



Degree Project in Computational Mathematics

Second cycle, 30 credits

# **Approximate Neutral Density in Physical Oceanography**

Formulation and Analysis of an Anisotropic Boundary Value  
Problem

**NORA R. ODELIUS**



# **Approximate Neutral Density in Physical Oceanography**

## **Formulation and Analysis of an Anisotropic Boundary Value Problem**

NORA R. ODELIUS

Master's Programme, Applied and Computational Mathematics, 120 credits  
Date: October 12, 2025

Supervisors: Fabien Roquet, Olof Runborg

Examiner: Olof Runborg

School of Engineering Sciences

Host organization: Gothenburg University

Swedish title: Approximation av den neutrala densitetsvariabeln inom fysisk oceanografi

Swedish subtitle: Utformning och analys av ett anisotropiskt randvärdesproblem



## Abstract

Neutral surfaces, on which the neutral density variable is constant, are widely used for understanding oceanic flow patterns and analysing hydrographic observations. The locally referenced normal vectors of a neutral surface are the so called diapycnal vectors. A neutral density variable for a given domain can be found through the diapycnal vectors and then used to construct neutral surfaces. Finding a realistic formulation of the neutral density variable and computing it for general hydrographic data present significant challenges in two key areas. Firstly, due to the non-zero helicity of the diapycnal vector field it is not possible to find neutral surfaces whose normals exactly align with the field. Secondly, evaluating neutral density is computationally heavy due to the size of global oceanic data. Given the current limitations, investigating alternative formulations of neutral density and developing computational methods for its evaluation remains an important area of ongoing research. This thesis is concerned with the construction of neutral density via a minimization problem that penalizes deviations from desirable properties of this variable. A finite difference scheme for two dimensional domains is suggested as an approach to solving the Euler-Lagrange equation corresponding to this optimization problem. The method is then investigated for two dimensional vector fields under varying rotational properties. The method's performance is investigated for conservative and non-conservative test fields as well as a diapycnal vector field. In this analysis, convergence is shown in the  $L_2$ -norm with respect to step-size for all fields and root mean square measures are analysed to evaluate consistency with theoretical results through numerical experiments. The results of this study indicate that the method works well when the vector field is aligned with the grid directions. As this is the case for the diapycnal vector field, the method shows promise for application in physical oceanography.

## Keywords

Neutral density, Neutral surfaces, Optimization, Variational calculus, Euler-Lagrange equation, Anisotropic diffusion equation, Finite difference method, Hydrographic data, Ocean flow patterns

## Sammanfattning

Neutrala ytor, där den neutrala densitetsvariabeln är konstant, används inom fysisk oceanografi för att förstå flödesmönster och analysera observationer i havet. Dessa ytor kan definieras av sina normalvektorer, som från ett lokalt perspektiv kallas de diapyknala vektorerna. Neutrala ytor kan konstrueras utifrån en neutral densitetsvariabel, men denna process är inte okomplicerad. För det första, på grund av att heliciteten av det diapyknala vektorfältet inte är noll, kommer man inte kunna hitta neutrala ytor vars normalvektorer exakt sammanfaller med vektorfältet. För det andra, på grund av havets storlek och därmed även storleken av hydrografisk data uppstår en stor beräkningskostnad för att uppskatta variabeln. Denna studie formulerar och undersöker en ny metod för att beräkna den neutrala densitetsvariabeln. Ett minimeringsproblem konstrueras som viktat variabeln baserat på önskvärda egenskaper. En finita differensmetod används därefter för att lösa den Euler-Lagrange ekvation som är associerad med optimeringsproblemet. Metoden testas sedan på vektorfält med olika egenskaper. I denna studie undersöks metodens prestanda för konservativa och icke-konservativa vektorfält samt ett verkligt diapyknalt vektorfält. Studien visar på konvergens i  $L_2$ -normen med avseende på steglängd för alla fält, och analyserar hur väl metoden följer teoretiska resultat genom numeriska experiment. Resultaten från denna studie indikerar att metoden fungerar väl när beräkningsnätets koordinatriktningar och riktningen hos vektorfältet överensstämmer. Det diapyknala vektorfältet har denna egenskap, vilket gör metoden särskilt lovande för tillämpningar inom fysisk oceanografi.

## Nyckelord

Neutral densitet, Neutrala ytor, Optimering, Variationskalkyl, Euler–Lagrange-ekvation, Anisotropisk diffusionsekvation, Finita differensmetoden, Hydrografisk data, Havsströmningsmönster

## Acknowledgments

I would like to thank my supervisor Fabien Roquet from the Department of Marine Sciences at the University of Gothenburg for his insights into the scientific process and for taking the time to meet and discuss my questions throughout the project. I am grateful for his warm welcome and enthusiasm in encouraging me to deepen my understanding of physical oceanography through seminars and other excursions. I am also grateful to Olof Runborg at KTH Royal Institute of Technology for sharing his extensive insight and intuition on numerical analysis and for showing patience and willingness to help whenever questions arose. Finally, it was a pleasure to discuss this topic with Trevor McDougall and Geoff Stanley, whose input, assistance with setting up the variational problem, and willingness to meet across difficult time zones were sincerely appreciated.

Stockholm, October 2025

Nora R. Odelius

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Ocean properties and neutral density</b>	<b>5</b>
2.1	The equation of state for seawater . . . . .	5
2.2	Oceanographic properties and stability . . . . .	6
2.2.1	Potential density . . . . .	6
2.2.2	Thermal expansion coefficient . . . . .	6
2.2.3	Haline contraction coefficient . . . . .	6
2.2.4	Adiabatic compressibility . . . . .	6
2.2.5	The buoyancy frequency . . . . .	7
2.3	Neutral density . . . . .	7
2.3.1	Neutral surfaces and the diapycnal vector . . . . .	7
2.3.2	The question of helicity . . . . .	9
2.3.3	Neutral density in 2D . . . . .	10
2.3.4	The neutral density variable . . . . .	12
<b>3</b>	<b>Mathematical model</b>	<b>13</b>
3.1	A new model for neutral density . . . . .	13
3.1.1	Neutral density as an optimization problem . . . . .	13
3.1.2	Variational calculus . . . . .	14
3.1.3	Reformulation of the optimization problem . . . . .	17
3.1.4	Expressions for the diffusion tensor in 2D and 3D . . . . .	20
3.1.5	Neutral density as a boundary value problem . . . . .	21
3.2	Wellposedness . . . . .	23
3.2.1	Variational (weak) formulation . . . . .	25
3.2.2	Wellposedness by Lax-Milgram . . . . .	26
3.3	Effect of $\mu$ on the boundary value problem . . . . .	30

<b>4</b>	<b>Finite differences</b>	<b>34</b>
4.1	Finite difference scheme . . . . .	35
4.2	Boundary conditions . . . . .	38
4.2.1	Natural Neumann boundary conditions . . . . .	38
4.2.2	Simple Neumann boundary condition . . . . .	39
4.3	Solving the linear system of equations . . . . .	40
<b>5</b>	<b>Results and analysis</b>	<b>42</b>
5.1	Software implementation and data . . . . .	42
5.2	Test cases and performance metrics . . . . .	44
5.3	Conservative fields . . . . .	45
5.3.1	Test cases . . . . .	46
5.3.2	Convergence . . . . .	49
5.3.3	Solution's dependence on $\mu$ . . . . .	53
5.4	Non-conservative fields . . . . .	57
5.4.1	Test cases . . . . .	57
5.4.2	Convergence . . . . .	58
5.4.3	Solution's dependence on $\mu$ . . . . .	60
5.5	Real diapycnal vector field . . . . .	63
5.5.1	Convergence . . . . .	63
5.5.2	Solution's dependence on $\mu$ . . . . .	68
<b>6</b>	<b>Conclusion</b>	<b>69</b>
6.1	Summary . . . . .	69
6.2	Limitations and future work . . . . .	71
	<b>References</b>	<b>73</b>
<b>A</b>	<b>Finite difference discretization</b>	<b>76</b>

# List of Figures

4.1	Illustration of the finite difference scheme utilized in this study (adapted from [20]). . . . .	37
5.1	Absolute salinity and conservative temperature from a simulated hydrographic dataset at 29.5 degrees east in zonal direction. . . . .	43
5.2	Thermal expansion and haline contraction coefficients from a simulated hydrographic dataset at 29.5 degrees east in zonal direction. . . . .	44
5.3	Vector field (5.4) and principal diffusion directions of the diffusion tensor $D$ at $\mu = 0.5$ . . . . .	46
5.4	Vector field (5.4) and principal diffusion directions of the diffusion tensor $D$ at $\mu = 1$ . . . . .	47
5.5	Vector field (5.5) and principal diffusion directions of the diffusion tensor $D$ at $\mu = 0.5$ . . . . .	47
5.6	Vector field (5.6) and principal diffusion directions of the diffusion tensor $D$ at $\mu = 0.5$ . . . . .	48
5.7	Vector field (5.7) and principal diffusion directions of the diffusion tensor $D$ at $\mu = 0.5$ . . . . .	48
5.8	Empirical convergence of conservative fields at $\mu = 0.5$ . . . . .	50
5.9	Vector field $\vec{A}$ and solution gradient $\nabla\gamma$ using natural boundary conditions with refined step size $h_1 = h_0 2^{-1}$ at $\mu = 0.5$ . . . . .	51
5.10	Vector field $\vec{A}$ and solution gradient $\nabla\gamma$ using natural boundary conditions with refined step size $h_3 = h_0 2^{-3}$ at $\mu = 0.5$ . . . . .	52
5.11	Angular and magnitude deviations between conservative vector field $\vec{A}$ and solution gradient $\nabla\gamma$ as a function of $\mu$ . . . . .	55
5.12	Angular and magnitude deviations between conservative vector field $\vec{A}$ and solution gradient $\nabla\gamma$ as a function of $\mu$ . . . . .	56

5.13	Vector field (5.8) and principal diffusion directions of the diffusion tensor $D$ at $\mu = 0.5, 0.1$ and $0.01$ for the non-conservative field with $\tau = 0.4$ . . . . .	58
5.14	Empirical convergence of the non-conservative field (5.8) with $\tau = 0.4$ at $\mu = 0.5, 0.1$ and $0.01$ . . . . .	59
5.15	Angular and magnitude deviations between the non-conservative vector field (5.8) with $\tau = 0.4$ and solution gradient $\nabla\gamma$ as a function of $\mu$ . . . . .	60
5.16	Vector field (5.8) with $\tau = 0.4$ and solution gradient $\nabla\gamma$ using natural boundary conditions with step size $h_0$ at $\mu = 0.5$ . . . . .	62
5.17	Vector field (5.8) with $\tau = 0.4$ and solution gradient $\nabla\gamma$ using natural boundary conditions with step size $h_0$ at $\mu = 0.1$ . . . . .	62
5.18	Vector field (5.8) with $\tau = 0.4$ and solution gradient $\nabla\gamma$ using natural boundary conditions with step size $h_0$ at $\mu = 0.01$ . . . . .	63
5.19	The real diapycnal vector field (2.6) and principal diffusion directions of the diffusion tensor $D$ with refined step sizes $\Delta y^{(1)} = \Delta y^{(0)}2^{-1}$ and $\Delta z^{(1)} = \Delta z^{(0)}2^{-1}$ . . . . .	64
5.20	Empirical convergence of the real diapycnal vector field (2.6) at $\mu = 0.5, 0.1$ and $0.01$ . . . . .	65
5.21	Vector field $\vec{A}$ and solution gradient $\nabla\gamma$ using natural boundary conditions with refined step sizes $\Delta y^{(1)} = \Delta y^{(0)}2^{-1}$ and $\Delta z^{(1)} = \Delta z^{(0)}2^{-1}$ . . . . .	66
5.22	Vector field $\vec{A}$ and solution gradient $\nabla\gamma$ using natural boundary conditions with refined step sizes $\Delta y^{(3)} = \Delta y^{(0)}2^{-3}$ and $\Delta z^{(3)} = \Delta z^{(0)}2^{-3}$ . . . . .	67
5.23	Angular and magnitude deviations between the real diapycnal vector field (2.6) and solution gradient $\nabla\gamma$ as a function of $\mu$ . . . . .	68

# List of Tables

5.1	Empirical convergence rates for conservative fields (5.4), (5.5), (5.6) and (5.7) at $\mu = 0.5$ where $h_i = h_0 2^{-i}$ . . . . .	49
5.2	Empirical convergence rates for non-conservative field (5.8) with $\tau = 0.4$ at $\mu = 0.5, 0.1$ and $0.01$ where $h_i = h_0 2^{-i}$ . . . . .	59
5.3	Empirical convergence rates for the real diapycnal vector field (2.6) at $\mu = 0.5, 0.1$ and $0.01$ where $\Delta y^{(i)} = \Delta y^{(0)} 2^{-i}$ and $\Delta z^{(i)} = \Delta z^{(0)} 2^{-i}$ . . . . .	64

# Chapter 1

## Introduction

Physical oceanography extends our understanding of the natural processes on Earth and serves as an integral part of research on climate change. While there are many topics worth discussing, a fundamental problem in physical oceanography is finding accurate ways of tracing water masses across our oceans to understand both spreading and mixing properties. Dating back to the 1920s, different kinds of *isopycnal* surfaces, that is surfaces on which some density variable is constant, have been frequently used to describe flow patterns [1]. In these models water parcels are assumed to flow along surfaces described by the chosen density variable. This thesis is concerned with approximating the *neutral density variable* which is used to define *neutral surfaces* on which water parcels are assumed to flow.

Density variables preceding neutral density have been proven insufficient for various reasons. While *in situ* density, which is the measured density of water, would seem like a natural choice, it proved to be unsuitable for tracing water masses as it is strongly affected by water compressibility. Surfaces of constant potential density, where a reference pressure is included, were therefore suggested as a response to this limitation. However, fixing a constant reference level induces approximations in the representation of compressibility of seawater causing inaccuracies that grow the further away from the chosen reference pressure we go. There are models using multiple reference pressures that reduces the effects caused by this simplification but, significant inaccuracies are still observed when working in the intermediate layers of the ocean or below [2]. Because of these limitations, neutral surfaces, on which the approximate neutral density variable is constant, is recommended to use. One reason being that it does not depend on a specific reference pressure [1].

Neutral surfaces are surfaces on which a water parcel moves short distances isentropically without experiencing buoyancy restoring forces [3]. This is a logical assumption of oceanic flow patterns because, if friction is disregarded, no force will be required to move the water parcel along such a surface. Formally, the neutral surface can be defined locally by its normal, the *diapycnal vector*  $\vec{A}$  which can be computed from temperature and salinity fields using the equation of state [4]. A scalar function  $\gamma$  is a neutral density variable if its gradient is everywhere collinear with the diapycnal vector, that is if there exists a scalar function  $b(\vec{x})$  such as  $\nabla\gamma(\vec{x}) = b(\vec{x})\vec{A}(\vec{x})$ . Such a variable  $\gamma(\vec{x})$  and factor  $b(\vec{x}) \neq 0$  only exist if the helicity  $H$  of  $\vec{A}(\vec{x})$  is zero, see [5], [6] and Section 2.3.2 of this thesis. The helicity of a vector field  $\vec{A}$  is defined through

$$H = \vec{A} \cdot (\nabla \times \vec{A}). \quad (1.1)$$

The main issue when looking to find neutral surfaces in the ocean is that the helicity of  $\vec{A}$  is non-zero and the neutral density variable  $\gamma$  will hence not be a well defined variable for the world's oceans [6]. Moreover, the factor  $b(\vec{x})$  is not known *a priori*. The general problem, and the goal of this study, is to define an approximate neutral density variable  $\gamma(\vec{x})$  that forms surfaces as neutral as possible, essentially trying to find  $b(\vec{x})$  and  $\gamma(\vec{x})$  such that  $\nabla\gamma(\vec{x}) \approx b(\vec{x})\vec{A}(\vec{x})$  over the entire domain. This study uses techniques in variational calculus to accomplish this without explicitly knowing  $b(x)$ . Before delving into more details on the methods deployed in this study, let us summarize the history of neutral surfaces.

Trevor J. McDougall was the first to give a systematic account of neutral surfaces for the worlds oceans, see [3] as well as [6]. In a paper by McDougall and Jackett [7], an algorithm is devised to arrive at a neutral density variable for general hydrographic data such that its level surfaces form neutral surfaces. Since then, there have been several attempts to both extend and redefine the neutral density variable to improve its definition in areas where it falls short. For example, Lang et al. (2020) [8] introduces a neutral density variable independent of vertical heave to reduce artificial mixing while still being practically comparable in its neutrality to the variable presented in [7]. Eden and Willebrand [9] noted that the neutral density variable in [7] was mostly suitable for analysing and diagnosing hydrographic observations and constructed a functional form of neutral density tuned for the North Atlantic or similar environments to be useful in layer models. In [10], Stanley et al. made significant improvements to reduce computational effort in the evaluation

of the neutral density variable. Further, Klocker et al. [11] have reduced the influence of artificial diapycnal diffusivity significantly which according to [10] results in the currently most accurate approximately neutral surface. Klocker et al. [11] do however emphasize that what constitutes as the *best* neutral density variable is dependent on its application.

In this thesis, the neutral density variable  $\gamma(\vec{x})$  is modelled as the solution to an optimization problem which penalizes deviations from desirable properties of the neutral density variable, in particular the collinearity of  $\vec{A}$  and  $\nabla\gamma(\vec{x})$ . Using variational calculus, the corresponding Euler-Lagrange equation is derived together with its natural Neumann boundary conditions. The neutral density variable can then be computed by solving an elliptic boundary value problem. A second order finite difference scheme is explored as possible solution method for two dimensional rectangular domains. Our method builds on the work of [9], who solved the simpler Helmholtz problem. In contrast to most previous attempts, the method suggested in this study would make it possible to compute the neutral density variable directly in 3D, possibly improving computational efforts. This is important because limited computational resources is one of the bigger bottlenecks in oceanographic modelling due to the scale of the ocean and consequently the size of hydrographic datasets.

While the possibility of reducing computational efforts by finding  $\gamma(\vec{x})$  everywhere simultaneously motivates the choice of method, this study is predominantly concerned with formulating a model for the neutral density variable so that neutral surfaces are physically accurate. Moreover, this study explores how the suggested method reacts to different properties of the input data, specifically characteristics of the diapycnal vector  $\vec{A}(\vec{x})$ .

The thesis is organized as follows. Chapter 2 includes a detailed explanation of neutral surfaces as well as information on other oceanographic properties necessary to understand what constitutes a physically accurate neutral density variable. This chapter also includes details on the importance of helicity and what changes in this context when turning to the simpler problem of 2D domains. Chapter 3 includes details on the model of neutral density beginning with formulation of the optimization problem, then moving on to derivation of the corresponding boundary value problem through variations of the objective function. This chapter also gives an account of some characteristics of the model including wellposedness and dependencies of the solution. Chapter 4 is concerned with application of finite differences to the boundary value problem stated in Chapter 3. Chapter 5 presents the test environment that is used to evaluate the behaviour of the scheme. It also

## 4 | Introduction

includes numerical examples, results and their implications. Finally, Chapter 6 includes a summary of the results as well as limitations of the study and suggestions on the focus of future work.

