}^{2}}{12 D_{i, j}}\left\{Q_{i, j}-\frac{1}{\Delta x_{i}}\left[J_{i+\frac{1}{2}, j}-J_{i-\frac{1}{2}, j}\right]\right\} \tag{3.41d}
\end{align*}
$$

These equations comprise the cell-based NIM discretization and we now proceed to prove their equivalence to the lowest-order NEM.

Subtracting $J_{i-\frac{1}{2}, j}^{+}$from $J_{i+\frac{1}{2}, j}^{-}$, where both expressions are given in (3.41a), and dividing through by $\Delta x_{i}$ we obtain,

$$
\begin{equation*}
\frac{1}{\Delta x_{j}}\left[J_{i+\frac{1}{2}, j}^{-}-J_{i-\frac{1}{2}, j}^{+}\right]=Q_{i, j}-\frac{1}{\Delta y_{j}}\left[J_{i, j+\frac{1}{2}}^{-}-J_{i, j-\frac{1}{2}}^{+}\right] \tag{3.42}
\end{equation*}
$$

which is the balance equation (3.8), and the first equation in Definition 3.4.2. Utilizing the balance equation (3.42) in the expression for the node average (3.41c) we obtain,

$$
\begin{equation*}
J_{i+\frac{1}{2}, j}^{-}-J_{i-\frac{1}{2}, j}^{+}=-6 \frac{D_{i, j}}{\Delta x_{i}}\left(\phi_{i+\frac{1}{2}, j}-2 \phi_{i, j}+\phi_{i-\frac{1}{2}, j}\right) \tag{3.43}
\end{equation*}
$$

Finally adding $J_{i-\frac{1}{2}, j}^{+}$and $J_{i+\frac{1}{2}, j}^{-}$[equation (3.41a)] yields,

$$
\begin{equation*}
J_{i+\frac{1}{2}, j}+J_{i-\frac{1}{2}, j}=-2 \frac{D_{i, j}}{\Delta x_{i}}\left(\phi_{i+\frac{1}{2}, j}-\phi_{i-\frac{1}{2}, j}\right) \tag{3.44}
\end{equation*}
$$

Hence adding (3.43) to (3.44) gives (3.23a) of Definition 3.4.2, while subtracting (3.43) from (3.44) we obtain (3.23b). Analogous expressions follow in $y$.

Thus we have established the equivalence of the interior equations, and the continuity of the normal current is enforced precisely as in NEM. The equivalence of the discrete boundary conditions, follows as a consequence of the cell-based treatment that is employed in both methods.

One implication of Theorem 3.1 is that in the conservative case, the flux formulations of the constant-constant NIM are obtained as the conservative limit of those given in Definition 3.4.4 and Definition 3.4.2. Specifically, with $\left(\Sigma_{a}\right)_{i, j} \rightarrow 0$ we have $\mu_{i, j}^{(x, y)}=1$.

### 3.5.3 The constant-constant NIM for Nonconservative Diffusion

Definition 3.5.1. The constant-constant NIM discretization of the one-group diffusion equation (2.12) is comprised of the cell balance equation (3.8) and the following cell-based normal currents,

$$
\begin{align*}
& J_{i+\frac{1}{2}, j}^{-}=-\frac{D_{i, j}}{\Delta x_{i}}\left\{\left(\alpha_{i, j}^{(x)}-\beta_{i, j}^{(x)}\right) \phi_{i-\frac{1}{2}, j}-2 \alpha_{i, j}^{(x)} \phi_{i, j}+\left(\alpha_{i, j}^{(x)}+\beta_{i, j}^{(x)}\right) \phi_{i+\frac{1}{2}, j}\right\}  \tag{3.45a}\\
& J_{i-\frac{1}{2}, j}^{+}=-\frac{D_{i, j}}{\Delta x_{i}}\left\{-\left(\alpha_{i, j}^{(x)}+\beta_{i, j}^{(x)}\right) \phi_{i-\frac{1}{2}, j}+2 \alpha_{i, j}^{(x)} \phi_{i, j}-\left(\alpha_{i, j}^{(x)}-\beta_{i, j}^{(x)}\right) \phi_{i+\frac{1}{2}, j}\right\} \tag{3.45b}
\end{align*}
$$

where we have defined,

$$
\begin{equation*}
\alpha_{i, j}^{(x)}=\frac{\left[\bar{\lambda}_{i, j}^{(x)}\right]^{2} \delta_{i, j}^{(x)}}{1-\delta_{i, j}^{(x)}}, \quad \beta_{i, j}^{(x)}=\frac{1}{\delta_{i, j}^{(x)}}, \quad \delta_{i, j}^{(x)}=\frac{1}{\bar{\lambda}_{i, j}^{(x)}} \tanh \left(\bar{\lambda}_{i, j}^{(x)}\right) \tag{3.46}
\end{equation*}
$$

with,

$$
\begin{equation*}
\lambda_{i, j}=\sqrt{\frac{\left(\Sigma_{a}\right)_{i, j}}{D_{i, j}}}, \quad \lambda_{i, j}^{(x)}=\lambda_{i, j} \Delta x_{i}, \quad \bar{\lambda}_{i, j}^{(x)}=\frac{1}{2} \lambda_{i, j}^{(x)} \tag{3.47}
\end{equation*}
$$

and where analogous expressions define $J_{i, j \pm \frac{1}{2}}^{ \pm}$. The continuity of the normal current is imposed across all interior edges, in the weak sense of the transverse average,

$$
\begin{equation*}
J_{i+\frac{1}{2}, j}^{-}=J_{i+\frac{1}{2}, j}^{+}, \quad J_{i, j+\frac{1}{2}}^{-}=J_{i, j+\frac{1}{2}}^{+} \tag{3.48}
\end{equation*}
$$

in conjunction with the discrete boundary equations to complete the discretization.

Deriving the constant-constant NIM Discretization. We begin by solving the transverse integrated ODEs. For example, consider equation (3.37a) whose solution may be written

$$
\begin{align*}
& \phi_{j}(x)=\frac{\sinh \left(\lambda_{i, j}\left[x_{i+\frac{1}{2}}-x\right]\right)}{\sinh \left(\lambda_{i, j}^{(x)}\right)} \phi_{i-\frac{1}{2}, j}+\frac{\sinh \left(\lambda_{i, j}\left[x-x_{i-\frac{1}{2}}\right]\right)}{\sinh \left(\lambda_{i, j}^{(x)}\right)} \phi_{i+\frac{1}{2}, j} \\
& +\frac{1}{\left(\Sigma_{a}\right)_{i, j}}\left[1-\frac{\cosh \left(\lambda_{i j}\left[x-x_{i}\right]\right)}{\cosh \left(\bar{\lambda}_{i, j}^{(x)}\right)}\right]\left\{Q_{i, j}-\mathcal{L}_{i, j}^{(y, 0)}\right\} \tag{3.49}
\end{align*}
$$

where $\lambda_{i, j}$ and $\bar{\lambda}_{i, j}^{(x)}$ are defined in (3.47). Applying the transverse integrated form of Fick's Law (3.15a) we obtain,

$$
\begin{align*}
J_{i+\frac{1}{2}, j}^{-} & =-\frac{D_{i, j} \lambda_{i, j}}{\sinh \left(\lambda_{i, j}^{(x)}\right)}\left\{\cosh \left(\lambda_{i, j}^{(x)}\right) \phi_{i+\frac{1}{2}, j}-\phi_{i-\frac{1}{2}, j}\right\}+\frac{\Delta x_{i}}{2} \delta_{i, j}^{(x)}\left\{Q_{i, j}-\mathcal{L}_{i, j}^{(y, 0)}\right\}  \tag{3.50a}\\
J_{i-\frac{1}{2}, j}^{+} & =-\frac{D_{i, j} \lambda_{i, j}}{\sinh \left(\lambda_{i, j}^{(x)}\right)}\left\{\phi_{i+\frac{1}{2}, j}-\cosh \left(\lambda_{i, j}^{(x)}\right) \phi_{i-\frac{1}{2}, j}\right\}-\frac{\Delta x_{i}}{2} \delta_{i, j}^{(x)}\left\{Q_{i, j}-\mathcal{L}_{i, j}^{(y, 0)}\right\} \tag{3.50b}
\end{align*}
$$

We continue by evaluating the consistency expressions of the node average,

$$
\begin{align*}
\phi_{i, j} & =\frac{1}{\Delta x_{i}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \phi_{j}(x) d x \\
& =\frac{1}{2} \delta_{i, j}^{(x)}\left\{\phi_{i+\frac{1}{2}, j}+\phi_{i-\frac{1}{2}, j}\right\}+\frac{1}{\left(\Sigma_{a}\right)_{i, j}}\left[1-\delta_{i, j}^{(x)}\right]\left\{Q_{i, j}-\mathcal{L}_{i, j}^{(y, 0)}\right\} \tag{3.51a}
\end{align*}
$$

and,

$$
\begin{align*}
\phi_{i, j} & =\frac{1}{\Delta y_{j}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \phi_{i}(y) d y \\
& =\frac{1}{2} \delta_{i, j}^{(y)}\left\{\phi_{i, j+\frac{1}{2}}+\phi_{i, j-\frac{1}{2}}\right\}+\frac{1}{\left(\Sigma_{a}\right)_{i, j}}\left[1-\delta_{i, j}^{(y)}\right]\left\{Q_{i, j}-\mathcal{L}_{i, j}^{(x, 0)}\right\} \tag{3.51b}
\end{align*}
$$

These equations comprise the cell-based NIM discretization and we now develop the expressions given in Definition 3.5.1.

We begin by subtracting (3.50b) from (3.50a) to obtain,

$$
\begin{equation*}
J_{i+\frac{1}{2}, j}^{-}-J_{i-\frac{1}{2}, j}^{+}=-D_{i, j} \lambda_{i, j} \tanh \left(\bar{\lambda}_{i, j}^{(x)}\right)\left\{\phi_{i+\frac{1}{2}, j}+\phi_{i-\frac{1}{2} . j}\right\}+\Delta x_{i} \delta_{i, j}^{(x)}\left\{Q_{i, j}-\mathcal{L}_{i, j}^{(y, 0)}\right\} \tag{3.52}
\end{equation*}
$$

From (3.51a) we obtain an expression for $\left(\phi_{i+\frac{1}{2}, j}+\phi_{i-\frac{1}{2}, j}\right)$ which upon substitution into (3.52) yields, after a little algebra, the balance equation,

$$
\begin{equation*}
\mathcal{L}_{i, j}^{(x, 0)}+\mathcal{L}_{i, j}^{(y, 0)}+\left(\Sigma_{a}\right)_{i, j} \phi_{i, j}=Q_{i, j} \tag{3.53}
\end{equation*}
$$

Now the balance equation (3.53) gives an expression for $\left\{Q_{i, j}-\mathcal{L}_{i, j}^{(y, 0)}\right\}$ hence substitution into the node average (3.51a) gives,

$$
\begin{equation*}
J_{i+\frac{1}{2}, j}^{-}-J_{i-\frac{1}{2}, j}^{+}=-\frac{2 D_{i, j}}{\Delta x_{i}} \alpha_{i, j}^{(x)}\left\{\phi_{i-\frac{1}{2}, j}-2 \phi_{i, j}+\phi_{i+\frac{1}{2}, j}\right\} \tag{3.54}
\end{equation*}
$$

where $\alpha_{i, j}^{(x)}$ is defined in (3.46). Adding (3.50b) to (3.50a) we find,

$$
\begin{equation*}
J_{i+\frac{1}{2}, j}^{+}+J_{i-\frac{1}{2}, j}^{+}=-2 \frac{D_{i, j}}{\Delta x_{i}} \beta_{i, j}^{(x)}\left\{\phi_{i+\frac{1}{2}, j}-\phi_{i-\frac{1}{2}, j}\right\} \tag{3.55}
\end{equation*}
$$

with $\beta_{i, j}^{(x)}$ defined in (3.46). Now we simply add (3.54) to (3.55) to obtain (3.45a). Conversely subtracting (3.54) from (3.55) yields (3.45b).

## The Flux Formulation

Definition 3.5.2. The flux formulation of the constant-constant NIM discretization [Definition 3.5.1] is defined by the following flux based balance equation

$$
\begin{equation*}
-2 \frac{D_{i, j}}{\Delta x_{i}^{2}} \alpha_{i, j}^{(x)}\left\{\phi_{i-\frac{1}{2}, j}+\phi_{i+\frac{1}{2}, j}\right\}+\frac{4 D_{i, j}}{\Delta x_{i} \Delta y_{j}} \gamma_{i, j}^{(x, y)} \phi_{i, j}-2 \frac{D_{i, j}}{\Delta y_{j}^{2}} \alpha_{i, j}^{(y)}\left\{\phi_{i, j-\frac{1}{2}}+\phi_{i, j+\frac{1}{2}}\right\}=Q_{i, j} \tag{3.56}
\end{equation*}
$$

where we have defined,

$$
\begin{equation*}
\gamma_{i, j}^{(x, y)}=\frac{\bar{\lambda}_{i, j}^{(x)} \bar{\lambda}_{i, j}^{(y)}\left[1-\delta_{i, j}^{(x)} \delta_{i, j}^{(y)}\right]}{\left[1-\delta_{i, j}^{(x)}\right]\left[1-\delta_{i, j}^{(y)}\right]} \tag{3.57}
\end{equation*}
$$

and one-dimensional stencils of the form,

$$
\begin{align*}
& \frac{D_{i}}{\Delta x_{i}}\left(\alpha_{i, j}^{(x)}-\beta_{i, j}^{(x)}\right) \phi_{i-\frac{1}{2}, j}-2 \frac{D_{i, j}}{\Delta x_{i}} \alpha_{i, j}^{(x)} \phi_{i, j} \\
&+ {\left[\frac { D _ { i , j } } { \Delta x _ { i } } \left(\alpha_{i, j}^{(x)}\right.\right.}  \tag{3.58}\\
&\left.\left.+\beta_{i, j}^{(x)}\right)+\frac{D_{i+1, j}}{\Delta x_{i+1}}\left(\alpha_{i+1, j}^{(x)}+\beta_{i+1, j}^{(x)}\right)\right] \phi_{i+\frac{1}{2}, j} \\
&-2 \frac{D_{i+1, j}}{\Delta x_{i+1}} \alpha_{i+1, j}^{(x)} \phi_{i+1, j}+\frac{D_{i+1, j}}{\Delta x_{i+1}}\left(\alpha_{i+1, j}^{(x)}-\beta_{i+1, j}^{(x)}\right) \phi_{i+\frac{3}{2}, j}=0
\end{align*}
$$

with a similar expression governing $\phi_{i, j+\frac{1}{2}}$.
Derivation of the Flux Formulation. The transverse leakage $\mathcal{L}_{i, j}^{(x, 0)}$ is readily obtained from (3.54) and an analogous expression for $\mathcal{L}_{i, j}^{(y, 0)}$ may be similarly derived. Substitution of these expressions into the balance equation (3.8) gives (3.56), although the coefficient of $\phi_{i, j}$ requires simplification. Specifically we have,

$$
\begin{align*}
2 D_{i, j}\left\{\frac{\alpha_{i, j}^{(x)}}{\Delta x_{i}^{2}}+\frac{\alpha_{i, j}^{(y)}}{\Delta y_{j}^{2}}\right\}+\left(\Sigma_{a}\right)_{i, j} & =D_{i, j} \lambda_{i, j}^{2}\left\{\frac{\delta_{i, j}^{(x)}}{\left[1-\delta_{i, j}^{(x)}\right]}+\frac{\delta_{i, j}^{(y)}}{\left[1-\delta_{i, j}^{(y)}\right]}+1\right\} \\
& =D_{i, j} \lambda_{i, j}^{2}\left\{\frac{1-\delta_{i, j}^{(x)} \delta_{i, j}^{(y)}}{\left[1-\delta_{i, j}^{(x)}\right]\left[1-\delta_{i, j}^{(y)}\right]}\right\} \tag{3.59}
\end{align*}
$$

thus the definition of $\dot{\gamma}_{i, j}^{(x, y)}$. Imposing the continuity of the normal current with (3.50a) and (3.50b) gives (3.58)

## Edge Based Flux Formulation

Definition 3.5.3. The edge-based flux formulation of the constant-constant NIM given in Definition 3.5.1 is defined by the following 7 -point nearest-neighbour flux based stencil,

$$
\begin{align*}
\hat{\vartheta}_{i, j}^{(x, y)} \phi_{i-\frac{1}{2}, j} & +\left[\hat{\zeta}_{i, j}^{(x, y)}+\hat{\zeta}_{i+1, j}^{(x, y)}\right] \phi_{i+\frac{1}{2}, j}+\hat{\vartheta}_{i+1, j}^{(x, y)} \phi_{i+\frac{3}{2}, j} \\
& +D_{i, j} \frac{\alpha_{i, j}^{(x)} \alpha_{i, j}^{(y)}}{\gamma_{i, j}^{(x, y)}}\left\{\delta_{i, j}^{(x)}\left(\phi_{i+\frac{1}{2}, j}+\phi_{i-\frac{1}{2}, j}\right)-\left(\phi_{i, j+\frac{1}{2}}+\phi_{i, j-\frac{1}{2}}\right)\right\} \\
& +D_{i+1, j} \frac{\alpha_{i+1, j}^{(x)} \alpha_{i+1, j}^{(y)}}{\gamma_{i+1, j}^{(x, y)}}\left\{\delta_{i+1, j}^{(x)}\left(\phi_{i+\frac{3}{2}, j}+\phi_{i+\frac{1}{2}, j}\right)-\left(\phi_{i+1, j+\frac{1}{2}}+\phi_{i+1, j-\frac{1}{2}}\right)\right\}  \tag{3.60}\\
& =\frac{\Delta y_{j}^{2}}{2}\left\{\frac{\alpha_{i, j}^{(x)}}{\gamma_{i, j}^{(x, y)}} Q_{i, j}+\frac{\alpha_{i+1, j}^{(x)}}{\gamma_{i, j}^{(x, y)}} Q_{i+1, j}\right\}
\end{align*}
$$

where

$$
\begin{align*}
& \hat{\vartheta}_{i, j}^{(x, y)}=D_{i, j} \frac{\Delta y_{j}}{\Delta x_{i}}\left(\left[1-\delta_{i, j}^{(x)}\right] \alpha_{i, j}^{(x)}-\beta_{i, j}^{(x)}\right)  \tag{3.61a}\\
& \hat{\zeta}_{i, j}^{(x, y)}=D_{i, j} \frac{\Delta y_{j}}{\Delta x_{i}}\left(\left[1-\delta_{i, j}^{(x)}\right] \alpha_{i, j}^{(x)}+\beta_{i, j}^{(x)}\right) \tag{3.61b}
\end{align*}
$$

with $\alpha_{i, j}^{(x)}, \beta_{i, j}^{(x)}$ and $\gamma_{i, j}^{(x)}$ given in Definition 3.5.2. The $y$-oriented analogue governs $\phi_{i, j+\frac{1}{2}}$.

Derivation of the Edge Based Flux Formulation. Solving the flux based balance equation (3.56) for the node average we obtain,

$$
\phi_{i, j}=\frac{1}{2 \gamma_{i, j}^{(x, y) .}}\left\{\alpha_{i, j}^{(x)} \frac{\Delta y_{j}}{\Delta x_{i}}\left(\phi_{i-\frac{1}{2}, j}+\phi_{i+\frac{1}{2}, j}\right)+\alpha_{i, j}^{(y)} \frac{\Delta x_{i}}{\Delta y_{j}}\left(\phi_{i, j-\frac{1}{2}}+\phi_{i, j+\frac{1}{2}}\right)+\frac{\Delta x_{i} \Delta y_{j}}{2 D_{i, j}} Q_{i, j}\right\}
$$

Hence substituting this expression into (3.45a) and multiplying through by $\Delta y_{j}$ gives,

$$
\begin{align*}
\Delta y_{j} J_{i+\frac{1}{2}, j}^{-}= & -D_{i, j} \frac{\Delta y_{j}}{\Delta x_{i}}\left\{\left(\alpha_{i, j}^{(x)}-\beta_{i, j}^{(x)}\right)-\frac{\Delta y_{j}}{\Delta x_{i}} \frac{\left[\alpha_{i, j}^{(x)}\right]^{2}}{\gamma_{i, j}^{(x, y)}}\right\} \phi_{i-\frac{1}{2}, j} \\
& -D_{i, j} \frac{\Delta y_{j}}{\Delta x_{i}}\left\{\left(\alpha_{i, j}^{(x)}+\beta_{i, j}^{(x)}\right)-\frac{\Delta y_{j}}{\Delta x_{i}} \frac{\left[\alpha_{i, j}^{(x)}\right]^{2}}{\left.\gamma_{i, j}^{(x, y)}\right\} \phi_{i+\frac{1}{2}, j}}\right.  \tag{3.62}\\
& +D_{i, j} \frac{\alpha_{i, j}^{(x)} \alpha_{i, j}^{(y)}}{\gamma_{i, j}^{(x, y)}}\left(\phi_{i, j+\frac{1}{2}}+\phi_{i, j-\frac{1}{2}}\right)+\frac{\Delta y_{j}^{2}}{2} \frac{\alpha_{i, j}^{(x)}}{\gamma_{i, j}^{(x, y)}} Q_{i, j}
\end{align*}
$$

Similarly from (3.45b) we obtain,

$$
\begin{align*}
\Delta y_{j} J_{i+\frac{1}{2}, j}^{-} & =D_{i, j} \frac{\Delta y_{j}}{\Delta x_{i}}\left\{\left(\alpha_{i, j}^{(x)}+\beta_{i, j}^{(x)}\right)-\frac{\Delta y_{j}}{\Delta x_{i}} \frac{\left[\alpha_{i, j}^{(x)}\right]^{2}}{\gamma_{i, j}^{(x, y)}}\right\} \phi_{i-\frac{1}{2}, j} \\
& +D_{i, j} \frac{\Delta y_{j}}{\Delta x_{i}}\left\{\left(\alpha_{i, j}^{(x)}-\beta_{i, j}^{(x)}\right)-\frac{\Delta y_{j}}{\Delta x_{i}} \frac{\left[\alpha_{i, j}^{(x)}\right]^{2}}{\gamma_{i, j}^{(x, y)}}\right\} \phi_{i+\frac{1}{2}, j}  \tag{3.63}\\
& -D_{i, j} \frac{\alpha_{i, j}^{(x)} \alpha_{i, j}^{(y)}}{\gamma_{i, j}^{(x, y)}}\left(\phi_{i, j+\frac{1}{2}}+\phi_{i, j-\frac{1}{2}}\right)-\frac{\Delta y_{j}^{2}}{2} \frac{\alpha_{i, j}^{(x)}}{\gamma_{i, j}^{(x, y)}} Q_{i, j}
\end{align*}
$$

Enforcing the continuity of the normal currents will yield (3.60), however this is not apparent without further simplification of the coefficients. Specifically we note that, after a little algebraic manipulation,

$$
\begin{equation*}
\alpha_{i, j}^{(x)}\left\{1-\frac{\alpha_{i, j}^{(x)}}{\gamma_{i, j}^{(x, y)}} \frac{\Delta y_{j}}{\Delta x_{i}}\right\}=\left[1-\delta_{i, j}^{(x)}\right] \alpha_{i, j}^{(x)}+\frac{\Delta x_{i}}{\Delta y_{j}} \frac{\alpha_{i, j}^{(x)} \alpha_{i, j}^{(y)}}{\gamma_{i, j}^{(x, y)}} \delta_{i, j}^{(x)} \tag{3.64}
\end{equation*}
$$

and thus the definition of $\hat{\zeta}_{i, j}^{(x, y)}$ in (3.61a) and $\hat{\vartheta}_{i, j}^{(x, y)}$ in (3.61b) complete the derivation.

### 3.6 Truncation Error Analysis

The second-order convergence of the lowest-order NEM and NIM discretizations is rigorously proven through their equivalence to certain FEMs [see Chapter 5]. However, such equivalences have not been established for all nodal methods, and hence a more rudimentary treatment of consistency and stability would be required to prove convergence in these cases. Unfortunately, a typical truncation analysis based on direct Taylor series expansions is inappropriate for nodal discretizations. This is a consequence of the cell- and edge-based unknowns, which are not associated with a particular point in their respective cell or along their respective edge. Hence a prudent treatment returns to the definitions of these unknowns, and considers the Taylor series expansion of the corresponding integrand. This technique is applicable to nodal methods in general, and is employed in the following discussion to verify the second order consistency of the lowest-order NEM and

NIM, under the assumption of piecewise constant coefficients. It is then used to examine the consistency of these methods for piecewise smooth coefficients.

We assume that the solution is sufficiently smooth on each cell $\Omega_{i, j}$ and write the Taylor series expansion about the cell centres $\left(x_{i}, y_{j}\right)$ in the form,

$$
\begin{aligned}
\mathbf{J}(x, y) \cdot \mathbf{n}_{x} & =p+p_{x}\left(x-x_{i}\right)+p_{y}\left(y-y_{j}\right)+\cdots \\
\mathbf{J}(x, y) \cdot \mathbf{n}_{y} & =q+q_{x}\left(x-x_{i}\right)+q_{y}\left(y-y_{j}\right)+\cdots \\
\phi(x, y) & =\varphi+\varphi_{x}\left(x-x_{i}\right)+\varphi_{y}\left(y-y_{j}\right)+\cdots
\end{aligned}
$$

where we have adopted a shorthand notation for the function and its derivatives evaluated at the centre of the cell (i.e. $\phi\left(x_{i}, y_{j}\right)=\varphi$ ). Thus, for example, the edge average $\phi_{i+\frac{1}{2}, j}$ has the expansion,

$$
\begin{align*}
\phi_{i+\frac{1}{2}, j} & =\frac{1}{\Delta y_{j}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}}\left\{\lim _{x \rightarrow x_{i+\frac{1}{2}}^{-}} \phi(x, y)\right\} d y \\
& =\frac{1}{\Delta y_{j}} \int_{y_{j-\frac{1}{2}}^{2}}^{y_{j+\frac{1}{2}}}\left\{\varphi+\varphi_{x}\left(x-x_{i}\right)+\varphi_{y}\left(y-y_{j}\right)+\cdots\right\} d y \\
& =\varphi+\frac{\Delta x_{i}}{2} \varphi_{x}+\frac{\Delta x_{i}^{2}}{8} \varphi_{x x}+\frac{\Delta y_{j}^{2}}{24} \varphi_{y y}+O\left(\Delta^{3}\right) \tag{3.65}
\end{align*}
$$

Although we assume that the source may be evaluated exactly, its analogous treatment is useful in the following discussion. Once again we adopt a shorthand notation,

$$
Q(x, y)=s+s_{x}\left(x-x_{i}\right)+s_{y}\left(y-y_{j}\right) \cdots
$$

such that the expansion of $Q_{i, j}$ may be written,

$$
\begin{equation*}
Q_{i, j}=s+\frac{\Delta x_{i}^{2}}{24} s_{x x}+\frac{\Delta y_{j}^{2}}{24} s_{y y}+O\left(\Delta^{4}\right) \tag{3.66}
\end{equation*}
$$

We note again that the proposed truncation analysis relates only to consistency. To bridge the gap between consistency and convergence, the stability of the lowest-order
nodal discretizations must be established, and we accomplish this indirectly through their equivalence to certain FEMs [see Chapter 5]. Alternatively, a local mode analysis could be employed to prove the stability of these nodal discretizations directly.

### 3.6.1 Piecewise Constant Coefficients

Using the approach outlined above we will now investigate the truncation error of the expressions in Definition 3.4.2 (NEM) and Definition 3.5.1 (NIM). In particular, it was noted in Section 3.1.1 that, under the assumption of piecewise constant coefficients and exact integration of the source, the balance equation (3.8) is an exact expression. Hence there is no truncation error,

$$
T_{\text {balance }}=0
$$

Next we consider the one-sided normal currents (3.23) of NEM, for which substitution of the necessary expansions gives

$$
T_{J_{i \pm \frac{1}{2}, j}^{\mp}}=-\frac{\Delta x_{i}^{2}}{12} D_{i, j} \varphi_{x x x}
$$

with an analogous expression for the normal components in $y$. This is in contrast to the truncation error of the one-sided normal currents (3.45) of NIM, which are given by

$$
T_{J_{i \pm \frac{1}{2}, j}^{\mp}}=\frac{\Delta x_{i}^{2}}{12}\left[-D_{i, j} \varphi_{x x x}+\left(\Sigma_{a}\right)_{i, j} \varphi_{x}\right]
$$

Differentiating the balance equation with respect to $x$ we see that, if the source was piecewise constant and $\varphi_{y y x}=0$ (i.e. $\varphi_{y y}=$ constant, a condition satisfied by an essentially one-dimensional problem) then this term would vanish, a property anticipated by the exact treatment of the transverse integrated ODEs. Moreover, in situations where $\varphi_{y y x}$ and $s_{x}$ are small, the truncation error of NIM will be significantly smaller than that of NEM.

Finally, for both NEM and NIM continuity of normal current is imposed exactly in the weak sense, but pointwise it is second order. Hence the lowest-order nodal methods are second order consistent on a non-uniform mesh. A similar approach may be used to treat the flux formulation and the edge-based flux formulation directly.

### 3.6.2 Piecewise Smooth Coefficients

To evaluate the potential of nodal discretizations in a more general, setting and to suggest possible improvements we extend the previous analysis to include piecewise smooth coefficients. In particular, we assume that the coefficients may be expanded about the centre of the cell,

$$
\begin{aligned}
\Sigma_{a}(x, y) & =\sigma+\sigma_{x}\left(x-x_{i}\right)+\sigma_{y}\left(y-y_{j}\right)+\cdots \\
D(x, y) & =\mathrm{D}+\mathrm{D}_{x}\left(x-x_{i}\right)+\mathrm{D}_{y}\left(y-y_{j}\right)+\cdots
\end{aligned}
$$

This spatial dependence introduces an error in the balance equation,

$$
T_{\text {balance }}=\frac{1}{24}\left\{\Delta x_{i}^{2} \sigma_{x} \varphi_{x}+\Delta y_{j}^{2} \sigma_{y} \varphi_{y}\right\}+\mathcal{O}\left(\Delta^{3}\right)
$$

But the truncation error of the one sided currents (both NEM and NIM) becomes first order,

$$
T_{J_{i \pm \frac{1}{2}, j}}= \pm \frac{\Delta x_{i}}{2} \mathrm{D}_{x} \varphi_{x}+\mathcal{O}\left(\Delta^{2}\right)
$$

with analogous expressions for the normal currents in $y$. The continuity of normal currents is unaffected. Thus the discretization is only first order consistent in this more general setting. However, the first order term of the error is proportional to the derivatives of the coefficients and may be small in practice.

The restoration of second order consistency appears feasible but introduces additional complexities into the discretization. Specifically, replacing the constant diffusion
coefficient with a diagonal tensor,

$$
D(x, y)=\left[\begin{array}{cc}
\mathrm{D}+\mathrm{D}_{x}\left(x-x_{i}\right) & 0  \tag{3.67}\\
0 & \mathrm{D}+\mathrm{D}_{y}\left(y-y_{j}\right)
\end{array}\right]
$$

would preserve the simple transverse average of Fick's law (3.15) while restoring second order truncation error in the one-sided currents. Note that the the use of NIM is still possible, although it may not be possible to solve the transverse integrated ODEs with more general polynomial coefficients.

Although the truncation error in the balance equation is second order these gradients may be large, and hence on coarse meshes introduce unacceptable errors. Eliminating these second order terms within the nodal procedure is nontrivial. For example, assume that only linear terms are present in the expansion of the absorption coefficient, and consider the removal term of the balance equation,

$$
\begin{aligned}
& \frac{1}{\Delta x_{i} \Delta y_{j}} \int_{x_{i-i}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \Sigma_{a}(x, y) \phi(x, y) d x d y \\
= & \frac{1}{\Delta x_{i} \Delta y_{j}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}^{2}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}}\left[\sigma+\sigma_{x}\left(x-x_{i}\right)+\sigma_{y}\left(y-y_{j}\right)\right] \phi(x, y) d x d y \\
= & \sigma \phi_{i, j}+\sigma_{x} \phi_{i, j}^{(1,0)}+\sigma_{y} \phi_{i, j}^{(0,1)}
\end{aligned}
$$

We have introduced the higher order cell-based moments of the flux (i.e. $\phi_{i, j}^{(1,0)}, \phi_{i, j}^{(0,1)}$ ) to accommodate the linear terms of $\Sigma_{a}(x, y)$. Within NEM a higher order expansion of the transverse averaged flux in conjunction with the weighted residual procedure [Section 3.4.1] may be used to derive the additional equations necessary to govern these new unknowns (e.g. [73]). Unfortunately, the application of NIM may not be feasible in this setting.

### 3.7 Numerical Experiments

To confirm the theoretical prediction of the convergence rate a simple test problem with known solution was contrived. The results of these computations are presented in Section 3.7.1. A more practical example from the reactor physics community, namely the ironwater benchmark problem [Dilber and Lewis [25]] is used to demonstrate the superior accuracy of nodal schemes, particularly NIM.

### 3.7.1 Example 1: A Simple Test

To verify the results of the truncation error analysis [Section 3.6.1], consider the square domain $[0,1] \times[0,1]$, over which we take the diffusivity and the absorption to be constant. In particular, we choose $D(x, y) \equiv 1, \Sigma_{a}(x, y) \equiv 1$, the source term,

$$
Q(x, y)=2+\left[1+\left(\alpha^{2}+\beta^{2}\right) \pi^{2}\right] \sin (2 \pi x) \sin (2 \pi y)
$$

and the Dirichlet boundary condition,

$$
\phi(x, y)=2 \quad \forall(x, y) \in \partial \Omega
$$

The exact solution is,

$$
\phi(x, y)=2+\sin (\alpha \pi x) \sin (2 \beta y)
$$

Thus, in the nodal discretizations we have $D_{i, j}=\left(\Sigma_{a}\right)_{i, j}=1$ with the cell-based source,

$$
Q_{i, j}=2\left[1+\left(\alpha^{2}+\beta^{2}\right) \pi^{2}\right]\left\{\frac{\sin \left(\pi \Delta x_{i}\right)}{\pi \Delta x_{i}}\right\}\left\{\frac{\sin \left(\pi \Delta y_{j}\right)}{\pi \Delta y_{j}}\right\} \sin (2 \pi x) \sin (2 \pi y)
$$

The results are displayed in Figure 3.5 where the transverse averages are displayed at their respective midpoints and the cell averages at the cell centres. Both surfaces
are in good qualitative agreement with the exact solution. Although this is encouraging we desire a quantitative estimate of the second order convergence of these nodal discretizations. Unfortunately, the transverse averaged quantities associated with a given mesh, for example $(\Delta x, \Delta y)$ do not exist on any finer mesh. This is also true of the cell averages. However, for certain fine mesh spacings, such as $(\Delta x / n, \Delta y / n)$, where $n$ is a positive integer, the coarse grid unknowns are, by definition, arithmetic means of the fine grid unknowns. Thus, the standard pointwise estimate of the convergence rate is complicated somewhat by the necessity to choose a coarse grid on which all errors will be computed. In order to obtain convincing asymptotic convergence behaviour without considering excessively fine meshes, errors were computed on a $10 \times 10$ grid. The results for the transverse average quantities are displayed in Table 3.1, while the cell averages appear in Table 3.3. In both cases second order convergence of the constant-constant NIM discretization is observed. Similar results are readily obtained for the lowest-order NEM.

It is also possible to obtain global convergence estimates by evaluating the error in an $L_{p}$-norm. The results for the $L_{\infty}$-norm, which are shown in Table 3.2 and Table 3.3, also confirm the second order convergence of the constant-constant NIM discretization.