## Chapter 2

# Ocean properties and neutral density

This chapter gives a formal account of neutral surfaces and describes desirable properties of the neutral density variable. To this end, preliminary oceanography is presented including some parts of the equation of state for seawater. This chapter also includes discussions of the existence of a neutral density variable  $\gamma$  in relation to the helicity of  $\vec{A}$  in 3D and the generalization of this argument to 2D.

### 2.1 The equation of state for seawater

In this study, oceanographic quantities are needed to compute the diapycnal vector  $\vec{A}$ . Researchers in oceanography take great care in defining and refining such quantities so that they are practical and accurate in application. The equation of state for seawater is a relation that gives *in situ* density as a function of temperature, salinity and pressure. Many definitions and standards are available for temperature and salinity, making it a somewhat confusing topic. The current standard for these quantities can be found in the International Thermodynamic Equation of State 2010, TEOS-10, see [12].

The equation of state distributed by TEOS-10 is defined in [4] where *in situ* density  $\rho$  ( $\text{kgm}^{-3}$ ) is defined as a function of absolute salinity  $S_A$ , conservative temperature  $\Theta$  and pressure  $p$  [12] such that

$$\rho = \rho(S_A, \Theta, p). \quad (2.1)$$

## 2.2 Oceanographic properties and stability

In this section, oceanographic quantities given by density, absolute salinity, conservative temperature and pressure are defined. These are relevant in the computation of the diapycnal vector  $\vec{A}$ . Moreover, the notion of stability for water masses is introduced as it is important in the discussion of the helicity of the diapycnal vector  $\vec{A}$ .

### 2.2.1 Potential density

Potential density is obtained by setting pressure to a reference pressure value,  $\rho_\theta = \rho(S_A, \Theta, p_{ref})$ . Typical potential density variables are  $\rho^0$ , which uses the surface pressure  $p_{ref} = 0$  dbar, and  $\rho^2$ , which uses  $p_{ref} = 2000$  dbar.

### 2.2.2 Thermal expansion coefficient

The thermal expansion coefficient  $\alpha$  ( $\text{K}^{-1}$ ) is derived through entropy by differentiation with respect to in-situ temperature  $t$  at fixed absolute salinity  $S_A$  and sea pressure  $p$ , see appendix A.15 of [12]. It is defined as the partial derivative of the density (2.1) with respect to conservative temperature  $\Theta$  such that

$$\alpha = -\frac{1}{\rho} \frac{\partial \rho}{\partial \Theta} \Big|_{S_A, p}. \quad (2.2)$$

### 2.2.3 Haline contraction coefficient

The haline contraction coefficient  $\beta$  ( $\text{kg/g}$ ) is derived using differentiation with respect to absolute salinity  $S_A$  at fixed  $t$  and  $p$ , see appendix A.15 of [12]. It is described in terms of the partial derivative of density (2.1) with respect to absolute salinity  $S_A$  such that

$$\beta = \frac{1}{\rho} \frac{\partial \rho}{\partial S_A} \Big|_{\Theta, p}. \quad (2.3)$$

### 2.2.4 Adiabatic compressibility

The adiabatic compressibility  $\kappa$  ( $\text{Pa}^{-1}$ ) is equivalently described in terms of the partial derivative of the density (2.1) with respect to sea pressure  $p$  such

that

$$\kappa = \left. \frac{1}{\rho} \frac{\partial \rho}{\partial p} \right|_{S_A, \Theta}. \quad (2.4)$$

## 2.2.5 The buoyancy frequency

The squared buoyancy frequency  $N^2$  ( $rad^2 s^{-2}$ ) is a measure of the stability of a fluid to vertical displacements and is given in terms of the vertical gradients of density and pressure, or in terms of vertical gradients of conservative temperature and absolute salinity, see [12]. We have,

$$g^{-1}N^2 = \alpha \frac{\partial \Theta}{\partial z} - \beta \frac{\partial S_A}{\partial z} = -\rho^{-1} \frac{\partial \rho}{\partial z} + \kappa \frac{\partial p}{\partial z} \quad (2.5)$$

where  $g \approx 9.81 m/s^2$  is the gravitational acceleration.

## 2.3 Neutral density

This section contains details on the definition of neutral surfaces and introduces the issue of non-zero helicity. The section then concludes by describing expected behaviour of the neutral density variable based on known properties of a neutral surface.

### 2.3.1 Neutral surfaces and the diapycnal vector

Neutral surfaces are defined as the direction along which a water parcel undergoing isentropic movement, that is with no exchange of heat or mass, does not experience buoyancy restoring forces [3]. The normal vectors of these surfaces will be the diapycnal vector  $\vec{A}$ . It is defined through density, absolute salinity and conservative temperature as well the haline contraction and thermal expansion coefficients such that

$$\vec{A} = \rho(\beta \nabla S_A - \alpha \nabla \Theta). \quad (2.6)$$

Note that the diapycnal vector depends implicitly on location, see Section 2.1 or the TEOS-10 manual [12].

To understand why this is the local normal of the neutral surface it is beneficial to consider the local stability of a water parcel. This will provide

an explanation of what directions the water parcel can move isentropically without being subjected to buoyancy forces that work to restore stability. To this end consider density changes and how they relate to the buoyancy frequency.

Density is explicitly dependent on absolute salinity, conservative temperature and pressure as seen equation (2.1) but only implicitly dependent on location  $\vec{x} = [x, y, z]^T$ . Consider the variation of the density  $\rho(S_A, \Theta, p)$  given as follows:

$$d\rho = \frac{\partial\rho}{\partial\Theta}d\Theta + \frac{\partial\rho}{\partial S_A}dS_A + \frac{\partial\rho}{\partial p}dp. \quad (2.7)$$

Using the definition for the thermal expansion coefficient  $\alpha$ , the haline contraction coefficient  $\beta$  and adiabatic compressibility  $\kappa$  we arrive at the following expression for the variation:

$$d\rho = \rho(\beta dS_A - \alpha d\Theta) + \rho\kappa dp. \quad (2.8)$$

Consider two points  $a$  and  $b$  which lie a small distance  $\delta x$  apart on the neutral surface. Between these points a water parcel is moved isentropically. When a water parcel is denser than its surrounding water it sinks and when it is less dense than its surrounding water it rises, in both cases creating a current to restore stable stratification. When  $N^2 > 0$ , the fluid is stably stratified and there exists a plane where a water parcel can be displaced small distances isentropically without experiencing buoyancy restoring forces. If the density of a water parcel is the same as its surroundings in the new position then there is no buoyancy restoring force acting on the parcel [13]. This will be the case if the density change is only due to pressure effects [3]. The neutral surface is hence defined such that, if  $a$  and  $b$  lies on the surface,

$$\lim_{\delta x \rightarrow 0} \frac{\rho(b) - \rho(a)}{\delta x} = \lim_{\delta x \rightarrow 0} \rho\kappa \frac{p(b) - p(a)}{\delta x}. \quad (2.9)$$

Consider the surface gradient operator  $\nabla_n u = \nabla u - \vec{n}(\vec{n} \cdot \nabla u)$  where  $\vec{n}$  is the unit normal of the surface. Essentially, the operator  $\nabla_n$  finds the projection of the gradient onto the surface defined by the unit normal vector  $\vec{n}$ . Suppose  $\vec{n}$  is the locally referenced unit normal vector to the neutral surface. Then an equivalent definition to (2.9) can describe the density change under isentropic

movement on the neutral surface, namely

$$\nabla_n \rho = \rho \kappa \nabla_n p. \quad (2.10)$$

Using equation (2.8), this is equivalently given as

$$\nabla_n \rho - \rho \kappa \nabla_n p = \rho (\beta \nabla_n S_A - \alpha \nabla_n \Theta) = 0. \quad (2.11)$$

From this expression it is clear why the diapycnal vector  $\vec{A}$  defines the neutral surface. Consider the definition of the surface gradient operator  $\nabla_n$ . This is applied to the second expression of equation (2.11) such that

$$\begin{aligned} & \rho [\beta (\nabla S_A - \vec{n}(\vec{n} \cdot \nabla S_A)) - \alpha (\nabla \Theta - \vec{n}(\vec{n} \cdot \nabla \Theta))] = 0 \\ \Leftrightarrow & \rho (\beta \nabla S_A - \alpha \nabla \Theta) - \vec{n}((\vec{n} \cdot \rho \beta \nabla S_A) - (\vec{n} \cdot \rho \alpha \nabla \Theta)) = 0 \\ \Leftrightarrow & \rho (\beta \nabla S_A - \alpha \nabla \Theta) - \vec{n}(\vec{n} \cdot \rho (\beta \nabla S_A - \alpha \nabla \Theta)) = 0 \\ \Leftrightarrow & \vec{A} - \vec{n}(\vec{n} \cdot \vec{A}) = 0 \end{aligned}$$

In other words, the projection of  $\vec{A}$  onto the neutral surface is 0 and  $\vec{A}$  is therefore its locally referenced normal vector.

### 2.3.2 The question of helicity

Neutral surfaces are found by constructing iso-surfaces of a neutral density variable  $\gamma$ , that is surfaces on which this variable is constant. The variable is constant on the surface if its gradient along the surface is zero which will be true if  $\nabla \gamma$  and  $\vec{A}$  are collinear. Let us consider two important conclusions about the existence of such a  $\gamma$  in three dimensional space.

1. If  $\vec{A}$  is a conservative field, that is if  $\nabla \times \vec{A} = 0$ , then there exists a potential  $\gamma$  such that  $\nabla \gamma = \vec{A}$  [14].
2. We are only interested in ensuring collinearity of  $\nabla \gamma$  and  $\vec{A}$ . Therefore, finding  $\gamma$  and  $b(\vec{x}) \neq 0$  such that  $\nabla \gamma = b \vec{A}$  everywhere is sufficient. Such a potential  $\gamma$  and scalar function  $b(\vec{x})$  exists if the helicity of  $\vec{A}$ , as defined in (1.1), is zero. See [5] for discussions of existence within the context of vector analysis or [7] which includes discussions of helicity in physical oceanography specifically.

The main problem we face is that the helicity of the diapycnal vector field  $\vec{A}$  is not zero in the world's oceans and existence is not ensured. Another problem is that even if the helicity were zero, the factor  $b(\vec{x})$  is not known making it

harder to find  $\gamma$ .

**Remark** In this chapter and in subsequent chapters, the diapycnal vector  $\vec{A} = \vec{A}(\vec{x})$  as well as the neutral density variable  $\gamma = \gamma(\vec{x})$  are denoted without reference to their positional dependence. This positional dependence should be kept in mind however, especially when considering derivations. Sometimes the proportionality factor  $b = b(\vec{x})$  is also referenced without explicitly stating its positional dependence.

To understand the issue of helicity beyond its definition, it is compelling to consider neutral surfaces in terms of their underlying building blocks, so called neutral trajectories. A displacement  $dr$  is on the neutral trajectory if

$$\vec{A}dr = 0. \quad (2.12)$$

A path is a neutral trajectory if the above condition is satisfied everywhere along the path. The set of small displacements  $dr$  from a single point satisfying the above condition defines the infinitesimally small neutral surface element. There is exactly one neutral surface element in each position and its normal vector is everywhere orthogonal to all possible directions  $dr$  [6]. A neutral surface can then be viewed as the envelope of locally defined neutral surface elements.

### 2.3.3 Neutral density in 2D

Our oceans are of course three dimensional. Sometimes however, it can be advantageous to model or investigate a slice of the ocean. In those cases, it is important to define behaviour in 2D where the problem is essentially fixed in one coordinate.

The idea of a surface does not exist in 2D, instead we are modelling neutral trajectories and computing iso-lines of neutral density. More importantly, helicity, as defined in equation (1.1), does not exist because the scalar product between the curl of a vector and the vector itself is not defined in two dimensions. However, it is possible to embed  $\vec{A}$  into three dimensions and then evaluate helicity to understand the characteristic of the field. Consider the diapycnal vector in three dimensions,  $\vec{A} = [A_x, A_y, A_z]^T$  and the curl

$\nabla \times \vec{A}$  defined as

$$\nabla \times \vec{A} = \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ \partial_x & \partial_y & \partial_z \\ A_x & A_y & A_z \end{vmatrix} = \begin{bmatrix} \frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z} \\ \frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x} \\ \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y} \end{bmatrix}. \quad (2.13)$$

Suppose  $A_x = 0$  and  $A_y$  as well as  $A_z$  are independent of  $x$ , that is  $\vec{A} = [0, A_y(y, z), A_z(y, z)]^T$ . Then

$$\nabla \times \vec{A} = \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ \partial_x & \partial_y & \partial_z \\ 0 & A_y & A_z \end{vmatrix} = \begin{bmatrix} \frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z} \\ 0 \\ 0 \end{bmatrix}. \quad (2.14)$$

Using this result the helicity of  $\vec{A}$  when embedded into 3D is

$$H = \vec{A} \cdot (\nabla \times \vec{A}) = \begin{bmatrix} 0 \\ A_y(y, z) \\ A_z(y, z) \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z} \\ 0 \\ 0 \end{bmatrix} = 0. \quad (2.15)$$

To conclude, if the diapycnal vector field  $\vec{A} = [A_y(y, z), A_z(y, z)]^T$  is embedded into three dimensions under the conditions above, then the helicity of  $\vec{A}$  is always 0. Consequently, a non-trivial potential  $\gamma$  and factor  $b \neq 0$  such that  $\nabla\gamma = b\vec{A}$  always exists in 2D. Because the factor  $b$  is unknown it is compelling to wonder if it can be ignored, essentially assuming  $b = 1$  globally. Unfortunately, such a potential only exists if  $\vec{A}$  is conservative, that is without rotation. In 3D, the rotation of a vector field is defined through the cross product. To understand rotation in 2D consider the cross product when the two dimensional vector field  $\vec{A}$  is embedded into 3D by the same procedure as above. Then there exists a potential  $\gamma$  such that  $\nabla\gamma = \vec{A}$  only if

$$\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z} = 0. \quad (2.16)$$

If this is true then  $\vec{A}$  is conservative otherwise the factor  $b$  must be considered even in two dimensions.

**Remark** Position in our oceans can be characterized by zonal, meridional and vertical directions. Zonal refers to directions along latitudinal lines, that is

from east to west while meridional refers to directions along longitudinal lines, that is from north to south. In this study, when considering neutral density in 2D, we fix the zonal direction, usually denoted by  $x$  and keep the meridional and vertical directions free. Because of this, our study will consider a 2D domain such that  $(y, z) \in \Omega$  where  $y$  denotes the meridional direction and  $z$  the vertical direction.

### 2.3.4 The neutral density variable

In this study we consider desirable conditions for a realistic neutral density variable, similar to those taken into account in C. Eden and J. Willebrand's paper from 1999 [9]. Suppose  $\vec{x} = [x, y, z]^T$  and let  $b(\vec{x})$  be a proportionality factor. Let also  $\nabla_h = \frac{\partial}{\partial x}\hat{i} + \frac{\partial}{\partial y}\hat{j}$  denote the horizontal gradient where  $\hat{i}$  and  $\hat{j}$  are unit vectors in the  $x$  and  $y$  directions. Then,  $\gamma$  should be constructed such that

1. The neutral density variable has iso-surfaces/-lines that coincide with the neutral surfaces/trajectories. To that end, its gradient should be everywhere collinear with the diapycnal vector  $\vec{A}$ , that is  $\nabla\gamma \times \vec{A} = 0$ .
2. The horizontal gradient of  $\gamma$  should be locally proportional to the horizontal components of the diapycnal vector  $\vec{A}_h = [A_x, A_y]^T$  with proportionality factor  $b(\vec{x})$ , that is  $\nabla_h\gamma = b(\vec{x})\vec{A}_h$ .
3. The vertical derivative of  $\gamma$  should be locally proportional to the vertical component of the diapycnal vector with the same proportionality factor  $b(\vec{x})$  as above, that is  $\frac{\partial\gamma}{\partial z} = b(\vec{x})\vec{A}_z$ .

The goal of this study is to find a neutral density variable  $\gamma$  which satisfies the above conditions as much as possible.

**Remark** In the 2D setting it is only the second condition that changes. In that case the horizontal gradient is simply defined by  $\nabla_h = \frac{\partial}{\partial y}$  and  $\vec{A}_h = [A_y]$ .

**Remark** If two of the above conditions are satisfied the third follows automatically.

# Chapter 3

## Mathematical model

In this chapter, an optimization problem is formulated to model appropriate behaviour of the neutral density variable  $\gamma$ . Moreover, the corresponding boundary value problem, consisting of a Euler-Lagrange equation and the natural Neumann boundary conditions, is derived. A proof of wellposedness is then presented. Finally, some characteristics of the boundary value problem is discussed including properties of the diffusion tensor  $D$  and the solution's dependence on a regularization parameter  $\mu$ .

### 3.1 A new model for neutral density

This study suggests a method for the construction of neutral surfaces where an optimization problem is formulated and rewritten as a boundary value problem using variational calculus.

#### 3.1.1 Neutral density as an optimization problem

The neutral density variable should aim to satisfy the conditions specified in Section 2.3.4. To this end, an optimization problem is constructed. Firstly, the neutral density variable should minimize angular deviations between the gradient of  $\gamma$  and the diapycnal vector  $\vec{A}$ . We construct the following minimization term:

$$\min_{\gamma} \int_{\Omega} |\nabla\gamma \times \vec{A}|^2 d\Omega.$$

The second and third condition in Section 2.3.4 aim to control the magnitude of the solution gradient. In doing so, they also determine the direction of the

solution gradient, constraining it to be parallel or antiparallel to  $\vec{A}$  depending on the proportionality factor  $b(\vec{x})$ . While the proportionality factor  $b(\vec{x})$  is unknown, it is generally close to 1. This study idealizes real ocean properties by setting  $b(\vec{x}) = 1$  over the entire domain.

The following minimization term enforces the second and third condition in Section 2.3.4 with the idealization  $b(\vec{x}) = 1$ :

$$\min_{\gamma} \int_{\Omega} (\vec{A} \cdot \nabla \gamma - |\vec{A}|^2)^2 d\Omega.$$

**Remark** If a different distribution of  $b(\vec{x})$  is desired it should be included as a factor to  $\vec{A}$  in this minimization term.

To regulate how important the two terms are in relation to each other, a parameter  $\mu \in (0, 1]$  is introduced in front of the second term. The parameter  $\mu > 0$  is chosen to be less than one because it is more important or equally important to penalize deviations from collinearity compared to deviations from magnitude equality. See Section 3.3 for a discussion of the influence of this parameter on the solution. The two terms are collected and we arrive at the optimization problem which we denote by  $(P)$ ,

$$(P) \quad \min_{\gamma} J[\gamma] = \frac{1}{2} \int_{\Omega} |\nabla \gamma \times \vec{A}|^2 + \mu (\vec{A} \cdot \nabla \gamma - |\vec{A}|^2)^2 d\Omega. \quad (3.1)$$

When finding the corresponding boundary value problem to this optimization problem through variational calculus it is necessary to use either Dirichlet boundary conditions or the natural Neumann boundary conditions coming out of the derivation of the Euler-Lagrange equation.

**Remark** As we will see in Section 3.2, the second term in  $(P)$  is necessary to ensure wellposedness. The number  $\mu$  can therefore be seen as a regularization parameter.

### 3.1.2 Variational calculus

Mathematical concepts have historically been brought forth gradually from many areas of science and when especially useful findings have been collected and given a name. Variation calculus emerged within the natural sciences by the philosophical argument that nature operates by means that are easiest and fastest [15]. An example in optics is Fermat's principle dating back to

1662 which informally states that a ray travelling between two points takes the path that can be travelled in the least time. In its early history, the argument was criticized for assuming intent of natural processes, suggesting that nature actively tries to find the most efficient path forward. The critique is valid but falls short in its claim that the argument inherently assumes intent. Rather, nature might do nothing, have no goal or purpose, but still be driven to do what exerts the least amount of energy.

Before giving an account of what now is called Fermat's principle, Fermat had laid some theoretical groundwork in *Methodus ad disquirendam maximam et minimam* [16] where he outlined fundamental principles of differential calculus, characterizing algebraic expressions near extrema and in doing so found strategies for the determination of such extrema. Variational calculus is an extension of similar arguments to functionals, that is functions of functions, and it includes extensive theoretical results that can be applied to many fields in the natural sciences.

In many areas, both natural and artificial, it has been shown to be advantageous to find quantities that obeys some minimization, maximization or saddle-point law. This study is not an exception and utilizes variational calculus to find the neutral density variable  $\gamma$  which minimizes the optimization problem (P) as defined in equation (3.1). Variational calculus is used to derive the so called Euler-Lagrange Equation and natural Neumann boundary condition corresponding to (3.1). To understand how this is done consider the general case.

An integral functional over an open and bounded domain  $\Omega \subset \mathbb{R}^n$  and  $u \in \mathbf{X}(\Omega; \mathbb{R})$ , where  $\mathbf{X}$  is some Banach space, is defined as

$$\mathcal{J}[u] = \int_{\Omega} h(\vec{x}, u, \nabla u) d\vec{x}. \quad (3.2)$$

Variational calculus, at its core, is results and tools of analysis related to this integral functional under additional conditions such as boundary conditions for  $u$ .

To find the Euler-Lagrange Equation, consider the formal definition of the first

variation  $\delta \mathcal{J} : \mathbf{X}(\Omega; \mathbb{R}^m) \rightarrow \mathbb{R}$  of  $\mathcal{J}[u]$  [17]

$$\delta \mathcal{J}[u; \psi] = \lim_{\epsilon \rightarrow 0} \frac{\mathcal{J}[u + \epsilon \psi] - \mathcal{J}[u]}{\epsilon}, \quad \psi \in \mathbf{X}(\Omega; \mathbb{R}^m). \quad (3.3)$$

where  $u \in \mathbf{X}(\Omega; \mathbb{R})$  and  $\epsilon \in \mathbb{R}$ . Assuming this limit exists, and noting that

$$\mathcal{J}[u + \epsilon \psi] = \int_{\Omega} h(\vec{x}, u + \epsilon \psi, \nabla(u + \epsilon \psi)) d\vec{x}, \quad (3.4)$$

the first variation of  $\mathcal{J}[u]$  is given by the chain rule. We have,

$$\begin{aligned} \delta \mathcal{J}[u] &= \frac{d}{d\epsilon} \mathcal{J}[u + \epsilon \psi] \Big|_{\epsilon=0} \\ &= \int_{\Omega} \frac{d}{d\epsilon} h(\vec{x}, u + \epsilon \psi, \nabla(u + \epsilon \psi)) d\vec{x} \Big|_{\epsilon=0} \\ &= \int_{\Omega} \frac{\partial h}{\partial u} \frac{d(u + \epsilon \psi)}{d\epsilon} \Big|_{\epsilon=0} + \frac{\partial h}{\partial \nabla u} \cdot \frac{d\nabla(u + \epsilon \psi)}{d\epsilon} \Big|_{\epsilon=0} d\vec{x} \\ &= \int_{\Omega} \frac{\partial h}{\partial u} \psi + \frac{\partial h}{\partial \nabla u} \cdot \nabla \psi d\vec{x}. \end{aligned}$$

The second term of the first variation is rewritten using integration by parts in multiple dimensions such that

$$\int_{\Omega} \frac{\partial h}{\partial \nabla u} \cdot \nabla \psi d\vec{x} = - \int_{\Omega} \nabla \cdot \left( \frac{\partial h}{\partial \nabla u} \right) \psi d\vec{x} + \int_{\partial \Omega} \left( \frac{\partial h}{\partial \nabla u} \cdot \vec{n} \right) \psi dS \quad (3.5)$$

where  $\vec{n}$  is the outward normal of the boundary. This yields the following expression for the first variation

$$\delta \mathcal{J}[u] = \int_{\Omega} \left( \frac{\partial h}{\partial u} - \nabla \cdot \left( \frac{\partial h}{\partial \nabla u} \right) \right) \psi d\vec{x} + \int_{\partial \Omega} \left( \frac{\partial h}{\partial \nabla u} \cdot \vec{n} \right) \psi dS. \quad (3.6)$$

For all  $\psi$ , every minimizer  $u^*$  of  $\mathcal{J}[u]$  is such that the first variation is zero;  $\delta \mathcal{J}[u^*] = 0$ . We can therefore find the minimizer by solving the following Euler–Lagrange equation with corresponding natural Neumann boundary condition:

$$\begin{aligned} \frac{\partial h}{\partial u} - \nabla \cdot \left( \frac{\partial h}{\partial \nabla u} \right) &= 0, & \vec{x} \in \Omega, \\ \frac{\partial h}{\partial \nabla u} \cdot \vec{n} &= 0, & \vec{x} \in \partial \Omega. \end{aligned}$$

### 3.1.3 Reformulation of the optimization problem

The procedure in Section 3.1.2 is now applied to our optimization problem (P). In this study we build two boundary value problems; one for two dimensional domains and one for three dimensional domains.

**Remark** While this study provide theoretical results that are valid for both two and three dimensions, the numerical method and numerical experiments are only constructed in the two dimensional setting (with the additional assumption of a rectangular domain).

Comparing (P) with the general optimization problem in equation (3.2) we see that  $h(\vec{x}, \gamma, \nabla\gamma)$  is defined through

$$h(\vec{x}, \gamma, \nabla\gamma) = \frac{1}{2} \left( |\nabla\gamma \times \vec{A}|^2 + \mu(\vec{A} \cdot \nabla\gamma - |\vec{A}|^2)^2 \right).$$

Following the results in Section 3.1.2 a boundary value problem is derived by finding the explicit expressions of  $\frac{\partial h}{\partial \nabla u}$  and  $\frac{\partial h}{\partial u}$  where  $u = \gamma$ . First we note that the derivative with respect to  $\gamma$  will be

$$\frac{\partial}{\partial \gamma} h(\vec{x}, \gamma, \nabla\gamma) = 0$$

for both two and three dimensional domains.

Conversely, the derivative with respect to  $\nabla\gamma$  will differ slightly between two and three dimensional domains. However, the structure will be similar and the term can be simplified before considering a specific dimension. We have,

$$\begin{aligned} \frac{\partial}{\partial \nabla\gamma} h(\vec{x}, \gamma, \nabla\gamma) &= \frac{\partial}{\partial \nabla\gamma} \left( \frac{1}{2} (|\nabla\gamma \times \vec{A}|^2 + \mu(\vec{A} \cdot \nabla\gamma - |\vec{A}|^2)^2) \right) \\ &= \frac{\partial}{\partial \nabla\gamma} \left( \underbrace{\frac{1}{2} (\nabla\gamma \times \vec{A}) \cdot (\nabla\gamma \times \vec{A})}_I \right) \\ &\quad + \frac{\partial}{\partial \nabla\gamma} \left( \underbrace{\frac{1}{2} \mu(\vec{A} \cdot \nabla\gamma - |\vec{A}|^2)^2}_II \right). \end{aligned}$$

The second term of this expression (II) can be simplified in the same way regardless of the dimension of the domain. We have,

$$\begin{aligned}\text{II} &= \mu(\vec{A} \cdot \nabla\gamma - |\vec{A}|^2) \left( D_{\nabla\gamma}(\vec{A} \cdot \nabla\gamma) - \mu D_{\nabla\gamma}(\vec{A} \cdot \vec{A}) \right) \\ &= \mu(\vec{A} \cdot \nabla\gamma - |\vec{A}|^2) (\vec{A} - \vec{0}) = \mu(\vec{A} \cdot \nabla\gamma - |\vec{A}|^2) (\vec{A} - \vec{0}) \\ &= \mu\vec{A}(\vec{A}^T \nabla\gamma - |\vec{A}|^2) = \mu\vec{A}\vec{A}^T \nabla\gamma - \mu\vec{A}|\vec{A}|^2.\end{aligned}$$

Now consider the first term (I) of the expression for  $\frac{\partial}{\partial \nabla\gamma} h(\vec{x}, \gamma, \nabla\gamma)$ . Let the cross product in the two dimensional setting be a 2D-curl so that the cross product of the first term (I) becomes

$$\nabla\gamma \times \vec{A} = \frac{\partial\gamma}{\partial y} A_z - \frac{\partial\gamma}{\partial z} A_y \quad (3.7)$$

in two dimensions where

$$\vec{A} = \begin{bmatrix} A_y \\ A_z \end{bmatrix}, \quad \nabla\gamma = \begin{bmatrix} \frac{\partial\gamma}{\partial y} \\ \frac{\partial\gamma}{\partial z} \end{bmatrix}.$$

Then, in both two and three dimensions, the cross product can be rewritten such that  $\nabla\gamma \times \vec{A} = M_A \nabla\gamma$  for some matrix  $M_A$  that depends on  $\vec{A}$  (see Section 3.1.4). Using this expression, the first term (I) can be simplified further for both two and three dimensional domains. Consider the part of (I) of which we take the derivative with respect to  $\nabla\gamma$ . For two and three dimensional domains we have,

$$\begin{aligned}\frac{1}{2}(\nabla\gamma \times \vec{A}) \cdot (\nabla\gamma \times \vec{A}) &= \frac{1}{2}(M_A \nabla\gamma)^T M_A \nabla\gamma \\ &= \frac{1}{2}\nabla\gamma^T M_A^T M_A \nabla\gamma.\end{aligned}$$

The derivative with respect to  $\nabla\gamma$  of this expression yields the first term (I). Because  $M_A^T M_A$  is symmetric we have,

$$\begin{aligned} \text{I} &= \frac{\partial}{\partial \nabla\gamma} \left( \frac{1}{2} (\nabla\gamma \times \vec{A}) \cdot (\nabla\gamma \times \vec{A}) \right) \\ &= \frac{\partial}{\partial \nabla\gamma} \left( \frac{1}{2} \nabla\gamma^T M_A^T M_A \nabla\gamma \right) \\ &= \frac{1}{2} (M_A^T M_A + (M_A^T M_A)^T) \nabla\gamma = M_A^T M_A \nabla\gamma =: X \nabla\gamma. \end{aligned}$$

What differs between the two and three dimensional models is  $M_A$  and therefore the matrix  $X$ . Before finding explicit expressions for  $X$  in two and three dimensions respectively let us construct the boundary value problem based on  $X$  and  $\vec{A}$ .

Collecting (I) and (II) yields an expression for the derivative of the integrand with respect to  $\nabla\gamma$  such that

$$\begin{aligned} \frac{\partial}{\partial \nabla\gamma} h(\vec{x}, \gamma, \nabla\gamma) &= \text{I} + \text{II} = X \nabla\gamma + \mu \vec{A} \vec{A}^T \nabla\gamma - \mu \vec{A} (\vec{A} \cdot \vec{A}) \\ &= (X + \mu \vec{A} \vec{A}^T) \nabla\gamma - \mu \vec{A} (\vec{A} \cdot \vec{A}) =: D \nabla\gamma - \mu \vec{A} |\vec{A}|^2. \end{aligned}$$

From this expression and based on the procedure described in Section 3.1.2 we arrive at the following Euler-Lagrange equation and the natural Neumann boundary condition:

$$\begin{aligned} \frac{\partial h}{\partial \gamma} - \nabla \cdot \left( \frac{\partial h}{\partial \nabla\gamma} \right) &= 0 \Leftrightarrow \nabla \cdot (D \nabla\gamma) = \nabla \cdot (\mu \vec{A} |\vec{A}|^2), & \vec{x} \in \Omega, \\ \frac{\partial h}{\partial \nabla\gamma} \cdot \vec{n} &= 0 \Leftrightarrow D \nabla\gamma \cdot \vec{n} = \mu \vec{A} |\vec{A}|^2 \cdot \vec{n}, & \vec{x} \in \partial\Omega \end{aligned}$$

where the diffusion tensor  $D$  is defined as

$$D = X + \mu \vec{A} \vec{A}^T. \quad (3.8)$$

The diffusion tensor  $D$  will differ between two and three dimensional domains because of the matrix  $X$ .

### 3.1.4 Expressions for the diffusion tensor in 2D and 3D

Let us now find expression for  $X$  and hence the diffusion tensor  $D$  in terms of  $\vec{A}$  for two and three dimensional domains respectively.

**2D** In two dimensions the diapycnal vector is defined by one horizontal component and one vertical component such that  $\vec{A} = [A_y, A_z]^T$ . By the 2D-curl defined above  $M_A$  is  $M_A = [A_z, -A_y]$ . This yields

$$\begin{aligned} X_{2D} &= M_A^T M_A = [A_z, -A_y]^T [A_z, -A_y] \\ &= \begin{bmatrix} A_z^2 & -A_z A_y \\ -A_y A_z & A_y^2 \end{bmatrix} = I|\vec{A}|^2 - \vec{A}\vec{A}^T \end{aligned}$$

where  $I$  is the identity matrix.

**3D** In three dimensions we have  $\vec{A} = [A_x, A_y, A_z]^T$ . Consider the evaluation of the first term (I) in this three dimensional setting. Let  $M_A$  be the matrix representing a cross-product with  $\vec{A}$ , so that

$$\nabla\gamma \times \vec{A} = -\vec{A} \times \nabla\gamma = - \begin{bmatrix} A_y \frac{\partial\gamma}{\partial z} - A_z \frac{\partial\gamma}{\partial y} \\ A_z \frac{\partial\gamma}{\partial x} - A_x \frac{\partial\gamma}{\partial z} \\ A_x \frac{\partial\gamma}{\partial y} - A_y \frac{\partial\gamma}{\partial x} \end{bmatrix} = M_A \nabla\gamma. \quad (3.9)$$

As seen from the above expression  $M_A$  is

$$M_A = - \begin{bmatrix} & -A_z & A_y \\ A_z & & -A_x \\ -A_y & A_x & \end{bmatrix} \quad (3.10)$$

in the three dimensional setting. The explicit expression for  $X_{3D}$  is

$$\begin{aligned} X_{3D} &= M_A^T M_A = \begin{bmatrix} & A_z & -A_y \\ -A_z & & A_x \\ A_y & -A_x & \end{bmatrix} \begin{bmatrix} & -A_z & A_y \\ A_z & & -A_x \\ -A_y & A_x & \end{bmatrix} \\ &= \begin{bmatrix} A_z^2 + A_y^2 & -A_y A_x & -A_z A_x \\ -A_x A_y & A_z^2 + A_x^2 & -A_z A_y \\ -A_x A_z & -A_y A_z & A_y^2 + A_x^2 \end{bmatrix} = I|\vec{A}|^2 - \vec{A}\vec{A}^T \end{aligned}$$

where  $I$  is the identity matrix.

We conclude that in both two and three dimensions  $X$  is given by

$$X = I|\vec{A}|^2 - \vec{A}\vec{A}^T. \quad (3.11)$$

Using this expression the diffusion tensor  $D$ , as defined in (3.8), is given by

$$D = I|\vec{A}|^2 - \vec{A}\vec{A}^T + \mu\vec{A}\vec{A}^T \quad (3.12)$$

for both two and three dimensional domains.

### 3.1.5 Neutral density as a boundary value problem

Collecting the above expressions yields the Euler-Lagrange equation and corresponding natural Neumann boundary condition in terms of  $\vec{A}$  and diffusion tensor  $D$ . The reformulation of the optimization problem ( $P$ ) results in a boundary value problem which we denote ( $BVP$ ), given as follows

$$\begin{aligned} (BVP) \quad \nabla \cdot (D\nabla\gamma) &= \nabla \cdot (\mu\vec{A}|\vec{A}|^2), \\ D\nabla\gamma \cdot \vec{n} &= \mu\vec{A}|\vec{A}|^2 \cdot \vec{n} \end{aligned} \quad (3.13)$$

where the explicit expressions of the diffusion tensor as defined in equation (3.12) is

$$\begin{aligned} D &= \begin{bmatrix} A_z^2 + \mu A_y^2 & (\mu - 1)A_y A_z \\ (\mu - 1)A_y A_z & A_y^2 + \mu A_z^2 \end{bmatrix}, & \text{in 2D,} \\ D &= \begin{bmatrix} A_z^2 + A_y^2 + \mu A_x^2 & (\mu - 1)A_y A_x & (\mu - 1)A_z A_x \\ (\mu - 1)A_x A_y & A_z^2 + A_x^2 + \mu A_y^2 & (\mu - 1)A_z A_y \\ (\mu - 1)A_x A_z & (\mu - 1)A_y A_z & A_y^2 + A_x^2 + \mu A_z^2 \end{bmatrix}, & \text{in 3D.} \end{aligned}$$

**Remark** The eigenvalues of the diffusion tensor  $D$  is  $|\vec{A}|^2$  and  $\mu|\vec{A}|^2$  for both two and three dimensions. We now provide a motivation for this statement. Consider  $D$  as defined in equation (3.12). By assuming the first eigenvector of  $D$  is  $\vec{A}$  we can find the corresponding eigenvalue  $\lambda_1$  through

$$\begin{aligned} D\vec{A} = \lambda_1\vec{A} &\Leftrightarrow I|\vec{A}|^2\vec{A} - \vec{A}\vec{A}^T\vec{A} + \mu\vec{A}\vec{A}^T\vec{A} = \lambda_1\vec{A} \\ &\Leftrightarrow \mu\vec{A}|\vec{A}|^2 = \lambda_1\vec{A} \Rightarrow \lambda_1 = \mu|\vec{A}|^2. \end{aligned}$$

It is confirmed that  $\vec{A}$  is an eigenvector of  $D$  with corresponding eigenvalue  $\mu|\vec{A}|^2$  for both two and three dimensional domains. We now want to find the

remaining eigenvalues for two and three dimensions respectively.