Table 3.1: Convergence results for Example 1: the transverse averaged scalar flux on a $10 \times 10$ uniform mesh.

| Computational Grid | $\left\|\phi_{\frac{13}{2}, 4}-\phi_{\frac{13}{2}, 4}^{\Delta x, \Delta y}\right\|$ | Ratio | $\left\|\phi_{5, \frac{9}{2}}-\phi_{5, \frac{9}{2}}^{\Delta x, \Delta y}\right\|$ | Ratio |
| :---: | :---: | :---: | :---: | :---: |
| $10 \times 10$ | $1.56802 \times 10^{-2}$ | - | $5.98930 \times 10^{-3}$ | - |
| $20 \times 20$ | $3.90134 \times 10^{-3}$ | 4.019 | $1.49018 \times 10^{-3}$ | 4.019 |
| $40 \times 40$ | $9.74176 \times 10^{-4}$ | 4.005 | $3.72102 \times 10^{-4}$ | 4.005 |
| $80 \times 80$ | $2.43472 \times 10^{-4}$ | 4.001 | $9.29979 \times 10^{-5}$ | 4.001 |
| $160 \times 160$ | $6.08634 \times 10^{-5}$ | 4.000 | $2.32477 \times 10^{-5}$ | 4.000 |

Table 3.2: Convergence results for Example 1 in the $L_{\infty}$-norm.

| Grid | $\left\\|\phi_{i+\frac{1}{2}, j}-\phi_{i+\frac{1}{2}, j}^{\Delta x y}\right\\|_{\infty}$ | Ratio | $\left\\|\phi_{i, j+\frac{1}{2}}-\phi_{i, j+\frac{1}{2}}^{\Delta x, \Delta y}\right\\|_{\infty}$ | Ratio |
| :---: | :---: | :---: | :---: | :---: |
| $10 \times 10$ | $3.13604 \times 10^{-2}$ | - | $3.13604 \times 10^{-2}$ | - |
| $20 \times 20$ | $8.20423 \times 10^{-3}$ | 3.822 | $8.20423 \times 10^{-3}$ | 3.822 |
| $40 \times 40$ | $2.07415 \times 10^{-3}$ | 3.955 | $2.07415 \times 10^{-3}$ | 3.955 |
| $80 \times 80$ | $5.19987 \times 10^{-4}$ | 3.989 | $5.19987 \times 10^{-4}$ | 3.989 |
| $160 \times 160$ | $1.30088 \times 10^{-4}$ | 3.997 | $1.30088 \times 10^{-4}$ | 3.997 |

Table 3.3: Convergence results for Example 1: the cell averages.

| Grid | $\left\|\phi_{4,3}-\phi_{4,3}^{\Delta x, \Delta y}\right\|$ | Ratio | $\left\\|\phi_{i, j}-\phi_{i, j}^{\Delta x, \Delta y}\right\\|_{\infty}$ | Ratio |
| :---: | :---: | :---: | :---: | :---: |
| $10 \times 10$ | $2.55358 \times 10^{-2}$ | - | $3.15640 \times 10^{-2}$ | - |
| $20 \times 20$ | $6.48514 \times 10^{-3}$ | 3.938 | $8.01607 \times 10^{-3}$ | 3.938 |
| $40 \times 40$ | $1.62752 \times 10^{-3}$ | 3.985 | $2.06219 \times 10^{-3}$ | 3.887 |
| $80 \times 80$ | $4.07269 \times 10^{-4}$ | 3.996 | $5.19237 \times 10^{-4}$ | 3.972 |
| $160 \times 160$ | $1.01841 \times 10^{-4}$ | 3.999 | $1.30041 \times 10^{-4}$ | 3.993 |



Figure 3.5: The transverse averaged scalar fluxes, $\left\{\phi_{i+\frac{1}{2}, j}, \phi_{i, j+\frac{1}{2}}\right\}$ obtained with the constant-constant NIM for this simple test case are shown in (a) while the corresponding node averages, $\phi_{i, j}$ are plotted in (b).

### 3.7.2 Example 2: The Iron-Water Benchmark

Since the inception of reactor modelling, organizations within the reactor physics community such as the International Atomic Energy Agency (IAEA) [4] have been designing and cataloguing benchmark problems. Many of these benchmarks have been used by a number of authors (see review by Lawrence [52]) to demonstrate the superior accuracy of nodal methods. Unfortunately, the majority of these benchmarks are physically meaningful complex models of various reactor assemblies, and as such are beyond the scope of this thesis. A reasonable compromise is offered by the iron-water benchmark problem first employed by Dilber and Lewis [25]. Specifically, this benchmark requires relatively little background in transport theory or reactor physics and yet provides compelling evidence of the nodal methods superiority, particularly highlighting the performance of NIM.

The governing equation in the iron-water benchmark is the one-group diffusion equation (2.12). We recall that in this setting the diffusion coefficient $D$, and the macroscopic absorption cross-section $\Sigma_{a}$, are defined by,

$$
D=\frac{1}{3 \Sigma_{t}}, \quad \Sigma_{a}=\Sigma_{t}-\Sigma_{s}
$$

where $\Sigma_{t}$ is the total macroscopic cross-section and $\Sigma_{s}$ is the macroscopic scattering cross-section. These macroscopic properties of iron and water are given in Table 3.4 along with the corresponding values of the diffusion coefficient and macroscopic removal crosssection obtained with the above definitions. The domain of this shielding experiment, $\Omega \equiv[0,30] \times[0,30]$, is divided into four regions which are shown schematically in Figure 3.6. The source is piecewise constant and may be written,

$$
Q(x, y)= \begin{cases}1 & (x, y) \in \Omega_{1} \\ 0 & (x, y) \in \Omega \backslash \Omega_{1}\end{cases}
$$

Table 3.4: The diffusivity $D$, and the absorption cross-section $\Sigma_{a}$, for Example 2 [25].

| Material | $\Sigma_{t}[\mathrm{~cm}]^{-1}$ | $\Sigma_{s}[\mathrm{~cm}]^{-1}$ | $D[\mathrm{~cm}]$ | $\Sigma_{a}[\mathrm{~cm}]^{-1}$ |
| :---: | :---: | :---: | :---: | :---: |
| Water | 3.33 | 3.31002 | 0.10010 | 0.01998 |
| Iron | 1.33 | 1.10523 | 0.25063 | 0.22477 |



Figure 3.6: Schematic of the iron-water benchmark problem [25]

Finally, to specify the boundary conditions it is convenient to define the following segments,

$$
\begin{aligned}
& \Gamma_{-,-}=\{(0 \leq x \leq 1, y=0) \bigcup(x=0,0 \leq y \leq 1)\} \\
& \Gamma_{+,+}=\{(0<x \leq 1, y=1) \bigcup(x=1,0<y \leq 1)\}
\end{aligned}
$$

such that on the lower left segment we have a reflecting or zero net current boundary condition (2.19b),

$$
\mathbf{J} \cdot \mathbf{n}=0 \quad \forall(x, y) \in \Gamma_{-,-}
$$

whereas on the upper right segment a vacuum boundary is assumed which implies that the inward directed partial current is set to zero as in (2.18),

$$
-\frac{1}{4} \phi+\frac{1}{2} \mathbf{J} \cdot \mathbf{n}=0 \quad \forall(x, y) \in \Gamma_{+,+}
$$

The homogenized coefficients which appear in the nodal discretizations are trivially defined from Table 3.4 as the grid is chosen such that both the diffusion coefficient $D(x, y)$ and the macroscopic absorption cross-section $\Sigma_{a}(x, y)$, which are piecewise constant over $\Omega$, are constant over each cell $\Omega_{i, j}$.

For the purposes of comparison the three discretizations developed in this Chapter were used to compute the solution of the iron-water benchmark problem. The first, which will be referred to as the finite volume discretization, is the classical 5-point cell-based scheme obtained in Section 3.1. The distinguishing feature of this discretization is the natural definition of the interface diffusion coefficient as the harmonic mean of its neighbours. The second discretization is the lowest-order NEM that was developed in Section 3.4. The fundamental improvement of this NEM over the finite volume discretization is that the currents are continuous and second order convergent. The constant-constant NIM of Section 3.5 is the third method and incorporates analytic information from the
transverse integrated ODEs. In the nonconservative case this results in a more costly stencil for which the truncation analysis [Section 3.6] suggests a significant improvement in accuracy may result. The NIM solution of the iron water benchmark problem on a $10 \times 10$ computational mesh is shown in Figure 3.7.

The use of an efficient multigrid solver [23] with the finite volume discretization enabled the generation of the reference solution on a $640 \times 640$ uniform mesh. The error associated with the reference solution is bounded in the maximum-norm by 0.01 . Hence the deviations given in Table 3.5 provide compelling evidence of the superior accuracy which may be obtained with nodal schemes. Specifically, in the $L_{2}$-norm, the NEM error is approximately 3 times less than that of the finite volume method, while the NIM error is nearly 7 times less than that of NEM. Similar reductions of the error in the $L_{\infty^{-}}$ norm are also shown. The exceptional performance of NIM is further highlighted in the cross-sections that are displayed in Figure 3.8, where the NIM cross-section is virtually indistinguishable from that of the reference solution. We also note that [Figure 3.8(b)] on this very coarse mesh the NEM discretization generates a solution with slightly negative, and therefore nonphysical, values near the layer. In contrast the NIM solution is positive everywhere. This observation is consistent with our forthcoming conjectures in Section 5.2.2 regarding NEM and in Section 5.2.3 regarding NIM.

Underlying the above comparisons of accuracy, is the computational cost of each method. In particular, the sparsity structure of the lowest-order NEM and constantconstant NIM are identical, and hence, the computational cost in these cases is identical. However, the computational cost of the finite volume method, on a given mesh, is lower. Taking into account the flop counts of these methods, it is possible to show that for this example, the work required to obtain comparable accuracy using the finite volume method is approximately 5 times that of NIM. Moreover, this computational savings is
independent of the mesh-size that is used to obtain the NIM approximation. This is critical aspect of a fair comparison, and is facilitated by the new family of fast iterative solvers that are developed in Chapter 6 and exhibit mesh-independent convergence.

Table 3.5: Errors for Example 2 on a $10 \times 10$ computational grid.

| Discretization | $\left\\|\phi_{i, j}-\phi_{i, j}^{\Delta x, \Delta y}\right\\|_{2}$ | $\left\\|\phi_{i, j}-\phi_{i, j}^{\Delta x, \Delta y}\right\\|_{\infty}$ |
| :--- | :---: | :---: |
| Finite Volume | 0.991 | 4.396 |
| lowest-order NEM | 0.315 | 1.126 |
| constant-constant NIM | 0.047 | 0.146 |



Figure 3.7: The constant-constant NIM solution of Example 2 computed on a $10 \times 10$ grid: (a) transverse averaged scalar fluxes (b) cell averages.


Figure 3.8: Solution cross-sections for Example 2 are compared to a reference solution that was obtained using the finite volume discretization on a $640 \times 640$ uniform mesh: $y_{j}=10.5$ (a) full domain (b) zoom on layer.

## Chapter 4

## Homogenization

In a practical reactor model, the coefficients of the multi-group diffusion equations exhibit strong variations on two significantly different length scales. Unfortunately, the computational cost associated with the fine-scale mesh is prohibitive, and yet the fine-scale behaviour of the coefficients significantly influences the coarse-scale properties of the solution. Generally, an efficient computational treatment of such two-scale problems relies on averaging the differential operator (in some sense) over the microscopic or fine scale to obtain an homogenized problem governing the macroscopic or coarse-scale solution. It is specifically this class of problems for which nodal methods were developed [Chapter 3] and their excellent performance (cf. Section 3.7) is tightly coupled to the corresponding homogenized coefficients. We note that the strong multi-scale dependence of coefficients arises in many applications outside of reactor modelling, including flow in porous media, heat transfer, and electromagnetism. Hence, the process of homogenization has received considerable attention both by the reactor physics community (see review by Smith [65]) and by others (cf. [11, 62]).

The most popular techniques within the reactor physics community pursue a physically motivated equivalence between certain coarse-scale properties (e.g. the coarse-scale cell average) of the solution of the two-scale problem, and the solution of the homogenized problem. A strict equivalence is not possible in general, and in Section 4.1 we review a popular approximation of Generalized Equivalence Theory. In other applications a rigorous mathematical technique based on a two-scale asymptotic analysis has been employed.

This technique has been particularly effective for periodic and nearly (i.e. non-uniformly) periodic structures and we review the relevant results of this work in Section 4.2. Finally, we present a new multilevel approximate homogenization algorithm in Section 4.3 that maintains many of the important properties of the rigorous asymptotic analysis and is considerably more efficient.

### 4.1 Equivalence Techniques in Reactor Modelling

There are, in fact, two separate issues of homogenization in a reactor core, and consequently, a two stage homogenization process. The first stage of homogenization utilizes various transport approximations and probabilistic methods to account for the breakdown of the $P_{1}$ approximation near interfaces, such as fuel pins, as well as the strong energy dependence of certain effects, such as absorption resonances [65]. These issues, while extremely important for the success of the model, are beyond the scope of the thesis. As our starting point we consider the output of this first stage homogenization procedure: the heterogeneous multi-group diffusion problem, which contains piecewise constant coefficients on the fine scale. In the interest of consistency and clarity we restrict our discussion to the one-group form, ${ }^{1}$ and explicitly define the heterogeneous and homogenized one-group fission-source diffusion problems [cf. Definition 2.14].

Definition 4.1.1. The heterogeneous one-group diffusion problem is defined by,

$$
\begin{align*}
\nabla \cdot \widetilde{\mathbf{J}}+\widetilde{\Sigma_{a}} \widetilde{\phi} & =\frac{\chi \nu}{k} \widetilde{\Sigma_{f}} \widetilde{\phi}  \tag{4.1a}\\
\widetilde{\mathbf{J}} & =-\widetilde{D} \nabla \widetilde{\phi} \tag{4.1b}
\end{align*}
$$

[^5]where the parameters $\widetilde{D}(x, y), \widetilde{\Sigma_{a}}(x, y), \widetilde{\Sigma_{f}}(x, y)$ are prescribed functions of $(x, y)$ that are piecewise constant on the fine-scale cells $F_{i, j}$ and subject to particular boundary conditions. The corresponding heterogeneous solution is $(\widetilde{\phi}, \widetilde{\mathbf{J}})$.

Definition 4.1.2. The homogeneous one-group diffusion problem is defined by,

$$
\begin{align*}
\nabla \cdot \widehat{J}+\widehat{\Sigma_{a}} \hat{\phi} & =\frac{\chi \nu}{k} \widehat{\Sigma_{f}} \widehat{\phi}  \tag{4.2a}\\
\widehat{J} & =-\widehat{D} \nabla \widehat{\phi} \tag{4.2~b}
\end{align*}
$$

where the parameters $\widehat{D}(x, y), \widehat{\Sigma_{a}}(x, y), \widehat{\Sigma_{f}}(x, y)$ are to be computed with a homogenization procedure (typically piecewise constant on the coarse-scale cells $C_{i, j}$ ) and ( $\widehat{\phi}, \widehat{J}$ ) is the corresponding solution.

The objective set out by the coarse-mesh nodal methods is for these homogenized parameters to be piecewise constant functions (i.e. constant on an assembly size coarsemesh cell $C_{i, j}$ ). In this situation we have shown that NEM and NIM lead to a discrete solution $\widehat{\phi}_{h}$ that is a second order approximation of $\widehat{\phi}$. In practice, errors of only $1-2 \%$ in average assembly powers are observed on meshes having only a single cell per assembly [65]. However, we demonstrated that for piecewise smooth coefficients these methods are only first order [Section 3.6] and in general, $\widehat{\phi}-\widetilde{\phi}=\mathcal{O}\left(\Delta^{c}\right)$ where $\Delta^{c}$ is the assembly size coarse-mesh spacing. This leads to errors of approximately $5-10 \%$ between the average assembly powers of $\widehat{\phi}_{h}$ and those of $\widetilde{\phi}$. Certainly a loss of fine-scale information is an inevitable consequence of the homogenization process. However, one would hope for better accuracy in such coarse-scale properties as the assembly averaged powers. Thus, the essential objective of the homogenization procedure is to reduce the errors associated with various prescribed coarse-scale quantities.

### 4.1.1 Coarse-Scale Properties

To motivate the homogenization techniques which follow we identify several integrated quantities that would ideally be preserved under homogenization. Studying the implication that this equivalence has in the definition of the homogenized parameters serves to demonstrate the extreme difficulties inherent in this process. A natural quantity to consider is the assembly averaged reaction rate whose preservation would require

$$
\begin{equation*}
\int_{C_{i, j}} \widehat{\Sigma_{\alpha}}(x, y) \widehat{\phi}(x, y) d x d y=\int_{C_{i, j}} \widetilde{\Sigma_{\alpha}}(x, y) \widetilde{\phi}(x, y) d x d y \tag{4.3}
\end{equation*}
$$

( $\alpha=a$ or $\alpha=f$ ) which under the assumption of piecewise constant homogenized parameters gives:

$$
\begin{equation*}
\left(\widehat{\Sigma_{\alpha}}\right)_{i, j}=\frac{\int_{C_{i, j}} \widetilde{\Sigma_{\alpha}}(x, y) \widetilde{\phi}(x, y) d x d y}{\int_{C_{i, j}} \widehat{\phi}(x, y) d x d y} \tag{4.4}
\end{equation*}
$$

Thus, we encounter the first well-known problem in defining an homogenization procedure: the ideal homogenized interaction coefficients depend upon the unknown heterogeneous solution as well as the homogeneous solution, the solution to the very problem we are working to define. Even if an efficient nonlinear iteration could be defined to solve for both homogenized quantities simultaneously, the fact remains that the heterogeneous solution is unknown.

The edge-averaged net current on each edge of an assembly is another important quantity and its preservation provides insight into the additional problems associated with defining an ideal homogenized diffusion coefficient. Specifically we would require,

$$
\begin{equation*}
\int_{\partial C_{i, j}^{k}} \widehat{J}(x, y) \cdot \mathbf{n} d s=\int_{\partial C_{i, j}^{k}} \widetilde{\mathbf{J}}(x, y) \cdot \mathbf{n} d s \tag{4.5}
\end{equation*}
$$

where $\partial C_{i, j}^{k}$ is the $k^{t h}$ edge of the coarse-mesh cell $C_{i, j}$. Utilizing the homogenized Fick's

Law (4.2b), we could define the homogenized diffusion coefficient as,

$$
\begin{equation*}
\widehat{D}_{i, j}=\frac{\int_{\partial C_{i, j}^{k}} \widetilde{\mathbf{J}}(x, y) \cdot \mathbf{n} d s}{\int_{\partial C_{i, j}^{k}} \nabla \widehat{\phi} \cdot \mathbf{n} d s} \tag{4.6}
\end{equation*}
$$

Therefore, overlooking the dependence on the unknown heterogeneous solution and the problem's own solution, each normal current equivalence (4.5) may result in a different value of $\widehat{D}_{i, j}$. Hence, a constant diffusion coefficient on each coarse-mesh cell is insufficient. This is not surprising: choosing $\mathbf{J} \cdot \mathbf{n}$ to be an edge-based quantity suggests, through Fick's law, that $D$ should also be edge-based, or at least be described by a corresponding number of degrees of freedom. For example, one might consider the diagonal linear tensor (3.67) or alternatively a constant, full tensor. Based on such suggestions it is clear that procedures which may preserve or reduce the errors associated with some desirable coarse-scale properties of the heterogeneous solution may necessitate modifications of the assumptions inherent in the derivation of standard nodal methods. Nevertheless, the most significant problem facing the definition of homogenized parameters is their dependence on the solutions of the heterogeneous and homogeneous problems.

### 4.1.2 A One Dimensional Study

To clarify the objectives of equivalence based homogenization and identify certain classes of problems for which equivalences may be determined independently of $\tilde{\phi}$, we begin with a discussion of the one-dimensional case. Specifically, in perhaps the first equivalence based homogenization procedure Selengut [64] suggested that the replacement of a heterogeneous cell with its homogenized counterpart should not perturb the solution outside the cell in any way. In one dimension this interpretation of the homogenization
of the cell $C_{i} \equiv\left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right]$ corresponds to the constraint that the boundary fluxes,

$$
\widehat{\phi}_{i-\frac{1}{2}}=\widehat{\phi}\left(x_{i-\frac{1}{2}}\right)=\widetilde{\phi}\left(x_{i-\frac{1}{2}}\right), \quad \widehat{\phi}_{i+\frac{1}{2}}=\widehat{\phi}\left(x_{i+\frac{1}{2}}\right)=\widetilde{\phi}\left(x_{i+\frac{1}{2}}\right)
$$

and the boundary currents,

$$
\widehat{J}_{i-\frac{1}{2}}=\widehat{J}\left(x_{i-\frac{1}{2}}\right)=\widetilde{J}\left(x_{i-\frac{1}{2}}\right), \quad \widehat{J}_{i+\frac{1}{2}}=\widehat{J}\left(x_{i+\frac{1}{2}}\right)=\widetilde{J}\left(x_{i+\frac{1}{2}}\right)
$$

of the heterogeneous and homogenized problems be equal. As an example consider the homogenization of a layered structure composed of infinite slabs: the cell $C_{b} \equiv[-b, b]$ for which the diffusion coefficient is given by

$$
D(x)= \begin{cases}D_{0} & |x|<a \\ D_{1} & a<|x|<b\end{cases}
$$

To this end, we first note that assuming a planar source at large negative $x$ with a constant diffusion coefficient the solution has the general form,

$$
\widehat{\phi}(x)=A+B x
$$

where $A$ and $B$ are constants determined by the strength and exact location of the source. If we now consider inserting the cell $C_{b}$ and enforce continuity of the scalar flux, we obtain

$$
\widetilde{\phi}(x)= \begin{cases}A+B \frac{\left(1+\epsilon_{0}\right)}{\left(1+\epsilon_{0} v\right)} x & |x|<a \\ A+B \frac{\left(x+\operatorname{sgn}(x) \epsilon_{0} a\right)}{\left(1+\epsilon_{0} v\right)} & a<|x|<b \\ A+B x & |x|>b\end{cases}
$$

where

$$
\epsilon_{0}=\frac{\left[D_{1}-D_{0}\right]}{D_{0}}
$$

and $v=a / b$ is the volume fraction of material 0 in the cell $C_{b}$. Thus the flux constraint becomes trivial, while the equivalence of net currents yields the well known result,

$$
\widehat{D}=\frac{D_{0} D_{1}}{(1-v) D_{0}+v D_{1}}=\left\{\frac{v}{D_{0}}+\frac{1-v}{D_{1}}\right\}^{-1}
$$

the harmonic average. An interesting benefit of this approach is that, by considering an infinite cylindrical geometry with azimuthal symmetry or an infinite spherical geometry, the corresponding two- or three-dimensional results are obtained. The general expression may be written,

$$
\begin{equation*}
\frac{(1+n v) D_{0}+n(1-v) D_{1}}{(1-v) D_{0}+(n+v) D_{1}} \tag{4.7}
\end{equation*}
$$

where $v=(a / b)^{n+1}$ and $n+1$ is the number of dimensions [64]. A comparison of the 2 D form of this result with commonly used averages, asymptotic analysis and the new multigrid procedure is made in Section 4.3.

Difficulties in the extension of this analysis to relevant, truly multidimensional problems was likely the reason that the impact of this research was primarily its underlying equivalence ideology. In fact, it wasn't until 1976 that Kollas and Henry [49] conducted a more detailed study of these ideas. To demonstrate their strategy we consider the constant coefficient source-free problem,

$$
-\frac{d}{d x}\left\{\widehat{D} \frac{d}{d x} \widehat{\phi}(x)\right\}=0, \quad a<x<b
$$

where, after a little algebra, one can deduce the relationship between the endpoint values,

$$
\left[\begin{array}{l}
\widehat{\phi}(a)  \tag{4.8}\\
\widehat{J}(a)
\end{array}\right]=\left[\begin{array}{cc}
1 & \Delta \\
0 & 1
\end{array}\right]\left[\begin{array}{l}
\widehat{\phi}(b) \\
\widehat{J}(b)
\end{array}\right]
$$

with $\Delta=b-a$. Now, consider a piecewise diffusion coefficient:

$$
D(x)= \begin{cases}D_{0} & a<x<x_{0} \\ D_{1} & x_{0}<x<b\end{cases}
$$

Using (4.8) we readily obtain,

$$
\left[\begin{array}{c}
\widetilde{\phi}(a) \\
\widetilde{J}(a)
\end{array}\right]=\left[\begin{array}{cc}
1 & \frac{\Delta_{1}}{D_{1}} \\
0 & 1
\end{array}\right]\left[\begin{array}{cc}
1 & \frac{\Delta_{2}}{D_{2}} \\
0 & 1
\end{array}\right]\left[\begin{array}{c}
\widetilde{\phi}(b) \\
\widetilde{J}(b)
\end{array}\right]=\left[\begin{array}{cc}
1 & \frac{\Delta_{1}}{D_{1}}+\frac{\Delta_{2}}{D_{2}} \\
0 & \cdot 1
\end{array}\right]\left[\begin{array}{c}
\widetilde{\phi}(b) \\
\widetilde{J}(b)
\end{array}\right]
$$

where $\Delta_{1}=x_{0}-a$ and $\Delta_{2}=b-x_{0}$. Thus, enforcing the equality of these boundary values yields,

$$
\begin{equation*}
\widehat{D}=\Delta\left\{\frac{\Delta_{1}}{D_{1}}+\frac{\Delta_{2}}{D_{2}}\right\}^{-1} \tag{4.9}
\end{equation*}
$$

which once again is the well known discrete result that, in one dimension, the harmonic average is considered the exact homogenization. However, in their paper Henry and Kollas specifically considered the use of this equivalence analysis for the one-group, one-dimensional diffusion equation. For symmetric regions they were able to determine homogenized parameters, $\left(\widehat{D}, \widehat{\Sigma_{\alpha}}\right)$ that were independent of the boundary values. However, more significantly they demonstrated that for a general region the homogenization process is overdetermined. We note that this distinction does not arise with the conservative, source-free, diffusion problem discussed above.

### 4.1.3 Multidimensional Equivalence Techniques

In more than one dimension a weaker equivalence is sought, as was suggested by the discussion in Section 4.1.1. In particular, a number of popular techniques have been implemented which endeavor to approximate the equivalences suggested by (4.3) and (4.5). The common thread among these methods is the use of locally defined auxiliary problems whose solution may be computed very accurately on a fine mesh. The most common of these considers each assembly with homogeneous Neumann boundary conditions and thus for convenient reference we provide the following definition.

Definition 4.1.3. The heterogeneous assembly one-group diffusion problem is defined by,

$$
\begin{align*}
& \nabla \cdot \widetilde{\mathcal{J}}+\widetilde{\Sigma_{a}} \widetilde{\varphi}=\frac{1}{k} \chi \nu \widetilde{\Sigma_{f}} \widetilde{\varphi}  \tag{4.10a}\\
& \widetilde{\mathcal{J}}=-\widetilde{D} \nabla \widetilde{\phi} \tag{4.10~b}
\end{align*}
$$

for $(x, y) \in C_{i, j}$ and is subject to the reflecting boundary condition [i.e. homogeneous Neumann BCs (2.19b)]

$$
\widetilde{\mathcal{J}} \cdot \mathbf{n}=0 \quad \forall(x, y) \in \partial C_{i, j}
$$

The parameters $\widetilde{D}(x, y), \widetilde{\Sigma_{a}}(x, y), \widetilde{\Sigma_{f}}(x, y)$ are prescribed functions of $(x, y)$ that are piecewise constant on the fine-scale cells $F_{i, j}$ The assembly solution is denoted $(\widetilde{\mathcal{J}}, \widetilde{\varphi})$.

The local solution $(\widetilde{\mathcal{J}}, \widetilde{\varphi})$ serves as an approximation of the heterogeneous solution in the desired equivalence relation. In the following discussion we present the first and perhaps most common of these techniques, flux weighting, as well as a prototypical example of its popular successors, an approximation of generalized equivalence theory.

## Flux Weighting

In the 1970's [41] and well into the 1980's [65] the most common approach to defining homogenized parameters was flux weighting. Motivated by (4.4) the heterogeneous solution $\widetilde{\phi}(x, y)$ on $C_{i, j}$ is approximated by $\widetilde{\varphi}(x, y)$ [Definition 4.1.3]. In addition, it is assumed that the node averages for this auxiliary problem and the homogenized problem are equal; that is,

$$
\int_{C_{i, j}} \tilde{\varphi}(x, y) d x d y=\int_{C_{i, j}} \widehat{\phi}(x, y) d x d y
$$

Direct substitution into (4.4) yields the flux weighted cross-sections ${ }^{2}$

$$
\begin{equation*}
\left(\widehat{\Sigma_{\alpha}}\right)_{i, j}^{(f w)}=\frac{\int_{C_{i, j}} \widetilde{\Sigma_{\alpha}}(x, y) \widetilde{\varphi}(x, y) d x d y}{\int_{C_{i, j}} \widetilde{\varphi}(x, y) d x d y} \tag{4.11}
\end{equation*}
$$

Typically the homogenized diffusion coefficient is taken from the transport definition [see Section 2.2.1], $D=\left[3 \Sigma_{t}\right]^{-1}$ hence the flux weighting of $\Sigma_{t}$ gives,

$$
\begin{equation*}
\widehat{D}_{i, j}^{(f w)}=\left\{\frac{\int_{C_{i, j}}[\widetilde{D}(x, y)]^{-1} \tilde{\varphi}(x, y) d x d y}{\int_{C_{i, j}} \widetilde{\varphi}(x, y) d x d y}\right\}^{-1} \tag{4.12}
\end{equation*}
$$

which interestingly is a weighted harmonic average.
The natural criticism of this approach is regarding the error associated with the zero net current boundary conditions. Not only can it be large but, perhaps more importantly, it is difficult to predict. The unfortunate implication is that none of the aforementioned integral quantities are preserved and the errors associated with them are unknown. Moreover, in the conservative case the solution $\widetilde{\varphi}(x, y)$ is unique only up to a constant, causing obvious problems with these definitions.

## Generalized Equivalence Theory

Koebke [48] demonstrated that if the exact heterogeneous solution $\widetilde{\phi}$ was known then relaxing the continuity assumption of the edge averaged scalar flux would facilitate the preservation of all node-integrated reaction rates and edge averaged net currents. In addition, the coarse-scale edge average of the heterogeneous scalar flux, may be recovered from the corresponding jump conditions. Naturally, in practice $\widetilde{\phi}$ is not known and only an approximate equivalence is possible. Two very similar approximate techniques have

[^6]emerged based on this equivalence theory, and we present an overview of Smith et al.'s [66] approximation of generalized equivalence theory.

We begin by writing the balance equation over a coarse-mesh cell $C_{i, j}$ for the heterogeneous problem in the form,

$$
\begin{align*}
\Delta y_{j}\left\{\widetilde{J}_{i+\frac{1}{2}, j}^{-}-\widetilde{J}_{i-\frac{1}{2}, j}^{+}\right\} & +\Delta x_{i}\left\{\widetilde{J}_{i, j+\frac{1}{2}}^{-}-\widetilde{J}_{i, j-\frac{1}{2}}^{+}\right\}  \tag{4.13}\\
& +\left(\widetilde{\Sigma_{t}}\right)_{i, j} \widetilde{\phi}_{i, j} \Delta x_{i} \Delta y_{j}=\frac{1}{k} \chi \nu\left({\widetilde{\Sigma_{f}}}^{)_{i, j}} \widetilde{\phi}_{i, j} \Delta x_{i} \Delta y_{j}\right.
\end{align*}
$$

where we have employed the following definitions of the node integrated flux and interaction coefficients,

$$
\begin{align*}
\tilde{\phi}_{i, j} & \equiv \frac{1}{\Delta x_{i} \Delta y_{j}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \widetilde{\phi}(x, y) d x d y  \tag{4.14a}\\
\left(\widetilde{\Sigma_{\alpha}}\right)_{i, j} & \equiv \frac{1}{\widetilde{\phi}_{i, j} \Delta x_{i} \Delta y_{j}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \widetilde{\Sigma_{\alpha}}(x, y) \widetilde{\phi}(x, y) d x d y \tag{4.14b}
\end{align*}
$$

$(\alpha=t$ or $\alpha=f)$. We note that this mean value definition of $\left(\widetilde{\Sigma_{a}}\right)_{i, j}$ also appears in (3.7), where it arose naturally during the derivation of the cell-balance equation. The edge averaged currents also follow the definitions of Chapter 3 and are written

$$
\widetilde{J}_{i \pm \frac{1}{2}, j}^{\mp} \equiv \frac{1}{\Delta y_{j}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \widetilde{\mathbf{J}}\left(x_{i \pm \frac{1}{2}}^{\mp}, y\right) \cdot \mathbf{n}_{x} d y
$$

with analogous definitions of $\widetilde{J}_{i, j \pm \frac{1}{2}}^{\mp}$. Similarly, we can write the balance equation over the coarse-mesh cell for the homogenized problem

$$
\begin{align*}
\Delta y_{j}\left\{\widehat{J}_{i+\frac{1}{2}, j}^{-}-\widehat{J}_{i-\frac{1}{2}, j}^{+}\right\} & +\Delta x_{i}\left\{\widehat{J}_{i, j+\frac{1}{2}}^{-}-\widehat{J}_{i, j-\frac{1}{2}}^{+}\right\}  \tag{4.15}\\
& +\left(\widehat{\Sigma_{t}}\right)_{i, j} \widehat{\phi}_{i, j} \Delta x_{i} \Delta y_{j}=\frac{\chi \nu}{k}\left(\widehat{\Sigma_{f}}\right)_{i, j} \widehat{\phi}_{i, j} \Delta x_{i} \Delta y_{j}
\end{align*}
$$

where we note that we have utilized the assumption that the homogenized interaction coefficients $\widehat{\Sigma_{\alpha}}$ are constant over each cell and all other definitions are analogous to the heterogeneous case. Assuming that the equivalence of the edge averaged normal currents
[i.e. enforcing (4.5)] is satisfied and selecting $\left(\widehat{\Sigma_{\alpha}}\right)_{i, j}=\left(\widetilde{\Sigma_{\alpha}}\right)_{i, j}$, subtracting (4.15) from (4.13) reveals that the node-average has been preserved,

$$
\widehat{\phi}_{i, j}=\int_{C_{i, j}} \widehat{\phi}(x, y) d x d y=\int_{C_{i, j}} \widetilde{\phi}(x, y) d x d y=\widetilde{\phi}_{i, j}
$$

and hence we preserve the reaction rates (4.3).
Thus, success of the equivalence process lies with the assumed equivalence of the edge integrated normal currents. Returning to the traditional nodal methodology of transverse integration we write the transverse integrated ODEs [cf. equation (3.16)] for the homogenized problem,

$$
\begin{align*}
&-\frac{\partial}{\partial x}\left\{\widehat{D}_{i, j} \frac{\partial}{\partial x} \widehat{\phi}_{j}(x)\right\}+\left\{\left(\widehat{\Sigma_{t}}\right)_{i, j}-\frac{1}{\lambda} \nu \chi\left(\widehat{\Sigma_{f}}\right)_{i, j}\right\} \widehat{\phi}_{j}(x)=\widehat{\mathcal{L}}_{i, j}^{x}(x)  \tag{4.16a}\\
&-\frac{\partial}{\partial y}\left\{\widehat{D}_{i, j} \frac{\partial}{\partial y} \widehat{\phi}_{i}(y)\right\}+\left\{\left(\widehat{\Sigma_{t}}\right)_{i, j}-\frac{1}{\lambda} \nu \chi\left(\widehat{\Sigma_{f}}\right)_{i, j}\right\} \widehat{\phi}_{i}(y)=\widehat{\mathcal{L}}_{i, j}^{y}(y) \tag{4.16b}
\end{align*}
$$

where $\widehat{\phi}_{j}(x), \widehat{\phi}_{i}(y)$ are the transverse averages defined in (3.2) and the pseudo source terms have been written in the form,

$$
\begin{aligned}
& \widehat{\mathcal{L}}_{i, j}^{x}(x)=\frac{1}{\Delta y_{j}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \widehat{D}_{i, j} \frac{\partial^{2} \widehat{\phi}}{\partial y^{2}} d y=\widehat{J}_{j+\frac{1}{2}}^{-}(x)-\widehat{J}_{j-\frac{1}{2}}^{+}(x) \\
& \widehat{\mathcal{L}}_{i, j}^{y}(y)=\frac{1}{\Delta x_{i}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \widehat{D}_{i, j} \frac{\partial^{2} \widehat{\phi}}{\partial x^{2}} d x=\widehat{J}_{i+\frac{1}{2}}^{-}(y)-\widehat{J}_{i-\frac{1}{2}}^{+}(y)
\end{aligned}
$$

It is at this point that we are forced to assume that the exact homogeneous solution $\widehat{\phi}(x, y)$ is known such that these transverse leakages may be treated as known.