**2D** In the two dimensional setting let  $\vec{A}^\perp \in \mathbb{R}^2$  be perpendicular to  $\vec{A}$ . Since  $\vec{A}^T \vec{A}^\perp = 0$  we have,

$$\begin{aligned} D\vec{A}^\perp = \lambda_2 \vec{A}^\perp &\Leftrightarrow I|\vec{A}|^2 \vec{A}^\perp - \vec{A}\vec{A}^T \vec{A}^\perp + \mu \vec{A}\vec{A}^T \vec{A}^\perp = \lambda_2 \vec{A}^\perp \\ &\Leftrightarrow |\vec{A}|^2 \vec{A}^\perp = \lambda_2 \vec{A}^\perp \Rightarrow \lambda_2 = |\vec{A}|^2. \end{aligned}$$

We have found that  $\vec{A}^\perp$  is indeed an eigenvector of  $D$  with corresponding eigenvalue  $\lambda_2 = |\vec{A}|^2$ .

**3D** In the three dimensional setting the plane orthogonal to  $\vec{A}$  is two-dimensional, so there exist two linearly independent vectors spanning it. Suppose  $v_2, v_3 \in \mathbb{R}^3$  are linearly independent column vectors both perpendicular to  $\vec{A}$ . Then,

$$Dv_j = I|\vec{A}|^2 v_j - \vec{A}\vec{A}^T v_j + \mu \vec{A}\vec{A}^T v_j = |\vec{A}|^2 v_j \quad (3.14)$$

for  $j = 2, 3$ . This shows that  $|\vec{A}|^2$  is an eigenvalue for both  $v_2$  and  $v_3$ .

To conclude, we have found that in both two and three dimensions the eigenvalues of  $D$  are  $|\vec{A}|^2$  and  $\mu|\vec{A}|^2$ . In three dimensions, the eigenvalue  $|\vec{A}|^2$  has an algebraic multiplicity of 2.

**Remark** With these eigenvalues of the diffusion tensor  $D$  we have that  $D$  is positive definite as long as  $|\vec{A}| > 0$  and  $\mu > 0$ . Moreover, the condition number of  $D$  is  $\kappa = \frac{\max \lambda}{\min \lambda} = \frac{1}{\mu}$  when  $\mu \in (0, 1]$ . Hence, the condition number  $\kappa \rightarrow \infty$  as  $\mu \rightarrow 0$ .

While it is not trivial to find the exact relationship between the condition number of  $D$  and the conditioning of the sparse matrix coming out of the finite differencing it is clear that the two are connected. That means that as  $\mu \rightarrow 0$  and the condition number of  $D$  worsens, the system is harder to solve numerically and it will affect the accuracy of the solution. As a result, even though the solution might be independent of  $\mu$  analytically the choice of  $\mu$  will affect the solution from a numerical perspective.

**Remark** If the outer unit normal vector  $\vec{n}$  is either parallel or perpendicular to  $\vec{A}$  at the boundary then the natural Neumann boundary condition can be simplified to  $\vec{n} \cdot \nabla \gamma = \vec{n} \cdot \vec{A}$ . To understand this, consider the boundary

condition of (BVP) and a reformulation of  $D\vec{A}$ . Since the diffusion tensor  $D$  has an eigenvector  $\vec{A}$  with corresponding eigenvalue  $\mu|\vec{A}|^2$  we have that

$$D\vec{A} = \mu|\vec{A}|^2\vec{A}.$$

for both two and three dimensional domains. Using this equality as well as the fact that  $D$  is symmetric the boundary condition can be rewritten as

$$\begin{aligned} D\nabla\gamma \cdot \vec{n} &= D\vec{A} \cdot \vec{n} \\ \Leftrightarrow D\vec{n} \cdot \nabla\gamma &= D\vec{n} \cdot \vec{A}. \end{aligned}$$

Vectors parallel or perpendicular to  $\vec{A}$ , as we assumed  $\vec{n}$  to be, are eigenvectors of the diffusion tensor  $D$  by the arguments of the first remark of this section. The boundary condition can therefore be simplified using  $D\vec{n} = \lambda\vec{n}$ , with  $\lambda \neq 0$ , to arrive at

$$\lambda\vec{n} \cdot \nabla\gamma = \lambda\vec{n} \cdot \vec{A} \Leftrightarrow \vec{n} \cdot \nabla\gamma = \vec{n} \cdot \vec{A}. \quad (3.15)$$

In conclusion, we have shown that if  $\vec{A}$  is either parallel or perpendicular to  $\vec{n}$  at the boundary then the natural Neumann boundary condition can be simplified to  $\vec{n} \cdot \nabla\gamma = \vec{n} \cdot \vec{A}$ .

When looking at numerical examples we will use this simplified boundary condition to analyse the method. When the test field is perpendicular and/or parallel to the normal vector at the boundary, the two boundary conditions should produce the same solution. The simpler boundary condition will be used even for test fields without this property for comparative purposes.

## 3.2 Wellposedness

It is well enough to setup a boundary value problem and motivate its representation of a physical system, but it is equally important to get some grip on its solution(s), if one exists. To this end Jacques Hadamard introduced the rational of wellposedness, emphasizing through example how existence, uniqueness and stability with regards to small changes in data can break down. This was done in *Sur les problèmes aux dérivées partielles et leur signification physique* [18] however, as comfort with the French language is required to fully grasp the concepts in this text, we follow Evans [19] and present the now formalized notion of Hadamard wellposedness.

**Definition** (Hadamard wellposedness). A partial differential equation with imposed boundary/initial conditions is well-posed if the following criteria are satisfied:

1. Existence: There exists a solution to the problem.
2. Uniqueness: This solution is unique.
3. Stability: The solution depends continuously on the data given in the problem.

Existence and uniqueness are important properties because they ensure that the solution we might get from the problem is in fact relevant and not another, equally valid solution, that do not enjoy any physical meaning. Moreover, the stability requirement is important to ensure that small changes in data put into the problem do not produce large changes in the solution. This ensures that measurement errors and numerical inaccuracies due to floating-point arithmetic do not lead to uncontrolled error amplification in the solution, i.e instability. For the boundary value problem in this study, data refers to boundary conditions, source term and diffusion tensor.

To ensure wellposedness of the boundary value problem, existence, uniqueness and stability are confirmed by the Lax-Milgram lemma.

**Theorem** (Lax-Milgram lemma). Let  $\mathcal{H}$  be a (real) Hilbert Space and  $\mathcal{H}^*$  be the dual space of  $\mathcal{H}$ . Let  $a : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$  be a bilinear form and  $L : \mathcal{H} \rightarrow \mathbb{R}$  be a linear functional i.e  $L \in \mathcal{H}^*$ . Assume there exists positive constants  $k_0 > 0$ ,  $k_1 > 0$  and  $k_2 > 0$  such that

1.  $a(w, v)$  is coercive;  $a(v, v) \geq k_0 \|v\|_{\mathcal{H}}^2$  for all  $v \in \mathcal{H}$ .
2.  $a(w, v)$  is bounded/continuous;  $|a(w, v)| \leq k_1 \|w\|_{\mathcal{H}} \|v\|_{\mathcal{H}}$  for all  $w, v \in \mathcal{H}$ .
3.  $L(v)$  is bounded/continuous;  $|L(v)| \leq k_2 \|v\|_{\mathcal{H}}$  for all  $v \in \mathcal{H}$

Then there exists a unique function  $u \in \mathcal{H}$  that is the solution to the variational (weak) problem

$$\text{find } u \in \mathcal{H} : a(u, v) = L(v) \quad \forall v \in \mathcal{H} \quad (3.16)$$

and the following stability measure holds

$$\|u\|_{\mathcal{H}} \leq \frac{1}{k_0} \|L\|_{\mathcal{H}^*}. \quad (3.17)$$

### 3.2.1 Variational (weak) formulation

To investigate wellposedness for the boundary value problem of this study a slight generalization of (BVP) is rewritten into a variational (weak) formulation such that  $a(\gamma, v) = L(v)$  where  $a(\gamma, v)$  is bilinear, i.e linear in each argument separately, and  $L(v)$  is linear. The generalization of (BVP), denoted by (BVP'), is

$$\begin{aligned} (BVP') \quad \nabla \cdot (K(\vec{x})\nabla\gamma) &= \nabla \cdot f(\vec{x}) \\ K(\vec{x})\nabla\gamma \cdot \vec{n} &= g(\vec{x}). \end{aligned}$$

Compared to (BVP) we consider here a general diffusion tensor  $K$ , source term  $f$  and boundary data  $g$ .

Coercivity and boundedness are argued, implying existence, uniqueness and stability through the Lax-Milgram lemma. From this result, wellposedness is guaranteed. Solutions and test functions will lie in the Hilbert space

$$\overline{H}(\Omega) = \{u \in H^1(\Omega) \mid \int_{\Omega} u dV = 0\}. \quad (3.18)$$

As will become clear throughout the arguments below, the space  $\overline{H}(\Omega)$  and the enforcement of zero mean of the solution is necessary to ensure wellposedness. Note also that the norm of this space is the same as the  $H^1(\Omega)$ -norm.

The boundary value problem (BVP') is rewritten to variational (weak) formulation by integrating over the volume and multiplying by the test function  $v \in \overline{H}(\Omega)$  on each side such that

$$\underbrace{\int_{\Omega} (\nabla \cdot (K\nabla\gamma)) v dV}_I = \underbrace{\int_{\Omega} (\nabla \cdot f) v dV}_II. \quad (3.19)$$

This expression is rewritten using integration by parts in multiple dimensions, that is the divergence theorem, such that

$$I = \int_{\Omega} (\nabla \cdot (K\nabla\gamma))v dV = \int_{\partial\Omega} (K\nabla\gamma)v \cdot \vec{n} dS - \int_{\Omega} \nabla v \cdot (K\nabla\gamma) dV,$$

$$II = \int_{\Omega} (\nabla \cdot f)v dV = \int_{\partial\Omega} fv \cdot \vec{n} dS - \int_{\Omega} \nabla v \cdot f dV,$$

$$I = II \Leftrightarrow \int_{\partial\Omega} (K\nabla\gamma)v \cdot \vec{n} dS - \int_{\Omega} \nabla v \cdot (K\nabla\gamma) dV = \int_{\partial\Omega} fv \cdot \vec{n} dS - \int_{\Omega} \nabla v \cdot f dV,$$

$$\Leftrightarrow \int_{\Omega} \nabla v \cdot (K\nabla\gamma) dV = - \int_{\partial\Omega} fv \cdot \vec{n} dS + \int_{\Omega} \nabla v \cdot f dV + \int_{\partial\Omega} (K\nabla\gamma)v \cdot \vec{n} dS.$$

Using the Neumann boundary condition  $K\nabla\gamma \cdot \vec{n} = g(\vec{x})$  over  $\partial\Omega$  this expression can be rewritten such that

$$\int_{\Omega} \nabla v \cdot (K\nabla\gamma) dV = \int_{\Omega} \nabla v \cdot f dV + \int_{\partial\Omega} gv dS - \int_{\partial\Omega} fv \cdot \vec{n} dS.$$

The general problem ( $BVP'$ ) have been successfully rewritten to variational form where

$$a(\gamma, v) = \int_{\Omega} \nabla v \cdot (K\nabla\gamma) dV \quad \text{is bilinear}$$

$$L(v) = \int_{\Omega} \nabla v \cdot f dV + \int_{\partial\Omega} v(g - f \cdot \vec{n}) dS \quad \text{is linear}$$

The variational (weak) formulation of the general problem ( $BVP'$ ) is therefore

$$\text{find } \gamma \in \overline{H}(\Omega) : a(\gamma, v) = L(v), \quad \forall v \in \overline{H}(\Omega). \quad (3.20)$$

### 3.2.2 Wellposedness by Lax-Milgram

Using the Lax-Milgram lemma we can now prove the following theorem:

**Theorem 3.2.1** *Suppose  $\Omega \subset \mathbb{R}^n$  is an open and bounded domain with Lipschitz boundary  $\partial\Omega$ . Let  $f \in H^1(\Omega)$ , and  $g \in L^2(\partial\Omega)$ . Moreover, assume  $D \in L^\infty(\Omega)$  is uniformly positive definite, that is*

$$\exists c > 0 \text{ s.t. } \xi^T K(\vec{x}) \xi \geq c|\xi|^2 \quad \forall \vec{x} \in \Omega. \quad (3.21)$$

*Then ( $BVP'$ ) has a unique solution  $u \in \overline{H}(\Omega)$ .*

This implies the following result regarding the non-generalized problem ( $BVP$ ).

**Corollary 3.2.1.1** *If  $\vec{A} \in H^1(\Omega) \cap L^\infty(\Omega)$ ,  $\mu > 0$  and  $|\vec{A}| \geq c > 0$  for all  $\vec{x} \in \mathbb{R}^n$  where  $n = 2, 3$  then (BVP) has a unique solution in  $\overline{H}(\Omega)$*

*Proof of Corollary 3.2.1.1*

The diffusion tensor  $K = D$ , as defined in equation (3.12) has eigenvalues  $|\vec{A}|^2$  and  $\mu|\vec{A}|^2$  for both two and three dimensions. Therefore, the diffusion tensor  $D$  satisfies (3.21) if  $|\vec{A}| \geq c > 0$  for all  $\vec{x} \in \Omega$  and if  $\mu > 0$ . Moreover,  $\vec{A} \in L^\infty(\Omega)$  implies that  $D \in L^\infty(\Omega)$ . Finally, we have  $f = \mu\vec{A}|\vec{A}|^2 \in H^1(\Omega)$  since  $\vec{A} \in H^1(\Omega) \cap L^\infty(\Omega)$  and therefore,  $g = f \cdot \vec{n} \in L^2(\partial\Omega)$  by the trace theorem (3.24).

*Proof of Theorem 3.2.1 by Lax-Milgram*

- **Proving coercivity of  $a(w, v)$** , that is  $\exists k_0 > 0$  such that  $a(v, v) \geq k_0 \|v\|_{\mathcal{H}}^2$  for all  $v \in \mathcal{H} = \overline{H}(\Omega)$

A lower limit for the bilinear form can be found because  $K$  is uniformly positive definite and  $\xi^T K \xi \geq c|\xi|^2$  implies  $\int_{\Omega} \xi^T K \xi dV \geq \int_{\Omega} c|\xi|^2 dV$ . Therefore,

$$a(v, v) = \int_{\Omega} \nabla v \cdot (K \nabla v) dV \geq c \int_{\Omega} |\nabla v|^2 dV = c \|\nabla v\|_{L^2(\Omega)}^2$$

Because  $\int_{\Omega} v dV = 0$  for all  $v \in \overline{H}(\Omega)$ , Poincaré's inequality can be applied which states that

$$\exists C > 0 \quad \|u\|_{L^2(\Omega)}^2 \leq C \|\nabla u\|_{L^2(\Omega)}^2 \quad \forall u \in H^1(\Omega).$$

The relationship between the squared  $\overline{H}(\Omega)$ -norm of  $v$  and the  $L^2(\Omega)$ -norm is  $\|v\|_{\overline{H}(\Omega)}^2 = \|v\|_{L^2(\Omega)}^2 + \|\nabla v\|_{L^2(\Omega)}^2 \Leftrightarrow \|v\|_{L^2(\Omega)}^2 = \|v\|_{\overline{H}(\Omega)}^2 - \|\nabla v\|_{L^2(\Omega)}^2$ . This expression can be used to rewrite Poincaré's inequality for  $u = v$ .

$$\begin{aligned} \|v\|_{\overline{H}(\Omega)}^2 - \|\nabla v\|_{L^2(\Omega)}^2 &\leq C \|\nabla v\|_{L^2(\Omega)}^2 \\ \Leftrightarrow \|v\|_{\overline{H}(\Omega)}^2 &\leq (C + 1) \|\nabla v\|_{L^2(\Omega)}^2 \\ \Leftrightarrow \|\nabla v\|_{L^2(\Omega)}^2 &\geq \frac{1}{(C + 1)} \|v\|_{\overline{H}(\Omega)}^2 \\ \Rightarrow a(v, v) &\geq c \|\nabla v\|_{L^2(\Omega)}^2 \geq \frac{c}{(C + 1)} \|v\|_{\overline{H}(\Omega)}^2 \end{aligned}$$

It has been proven that for  $k_0 = \frac{c}{(C+1)}$  then  $a(v, v) \geq k_0 \|v\|_{\overline{H}(\Omega)}^2$  for all

$v \in \overline{H}(\Omega)$ , which means  $a(u, v)$  is coercive on  $\overline{H}(\Omega)$ .

- **Proving boundedness of  $a(w, v)$** , that is  $\exists k_1 > 0$  such that  $|a(w, v)| \leq k_1 \|w\|_{\mathcal{H}} \|v\|_{\mathcal{H}}$  for all  $w, v \in \mathcal{H} = \overline{H}(\Omega)$ .

The Cauchy-Schwarz inequality gives for  $u_1, u_2 \in L^2$

$$|\langle u_1, u_2 \rangle_{L^2(\Omega)}| \leq \|u_1\|_{L^2(\Omega)} \|u_2\|_{L^2(\Omega)}. \quad (3.22)$$

For  $w, v \in \overline{H}(\Omega)$  we have  $\nabla w, \nabla v \in L^2(\Omega)$ . Because  $K \in L^\infty(\Omega)$  then  $K\nabla v \in L^2(\Omega)$ . Cauchy-Schwarz inequality now can be used to find a bound on the bilinear term.

$$\begin{aligned} |a(w, v)| &= \left| \int_{\Omega} \nabla w \cdot (K\nabla v) dV \right| = |\langle \nabla w, K\nabla v \rangle_{L^2(\Omega)}| \\ &\leq \|\nabla w\|_{L^2(\Omega)} \|K\nabla v\|_{L^2(\Omega)} \\ &\leq \|K\|_{L^\infty(\Omega)} \|\nabla w\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)}. \end{aligned}$$

Again, consider the relationship between the squared  $\overline{H}(\Omega)$ -norm on  $\zeta$  and the  $L^2(\Omega)$ -norm  $\|\zeta\|_{\overline{H}(\Omega)}^2 = \|\zeta\|_{L^2(\Omega)}^2 + \|\nabla \zeta\|_{L^2(\Omega)}^2$ . Through this the following inequality holds,

$$\|\zeta\|_{\overline{H}(\Omega)} \geq \|\nabla \zeta\|_{L^2(\Omega)}. \quad (3.23)$$

The bound for the bilinear form can be written in terms of  $\overline{H}(\Omega)$ -norms.

$$\begin{aligned} |a(w, v)| &\leq \|K\|_{L^\infty(\Omega)} \|\nabla w\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} \\ &\leq \|K\|_{L^\infty(\Omega)} \|w\|_{\overline{H}(\Omega)} \|v\|_{\overline{H}(\Omega)} \\ &= k_1 \|w\|_{\overline{H}(\Omega)} \|v\|_{\overline{H}(\Omega)}. \end{aligned}$$

It has been proven that  $|a(w, v)| \leq k_1 \|w\|_{\mathcal{H}} \|v\|_{\mathcal{H}}$  for all  $w, v \in \mathcal{H} = \overline{H}(\Omega)$  with  $k_1 = \|K\|_{L^\infty(\Omega)} > 0$ .

- **Proving boundedness of  $L(v)$** , that is  $\exists k_2 > 0$  such that  $|L(v)| \leq k_2 \|v\|_{\mathcal{H}}$  for all  $v \in \mathcal{H} = \overline{H}(\Omega)$ .

We have,

$$\begin{aligned} |L(v)| &= \left| \int_{\Omega} \nabla v \cdot f dV + \int_{\partial\Omega} v(g - f \cdot \vec{n}) dS \right| \\ &\leq \underbrace{\left| \int_{\Omega} \nabla v \cdot f dV \right|}_I + \underbrace{\left| \int_{\partial\Omega} v(g - f \cdot \vec{n}) dS \right|}_{II}. \end{aligned}$$

The treatments of  $I$  and  $II$  differ a bit due to the later integral being evaluated on the boundary. The first term is treated using Cauchy-Schwarz inequality (3.22) with  $\nabla v, f \in L^2(\Omega)$ .

$$\begin{aligned} I &= \left| \int_{\Omega} \nabla v \cdot f dV \right| = |\langle \nabla v, f \rangle_{L^2(\Omega)}| \\ &\leq \|\nabla v\|_{L^2(\Omega)} \|f\|_{L^2(\Omega)}. \end{aligned}$$

Through the relationship between  $\overline{H}(\Omega)$ -norm and  $L^2(\Omega)$ -norm expressed in (3.23), this inequality can be rewritten in terms of the  $\overline{H}(\Omega)$ -norm of  $v$ .

$$I \leq \|f\|_{L^2(\Omega)} \|v\|_{\overline{H}(\Omega)}.$$

Consider the second term  $II$  where  $g \in L^2(\partial\Omega)$  and  $f, v \in \overline{H}(\Omega)$ . The second term can be bounded above by Cauchy-Schwarz inequality (3.22) such that

$$\begin{aligned} II &= \left| \int_{\partial\Omega} v(g - f \cdot \vec{n}) dS \right| = |\langle v, g - f \cdot \vec{n} \rangle_{L^2(\partial\Omega)}| \\ &\leq \|v\|_{L^2(\partial\Omega)} \|g - f \cdot \vec{n}\|_{L^2(\partial\Omega)} \leq \|v\|_{L^2(\partial\Omega)} \left( \|g\|_{L^2(\partial\Omega)} + \|f\|_{L^2(\partial\Omega)} \right). \end{aligned}$$

The trace theorem states that there exists a  $\bar{c}$  such that

$$\|\zeta\|_{L^2(\partial\Omega)} \leq \bar{c} \|\zeta\|_{H^1(\Omega)} \quad \forall \zeta \in H^1(\Omega) \quad (3.24)$$

if the boundary of  $\Omega$  is Lipschitz. This theorem can be applied to  $\|f\|_{L^2(\partial\Omega)}$  and  $\|v\|_{L^2(\partial\Omega)}$  to arrive at an appropriate bound for the second term. Recall that the norm of  $\overline{H}$  is the same as the one on  $H^1$  and we have,

$$II \leq \bar{c} \left[ \|g\|_{L^2(\partial\Omega)} + \|f\|_{H^1(\Omega)} \right] \|v\|_{\overline{H}(\Omega)}.$$

Using the bounds on the first and second term,  $I$  and  $II$  we have,

$$\begin{aligned} L(v) = I + II &\leq \|f\|_{L^2(\Omega)} \|v\|_{\overline{H}(\Omega)} + \bar{c} \left[ \|g\|_{L^2(\partial\Omega)} + \|f\|_{H^1(\Omega)} \right] \|v\|_{\overline{H}(\Omega)} \\ &= \left( (1 + \bar{c}) \|f\|_{H^1(\Omega)} + \bar{c} \|g\|_{L^2(\partial\Omega)} \right) \|v\|_{\overline{H}(\Omega)}. \end{aligned}$$

It has been proven that  $\exists k_2 > 0$  such that  $|L(v)| \leq k_2 \|v\|_{\mathcal{H}}$  for all  $v \in \mathcal{H} = \overline{H}(\Omega)$  with  $k_2 = (1 + \bar{c}) \|f\|_{H^1(\Omega)} + \bar{c} \|g\|_{L^2(\partial\Omega)} > 0$ .

This concludes the proof.

### 3.3 Effect of $\mu$ on the boundary value problem

In this section we discuss how the choice of the parameter  $\mu$  might affect the boundary value problem (*BVP*) and its solution for two dimensional domains. First we conclude that for conservative vector fields  $\vec{A}$  in 2D the solution is independent of  $\mu$ . Then we conclude that  $\|\nabla\gamma \times \vec{A}\|_{L^2(\Omega)} \rightarrow 0$  as  $\mu \rightarrow 0$  in both 3D and 2D. Finally, we find that for  $\mu = 1$ , (*BVP*) is isotropic in diffusion. This last statement will be true for both two and three dimensional domains. In the discussion of how (*BVP*) behaves at  $\mu = 1$ , a connection is also found between the Euler-Lagrange equation of (*BVP*) in 2D and the Poisson equation.

1. For conservative vector fields  $\vec{A}$  in 2D, the solution is independent of  $\mu$ . To understand this dependence we examine the optimization problem (*P*). The integrand in (*P*) is designed to penalizes deviations of the gradient to the direction and magnitude of the diapycnal vector  $\vec{A}$ ;  $I = |\nabla\gamma \times \vec{A}|^2$  and  $II = \mu(\vec{A} \cdot \nabla\gamma - |\vec{A}|^2)^2$ .

All solutions gradients can be written as a linear combination of two linearly independent vectors. Consider the diapycnal vector  $\vec{A}$ , the vector perpendicular the diapycnal vector  $\vec{A}^\perp = [A_z, -A_y]^T$  where  $|\vec{A}^\perp| = |\vec{A}|$ . Consider the gradient of neutral density as a linear combination of  $\vec{A}^\perp$  and  $\vec{A}$  such that

$$\nabla\gamma = c_0(\vec{x})\vec{A} + c_1(\vec{x})\vec{A}^\perp. \quad (3.25)$$

The second term  $II$  can be simplified using this expression. We have,

$$\begin{aligned} II &= \mu(\vec{A} \cdot \nabla\gamma - |\vec{A}|^2)^2 = \mu(\vec{A} \cdot (c_0(\vec{x})\vec{A} + c_1(\vec{x})\vec{A}^\perp) - |\vec{A}|^2)^2 \\ &= \mu(c_0(\vec{x})|\vec{A}|^2 + c_1(\vec{x})0 - |\vec{A}|^2)^2 \\ &= \mu(c_0(\vec{x}) - 1)^2|\vec{A}|^4. \end{aligned}$$

Using the 2D-curl for the cross product as defined in equation (3.7) and the explicit expression for  $\vec{A}^\perp$  we have,

$$\begin{aligned} \nabla\gamma \times \vec{A} &= (c_0(\vec{x})A_y + c_1(\vec{x})A_z)A_z - (c_0(\vec{x})A_z - c_1(\vec{x})A_y)A_y \\ &= c_1(\vec{x})(A_z^2 + A_y^2) = c_1(\vec{x})|\vec{A}|^2 \\ \Rightarrow I &= |\nabla\gamma \times \vec{A}|^2 = c_1(\vec{x})^2|\vec{A}|^4. \end{aligned}$$

To summarize, we have found that the first term  $I$  is given as  $I = c_1(\vec{x})^2|\vec{A}|^4$  and the second term  $II$  is given as  $II = \mu(c_0(\vec{x}) - 1)^2|\vec{A}|^4$ . If  $\vec{A}$  is conservative then a solution  $\gamma$  exists such that  $\nabla\gamma = \vec{A}$ . For this solution we have that  $c_0(\vec{x}) = 1$  and  $c_1(\vec{x}) = 0$  which means both  $I$  and  $II$  will be zero. Consequently, this  $\gamma$  is the minimizer of  $(P)$  for all  $\mu$  since we have a well-posed problem, see Section 3.2. This means the solution is independent of  $\mu$ . If  $\vec{A}$  is not conservative then such a solution does not exist. Deviations from either magnitude or direction will be present in the solution and independence to  $\mu$  is not guaranteed.

2.  $\|\nabla\gamma \times \vec{A}\|_{L^2(\Omega)} \rightarrow 0$  as  $\mu \rightarrow 0$  in 2D and in 3D when the helicity  $H$  is zero. To understand this, let us consider a candidate  $\gamma^*$  for  $(P)$  chosen such that  $|\nabla\gamma^* \times \vec{A}| = 0$ . This candidate exists when  $H = 0$ . We define  $C_{\gamma^*}$  by,

$$\int_{\Omega} |\nabla\gamma^* \times \vec{A}|^2 d\Omega + \mu \int_{\Omega} (\vec{A} \cdot \nabla\gamma^* - |\vec{A}|^2)^2 d\Omega = \mu C_{\gamma^*}.$$

A solution  $\gamma$ , which exists uniquely as proven in Section 3.2, gives an objective value smaller or equals to that of the candidate  $\gamma^*$ . Because  $(\vec{A} \cdot \nabla\gamma^* - |\vec{A}|^2)^2 > 0$  we have,

$$\begin{aligned} &\int_{\Omega} |\nabla\gamma \times \vec{A}|^2 d\Omega + \mu \int_{\Omega} (\vec{A} \cdot \nabla\gamma - |\vec{A}|^2)^2 d\Omega \leq \mu C_{\gamma^*} \\ \Rightarrow &\int_{\Omega} |\nabla\gamma \times \vec{A}|^2 d\Omega \leq \mu C_{\gamma^*} \\ \Leftrightarrow &\|\nabla\gamma \times \vec{A}\|_{L^2(\Omega)}^2 \leq \mu C_{\gamma^*}. \end{aligned}$$

From this upper bound on the squared-norm we can conclude that

$$\|\nabla\gamma \times \vec{A}\|_{L^2(\Omega)} \rightarrow 0 \text{ as } \mu \rightarrow 0$$

*Conjecture* When the helicity  $H$ , as defined in 1.1, is zero in 3D or 2D, we have  $\nabla\gamma \rightarrow b\vec{A}$  for  $b \neq 0$ . However, when  $H \neq 0$  in 3D, then  $\gamma \rightarrow 0$ .

3. If  $\mu = 1$ , then (BVP) is isotropic in diffusion. The diffusion tensor  $D$  of (BVP) as given by equation (3.12) and holds for both two and three dimensional domains. For  $\mu = 1$  we get,

$$D = I|\vec{A}|^2 - \vec{A}\vec{A}^T + \mu\vec{A}\vec{A}^T = I|\vec{A}|^2 - \vec{A}\vec{A}^T + \vec{A}\vec{A}^T = |\vec{A}|^2 I.$$

Putting this into (BVP) using  $\mu = 1$  yields

$$\nabla \cdot \left( |\vec{A}|^2 \nabla \gamma \right) = \nabla \cdot \left( \vec{A} |\vec{A}|^2 \right)$$

which is an isotropic diffusion equation.

Let us consider the details of this expression in the two dimensional setting where  $\vec{A} = [A_y, A_z]^T$ . The diapycnal vector  $\vec{A}$  depends on quantities such as absolute salinity and conservative temperature. These in turn depend on location. Therefore, evaluation of the divergence in the analytic setting must be done using the product rule.

$$\begin{aligned} \nabla \cdot \left( |\vec{A}|^2 \nabla \gamma \right) &= \frac{\partial(|\vec{A}|^2 \frac{\partial \gamma}{\partial y})}{\partial y} + \frac{\partial(|\vec{A}|^2 \frac{\partial \gamma}{\partial z})}{\partial z} \\ &= \frac{\partial(|\vec{A}|^2)}{\partial y} \frac{\partial \gamma}{\partial y} + |\vec{A}|^2 \frac{\partial^2 \gamma}{\partial y^2} + \frac{\partial(|\vec{A}|^2)}{\partial z} \frac{\partial \gamma}{\partial z} + |\vec{A}|^2 \frac{\partial^2 \gamma}{\partial z^2} \\ &= |\vec{A}|^2 \nabla^2 \gamma + 2A_y \frac{\partial A_y}{\partial y} \frac{\partial \gamma}{\partial y} + 2A_z \frac{\partial A_z}{\partial z} \frac{\partial \gamma}{\partial z} \\ &= |\vec{A}|^2 \nabla^2 \gamma + 2(\vec{A} \circ \left[ \begin{array}{c} \frac{\partial A_y}{\partial y} \\ \frac{\partial A_z}{\partial z} \end{array} \right]) \cdot \nabla \gamma \end{aligned} \quad (3.26)$$

where  $\nabla^2$  is the Laplace operator.

From this expression, we note that if the variation of the components of  $\vec{A}$  in the two grid directions respectively is sufficiently small then the

Euler Lagrange equation reduces approximately to the Poisson equation at  $\mu = 1$ . Only if  $\vec{A}$  is constant in space does the problem reduce to the Poisson equation exactly at  $\mu = 1$ . The second term of (3.26) is zero in that case.

Even though the problem does not exactly reduce to the well known Poisson equation when  $\mu = 1$ , choosing  $\mu = 1$  still has its advantages since the problem becomes isotropic, rather than anisotropic, in diffusion. In isotropic diffusion, the partial differential equation does not include mixed derivatives as can be seen in the simplified expression of  $\nabla(|\vec{A}|^2 \nabla \gamma)$  at  $\mu = 1$ , see equation (3.26). An equation without mixed partials is simpler to handle because the stencil can be smaller. In this study a 9-point stencil is used however, if we were to only consider isotropic diffusion a 5 point stencil would suffice. This would ease implementation as well as handling of boundary contributions. A second advantage of choosing  $\mu = 1$  is that having mixed derivatives results in diffusion directions that are rotated relative to the grid. Such behaviour is harder to capture numerically if extra care is not taken as discussed in [20].

# Chapter 4

## Finite differences

In this study, the neutral density variable is found in 2D on a regular rectangular domain such that  $(y, z) \in \Omega \subset \mathbb{R}^2$  by solving (BVP) using finite differences. To formalize this method, consider the generalized version (BVP') with  $g = f \cdot \vec{n}$  on a 2D domain such as

$$\begin{aligned}\nabla \cdot (K(y, z) \nabla \gamma) &= \nabla \cdot f(y, z), \\ K(y, z) \nabla \gamma \cdot \vec{n} &= f(y, z) \cdot \vec{n}.\end{aligned}$$

In the discrete setting, the following notation is used to define cell centre points and faces. Suppose the centre points in the domain is characterized by indices  $(i, j) \in \mathcal{I}_y \times \mathcal{I}_z = [1, \dots, n_y] \times [1, \dots, n_z] \subset \mathbb{N}_+^2$ . Then cell centres and cell faces in the discrete setting with step-sizes  $\Delta y$  and  $\Delta z$  are given by

$$\begin{aligned}\text{Cell centres} &= \{(y_i, z_j) \mid i \in \mathcal{I}_y, j \in \mathcal{I}_z\}, \\ \text{Cell edges} &= \{(y_{i+\frac{1}{2}}, z_{j+\frac{1}{2}}) \mid i \in \{0\} \cup \mathcal{I}_y, j \in \{0\} \cup \mathcal{I}_z\}\end{aligned}$$

where

$$\begin{aligned}(y_i, z_j) &= (y_0 + (i - 1)\Delta y, z_0 + (j - 1)\Delta z), \\ (y_{i+\frac{1}{2}}, z_{j+\frac{1}{2}}) &= (y_0 + (i - \frac{1}{2})\Delta y, z_0 + (j - \frac{1}{2})\Delta z).\end{aligned}$$

In this thesis, discrete quantities  $q$  living on the centre points are denoted by  $q_{i,j}$ . Analogously, on faces they are denoted by  $q_{i\pm\frac{1}{2},j}$  or  $q_{i,j\pm\frac{1}{2}}$ . The neutral density variable found in this study is an approximate variable living on cell

centres. The neutral density variable  $\gamma$  has the following discrete notation

$$\gamma(y_i, z_j) \approx \gamma_{i,j} \quad i \in \mathcal{I}_y, j \in \mathcal{I}_z.$$

When a quantity  $q(y, z)$  is to be considered without approximation at a specific discrete point it will have the following notation:

$$q(y_i, z_j) = q_{i,j}. \quad (4.1)$$

The finite difference scheme for  $(BVP')$  will result in a linear system of equations such that  $A_m \vec{\gamma}_{flat} = \vec{b}_{rhs}$  where

$$A_m \in \mathbb{R}^{(n_y+1) \times (n_z+1)}, \quad \vec{\gamma}_{flat} \in \mathbb{R}^{n_y n_z + 1}, \quad \vec{b}_{rhs} \in \mathbb{R}^{n_y n_z + 1}.$$

$A_m$  is the system matrix,  $\vec{b}_{rhs}$  is the right hand side vector containing source and boundary data and  $\vec{\gamma}_{flat}$  contains the approximations  $\{\gamma_{i,j}\}$  flattened into a vector over the entire domain. This notation is used to eliminate confusion with the diapycnal vector  $\vec{A}$ , factor  $b(\vec{x})$  and scalar valued function  $\gamma$  when working in the discrete setting.

**Remark** The linear system of equation coming out of the finite difference discretization will be modified slightly, see Section 4.3. This is done so that requirements for wellposedness are satisfied and it is the reason for the dimensions of the linear system of equations given above.

## 4.1 Finite difference scheme

The building blocks of finite difference schemes are presented in this section. We consider discretization in the interior, that is indices  $(i, j)$  such that  $(y_i, z_j) \in \Omega$ . For the interested reader, the complete derivation of the finite difference discretization of Euler Lagrange equation can be found in Appendix A.

Consider the generalized problem  $(BVP')$  and suppose  $F := K \nabla \gamma$  where

$$F := \begin{bmatrix} F^{(y)} \\ F^{(z)} \end{bmatrix}, \quad K := \begin{bmatrix} K^{(yy)} & K^{(yz)} \\ K^{(zy)} & K^{(zz)} \end{bmatrix}, \quad \nabla \gamma = \begin{bmatrix} \frac{\partial \gamma}{\partial y} \\ \frac{\partial \gamma}{\partial z} \end{bmatrix}.$$

There is not one unique way of handling the discretization of these terms. In this study an asymmetric scheme described by van Es et al. [20] is used. This article also includes other alternative schemes and discusses their effectiveness to handle anisotropic diffusion.