Thus, all that remains is the selection of the appropriate boundary conditions. Recalling that the objective of the analysis is the preservation of the edge integrated normal currents we choose these as the boundary conditions,

$$
\widehat{J}_{i \pm \frac{1}{2}, j}^{\mp}=\widetilde{J}_{i \pm \frac{1}{2}, j}^{\mp} \quad \widehat{J}_{i, j \pm \frac{1}{2}}^{\mp}=\widetilde{J}_{i, j \pm \frac{1}{2}}^{\mp}
$$

This choice not only satisfies our objective but automatically guarantees the continuity of the edge integrated normal currents within the global homogenized problem. Moreover, the homogenized diffusion coefficient is arbitrary, as we have satisfied all desired relations between the heterogeneous and homogeneous problems. This indicates an unfortunate inconsistency between the formulation of the continuous homogenized problem and the homogenized discretization. The manifestation of this inconsistency is the loss of continuity in the edge averaged scalar flux of the discrete homogenized problem.

Allowing the discontinuity of the edge-averaged scalar flux introduces the following jump conditions,

$$
f_{i, j}^{(x)-} \equiv \frac{\widehat{\phi}_{i-\frac{1}{2}, j}^{+}}{\widetilde{\phi}_{i-\frac{1}{2}, j}} \quad f_{i, j}^{(x)+} \equiv \frac{\widehat{\phi}_{i+\frac{1}{2}, j}^{-}}{\widetilde{\phi}_{i-\frac{1}{2}, j}}
$$

with analogous conditions defining $f_{i, j}^{(y) \pm}$. Once again application of the equivalence technique requires an approximation of $\widetilde{\phi}(x, y)$, for which they chose the local assembly solution $\widetilde{\varphi}(x, y)$. This approximation implies that the equivalence theory cross-sections and diffusion coefficient are precisely the FWCs,

$$
\left(\widehat{\Sigma_{\alpha}}\right)_{i, j}=\left(\widehat{\Sigma_{\alpha}}\right)_{i, j}^{(f w)} \quad \widehat{D}_{i, j}=\widehat{D}_{i, j}^{(f w)}
$$

The crucial difference is that the specific definition of $\widehat{D}_{i, j}$ is irrelevant as the computation of the discontinuity factors represent all necessary degrees of freedom. However, evaluation of the discontinuity factors requires an additional approximation. Specifically, we solve the local assembly problem [i.e. Definition 4.1.3] for the homogenized assembly and note that with constant cross sections and zero net current boundary conditions the solution $\widehat{\varphi}(x, y)$ is spatially constant,

$$
\widehat{\varphi}_{i+\frac{1}{2}, j}^{-}=\widehat{\varphi}_{i-\frac{1}{2}, j}^{+}=\widehat{\varphi}_{i+\frac{1}{2}, j}=\widehat{\varphi}_{i, j-\frac{1}{2}}^{+}=\widehat{\varphi}_{i, j}
$$

but we have $\widetilde{\varphi}_{i, j}=\widehat{\varphi}_{i, j}$ hence we define the assembly discontinuity factors (ADFs) as,

$$
f_{i, j}^{(x) \pm}=\frac{\widetilde{\varphi}_{i, j}}{\widetilde{\varphi}_{i \pm \frac{1}{2}, j}^{\mp}} \quad f_{i, j}^{(y) \pm}=\frac{\widetilde{\varphi}_{i, j}}{\widetilde{\varphi}_{i, j \pm \frac{1}{2}}^{\mp}}
$$

The use of ADFs in GET offers some improvement over the direct application of the FWCs in that if the errors associated with the zero net current assumption are small reaction rates will be very accurate. But clearly the same problems arise here as well: the errors in this assumption are difficult to predict and may lead to significant errors in the solution.

### 4.2 Asymptotic Homogenization Techniques

There are several asymptotic methods which consider the problem of homogenization in periodic or nearly periodic composite structures (cf. [11, 62]). Perhaps the most popular of these is based on a two-scale treatment that is facilitated by the introduction of a fine-scale parameter, $\epsilon$. Specifically one considers the coarse-scale dependence $\mathbf{r}$, and the fine-scale dependence $\rho=\mathrm{r} / \epsilon$, to be independent variables. Then the second order form of the heterogeneous problem may be written,

$$
\mathrm{L}_{\epsilon}\left(\phi_{\epsilon}\right)=Q(\mathbf{r})
$$

$\phi_{\epsilon}$ subject to suitable BCs
where $L_{\epsilon}$ is uniformly elliptic in $\epsilon$. Note that to simplify our discussion we have made the common assumption that the source is independent of the fine scale.

An asymptotic expansion of $\phi_{\epsilon}$ is then introduced,

$$
\phi_{\epsilon}=\phi_{0}(\mathbf{r}, \boldsymbol{\rho})+\epsilon \phi_{1}(\mathbf{r}, \boldsymbol{\rho})+\epsilon^{2} \phi_{2}(\mathbf{r}, \boldsymbol{\rho})+\cdots
$$

such that one may investigate the behaviour of $\phi_{\epsilon}$ as $\epsilon \rightarrow 0^{+}$. For the fine-scale periodic and near periodic cases of linear second-order elliptic operators, one finds that (in a weak
sense) $\phi_{\epsilon}$ converges to $\phi_{0}$ as $\epsilon \rightarrow 0^{+}$, where $\phi_{0}$ is the solution of the homogenized problem,

$$
\begin{aligned}
\widehat{\mathrm{L}}\left(\phi_{0}\right) & =Q(\mathbf{r}) \\
\phi_{0} & \text { subject to corresponding BCs }
\end{aligned}
$$

and the second-order elliptic operator, $\widehat{L}$, is referred to as the homogenized operator. The coefficients of the homogenized problem vary on the coarse scale in the nearly periodic case, and are in fact constant for fine-scale periodic problems. It is important to note that in general, $\hat{L}$ is not related to the operator $L_{o}$ that arises from evaluating the limit $\epsilon \rightarrow 0^{+}$of the problem's coefficients.

In the following discussion we review the two-scale asymptotic analysis of Bensoussan et al. [11] and, in particular, discuss the relevant results for the fine-scale periodic and nearly (i.e. non-uniformly) periodic cases. An excellent introduction to these concepts may be found in [42].

### 4.2.1 Multidimensional Periodic Analysis

We consider the fine-scale periodic family of linear differential operators,

$$
\begin{equation*}
\mathrm{L}_{\epsilon}=\nabla_{\mathbf{r}} \cdot\left(\mathcal{D}\left(\frac{\mathbf{r}}{\epsilon}\right) \nabla_{\mathbf{r}}\right)+\Sigma_{a}\left(\frac{\mathbf{r}}{\epsilon}\right) \tag{4.17}
\end{equation*}
$$

where $D(\boldsymbol{\rho}), \boldsymbol{\rho}=\mathbf{r} / \epsilon$ is a diffusion tensor defined such that $\mathrm{L}_{\epsilon}$ is uniformly elliptic in $\epsilon$. In addition utilizing the fine-scale periodicity we introduce the following local fine-scale diffusion problem,

$$
\begin{array}{r}
-\nabla_{\rho} \cdot \mathcal{D}(\boldsymbol{\rho}) \nabla_{\rho} \phi=0, \quad \text { in } F \\
\phi \text { periodic in } F \tag{4.18b}
\end{array}
$$

The corresponding weak variational formulation (cf. Section 5.1.1) may be written as:

Chapter 4. Homogenization
find $\phi \in W(F)$ such that

$$
\begin{equation*}
a_{\boldsymbol{\rho}}(\phi, \psi) \equiv\left(D(\boldsymbol{\rho}) \nabla_{\boldsymbol{\rho}} \phi, \nabla_{\boldsymbol{\rho}} \psi\right)=0 \quad \forall \psi \in W(F) \tag{4.19}
\end{equation*}
$$

where the space of admissible functions is $W(F) \equiv\left\{\phi \mid \phi \in H^{1}(F), \phi\right.$ periodic $\}$. The results of a two-scale asymptotic analysis [11] are summarized in the following theorem.

Theorem 4.1. For the globally periodic problem with $\mathrm{L}_{\epsilon}$ given in (4.17) the homogenized operator may be written,

$$
\widehat{\mathrm{L}}=\nabla_{\mathbf{r}} \cdot\left(\widehat{\mathcal{D}} \nabla_{\mathbf{r}}\right)+\widehat{\Sigma_{a}}
$$

with the homogenized absorption cross-section given by,

$$
\begin{equation*}
\widehat{\Sigma_{a}}=\frac{1}{\Delta \boldsymbol{\rho}} \int_{F} \Sigma_{a}(\boldsymbol{\rho}) d \boldsymbol{\rho} \tag{4.20a}
\end{equation*}
$$

where $F$ is a fine-scale cell and $\Delta \rho$ its corresponding volume. In addition, the elements of the homogenized diffusion tensor are written,

$$
\begin{equation*}
\mathcal{D}^{(i, j)}=\frac{1}{\Delta \boldsymbol{\rho}} a_{\boldsymbol{\rho}}\left(\chi^{j}(\boldsymbol{\rho})-\rho_{j}, \chi^{i}(\boldsymbol{\rho})-\rho_{i}\right) \tag{4.20b}
\end{equation*}
$$

with $\chi^{j}(\boldsymbol{\rho})$ the solution of,

$$
\begin{equation*}
a_{\rho}\left(\chi^{j}(\boldsymbol{\rho})-\rho_{j}, \psi\right)=0 \quad \forall \psi \in W(F) \tag{4.21}
\end{equation*}
$$

and the bilinear form, $a_{\rho}(\cdot, \cdot)$ is given in (4.19).

We note that (4.21) is a local elliptic boundary value problem that must be solved to obtain the homogenized diffusion tensor. This situation is very similar to that found in the equivalence analysis of the previous section. However, an apparent difference is that the homogenization of the absorption coefficient (4.20a) does not depend on the fine-scale solution (i.e. $\chi^{j}$ in this case), and instead is simply the integral average. But
here a strict fine-scale periodicity has been assumed, yielding $\phi_{0}(\mathbf{r}, \boldsymbol{\rho})=\phi_{0}(\mathbf{r})$. Thus substitution of $\phi_{\epsilon}$ into (4.4) gives $\left(\widehat{\Sigma_{a}}\right)_{i, j}=\widehat{\Sigma_{a}}+\mathcal{O}(\epsilon)$, indicating that these methods are consistent and that the strict equivalence approach implicitly captures higher order terms. Unfortunately, as we noted in Section 4.1.1 enforcing this strict equivalence is not possible in practice. Moreover, the error associated with the asymptotic treatment is bounded, $\left\|\phi_{\epsilon}-\phi_{0}\right\|_{L^{\infty}(\Omega)} \leq c \epsilon$.

### 4.2.2 Multidimensional Nearly Periodic Analysis

In many applications a strict fine-scale periodicity does not exist, and hence we consider the extension to a fine-scale non-uniformly periodic family of linear differential operators,

$$
\begin{equation*}
\mathrm{L}_{\epsilon}=\nabla_{\mathbf{r}} \cdot\left(\mathcal{D}\left(\mathbf{r}, \frac{\mathbf{r}}{\epsilon}\right) \nabla_{\mathbf{r}}\right)+\Sigma_{a}\left(\mathbf{r}, \frac{\mathbf{r}}{\epsilon}\right) \tag{4.22}
\end{equation*}
$$

where $D(\mathbf{r}, \boldsymbol{\rho})$ is a diffusion tensor defined such that $\mathrm{L}_{\varepsilon}$ is uniformly elliptic in $\epsilon$.
Once again we summarize the relevant results of [11].
Theorem 4.2. For the globally periodic problem with $\mathrm{L}_{\epsilon}$ given in (4.22) the homogenized operator may be written,

$$
\widehat{\mathrm{L}}=\nabla_{\mathbf{r}}\left(\widehat{\mathcal{D}}(\mathbf{r}) \cdot \nabla_{\mathbf{r}}\right)+\widehat{\Sigma_{a}}(\mathbf{r})
$$

The homogenized absorption cross-section is given by,

$$
\widehat{\Sigma_{a}}(\mathbf{r})=\frac{1}{\Delta \boldsymbol{\rho}} \int_{F} \Sigma_{a}(\mathbf{r}, \boldsymbol{\rho}) d \boldsymbol{\rho}
$$

and the elements of the homogenized diffusion tensor are written,

$$
\mathcal{D}^{(i, j)}(\mathbf{r})=\frac{1}{\Delta \boldsymbol{\rho}} a_{\rho}\left(\chi^{j}(\mathbf{r}, \boldsymbol{\rho})-\rho_{j}, \chi^{i}(\mathbf{r}, \boldsymbol{\rho})-\rho_{i}\right)
$$

with $\chi^{j}(\mathbf{r}, \boldsymbol{\rho})$ the solution of,

$$
\begin{equation*}
-\nabla_{\boldsymbol{\rho}} \cdot\left(\mathcal{D}(\mathrm{r}, \boldsymbol{\rho}) \nabla_{\boldsymbol{\rho}}\right)\left(\chi^{j}(\mathrm{r}, \boldsymbol{\rho})-\rho_{j}\right)=0 \tag{4.23}
\end{equation*}
$$

subject to periodic boundary conditions.

Recall that in the fine-scale periodic case the homogenized diffusion tensor is defined by only a single elliptic boundary value problem, whereas in the nearly periodic case we have an elliptic boundary problem (4.23) for each point $\mathbf{r}$, in the global domain $\Omega$. In general, a practical application of this result considers the solution of (4.23), and hence the computation of the homogenized diffusion tensor, at only a few points within each coarse-mesh cell. A piecewise smooth function may then be defined to interpolate these values, once again suggesting the generalization of nodal discretizations to piecewise smooth coefficients [cf. Section 3.6].

### 4.3 Multigrid Homogenization

The multigrid solution of second order elliptic PDEs has been studied extensively since Brandt's [13] seminal paper appeared in 1977. Several robust multigrid methods have been developed which, even in the presence of very rough coefficients, demonstrate optimal convergence rates in many practical situations. The efficiency of any multigrid algorithm depends heavily on the coarse grid operators' approximation of the coarse-scale properties of the fine-grid discretization. This strongly suggests that a robust multigrid algorithm for solving such heterogeneous problems [equation (4.1)] must employ coarse grid operators that are representative of the discretization of the corresponding homogenized problem [equation (4.2)]. We asked the following question in [24]: what information, if any, regarding the homogenized diffusion coefficient (or possibly tensor) could be extracted from these coarse grid operators? The results, summarized in this section, are very promising. Recently Knapek [46, 47] addressed this question in a complementary manner and we comment on his approach in the following discussion.

### 4.3.1 Obtaining the Diffusion Tensor

In the periodic case the asymptotic analysis of problems with rapidly varying coefficients demonstrated that the homogenization process leads to a full diffusion tensor. Intuitively, this is the case if we consider an essentially one-dimensional heterogeneous material (e.g. shale) which is not aligned with the coordinate axes. Clearly the primary axes of the homogenized diffusion problem should be aligned with those of the essentially one-dimensional structure. In general, this is only possible if the homogenization yields a full diffusion tensor. Thus, our objective is to develop a local technique to compute a $2 \times 2$ constant tensor for each cell of a coarse grid, based solely on the coarse grid operator itself. To this end we adopt the compass based notation displayed in Figure 4.1 (i.e. $S_{l, k}^{(*)}$ denotes the stencil weight in the direction "*", negative signs are treated explicitly), and generate the 9-point coarse grid operators with the operator-induced variational coarsening of Dendy's black box multigrid [23]. Then observing that any diffusion stencil may be viewed as a superposition of currents, an analysis of the current passing through a cell centred reference frame yields the following results.


Figure 4.1: A compass based definition of an arbitrary 9-point stencil

Theorem 4.3. Consider the conforming bilinear finite element stencil for conservative anisotropic diffusion subject to periodic boundary conditions and under the assumption of a piecewise constant diffusion tensor. A second-order approximation of the coarse-scale diffusion tensor $\mathcal{D}_{i, j}$ on $\Omega_{i, j}$ is given by,

$$
\widehat{\mathcal{D}_{i, j}}=\left[\begin{array}{cc}
\frac{\Delta x}{\Delta y}\left\{\bar{S}_{l, k}^{E}+S_{l, k}^{N E}+S_{l+1, k}^{N W}\right\} & \left(S_{l, k}^{N E}-S_{l+1, k}^{N W}\right)  \tag{4.24}\\
\left(S_{l, k}^{N E}-S_{l+1, k}^{N W}\right) & \frac{\Delta y}{\Delta x}\left\{\bar{S}_{l, k}^{N}+S_{l, k}^{N E}+S_{l+1, k}^{N W}\right\}
\end{array}\right]
$$

where $(\Delta x, \Delta y)$ denotes the grid spacing, $(l, k)=\left(i-\frac{1}{2}, j-\frac{1}{2}\right)$ and we have defined,

$$
\begin{aligned}
& \bar{S}_{l, k}^{E}=\frac{1}{2}\left(S_{l, k}^{E}+S_{l, k+1}^{E}\right) \\
& \bar{S}_{l, k}^{N}=\frac{1}{2}\left(S_{l, k}^{N}+S_{l+1, k}^{N}\right)
\end{aligned}
$$

For a strictly constant diffusion tensor (i.e. $\mathcal{D}_{i, j}=\mathcal{D}$ ), equation (4.24) is an exact expression.

Proof. A detailed proof may be found in [24].
Corollary 4.1. Given a coarse grid operator generated by operator-induced coarsening of the conforming bilinear stencil described in Theorem 4.3, the diffusion tensor $\widehat{\mathcal{D}_{i, j}}$ given in (4.24) is a second-order approximation of the multigrid homogenized diffusion tensor. Moreover, if the stencil weights are spatially constant then (4.24) gives the multigrid homogenized tensor exactly.

Proof. For this purpose the two key properties of the 9-point stencil are that it is symmetric and conservative. Operator-induced coarsening preserves these properties, and hence, the proof follows from that of Theorem 4.3.

Remark 4.1. We note that, in Theorem 4.3, assuming a conforming bilinear basis ensures that the only error is associated with $\bar{S}_{l, k}^{E}$ and $\bar{S}_{l, k}^{N}$. However, in general a second-order error also arises with the approximation of first derivatives (i.e. for the current) at the
cell centre. These two errors manifest themselves very differently. Specifically, $\bar{S}_{l, k}^{E}$ and $\bar{S}_{l, k}^{N}$ introduce a second order error in $\widehat{\mathcal{D}_{i, j}}$ which vanishes when the stencil is spatially constant. In contrast the errors associated with the solution do not influence $\widehat{\mathcal{D}_{i, j}}$ but lead to an error in the resulting stencil. Numerically this error has been demonstrated to be second-order in [47], although a proof remains elusive.

Based on Corollary 4.1 our multigrid homogenization algorithm consists of first defining the conforming bilinear discretization on a fine grid capable of resolving the fine-scale structure of the problem. Then, the operator-induced variational coarsening of black box multigrid is applied until a coarse-mesh spacing suitable for efficient computation is reached. From this coarse-mesh problem we compute the cell-based multigrid homogenized diffusion tensors using the cell-based expression (4.24). An alternative vertex-based approach is considered in [47] which inverts a $9 \times 9$ system that is defined over a group of 4 cells. These methods result in equivalent homogenized diffusion tensors when the stencil is spatially constant, in which case it is natural to assume that 4 neighbouring cells will have identical properties. However, this assumption may be too restrictive in a more general setting and hence we feel that the local technique is preferable.

Remark 4.2. Note that we have not entirely escaped the dependence that the homogenization process has on the heterogeneous solution. Instead, the multigrid homogenization approach shifts this information from the actual heterogeneous solution to the finite dimensional subspace of the discrete solution. Choosing a different fine grid discretization corresponds to choosing a different subspace and will yield different results. This is evident if we consider a simple constant coefficient diffusion problem and employ the standard 5-point vertex based discrete operator. The homogenized coefficients determined with multigrid homogenization will be in error, approaching the correct diffusion coefficient only in the limit of infinite coarsening. This suggests that for a given discretization
it may be necessary to utilize a compatible sequence of subspaces in the definition of the coarse grid operators and hence the homogenization process.

### 4.3.2 A Numerical Homogenization Suite

Unfortunately, there are very few exact theoretical homogenization results, with the vast majority restricted to one-dimensional problems. However, based on the asymptotic analysis outlined in Section 4.2, numerical computations may be performed for the case of periodic boundary conditions which converge to the homogenized diffusion tensor [12]. Thus, we have conducted a series of tests in the periodic case to gauge the potential of this new technique. To accomplish the homogenization of a representative cell we acknowledge that the smallest plausible computational grid is $3 \times 3$ and hence define the corresponding problem as a $3 \times 3$ tiling of representative cells. A fine grid is selected which adequately represents the fine-scale structure of the representative cell and the operator induced coarsening is applied. Finally, following Corollary 4.1, the homogenized diffusion tensors are given by (4.24).

Several rudimentary tests, which are not included here, were conducted to verify the proper functioning of the code [24]. These included, the demonstration that the algorithm reproduces a constant tensor and that the homogenization of essentially one-dimensional problems is exact. Unfortunately it was also revealed that this homogenization technique is not infallible: $\widehat{\mathcal{D}_{i, j}}$ is simply the arithmetic average for a checkerboard problem. We believe that this pathology, which is linked to the lumping inherent in operator induced coarsening, is an isolated example that is overshadowed by the method's performance in other cases. In particular, the tests that we present in the following subsections, which characterize the dependence of the homogenized diffusion tensor for a representative cell $\Omega_{r}$ on the shape and diffusivity of an internal inhomogeneity $\Omega_{1}$ (i.e. $\Omega_{1} \in \Omega_{r}$ ) [12], clearly highlight the strengths of this new homogenization technique.

## Shape Dependence

An evaluation of the geometric dependence of the homogenized diffusion tensor is accomplished with the set of representative cells shown in Figure 4.2(a)-(c). The diffusion tensor on these representative cells is defined by,

$$
\mathcal{D}(x, y)=\left\{\begin{array}{cl}
I_{2} & \forall(x, y) \in \Omega_{0} \\
10 I_{2} & \forall(x, y) \in \Omega_{\mathbf{1}}
\end{array}\right.
$$

where $\Omega_{0}=\Omega_{r} \backslash \Omega_{1}$. In all cases the area of $\Omega_{1}$ is $1 / 4$, implying the standard averaging techniques (e.g. arithmetic, harmonic) would yield the same homogenized diffusion coefficient independently of the inhomogeneity's shape. Based on their symmetry we anticipate that the homogenized diffusion tensors will in fact be scalar multiples of the identity and this was confirmed numerically. A comparison of the blackbox homogenized diffusion coefficient with Bourgat's asymptotic computation is provided in Table 4.1, where percentage differences, relative to the square inhomogeneity, are also included. These results demonstrate that the relative sensitivity of blackbox homogenization to the shape of the heterogeneity is similar to the rigorous treatment of Bourgat. In a direct comparison we find that the blackbox results consistently overestimate the asymptotic value by approximately $2-3 \%$. This is impressive when a commonly employed alternative such as the two-dimensional harmonic average underestimates the asymptotic value by approximately $17 \%$.

## Dependence on the Diffusivity

The second case is simply a square inhomogeneity [Figure 4.3] defined by,

$$
\mathcal{D}(x, y)=\left\{\begin{array}{cc}
I_{2} & \forall(x, y) \in \Omega_{0} \\
\lambda I_{2} & \forall(x, y) \in \Omega_{1}
\end{array}\right.
$$

to evaluate the dependence of the homogenized diffusion tensor on the parameter $\lambda$. Symmetry once again guarantees that the homogenized diffusion tensor is simply a scalar


Figure 4.2: Three inhomogeneities with the same area of $1 / 4$, but different shape.

Table 4.1: A comparison of the shape dependence of the diffusivity relative to the square (i.e. the percent relative difference - \% RD) is presented for the results of Bourgat [12] and the black box homogenization.

| Shape | Bourgat | \% RD | Black Box | \% RD |
| :--- | :---: | :---: | :---: | :---: |
| Square | 1.548 | - | 1.598 | - |
| Disk | 1.516 | -2.06 | 1.563 | -2.19 |
| Lozenge | 1.573 | +1.69 | 1.608 | +0.63 |

multiple of the identity. The multigrid homogenization results are presented in Figure 4.4 along with the asymptotic results of Bourgat, the aforementioned 2D result of Selengut (4.7) and the commonly used means,

$$
\begin{aligned}
& \mathcal{D}^{(a m)}=\int_{0}^{1} \int_{0}^{1} \mathcal{D}(x, y) d x d y=\frac{1}{9}(\lambda+8) I_{2} \\
& \mathcal{D}^{(h m)}=\left[\int_{0}^{1} \int_{0}^{1}[\mathcal{D}(x, y)]^{-1} d x d y\right]^{-1}=\frac{9 \lambda}{(1+8 \lambda)} I_{2}
\end{aligned}
$$

We note the excellent agreement of the blackbox homogenized diffusion coefficient with the asymptotic results over 8 orders of magnitude in $\lambda$. We also observe the catastrophic failure of the harmonic mean for $\lambda \rightarrow 0^{+}$is contrasted by an overestimation of approximately $10 \%$ in the arithmetic mean. Whereas for $\lambda \rightarrow+\infty$ the harmonic mean yields approximately a $10 \%$ underestimation, while the arithmetic mean grows linearly, displaying an arbitrarily large error. Having been derived in an infinite cylindrical geometry the
result of Selengut (4.7) is very interesting to include because it demonstrates the effect of the geometry. Here we see that for $\lambda \rightarrow 0^{+}$the limiting value is very insensitive to the geometry while for $\lambda \rightarrow+\infty$ it is quite sensitive, underestimating the final result by approximately $4 \%$. Nevertheless, it is the only closed form expression to mimic the behaviour of the numerical calculations over the entire range of $\lambda$.


Figure 4.3: A square inhomogeneity with diffusivity, $\lambda$ and an area of $1 / 9^{t h}$.


Figure 4.4: A comparison of homogenized diffusivities.

## A Nonsymmetric Inhomogeneity

Consider the L shaped region shown in Figure 4.5 with the diffusion tensor,

$$
\mathcal{D}(x, y)=\left\{\begin{array}{cl}
I_{2} & \forall(x, y) \in \Omega_{0} \\
10 I_{2} & \forall(x, y) \in \Omega_{1}
\end{array}\right.
$$

from which the asymptotic computation yields [12],

$$
\mathcal{D}_{\mathcal{H}}^{(a s)}=\left[\begin{array}{cc}
1.915 & -0.101 \\
-0.101 & 1.915
\end{array}\right]=Q\left[\begin{array}{cc}
2.016 & 0 \\
0 & 1.814
\end{array}\right] Q^{T}
$$

where the matrix of eigenvectors $Q$ is given by,

$$
Q=\frac{1}{\sqrt{2}}\left[\begin{array}{rr}
-1 & 1 \\
1 & 1
\end{array}\right]
$$

and defines the principal axes of diffusion, in this case a rotation of $45^{\circ}$. Blackbox homogenization also gives a full tensor

$$
\mathcal{D}_{\mathcal{H}}^{(b b)}=\left[\begin{array}{cc}
1.959 & -0.153 \\
-0.153 & 1.959
\end{array}\right]=Q\left[\begin{array}{cc}
2.113 & 0 \\
0 & 1.806
\end{array}\right] Q^{T}
$$

Moreover, we remarkably obtain the exact principal axes of diffusion in this case. The only error is the scaling in each of these directions, approximately $5 \%$ and $0.4 \%$ respectively.


Figure 4.5: The homogenization of an $L$ shaped inhomogeneity leads to a dense tensor.

## Chapter 5

## Finite Element Methods

The development of modern nodal methods is based almost entirely on physical arguments. However, traditionally no assurance was given of a correspondingly solid mathematical foundation. Over the years there has been a continuing effort to remedy this situation with definite advances in some areas. One such vein of research, pursued by Hennart [35, 36], has established the equivalence of certain nodal and nonconformal finite element methods (FEMs). Equivalence results such as this form important contributions to the mathematical foundation of nodal methods, because the related FEM theoretical results are immediately applicable to the corresponding nodal methods. In particular, stability and optimality results for the corresponding nodal discretization may be proven in this way.

We explore this equivalence for the lowest-order nodal methods that were introduced in Chapter 3 and, in particular, we utilize a one-dimensional study to highlight information that is present within these nonconformal FEMs but was previously obscured. For example, in [36] a nonstandard Radau quadrature is suggested for the treatment of the mass matrix, however we show that a simple lumping [69] is all that is required in the polynomial case [Section 5.2.2], and moreover, the exact evaluation is necessary in the analytic case [Section 5.2.3]. The primary affect of this lumping procedure is to restore the cell-based balance relation. In addition, motivated by the well known equivalences between some nonconformal and mixed-hybrid finite element methods [14], Hennart et al. [39] have also studied the equivalence between mixed-hybrid FEMs and various nodal
methods. In Section 5.3 we demonstrate this equivalence for the polynomial case and discuss the difficulty associated with extending it to the analytic case. Specifically, we show that although it is possible to cast the constant-constant NIM in the indefinite form characteristic of mixed-hybrid FEMs, the appropriate basis functions are elusive.

### 5.1 Primal Conforming FEM

In its second-order form the finite element treatment of the one-group diffusion equation commonly employs a piecewise bilinear basis. The use of this'popular basis is motivated by two essential properties, namely, it is a low order polynomial that leads to a second order discretization and it is conforming. However, this basis performs poorly for problems with piecewise constant coefficients that exhibit severe variations. Thus, to highlight the strengths of the nodal discretizations, and hence the equivalent FEMs, we review this conforming case.

### 5.1.1 Variational Formulation

For completeness we briefly review the variational formulation and establish the notation that we employ in the following sections. Consider the second order divergence form of the one-group diffusion equation in $\mathbb{R}^{n}$,

$$
-\nabla \cdot(D(\mathbf{r}) \nabla \phi)+\Sigma_{a}(\mathbf{r}) \phi=Q(\mathbf{r}) \quad \forall \mathbf{r} \in \Omega
$$

with $D(\mathbf{r}) \geq D_{\text {min }}>0$ and $\Sigma_{a}(\mathbf{r}) \geq 0$. To avoid unnecessary complication we will restrict this theoretical discussion to homogeneous Dirichlet boundary conditions,

$$
\phi(\mathbf{r})=0, \quad \forall \mathbf{r} \in \partial \Omega
$$

It is well known (e.g. [15]) that the weak variational form of this diffusion problem may be written as:
find $\phi \in V$ such that,

$$
\begin{equation*}
a(\phi, v) \equiv(D \nabla \phi, \nabla v)+\left(\Sigma_{a} \phi, v\right)=(Q, v), \quad \forall v \in V \tag{5.1}
\end{equation*}
$$

The space of admissible functions $V \equiv H_{0}^{1}(\Omega)$,

$$
\begin{aligned}
H^{1}(\Omega) & =\left\{v \mid\|v\|_{1}<\infty\right\} \\
H_{0}^{1}(\Omega) & =\{v \mid v=0 \forall \mathbf{r} \in \partial \Omega\} \cap H^{1}(\Omega)
\end{aligned}
$$

where $\|v\|_{1}=\left(\|\nabla v\|_{0}^{2}+\|v\|_{0}^{2}\right)^{\frac{1}{2}}$ and $\|v\|_{0}=\|v\|_{L_{2}(\Omega)}$. The conformal approach to numerically approximating this variational problem is to replace $V$ by some finite-dimensional subspace $V_{h}$ (i.e. $V_{h} \subset V$ ). A basis for $V_{h}$ is chosen,

$$
V_{h}=\operatorname{span}\left\{\Phi_{1}, \Phi_{2}, \ldots, \Phi_{N}\right\} \subset V
$$

such that the approximate solution $\phi_{h}$ is given by,

$$
\phi_{h}(\mathbf{r})=\sum_{i=1}^{N} \alpha_{i} \Phi_{i}(\mathbf{r})
$$

In the Galerkin approximation the corresponding linear system that determines the coefficients $\alpha_{i}$ is obtained by direct substitution into (5.1),

$$
\begin{align*}
a\left(\sum_{i=1}^{N} \alpha_{i} \Phi_{i}, \Phi_{j}\right) & =\left(Q, \Phi_{j}\right), & j=1, \ldots, N \\
\Rightarrow \sum_{i=1}^{N} \alpha_{i} a\left(\Phi_{i}, \Phi_{j}\right) & =\left(Q, \Phi_{j}\right) & j=1, \ldots, N
\end{align*}
$$

which is a linear system $\mathcal{A} \boldsymbol{\alpha}=\boldsymbol{b}$ where the components of the load vector $\boldsymbol{b}$ are simply,

$$
b_{j}=\left(Q, \Phi_{j}\right)
$$

while the matrix $\mathcal{A}=\mathcal{A}^{(s)}+\mathcal{A}^{(m)}$ is decomposed into the stiffness matrix $\mathcal{A}^{(s)}$ and the mass matrix $\mathcal{A}^{(m)}$ with elements defined by

$$
\left[\mathcal{A}^{(s)}\right]_{i, j}=\left(D \nabla \Phi_{j}, \nabla \Phi_{i}\right), \quad\left[\mathcal{A}^{(m)}\right]_{i, j}=\left(\Sigma_{a} \Phi_{j}, \Phi_{i}\right)
$$

### 5.1.2 Conforming Polynomial Basis

To provide the necessary background for the one-dimensional discussions in Sections 5.2.2 and 5.2 .3 we recall the one-dimensional conforming piecewise linear FEM and its properties. The well-established extensions to a two-dimensional tensor product mesh are noted at the end of this discussion.

We begin by stating the global piecewise linear basis functions,

$$
\Psi_{i+\frac{1}{2}}(x)= \begin{cases}\frac{1}{\Delta x_{i}}\left(x-x_{i-\frac{1}{2}}\right) & x \in \Omega_{i}  \tag{5.3}\\ \frac{1}{\Delta x_{i+1}}\left(x_{i+\frac{3}{2}}-x\right) & x \in \Omega_{i+1} \\ 0 & x \in \Omega_{h} \backslash \Omega_{i}\end{cases}
$$

which leads to the following discretization.

Definition 5.1.1. The piecewise linear FEM for one-dimensional one-group diffusion is defined by the following stencil,

$$
\begin{equation*}
-c_{i} \phi_{i-\frac{1}{2}}+\left(d_{i}+d_{i+1}\right) \phi_{i+\frac{1}{2}}-c_{i+1} \phi_{i+\frac{3}{2}}=f_{i}+f_{i+1} \tag{5.4a}
\end{equation*}
$$

where,

$$
\begin{equation*}
c_{i}=\frac{D_{i}}{\Delta x_{i}}\left\{1-\frac{2}{3}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}\right\}, \quad d_{i}=\frac{D_{i}}{\Delta x_{i}}\left\{1+\frac{4}{3}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}\right\}, \quad f_{i}=\frac{\Delta x_{i}}{2} Q_{i} \tag{5.4~b}
\end{equation*}
$$

and,

$$
\lambda_{i, j}=\sqrt{\frac{\left(\Sigma_{a}\right)_{i, j}}{D_{i, j}}}, \quad \lambda_{i, j}^{(x)}=\lambda_{i, j} \Delta x_{i}, \quad \bar{\lambda}_{i, j}^{(x)}=\frac{1}{2} \lambda_{i, j}^{(x)}
$$

was defined in (3.47).
This is a well-known second-order discretization (see, e.g. [69]), for which the variational form implies that the discrete error

$$
e_{h}(x)=\phi(x)-\phi_{h}(x)
$$

is minimized over the trial space $V_{h}$ in the $H^{\mathbf{1}}(\Omega)$ norm. However, several relevant physical properties of the discretization are seldom referenced, hence we recount them here. To this end we first recall the definition of an M-matrix given in Varga [70] (pg. 85).

Definition 5.1.2. A real $n \times n$ matrix $A=\left(a_{i, j}\right)$ with $a_{i, j} \leq 0$ for all $i \neq j$ is an $M$-matrix if $A$ is nonsingular, and $A^{-1} \geq 0$.

Here the notation $A^{-1} \geq 0$ indicates that the individual elements of the inverse are nonnegative, and thus, if $\mathcal{A}$ is an M-matrix and the source $Q(x) \geq 0$ for all $x \in \Omega$ then a nonnegative (and hence physical) solution $\phi_{h}(x) \geq 0$ for all $x \in \Omega$ is guaranteed.

The following Corollaries of Varga [70] (pg. 85) are employed throughout our discussion to prove that certain global matrices $\mathcal{A}$ satisfy the above definition.

Corollary 5.1. If $A=\left(a_{i, j}\right)$ is a real, irreducibly diagonally dominant $n \times n$ matrix with $a_{i, j} \leq 0$ for all $i \neq j$, and $a_{i, i}>0$ for all $1 \leq i \leq n$ then $A^{-1}>0$.

Corollary 5.2. If $A=\left(a_{i, j}\right)$ is a real, symmetric and nonsingular $n \times n$ irreducible matrix, where $a_{i, j} \leq 0$ for all $i \neq j$, then $A^{-1}>0$ if and only if $A$ is positive definite.

From Corollary 5.1 it is trivial to conclude that $\mathcal{A}$ of Definition 5.1.1 is an $M$-matrix if $\left[\bar{\lambda}_{i}^{(x)}\right]^{2}<2 / 3$. In applications where the coefficients vary by orders of magnitude this restriction may be quite severe. Hence, even though Corollary 5.1 only gives a sufficient condition for $A^{-1}>0$, this breakdown may deter the use of this FEM discretization.

The computation of the current is also required in a number of modelling problems. Typically this is obtained through the application of Fick's law to the finite element solution,

$$
\begin{equation*}
J_{h}(x)=-D_{i} \frac{\partial}{\partial x}\left[\phi_{h}(x)\right] \tag{5.5}
\end{equation*}
$$

In this manner we find,

$$
J_{i+\frac{1}{2}}^{+}-J_{i+\frac{1}{2}}^{-}=-\frac{D_{i}}{\Delta x_{i}} \phi_{i-\frac{1}{2}}+\left(\frac{D_{i}}{\Delta x_{i}}+\frac{D_{i+1}}{\Delta x_{i+1}}\right) \phi_{i+\frac{1}{2}}-\frac{D_{i+1}}{\Delta x_{i+1}} \phi_{i+\frac{3}{2}}
$$

and hence from Definition 5.1.1 it is apparent that the normal currents are discontinuous. Furthermore, the current is constant on each cell and hence it is not possible to satisfy a cell-based balance equation. Instead we find that the primal conforming FEM approximates a vertex-based balance relation.

The common procedure of lumping the mass matrix [69] may be used to alter the properties of this discretization. In particular, lumping modifies the coefficients of (5.4a) to read,

$$
\begin{equation*}
c_{i}=\frac{D_{i}}{\Delta x_{i}}, \quad d_{i}=\frac{D_{i}}{\Delta x_{i}}\left\{1+2\left[\bar{\lambda}_{i}^{(x)}\right]^{2}\right\}=\frac{D_{i}}{\Delta x_{i}}+\frac{\Delta x_{i}}{2}\left(\Sigma_{a}\right)_{i} \tag{5.6}
\end{equation*}
$$

while the source term is unaffected. Then it is apparent from Corollary 5.1 that the global matrix $\mathcal{A}$ is unconditionally an M-matrix. Moreover, if the coefficients are constant on each cell, then the discretization satisfies the following vertex based balance equation,

$$
\begin{equation*}
J_{i+1}-J_{i}+\frac{1}{2}\left\{\Delta x_{i}\left(\Sigma_{a}\right)_{i}+\Delta x_{i+1}\left(\Sigma_{a}\right)_{i+1}\right\} \phi_{i+\frac{1}{2}}=\frac{1}{2}\left\{\Delta x_{i} Q_{i}+\Delta x_{i+1} Q_{i+1}\right\} \tag{5.7}
\end{equation*}
$$

However, the lumping process also destroys the variational form, and hence we no longer $\operatorname{minimize}\left\|e_{h}(x)\right\|_{1}$ over $V_{h}$.

We summarize these observations in the following proposition.

Proposition 5.1. The piecewise linear FEM with the coefficients constant on each cell and exact integration of the stiffness and mass matrices is a second order discretization of the one-group diffusion equation that is characterized by the following properties

1. $\left\|e_{h}(x)\right\|_{1}$ is minimized over the trial space $V_{h}$
2. $\phi_{h}(x) \geq 0$ for all $x \in \Omega$ if $\left[\bar{\lambda}_{i}^{(x)}\right]^{2}<2 / 3$ for all $i \in[1, L]$
3. $J_{i+\frac{1}{2}}^{+}-J_{i+\frac{1}{2}}^{-}=\mathcal{O}(\Delta x)$
4. A cell balance equation does not exist
5. A vertex based balance equation is approximated

Lumping the mass matrix also produces a second order discretization which possesses the following properties

1. $\left\|e_{h}(x)\right\|_{1}$ is no longer minimized over the trial space $V_{h}$
2. $\phi_{h}(x) \geq 0$ for all $x \in \Omega$ independent of the mesh
3. $J_{i+\frac{1}{2}}^{+}-J_{i+\frac{1}{2}}^{-}=\mathcal{O}(\Delta x)$
4. A cell balance equation does not exist
5. The vertex based balance equation (5.7) is satisfied exactly

Thus, it is apparent that a lumping of the mass matrix (i.e. the lower-order terms) may be used to alter certain properties of the discretization. Specifically, we demonstrated that one may sacrifice the variational principle in favour of certain physical properties, such as positivity of the FEM solution and a balance relation. Ideally, however, such a sacrifice should be unnecessary, and indeed all of the desired properties are obtained with the use of analytic basis functions (see Section 5.2.3).

Finally, we note that the extension of these observations to the two-dimensional case on a tensor product mesh is, apart from the vertex-based balance relation, quite natural. The control volume for the lumped approximation of the bilinear finite elements is nonstandard and is discussed in some detail in [19].

### 5.2 Primal Nonconforming FEM

In Section 5.1 .2 it was demonstrated that the popular conforming piecewise bilinear basis is not necessarily the best choice, particularly if the continuous problem has piecewise constant coefficients which jump by orders of magnitude. In many situations it has been demonstrated that relaxing the conformity restriction permits the choice of more physically motivated basis functions which are ultimately more accurate. We will demonstrate that this is precisely the case for the nodal discretizations whose corresponding basis functions are based on average quantities, and hence, enforce continuity in only a weak sense. To simplify the development and to highlight the underlying influence of transverse integration we first derive the discretizations in 1D, and then demonstrate their natural extension to a tensor product mesh in 2D.

### 5.2.1 The Edge and Cell Basis

The underlying commonality between the various nodal methods is the choice of cell- and edge-based unknowns. To develop primal nonconforming finite element methods that are equivalent to specific nodal methods we first introduce a notation suitable for the cell and edge bases.

## One-Dimensional Basis

The distinction between cell and edge unknowns in 1D simply states that for each cell, centred at $x_{i}$, we have an associated unknown, $\phi_{i}$, which is defined as,

$$
\phi_{i}=\frac{1}{\Delta x_{i}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \phi(x) d x
$$

while the edges are degenerate, being simply point values,

$$
\phi_{i+\frac{1}{2}}=\phi\left(x_{i+\frac{1}{2}}\right), \quad i=0, \ldots, L .
$$

The implication of this degeneracy is that these methods are actually conforming in 1D. Considering the canonical element $\Omega_{c} \equiv[-1,1]$ we adopt the following notation,

$$
\psi_{\left(\xi_{k}\right)}^{\left(l_{k}\right)}(\xi)
$$

where the superscript $l_{k}$ indicates the moment that normalizes this basis function and the subscript $\xi_{k}$ defines at which point in $\Omega_{c}$ the function is normalized to 1 . In the event that either script does not apply to a particular basis function, an asterisk "*" is used. Moreover, all other relevant moments and point values of a particular basis function are set to zero. Although this notation may seem excessive in 1D, it is a notation that extends well to higher dimensions and higher-order methods. The corresponding global

Table 5.1: Canonical Definitions in 1D

$$
\begin{aligned}
\psi_{(1)}^{(*)}(\xi) & \Rightarrow \quad \psi_{(-1)}^{(*)}(-1)=1 \\
\psi_{(-1)}^{(*)}(\xi) & \Rightarrow \quad \psi_{(1)}^{(*)}(1)=1 \\
\psi_{(*)}^{(0)}(\xi) & \Rightarrow \quad \frac{1}{2} \int_{-1}^{1} \psi_{(*)}^{(0)}(\xi) d \xi=1
\end{aligned}
$$

basis functions are simply,

$$
\begin{aligned}
& \Psi_{i+\frac{1}{2}}^{(*)}(x)= \begin{cases}\psi_{(-1)}^{(*)}\left[\xi_{i}(x)\right] & x_{i-\frac{1}{2}} \leq x \leq x_{i+\frac{1}{2}} \\
\psi_{(1)}^{(*)}\left[\xi_{i+1}(x)\right] & x_{i+\frac{1}{2}} \leq x \leq x_{i+\frac{3}{2}} \\
0 & \text { otherwise }\end{cases} \\
& \Psi_{i}^{(0)}(x)= \begin{cases}\psi_{(*)}^{(0)}\left[\xi_{i}(x)\right] & x_{i-\frac{1}{2}} \leq x \leq x_{i+\frac{1}{2}} \\
0 & \text { otherwise }\end{cases}
\end{aligned}
$$

where $\xi_{i}(x)=2\left(x-x_{i}\right) / \Delta x_{i}$. Thus the global expansion of $\phi_{h}(x)$ is given by

$$
\phi_{h}(x)=\sum_{i=1}^{L-1} \phi_{i+\frac{1}{2}} \Psi_{i+\frac{1}{2}}^{(*)}(x)+\sum_{i=1}^{L} \phi_{i} \Psi_{i}^{(0)}(x)
$$

With the definitions of the basis functions on the canonical element it is possible to perform an element by element construction of the stiffness and mass matrices (cf. [15]). Specifically we consider the contribution from the element $\Omega_{i}$ which is described by the following $3 \times 3$ block,

$$
\mathrm{S}_{i}=\left[\begin{array}{c|cc}
\mathrm{s}_{i}\left(\psi_{(*)}^{(0)}, \psi_{(*)}^{(0)}\right) & \mathbf{s}_{i}\left(\psi_{(*)}^{(0)}, \psi_{(-1)}^{(*)}\right) & \mathrm{s}_{i}\left(\psi_{(*)}^{(0)}, \psi_{(1)}^{(*)}\right)  \tag{5.8}\\
\hline \mathbf{s}_{i}\left(\psi_{(-1)}^{(*)}, \psi_{(*)}^{(0)}\right) & \mathrm{s}_{i}\left(\psi_{(-1)}^{(*)}, \psi_{(-1)}^{(*)}\right) & \mathbf{s}_{i}\left(\psi_{(-1)}^{(*)}, \psi_{(1)}^{(*)}\right) \\
\mathrm{s}_{i}\left(\psi_{(1)}^{(*)}, \psi_{(*)}^{(0)}\right) & \mathrm{s}_{i}\left(\psi_{(1)}^{(*)}, \psi_{(-1)}^{(*)}\right) & \mathbf{s}_{i}\left(\psi_{(1)}^{(*)}, \psi_{(1)}^{(*)}\right)
\end{array}\right]
$$

where

$$
\mathbf{s}_{i}\left(\psi_{\left(\xi_{1}\right)}^{\left(m_{1}\right)}, \psi_{\left(\xi_{2}\right)}^{\left(m_{2}\right)}\right)=\frac{2 D_{i}}{\Delta x_{i}} \int_{-1}^{1}\left\{\left[\psi_{\left(\xi_{1}\right)}^{\left(m_{1}\right)}(\xi)\right]_{\xi} \cdot\left[\psi_{\left(\xi_{2}\right)}^{\left(m_{2}\right)}(\xi)\right]_{\xi}\right\} d \xi
$$

Similarly we have the elemental mass matrix,

$$
\mathbf{M}_{i}=\left[\begin{array}{l|ll}
\mathrm{m}_{i}\left(\psi_{(*)}^{(0)}, \psi_{(*)}^{(0)}\right) & \mathrm{m}_{i}\left(\psi_{(*)}^{(0)}, \psi_{(-1)}^{(*)}\right) & \mathrm{m}_{i}\left(\psi_{(*)}^{(0)}, \psi_{(1)}^{(*)}\right)  \tag{5.9}\\
\hline \mathrm{m}_{i}\left(\psi_{(-1)}^{(*)}, \psi_{(*)}^{(0)}\right) & \mathrm{m}_{i}\left(\psi_{(-1)}^{(*)}, \psi_{(-1)}^{(*)}\right) & \mathrm{m}_{i}\left(\psi_{(-1)}^{(*)}, \psi_{(1)}^{(*)}\right) \\
\mathrm{m}_{i}\left(\psi_{(1)}^{(*)}, \psi_{(*)}^{(0)}\right) & \mathrm{m}_{i}\left(\psi_{(1)}^{(*)}, \psi_{(-1)}^{(*)}\right) & \mathrm{m}_{i}\left(\psi_{(1)}^{(*)}, \psi_{(1)}^{(*)}\right)
\end{array}\right]
$$

where,

$$
\mathrm{m}_{i}\left(\psi_{\left(\xi_{1}\right)}^{\left(m_{1}\right)}, \psi_{\left(\xi_{2}\right)}^{\left(m_{2}\right)}\right)=\frac{\sigma_{i} \Delta x_{i}}{2} \int_{-1}^{1}\left\{\psi_{\left(\xi_{1}\right)}^{\left(m_{1}\right)}(\xi) \cdot \psi_{\left(\xi_{2}\right)}^{\left(m_{2}\right)}(\xi)\right\} d \xi
$$

Hence, the total elemental matrix is $A_{i}=S_{i}+M_{i}$. Finally the elemental contribution to the load vector is

$$
\left.\mathrm{b}_{i}=\frac{\Delta x_{i}}{2}\left[\frac{\left(Q, \psi_{(*)}^{(0)}\right)}{\left(Q, \psi_{(-1)}^{(*)}\right)}\right]\left(Q, \psi_{(1)}^{(*)}\right)\right]
$$

which, under the assumption of a piecewise constant source gives, $\mathbf{b}_{i}=\left[Q_{i}, 0,0\right]^{T}$.

## Two-Dimensional Basis

The two-dimensional cell- and edge-based unknowns were introduced in Chapter 3, and provide 5 degrees of freedom on the canonical element $\Omega_{c} \equiv[-1,1] \times[-1,1]$. The corresponding extension of the one-dimensional basis function notation is summarized in Table 5.2. In addition, the two-dimensional bases that we consider are unions

Table 5.2: Canonical Definitions in 2D

$$
\begin{aligned}
\psi_{(-1, *)}^{(*, 0)}(\xi, \eta) & \Rightarrow \frac{1}{2} \int_{-1}^{1} \psi_{(-1, *)}^{(*, 0)}(-1, \eta) d \eta=1 \\
\psi_{(1, *)}^{(*, 0)}(\xi, \eta) & \Rightarrow \frac{1}{2} \int_{-1}^{1} \psi_{(1, *)}^{(*, 0)}(1, \eta) d \eta=1 \\
\psi_{(*,-1)}^{(0, *)}(\xi, \eta) & \Rightarrow \frac{1}{2} \int_{-1}^{1} \psi_{(*,-1)}^{(0, *)}(\xi,-1) d \xi=1 \\
\psi_{(*, 1)}^{(0, *)}(\xi, \eta) & \Rightarrow \frac{1}{2} \int_{-1}^{1} \psi_{(*, 1)}^{(0, *)}(\xi, 1) d \xi=1 \\
\psi_{(*, *)}^{(0,0)}(\xi, \eta) & \Rightarrow \frac{1}{4} \int_{-1}^{1} \int_{-1}^{1} \psi_{(*, *)}^{(0,0)}(\xi, \eta) d \xi d \eta=1
\end{aligned}
$$

of one-dimensional bases. This property significantly simplifies the development and implementation of these methods and is stated in the following Corollary.

Corollary 5.3. The lowest-order two-dimensional cell and edge basis functions that we consider span a union of one-dimensional spaces, and hence, may be written,

$$
\begin{align*}
\psi_{(*, *)}^{(0,0)}(\xi, \eta) & =\psi_{(*)}^{(0)}(\xi)+\psi_{(*)}^{(0)}(\eta)-1  \tag{5.10a}\\
\psi_{(1, *)}^{(*, 0)}(\xi, \eta) & =\psi_{(1)}^{(*)}(\xi)  \tag{5.10b}\\
\psi_{(-1, *)}^{(*, 0)}(\xi, \eta) & =\psi_{(-1)}^{(*)}(\xi)  \tag{5.10c}\\
\psi_{(*, 1)}^{(0, *)}(\xi, \eta) & =\psi_{(1)}^{(*)}(\eta)  \tag{5.10d}\\
\psi_{(*,-1)}^{(0, *)}(\xi, \eta) & =\psi_{(-1)}^{(*)}(\eta) \tag{5.10e}
\end{align*}
$$

where the one-dimensional basis functions are characterized in Table 5.1 and span the corresponding one-dimensional space.

Proof. A two-dimensional canonical basis function may be written

$$
\psi(\xi, \eta)=c_{0}+a_{0} \psi_{(*)}^{(1)}(\xi)+a_{1} \psi_{(-1)}^{(*)}(\xi)+a_{2} \psi_{(1)}^{(*)}(\xi)+b_{0} \psi_{(0)}^{(*)}(\eta)+b_{1} \psi_{(*)}^{(-1)}(\eta)+b_{2} \psi_{(*)}^{(1)}(\eta)
$$

and hence the correspondence with cell and edge unknowns yields the following underdetermined system,

$$
\left[\begin{array}{lllllll}
1 & 0 & 0 & 1 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 1
\end{array}\right]\left[\begin{array}{c}
a_{0} \\
a_{1} \\
a_{2} \\
b_{0} \\
b_{1} \\
b_{2} \\
c_{0}
\end{array}\right]=e_{k}
$$

where $e_{k}=\left[\delta_{k, 1}, \delta_{k, 2}, \delta_{k, 3}, \delta_{k, 4}, \delta_{k, 5}\right]^{T}$, and $\delta_{k, l}$ is the Kronecker delta function. It is a simple matter to verify that (5.10) is a solution.

Utilizing (5.10) the global basis functions are simply,

$$
\begin{align*}
& \Psi_{i+\frac{1}{2}, j}^{(*, 0)}(x, y)= \begin{cases}\psi_{(1)}^{(*)}\left[\xi_{i}(x)\right] & x \in \Omega_{i, j} \\
\psi_{(-1)}^{(*)}\left[\xi_{i+1}(x)\right] & x \in \Omega_{i+1, j} \\
0 & x \in \Omega_{h} \backslash\left\{\Omega_{i, j} \cup \Omega_{i+1, j}\right\}\end{cases}  \tag{5.11a}\\
& \Psi_{i, j+\frac{1}{2}}^{(0, *)}(x, y)= \begin{cases}\psi_{(1)}^{(*)}\left[\eta_{j}(y)\right] & x \in \Omega_{i, j} \\
\psi_{(-1)}^{(*)}\left[\eta_{j+1}(y)\right] & x \in \Omega_{i, j+1} \\
0 & x \in \Omega_{h} \backslash\left\{\Omega_{i, j} \cup \Omega_{i, j+1}\right\}\end{cases}  \tag{5.11b}\\
& \Psi_{i, j}^{(0,0)}(x, y)= \begin{cases}\psi_{(*)}^{(0)}\left[\xi_{i}(x)\right]+\psi_{(*)}^{(0)}\left[\eta_{j}(y)\right]-1 & (x, y) \in \Omega_{i, j} \\
0 & (x, y) \in \Omega_{h} \backslash \Omega_{i, j}\end{cases} \tag{5.11c}
\end{align*}
$$

where $\left\{\xi_{i}(x), \eta_{j}(y)\right\}$ define the affine mapping of $\Omega_{i, j}$ onto $\Omega_{c}$. The global expansion is $\phi_{h}(x, y)=\sum_{i=1}^{L-1} \sum_{j=1}^{M} \phi_{i+\frac{1}{2}, j} \Psi_{i+\frac{1}{2}, j}^{(*, 0)}(x, y)+\sum_{i=1}^{L} \sum_{j=1}^{M-1} \phi_{i, j+\frac{1}{2}} \Psi_{i, j+\frac{1}{2}}^{(0, *)}(x, y)+\sum_{i=1}^{L} \sum_{j=1}^{M} \phi_{i, j} \Psi_{i, j}^{(0,0)}(x, y)$

The elemental construction of the stiffness and mass matrices is now a trivial extension of the 1D case, with the elements of the elemental stiffness matrix $S_{i, j}$ given by,

$$
\begin{aligned}
\mathbf{s}_{i, j}\left(\psi_{\left(\xi_{1}, \eta_{1}\right)}^{\left(l_{1}, m_{1}\right)}, \psi_{\left(\xi_{2}, \eta_{2}\right)}^{\left(l_{2}, m_{2}\right)}\right) & =D_{i, j} \frac{\Delta y_{j}}{\Delta x_{i}} \int_{-1}^{1} \int_{-1}^{1}\left\{\left[\psi_{\left(\xi_{1}, \eta_{1}\right)}^{\left(l_{1}, m_{1}\right)}(\xi, \eta)\right]_{\xi} \cdot\left[\psi_{\left(\xi_{2}, \eta_{2}\right)}^{\left(l_{2}, m_{2}\right)}(\xi, \eta)\right]_{\xi}\right\} d \xi d \eta \\
& +D_{i, j} \frac{\Delta x_{i}}{\Delta y_{j}} \int_{-1}^{1} \int_{-1}^{1}\left\{\left[\psi_{\left(\xi_{1}, \eta_{1}\right)}^{\left(l_{1}, m_{1}\right)}(\xi, \eta)\right]_{\eta} \cdot\left[\psi_{\left(\xi_{2}, \eta_{2}\right)}^{\left(l_{2}, m_{2}\right)}(\xi, \eta)\right]_{\eta}\right\} d \xi d \eta
\end{aligned}
$$

and similarly the elements of the elemental mass matrix are given by,

$$
\mathbf{m}_{i, j}\left(\psi_{\left(\xi_{1}, \eta_{1}\right)}^{\left(l_{1}, m_{1}\right)}, \psi_{\left(\xi_{2}, \eta_{2}\right)}^{\left(l_{2}, m_{2}\right)}\right)=\frac{1}{2} \sigma_{i} \Delta x_{i} \Delta y_{j} \int_{-1}^{1} \int_{-1}^{1}\left\{\left[\psi_{\left(\xi_{1}, \eta_{1}\right)}^{\left(l_{1}, m_{1}\right)}(\xi, \eta)\right] \cdot\left[\psi_{\left(\xi_{2}, \eta_{2}\right)}^{\left(l_{2}, m_{2}\right)}(\xi, \eta)\right]\right\} d \xi d \eta
$$

We note that the evaluation of these integrals is greatly simplified with the relationships presented in (5.10)

### 5.2.2 Polynomial Basis Functions

The objective is to demonstrate that a simple tensor product polynomial basis may be used to derive a nonconformal finite element method which is equivalent to the NEM discretization. Recall that NEM and NIM are equivalent for conservative diffusion. Hence for this case we investigate the implication of this additional equivalence on the interpretation of the nonconformal FEM.

## A One-Dimensional Study

Although nodal methods lead to conforming finite elements in the simple 1D case, the use of transverse integration in modern nodal methods effectively decouples the coordinate dependence and hence most treatments are essentially one-dimensional in nature. Thus, a study of the one-dimensional case is an informative starting point, in which the basis functions and integrals are significantly less complicated.

Definition 5.2.1. In one dimension the cell and edge quadratic basis functions on the canonical element $\Omega_{c}$ may be written,

$$
\begin{align*}
\psi_{(-1)}^{(*)}(\xi) & =-\frac{1}{4}-\frac{1}{2} \xi+\frac{3}{4} \xi^{2}  \tag{5.12a}\\
\psi_{(1)}^{(*)}(\xi) & =-\frac{1}{4}+\frac{1}{2} \xi+\frac{3}{4} \xi^{2}  \tag{5.12b}\\
\psi_{(*)}^{(0)}(\xi) & =\frac{3}{2}\left(1-\xi^{2}\right) \tag{5.12c}
\end{align*}
$$

and are displayed in Figure 5.1. The corresponding global quadratic basis functions are given by,

$$
\begin{align*}
& \Psi_{i+\frac{1}{2}}^{(*)}(x)= \begin{cases}1-\frac{4}{\Delta x_{i}}\left(x_{i+\frac{1}{2}}-x\right)+\frac{3}{\Delta x_{i}^{2}}\left(x_{i+\frac{1}{2}}-x\right)^{2} & x \in \Omega_{i} \\
1-\frac{4}{\Delta x_{i+1}}\left(x-x_{i+\frac{1}{2}}\right)+\frac{3}{\Delta x_{i+1}^{2}}\left(x-x_{i+\frac{1}{2}}\right)^{2} & x \in \Omega_{i+1} \\
0 & x \in \Omega_{h} \backslash\left\{\Omega_{i} \cup \Omega_{i+1}\right\}\end{cases}  \tag{5.13a}\\
& \Psi_{i}^{(0)}(x)= \begin{cases}\frac{6}{\Delta x_{i}^{2}}\left(x_{i+\frac{1}{2}}-x\right)\left(x-x_{i-\frac{1}{2}}\right) & x \in \Omega_{i} \\
0 & x \in \Omega_{h} \backslash \Omega_{i}\end{cases} \tag{5.13b}
\end{align*}
$$

Remark 5.1. Basis functions such as $\Psi_{i}^{(0)}(x)$ are often referred to as bubble functions as their support is restricted to a single cell and hence their appearance is bubble-like.

Derivation of the $1 D$ Polynomial Basis. Writing a general quadratic as

$$
\psi(\xi)=a_{0}+a_{1} \xi+a_{2} \xi^{2}
$$

the correspondence with cell and edge unknowns yields the following linear system governing the expansion coefficients,

$$
\left[\begin{array}{rrr}
1 & -1 & 1 \\
1 & 1 & 1 \\
1 & 0 & \frac{1}{3}
\end{array}\right]\left[\begin{array}{l}
a_{0} \\
a_{1} \\
a_{2}
\end{array}\right]=e_{k}
$$



Figure 5.1: The quadratic canonical cell and edge basis functions.
where $e_{k}=\left[\delta_{k, 1}, \delta_{k, 2}, \delta_{k, 3}\right]$. With a little help from MAPLE [17] we obtain the canonical basis functions given in (5.12). If we note that,

$$
\begin{equation*}
\xi=\frac{2}{\Delta x_{i}}\left(x-x_{i}\right)=1-\frac{2}{\Delta x_{i}}\left(x_{i+\frac{1}{2}}-x\right)=\frac{2}{\Delta x_{i}}\left(x-x_{i-\frac{1}{2}}\right)-1 \tag{5.14}
\end{equation*}
$$

then the global basis functions (5.13) follow.
Being low order polynomials the exact evaluation of the elemental stiffness and mass matrices (5.8) and (5.9) respectively, is simple and gives,

$$
\mathrm{S}_{i}=2 \frac{D_{i}}{\Delta x_{i}}\left[\begin{array}{c|rr}
6 & -3 & -3  \tag{5.15}\\
\hline-3 & 2 & 1 \\
-3 & 1 & 2
\end{array}\right], \quad \mathrm{M}_{i}=\frac{\left(\Sigma_{a}\right)_{i} \Delta x_{i}}{2}\left[\begin{array}{r|rr}
\frac{12}{5} & -\frac{1}{5} & -\frac{1}{5} \\
\hline-\frac{1}{5} & \frac{4}{15} & -\frac{1}{15} \\
-\frac{1}{5} & -\frac{1}{15} & \frac{4}{15}
\end{array}\right]
$$

Furthermore recalling that,

$$
\frac{\left(\Sigma_{a}\right)_{i} \Delta x_{i}}{2}=2 \frac{D_{i}}{\Delta x_{i}}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}
$$

we find that the total elemental matrix is simply,

$$
\mathrm{A}_{i}=\frac{D_{i}}{\Delta x_{i}}\left[\begin{array}{r|rr}
12\left\{1+\frac{2}{5}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}\right\} & -6\left\{1+\frac{1}{15}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}\right\} & -6\left\{1+\frac{1}{15}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}\right\}  \tag{5.16}\\
\hline-6\left\{1+\frac{1}{15}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}\right\} & 4\left\{1+\frac{2}{15}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}\right\} & 2\left\{1-\frac{1}{15}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}\right\} \\
-6\left\{1+\frac{1}{15}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}\right\} & 2\left\{1-\frac{1}{15}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}\right\} & 4\left\{1+\frac{2}{15}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}\right\}
\end{array}\right]
$$

To highlight the properties of this discretization we present the corresponding stencils.

Definition 5.2.2. The cell- and edge-based polynomial FEM for one-dimensional onegroup diffusion is defined by the following local cell-based equation,

$$
\begin{align*}
-6 \frac{D_{i}}{\Delta x_{i}}\left\{1+\frac{1}{15}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}\right\} \phi_{i-\frac{1}{2}}+12 \frac{D_{i}}{\Delta x_{i}}\{1 & \left.+\frac{2}{5}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}\right\} \phi_{i}  \tag{5.17}\\
& -6 \frac{D_{i}}{\Delta x_{i}}\left\{1+\frac{1}{15}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}\right\} \phi_{i+\frac{1}{2}}=\Delta x_{i} Q_{i}
\end{align*}
$$

and an edge-based stencil of the form,

$$
\begin{align*}
2 \frac{D_{i}}{\Delta x_{i}}\{1 & \left.-\frac{1}{15}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}\right\} \phi_{i-\frac{1}{2}}-6 \frac{D_{i}}{\Delta x_{i}}\left\{1+\frac{1}{15}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}\right\} \phi_{i} \\
& +\left(4 \frac{D_{i}}{\Delta x_{i}}\left\{1+\frac{2}{15}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}\right\}+4 \frac{D_{i+1}}{\Delta x_{i+1}}\left\{1+\frac{2}{15}\left[\bar{\lambda}_{i+1}^{(x)}\right]^{2}\right\}\right) \phi_{i+\frac{1}{2}}  \tag{5.18}\\
& -6 \frac{D_{i+1}}{\Delta x_{i+1}}\left\{1+\frac{1}{15}\left[\bar{\lambda}_{i+1}^{(x)}\right]^{2}\right\} \phi_{i+1}+2 \frac{D_{i+1}}{\Delta x_{i+1}}\left\{1-\frac{1}{15}\left[\bar{\lambda}_{i+1}^{(x)}\right]^{2}\right\} \phi_{i+\frac{3}{2}}=0
\end{align*}
$$

Definition 5.2.3. The reduced edge-based polynomial FEM for one-group diffusion is given by the following edge-based equation,

$$
\begin{equation*}
-c_{i} \phi_{i-\frac{1}{2}}+\left(d_{i}+d_{i+1}\right) \phi_{i+\frac{1}{2}}-c_{i+1} \phi_{i+\frac{3}{2}}=f_{i}+f_{i+1} \tag{5.19a}
\end{equation*}
$$

where,

$$
\begin{align*}
& c_{i}=\frac{D_{i}}{\Delta x_{i}}\left\{\frac{1-\frac{4}{15}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}+\frac{1}{15}\left[\bar{\lambda}_{i}^{(x)}\right]^{4}}{1+\frac{2}{5}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}}\right\}  \tag{5.19b}\\
& d_{i}=\frac{D_{i}}{\Delta x_{i}}\left\{\frac{1+\frac{26}{15}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}+\frac{1}{5}\left[\bar{\lambda}_{i}^{(x)}\right]^{4}}{1+\frac{2}{5}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}}\right\}  \tag{5.19c}\\
& f_{i}=\frac{1}{2}\left\{\frac{1+\frac{1}{15}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}}{1+\frac{2}{5}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}}\right\} \Delta x_{i} Q_{i} \tag{5.19d}
\end{align*}
$$

Derivation of the $1 D$ polynomial FEMs. The coefficients in equations (5.17) and (5.18) are simply read from the total elemental matrix (5.16). The support of $\Psi_{i}^{(0)}(x)$ is $\Omega_{i}$ and hence $\phi_{i}$ may be eliminated locally. Solving for the cell average in (5.17) gives,

$$
\begin{equation*}
\phi_{i}=\frac{1}{2} \frac{\left\{1+\frac{1}{15}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}\right\}}{\left\{1+\frac{2}{5}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}\right\}}\left(\phi_{i+\frac{1}{2}}+\phi_{i-\frac{1}{2}}\right)+\frac{\Delta x_{i}^{2}}{12 D_{i}\left\{1+\frac{2}{5}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}\right\}} Q_{i} \tag{5.20}
\end{equation*}
$$

and hence, substitution into (5.18) yields (5.19a).

Lemma 5.1. The global coefficient matrix $\mathcal{A}$ of the $1 D$ reduced edge-based polynomial FEM (5.19a) is an M-matrix.

Proof. The coefficients $c_{i}$ are strictly positive, with the numerator attaining an absolute minimum at $\bar{\lambda}_{i}^{(x)}=\sqrt{2}$ of $\frac{11}{15}>0$, and hence the elements $[\mathcal{A}]_{l, k} \leq 0 \forall l \neq k$. In addition for $\bar{\lambda}_{i}^{(x)}>0$ we have $d_{i}>c_{i}$ hence $\mathcal{A}$ is diagonally dominant. Thus by Corollary 5.2 it is an M-matrix.

Obtaining an M-matrix, independently of the off-diagonal terms that arise from the mass matrix, appears to be a significant improvement over what is obtained with the piecewise linear basis. However, the piecewise quadratic basis functions are not nonnegative, and hence, we obtain unconditional positivity of only the point values, $\phi_{h}\left(x_{i+\frac{1}{2}}\right)$ for $i=1, \ldots, L-1$. In addition, upon evaluating the current (5.5) it is apparent that (5.17) is not the balance equation over $\Omega_{i}$. Moreover, the current is not continuous at the cell edges. Hence, the quadratic basis does share some of the physical deficiencies that appeared previously with the piecewise linear basis functions. Ultimately, however, we seek to establish equivalences between certain nodal methods and certain FEM methods, and for this both the current continuity and the balance equation are essential. Thus, we consider the approximate treatment of the elemental mass matrix. In particular, lumping the mass matrix gives,

$$
\mathfrak{L}\left(\mathrm{M}_{i}\right)=\widetilde{M}_{i}=\frac{\left(\Sigma_{a}\right)_{i} \Delta x_{i}}{2}\left[\begin{array}{l|ll}
2 & 0 & 0 \\
\hline 0 & 0 & 0 \\
0 & 0 & 0
\end{array}\right]
$$

Hence the total elemental matrix becomes,

$$
\widetilde{A}_{i}=S_{i}+\widetilde{\mathrm{M}_{i}}=\frac{D_{i}}{\Delta x_{i}}\left[\left.\begin{array}{c|cc}
12\left\{1+\frac{1}{3}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}\right\} & -6 & -6  \tag{5.21}\\
\hline-6 \\
-6
\end{array} \right\rvert\, \begin{array}{rr}
4 & 2 \\
2 & 4
\end{array}\right]
$$

such that the resulting discretization not only preserves the cell based balance equation but also restores the continuity of the current. However, the corresponding coefficients of the reduced edge formulation become,

$$
\begin{align*}
& \widetilde{c}_{i}=\frac{D_{i}}{\Delta x_{i}}\left\{\frac{1-\frac{2}{3}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}}{1+\frac{1}{3}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}}\right\}  \tag{5.22a}\\
& \widetilde{d}_{i}=\frac{D_{i}}{\Delta x_{i}}\left\{\frac{1+\frac{4}{3}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}}{1+\frac{1}{3}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}}\right\}  \tag{5.22b}\\
& \widetilde{f}_{i}=\frac{1}{2}\left\{\frac{\Delta x_{i} Q_{i}}{1+\frac{1}{3}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}}\right\} \tag{5.22c}
\end{align*}
$$

and hence the lumped global matrix $\widetilde{\mathcal{A}}$ is not an M-matrix unless $\left[\bar{\lambda}_{i}^{(x)}\right]^{2}<2 / 3$. Thus we have exchanged the property that $\mathcal{A}$ is unconditionally an M -matrix, for the adherence to cell balance and current continuity. This is in distinct contrast to the piecewise linear basis, for which lumping gives an M-matrix unconditionally but produces an edge- as opposed to cell-based balance relation. We also note that this one-dimensional cell- and edge-based FEM (5.21) foreshadows the equivalence with NEM that we will verify in two dimensions.