In practice, hydrographic data, from which the diapycnal vector  $\vec{A}$  and therefore the diffusion tensor  $K(\vec{x})$  and source  $f(\vec{x})$  are built, are only known at cell centres. In the finite difference scheme however, these quantities are needed at cell faces. Therefore, a second order interpolation method is utilized to get data on cell faces when the original data is given at cell centres. For an arbitrary quantity  $q$  we write

$$q(y_{i+\frac{1}{2}}, z_j) \approx \bar{q}_{i+\frac{1}{2},j} = \frac{q_{i+1,j} + q_{i,j}}{2},$$

$$q(y_i, z_{j+\frac{1}{2}}, z_j) \approx \bar{q}_{i,j+\frac{1}{2}} = \frac{q_{i,j+1} + q_{i,j}}{2}.$$

Formulas for  $q_{i-\frac{1}{2},j}$  and  $q_{i,j-\frac{1}{2}}$  are analogous.

Furthermore, second order approximations of partial derivatives are used in an asymmetric manner. The partial derivatives of the variable  $\gamma$  are approximated on cell faces as follows:

$$\frac{\partial \gamma}{\partial y}(y_{i+\frac{1}{2}}, z_j) \approx d_y \gamma_{i+\frac{1}{2},j} = \frac{\gamma_{i+1,j} - \gamma_{i,j}}{\Delta y},$$

$$\frac{\partial \gamma}{\partial z}(y_i, z_{j+\frac{1}{2}}) \approx d_z \gamma_{i,j+\frac{1}{2}} = \frac{\gamma_{i,j+1} - \gamma_{i,j}}{\Delta z},$$

$$\frac{\partial \gamma}{\partial y}(y_i, z_{j+\frac{1}{2}}) \approx d_y \gamma_{i,j+\frac{1}{2}} = \frac{\gamma_{i+1,j+1} + \gamma_{i+1,j} - \gamma_{i-1,j+1} - \gamma_{i-1,j}}{4\Delta y},$$

$$\frac{\partial \gamma}{\partial z}(y_{i+\frac{1}{2}}, z_j) \approx d_z \gamma_{i+\frac{1}{2},j} = \frac{\gamma_{i+1,j+1} + \gamma_{i,j+1} - \gamma_{i,j-1} - \gamma_{i+1,j-1}}{4\Delta z}.$$

Analogous formulas are applied to form  $d_y \gamma_{i-\frac{1}{2},j}$ ,  $d_z \gamma_{i,j-\frac{1}{2}}$ ,  $d_y \gamma_{i,j-\frac{1}{2}}$  and  $d_z \gamma_{i-\frac{1}{2},j}$ .

The diffusion tensor at cell faces is applied to the gradient of  $\gamma$  approximated by the above formulas to arrive at the flux  $F = K(y, z)\nabla\gamma$  at cell faces. We

have,

$$\begin{aligned}
 F(y_{i+\frac{1}{2}}, z_j) &= K(y_{i+\frac{1}{2}}, z_j) \nabla \gamma(y_{i+\frac{1}{2}}, z_j) \\
 &\approx F_{i+\frac{1}{2}, j} = \bar{K}_{i+\frac{1}{2}, j} \begin{bmatrix} d_y \gamma_{i+\frac{1}{2}, j} \\ d_z \gamma_{i+\frac{1}{2}, j} \end{bmatrix} =: \begin{bmatrix} F_{i+\frac{1}{2}, j}^{(y)} \\ F_{i+\frac{1}{2}, j}^{(z)} \end{bmatrix}.
 \end{aligned}$$

Formulas for  $F_{i-\frac{1}{2}, j}$ ,  $F_{i, j+\frac{1}{2}}$  and  $F_{i, j-\frac{1}{2}}$  are analogous.

The divergence of the flux,  $\nabla \cdot F$ , drives the flux term onto cell centres and is treated as follows:

$$(\nabla \cdot F)(y_i, z_j) \approx \frac{F_{i+\frac{1}{2}, j}^{(y)} - F_{i-\frac{1}{2}, j}^{(y)}}{\Delta y} + \frac{F_{i, j+\frac{1}{2}}^{(z)} - F_{i, j-\frac{1}{2}}^{(z)}}{\Delta z}.$$

In Figure 4.1, the finite difference scheme of this study is illustrated by tracing how discrete values of  $\gamma$  are used to arrive at components of the flux term and subsequently, how the components of the flux term are used to arrive at an approximate expression for the full divergence term. To understand the finite difference discretization better, compare Figure 4.1 with the derivations in Appendix A.

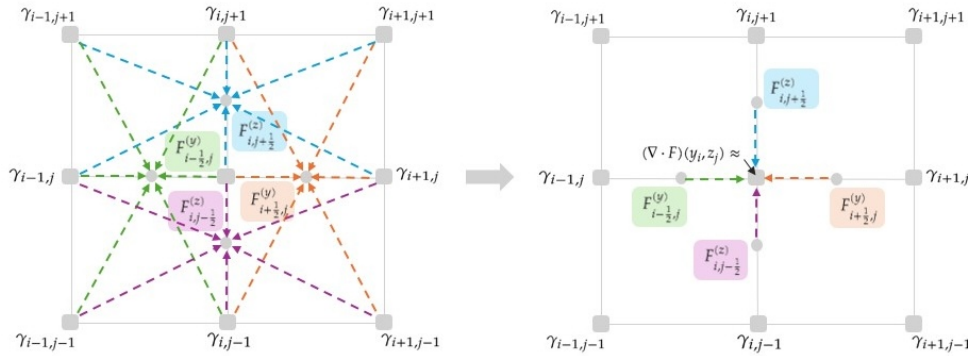


Figure 4.1: Illustration of the finite difference scheme utilized in this study (adapted from [20]).

## 4.2 Boundary conditions

In this section we consider discretization on the boundary, that is indices  $(i, j)$  such that  $(y_i, z_j) \in \partial\Omega$ . We use ghost points to approximate the boundary conditions. These ghost points are grid points located just outside the domain  $\Omega$ . The domain in consideration in this study has four straight edges aligning with axis directions. Therefore, the outer unit normal vectors of each boundary are trivial and the boundary condition can be simplified such that

$$\begin{aligned} \text{left: } \vec{n} &= [-1, 0]^T, \text{ right: } \vec{n} = [1, 0]^T \Rightarrow K^{(yy)} \frac{\partial\gamma}{\partial y} + K^{(yz)} \frac{\partial\gamma}{\partial z} = g, \\ \text{upper: } \vec{n} &= [0, 1]^T, \text{ lower: } \vec{n} = [0, -1]^T \Rightarrow K^{(zy)} \frac{\partial\gamma}{\partial y} + K^{(zz)} \frac{\partial\gamma}{\partial z} = g. \end{aligned}$$

We can now discretize the boundary condition to get an expression for  $\gamma_{i,j}$  at ghost points in terms of  $\gamma_{i,j}$  at grid points in  $\Omega$ .

### 4.2.1 Natural Neumann boundary conditions

To get expressions for ghost point values in the finite difference scheme we consider the possible scenarios. Suppose  $(i \pm 1, j)$  for some  $i \in \mathcal{I}_y$  and  $j \in \mathcal{I}_z$  does not exist in the domain, that is the ghost point to consider is beyond one of the two boundaries in the  $y$ -axis. By using central differences to approximate the boundary condition

$$K^{(yy)} \frac{\partial\gamma}{\partial y} + K^{(yz)} \frac{\partial\gamma}{\partial z} = g,$$

expressions for these ghost points can be found. We get

$$\begin{aligned} K_{i,j}^{(yy)} \frac{\gamma_{i+1,j} - \gamma_{i-1,j}}{2\Delta y} + K_{i,j}^{(yz)} \frac{\gamma_{i,j+1} - \gamma_{i,j-1}}{2\Delta z} &= g_{i,j} \\ \Leftrightarrow \gamma_{i+1,j} &= -\frac{\Delta y}{\Delta z} \frac{K_{i,j}^{(yz)}}{K_{i,j}^{(yy)}} (\gamma_{i,j+1} - \gamma_{i,j-1}) + \gamma_{i-1,j} + \frac{2\Delta y}{K_{i,j}^{(yy)}} g_{i,j} \\ \Leftrightarrow \gamma_{i-1,j} &= \frac{\Delta y}{\Delta z} \frac{K_{i,j}^{(yz)}}{K_{i,j}^{(yy)}} (\gamma_{i,j+1} - \gamma_{i,j-1}) + \gamma_{i+1,j} - \frac{2\Delta y}{K_{i,j}^{(yy)}} g_{i,j}. \end{aligned}$$

Ghost point values are thus expressed as a linear combination of interior points and boundary values. These linear combinations are then entered into the finite

difference scheme.

Analogously, suppose  $(i, j \pm 1)$  for some  $i \in \mathcal{J}_y$  and  $j \in \mathcal{J}_z$  does not exist in the domain, that is the ghost point to consider is beyond one of the two boundaries in the z-axis. Expressions for these ghost points can be found by approximating the boundary condition

$$K^{(zy)} \frac{\partial \gamma}{\partial y} + K^{(zz)} \frac{\partial \gamma}{\partial z} = g$$

with central differences,

$$\begin{aligned} K_{i,j}^{(zy)} \frac{\gamma_{i+1,j} - \gamma_{i-1,j}}{2\Delta y} + K_{i,j}^{(zz)} \frac{\gamma_{i,j+1} - \gamma_{i,j-1}}{2\Delta z} &= g_{i,j} \\ \Leftrightarrow \gamma_{i,j+1} &= -\frac{\Delta z}{\Delta y} \frac{K_{i,j}^{(zy)}}{K_{i,j}^{(zz)}} (\gamma_{i+1,j} - \gamma_{i-1,j}) + \gamma_{i,j-1} + \frac{2\Delta z}{K_{i,j}^{(zz)}} g_{i,j} \\ \Leftrightarrow \gamma_{i,j-1} &= \frac{\Delta z}{\Delta y} \frac{K_{i,j}^{(zy)}}{K_{i,j}^{(zz)}} (\gamma_{i+1,j} - \gamma_{i-1,j}) + \gamma_{i,j+1} - \frac{2\Delta z}{K_{i,j}^{(zz)}} g_{i,j}. \end{aligned}$$

Again, ghost point values can be expressed as a linear combination of interior points and boundary values, which are entered into the finite difference scheme.

## 4.2.2 Simple Neumann boundary condition

For comparative purposes, a simplified version of the Neumann boundary conditions is considered as well. This simple Neumann boundary condition is given by setting

$$K = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad g = \vec{A} \cdot \vec{n} \Rightarrow \nabla \gamma \cdot \vec{n} = \vec{A} \cdot \vec{n}$$

in the formulas above where

$$\vec{A} := \begin{bmatrix} A^{(y)} \\ A^{(z)} \end{bmatrix}.$$

**Remark** Previously  $\vec{A} = [A_y, A_z]^T$  has been used to denote components of the diapycnal vector for simplicity. During discretization of the equation this notation is changes so that subscripts can be used to describe positioning in

the domain.

The formulas for ghost points can then be simplified and we arrive at

$$\begin{aligned}\gamma_{i+1,j} &= \gamma_{i-1,j} + 2\Delta y A_{i,j}^{(y)}, \\ \gamma_{i-1,j} &= \gamma_{i+1,j} - 2\Delta y A_{i,j}^{(y)},\end{aligned}$$

$$\begin{aligned}\gamma_{i,j+1} &= \gamma_{i,j-1} + 2\Delta z A_{i,j}^{(z)}, \\ \gamma_{i,j-1} &= \gamma_{i,j+1} - 2\Delta z A_{i,j}^{(z)}.\end{aligned}$$

### 4.3 Solving the linear system of equations

To ensure wellposedness, zero mean of  $\gamma$  must be ensured over the domain. This is done by introducing Lagrange multipliers which means briefly returning to the point of view of optimization.

The linear system of equations from the finite difference discretization can be written as a quadratic minimization problem. Suppose we have the following quadratic minimization problem

$$\min h(\vec{\gamma}_{flat}) = \frac{1}{2} \vec{\gamma}_{flat} A_m \vec{\gamma}_{flat} - \vec{b}_{rhs} \vec{\gamma}_{flat} \quad (4.2)$$

$$\vec{\gamma}_{flat} \in \mathbb{R}^N \quad (4.3)$$

A unique solution to this problem exists if  $A_m$  is positive definite and that solution is given by  $\nabla h(\vec{\gamma}_{flat}) = A_m \vec{\gamma}_{flat} - \vec{b}_{rhs} = 0$ . This is exactly the linear system of equations to be solved. Therefore, to impose the additional condition of zero mean on the original linear system of equations, Lagrange Multipliers can be introduced to the quadratic minimization problem and its optimality conditions will result in a slightly modified linear system of equations.

The condition of zero mean is such that

$$I = \int_{\Omega} \gamma dV = 0. \quad (4.4)$$

To achieve second order accurate numerical integration, the trapezoidal rule in 2D can be utilized such that

$$I \approx \sum_i^{n_y} \sum_j^{n_z} \Delta y \Delta z w_{i,j} \gamma_{i,j} = 0$$

$$\Leftrightarrow \vec{w}^T \vec{\gamma}_{flat} = 0$$

where the weight  $w_{i,j}$  is 1 on interior points,  $\frac{1}{2}$  on the edges and  $\frac{1}{4}$  in the corners.

Using this numerical integration, the quadratic problem is minimized under the additional condition that  $\vec{w}^T \vec{\gamma}_{flat} = 0$ . The Lagrangian therefore becomes

$$L(\vec{\gamma}_{flat}, \lambda) = \frac{1}{2} \vec{\gamma}_{flat}^T A_m \vec{\gamma}_{flat} - \vec{b}_{rhs}^T \vec{\gamma}_{flat} - \lambda \vec{w}^T \vec{\gamma}_{flat} \quad (4.5)$$

with corresponding optimality conditions

$$\nabla_{\vec{\gamma}_{flat}} L = 0, \quad \frac{\partial L}{\partial \lambda} = 0$$

$$\Leftrightarrow A_m \vec{\gamma}_{flat} - \vec{b}_{rhs} - \lambda \vec{w} = 0$$

$$\vec{w}^T \vec{\gamma}_{flat} = 0.$$

After finite difference discretization and when including the additional constraint of zero mean the system of equation to be solved is

$$\begin{bmatrix} A_m & \vec{w} \\ \vec{w}^T & 0 \end{bmatrix} \begin{bmatrix} \vec{\gamma}_{flat} \\ \lambda \end{bmatrix} = \begin{bmatrix} \vec{b}_{rhs} \\ 0 \end{bmatrix}. \quad (4.6)$$

# Chapter 5

## Results and analysis

In this chapter, we give details on the implementation of the method presented in this thesis and an explanation of the test environment. Numerical examples are formalized and their characteristics introduced. Moreover, results are presented after each test case is introduced and include convergence behaviour in the  $L_2$ -norm as well as root mean square measures that quantify the accuracy of the solutions.

### 5.1 Software implementation and data

The scheme was implemented in python to solve  $(BVP')$ . The program was then used to solving the case-specific problem  $(BVP)$  with boundary conditions

$$\begin{aligned} D\nabla\gamma \cdot \vec{n} &= f \cdot \vec{n} && \text{on } \partial\Omega && \text{(Natural boundary conditions)} \\ \text{or } \nabla\gamma \cdot \vec{n} &= \vec{A} \cdot \vec{n} && \text{on } \partial\Omega && \text{(Simple boundary conditions).} \end{aligned}$$

The resulting linear system of equations (4.6) was solved using built in *spsolve* from the Scipy library. Both the natural and simple boundary conditions are possible choices in this implementation to modify the linear system of equations and include boundary contributions.

Hydrographic datasets containing values of salinity and temperature are necessary when testing the finite difference methods on the real diapycnal vector field  $\vec{A}$  as it is defined through oceanographic quantities. In this study, a simulated hydrographic dataset that contains values of absolute salinity and conservative temperature is used, see Figure 5.1. As this study does not

delve into the question of neutrality of the solutions or the realism of the neutral trajectories, the similarity between these simulations and real oceanic measurements will not be discussed. In this dataset,  $z$  decreases downwards and is zero at the surface. Moreover, while the meridional direction is regular in  $\Delta y$  it has varying steps sizes  $\Delta z$  in the vertical direction.

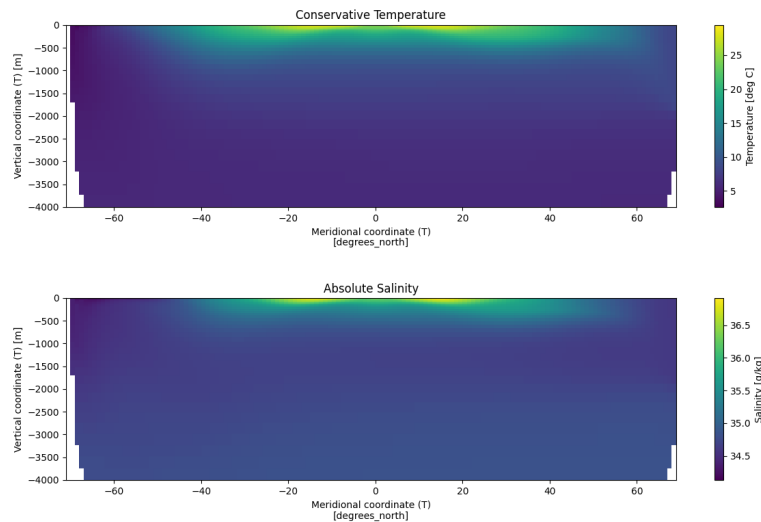


Figure 5.1: Absolute salinity and conservative temperature from a simulated hydrographic dataset at 29.5 degrees east in zonal direction.

The finite difference discretization of this study assumes regularly spaced grid points in both  $y$ - and  $z$ -direction when looking at trivial test cases and the more realistic oceanic setting. Generally, hydrographic datasets containing salinity, temperature and pressure values are not that simple. The simulated dataset used to evaluate the method in this study is not an exception since  $\Delta z$  varies. Because this study does not consider the realism or neutrality of the generated solutions but is more concerned with the behaviour of the method when exposed to different scenarios, the simulated hydrographic dataset is simply mapped to a regular domain as an idealization when looking at the real oceanic setting and diapycnal vector  $\vec{A}$ .

The Gibbs-SeaWater Oceanographic Toolbox (the `gsw` library in python) is used to derive quantities such as the haline contraction coefficient and the thermal expansion coefficients from the state variables in the simulated

hydrographic dataset, see Figure 5.2. Finally, python libraries *xarray* and *xgcm* were used to organize data on cell centres and cell edges in a more systematic and safe manner.