Two alternative treatments of the mass matrix have appeared in the literature. Specifically, the nonstandard Radau quadrature suggested by Hennart [30, 36] also yields $\widetilde{\mathrm{M}_{i}}$. The work of Arbogast and Chen [3] suggests that the required approximation can be viewed as a projection,

$$
\widetilde{\mathrm{m}_{i}}=\left(\mathcal{P}_{W_{h}}\left(\psi_{\left(\xi_{1}\right)}^{\left(m_{1}\right)}\right), \mathcal{P}_{W_{h}}\left(\psi_{\left(\xi_{2}\right)}^{\left(m_{2}\right)}\right)\right)
$$

In this case the projection operator $\mathcal{P}()$ is simply the integral average.
A treatment of the mass terms is entirely avoided with conservative diffusion, and moreover, the previously obscured correspondence of the cell- and edge-based FEM with NIM becomes evident. For completeness we first state the conservative forms of Definition 5.2.2 and Definition 5.2 .3 (i.e. $\bar{\lambda}_{i}^{(x)} \rightarrow 0$ ).

Corollary 5.4. The cell- and edge-based polynomial FEM for conservative diffusion yields the following flux based balance equation,

$$
\begin{equation*}
-6 \frac{D_{i}}{\Delta x_{i}} \phi_{i-\frac{1}{2}}+12 \frac{D_{i}}{\Delta x_{i}} \phi_{i}-6 \frac{D_{i}}{\Delta x_{i}} \phi_{i+\frac{1}{2}}=\Delta x_{i} Q_{i} \tag{5.23}
\end{equation*}
$$

while at the edges $x=x_{i+\frac{1}{2}}$ we have,

$$
2 \frac{D_{i}}{\Delta x_{i}} \phi_{i-\frac{1}{2}}-6 \frac{D_{i}}{\Delta x_{i}} \phi_{i}+4\left(\frac{D_{i}}{\Delta x_{i}}+\frac{D_{i+1}}{\Delta x_{i+1}}\right) \phi_{i+\frac{1}{2}}-6 \frac{D_{i+1}}{\Delta x_{i+1}} \phi_{i+1}+2 \frac{D_{i+1}}{\Delta x_{i+1}} \phi_{i+\frac{3}{2}}=0
$$

Corollary 5.5. The reduced edge-based polynomial FEM for conservative one-group diffusion is given by the following edge-based equation,

$$
\begin{equation*}
-\frac{D_{i}}{\Delta x_{i}} \phi_{i-\frac{1}{2}}+\left(\frac{D_{i}}{\Delta x_{i}}+\frac{D_{i+1}}{\Delta x_{i+1}}\right) \phi_{i+\frac{1}{2}}-\frac{D_{i+1}}{\Delta x_{i+1}} \phi_{i+\frac{3}{2}}=\frac{1}{2} \Delta x_{i} Q_{i}+\frac{1}{2} \Delta x_{i+1} Q_{i+1} \tag{5.24}
\end{equation*}
$$

Lemma 5.2. The finite element solution $\phi_{h}(x)$ obtained with the one-dimensional polynomial basis [Definition 5.2.1] is the exact solution of the one-dimensional, one-group, conservative diffusion equation if both the diffusion coefficient $D(x)$ and the source $Q(x)$ are constant over each cell.

Proof. The exact solution is piecewise quadratic and is, therefore, included in the trial space. Since the variational formulation chooses the optimal $H^{1}(\Omega)$ element of the trial space, the FEM solution will also be the exact solution.

Remark 5.2. It is straightforward to obtain the exact solution of the problem defined in Lemma 5.2 in explicit form:

$$
\begin{equation*}
\phi_{h}(x)=\frac{\left(x_{i+\frac{1}{2}}-x\right)}{\Delta x_{i}} \phi_{i-\frac{1}{2}}+\frac{\left(x-x_{i-\frac{1}{2}}\right)}{\Delta x_{i}} \phi_{i+\frac{1}{2}}+\frac{1}{2 D_{i}}\left(x-x_{i-\frac{1}{2}}\right)\left(x_{i+\frac{1}{2}}-x\right) Q_{i} \tag{5.25}
\end{equation*}
$$

where $x_{i-\frac{1}{2}} \leq x \leq x_{i+\frac{1}{2}}$. This expression (5.25) clearly demonstrates that having $Q_{i} \geq 0$ for $i=1, \ldots, L$, and $\phi_{h}\left(x_{i+\frac{1}{2}}\right) \geq 0$ for $i=1, \ldots, L-1$, are the necessary and sufficient conditions to have $\phi_{h}(x) \geq 0 \forall x \in \Omega$. Therefore, having unconditionally that $\mathcal{A}$ is an Mmatrix is sufficient to ensure that for $Q(x) \geq 0$ we obtain, unconditionally, a nonnegative FEM solution.

## Equivalence in Two Dimensions

We extend the one-dimensional analysis and demonstrate the equivalence between the following cell- and edge-based nonconforming finite element method and the NEM discretization. The two-dimensional cell and edge basis functions on the canonical element are defined by (5.10) using the one-dimensional basis functions of Definition 5.2.1. The corresponding global basis functions are defined according to (5.11). This relationship with the one- dimensional basis greatly simplifies the evaluation of the elemental stiffness and mass matrices, reducing most integrals to those of the previous section. It also gives a clear indication of the relevance of the one-dimensional study.

Evaluating the elemental stiffness matrix we obtain,

$$
S_{i, j}=2 D_{i, j}\left[\begin{array}{c|cc|cc}
6\left[\varrho_{i, j}^{(x, y)}\right]^{-1} & -3 r_{i, j} & -3 r_{i, j} & -3 \bar{r}_{i, j} & -3 \bar{r}_{i, j}  \tag{5.26}\\
\hline-3 r_{i, j} & 2 r_{i, j} & r_{i, j} & 0 & 0 \\
-3 r_{i, j} & r_{i, j} & 2 r_{i, j} & 0 & 0 \\
\hline-3 \bar{r}_{i, j} & 0 & 0 & 2 \bar{r}_{i, j} & \bar{r}_{i, j} \\
-3 \bar{r}_{i, j} & 0 & 0 & \bar{r}_{i, j} & 2 \bar{r}_{i, j}
\end{array}\right]
$$

where $\varrho_{i, j}^{(x, y)}=\Delta x_{i} \Delta y_{j} /\left(\Delta x_{i}^{2}+\Delta y_{j}^{2}\right)$ is given in Definition 3.4.4, and we have defined, $r_{i, j}=\Delta y_{j} / \Delta x_{i}$ and $\bar{r}_{i, j}=\Delta x_{i} / \Delta y_{j}$. Exact integration of the elemental mass terms gives,

$$
M_{i, j}=\frac{\left(\Sigma_{a}\right)_{i, j} \Delta x_{i} \Delta y_{j}}{2}\left[\begin{array}{c|cc|cc}
\frac{14}{5} & -\frac{1}{5} & -\frac{1}{5} & -\frac{1}{5} & -\frac{1}{5}  \tag{5.27}\\
\hline-\frac{1}{5} & \frac{4}{15} & -\frac{1}{15} & 0 & 0 \\
-\frac{1}{5} & -\frac{1}{15} & \frac{4}{15} & 0 & 0 \\
\hline-\frac{1}{5} & 0 & 0 & \frac{4}{15} & -\frac{1}{15} \\
-\frac{1}{5} & 0 & 0 & -\frac{1}{15} & \frac{4}{15}
\end{array}\right]
$$

Observe that there is no direct coupling of horizontal and vertical edges. In addition, the failure to satisfy the cell balance equation (3.8) and the continuity of normal currents first
identified in the 1D case is also found here. Thus, equivalence with the nodal expansion method relies on an approximate treatment of the mass terms, and in particular lumping gives

$$
\widetilde{\mathbf{M}_{i, j}}=\frac{\left(\Sigma_{a}\right)_{i, j} \Delta x_{i} \Delta y_{j}}{2}\left[\begin{array}{c|cc|cc}
2 & 0 & 0 & 0 & 0  \tag{5.28}\\
\hline 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
\hline 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{array}\right]
$$

and leads to the following conclusion.

Theorem 5.1. The cell and edge polynomial basis yields a nonconforming finite element discretization of the one-group diffusion equation that is equivalent to the NEM discretization [Definition 3.4.4] if the mass matrix is lumped.

Proof. The coefficient of $\phi_{i, j}$ in the flux based balance equation is simply,

$$
\frac{12 D_{i, j}}{\varrho_{i, j}^{(x, y)}}+\left(\Sigma_{a}\right)_{i, j} \Delta x_{i} \Delta y_{j}=\frac{12 D_{i, j}}{\varrho_{i, j}^{(x, y)}}\left\{1+\frac{\left(\Sigma_{a}\right)_{i, j} \Delta x_{i} \Delta y_{j}}{12 D_{i, j}} \varrho_{i, j}^{(x, y)}\right\}=12 D_{i, j} \frac{\mu_{i, j}^{(x, y)}}{\varrho_{i, j}^{(x, y)}}
$$

All other coefficients are read directly from (5.26) and thus, dividing through by $\Delta x_{i} \Delta y_{j}$, equivalence with Definition 3.4.4 follows.

The properties of the exact and lumped discretizations parallel those of the onedimensional case. However, there is one significant exception: $\mathcal{A}$ is not an M-matrix. Yet these discretizations have been used extensively in reactor modelling, an application in which it is required that $\phi_{h}(x, y) \geq 0$, with great success. Therefore, we investigated the inverse of a small constant coefficient conservative Dirichlet problem using MAPLE [17] and found that indeed $\mathcal{A}^{-1}>0$. We conjecture that with the exact evaluation of the stiffness and mass matrices, $\mathcal{A}^{-1}>0$ unconditionally, while for the lumped case
a condition similar to that found in 1D exists. Note that nonhomogeneous boundary conditions could still produce a nonphysical solution, but in many practical cases the BCs are homogeneous.

### 5.2.3 Analytic Basis Functions

The motivation of enforcing nodal balance through the naturally arising edge and cell based unknowns is taken one step further in analytic nodal methods. Specifically, we recall that, under the assumption of piecewise constant coefficients and source, the transverse integrated ODEs are solved exactly. In terms of the nonconformal finite element methods this suggests that the basis be spanned by the homogeneous and particular solutions of these transverse integrated ODEs [37]. Thus, we note that in the case of conservative diffusion (i.e. $\Sigma_{a} \equiv 0$ ), the polynomial and analytic bases are equivalent independent of dimension.

## A One-Dimensional Study

Following the treatment of the polynomial FEMs we begin with an investigation of the analytic basis functions in one dimension.

Definition 5.2.4. In one dimension the cell and edge analytic basis on the canonical element $\Omega_{c}$ are given by,

$$
\begin{align*}
\psi_{(-1)}^{(*)}(\xi) & =-\frac{1}{2} \frac{\delta_{i}^{(x)}}{\left[1-\delta_{i}^{(x)}\right]}-\frac{1}{2} \frac{\sinh \left(\bar{\lambda}_{i}^{(x)} \xi\right)}{\sinh \left(\bar{\lambda}_{i}^{(x)}\right)}+\frac{1}{2} \frac{1}{\left[1-\delta_{i}^{(x)}\right]} \frac{\cosh \left(\bar{\lambda}_{i}^{(x)} \xi\right)}{\cosh \left(\bar{\lambda}_{i}^{(x)}\right)}  \tag{5.29a}\\
\psi_{(1)}^{(*)}(\xi) & =-\frac{1}{2} \frac{\delta_{i}^{(x)}}{\left[1-\delta_{i}^{(x)}\right]}+\frac{1}{2} \frac{\sinh \left(\bar{\lambda}_{i}^{(x)} \xi\right)}{\sinh \left(\bar{\lambda}_{i}^{(x)}\right)}+\frac{1}{2} \frac{1}{\left[1-\delta_{i}^{(x)}\right]} \frac{\cosh \left(\bar{\lambda}_{i}^{(x)} \xi\right)}{\cosh \left(\bar{\lambda}_{i}^{(x)}\right)}  \tag{5.29b}\\
\psi_{(*)}^{(0)}(\xi) & =\frac{1}{\left[1-\delta_{i}^{(x)}\right]}\left\{1-\frac{\cosh \left(\bar{\lambda}_{i}^{(x)} \xi\right)}{\cosh \left(\bar{\lambda}_{i}^{(x)}\right)}\right\} \tag{5.29c}
\end{align*}
$$

where we recall $\delta_{i}^{(x)}=\left[\bar{\lambda}_{i}^{(x)}\right]^{-1} \tanh \left(\bar{\lambda}_{i}^{(x)}\right)$.

Derivation of the $1 D$ Analytic Basis. We consider the basis $\left\{1, \sinh \left(\bar{\lambda}_{i}^{(x)} \xi\right), \cosh \left(\bar{\lambda}_{i}^{(x)} \xi\right)\right\}$, which contains the homogeneous and particular solutions of the piecewise constant ODE on the canonical element $\Omega_{c}$. Thus a general canonical basis function is written,

$$
\psi(\xi)=a_{0}+a_{1} \sinh \left(\bar{\lambda}_{i}^{(x)} \xi\right)+a_{2} \cosh \left(\bar{\lambda}_{i}^{(x)} \xi\right)
$$

yielding the system,

$$
\left[\begin{array}{ccc}
1 & -\sinh \left(\bar{\lambda}_{i}^{(x)}\right) & \cosh \left(\bar{\lambda}_{i}^{(x)}\right) \\
1 & \sinh \left(\bar{\lambda}_{i}^{(x)}\right) & \cosh \left(\bar{\lambda}_{i}^{(x)}\right) \\
1 & 0 & \frac{1}{\bar{\lambda}_{i}^{(x)}} \sinh \left(\bar{\lambda}_{i}^{(x)}\right)
\end{array}\right]\left[\begin{array}{l}
a_{0} \\
a_{1} \\
a_{2}
\end{array}\right]=e_{i}
$$

With a little help from MAPLE [17] we obtain the explicit formulae given in (5.29).

Evaluating the integrals of the stiffness matrix we obtain,

$$
\mathrm{S}_{i}=2 \frac{D_{i}}{\Delta x_{i}}\left[\begin{array}{c|cc}
\gamma_{i}^{(x, s)} & -\frac{1}{2} \alpha_{i}^{(x, s)} & -\frac{1}{2} \alpha_{i}^{(x, s)} \\
\hline-\frac{1}{2} \alpha_{i}^{(x, s)} & \frac{1}{4}\left\{\alpha_{i}^{(x, s)}+\beta_{i}^{(x, s)}\right\} & \frac{1}{4}\left\{\alpha_{i}^{(x, s)}-\beta_{i}^{(x, s)}\right\} \\
-\frac{1}{2} \alpha_{i}^{(x, s)} & \frac{1}{4}\left\{\alpha_{i}^{(x, s)}-\beta_{i}^{(x, s)}\right\} & \frac{1}{4}\left\{\alpha_{i}^{(x, s)}+\beta_{i}^{(x, s)}\right\}
\end{array}\right]
$$

where we have conveniently used the following definitions,

$$
\begin{aligned}
& \alpha_{i}^{(x, s)}=\gamma_{i}^{(s)}=\left[\bar{\lambda}_{i}^{(x)}\right]^{3} \frac{\left\{\sinh \left(\bar{\lambda}_{i}^{(x)}\right) \cosh \left(\bar{\lambda}_{i}^{(x)}\right)-\bar{\lambda}_{i}^{(x)}\right\}}{\left\{\bar{\lambda}_{i}^{(x)} \cosh \left(\bar{\lambda}_{i}^{(x)}\right)-\sinh \left(\bar{\lambda}_{i}^{(x)}\right)\right\}^{2}} \\
& \beta_{i}^{(x, s)}=\left[\bar{\lambda}_{i}^{(x)}\right] \frac{\left\{\sinh \left(\bar{\lambda}_{i}^{(x)}\right) \cosh \left(\bar{\lambda}_{i}^{(x)}\right)+\bar{\lambda}_{i}^{(x)}\right\}}{\sinh ^{2}\left(\bar{\lambda}_{i}^{(x)}\right)}
\end{aligned}
$$

Similarly, the elemental construction of the mass matrix may be expressed in the form,

$$
\mathrm{M}_{i}=\frac{\sigma_{i} \Delta x_{i}}{2}\left[\begin{array}{c|cc}
\gamma_{i}^{(x, m)} & -\frac{1}{2} \alpha_{i}^{(x, m)} & -\frac{1}{2} \alpha_{i}^{(x, m)} \\
\hline-\frac{1}{2} \alpha_{i}^{(x, m)} & \frac{1}{4}\left\{\alpha_{i}^{(x, m)}+\beta_{i}^{(x, m)}\right\} & \frac{1}{4}\left\{\alpha_{i}^{(x, m)}-\beta_{i}^{(x, m)}\right\} \\
-\frac{1}{2} \alpha_{i}^{(x, m)} & \frac{1}{4}\left\{\alpha_{i}^{(x, m)}-\beta_{i}^{(x, m)}\right\} & \frac{1}{4}\left\{\alpha_{i}^{(x, m)}+\beta_{i}^{(x, m)}\right\}
\end{array}\right]
$$

where,

$$
\begin{aligned}
& \alpha_{i}^{(x, m)}=\frac{\sinh \left(\bar{\lambda}_{i}^{(x)}\right)}{\left\{\bar{\lambda}_{i}^{(x)} \cosh \left(\bar{\lambda}_{i}^{(x)}\right)-\sinh \left(\bar{\lambda}_{i}^{(x)}\right)\right\}}+\frac{\left[\bar{\lambda}_{i}^{(x)}\right]^{2}-\sinh ^{2}\left(\bar{\lambda}_{i}^{(x)}\right)}{\left\{\bar{\lambda}_{i}^{(x)} \cosh \left(\bar{\lambda}_{i}^{(x)}\right)-\sinh \left(\bar{\lambda}_{i}^{(x)}\right)\right\}^{2}} \\
& \beta_{i}^{(x, m)}=\frac{\sinh \left(\bar{\lambda}_{i}^{(x)}\right) \cosh \left(\bar{\lambda}_{i}^{(x)}\right)-\bar{\lambda}_{i}^{(x)}}{\bar{\lambda}_{i}^{(x)} \sinh ^{2}\left(\bar{\lambda}_{i}^{(x)}\right)} \\
& \gamma_{i}^{(x, m)}=\frac{2 \bar{\lambda}_{i}^{(x)} \cosh \left(\bar{\lambda}_{i}^{(x)}\right)}{\left\{\bar{\lambda}_{i}^{(x)} \cosh \left(\bar{\lambda}_{i}^{(x)}\right)-\sinh \left(\bar{\lambda}_{i}^{(x)}\right)\right\}}+\frac{\left\{\left[\bar{\lambda}_{i}^{(x)}\right]^{2}-\bar{\lambda}_{i}^{(x)} \cosh \left(\bar{\lambda}_{i}^{(x)}\right) \sinh \left(\bar{\lambda}_{i}^{(x)}\right)\right\}}{\left\{\bar{\lambda}_{i}^{(x)} \cosh \left(\bar{\lambda}_{i}^{(x)}\right)-\sinh \left(\bar{\lambda}_{i}^{(x)}\right)\right\}^{2}}
\end{aligned}
$$

hence the total elemental matrix may be written,

$$
\mathrm{A}_{i}=\frac{D_{i}}{\Delta x_{i}}\left[\begin{array}{c|cc}
4 \gamma_{i}^{(x)} & -2 \alpha_{i}^{(x)} & -2 \alpha_{i}^{(x)}  \tag{5.30}\\
\hline-2 \alpha_{i}^{(x)} & \left\{\alpha_{i}^{(x)}+\beta_{i}^{(x)}\right\} & \left\{\alpha_{i}^{(x)}-\beta_{i}^{(x)}\right\} \\
-2 \alpha_{i}^{(x)} & \left\{\alpha_{i}^{(x)}-\beta_{i}^{(x)}\right\} & \left\{\alpha_{i}^{(x)}+\beta_{i}^{(x)}\right\}
\end{array}\right]
$$

where,

$$
\begin{aligned}
& \alpha_{i}^{(x)}=\frac{1}{2}\left(\alpha_{i}^{(x, s)}+\left[\bar{\lambda}_{i}^{(x)}\right]^{2} \alpha_{i}^{(x, m)}\right)=\frac{\left[\bar{\lambda}_{i}^{(x)}\right]^{2} \delta_{i}^{(x)}}{\left[1-\delta_{i}^{(x)}\right]} \\
& \beta_{i}^{(x)}=\frac{1}{2}\left(\beta_{i}^{(x, s)}+\left[\bar{\lambda}_{i}^{(x)}\right]^{2} \beta_{i}^{(x, m)}\right)=\frac{1}{\delta_{i}^{(x)}} \\
& \gamma_{i}^{(x)}=\frac{1}{2}\left(\gamma_{i}^{(x, s)}+\left[\bar{\lambda}_{i}^{(x)}\right]^{2} \gamma_{i}^{(x, m)}\right)=\frac{\left[\bar{\lambda}_{i}^{(x)}\right]^{2}}{\left[1-\delta_{i}^{(x)}\right]}
\end{aligned}
$$

Although the elemental matrices provide a compact representation of the discretization they obscure many of the interesting properties. Thus, in the following discussion we first present the corresponding stencils.

Definition 5.2.5. The cell- and edge-based analytic FEM for one-dimensional, onegroup, diffusion is defined by the following local cell-based balance equation,

$$
\begin{equation*}
-2 \frac{D_{i}}{\Delta x_{i}} \alpha_{i}^{(x)} \phi_{i-\frac{1}{2}}+4 \frac{D_{i}}{\Delta x_{i}} \gamma_{i}^{(x)} \phi_{i}-2 \frac{D_{i}}{\Delta x_{i}} \alpha_{i}^{(x)} \phi_{i+\frac{1}{2}}=\Delta x_{i} Q_{i} \tag{5.31}
\end{equation*}
$$

and edge-based stencils of the form,

$$
\begin{align*}
{\left[\frac{D_{i}}{\Delta x_{i}}\left(\alpha_{i}^{(x)}-\beta_{i}^{(x)}\right)\right] \phi_{i-\frac{1}{2}} } & -2 \frac{D_{i}}{\Delta x_{i}} \alpha_{i}^{(x)} \phi_{i} \\
+\left[\frac { D _ { i } } { \Delta x _ { i } } \left(\alpha_{i}^{(x)}\right.\right. & \left.\left.+\beta_{i}^{(x)}\right)+\frac{D_{i+1}}{\Delta x_{i+1}}\left(\alpha_{i+1}^{(x)}+\beta_{i+1}^{(x)}\right)\right] \phi_{i+\frac{1}{2}}  \tag{5.32}\\
& -2 \frac{D_{i+1}}{\Delta x_{i+1}} \alpha_{i+1}^{(x)} \phi_{i+1}+\left[\frac{D_{i+1}}{\Delta x_{i+1}}\left(\alpha_{i+1}^{(x)}-\beta_{i+1}^{(x)}\right)\right] \phi_{i+\frac{3}{2}}=0
\end{align*}
$$

Definition 5.2.6. The reduced edge-based analytic FEM for one-group diffusion is given by the following edge-based equation,

$$
\begin{equation*}
-c_{i} \phi_{i-\frac{1}{2}}+\left(d_{i}+d_{i+1}\right) \phi_{i+\frac{1}{2}}-c_{i+1} \phi_{i+\frac{3}{2}}=f_{i}+f_{i+1} \tag{5.33a}
\end{equation*}
$$

where,

$$
\begin{align*}
& c_{i}=\frac{D_{i}}{\Delta x_{i}}\left\{\left[1-\delta_{i}^{(x)}\right] \alpha_{i}^{(x)}-\beta_{i}^{(x)}\right\}=\frac{D_{i}}{\Delta x_{i}} \frac{\lambda_{i}^{(x)}}{\sinh \left(\lambda_{i}^{(x)}\right)}  \tag{5.33b}\\
& d_{i}=\frac{D_{i}}{\Delta x_{i}}\left\{\left[1-\delta_{i}^{(x)}\right] \alpha_{i}^{(x)}+\beta_{i}^{(x)}\right\}=\frac{D_{i}}{\Delta x_{i}} \frac{\lambda_{i}^{(x)} \cosh \left(\lambda_{i}^{(x)}\right)}{\sinh \left(\lambda_{i}^{(x)}\right)}  \tag{5.33c}\\
& f_{i}=\frac{1}{2} \delta_{i}^{(x)} \Delta x_{i} Q_{i} \tag{5.33d}
\end{align*}
$$

Derivation of the $1 D$ analytic FEMs. The coefficients in equations (5.31) and (5.32) are simply read from the total elemental matrix (5.30). The support of $\Psi_{i}^{(0)}(x)$ is $\Omega_{i}$ and hence $\phi_{i}$ may be eliminated locally. Solving for the cell average in (5.31) gives,

$$
\begin{equation*}
\phi_{i}=\frac{1}{2} \delta_{i}^{(x)}\left(\phi_{i-\frac{1}{2}}+\phi_{i+\frac{1}{2}}\right)+\frac{\left[1-\delta_{i}^{(x)}\right]}{\left[\bar{\lambda}_{i}^{(x)}\right]^{2}} \frac{\Delta x_{i}^{2} Q_{i}}{4 D_{i}} \tag{5.34}
\end{equation*}
$$

and hence substitution into (5.32) yields (5.33a). The simplified form of the coefficients is obtained by applying the definitions of $\alpha_{i}^{(x)}, \beta_{i}^{(x)}$ and $\delta_{i}^{(x)}$ along with a few hyperbolic trigonometric identities.

Similarly to Lemma 5.1 we obtain

Lemma 5.3. The global coefficient matrix $\mathcal{A}$ of the $1 D$ reduced edge-based analytic FEM (5.33a) is an M-matrix independent of $\bar{\lambda}_{i}^{(x)}$.

Without any special treatment of the mass terms the analytic basis functions unconditionally yield an M-matrix. In addition, Fick's law (5.5) verifies that the cell balance equation is satisfied and that the currents are continuous over the entire domain. Thus, we have attained both the physical and mathematical objectives of the discretization. This is readily understood with the consideration of the following Lemma.

Lemma 5.4. The finite element solution $\phi_{h}(x)$ obtained with the one-dimensional analytic basis [Definition 5.2.4] is the exact solution of the one-dimensional, one-group diffusion equation if the coefficients $D(x), \Sigma_{a}(x)$ and the source $Q(x)$ are constant over each cell.

Proof. As in Lemma 5.2 the trial space includes the exact solution.

Remark 5.3. It is possible to show that the exact solution to the problem defined in Lemma 5.4 may be written as

$$
\begin{align*}
\phi_{h}(x) & =\frac{\sinh \left(\lambda_{i}^{(x)}\left[x_{i+\frac{1}{2}}-x\right]\right)}{\sinh \left(\lambda_{i}^{(x)}\right)} \phi_{i-\frac{1}{2}}  \tag{5.35}\\
& +\frac{\sinh \left(\lambda_{i}^{(x)}\left[x-x_{i-\frac{1}{2}}\right]\right)}{\sinh \left(\lambda_{i}^{(x)}\right)} \phi_{i+\frac{1}{2}}+\left\{1-\frac{\cosh \left(\lambda_{i}^{(x)}\left[x-x_{i}\right]\right)}{\cosh \left(\lambda_{i}^{(x)}\right)}\right\} \frac{Q_{i}}{\sigma_{i}}
\end{align*}
$$

where $x_{i-\frac{1}{2}} \leq x \leq x_{i+\frac{1}{2}}$. This provides a straightforward verification that having $Q_{i} \geq 0$ for $i=1, \ldots, L$, and $\phi_{h}\left(x_{i+\frac{1}{2}}\right) \geq 0$ for $i=1, \ldots, L-1$, are the necessary and sufficient conditions to have $\phi_{h}(x) \geq 0 \forall x \in \Omega$. Therefore, having unconditionally that $\mathcal{A}$ is an Mmatrix is sufficient to ensure that for $Q(x) \geq 0$ we obtain, unconditionally, a nonnegative FEM solution. Thus, we see that the analytic basis functions yield a discretization that
captures the desirable physical and mathematical properties, even in the nonconservative case, while the polynomial basis does not.

Finally we examine the conservative limit of the total elemental matrix $\mathcal{A}$ to verify that it corresponds to the total elemental matrix of the polynomial FEM (5.16). We also demonstrate that a definite relationship exists between the analytic and polynomial FEMs. Specifically, performing the Taylor expansion of the elements of $A_{i}$ (5.30) we find,

$$
\begin{aligned}
\alpha_{i}^{(x)} & \approx\left\{3+\frac{1}{5}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}-\frac{1}{175}\left[\bar{\lambda}_{i}^{(x)}\right]^{4}+\mathcal{O}\left[\bar{\lambda}_{i}^{(x)}\right]^{6}\right\} \\
\gamma_{i}^{(x)} & \approx\left\{3+\frac{6}{5}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}-\frac{1}{175}\left[\bar{\lambda}_{i}^{(x)}\right]^{4}+\mathcal{O}\left[\bar{\lambda}_{i}^{(x)}\right]^{6}\right\} \\
\frac{1}{2}\left(\alpha_{i}^{(x)}+\beta_{i}^{(x)}\right) & \approx\left\{2+\frac{4}{15}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}-\frac{22}{1575}\left[\bar{\lambda}_{i}^{(x)}\right]^{4}+\mathcal{O}\left[\bar{\lambda}_{i}^{(x)}\right]^{6}\right\} \\
\frac{1}{2}\left(\alpha_{i}^{(x)}-\beta_{i}^{(x)}\right) & \approx\left\{1-\frac{1}{15}\left[\bar{\lambda}_{i}^{(x)}\right]^{2}-\frac{13}{1575}\left[\bar{\lambda}_{i}^{(x)}\right]^{4}+\mathcal{O}\left[\bar{\lambda}_{i}^{(x)}\right]^{6}\right\}
\end{aligned}
$$

and hence truncating the expansions at second order in $\bar{\lambda}_{i}^{(x)}$ gives precisely the elemental matrix obtained with the polynomial basis functions (5.16). Furthermore, evaluating the limit $\bar{\lambda}_{i}^{(x)} \rightarrow 0$ we obtain identical systems from the polynomial and analytic FEMs.

## Equivalences in Two Dimensions

Following the polynomial case [Section 5.2.2] we now extend the one-dimensional analysis and demonstrate the equivalence between the following cell and edge-based nonconforming finite element method and the NIM discretization. The two-dimensional cell and edge analytic basis functions on the canonical element are defined by (5.10) using the one-dimensional basis given in Definition 5.2.4. Once again the corresponding global functions are defined according to (5.11). The relationship between the one- and twodimensional basis functions that was first utilized in the polynomial case facilitates the exact evaluation of the elemental stiffness and mass matrices. Specifically, utilizing the

1D results the stiffness matrix may be written,

$$
S_{i, j}=2 D_{i, j}\left[\begin{array}{c|cc|cc}
\Lambda_{i, j}^{(s)} & -\frac{1}{2} r_{i, j} \alpha_{i}^{(x, s)} & -\frac{1}{2} r_{i, j} \alpha_{i}^{(x, s)} & -\frac{1}{2} \bar{r}_{i, j} \alpha_{i}^{(y, s)} & -\frac{1}{2} \bar{r}_{i, j} \alpha_{i}^{(y, s)}  \tag{5.36}\\
\hline-\frac{1}{2} r_{i, j} \alpha_{i, j}^{(x, s)} & \frac{1}{4} r_{i, j} \Theta_{i, j}^{(x, s)} & \frac{1}{4} r_{i, j} \Xi_{i, j}^{(x, s)} & 0 & 0 \\
-\frac{1}{2} r_{i, j} \alpha_{i, j}^{(x, s)} & \frac{1}{4} r_{i, j} \Xi_{i, j}^{(x, s)} & \frac{1}{4} r_{i, j} \Theta_{i, j}^{(x, s)} & 0 & 0 \\
\hline-\frac{1}{2} \bar{r}_{i, j} \alpha_{i}^{(y, s)} & 0 & 0 & \frac{1}{4} \bar{r}_{i, j} \Theta_{i, j}^{(y, s)} & \frac{1}{4} \bar{r}_{i, j} \Xi_{i, j}^{(y, s)} \\
-\frac{1}{2} \bar{r}_{i, j} \alpha_{i}^{(y, s)} & 0 & 0 & \frac{1}{4} \bar{r}_{i, j} \Xi_{i, j}^{(y, s)} & \frac{1}{4} \bar{r}_{i, j} \Theta_{i, j}^{(y, s)}
\end{array}\right]
$$

where we have defined,

$$
\Lambda_{i, j}^{(s)}=r_{i, j} \gamma_{i, j}^{(x, s)}+\bar{r}_{i, j} \gamma_{i, j}^{(y, s)}, \quad \Theta_{i, j}^{(*, s)}=\alpha_{i, j}^{(*, s)}+\beta_{i, j}^{(*, s)}, \quad \Xi_{i, j}^{(*, s)}=\alpha_{i, j}^{(*, s)}-\beta_{i, j}^{(*, s)}
$$

with $*=x$, or $*=y$. Similarly the mass matrix is given by

$$
\mathrm{M}_{i, j}=\frac{\left(\Sigma_{a}\right)_{i, j} \Delta x_{i} \Delta y_{j}}{2}\left[\begin{array}{c|cc|cc}
\Lambda_{i, j}^{(m)} & -\frac{1}{2} \alpha_{i}^{(x, m)} & -\frac{1}{2} \alpha_{i}^{(x, m)} & -\frac{1}{2} \alpha_{i}^{(y, m)} & -\frac{1}{2} \alpha_{i}^{(y, m)}  \tag{5.37}\\
\hline-\frac{1}{2} \alpha_{i, j}^{(x, m)} & \frac{1}{4} \Theta_{i, j}^{(x, m)} & \frac{1}{4} \Xi_{i, j}^{(x, m)} & 0 & 0 \\
-\frac{1}{2} \alpha_{i, j}^{(x, m)} & \frac{1}{4} \Xi_{i, j}^{(x, m)} & \frac{1}{4} \Theta_{i, j}^{(x, m)} & 0 & 0 \\
\hline-\frac{1}{2} \alpha_{i}^{(y, m)} & 0 & 0 & \frac{1}{4} \Theta_{i, j}^{(y, m)} & \frac{1}{4} \Xi_{i, j}^{(y, m)} \\
-\frac{1}{2} \alpha_{i}^{(y, m)} & 0 & 0 & \frac{1}{4} \Xi_{i, j}^{(y, m)} & \frac{1}{4} \Theta_{i, j}^{(y, m)}
\end{array}\right]
$$

where we have defined,

$$
\Lambda_{i, j}^{(m)}=\gamma_{i, j}^{(x, m)}+\gamma_{i, j}^{(y, m)}, \quad \Theta_{i, j}^{(*, m)}=\alpha_{i, j}^{(*, m)}+\beta_{i, j}^{(*, m)}, \quad \Xi_{i, j}^{(*, m)}=\alpha_{i, j}^{(*, m)}-\beta_{i, j}^{(*, m)}
$$

Finally the total elemental matrix may be written,

$$
\mathrm{A}_{i, j}=D_{i, j}\left[\begin{array}{c|cc|cc}
\gamma_{i, j}^{(x, y)} & -2 r_{i, j} \alpha_{i}^{(x)} & -2 r_{i, j} \alpha_{i}^{(x)} & -2 \bar{r}_{i, j} \alpha_{i}^{(y)} & -2 \bar{r}_{i, j} \alpha_{i}^{(y)}  \tag{5.38}\\
\hline-2 r_{i, j} \alpha_{i, j}^{(x)} & r_{i, j} \Theta_{i, j}^{(x)} & r_{i, j} \Xi_{i, j}^{(x)} & 0 & 0 \\
-2 r_{i, j} s \alpha_{i, j}^{(x)} & r_{i, j} \Xi_{i, j}^{(x)} & r_{i, j} \Theta_{i, j}^{(x)} & 0 & 0 \\
\hline-2 \bar{r}_{i, j} \alpha_{i}^{(y)} & 0 & 0 & \bar{r}_{i, j} \Theta_{i, j}^{(y)} & \bar{r}_{i, j} \Xi_{i, j}^{(y)} \\
-2 \bar{r}_{i, j} \alpha_{i}^{(y)} & 0 & 0 & \bar{r}_{i, j} \Xi_{i, j}^{(y)} & \bar{r}_{i, j} \Theta_{i, j}^{(y)}
\end{array}\right]
$$

where

$$
\Theta_{i, j}^{(*, s)}=\alpha_{i, j}^{(*)}+\beta_{i, j}^{(*)}, \quad \Xi_{i, j}^{(*, s)}=\alpha_{i, j}^{(*)}-\beta_{i, j}^{(*)}
$$

which trivially yields the following theorem.

Theorem 5.2. The cell and edge analytic basis yields a nonconformal finite element discretization of the one-group diffusion equation that is equivalent to the constant-constant NIM discretization [Definition 3.4.4].

We note that this equivalence is stronger than that stated by Hennart [37] as it doesn't require the application of a nonstandard Radau quadrature to the mass matrix, and hence, assures that $\left\|e_{h}(x)\right\|_{1}$ is minimized over the trial space $V_{h}$. In addition equivalence indicates that the continuity of the normal current and the cell-based balance equation follow. However, we do not have an M-matrix. Similar to the conjecture in Section 5.2.2 our one-dimensional analysis suggests that $\mathcal{A}^{-1}>0$ unconditionally. Unfortunately, thus far we have been unsuccessful in proving this conjecture.

### 5.3 Mixed Hybrid Finite Element Methods

An apparent weakness of the primal approach is that it abandons the rigorous treatment of cell balance which is central to the nodal view. In contrast a mixed FEM treats
the first order form (2.12), viewing Fick's law (2.12b) as a constraint of the balance equation (2.12a) and hence explicitly enforcing cell balance. Hybrid FEMs are a special class of mixed FEMs which temporarily relax certain inter-element continuity conditions, ultimately enforcing them in the weak variational sense. Thus, the hybridization of a mixed method captures the essence of nodal ideology and strongly suggests that an equivalence with NEM and NIM exists. Indeed such an equivalence has been established for the lowest order NEM by Hennart and del Valle [39] and we outline their results in the following section. A discussion of the issues surrounding a potential equivalence with NIM is also presented.

### 5.3.1 Variational Formulation

The typical (e.g. [5, 3]) introduction of mixed-hybrid FEMs is facilitated by first presenting the mixed formulation followed by its hybridization. However, a stronger connection to nodal discretizations is apparent if we define the hybridization of the first-order system (2.12) and then proceed directly to the mixed-hybrid variational formulation. To this end we denote the set of edges $E$ of the rectangles $\Omega_{i, j}$ as $\mathcal{E}_{h}$. Then the sets of boundary and interior edges are defined by $\mathcal{E}_{h}^{\partial}=\left\{E \in \mathcal{E}_{h} \mid E \subset \partial \Omega\right\}$ and $\mathcal{E}_{h}^{0}=\mathcal{E}_{h} \backslash \mathcal{E}_{h}^{\partial}$, respectively. The strong formulation of the corresponding hybrid one-group diffusion equation is given by,

$$
\begin{array}{rlr}
{[D(\mathbf{r})]^{-1} \mathbf{J}+\nabla \phi=0} & \forall \Omega_{i, j} \in \Omega_{h} \\
\nabla \cdot \mathbf{J}+\Sigma_{a}(\mathbf{r}) \phi=Q & \forall \Omega_{i, j} \in \Omega_{h} \\
\left.\mathbf{J} \cdot \mathbf{n}\right|_{E^{+}}-\left.\mathbf{J} \cdot \mathbf{n}\right|_{E^{-}}=0 & \forall E \in \mathcal{E}_{h}^{0} \tag{5.39c}
\end{array}
$$

subject to boundary conditions. To avoid unnecessary complications in the following
theoretical discussion we consider homogeneous Dirichlet boundary conditions,

$$
\phi(\mathbf{r})=0, \quad \forall \mathbf{r} \in \partial \Omega
$$

We have adopted the standard ordering of these equations, with Fick's law (5.39a) stated first, followed by the balance equation (5.39b), and finally the constraint that enforces the continuity of normal currents (5.39c). In addition, note that Fick's law has been divided through by the diffusion coefficient so that the harmonic average will arise from subsequent integration. In Chapter 4 it was demonstrated that in one dimension this is the correct averaging, although no rigorous extension to two dimensions exists. Moreover, this is the form of Fick's law [equation (2.10)] that was obtained with the $P_{1}$ approximation of the transport equation [Section 2.2.1].

We define the following function spaces:

$$
\begin{aligned}
R_{-1}\left(\Omega_{h}\right) & =\left\{\mathbf{q} \mid \mathbf{q} \in L^{2}(\Omega) \times L^{2}(\Omega), \nabla \cdot \mathbf{q} \in L^{2}\left(\Omega_{i, j}\right) \forall \Omega_{i, j} \in \Omega_{h}\right\} \\
R_{0}\left(\Omega_{h}\right) & =\left\{\mathbf{q} \mid \mathbf{q} \in R_{-1}\left(\Omega_{h}\right),\langle\mathbf{q} \cdot \mathbf{n}\rangle_{E^{+}}=\langle\mathbf{q} \cdot \mathbf{n}\rangle_{E^{-}} \forall E \in \mathcal{E}_{h}^{0}\right\} \\
U_{-1}\left(\Omega_{h}\right) & =\left\{\varphi \mid \varphi \in L^{2}\left(\Omega_{i, j}\right), \quad \forall \Omega_{i, j} \in \Omega_{h}\right\} \\
M_{-1}\left(\mathcal{E}_{h}\right) & =\left\{\mu\left|\mu \in L^{2}\left(\mathcal{E}_{h}\right), \mu\right|_{E}=0 \quad \forall E \in \mathcal{E}_{h}^{\partial}\right\}
\end{aligned}
$$

to facilitate the statement of the well known (cf. [14, 15]) weak mixed-hybrid variational formulation of (5.39) as:
find $\left(\mathbf{J}_{h}, \phi_{h}, \mu_{h}\right) \in R\left(\Omega_{h}\right) \times U\left(\Omega_{h}\right) \times M\left(\mathcal{E}_{h}^{0}\right)$ such that

$$
\begin{array}{cl}
\left(D^{-1} \mathbf{J}_{h}, \mathbf{q}_{h}\right)-\sum_{i, j}\left\{\left(\phi_{h}, \nabla \cdot \mathbf{q}_{h}\right)_{\Omega_{i, j}}-\left\langle\mu_{h}, \mathbf{q}_{h} \cdot \mathbf{n}\right\rangle_{\partial \Omega_{i, j}}\right\}=-\left\langle g, \mathbf{q}_{h} \cdot \mathbf{n}\right\rangle . & \forall \mathbf{q}_{h} \in R\left(\Omega_{h}\right) \\
-\sum_{i, j}\left(\nabla \cdot \mathbf{J}_{h}, \varphi_{h}\right)_{\Omega_{i, j}}-\left(\Sigma_{a} \phi_{h}, \varphi_{h}\right)=-\left(Q, \varphi_{h}\right) & \forall \varphi_{h} \in U\left(\Omega_{h}\right) \\
\sum_{i, j}\left\langle\nu, \mathbf{J}_{h} \cdot \mathbf{n}\right\rangle_{\partial \Omega_{i, j}}=0 & \forall \nu_{h} \in M\left(\mathcal{E}_{h}\right)
\end{array}
$$

where $R \subset R_{-1}, U \subset U_{-1}, M \subset M_{-1}$ are particular choices of finite dimensional subspaces. We further restrict the choice of subspaces such that the (weak) continuity
constraint for the normal currents ensures that we find $\mathbf{J}_{h} \in R_{0}\left(\Omega_{h}\right) \subset H(\operatorname{div} ; \Omega)$, a conforming approximation of the corresponding mixed variational problem. Selecting a particular set of basis functions and performing the integration yields a system of the following form,

$$
\left[\begin{array}{ccc}
A & B^{T} & C^{T}  \tag{5.40}\\
B & -E & 0 \\
C & 0 & 0
\end{array}\right]\left[\begin{array}{l}
\boldsymbol{J}_{h} \\
\boldsymbol{\phi}_{h} \\
\boldsymbol{\mu}_{h}
\end{array}\right]=-\left[\begin{array}{c}
\boldsymbol{Q}_{\mathbf{J}} \\
\boldsymbol{Q}_{\phi} \\
0
\end{array}\right]
$$

Like the mixed FEM, this system is indefinite; however, unlike the mixed FEM, $A$ is block diagonal with each block corresponding to a cell $\Omega_{i, j}$. The implications of this property on the sparsity of the reduced system are discussed in detail in Chapter 6 .

### 5.3.2 Polynomial Basis Functions

Recall the lowest-order Raviart-Thomas bases [60], which may be written

$$
\begin{aligned}
R T_{-1}^{0}\left(\Omega_{h}\right) & =\left\{\left.\mathrm{q}\right|_{\Omega_{i, j}} \in\{\{1, \xi\},\{1, \eta\}\} \forall \Omega_{i, j} \in \Omega_{h}\right\} \bigcap R_{-1} \\
U_{-1}^{0}\left(\Omega_{h}\right) & =\left\{\left.\varphi\right|_{\Omega_{i, j}} \in\{1\} \quad \forall \Omega_{i, j} \in \Omega_{h}\right\} \bigcap U_{-\mathbf{1}} \\
M_{-1}^{0}\left(\Omega_{h}\right) & =\left\{\left.\mu\right|_{E} \in\{1\} \quad \forall E \in \mathcal{E}_{h}^{0}\right\} \bigcap M_{-1}
\end{aligned}
$$

The following theorem is due to Hennart [39].

Theorem 5.3. The lowest-order Raviart-Thomas bases yield a mixed-hybrid finite element discretization of the one-group diffusion equation that is equivalent to the NEM discretization [Definition 3.4.4].

Proof. Utilizing the lowest-order Raviart-Thomas bases and assuming piecewise constant coefficients it is straightforward to perform the integration of the variational form exactly and obtain the following matrices. The matrix $A$ is cell-based and hence has a block
diagonal structure with each block itself being block diagonal. These blocks are given by,

$$
\left[A_{i, j}\right]_{k, k}=\frac{\Delta x_{i} \Delta y_{j}}{4 D_{i, j}}\left[\begin{array}{cc}
A_{i, j}^{(x)} & 0  \tag{5.41}\\
0 & A_{i, j}^{(y)}
\end{array}\right], \quad A_{i, j}^{(x)}=A_{i, j}^{(y)}=\left[\begin{array}{ll}
2 & 1 \\
1 & 2
\end{array}\right]
$$

where the index $k$ references the diagonal block of $A$. The matrix $B$ is also cell-based, and block diagonal, although the blocks are degenerate in the lowest-order case (i.e. [ $1 \times 4$ ]),

$$
\left[B_{i, j}\right]_{k, k}=\left[\Delta y_{j},-\Delta y_{j}, \Delta x_{i},-\Delta x_{i}\right]
$$

The last cell-based entry in (5.40) is $E$ which in this lowest-order case is strictly diagonal with entries,

$$
\left[E_{i, j}\right]_{k, k}=\left(\Sigma_{a}\right)_{i, j}
$$

Finally, the matrix $C$ is edge-based, and is composed of a two blocks corresponding to the horizontal and vertical edges, $C=\left[C^{(x)} \mid C^{(y)}\right]^{T}$. Each of these blocks is bidiagonal, composed of the following cell-based $1 \times 4$ blocks,

$$
\begin{array}{ll}
C_{i, j}^{(x,+)}=\left[-\Delta y_{j}, 0,0,0\right], & C_{i, j}^{(x,-)}=\left[0, \Delta y_{j}, 0,0\right] \\
C_{i, j}^{(y,+)}=\left[0,0,-\Delta x_{i}, 0\right], & C_{i, j}^{(y,-)}=\left[0,0,0, \Delta x_{i}\right]
\end{array}
$$

where the $\pm$ signs indicate which side of the edge lies within $\Omega_{i, j}$.
The block bidiagonal structure of $C$ enforces the continuity of normal currents (3.24). Moreover, a direct evaluation of $B \mathbf{J}_{h}-E \phi_{h}=-Q_{h}$ gives the balance equation (3.8). The definition of the one-sided currents (3.23) is obtained from $A \mathbf{J}_{h}+B^{T} \phi_{h}+C \mu_{h}=0$ with the diagonalization of $A$ and the realization that the Lagrange multipliers are the edge unknowns of the NEM.

### 5.3.3 Analytic Basis Functions

When developing preconditioners for the nodal discretizations [Chapter 6] we make extensive use of the indefinite formulation (5.40). Unfortunately, the equivalence of NIM with a particular mixed-hybrid FEM has not been established. Although a strict equivalence would be ideal, for our purposes it is sufficient to cast the NIM system in this indefinite form.

Definition 5.3.1. The indefinite formulation of the constant-constant NIM discretization for nonconservative one-group diffusion [Definition 3.5.1] is given by (5.40) with,

$$
A_{i, j}=\frac{\Delta x_{i} \Delta y_{j}}{4 D_{i, j}}\left[\begin{array}{cc}
A_{i, j}^{(x)} & 0  \tag{5.42}\\
0 & A_{i, j}^{(y)}
\end{array}\right]
$$

where

$$
A_{i, j}^{(x)}=\frac{1}{\alpha_{i, j}^{(x)} \beta_{i, j}^{(x)}}\left[\begin{array}{ll}
\left(\alpha_{i, j}^{(x)}+\beta_{i, j}^{(x)}\right) & \left(\alpha_{i, j}^{(x)}-\beta_{i, j}^{(x)}\right)  \tag{5.43}\\
\left(\alpha_{i, j}^{(x)}-\beta_{i, j}^{(x)}\right) & \left(\alpha_{i, j}^{(x)}+\beta_{i, j}^{(x)}\right)
\end{array}\right]
$$

and $A_{i, j}^{(y)}$ is similarly defined. The matrices $B, C$ and $E$ are defined in Section 5.3.2.
Derivation of the Indefinite Formulation. We readily obtain the indefinite form of the constant-constant NIM discretization [Definition 3.5.1] from the one-sided net currents, (3.45). Specifically, we consider adding $\left(\alpha_{i, j}^{(x)}+\beta_{i, j}^{(x)}\right)(3.45 \mathrm{~b})$ and $\left(\alpha_{i, j}^{(x)}-\beta_{i, j}^{(x)}\right)(3.45 \mathrm{a})$ to obtain,

$$
\left(\alpha_{i, j}^{(x)}+\beta_{i, j}^{(x)}\right) J_{i-\frac{1}{2}, j}^{+}+\left(\alpha_{i, j}^{(x)}-\beta_{i, j}^{(x)}\right) J_{i+\frac{1}{2}, j}^{-}=-4 \frac{D_{i, j}}{\Delta x_{i}} \alpha_{i, j}^{(x)} \beta_{i, j}^{(x)}\left(\phi_{i,}-\phi_{i-\frac{1}{2}, j}\right)
$$

Thus multiplying through by $\left[\Delta x_{i} \Delta y_{j}\right]\left[4 D_{i, j}\right]^{-1}$ gives the first row of $A_{i, j}^{(x)}$ defined in (5.43). Similarly adding $\left(\alpha_{i, j}^{(x)}-\beta_{i, j}^{(x)}\right)(3.45 \mathrm{~b})$ and $\left(\alpha_{i, j}^{(x)}+\beta_{i, j}^{(x)}\right)$ (3.45a) yields the second row $A_{i, j}^{(x)}$ given in (5.43). Finally, $B \boldsymbol{J}_{h}-E \boldsymbol{\phi}_{h}=-\boldsymbol{Q}_{h}$ is simply the balance equation (3.8), while $C \boldsymbol{J}_{h}=0$ enforces the continuity of the normal currents.

## Chapter 6

## Reduced Systems and Approximate Schur Complements

An important objective of this research is the development of fast iterative solvers for the nodal discretizations (equivalently the mixed-hybrid FEM discretization). One obstacle in this pursuit has become apparent through the abstract formulation of the mixed-hybrid FEMs, namely the linear system is indefinite. Fortunately, this indefiniteness is readily circumvented with the block elimination of $\boldsymbol{J}_{h}$ in (5.40) which yields

$$
\mathcal{S}_{(\phi, \mu)}\left[\begin{array}{l}
\phi_{h}  \tag{6.1}\\
\boldsymbol{\mu}_{h}
\end{array}\right]=\boldsymbol{Q}_{\mathcal{S}_{(\phi, \mu)}}
$$

with

$$
\mathcal{S}_{(\phi, \mu)}=\left[\begin{array}{cc}
\mathcal{S}_{B} & B A^{-1} C^{T} \\
C A^{-1} B^{T} & \mathcal{S}_{C}
\end{array}\right], \quad \boldsymbol{Q}_{\mathcal{S}_{(\phi, \mu)}}=-\left[\begin{array}{c}
B A^{-1} \boldsymbol{Q}_{\mathbf{J}}-\boldsymbol{Q}_{\phi} \\
C A^{-1} \boldsymbol{Q}_{\mathbf{J}}
\end{array}\right]
$$

where $\mathcal{S}_{B}=B A^{-1} B^{T}+E$ and $\mathcal{S}_{C}=C A^{-1} C^{T}$. Thus far, the elementwise block diagonal structure of $A$, and hence $A^{-1}$, has preserved the relative sparsity of the system. In fact, this is an abstract representation of the flux formulation that was given for NEM and NIM in Definitions 3.4.4 and 3.5.2 respectively. Thus, the exact sparsity structure of $\mathcal{S}_{(\phi, \mu)}$ is displayed in Figure 3.3. Perhaps more importantly, the Schur complement $\mathcal{S}_{(\phi, \mu)}$ is symmetric positive definite indicating that it is an ideal candidate for iterative methods. In addition, the equivalence of this formulation with nonconformal FEMs was established in Section 5.2. However, the majority of reported works using the nodal, mixed-hybrid and nonconformal methods proceed to eliminate the cell-based unknowns $\{$ e.g. $[38,50,18]\}$, so we consider this case first.

### 6.1 Lumping The Edge-Based Schur Complement

Proceeding to eliminate the cell-based unknowns $\phi_{h}$ from (6.1) leads to the Schur complement of the edge unknowns,

$$
\begin{equation*}
\mathcal{S}_{\mu} \boldsymbol{\mu}_{h}=\boldsymbol{Q}_{\mathcal{S}_{\mu}} \tag{6.2}
\end{equation*}
$$

where

$$
\begin{aligned}
\mathcal{S}_{\mu} & =C\left(A^{-1}-A^{-1} B^{T} \mathcal{S}_{B}^{-1} B A^{-1}\right) C^{T} \\
\boldsymbol{Q}_{\mathcal{S}_{\mu}} & =C\left(A^{-1} B^{T} \mathcal{S}_{B}^{-1} B A^{-1}-A^{-1}\right) \boldsymbol{Q}_{\mathbf{J}}-C A^{-1} B^{T} \mathcal{S}_{B}^{-1} \boldsymbol{Q}_{\phi}
\end{aligned}
$$

Note that (6.2) is precisely the edge-based formulation that was defined in Definitions 3.4.5, 3.5.3 and shown schematically in Figure 3.4. The relative sparsity of the original system (5.40) continues to be preserved in $\mathcal{S}_{\mu}$ because $\mathcal{S}_{B}$ is strictly diagonal (Figure 3.3). Yet, once again the most important property of this reduced system is that it is symmetric positive definite. In fact, it is this combination that has made (6.2) the most widely studied formulation of the mixed-hybrid finite element system $\{$ e.g. $[50,18]\}$. However, as we shall see, it is not without its disadvantages.

To understand the inherent complexity of creating fast iterative solvers for (6.2) and to motivate our approach we partition the edge unknowns by orientation, vertical and horizontal, $\boldsymbol{\mu}_{h}=\{\boldsymbol{u}, \boldsymbol{v}\}$ such that (6.2) may be rewritten,

$$
\left[\begin{array}{ll}
A_{u u} & A_{u v}  \tag{6.3}\\
A_{v u} & A_{v v}
\end{array}\right]\left[\begin{array}{l}
\boldsymbol{u} \\
\boldsymbol{v}
\end{array}\right]=\left[\begin{array}{l}
\boldsymbol{Q}_{u} \\
\boldsymbol{Q}_{v}
\end{array}\right]
$$

where $A_{u u}, A_{v v}$ are tridiagonal and $A_{u v}$ is block bidiagonal, with each block itself bidiago$\operatorname{nal}\left(A_{v u}=A_{u v}^{T}\right)$. ADI methods appear to be a natural choice and have been investigated by a number of authors $[38,8]$. In addition, most standard preconditioners such as polynomial and incomplete factorizations have been studied [56]. However, none of these
methods can attain the efficiency of multi-level solvers, the development of which has been hindered by the difficult hierarchy of grids and inter-grid transfer operators which must be defined for these edge-based stencils. Certainly multi-level solvers would be much easier to develop if only a single orientation of the edge unknowns remained. However, the choice to eliminate either $\boldsymbol{u}$ or $\boldsymbol{v}$ is arbitrary and moreover it is apparent that the respective Schur Complements,

$$
\mathcal{S}_{u}=A_{u u}-A_{v u}\left(A_{v v}\right)^{-1} A_{u v}, \quad \mathcal{S}_{v}=A_{v v}-A_{u v}\left(A_{u u}\right)^{-1} A_{v u}
$$

suffer an equivalent loss of sparsity as a result of the appearance of $\left(A_{v v}\right)^{-1}$ and $\left(A_{u u}\right)^{-1}$, respectively. Thus, posed purely as a problem in linear algebra, a suitable objective is to minimize the fill generated during the formation of $\mathcal{S}_{u}$ (or $\mathcal{S}_{v}$ ). Further noting that each diagonal block of $\left(A_{u u}\right)^{-1}$ is pre and post multiplied by bidiagonal matrices, the minimal fill attainable through approximations of $\left(A_{u u}\right)^{-1}$ results when it is diagonal. Determining the optimal diagonal approximation is a foreboding task. However, recalling that the best incomplete Cholesky factorizations preserve row sums it is reasonable to construct a diagonal approximation, $\widetilde{A_{u u}}$, which is composed simply of row sums of $A_{u u}$. A simple inversion then gives $\left(\widetilde{A_{u u}}\right)^{-1} \approx\left(A_{u u}\right)^{-1}$. Hence the preconditioner may be written as

$$
\widetilde{\mathcal{S}_{\mu ; u}}=\left[\begin{array}{cc}
\widetilde{A_{u z}} & A_{u v}  \tag{6.4}\\
A_{v u} & A_{v v}
\end{array}\right]
$$

and the reduced system $\widetilde{\mathcal{S}_{v}}=A_{v v}-A_{v u}\left(\widetilde{A_{u u}}\right)^{-1} A_{u v}$ is symmetric positive definite. In addition, we note that it is straightforward to prove that $\mathcal{S}_{v}$ is a second-order 9-point approximation of an elliptic PDE. Thus, efficient multigrid algorithms exist to "invert" it. In practice we use Dendy's black box multigrid code [23], a robust solver which requires only the fine grid stencil as input.

### 6.1.1 A Condition Bound

We prove two results independently. The first is a proposition which makes the connection between the eigenvalues of the preconditioner and the difference between $A_{u u}$ and its approximation.

Proposition 6.1. The eigenvalues of the preconditioned system, $\lambda\left(\left[\widetilde{\mathcal{S}_{\mu ; u}}\right]^{-1} \mathcal{S}_{\mu}\right)$ may be written in the following form,

$$
\begin{aligned}
& \lambda=1, \text { of multiplicity } N_{v} \\
& \lambda=\lambda\left(\mathcal{S}_{\tilde{u}}^{-1} \mathcal{S}_{u}\right)=1+\lambda\left(\mathcal{S}_{\tilde{u}}^{-1} A_{u u}^{*}\right)
\end{aligned}
$$

where $\mathcal{S}_{\tilde{u}}=\widetilde{A_{u u}}-A_{u v}\left(A_{v v}\right)^{-1} A_{v u}$ and $A_{u u}^{*}=A_{u u}-\widetilde{A_{u u}}$.
Proof. Observe that the inverse of the preconditioner $\widetilde{\mathcal{S}_{\mu ; u}}$ may be expressed as

$$
\left[\widetilde{\mathcal{S}_{\mu ; u}}\right]^{-1}=\left[\begin{array}{ll}
\widetilde{A_{u u}} & A_{u v} \\
A_{v u} & A_{v v}
\end{array}\right]^{-1}=\left[\begin{array}{cc}
\mathcal{S}_{\tilde{u}}^{-1} & -\mathcal{S}_{\tilde{u}}^{-1} A_{u v} A_{v v}^{-1} \\
-\left[\widetilde{\mathcal{S}_{v}}\right]^{-1} A_{v u}\left[\widetilde{A_{u u}}\right]^{-1} & {\left[\widetilde{\mathcal{S}_{v}}\right]^{-1}}
\end{array}\right]
$$

where $\widetilde{\mathcal{S}_{v}}=A_{v v}-A_{v u}\left(\widetilde{A_{u u}}\right)^{-1} A_{u v}$. Block multiplication and simplification give

$$
\left[\widetilde{\mathcal{S}_{\mu ; u}}\right]^{-1} \mathcal{S}_{\mu}=\left[\begin{array}{cc}
\mathcal{S}_{\tilde{u}}^{-1} \mathcal{S}_{u} & 0 \\
-A_{v v}^{-1} A_{v u} \mathcal{S}_{\tilde{u}}^{-1} A_{u u}^{*} & I
\end{array}\right]
$$

where $A_{u u}^{*}=A_{u u}-\widetilde{A_{u u}}$. Hence the eigenvalues are, $\lambda=1$, with multiplicity $N_{v}$ and $\lambda=\lambda\left(\mathcal{S}_{\tilde{u}}^{-1} \mathcal{S}_{u}\right)$. Since $\mathcal{S}_{u}=\mathcal{S}_{\tilde{u}}+A_{u u}^{*}$ we have $\mathcal{S}_{\tilde{u}}^{-1} \mathcal{S}_{u}=I+\mathcal{S}_{\bar{u}}^{-1} A_{u u}^{*}$.

Remark 6.1. Note that the preconditioner is symmetric positive definite, and hence may be factored: $\widetilde{\mathcal{S}_{\mu ; u}}=\mathcal{F} \mathcal{F}^{T}$. Thus, the preconditioned system is $\mathcal{G}=\mathcal{F}^{-1} \mathcal{S}_{\mu} \mathcal{F}^{-T}$, and moreover, it is similar to $\left[\widetilde{\mathcal{S}_{\mu ; u}}\right]^{-1} \mathcal{S}_{\mu}$. Consequently, to bound $\kappa(\mathcal{G})$ it is sufficient to bound the eigenvalues of $\left[\widetilde{\mathcal{S}_{\mu ; u}}\right]^{-1} \mathcal{S}_{\mu}$, and in fact, this is standard practice in CG related estimates.

Next we bound the condition number of the edge-based preconditioned system, independently of the mesh size, the diffusion coefficient and the absorption cross-section. We begin with the conservative case and consider the nonconservative case in subsequent Corollaries.

Theorem 6.1. Let $r_{i, j}=\Delta y_{j} / \Delta x_{i}$. Then the condition number of the preconditioned system with $\Sigma_{a} \equiv 0$ is bounded by

$$
\begin{equation*}
\kappa\left(\left[\widetilde{\mathcal{S}_{\mu ; u}}\right]^{-1} \mathcal{S}_{\mu}\right) \leq \max _{r_{i, j}}\left\{\left(1+r_{i, j}^{2}\right), 3\right\} . \tag{6.5}
\end{equation*}
$$

With a uniform grid spacing $r=r_{i, j}$, an even tighter bound is realized:

$$
\begin{equation*}
\kappa\left(\left[\widetilde{\mathcal{S}_{\mu ; u}}\right]^{-1} \mathcal{S}_{\mu}\right) \leq \max \left\{\frac{1}{3}\left(1+r^{2}\right), \frac{3}{\left(1+r^{2}\right)}\right\} \tag{6.6}
\end{equation*}
$$

Proof. We employ a superelement analysis as proposed by Kuznetsov [50] to obtain a local bound on the condition number $\kappa$. Consider first the elemental construction of the edge-based system by either utilizing its equivalence with a nonconformal method or by elementwise elimination of the currents and the cell-based unknowns from the original system (5.40). On a single cell $\Omega_{i, j}$ we have

$$
\mathcal{S}_{\mu}^{\mathrm{s}}=\left[\begin{array}{cc}
A_{u u}^{\mathrm{s}} & A_{u v}^{\mathrm{s}} \\
A_{v u}^{\mathrm{s}} & A_{v v}^{\mathrm{s}}
\end{array}\right]=\left[\begin{array}{cc|cc}
\beta+\alpha & \beta-\alpha & -\beta & -\beta \\
\beta-\alpha & \beta+\alpha & -\beta & -\beta \\
\hline-\beta & -\beta & \beta+\gamma & \beta-\gamma \\
-\beta & -\beta & \beta-\gamma & \beta+\gamma
\end{array}\right]
$$

with $\alpha=D_{i, j} r_{i, j}, \beta=3 D_{i, j} \varrho_{i, j}^{(x, y)}, \gamma=D_{i, j} \bar{r}_{i, j}, \bar{r}_{i, j}=\left[r_{i, j}\right]^{-1}$ and where the unknowns on $\Omega_{i, j}$ have been ordered as $\left[\phi_{i-\frac{1}{2}, j}, \phi_{i+\frac{1}{2}, j}, \phi_{i, j-\frac{1}{2}}, \phi_{i, j+\frac{1}{2}}\right]$. The lumped approximation gives

$$
\left.\left[\widetilde{\mathcal{S}_{\mu ; u}}\right]^{\mathrm{s}}=\left[\begin{array}{cc}
{\left[\widetilde{A_{u u}}\right.}
\end{array}\right]^{\mathrm{s}} A_{u v}^{\mathrm{s}}, \quad\left[\widetilde{A_{u u}}\right]^{\mathrm{s}}=\left[\begin{array}{cc}
2 \beta & 0 \\
A_{v u}^{\mathrm{s}} & A_{v v}^{\mathrm{s}}
\end{array}\right], \quad 2 \beta\right]
$$

It is straightforward to verify $\mathcal{K}=\operatorname{ker}\left(\mathcal{S}_{\mu}^{\mathbf{s}}\right)=\operatorname{ker}\left(\left[\widetilde{\mathcal{S}_{\mu ; u}}\right]^{\mathrm{s}}\right)=[1,1,1,1]^{T}$. This kernel is to be expected as $\mathcal{S}_{\mu}^{\mathrm{s}}$ essentially defines a pure Neumann problem on the cell, $\Omega_{i, j}$. Finally
we consider the generalized eigenvalue problem,

$$
\mathcal{S}_{\mu}^{\mathrm{s}} \boldsymbol{\mu}_{\mathrm{s}}=\lambda_{\mathrm{s}}\left[\widetilde{\mathcal{S}_{\mu ; u}}\right]^{\mathrm{s}} \boldsymbol{\mu}_{\mathrm{s}} \quad \text { with } \boldsymbol{\mu}_{\mathrm{s}} \perp \mathcal{K}
$$

Applying Theorem 6.3.1 of Rao and Mitra [59] we find the that three "proper" eigenvalues are $\lambda_{\mathrm{s}}=\left[1,1, \frac{1}{3}\left(1+r_{i, j}^{2}\right)\right]$. If $r_{i, j}>\sqrt{2} \Rightarrow \lambda_{\mathrm{s}} \in\left[1, \frac{1}{3}\left(1+r_{i, j}^{2}\right)\right]$ while if $r_{i, j}<\sqrt{2} \Rightarrow \lambda_{\mathrm{s}} \in$ $\left[\frac{1}{3}\left(1+r_{i, j}^{2}\right), 1\right]$. Hence, if the mesh spacing is constant superelement analysis trivially yields the bound given in (6.6). Whereas on a spatially dependent mesh we consider the following,

$$
\begin{aligned}
& r_{i, j}>\sqrt{2} \forall(i, j) \Rightarrow \kappa \leq \max _{r_{i, j}}\left\{\frac{1}{3}\left(1+r_{i, j}^{2}\right)\right\} \\
& r_{i, j}<\sqrt{2} \forall(i, j) \Rightarrow \kappa \leq \max _{r_{i, j}}\left\{\frac{3}{\left(1+r_{i, j}^{2}\right)}\right\}<3 \\
& r_{i, j}>\sqrt{2} \text { and } r_{k, l}<\sqrt{2} \Rightarrow \kappa \leq \max _{r_{i, j}, r_{k, l}}\left\{\frac{1+r_{i, j}^{2}}{1+r_{k, l}^{2}}\right\}<\max _{r_{i, j}}\left\{1+r_{i, j}^{2}\right\}
\end{aligned}
$$

Combining these bounds yields (6.5).

Remark 6.2. The form of the bound given in (6.5) is unchanged when the analysis is extended to a diagonal diffusion tensor. The crucial difference is the definition of $r_{i, j}$, which is modified to read

$$
r_{i, j}=\sqrt{\frac{D_{i, j}^{(x)}}{D_{i, j}^{(y)}}}\left[\frac{\Delta y_{j}}{\Delta x_{i}}\right]
$$

Hence we see that the implication of poor conditioning on a spatially dependent high aspect ratio grid is also indicative of poor conditioning in the presence of strong anisotropy.

We now address the nonconservative case, for which the lowest-order NEM and constant-constant NIM are no longer equivalent.

Corollary 6.1. The condition number of the preconditioned NEM system [Definition 3.4.5] for nonconservative diffusion satisfies the bound given in (6.5) of Theorem 6.1.

Proof. Following the proof of Theorem 6.1 we consider a superelement analysis. The blocks of the elemental NEM system may be written,

$$
A_{u u}^{\mathrm{s}}=\left[\begin{array}{ll}
\zeta_{i, j}^{(x, y)}+\theta & \vartheta_{i, j}^{(x, y)}+\theta \\
\vartheta_{i, j}^{(x, y)}+\theta & \zeta_{i, j}^{(x, y)}+\theta
\end{array}\right], \quad A_{u v}^{\mathrm{s}}=A_{v u}^{\mathrm{s}}=-\left[\begin{array}{cc}
\theta & \theta \\
\theta & \theta
\end{array}\right]
$$

where $\theta=3 D_{i, j} \varrho_{i, j}^{(x, y)}\left[\mu_{i, j}^{(x, y)}\right]^{-1}$ and $A_{v v}^{\mathrm{s}}$ is defined analogously. Utilizing the definitions of $\zeta_{i, j}^{(x, y)}$ and $\vartheta_{i, j}^{(x, y)}$ we find that the lumped approximation gives,

$$
\left[\widetilde{A_{u u}}\right]^{\mathrm{s}}=\left[\begin{array}{cc}
6 D_{i, j} r_{i, j}\left\{1-\left[\mu_{i, j}^{(x, y)}\right]^{-1}\right\}+2 \theta & 0 \\
0 & 6 D_{i, j} r_{i, j}\left\{1-\left[\mu_{i, j}^{(x, y)}\right]^{-1}\right\}+2 \theta
\end{array}\right]
$$

In contrast to the conservative case, the kernel is the null space, a property that follows from the well-posedness of the Neumann problem for $\Sigma_{a}>0$ and greatly simplifies the specification and treatment of the generalized eigenvalue problem. Specifically, we find that the eigenvalues may be written,

$$
\lambda_{\mathbf{s}}=\left[1,1,1, \frac{1}{3}\left\{1+\frac{r_{i, j}^{2}}{1+\frac{1}{3}\left[\bar{\lambda}_{i, j}^{(x)}\right]^{2}}\right\}\right]
$$

Thus the only non-unity eigenvalue is bounded by

$$
\frac{1}{3}<\lambda_{\mathrm{s}}<\frac{1}{3}\left(1+r_{i, j}^{2}\right)
$$

and hence the bound in (6.5) follows.