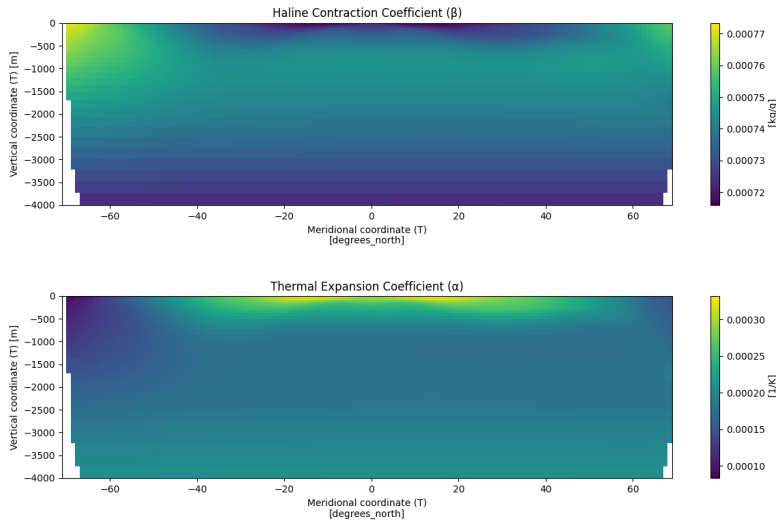


Figure 5.2: Thermal expansion and haline contraction coefficients from a simulated hydrographic dataset at 29.5 degrees east in zonal direction.

## 5.2 Test cases and performance metrics

The finite difference method of this study is evaluated on simple rectangular domains for conservative fields and non-conservative fields. The real diapycnal vector field lives on a domain where the meridional direction is much wider than the depth of the vertical direction, see Figure 5.1 and 5.2 and consider that one degree in meridional direction corresponds to approximately 110 km. For conservative testcases, the factor  $b(\vec{x})$  is known to be 1 everywhere. A test case which is non-conservative is also constructed so that the behaviour of the method can be evaluated when the solution  $\gamma$  follows  $\nabla\gamma = bA$  for some unknown  $b \neq 1$ .

**Remark** For artificial test cases,  $\vec{A}$  will denote the vector field in question. However, this  $\vec{A}$  is not to be confused with the real diapycnal vector field defined through hydrographic data.

To understand the behaviour of our method, we consider empirical convergence in the  $L_2$ -norm for all test cases. For conservative test cases, where a potential  $\gamma$  such that  $\vec{A} = \nabla\gamma$  exists, the analytic solution is used to find the empirical convergence through the discrete  $L_2$ -norm defined as

$$E_h^{exact} = \|\gamma - \gamma^{exact}\|_2 = \sqrt{\Delta y \Delta z \sum_i^{n_y} \sum_j^{n_z} (\gamma_{i,j} - \gamma_{i,j}^{exact})^2} \quad (5.1)$$

where  $\Delta y, \Delta z \sim h$ .

In cases where there is no such analytical solution, the convergence is evaluated using successive numerical solutions and  $E_h$  is instead computed as

$$E_h^{approx} = \|\gamma_h - \gamma_{\frac{h}{2}}^{sample}\|_2 \quad (5.2)$$

where  $\gamma_{\frac{h}{2}}^{sample}$  is the solution found on a twice as fine grid but sampled at the coarser grid points.

The empirical convergence is found through iteratively refining the grid by a factor of two in each direction and then looking at the rates

$$r_h = \frac{\log(E_h/E_{\frac{h}{2}})}{\log(2)}. \quad (5.3)$$

### 5.3 Conservative fields

In this section four conservative fields of  $\vec{A}$  are investigated. These fields are constructed from a given solution  $\gamma$  as  $\vec{A} = \nabla\gamma$ . They are chosen such that  $|\vec{A}| > 0$  everywhere. The first two fields (5.4) and (5.5) gives isotropic diffusion as one component of the vector field is zero everywhere. These fields are also parallel or perpendicular to the normal  $\vec{n}$  at the boundary. Conversely, the second two fields (5.6) and (5.7) gives anisotropic diffusion and does not enjoy the property of being parallel or perpendicular to  $\vec{n}$  at the boundary.

### 5.3.1 Test cases

The first test case is

$$\gamma(y, z) = \cos(\pi z) - 4z - C \quad \Rightarrow \quad \vec{A}(y, z) = \begin{bmatrix} 0 \\ -\pi \sin(\pi z) - 4 \end{bmatrix}. \quad (5.4)$$

**Remark** The solver will find a solution with zero mean over the domain. Suppose  $\gamma = \hat{\gamma} - C$ . By setting  $C$  as the mean of  $\hat{\gamma}$  over the domain we ensure the real solution specified is the one being produced by the solver.

**Remark** The term  $-4z$  is included to ensure that the vector  $\vec{A}$  has a non-zero magnitude everywhere,  $|\vec{A}| > 0$ .

Figure 5.3 illustrates the vector field (5.4) with principal diffusion directions of the diffusion tensor  $D$  at  $\mu = 0.5$ . As  $\mu \rightarrow 0$ , the strength of the principal diffusion direction will grow. For  $\mu = 1$  diffusion is uniform. To see this, compare Figure 5.3 and Figure 5.4. In the latter figure, the vector field (5.4) is shown together with diffusion directions at  $\mu = 1$ .

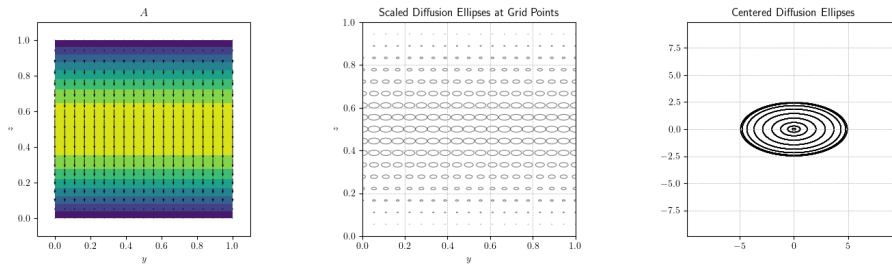


Figure 5.3: Vector field (5.4) and principal diffusion directions of the diffusion tensor  $D$  at  $\mu = 0.5$ .

The second test case is

$$\gamma(y, z) = \sin(\pi z) - 4z - C \quad \Rightarrow \quad \vec{A}(y, z) = \begin{bmatrix} 0 \\ \pi \cos(\pi z) - 4 \end{bmatrix} \quad (5.5)$$

where  $C$  is set as the mean of  $\sin(\pi z) - 4z$ .

Figure 5.5 illustrates the vector field (5.5) with principal diffusion directions of the diffusion tensor  $D$  at  $\mu = 0.5$ .

Test cases (5.4) and (5.5) have the same principal diffusion direction over the

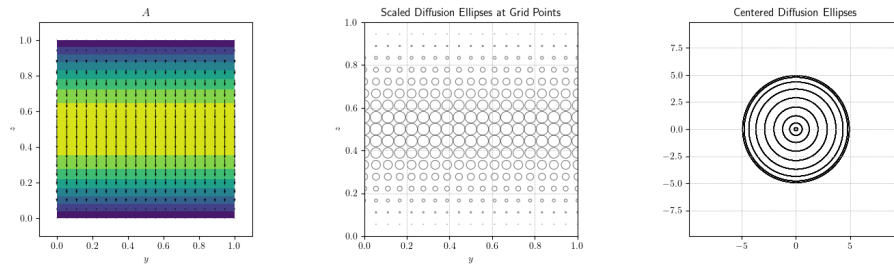


Figure 5.4: Vector field (5.4) and principal diffusion directions of the diffusion tensor  $D$  at  $\mu = 1$ .

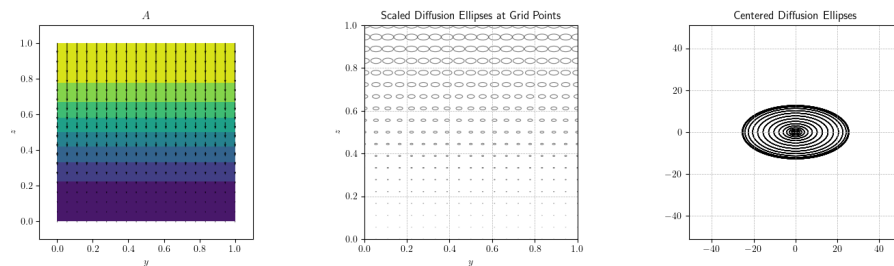


Figure 5.5: Vector field (5.5) and principal diffusion directions of the diffusion tensor  $D$  at  $\mu = 0.5$ .

entire domain. Moreover, the vector field (5.4) is zero in the direction of the outer unit normal  $\vec{n}$  over the boundary  $\partial\Omega$  while (5.5) is not zero at the top boundary in a direction parallel to  $\vec{n}$ .

The third test case is

$$\begin{aligned} \gamma(y, z) &= \cos(\pi(y + 1))\cos(\pi(y + 1)) - 4z - C \\ \Rightarrow \vec{A}(y, z) &= \begin{bmatrix} -\pi\sin(\pi(y + 1))\cos(\pi(y + 1)) \\ -\pi\sin(\pi(y + 1))\cos(\pi(y + 1)) - 4 \end{bmatrix} \end{aligned} \quad (5.6)$$

where  $C$  is set as the mean of  $\cos(\pi y)\cos(\pi z) - 4z$ .

Figure 5.6 illustrates the vector field (5.6) with principal diffusion directions of the diffusion tensor  $D$  at  $\mu = 0.5$ .

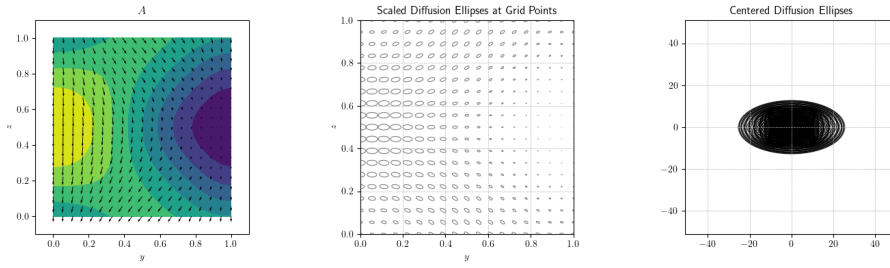


Figure 5.6: Vector field (5.6) and principal diffusion directions of the diffusion tensor  $D$  at  $\mu = 0.5$ .

The fourth conservative test case is

$$\begin{aligned} \gamma(y, z) &= \sin(\pi(y + 1))\sin(\pi(z + 1)) - 6z - C \\ \Rightarrow \vec{A} &= \begin{bmatrix} \pi\cos(\pi(y + 1))\sin(\pi(z + 1)) \\ \pi\cos(\pi(z + 1))\sin(\pi(y + 1)) - 6 \end{bmatrix} \end{aligned} \quad (5.7)$$

where  $C$  is set as the mean of  $\sin(\pi y)\sin(\pi z) - 6z$ .

Figure 5.7 illustrates the vector field (5.7) with principal diffusion directions of the diffusion tensor  $D$  at  $\mu = 0.5$ .

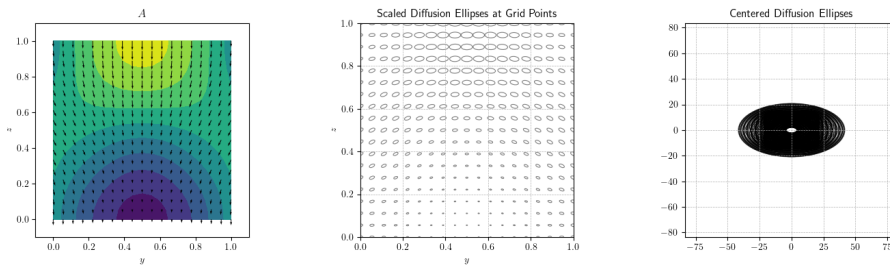


Figure 5.7: Vector field (5.7) and principal diffusion directions of the diffusion tensor  $D$  at  $\mu = 0.5$ .

Test cases (5.6) and (5.7) results in diffusion directions that are not the same in each position for  $\mu \neq 1$ . Moreover, vector field (5.6) is zero in the direction parallel to the outer unit normal  $\vec{n}$  over the left and right boundary but not at the top and bottom. Conversely, vector field (5.7) is non-zero over the top and bottom boundary in directions parallel to  $\vec{n}$ .

### 5.3.2 Convergence

Empirical convergence is found using the analytic solutions as described in Section 5.2. The grid is refined five times starting at  $i = 0$  such that  $\Delta y^{(0)} = \frac{1}{n_y - 1}$  and  $\Delta z^{(0)} = \frac{1}{n_z - 1}$  where  $n_y = n_z = 10$ . Because  $n_y = n_z$  a general step size for both directions can be denoted  $h_0 = \Delta y^{(0)} = \Delta z^{(0)}$ . This step size decreases in each refinement  $i$  such that  $h_i = h_0 2^{-i}$ . Figure 5.8 illustrates second order convergence for the conservative vector fields (5.4), (5.4), (5.6) and (5.7) for both natural and simple boundary conditions at  $\mu = 0.5$ . Figure 5.8 also suggest there is no significant difference between natural and simple boundary conditions however, for vector field (5.6) it is noticeable. These observation can also be confirmed by looking at the convergence rates in Table 5.1.

i	BC	0	1	2	3	4	5
Vector field (5.4)	Natural	2.51	2.29	2.16	2.08	2.04	2.02
	Simple	2.51	2.29	2.16	2.08	2.04	2.02
Vector field (5.5)	Natural	2.15	2.06	2.03	2.02	2.01	2.00
	Simple	2.15	2.06	2.03	2.02	2.01	2.00
Vector field( 5.6)	Natural	2.01	2.00	2.00	2.00	2.00	2.00
	Simple	1.92	1.96	1.98	1.99	2.00	2.00
Vector field (5.7)	Natural	2.27	2.15	2.08	2.04	2.02	2.01
	Simple	2.24	2.13	2.07	2.04	2.02	2.01

Table 5.1: Empirical convergence rates for conservative fields (5.4), (5.5), (5.6) and (5.7) at  $\mu = 0.5$  where  $h_i = h_0 2^{-i}$ .

Figure 5.9 shows  $\vec{A}$  and solution gradient  $\nabla\gamma$  for test cases (5.4), (5.4), (5.6) and (5.7) as well as the norm of the difference between  $\vec{A}$  and  $\nabla\gamma$  which quantifies both angular and magnitude deviations between the two. The difference  $|\vec{A} - \nabla\gamma|$  is around  $10^{-2}$  for all test cases. This difference is noteworthy however, as indicated by the convergence rates the accuracy can be improved by increasing resolution. This can be seen by comparing the measured accuracy in the right most plots of subfigures in Figure 5.9 and 5.10.

**Remark** There is no noteworthy difference between the solutions produced by the natural Neumann boundary condition and the simplified boundary condition. We therefore refrain from illustrating the solutions given when using simple boundary conditions in the same way as Figure 5.9 and 5.10 where natural Neumann boundary condition were used.

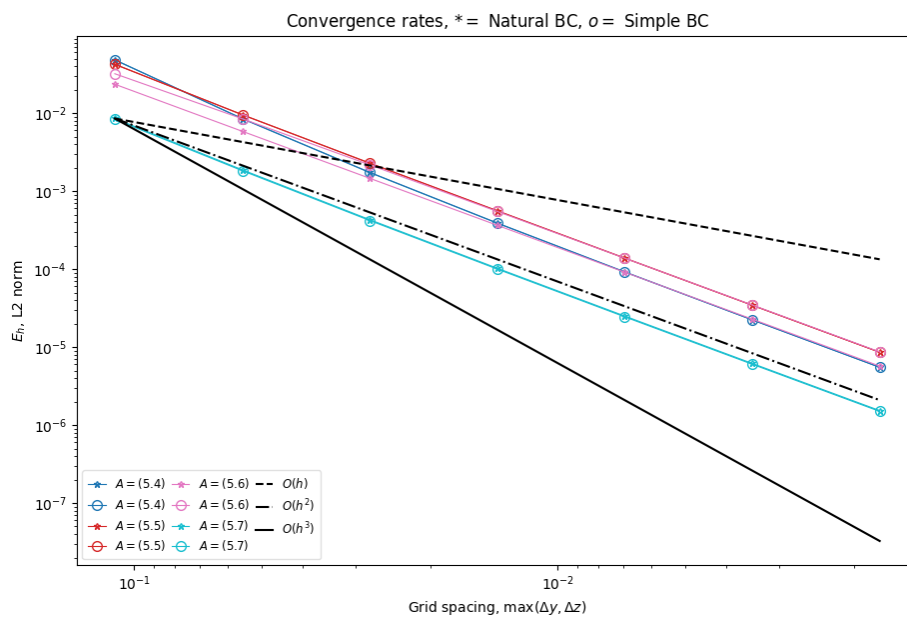
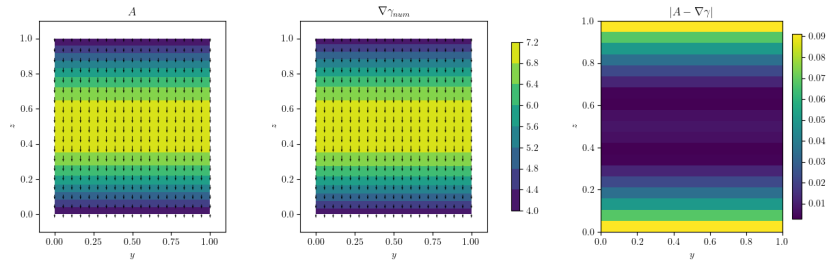
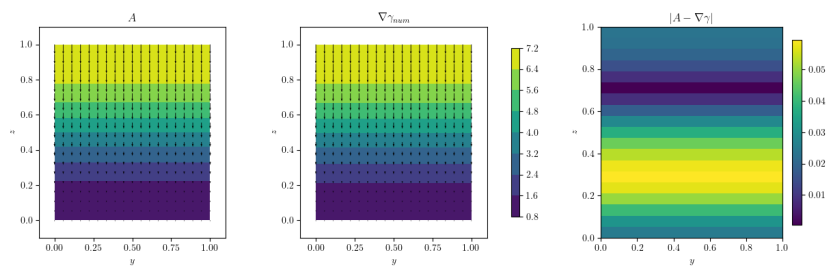


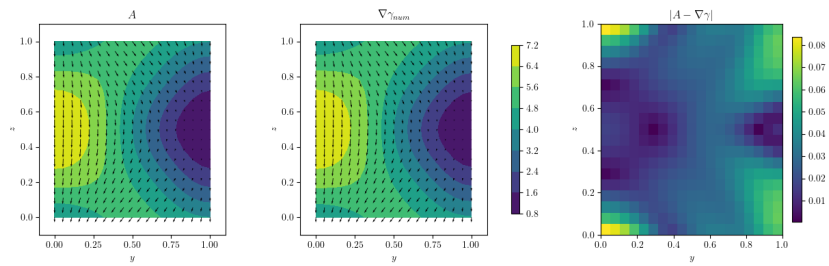
Figure 5.8: Empirical convergence of conservative fields at  $\mu = 0.5$ .