Corollary 6.2. The condition number of the preconditioned NIM system for nonconservative diffusion [Definition 3.5.3] is bounded by,

$$
\begin{equation*}
\kappa\left(\left[\widetilde{\mathcal{S}_{\mu ; j}}\right]^{-1} \mathcal{S}_{\mu}\right) \leq \max _{r_{i, j}}\left\{\frac{10}{9}\left(1+r_{i, j}^{2}\right), \frac{10}{3}\right\} \tag{6.7}
\end{equation*}
$$

Proof. Following the proof of Theorem 6.1 we consider a superelement analysis. The blocks of the elemental NIM system may be written,

$$
A_{u u}^{\mathrm{s}}=\left[\begin{array}{cc}
\hat{\zeta}_{i, j}^{(x, y)}+\theta & \hat{\vartheta}_{i, j}^{(x, y)}+\theta \\
\hat{\vartheta}_{i, j}^{(x, y)}+\theta & \hat{\zeta}_{i, j}^{(x, y)}+\theta
\end{array}\right], \quad A_{u v}^{\mathrm{s}}=A_{v u}^{\mathrm{s}}=-\left[\begin{array}{ll}
\theta & \theta \\
\theta & \theta
\end{array}\right]
$$

where $\theta=D_{i, j} \delta_{i, j}^{(x)} \alpha_{i, j}^{(x)} \alpha_{i, j}^{(y)}\left[\gamma_{i, j}^{(x, y)}\right]^{-1}$ and $A_{v v}^{s}$ is defined analogously. Utilizing the definitions of $\hat{\zeta}_{i, j}^{(x, y)}$ and $\hat{\vartheta}_{i, j}^{(x, y)}$ we find the lumped approximation gives,

$$
\left[\widetilde{A_{u u}}\right]^{\mathrm{s}}=\left[\begin{array}{cc}
2\left(D_{i, j} r_{i, j}\left[1-\delta_{i, j}^{(x)}\right] \alpha_{i, j}^{(x)}+\theta\right) & 0 \\
0 & 2\left(D_{i, j} r_{i, j}\left[1-\delta_{i, j}^{(x)}\right] \alpha_{i, j}^{(x)}+\theta\right)
\end{array}\right]
$$

Hence the generalized eigenvalues are given by

$$
\lambda_{\mathrm{s}}=\left[1,1,1,\left\{\frac{1-\delta_{i, j}^{(x)} \delta_{i, j}^{(y)}}{\left[\bar{\lambda}_{i, j}^{(x)} \delta_{i, j}^{(x)}\right]^{2}\left[1+\delta_{i, j}^{(y)}\left(1-\delta_{i, j}^{(x)}\right)\right]}\right\}\right]
$$

To bound the non-unity eigenvalue we first note that $\bar{\lambda}_{i, j}^{(y)}=r_{i, j} \bar{\lambda}_{i, j}^{(x)}$ and hence $\lambda_{\mathrm{s}}=$ $\lambda_{\mathrm{s}}\left(\bar{\lambda}_{i, j}^{(x)}, r_{i, j}\right)$. It is then straightforward to show that,

$$
\frac{3}{10}<\lambda_{\mathbf{s}}(2,0) \leq \lambda_{\mathbf{s}} \leq \lambda_{\mathbf{s}}(0, r)=\frac{1}{3}\left(1+r_{i, j}\right)
$$

and hence we obtain (6.7).

### 6.1.2 Bounding the elements of $A_{u u}^{-1}$

The futility of searching for diagonal approximations of $A_{u u}$ which eliminate the dependence on $r_{i, j}$ may be understood if we employ Theorem 2.4 of Demko et al. [22] which bounds the exponential decay rate of elements in the inverse of a banded matrix. Specifically, we consider constant mesh spacing and constant diffusivity with Dirichlet boundary conditions for conservative diffusion such that the diagonal blocks of $A_{u u}$ are given by

$$
\begin{equation*}
\left(A_{u u}\right)_{j j}=D\left(\frac{r}{1+r^{2}}\right) \operatorname{tri}\left[2-r^{2}, 8+2 r^{2}, 2-r^{2}\right], \quad j=1, \ldots, M \tag{6.8}
\end{equation*}
$$

and of dimension $(L-1) \times(L-1)$. A bound on the elemental decay rate may be written in the form

$$
\begin{equation*}
\left|\left\{\left(A_{u u}\right)_{j j}^{-1}\right\}_{k l}\right| \leq C \lambda(r)^{|k-l|} \tag{6.9}
\end{equation*}
$$

where the decay rate is itself bounded,

$$
\begin{equation*}
\lambda(r) \leq \lambda_{\infty}(r)=\frac{\alpha^{+}-\alpha^{-}}{\alpha^{+}+\alpha^{-}}, \quad \alpha^{ \pm}=\sqrt{\left(4+r^{2}\right) \pm\left|2-r^{2}\right|} \tag{6.10}
\end{equation*}
$$

We note that the bound $\lambda_{\infty}$ is consistent with Theorem 6.1, clearly capturing its dependence on $r$. In particular, the exact nature of the approximation for $r=\sqrt{2}$ is observed with $\lambda(\sqrt{2})=0$. But perhaps most importantly, Figure 6.1 reveals that $\lambda_{\infty}(r) \rightarrow 1$ as $r \rightarrow \infty$, implying that the diagonal blocks of $A_{u u}^{-1}$ are not only dense but all the elements are of approximately the same magnitude. Thus, it is not surprising that a diagonal approximation is inadequate.



Figure 6.1: Upper bound of the condition number, $\kappa(6.6)$ and of the decay rate $\lambda_{\infty}(r)$.

### 6.2 A Two-Step Lumped Approximation

Eliminating $\boldsymbol{u}$ to obtain the Schur complement for $\boldsymbol{v}$ is an arbitrary choice and it may not always be the natural one. Given this preconditioner's dependence on the ordering of unknowns we postulate that it can be improved, especially for nonuniform grids, in much the same manner that symmetric SOR (SSOR) improves upon SOR,(i.e. by composing the two possible orderings). To this end consider two splittings of $\mathcal{S}_{\mu}$,

$$
\mathcal{S}_{\mu}=E_{1}+F_{1}, \quad \mathcal{S}_{\mu}=E_{2}+F_{2}
$$

which are employed in the following two stage iteration of the system $\mathcal{S}_{\mu} \boldsymbol{\mu}_{h}=\boldsymbol{Q}_{\mathcal{S} \mu}$,

$$
\begin{align*}
E_{1} \boldsymbol{\mu}_{h}^{\left(k+\frac{1}{2}\right)} & =-F_{1} \boldsymbol{\mu}_{h}^{(k)}+\boldsymbol{Q}_{\mathcal{S}_{\mu}}  \tag{6.11a}\\
E_{2} \boldsymbol{\mu}_{h}^{(k+1)} & =-F_{2} \boldsymbol{\mu}_{h}^{\left(k+\frac{1}{2}\right)}+\boldsymbol{Q}_{\mathcal{S}_{\mu}} \tag{6.11b}
\end{align*}
$$

Defining the residual in the standard manner, $\boldsymbol{r}^{(k)}=\boldsymbol{Q}_{\mathcal{S}_{\mu}}-\mathcal{S}_{\mu} \boldsymbol{\mu}_{h}^{(k)}$, allows (6.11a) to be rewritten in the form

$$
\begin{equation*}
\boldsymbol{\mu}_{h}^{\left(k+\frac{1}{2}\right)}=\boldsymbol{\mu}_{h}^{(k)}+E_{1}^{-1} \boldsymbol{r}^{(k)} \tag{6.12}
\end{equation*}
$$

such that substitution of (6.12) into (6.11b) gives

$$
E_{2} \boldsymbol{\mu}_{h}^{(k+1)}=-F_{2}\left(\boldsymbol{\mu}_{h}^{(k)}+E_{1}^{-1} \boldsymbol{r}^{(k)}\right)+\boldsymbol{Q}_{\mathcal{S}_{\mu}}=E_{2} \boldsymbol{\mu}_{h}^{(k)}+\left(E_{1}-F_{2}\right) E_{1}^{-1} \boldsymbol{r}^{(k)}
$$

Hence the two-stage iteration is readily expressed as a single residual equation,

$$
\begin{equation*}
\left\{E_{1}\left(E_{1}-F_{2}\right)^{-1} E_{2}\right\} \Delta \boldsymbol{\mu}_{h}^{(k+1)}=\boldsymbol{r}^{(k)} \tag{6.13}
\end{equation*}
$$

To apply this to the nodal equations we need only realize that a splitting exists for the lumped preconditioner. The splitting is described by

$$
E_{1}=\left[\begin{array}{cc}
\widetilde{A_{u u}} & A_{u v} \\
A_{v u} & A_{v v}
\end{array}\right], \quad F_{1}=\left[\begin{array}{cc}
A_{u u}^{*} & 0 \\
0 & 0
\end{array}\right]
$$

with $A_{u u}=\widetilde{A_{u u}}+A_{u u}^{*}$. Similarly we introduce the lumping in $\boldsymbol{v}$ writing $A_{v v}=\widetilde{A_{v v}}+A_{v v}^{*}$. Substitution into (6.13) yields a two step lumped preconditioner,

$$
\left.\begin{array}{rl}
\widetilde{\mathcal{S}_{\mu ; u, v}} & =\left\{\widetilde{\mathcal{S}_{\mu ; u}}\left[\widetilde{\mathcal{S}_{\mu}}\right]^{-1} \widetilde{\mathcal{S}_{\mu ; v}}\right\} \\
& =\left[\begin{array}{ll}
\widetilde{A_{u u}} & A_{u v} \\
A_{v u} & A_{v v}
\end{array}\right]\left[\begin{array}{ll}
\widetilde{A_{u u}} & \frac{A_{u v}}{A_{v u}}
\end{array} \widetilde{A}_{v v}\right.
\end{array}\right]^{-1}\left[\begin{array}{cc}
A_{u u} & \frac{A_{u v}}{A_{v u}}  \tag{6.14}\\
\widetilde{A_{v v}}
\end{array}\right]
$$

Observe that $\widetilde{\mathcal{S}_{\mu ; v}} \neq{\widetilde{\mathcal{S}_{\mu ; u}}}^{T}$ and hence $\widetilde{\mathcal{S}_{\mu ; u, v}}$ will be asymmetric in general. This is unfortunate, as it precludes the use of preconditioned conjugate gradients used elsewhere in this
work. However, there are many alternative, transform free Krylov subspace solvers available. In our numerical tests we employ preconditioned Orthores [75] with a truncation length of only five and observe excellent performance.

The following result again relates the eigenvalues of the preconditioned system to the approximation difference in the Schur complement, and like Proposition 3.1 can be easily proved.

Proposition 6.2. The eigenvalues of the preconditioned system arising from the application of $\widetilde{\mathcal{S}_{\mu ; u, v}}$ are given by,

$$
\begin{aligned}
\lambda & =1, \text { of multiplicity } N_{v} \\
\lambda & =\lambda\left(\left[\widetilde{\mathcal{S}}_{\tilde{u}}^{-1} \widetilde{\mathcal{S}}_{u}\right]^{-1}\left[\mathcal{S}_{\tilde{u}}^{-1} \mathcal{S}_{u}\right]\right) \\
& =1-\lambda\left(\left[\widetilde{\mathcal{S}}_{u}^{-1} A_{u v}\left(\widetilde{A_{v v}}\right)^{-1} A_{v v}^{*}\right]\left[\widetilde{\mathcal{S}}_{v}^{-1} A_{v u}\left(\widetilde{A_{u u}}\right)^{-1} A_{u u}^{*}\right]\right)
\end{aligned}
$$

where $\widetilde{\mathcal{S}_{\tilde{u}}}=\widetilde{A_{u u}}-A_{u v}\left(\widetilde{A_{v v}}\right)^{-1} A_{v u}$.
Unfortunately, superelement analysis is not directly applicable to this preconditioner. Moreover, an easily obtainable bound such as

$$
\lambda_{\max }\left(\left[\widetilde{\mathcal{S}_{\mu ; u, v}}\right]^{-1}\right) \leq \lambda_{\max }\left(\left[\left(\widetilde{\mathcal{S}_{\tilde{u}}}\right)^{-1} \widetilde{\mathcal{S}_{u}}\right]^{-1}\right) \lambda_{\max }\left(\left[\mathcal{S}_{\tilde{u}}\right]^{-1} \mathcal{S}_{u}\right)
$$

is too crude to be of use. However, it is very encouraging that the eigenvalues may be expressed as a perturbation from unity, with the perturbation a product involving the errors associated with the lumped approximations. These errors possess a cell-based block diagonal structure hence for every cell at least one of $A_{u u}^{*}$ and $A_{v v}^{*}$ is associated with a good approximation. Thus we anticipate $\widetilde{\mathcal{S}_{\mu ; u, v}}$ to be robust for high aspect ratio problems and we demonstrate this computationally in Section 6.5.

### 6.3 Lumping the Cell-Based Schur Complement

Eliminating $\boldsymbol{\mu}_{h}$ from (6.1) we obtain the Schur complement for the cell-based unknowns

$$
\begin{equation*}
\mathcal{S}_{\phi} \phi_{h}=\boldsymbol{Q}_{\mathcal{S}_{\phi}} \tag{6.15}
\end{equation*}
$$

where

$$
\begin{aligned}
\mathcal{S}_{\phi} & =B\left(A^{-1}-A^{-1} C^{T} \mathcal{S}_{C}^{-1} C A^{-1}\right) B^{T} \\
\boldsymbol{Q}_{\mathcal{S}_{\phi}} & =\boldsymbol{Q}_{\phi}-B\left(A^{-1}-A^{-1} C^{T} \mathcal{S}_{C}^{-1} C A^{-1}\right) \boldsymbol{Q}_{\mathbf{J}}
\end{aligned}
$$

The Schur complement $\mathcal{S}_{\phi}$ is symmetric positive definite and hence we consider the development of a preconditioner for it. Of particular importance to the robustness of the resulting solvers is that this reduced system (6.15) governs only the cell-based unknowns. However, the relative sparsity structure of the system has been compromised because $\mathcal{S}_{C}$ is block diagonal, with each block itself tridiagonal. Thus the matrix-vector multiplication, $\mathcal{S}_{\phi} \phi_{h}$ will involve the inversion of $M,(L-1) \times(L-1)$ and $L,(M-1) \times(M-1)$ tridiagonal matrices.

To develop a preconditioner for $\mathcal{S}_{\phi}$, and in particular one which is suitable for multigrid inversion, we return to the full system (5.40) and note that $\mathcal{S}_{\phi}$ is positive definite if $A$ is positive definite and range $(B) \bigcap$ range $(C)=\{\emptyset\}$. Hence, the primary objective is the minimization of fill through symmetric positive definite approximations of $A$. To this end we continue the idea of the previous section and replace $A$, a block diagonal matrix, with $\tilde{A}$ a strictly diagonal lumped approximation. For NEM (and NIM with $\Sigma_{a} \equiv 0$ ) lumping (5.41) we obtain,

$$
\tilde{A}_{i, j}=\frac{\Delta x_{i} \Delta y_{i}}{2 D_{i, j}}\left[\begin{array}{cc|cc}
1 & 0 & 0 & 0  \tag{6.16}\\
0 & 1 & 0 & 0 \\
\hline 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{array}\right]
$$

This simple approximation transforms $\mathcal{S}_{\phi} \rightarrow \widetilde{\mathcal{S}_{\phi}}$ where $\widetilde{\mathcal{S}_{\phi}}$ is in fact the well known 5point cell-centred discretization. In the nonconservative case lumping $A_{i, j}$ of the NIM discretization (5.42) gives,

$$
\tilde{A}_{i, j}=\frac{\Delta x_{i} \Delta y_{i}}{2 D_{i, j}}\left[\begin{array}{cc|cc}
\delta_{i, j}^{(x)} & 0 & 0 & 0  \tag{6.17}\\
0 & \delta_{i, j}^{(x)} & 0 & 0 \\
\hline 0 & 0 & \delta_{i, j}^{(y)} & 0 \\
0 & 0 & 0 & \delta_{i, j}^{(y)}
\end{array}\right]
$$

and hence we still obtain a symmetric 5-point cell-centred discretization that is an Mmatrix, although it has nonstandard stencil weights. Thus we have derived an approximate Schur complement that may be used as a preconditioner and which may again be efficiently inverted using a standard multigrid method.

### 6.3.1 A Condition Bound

The use of tridiagonal solves in the matrix-vector product $\mathcal{S}_{\phi} \boldsymbol{\phi}_{h}$ makes a cell-based superelement analysis impossible. Certainly using superelements defined over entire rows and columns of the mesh is possible but obtaining analytic expressions for the corresponding eigenvalues seems improbable. It is, in fact, more promising to consider the full scalar system (i.e. the lumping of $A$ in (5.40) transforms $\mathcal{S}_{(\phi, \mu)}$ of $(6.1)$ to $\widetilde{\mathcal{S}_{(\phi, \mu)}}$ ) for which we bound the condition number of the correspondingly preconditioned system by a constant independent of the mesh, the diffusion coefficient and the absorption crosssection. We begin with the conservative case, which defines a bound that will be satisfied in the nonconservative case.

Theorem 6.2. For conservative diffusion the preconditioner $\widetilde{\mathcal{S}_{(\phi, \mu)}}$ yields an effective condition number which is bounded by,

$$
\begin{equation*}
\kappa\left(\left[\widetilde{\mathcal{S}_{(\phi, \mu)}}\right]^{-1} \mathcal{S}_{(\phi, \mu)}\right) \leq 3 \tag{6.18}
\end{equation*}
$$

independently of the the mesh and the diffusion coefficient.

Proof. Once again we employ superelement analysis to relate local bounds, which are readily obtained, to global ones. We begin by writing down the elemental problem,

$$
\mathcal{S}_{(\phi, \mu)}^{\mathrm{s}}=\left[\begin{array}{c|cc|cc}
12(\alpha+\gamma) & -6 \alpha & -6 \alpha & -6 \gamma & -6 \gamma \\
\hline-6 \alpha & 4 \alpha & 2 \alpha & 0 & 0 \\
-6 \alpha & 2 \alpha & 4 \alpha & 0 & 0 \\
\hline-6 \gamma & 0 & 0 & 4 \gamma & 2 \gamma \\
-6 \gamma & 0 & 0 & 2 \gamma & 4 \gamma
\end{array}\right]
$$

with $\alpha=D_{i, j} \Delta y_{j} / \Delta x_{i}$ and $\gamma=D_{i, j} \Delta x_{i} / \Delta y_{j}$ and where the unknowns have been ordered, $\boldsymbol{\Phi}_{\mathrm{s}}=\left[\phi_{i, j}, \phi_{i-\frac{1}{2}, j}, \phi_{i+\frac{1}{2}, j}, \phi_{i, j-\frac{1}{2}}, \phi_{i, j+\frac{1}{2}}\right]^{T}$. Approximating $A$ by $\widetilde{A}$ gives,

$$
\widetilde{\mathcal{S}}_{(\phi, \mu)}^{\mathrm{s}}=\left[\begin{array}{c|cc|cc}
4(\alpha+\gamma) & -2 \alpha & -2 \alpha & -2 \gamma & -2 \gamma \\
\hline-2 \alpha & 2 \alpha & 0 & 0 & 0 \\
-2 \alpha & 0 & 2 \alpha & 0 & 0 \\
\hline-2 \gamma & 0 & 0 & 2 \gamma & 0 \\
-2 \gamma & 0 & 0 & 0 & 2 \gamma
\end{array}\right]
$$

where it is apparent that we have preserved the kernel,

$$
\mathcal{K}=\operatorname{ker}\left(\mathcal{S}_{(\phi, \mu)}^{\mathbf{s}}\right)=\operatorname{ker}\left(\widetilde{\mathcal{S}}_{(\phi, \mu)}^{\mathbf{s}}\right)=[1,1,1,1,1]^{T}
$$

Solving the corresponding generalized eigenvalue problem,

$$
\mathcal{S}_{(\phi, \mu)}^{\mathrm{s}} \boldsymbol{\Phi}_{\mathrm{s}}=\lambda_{\mathrm{s}} \widetilde{\mathcal{S}}_{(\phi, \mu)}^{\mathrm{s}} \boldsymbol{\Phi}_{\mathrm{s}} \quad \text { with } \quad \boldsymbol{\Phi}_{\mathrm{s}} \perp \mathcal{K}
$$

we obtain, $\lambda_{\mathrm{s}}=[1,1,3,3]$. Since these eigenvalues are independent of the cell indices $(i, j)$, the global bound follows immediately.

Remark 6.3. The bound (6.18) is unchanged if we extend the analysis to a diagonal diffusion tensor. The implication is that this preconditioner is robust with respect to anisotropy as well as spatial variations in the diffusion tensor.

Corollary 6.3. . For the lowest-order NEM discretization with nonconservative diffusion [Definition 3.4.4] the condition number of the preconditioned system satisfies the bound given in (6.18) of Theorem 6.2.

Proof. In the case of NEM the addition of absorption requires only a trivial change in the proof of Theorem 6.2. Specifically the $(1,1)$ elements of the superelement matrices are perturbed as follows,

$$
\begin{aligned}
12(\alpha+\gamma) & \rightarrow 12(\alpha+\gamma)+\epsilon \\
4(\alpha+\gamma) & \rightarrow 4(\alpha+\gamma)+\epsilon
\end{aligned}
$$

where $\epsilon=\left(\Sigma_{a}\right)_{i, j} \Delta x_{i} \Delta y_{j}$. This perturbation makes the corresponding pure Neumann problem well posed and thus eliminates the kernel. The eigenvalues are easily evaluated with $\lambda_{\mathrm{s}}=[1,1,1,3,3]$ and hence we obtain the bound (6.18).

Corollary 6.4. For the constant-constant NIM discretization with nonconservative diffusion the condition number of the preconditioned system satisfies the bound given in (6.18) of Theorem 6.2.

Proof. Following the proof of Theorem 6.2 we employ superelement analysis to obtain a global bound on the condition number. We begin by writing down the elemental problem,

$$
\mathcal{S}_{(\phi, \mu)}^{\mathbf{s}}=\left[\begin{array}{c|cc|cc}
4(\alpha+\gamma)+\epsilon & -2 \alpha & -2 \alpha & -2 \gamma & -2 \gamma \\
\hline-2 \alpha & (\alpha+\beta) & (\alpha-\beta) & 0 & 0 \\
-2 \alpha & (\alpha-\beta) & (\alpha+\beta) & 0 & 0 \\
\hline-2 \gamma & 0 & 0 & (\gamma+\varsigma) & (\gamma-\varsigma) \\
-2 \gamma & 0 & 0 & (\gamma-\varsigma) & (\gamma+\varsigma)
\end{array}\right]
$$

with $\alpha=D_{i, j} r_{i, j} \alpha_{i, j}^{(x)}, \gamma=D_{i, j} \bar{r}_{i, j} \alpha_{i, j}^{(y)}, \beta=D_{i, j} r_{i, j} \beta_{i, j}^{(x)}$ and $\varsigma=D_{i, j} \bar{r}_{i, j} \beta_{i, j}^{(y)}$. Approximating $A$ by $\widetilde{A}(6.17)$ gives,

$$
\widetilde{\mathcal{S}}_{(\phi, \mu)}^{\mathrm{s}}=\left[\begin{array}{c|cc|cc}
4(\alpha+\gamma) & -2 \beta & -2 \beta & -2 \varsigma & -2 \varsigma \\
\hline-2 \beta & 2 \beta & 0 & 0 & 0 \\
-2 \beta & 0 & 2 \beta & 0 & 0 \\
\hline-2 \varsigma & 0 & 0 & 2 \varsigma & 0 \\
-2 \varsigma & 0 & 0 & 0 & 2 \varsigma
\end{array}\right]
$$

Solving the corresponding generalized eigenvalue problem, we obtain,

$$
\lambda_{\mathbf{s}}=\left\{1,1,1, \frac{\alpha_{i, j}^{(x)}}{\beta_{i, j}^{(x)}}, \frac{\alpha_{i, j}^{(y)}}{\beta_{i, j}^{(y)}}\right\}
$$

Consider the eigenvalue,

$$
\lambda_{\mathrm{s}}^{(x)}=\frac{\alpha_{i, j}^{(x)}}{\beta_{i, j}^{(x)}}=\frac{\bar{\lambda}_{i, j}^{(x)} \tanh \left(\bar{\lambda}_{i, j}^{(x)}\right)}{\left[\bar{\lambda}_{i, j}^{(x)} \operatorname{coth}\left(\bar{\lambda}_{i, j}^{(x)}\right)-1\right]}
$$

which is a symmetric function of $\bar{\lambda}_{i, j}^{(x)}$. It is possible to show that $\lambda_{\mathrm{s}}^{(x)}$ has a horizontal asymptote at $\lambda_{\mathrm{s}}=1$ and a local maximum of $\lambda_{\mathrm{s}}^{(x)}(0)=3$. Moreover, it is monotonically decreasing for $\bar{\lambda}_{i, j}^{(x)}>0$ implying that it satisfies the bound $\lambda_{\mathrm{s}}^{(x)} \in(1,3)$. The other non-unity eigenvalue may be bounded in the same manner and thus we find that the condition number is bounded by (6.18)

Based on the equivalence of mixed and mixed-hybrid FEMs it is apparent that $\mathcal{S}_{\phi}$ is equivalently the reduced system obtained in mixed FEMs. Generally, it is problematic to solve this system because of the loss of sparsity. Here we are able to take advantage of the restricted geometry, incurring only the additional cost of tridiagonal solves. A sensitivity to highly variable coefficients has also been a problem for some methods. Allan et al. [2] observed this and employed a lumped diagonal preconditioner to the mixed formulation, and in taking advantage of the assumed rectangular geometry obtained results comparable to ours. However, with the use of the mixed-hybrid discretization we have found a simpler approach to a robust cell-based preconditioner which in the form $\widetilde{\mathcal{S}_{(\phi, \mu)}}$ has the potential to be extended to more general geometries.

### 6.4 Approximate Multi-level Inversion of the Preconditioner

Using either of the aforementioned preconditioners would be clearly impractical if we actually intended to "invert" them to some strict level of accuracy. However, this is unnecessary because the preconditioners are standard discretizations of (2.12) and hence standard multigrid methods yield an error reduction per iteration which is independent of the grid size. This implies a spectral equivalence which allows us to employ single

V- or W-cycle in lieu of a complete multigrid solve. We state and prove the following Lemma for completeness.

Lemma 6.1. Let $\mathcal{A}, \mathcal{B}, \mathcal{C}$ be $n \times n$ symmetric positive definite matrices, and assume

$$
\begin{equation*}
\rho\left(I-\mathcal{C}^{-1} \mathcal{B}\right) \leq c<1 \tag{6.19}
\end{equation*}
$$

where $\rho$ denotes the spectral radius. Then

$$
\begin{equation*}
\kappa\left(\mathcal{C}^{-1} \mathcal{A}\right) \leq\left\{\frac{1+c}{1-c}\right\} \kappa\left(\mathcal{B}^{-1} \mathcal{A}\right) \tag{6.20}
\end{equation*}
$$

Proof. Let $\lambda$ be an eigenvalue of $\mathcal{C}^{-1} \mathcal{B}$. Then (6.19) implies $1-c \leq \lambda \leq 1+c$ and hence we readily obtain (6.20).

The practical implication of Theorem 6.1 follows if we let $\mathcal{A}$ be the edge-based $\mathcal{S}_{\mu}$ or the cell-based $\mathcal{S}_{\phi}$, and let $\mathcal{B}$ be its symmetric preconditioner. $\mathcal{B}$ is further approximated by $\mathcal{C}$, a single multigrid cycle for which we have $c$ independent of grid size, typically $c \approx 0.1$. Thus, $\frac{1+c}{1-c} \leq 1.3$ implying that $\mathcal{C}$ is (almost) as effective as $\mathcal{B}$.

### 6.5 Numerical Experiments

A progressive test suite is presented which systematically highlights the strengths and weaknesses of each approximate Schur complement preconditioner. Beginning with a constant coefficient Dirichlet problem on the unit square [Section 6.5.1] we verify numerically that on a uniform mesh $(r=1)$ all three preconditioners exhibit mesh independent convergence. Continuing with this simple example we vary the aspect ratio of a constant mesh to confirm the breakdown of the simple edge based preconditioner $\widetilde{\mathcal{S}_{\mu ; u}}$ and the robustness of $\widetilde{\mathcal{S}_{\mu ; u, v}}$ and $\widetilde{\mathcal{S}_{\phi}}$. In the next test we consider a groundwater flow problem [Section 6.5.2] with significant jumps in the diffusion coefficient. To isolate the influence of this spatial dependence we solve this problem on a uniform grid, once again
observing mesh independent convergence. To evaluate the relative cost of each preconditioner machine timings are also presented. Finally, in Section 6.5 . 3 we present a diffusive checkerboard problem which combines significant jumps in $D(x, y)$ with a spatially dependent grid containing high aspect ratio cells. The inevitable breakdown of $\widetilde{\mathcal{S}_{\mu ; u}}$ is observed along with the robustness of $\widetilde{\mathcal{S}_{\mu ; u, v}}$ and $\widetilde{\mathcal{S}_{\phi}}$.

Note that the following relative residual convergence criteria,

$$
\frac{\left\|\boldsymbol{r}^{(k)}\right\|}{\left\|\boldsymbol{r}^{(0)}\right\|}<10^{-6}
$$

is employed in all of the numerical tests that appear in this section.

### 6.5.1 A Toy Problem

We begin with the Dirichlet problem of Section 3.7.1 in the conservative case (i.e. $\Sigma_{a} \equiv 0$ ). A comparison of the two edge-based preconditioners, and the cell-based preconditioner is presented in Table 6.1 for the domain $[0,1] \times[0,1]$ with $\alpha=\beta=2$. It is clear from these results that all preconditioners generate an average residual reduction which is independent of the mesh size. Moreover, there is no difference in the iteration counts for those runs which inverted the preconditioner exactly (i.e. columns marked C ) and those which only used a single $V(1,1,1)$ cycle (i.e. columns marked $V$ ). Thus, these solvers offer an efficiency for the solution of the nodal equations which is comparable to standard multigrid algorithms applied to standard discretizations of Poisson equations. Moreover, these results are consistent with the theoretical analysis performed in Chapter 6.

Unfortunately, we must also demonstrate the vulnerability of $\widetilde{\mathcal{S}_{\mu ; u}}$ regarding high aspect ratio cells and hence we conducted several runs in which the size of the physical domain ( $[0, a] \times[0, b]$ ) was varied while the mesh size remained fixed. The results for the preconditioned edge-based solver $\widetilde{\mathcal{S}_{\mu ; u}}$ are summarized in Table 6.2. Here it is apparent that for $r$ sufficiently close to 1 , this solver demonstrates excellent efficiency. Performance
is still excellent as $r$ approaches zero. However, just as the bound on $\kappa$ [Theorem 6.1, equation (6.6)] predicts, the efficiency is degraded as $r$ increases. Conversely, the two step lumped preconditioner $\widetilde{\mathcal{S}_{\mu ; u, v}}$ displays perfectly symmetric iteration counts on the $40 \times 40$ mesh [Table 6.3]. This symmetry in $r$ is very encouraging, particularly in light the preconditioner's asymmetry. Similarly, Table 6.4 clearly displays the $r$-independent convergence of $\widetilde{\mathcal{S}_{\phi}}$.

Table 6.1: Iteration counts for the preconditioners, $\widetilde{\mathcal{S}_{\mu ; u}}$ and $\widetilde{\mathcal{S}_{\phi}}$ for the toy problem with constant mesh spacing on the domain $[0,1] \times[0,1]$.

| Mesh Size | $\widetilde{\mathcal{S}_{\mu ; u}}$ |  | $\widetilde{\mathcal{S}_{\mu ; i, v}}$ |  | $\widetilde{\mathcal{S}_{\phi}}$ |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | C | V | C | V | C | V |
| $20 \times 20$ | 6 | 6 | 4 | 4 | 11 | 11 |
| $40 \times 40$ | 6 | 6 | 4 | 4 | 11 | 11 |
| $80 \times 80$ | 6 | 6 | 4 | 4 | 11 | 11 |

Table 6.2: Iteration counts for the lumped preconditioner, $\widetilde{\mathcal{S}_{\mu ; u}}$ with constant mesh spacing on the domain $[0, a] \times[0, b]$. $\{$ Note: $\alpha=2 / a, \beta=2 / b\}$

| Mesh Size | Aspect ratio: $r=b / a$ |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $1 / 8$ | $1 / 4$ | $1 / 2$ | 1 | 2 | 4 | 8 |
| $10 \times 10$ | 9 | 9 | 9 | 6 | 7 | 15 | 19 |
| $20 \times 20$ | 9 | 10 | 9 | 6 | 7 | 16 | 30 |
| $40 \times 40$ | 9 | 10 | 9 | 6 | 7 | 15 | 29 |

Table 6.3: Iteration counts for the two-step lumped preconditioner, $\widetilde{\mathcal{S}_{\mu ; u, v}}$ with constant mesh spacing on the domain $[0, a] \times[0, b]$. $\{$ Note: $\alpha=2 / a, \beta=2 / b\}$

| Mesh Size | Aspect ratio: $r=b / a$ |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $1 / 8$ | $1 / 4$ | $1 / 2$ | 1 | 2 | 4 | 8 |
| $10 \times 10$ | 7 | 7 | 5 | 4 | 5 | 6 | 7 |
| $20 \times 20$ | 9 | 7 | 5 | 4 | 5 | 6 | 8 |
| $40 \times 40$ | 9 | 7 | 5 | 4 | 5 | 7 | 9 |

Table 6.4: Iteration counts for the cell-based preconditioner, $\widetilde{\mathcal{S}_{\phi}}$ with constant mesh spacing on the domain $[0, a] \times[0, b]$. $\{$ Note: $\alpha=2 / a, \beta=2 / b\}$

| Mesh Size | Aspect ratio: $r=b / a$ |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $1 / 8$ | $1 / 4$ | $1 / 2$ | 1 | 2 | 4 | 8 |  |
| $10 \times 10$ | 9 | 10 | 10 | 10 | 10 | 10 | 9 |  |
| $20 \times 20$ | 10 | 10 | 11 | 11 | 11 | 11 | 10 |  |
| $40 \times 40$ | 11 | 11 | 11 | 11 | 11 | 11 | 11 |  |

### 6.5.2 Ground Water Flow

The saturated flow problem of Mosé et al. [56] serves both as an excellent test of the preconditioners developed in Section 6 and as a showcase for the discretizations themselves. The problem, shown in Figure 6.2 models the flow of a ground water through the channel formed by the impervious sides,

$$
\mathbf{J} \cdot \mathbf{n}=0 \quad \forall(x, y) \in \Gamma_{I}=\{x=0, x=100,0 \leq y \leq 100\}
$$

and driven by the externally imposed gradient,

$$
\begin{array}{ll}
\phi(x, y)=100 & \forall(x, y) \in \Gamma_{T}=\{0 \leq x \leq 100, y=100\} \\
\phi(x, y)=99 & \forall(x, y) \in \Gamma_{B}=\{0 \leq x \leq 100, y=0\} \tag{6.22}
\end{array}
$$

The flow is impeded by two regions of extremely low hydraulic conductivity. Solutions computed on a very coarse mesh of $25 \times 25$ are shown in Figure 6.3. Particularly impressive are the streamlines which are significantly more accurate than those obtained even with careful postprocessing of conforming methods [56]. Computations were performed on a uniform mesh with the resulting iteration counts presented in Table 6.5. The columns marked $C G\left(\mathcal{S}_{\mu}\right)$ and $C G\left(\mathcal{S}_{\phi}\right)$ record the performance of conjugate gradients applied directly to the reduced systems $\mathcal{S}_{\mu}$ and $\mathcal{S}_{\phi}$, respectively. This provides an indication of the problem's conditioning. For the purposes of comparison iteration counts for diagonal preconditioning of the edge-based system, denoted $\mathcal{S}_{\mu}^{D}$ are also included. Consistent
with the theoretical bounds, the entries for the three approximate Schur complement preconditioners are identical to the constant coefficient Dirichlet problem [Table 6.1].

In Table 6.6 we present machine timings for the three preconditioners ${ }^{1}$. Not surprisingly the simplest preconditioner, $\widetilde{\mathcal{S}_{\mu ; u}}$ which in this case also yields the smallest bound on the condition of the system, is the fastest. However, we have established the lack of robustness in $\widetilde{\mathcal{S}_{\mu ; u}}$ and it is encouraging to see that only approximately $50 \%$ more time is required by the robust preconditioners $\widetilde{\mathcal{S}_{\mu ; u, v}}$ and $\widetilde{\mathcal{S}_{\phi}}$. It is also interesting to note that despite having approximately twice as many unknowns in the iterative process and needing OrthoRes in place of a simple conjugate gradient solver, $\widetilde{\mathcal{S}_{\mu ; u, v}}$ is competitive with $\widetilde{\mathcal{S}_{\phi}}$. This is of particular interest to people with existing codes designed to solve the edge-based system. Finally, we note that on the $200 \times 200$ problem the approximate Schur complement preconditioners are approximately 16 times faster than the simple diagonal preconditioner.


Figure 6.2: A schematic of Mosé et al.'s [56] first example.

[^7]

Figure 6.3: Solution $\phi(x, y)$ and streamlines of Mosé et al.'s [56] first example (above)

Table 6.5: Iteration Counts with single V $(1,1,1)$ cycles

| Mesh Size | $C G\left(\mathcal{S}_{\mu}\right)$ | $\mathcal{S}_{\mu}^{D}$ | $\widetilde{\mathcal{S}_{\mu ; u}}$ | $\widetilde{\mathcal{S}_{\mu ; u, v}}$ | $\widetilde{\mathcal{S}_{\phi}}$ | $C G\left(\mathcal{S}_{\phi}\right)$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $50 \times 50$ | 795 | 342 | 6 | 4 | 11 | 667 |
| $100 \times 100$ | 1,675 | 600 | 6 | 4 | 11 | 1,709 |
| $200 \times 200$ | 3,049 | 1,027 | 6 | 4 | 11 | 4,486 |

Table 6.6: Iteration Timings, [s] for the flow problem with single $\mathrm{V}(1,1,1)$ cycles

| Mesh Size | $C G\left(\mathcal{S}_{\mu}\right)$ | $\mathcal{S}_{\mu}^{D}$ | $\widetilde{\mathcal{S}_{\mu ; u}}$ | $\widetilde{\mathcal{S}_{\mu ; u, v}}$ | $\widetilde{\mathcal{S}_{\phi}}$ | $C G\left(\mathcal{S}_{\phi}\right)$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $50 \times 50$ | 7.62 | 3.50 | 0.430 | 0.610 | 0.450 | 5.95 |
| $100 \times 100$ | 72.3 | 32.4 | 2.13 | 3.26 | 2.51 | 71.5 |
| $200 \times 200$ | 601 | 243 | 10.4 | 15.53 | 15.8 | 939 |

### 6.5.3 A Diffusive Checkerboard

The final test of the preconditioners combines both huge jumps in the diffusion coefficient with a spatially dependent grid containing high aspect ratio cells. Shown schematically in Figure 6.4 we consider a single checkerboard on the square domain $[0,24] \times[0,24]$ with $\Omega_{1}=\{(0,12) \times(0,12)\} \cup\{(12,24) \times(12,24)\}, \Omega_{2}=\{(0,12) \times(0,12)\} \cup\{(12,24) \times$ $(12,24)\}$. We begin by introducing reflective boundary conditions (2.19b),

$$
\mathbf{J} \cdot \mathbf{n}=0 \quad \forall(x, y) \in \Gamma_{R}
$$

along $\Gamma_{R}=\{x=0,0<y<24\} \cup\{0<x<24, y=0\}$. Vacuum boundary conditions (2.18),

$$
\frac{1}{4} \phi-\frac{1}{2} \mathbf{J} \cdot \mathbf{n}=0 \quad \forall(x, y) \in \Gamma_{V}
$$

are specified along $\Gamma_{V}=\{x=24,0<y<24\} \cup\{0<x<24, y=24\}$. In the constant coefficient case with a constant source this imposes as natural flow of neutrons along streamlines $y-x=$ constant. Introducing a piecewise constant definition of the diffusion coefficient and the source

$$
D(x, y)=\left\{\begin{array}{ll}
1000 & \forall(x, y) \in \Omega_{1} \\
1 & \forall(x, y) \in \Omega_{2}
\end{array}, \quad Q(x, y)= \begin{cases}1 & \forall(x, y) \in \Omega_{1} \\
0 & \forall(x, y) \in \Omega_{2}\end{cases}\right.
$$

creates a significant perturbation of this idealized flow. Particularly, the relatively low diffusion coefficient in $\Omega_{2}$ serves to channel the neutrons through a single point, $(12,12)$ creating the large gradients observed in Figure 6.5. Thus, this problem introduces a mixed boundary condition and a significantly worse singularity than was present in the ground water flow problem Section 6.5.2. To further test the preconditioners and to accommodate internal layers near the interfaces of $\Omega_{1}$ and $\Omega_{2}$ we introduce an exponential


Figure 6.4: A diffusive checkerboard configuration based on Dendy's [23] example.
mesh. Specifically on the region $0 \leq x \leq 12$ we consider a grid spacing which is uniform in area such that

$$
\Delta x=\frac{2}{L} \int_{0}^{12} e^{\gamma_{x} x} d x, \quad x_{i+1}=\frac{1}{\gamma_{x}} \log \left(e^{\gamma_{x} x}+\gamma_{x} \Delta x\right)
$$

similar definitions are employed for $12 \leq x \leq 24$ and for the grid in $y$. A typical mesh generated with $\gamma_{x}=\gamma_{y}=2 / 10$ is displayed in Figure 6.6.

The iteration counts for this grid are given in Table 6.7 where column $r$ indicates the maximum aspect ratio present. We first note that the iteration counts for $C G\left(\mathcal{S}_{\mu}\right)$ and $C G\left(\mathcal{S}_{\phi}\right)$ are comparable and extremely large, ( 16,055 and 19,054 respectively on the $96 \times 96$ mesh) giving a clear indication of the conditioning of the problem. The influence of $\mathcal{S}_{\mu}^{D}$ is quite impressive given its simplicity, achieving iteration counts which are only $25 \%$ higher than for the ground water flow problem. As expected the iteration counts for $\widetilde{\mathcal{S}_{\mu ; u}}$ are extremely poor, remaining consistent with the bound given in Theorem 6.1. Conversely the results for both $\widetilde{\mathcal{S}_{\mu ; u, v}}$ and $\widetilde{\mathcal{S}_{\phi}}$ remain excellent. We also note that a comparison of the iteration counts for $\widetilde{\mathcal{S}_{\mu ; u, v}}$ in Table 6.7 with those for the constant coefficient Dirichlet problem in Table 6.3 with $r=8$ (or $r=1 / 8$ ) reveals a comparable


Figure 6.5: The solution, $\phi(x, y)$ of the diffusive checkerboard computed on a $24 \times 24$ uniform mesh


Figure 6.6: 2D Exponential Grid with the parameters $\gamma_{x}=\gamma_{y}=2 / 10$.
performance on this significantly more difficult problem. The equivalent comparison for $\widetilde{\mathcal{S}_{\phi}}$ leads to the same conclusion.

Table 6.7: Iteration counts for $\gamma_{x}=\gamma_{y}=2 / 10$.

| Mesh Size | $r$ | $C G\left(\mathcal{S}_{\mu}\right)$ | $\mathcal{S}_{\mu}^{D}$ | $\widetilde{\mathcal{S}_{\mu ; u}}$ | $\widetilde{\mathcal{S}_{\mu ; u, v}}$ | $\widetilde{\mathcal{S}_{\phi}}$ | $C G\left(\mathcal{S}_{\phi}\right)$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $24 \times 24$ | 7.71 | 2,129 | 181 | 59 | 9 | 13 | 2,063 |
| $48 \times 48$ | 9.04 | 6,190 | 374 | 74 | 10 | 13 | 6,932 |
| $96 \times 96$ | 9.91 | 16,055 | 759 | 86 | 11 | 14 | 19,054 |

## Chapter 7

## Conclusions

### 7.1 Research Summary

In this research we have addressed the fundamental issues that have restricted the practical use of nodal methods to the reactor physics community. Specifically, we have enhanced the mathematical foundation of the lowest-order nodal methods by establishing new equivalences with certain FEMs and by providing a generally applicable truncation error analysis. In addition, we have developed a new family of preconditioners for the iterative solution of the resulting system of equations, which yield optimally fast solvers. Finally, we have introduced a new efficient multi-level homogenization procedure.

Beginning in Section 3.6 we introduced a rudimentary truncation analysis that may be used to verify the consistency of nodal discretizations. The incompleteness of established equivalences between all nodal methods and corresponding FEM discretizations (e.g. utilization of higher order transverse leakage approximations [Section 3.4.1]) makes this approach necessary. We demonstrated the utility of this technique with the analysis of the NEM [Definition 3.4.2] and NIM [Definition 3.5.1] discretizations. Specifically, under the assumption of piecewise constant coefficients we verified the second-order consistency of these popular nodal methods. This analysis also served to highlight the exact treatment of both the cell balance relation, and the continuity (in the weak sense) of normal currents, which is instrumental in their performance. In addition, the truncation errors for the corresponding one-sided normal currents were used to characterize
situations in which NIM is anticipated to be more accurate than NEM. This was demonstrated computationally in Section 3.7. Unfortunately, in the general setting of piecewise smooth coefficients [i.e. $D_{i, j}(\mathbf{r}),\left(\Sigma_{a}\right)_{i, j}(\mathbf{r})$ ] this analysis revealed that the one-sided normal currents of both NIM and NEM are only first order consistent, with the leading term depending on the derivative of the diffusion coefficient. Moreover, we determined that the balance equation is no longer exact, exhibiting a second-order error that depends on the first derivatives of the absorption coefficient. These properties are an asset in applications which rely on homogenization procedures. Nevertheless, a discussion of the key issues surrounding the restoration of second-order consistency and exact cell balance was presented.

This rudimentary analysis considered the nodal discretization in isolation, while ignoring the influence of the homogenization procedure. However, the approximations of Equivalence Theory [Section 4.1.3], which are the most popular homogenization techniques employed in reactor physics, introduce discontinuities in the flux. Discontinuous FEMs [1] might offer a formal treatment of this scenario, but the jumps in the flux are determined through local auxiliary fine scale computations. This latter aspect of Equivalence Theory complicates the analysis and discourages its use in more general settings. Moreover, the impressive accuracy that is obtained on assembly size meshes in practical reactor computations reflects not only the aforementioned properties of the nodal discretization, but its integration with the approximations of Equivalence Theory. An unfortunate consequence of this interdependence is that the discretization itself has received scant consideration outside the reactor physics community. Efforts to adopt a more rigorous treatment of homogenization, which may improve this situation, are considered in [76]. Unfortunately, even with an asymptotic treatment of the nearly periodic case, practical computations still rely on the solution of local auxiliary problems.

In Section 4.3 we introduced a new multigrid homogenization procedure which circumvents many of these problems. In particular, the operator-induced coarsening of black box multigrid is used to construct coarse-scale operators from a specified fine-scale discretization. Since this coarsening procedure is a generalization of variational coarsening, a bilinear finite element discretization is used on the fine grid and assumed on the coarse grid. The diffusion tensor is then extracted from the coarse-grid operator according to Theorem 4.3. Thus we shift the assumption of a priori knowledge of the exact solution to a specification of the solution space. We believe that in this respect our technique marks a considerable step forward in practical homogenization. We demonstrated the excellent performance of our technique for a series of periodic test problems that included a computation involving nonzero off-diagonal terms. This ability is crucial for modelling layered structures that are not aligned with the coordinate axes.

Nevertheless, there are many applications for which homogenization procedures already exist, and frequently these do not address the possible need of off-diagonal entries in the homogenized diffusion tensor. In this vein a few researchers have focused on the need for a stronger mathematical foundation to assess the applicability of nodal discretizations for general diffusion problems. This research established the equivalence of various nodal methods with certain nonconforming and mixed-hybrid FEMs [36, 39]. Our investigation of these equivalences in the lowest-order case produced several new results. In particular, we found that the nodal integration method (NIM) was equivalent to a primal nonconforming method, which employed an analytic basis, if exact integration is used to evaluate the elements of the stiffness and mass matrices [Theorem 5.2]. Moreover, utilizing a quadratic polynomial basis in a primal nonconforming FEM is equivalent to the lowest-order nodal expansion method (NEM) if a special treatment of the mass
matrix is employed. ${ }^{1}$ We demonstrated that a simple mass lumping is sufficient, and corresponds to a restoration of a cell balance relation and the continuity of normal currents. Recalling that the primal nonconformal methods do not rigorously enforce cell balance, an apparent advantage of the analytic basis is that it intrinsically embodies this property. However, a rigorous treatment of cell balance does appear in mixed-hybrid FEMs. In fact, Hennart and del Valle [39] have established the equivalence of the lowest-order NEM with the Raviart-Thomas mixed-hybrid methods [Section 5.3.2].

The one-dimensional analysis of Section 5.2 was instrumental in identifying the subtle aspects of the nonconformal equivalences. It has also provided important information regarding the one-dimensional properties of these discretizations that do not generalize completely to two dimensions. Specifically, in the one-dimensional case elimination of the cell unknowns yields an edge-based system which, in certain cases, is an M-matrix. In particular, with the analytic basis functions (i.e. corresponding to NIM) we found that the edge-based system is unconditionally an M-matrix. Similarly, the polynomial basis unconditionally yields an M-matrix when the exact stiffness and mass matrices are used. However, once lumping is employed to establish a cell balance relation and continuity of the current (i.e. corresponding to NEM), we obtain an M-matrix subject to a constraint on the mesh size that depends on $\sqrt{\left(\Sigma_{a}\right)_{i, j} / D_{i, j}}$. Thus, the analytic basis yields a discretization that appears to capture the essential physical and mathematical properties of the differential problem while the polynomial basis does not. Moreover, the implication of the edge-based system corresponding to an M-matrix is that the global celland edge-based system satisfies a $\mathcal{A}^{-1} \geq 0$. We conjecture that although an M-matrix does not arise in two dimensions, the global cell- and edge-based system also satisfies $\mathcal{A}^{-1} \geq 0$ in a manner consistent with the one-dimensional case. If proven, this would

[^8]establish an important advantage for NIM: in difficult problems with severe variations in the coefficients no restriction of mesh spacing would be necessary.

Despite the many desirable properties that the families of nodal methods and mixedhybrid finite element methods embody, their very design makes the solution of the resulting equations awkward and costly, thwarting many potential users. Yet, inherent in this structure is a natural partitioning of the system which can be utilized in the development of preconditioners. The existence of such a partitioning suggests the investigation of various reduced systems (e.g. Schur complements) in conjunction with suitably sparse approximations. In particular, we presented three such approximations, $\widetilde{\mathcal{S}_{\mu ; u}}, \widetilde{\mathcal{S}_{\mu ; u, v}}$ and $\widetilde{\mathcal{S}_{\phi}}$, and demonstrated that the preconditioning which resulted was optimal in the sense that a fixed number of iterations, independent of the mesh size, was required to reduce the residual by a fixed amount. These solvers are competitive with multi-level methods because the preconditioner is only approximately inverted with a single V - or W -multigrid cycle.

Unfortunately, $\widetilde{\mathcal{S}_{\mu ; u}}$ which preconditions the most popular form of the discretization, is sensitive to high aspect ratio cells having $r>1$. This shortcoming, an artifact of arbitrarily approximating $A_{u u}$ as opposed to $A_{v v}$, was alleviated by the two step preconditioner $\widetilde{\mathcal{S}_{\mu ; u, v}}$. Conversely, an identical approximation is made in both coordinate directions during the development of $\widetilde{\mathcal{S}_{\phi}}$ and hence its preconditioning is insensitive to high aspect ratio cells. Moreover, the bound obtained in Theorem 6.2 suggests that in a more general setting the system $\mathcal{S}_{(\phi, \mu)}$ will be of significant practical interest. In all cases the preconditioners were found to be robust with respect to spatial variations in the diffusion coefficient.

### 7.2 Future Work

Our analysis of nodal discretizations considered their most common treatment in reactor physics, assuming a tensor product mesh with piecewise constant homogenized coefficients. In fact, this restricted setting represents a significant class of modelling problems. Nevertheless, these assumptions may be inadequate in more general settings. In Section 3.6 we discussed the issues surrounding the use of piecewise smooth coefficients and concluded that this is a feasible extension to pursue. In addition, the asymptotic homogenization techniques indicated that a piecewise smooth diffusion tensor is necessary to accurately model general problems on a coarse mesh. Unfortunately, neither NEM nor NIM facilitates a straightforward extension in this case, as both the tangential and normal components of the current are required at the cell interface. A similar situation arises with a logically rectangular mesh, particularly as the application of the transverse integration procedure becomes unclear. We feel that in each of these cases examination of the corresponding mixed-hybrid FEM offers an excellent starting point for the investigation.

The incomplete analysis found with the lowest-order nodal discretizations is compounded for higher-order methods. Many methods are referred to in the literature as higher-order because a particular aspect of the discretization utilizes a higher-order approximation, yet a correspondingly higher order truncation error is not necessarily attained [36]. Nevertheless, the construction of higher-order nodal methods that adhere to the principles outlined in this research, is a compelling topic for future consideration.

The results of the new multigrid homogenization technique presented in Section 4.3 were restricted to the periodic case. The extension of the local result given in Theorem 4.3 to the nonperiodic case is straightforward. Specifically, no change is required for Neumann boundary conditions, as these are the natural BCs of the primal variational
formulation, while Dirichlet BCs may be incorporated directly. We note that if the stencil weights vary spatially on the coarse mesh this local result is a second order approximation of the multigrid homogenized diffusion tensor. Although this may be unavoidable in the periodic case, for Neumann and Dirichlet BCs it seems possible to extract the exact multigrid homogenized diffusion tensor through a series of one-dimensional leastsquares problems. However, the exact multigrid homogenized diffusion tensor is itself an approximation that reflects the operator induced coarsening procedure. An analysis of this procedure is required to understand the errors that arise in truly multidimensional problems, particularly in pathological cases such as the checkerboard. In addition to a theoretical analysis of the errors, it will be very informative to apply the multigrid homogenization technique to realistic modelling problems.

The ability of the nodal integration method to build analytic information into the discretization suggests that it may be a very powerful technique for discretizing other differential equations. Research in this area, while still focused on aspects reactor simulation, include applications of NIM to the transport equation [52] and problems in fluid flow [ $28,43,7]$. In a preliminary investigation we have applied NIM to the convection-diffusion equation (see also [55, 28]) with encouraging results.

Specifically, the use of an analytic one-dimensional solution to the transverse integrated ODEs provides an intrinsic adaptive upwinding. Moreover, it may be viewed as a two-dimensional generalization of the well-known Scharfetter-Gummel discretization employed in one-dimensional semiconductor modelling [63]. Our results for the circulating flow of Smith and Hutton [67], a popular convection diffusion benchmark, are displayed in Figure 7.1. Note the impressive outflow profile, which is exact in the convective limit. However, there are also small oscillations observed near the layer [Figure 7.2]. Through further study of the convective limit we identified the primary source of this instability:
the violation of the assumption that the transverse normal currents are approximately constant. The prospect of employing assumptions which are compatible with both the hyperbolic and elliptic limits of the convection-diffusion equation represents an interesting approach for new discretizations.


Figure 7.1: Strearn lines for the convection dominated flow of Smith and Hutton [67].


Figure 7.2: The convection dominated flow of [67] with $P e=[D]^{-1}=500$ computed using a constant-constant NIM: (a) cell averages (d) inflow and outflow profiles.

The solvers that we considered in Chapter 6 were based on preconditioning Krylov subspace methods. However, it is possible to view these preconditioners as coarse-grid operators for which the specification of intergrid transfer operators and a fine-grid smoother would yield a multi-level algorithm. Our investigation of this scenario for the cell-based coarse-grid operator $\widetilde{\mathcal{S}_{\phi}}$ is very encouraging. Ultimately this approach may play a significant role in the iterative solution of higher-order methods, since in this setting the lowest-order nodal discretizations may be viewed as coarse-grid operators themselves.

Changes to the underlying assumptions of the discretization present new challenges for an iterative solver. It was noted in Chapter 6 that the extension to a piecewise constant diagonal tensor was straightforward and that both $\widetilde{\mathcal{S}_{\mu ; u, v}}$ and $\widetilde{\mathcal{S}_{\phi}}$ are expected to perform well. The treatment of piecewise smooth coefficients is unclear as it may necessitate the use of a higher-order discretization, in which case we may appeal to the multi-level approach described above. In addition, both the extension to a full tensor and logically rectangular meshes pose a formidable challenge. Specifically, if the ultimate form of a nodal discretization in these cases is assumed to be that of the corresponding mixed-hybrid FEM, it is apparent that $A$ remains block diagonal, but the blocks are dense. Nevertheless we believe an extension of our technique is possible in these cases.

## Bibliography

[1] M. L. Adams and W. R. Martin, Diffusion synthetic acceleration of discontinuous finite element transport iterations, Nucl. Sci. Eng., 111 (1992), pp. 145-167.
[2] M. B. Allen, R. E. Ewing, and P. Lu, Well-conditioned iterative schemes for mixed finite-element models of porous-media flows, SIAM J. Sci. Stat. Comput., 13 (1992), pp. 794-814.
[3] T. Arbogast and Z. Chen, On the implementation of mixed methods as nonconforming methods for second-order elliptic problems, Math. Comput., 64 (1995), pp. 943-972.
[4] Argonne National Laboratory, Benchmark Problem Book: ANL-7416, Supplement 2, 1977.
[5] D. N. Arnold and F. Brezzi, Mixed and nonconforming finite element methods: Implementation, postprocessing and error estimates, Math. Model. Numer. Anal., 19 (1985), pp. 7-32.
[6] Y. Y. Azmy and J. J. Dorning, A nodal integral approach to the numerical solution of partial differential equations, in Advances in Reactor Computations, vol. II, Salt Lake City, UT, March 28-30, 1983, American Nuclear Society, ANS, pp. 893909.
[7] ——, Nodal integral method solutions of cavity thermal convection problems, in $\mathrm{Nu}-$ merical Methods in Laminar and Turbulent Flow, IV, C. Taylor, M. D. Olson, P. M.

Gresho, and W. G. Habaski, eds., Swansea, Wales, UK, July 9-12 1985, Pineridge Press, UK, pp. 753-767.
[8] Y. Y. Azmy and B. L. Kirk, Performance modeling of parallel algorithms for solving neutron diffusion problems, Nucl. Sci. Eng., 120 (1995), pp. 1-17.
[9] G. I. Bell and S. Glasstone, Nuclear Reactor Theory, Van Nostrand Reinhold Company, New York, 1970.
[10] F. Bennewitz, H. Finnemann, and M. R. Wagner, Higher order corrections in nodal reactor calculations, Trans. Am. Nucl. Soc., 22 (1975), pp. 250-251.
[11] A. Bensoussan, J.-L. Lions, and G. Papanicolaou, Asymptotic Analysis For Periodic Structures, vol. 5 of Studies in Mathematics and its Applications, NorthHolland, New York, 1978.
[12] J. F. Bourgat, Numerical experiments of the homogenization method for operators with periodic coefficients, in Computing Methods in Applied Science and Engineering, I, R. Glowinski and J.-L. Lions, eds., Versailles, December 5-9 1977, Springer, pp. 330-356.
[13] A. Brandt, Multi-level adaptive solutions to boundary-value problems, Math. Comput., 31 (1977), pp. 333-390.
[14] F. Brezzi and M. Fortin, Mixed and Hybrid Finite Element Methods, no. 15 in Springer Series in Computational Mathematics, Springer-Verlag, New York, 1991.
[15] G. F. Carey and J. T. Oden, Finite Elements: A Second Course, vol. II of The Texas Finite Element Series, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1983.
[16] K. M. Case and P. F. Zweifel, Linear Transport Theory, Addison-Wesley, Reading, Massachusetts, 1967.
[17] B. Char, K. Geddes, G. Gonnet, B. Leong, M. Monagan, and S. Watt, Maple V: Language Reference Manual, Spinger-Verlag, New York, 1991.
[18] Z. Chen and D. Kwak, The analysis of multigrid algorithms for nonconforming and mixed methods for second order elliptic problems, IMA Preprint 1277, Institute for Mathematics and its Applications, Minneapolis, MN, 1994.
[19] C. Cordes and W. Kinzelbach, Continuous groundwater velocity fields and path lines in linear, bilinear and trilinear finite elements, Water Resour. Res., 28 (1992), pp. 2903-2911.
[20] B. Davison, Neutron Transport Theory, Oxford University Press, London, 1957.
[21] D. L. Delp, D. L. Fischer, J. M. Harriman, and M. J. Stidwell, FLare: A three-dimensional boiling reactor water simulator, Atomic Energy Commission Research and Development Report GEAP-4598, Atomic Power Equipment Department, General Electric, July 1964.
[22] S. Demko, W. F. Moss, and P. W. Smith, Decay rates for inverses of band matrices, Math. Comput., 43 (1984), pp. 491-499.
[23] J. E. Dendy, Black box multigrid, J. Comput. Phys., 48 (1982), pp. 366-386.
[24] J. E. Dendy, J. M. Hyman, and J. D. Moulton, The black box multigrid numerical homogenization algorithm, Tech. Rep. LAUR-96-3588, Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM, March 1994.
[25] I. Dilber and E. E. Lewis, Variational nodal methods for neutron transport, Nucl. Sci. Eng., 91 (1985), pp. 132-142.
[26] J. J. Duderstadt and L. J. Hamilton, Nuclear Reactor Analysis, John Wiley \& Sons, Inc., New York, 1976.
[27] J. J. Duderstadt and W. R. Martin, Transport Theory, John Wiley \& Sons, Inc., New York, 1979.
[28] P. D. Esser and R. J. Witt, An upwind nodal integral method for incompressible fluid flow, Nucl. Sci. Eng., 114 (1993), pp. 20-35.
[29] R. E. Ewing, Y. Kuznetsov, R. D. Lazarov, and S. Maliassov, Finite Element Modeling of Environmental Problems, John Wiley \& Sons, New York, 1995, ch. Substructure Preconditioning for Porous Flow Problems, pp. 303-329.
[30] C. Fedon-Magnaud, J. P. Hennart, and J. J. Lautard, On the relationship between some nodal schemes and the finite element method in static diffusion calculations, in Proceedings of A Topical Meeting on Advances in Reactor Computations, vol. 2, Salt Lake City, Utah, March 28-30, 1983, American Nuclear Society, Mathematics and Computation Division, pp. 987-1000.
[31] H. Finnemann, F. Bennewitz, and M. R. Wagner, Interface current techniques for multidimensional reactor calculations, Atomkernenergie, 30 (1977), pp. 123-128.
[32] H. D. Fischer and H. Finnemann, The nodal integration method - A diverse solver for neutron diffusion problems, Atomkernenergie-Kerntechnik, 39 (1981), pp. 229-236.
[33] G. Greenman, K. Smith, and A. F. Henry, Recent advances in an analytic nodal method for static and transient reactor analysis, in Computational Methods in Nuclear Engineering, vol. I, Williamsburg, Virginia, April 23-25, 1979, American Nuclear Society, ANS, pp. 3-49-3-72.
[34] N. K. Gupta, Nodal methods for three-dimensional simulators, Progress in Nuclear Energy, 7 (1981), pp. 127-149.
[35] J. P. Hennart, A general finite element framework for nodal methods, in The Mathematics of Finite Elements and Applications V, J. R. Whiteman, ed., London, 1985, Academic Press, pp. 309-316.
[36] __, A general family of nodal schemes, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 264-287.
[37] __, On the numerical analysis of analytical nodal methods, Numerical Methods for Partial Differential Equations, 4 (1988), pp. 233-254.
[38] J. P. Hennart and E. del Valle, Fast elliptic solvers using mixed-hybrid nodal finite elements, in Advances in Numerical PDE's and Optimization, S. Gómez, J. P. Hennart, and R. A. Tapia, eds., Philadelphia, 1991, SIAM, pp. 171-184.
[39] __, On the relationship between nodal schemes and mixed-hybrid finite elements, Numerical Methods for Partial Differential Equations, 9 (1993), pp. 411-430.
[40] A. F. Henry, Refinements in accuracy of coarse-mesh finite-difference solutions of the group-diffusion equations, in Numerical Reactor Calculations, no. IAEA-SM154/21 in Proceedings Series, Vienna, January 17-21, 1972, Internaltional Atomic Energy Agency, IAEA, pp. 447-471.
[41] A. F. Henry, Nuclear Reactor Analysis, The MIT Press, Cambridge, Massachusetts, 1975.
[42] M. H. Holmes, Introduction to Perturbation Methods, no. 20 in Texts in Applied Mathematics, Springer-Verlag, New York, 1995.
[43] W. C. Horak and J. J. Dorning, A nodal coarse-mesh method for the efficient numerical solution of laminar flow problems, J. Comput. Phys., 59 (1985), pp. 405440.
[44] J. J. Douglas, R. Durán, and P. Pietra, Numerical Approximation of Partial Differential Equations, North-Holland, Amsterdam, 1987, ch. Formulation of Alternating-Direction Iterative Methods for Mixed Methods in Three Space, pp. 2130.
[45] H. K. Joo and C. H. Kim, A new approach to core-reflector boundary conditions for nodal reactor computations, Nucl. Sci. Eng., 116 (1994), pp. 300-312.
[46] S. Knapek, Matrix-dependent multigrid-homogenization for diffusion problems, in The Preliminary Proceedings of the Copper Mountain Conference on Iterative Methods, T. Manteuffal and S. McCormick, eds., viol. I, SIAM Special Interest Group on Linear Algebra, Cray Research, April 9-13 1996.
[47] __, Matrix-dependent multigrid-homogenization for diffusion problems, Preprint: submitted to SIAM J. Sci. Stat. Comput., (1996).
[48] K. Koebke, A new approach to homogenization and group condensation, in Homogenization Methods in Reactor Physics, Lugano, Switzerland, November 13-15, 1978, International Atomic Energy Agency, IAEA, pp. 303-323.
[49] J. G. Kollas and A. F. Henry, The determination of homogenized group diffusion theory parameters, Nucl. Sci. Eng., 60 (1976), pp. 464-471.
[50] Y. A. Kuznetsov and S. Y. Maliassov, Substructuring preconditioners for nonconforming finite element approximations of second order elliptic problems with anisotropy, Russ. J. Numer. Anal. Model., 10 (1995), pp. 511-533.
[51] E. W. Larsen and G. C. Pomraning, The $P_{N}$ theory as an asymptotic limit of transport theory in planar geometry-I: Analysis, Nucl. Sci. Eng., 109 (1991), pp. 4975.
[52] R. D. Lawrence, Progress in nodal methods for the solution of the neutron diffusion and transport equations, Progress in Nuclear Energy, 17 (1986), pp. 271-301.
[53] _-, Three-dimensional nodal diffusion and transport methods for the analysis of fast-reactor critical experiments, Progress in Nuclear Energy, 18 (1986), pp. 101-111.
[54] R. D. Lawrence and J. J. Dorning, A nodal green's function method for multidimensional neutron diffusion calculations, Nucl. Sci. Eng., 76 (1980), pp. 218-231.
[55] E. P. E. Micheal, J. J. Dorning, E. M. Gelbard, and Rizwan-uddin, Nodal integration method for the convection-diffusion heat equation, Trans. Am. Nucl. Soc., 69 (1993), pp. 239-241.
[56] R. Mosé, P. Siegel, P. Ackerer, and G. Chavent, Application of the mixed hybrid finite element approximation in a groundwater flow model: Luxury or necessity, Water Resour. Res., 30 (1994), pp. 3001-3012.
[57] J. D. Moulton, Diffusion modelling of picosecond pulse propagation in turbid media, Master's thesis, McMaster University, Hamilton, Ontario, August 1990.
[58] J. M. Noh and N. Z. Cho, A new approach of analytic basis function expansion to neutron diffusion nodal calculation, Nucl. Sci. Eng., 116 (1994), pp. 165-180.
[59] C. R. Rao and S. K. Mitra, Generalized Inverse of Matrices and its Applications, John Wiley \& Sons, Inc., New York, 1971.
[60] P. A. Raviart and J. M. Thomas, A mixed finite element method for 2-nd order elliptic problems, in Mathematical Aspects of the Finite Element Method, I. Galligani and E. Magenes, eds., vol. 606 of Lecture Notes in Mathematics, Springer Verlag, 1977, pp. 292-315.
[61] T. Rusten and R. Winther, A preconditioned iterative method for saddlepoint problems, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 887-904.
[62] E. Sanchez-Palencia, Non-Homogeneous Media and Vibration Theory, no. 127 in Lecture Notes in Physics, Springer-Verlag, Berlin, 1980.
[63] D. L. Scharfetter and H. K. Gummel, Large signal analysis of a silicon read diode oscillator, IEEE Trans. Electron Devices, 16 (1969), pp. 64-77.
[64] D. S. Selengut, Diffusion coefficients for heterogenous systems, Nuclear Physics Research Quarterly Report HW-60220, Hanford Laboratories, General Electric, Richland, Washington, April 1959.
[65] K. S. Smith, Assembly homogenization techniques for light water reactor analysis, Progress in Nuclear Energy, 17 (1986), pp. 303-335.
[66] K. S. Smith, A. F. Henry, and R. A. Loretz, The determination of homogenized diffusion theory parameters for coarse mesh nodal analysis, in Advances in Reactor Physics and Shielding, Sun Valley, September 14-19, 1980, American Nuclear Society, ANS, pp. 294-308.
[67] R. M. Smith and A. G. Hutton, The numerical treatment of advection: A performance comparison of current methods, Numer. Heat Transf., 5 (1982), pp. 439461.
[68] J. W. Song and J. K. Kim, An efficient nodal method for transient calculations in light water reactors, Nuclear Technology, 103 (1993), pp. 157-167.
[69] G. Strang and G. J. Fix, An Analysis of the Finite Element Method, Series in Automatic Computation, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1973.
[70] R. S. Varga, Matrix Iterative Analysis, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1962.
[71] E. L. Wachspress, R. D. Burgess, and S. Baron, Multichannel flux synthesis, Nucl. Sci. Eng., 12 (1962), pp. 381-389.
[72] M. R. Wagner, H. Finnemann, K. Koebke, and H.-J. Winter, Validation of the nodal expansion method and the depletion program MEDIUM-2 by benchmark calculations and direct comparison with experiment, Atomkernenergie, 30 (1977), pp. 129-135.
[73] M. R. Wagner, K. Koebke, and H.-J. Winter, A nonlinear extension of the nodal expansion method, in Advances in Mathematical Methods for the Solution of Nuclear Engineering Problems, vol. 2, Munich, Germany, April 27-29 1981, American Nuclear Society, ANS, Mathematics and Computation Division, pp. 43-58.
[74] A. Weiser and M. F. Wheeler, On convergence of block-centered finite differences for elliptic problems, SIAM J. Numer. Anal., 25 (1988), pp. 351-375.
[75] D. M. Young and K. C. Jea, Generalized conjugate-gradient acceleration of nonsymmetrizable iterative methods, Linear Algebr. Appl., 34 (1980), pp. 159-194.
[76] H. Zhang, Rizwan-uddin, and J. J. Dorning, Systematic homogenization and self-consistent flux and pin power reconstruction for nodal diffusion methods - I: Diffusion equation-based theory, Nucl. Sci. Eng., 121 (1995), pp. 226-244.


[^0]:    ${ }^{1}$ To avoid confusion with the summation symbol, interaction coefficients will always appear with a subscript that denotes their classification. In definitions and statements which apply over all classes, the subscript $\alpha$ will be used.

[^1]:    ${ }^{2}$ In hopes of distinguishing it from any similar or related vector interpretation, and not merely to create an oxymoron, the flux is sometimes referred to as the scalar flux.

[^2]:    ${ }^{3}$ The vacuum boundary is also referred to as a free boundary and a purely absorbing boundary.

[^3]:    ${ }^{1}$ In this context accuracy is relative to in core measurements and does not necessarily correspond to the associated discretization error.

[^4]:    ${ }^{2}$ The diffusion length $L_{i, j}$ in the one-group diffusion model is defined by $L_{i, j}^{2}=\left(\Sigma_{a}\right)_{i, j} / D_{i, j}$, and is a measure of how far neutrons will diffuse before being absorbed. The migration area $M_{i, j}^{2}$ is an improved approximation of the true diffusion length (squared), that incorporates multigroup information.

[^5]:    ${ }^{1}$ The energy dependence as well as its potential homogenization, often referred to as condensation or collapse, play an important role in reactor modelling. This dependence is readily included in the analysis and appears in most of the methods discussed here.

[^6]:    ${ }^{2}$ In his review Smith [65] refers to these as assembly homogenized cross sections (AXSs)

[^7]:    ${ }^{1}$ Timings were performed on an HP Apollo $9000 / 735$ ( 99 MHz PA-RISC 7100)

[^8]:    ${ }^{1}$ Note that in the conservative case the analytic and polynomial bases (i.e. NIM and NEM) are equivalent and hence quite fortuitously the polynomial basis inherits its advantageous properties.