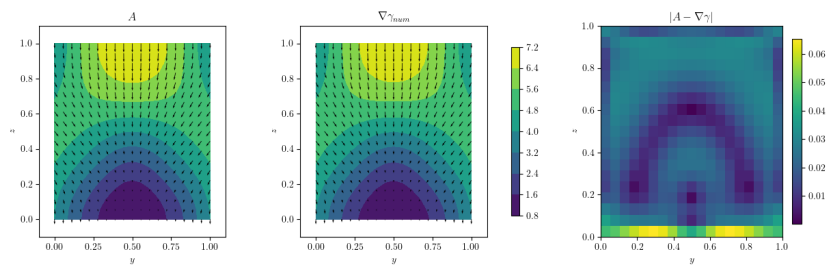
(a) Vector field (5.4).



(b) Vector field (5.5).

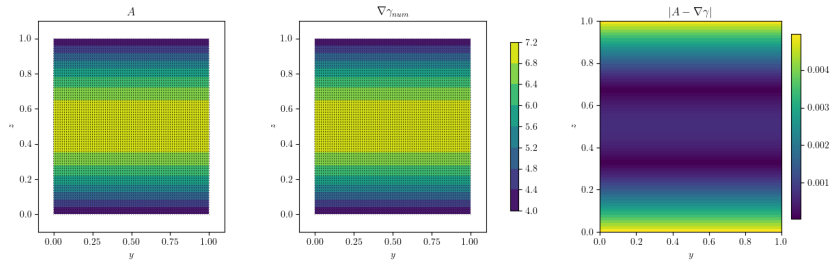


(c) Vector field (5.6).

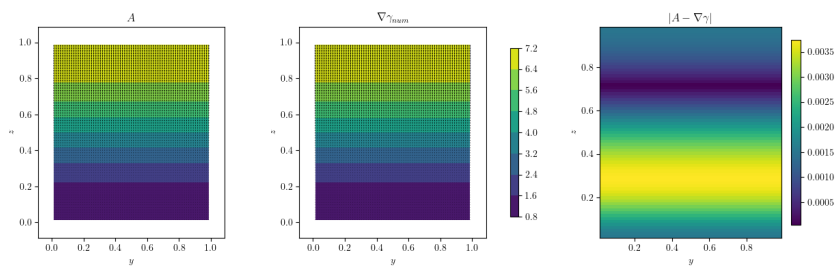


(d) Vector field (5.7).

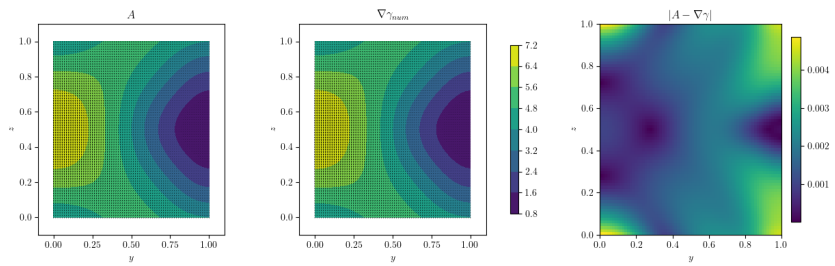
Figure 5.9: Vector field  $\vec{A}$  and solution gradient  $\nabla\gamma$  using natural boundary conditions with refined step size  $h_1 = h_0 2^{-1}$  at  $\mu = 0.5$ .



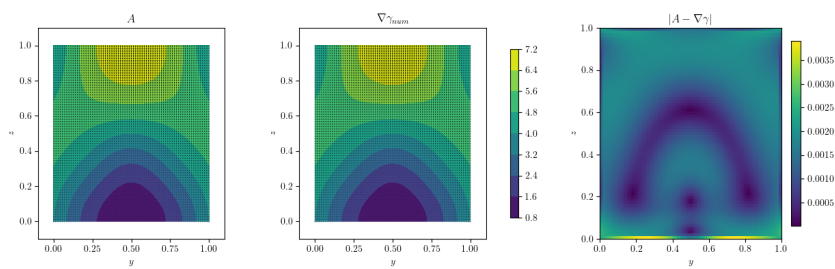
(a) Vector field (5.4).



(b) Vector field (5.5).



(c) Vector field (5.6).



(d) Vector field (5.7).

Figure 5.10: Vector field  $\vec{A}$  and solution gradient  $\nabla\gamma$  using natural boundary conditions with refined step size  $h_3 = h_0 2^{-3}$  at  $\mu = 0.5$ .

### 5.3.3 Solution's dependence on $\mu$

Figure 5.11 and 5.12 illustrates solution's dependence on  $\mu \in (0, 1]$  by looking at the angular deviation measured by the root mean square of  $|\vec{A} \times \nabla\gamma|$  as well as a root mean square measure of the difference between  $\vec{A}$  and solution gradient  $\nabla\gamma$ . This is done for four refinements of the grid to analyse convergence behaviour. For each refinement we have  $h = \Delta y = \Delta z$  where  $h$  is based on the initial resolution  $h_0$  as described and given above.

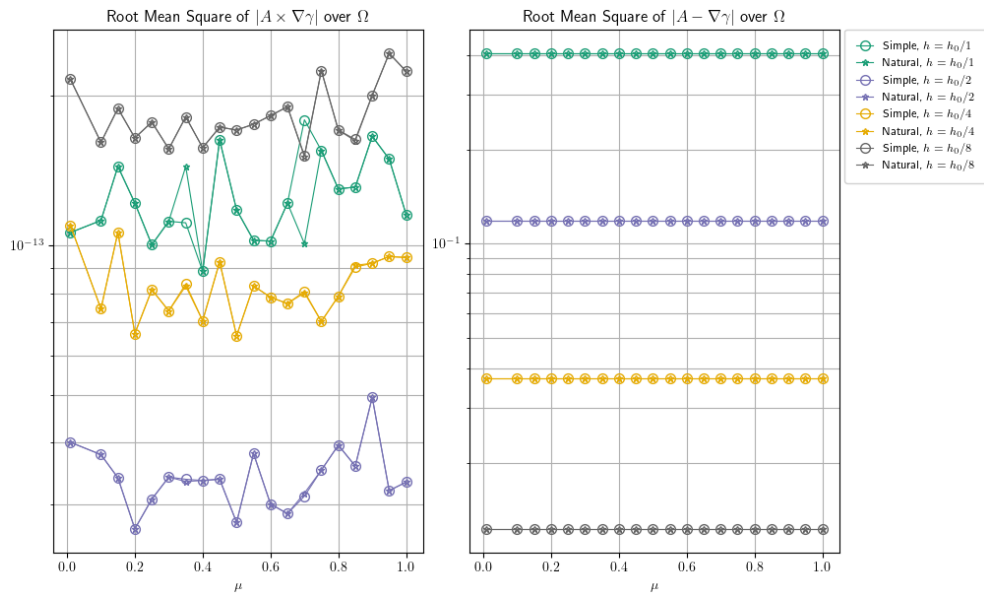
Figure 5.11 includes root mean square measures for vector fields (5.4) and (5.5). The left most subplots of Figure 5.11 shows that the angular deviation is almost zero for all  $\mu$ . Moreover, the right most subplots of Figure 5.11 shows that the difference measure for the same test fields stays constant with decreasing  $\mu$ . This result suggests an independence of the solution to  $\mu$  which is consistent with the theoretical result given in Section 3.3. In Figure 5.11 we also observe appropriate convergence behaviour for test fields (5.4) and (5.5) as resolution increases.

Figure 5.12 illustrate's the root mean square measures of angular and magnitude deviations for vector fields (5.6) and (5.7). For both of these test fields we observe a slight decrease in  $|\vec{A} - \nabla\gamma|$  as  $\mu \rightarrow 0.5$ . Conversely, test field (5.7) results in a gradual increase in  $|\vec{A} - \nabla\gamma|$  as  $\mu < 0.5 \rightarrow 0$ . Such an increase in the difference measure is present for field (5.6) as well but appears when  $\mu$  is closer to 0. Since change is gradual this is an indication of some numerical artifact as  $\mu \rightarrow 0$ . Both vector field (5.6) and (5.7) are initially constant in angular deviation however, as  $\mu \rightarrow 0$  the test field (5.7) results in increased angular deviation. This is especially noticeable when the simplified boundary conditions were used but we do observe convergence towards zero as  $\Delta y$  and  $\Delta z \rightarrow 0$ .

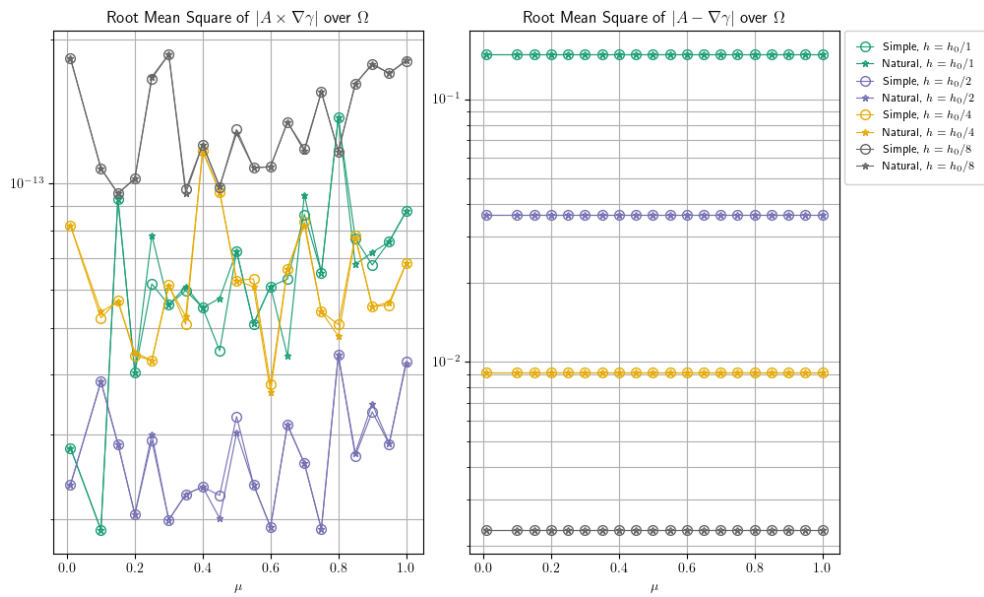
To summarize, Figure 5.12 shows that the method becomes less accurate as  $\mu \rightarrow 0$  for test field (5.7) especially. This is inconsistent with the theoretical result given in Section 3.3. Two observations can be contributing factors to this behaviour. Firstly, when  $\mu$  is zero, wellposedness is not guaranteed and while Figure 5.12 shows stable convergence behaviour for all  $\mu$  we are not guaranteed to converge to a solution such that  $\nabla\gamma = \vec{A}$  when  $\mu \rightarrow 0$ . Consequently, even though we are sure that  $\gamma$  such that  $\nabla\gamma = \vec{A}$  exists since the field is conservative we might not be finding that solution for  $\mu$  closer to 0. If we are not finding that solution it is expected that the root mean square measures grow. Secondly, as

$\mu \rightarrow 0$  the condition number of the diffusion tensor will increase. This causes an increased condition number of the matrix of the linear system of equations which makes it harder to solve. Finally, the independence of the solution with respect to  $\mu$  should hold when  $\Delta y, \Delta z \rightarrow 0$ . While the discussion here concerns finite  $\Delta y, \Delta z > 0$ , this last statement is confirmed by looking at Figure 5.11 and 5.12 and noting that the root mean square measures decrease gradually as the grid is refined.

In the last remark of Section 3.1.3 we concluded that if the vector field is perpendicular or parallel to the normal at the boundary the simple boundary condition is equivalent to the natural Neumann boundary condition. The root means square measures of Figure 5.11 and 5.12 as well as the convergence rates of Table 5.1 are confirmations of this. For test fields (5.4) and (5.5) the two boundary conditions are equivalent and the rates and root mean square measures are the same. The opposite is true for test fields (5.6) and (5.7) where the two boundary conditions are not necessarily equivalent and measures differ slightly.

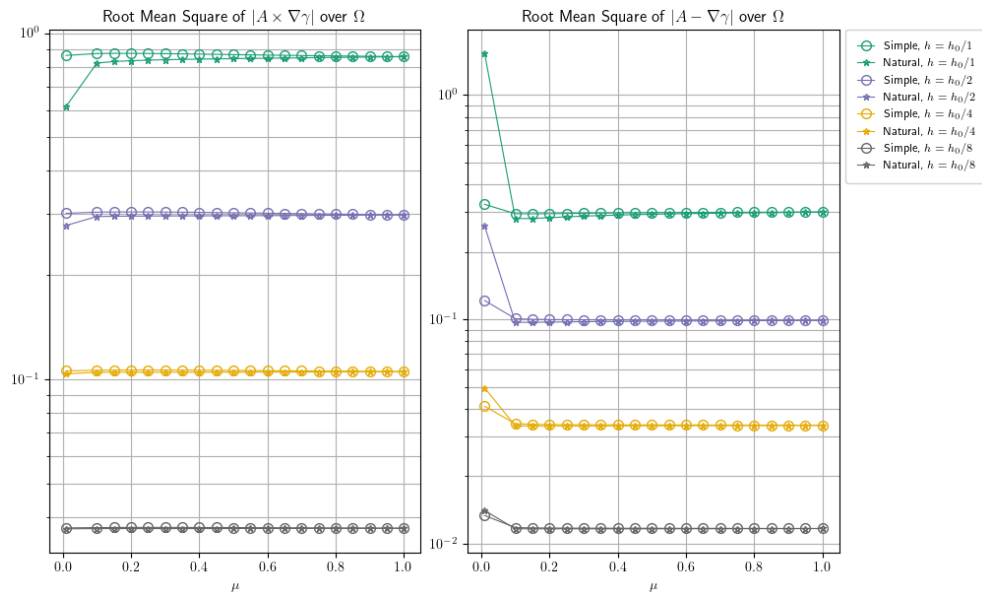


(a) Vector field (5.4).

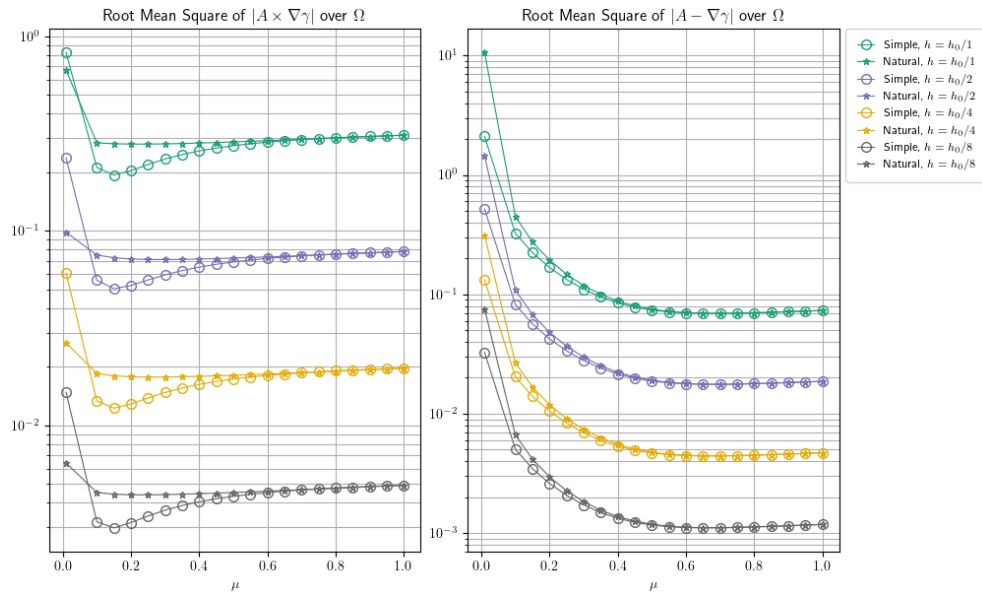


(b) Vector field (5.5).

Figure 5.11: Angular and magnitude deviations between conservative vector field  $\vec{A}$  and solution gradient  $\nabla \gamma$  as a function of  $\mu$ .



(a) Vector field (5.6).



(b) Vector field (5.7).

Figure 5.12: Angular and magnitude deviations between conservative vector field  $\vec{A}$  and solution gradient  $\nabla \gamma$  as a function of  $\mu$ .

## 5.4 Non-conservative fields

In this section a non-conservative field of  $\vec{A}$  is investigated. This field is built such that  $|\vec{A}| > 0$  everywhere. Moreover, diffusion is anisotropic and the field is parallel or perpendicular to the normal  $\vec{n}$  at the boundary.

### 5.4.1 Test cases

A test case  $\vec{A}$  is constructed such that its rotation can be tuned by a parameter  $\tau$ . Using one conservative component and one non-conservative component given as

$$\vec{A}_{con} = \text{Vector field(5.5)}, \quad \vec{A}_{non} = \begin{bmatrix} \cos(\pi z) \\ 50y(1-y)z(1-z) \end{bmatrix}$$

we arrive at an expression for the non-conservative field

$$\vec{A} = \tau \vec{A}_{con} + (1 - \tau) \vec{A}_{non} - \begin{bmatrix} 0 \\ 4 \end{bmatrix}. \quad (5.8)$$

To understand the relationship between rotation and  $\tau$  consider the rotation of this vector field using the expression derived in Section 2.3.3. We get,

$$\begin{aligned} \vec{A} &= \begin{bmatrix} (1 - \tau) \cos(\pi z) \\ \tau \pi \cos(\pi z) - (1 - \tau) 50y(1 - y)z(1 - z) - 4 \end{bmatrix} \\ \Rightarrow \nabla \times \vec{A} &= -(1 - \tau) \pi \sin(\pi z) + 50(1 - \tau)z(1 - z) - 100(1 - \tau)yz(1 - z) \\ &= (1 - \tau) [-\pi \sin(\pi z) + (50 - 100y)z(1 - z)]. \end{aligned}$$

For  $\tau \rightarrow 0$  the rotation increases and for  $\tau = 1$  the rotation is zero.

This field is tested for  $\tau = 0.4$ . Figure 5.13 illustrates the vector field (5.8) at  $\tau = 0.4$  with principal diffusion directions of the diffusion tensor  $D$  at  $\mu = 0.5, 0.1$  and  $0.01$ .

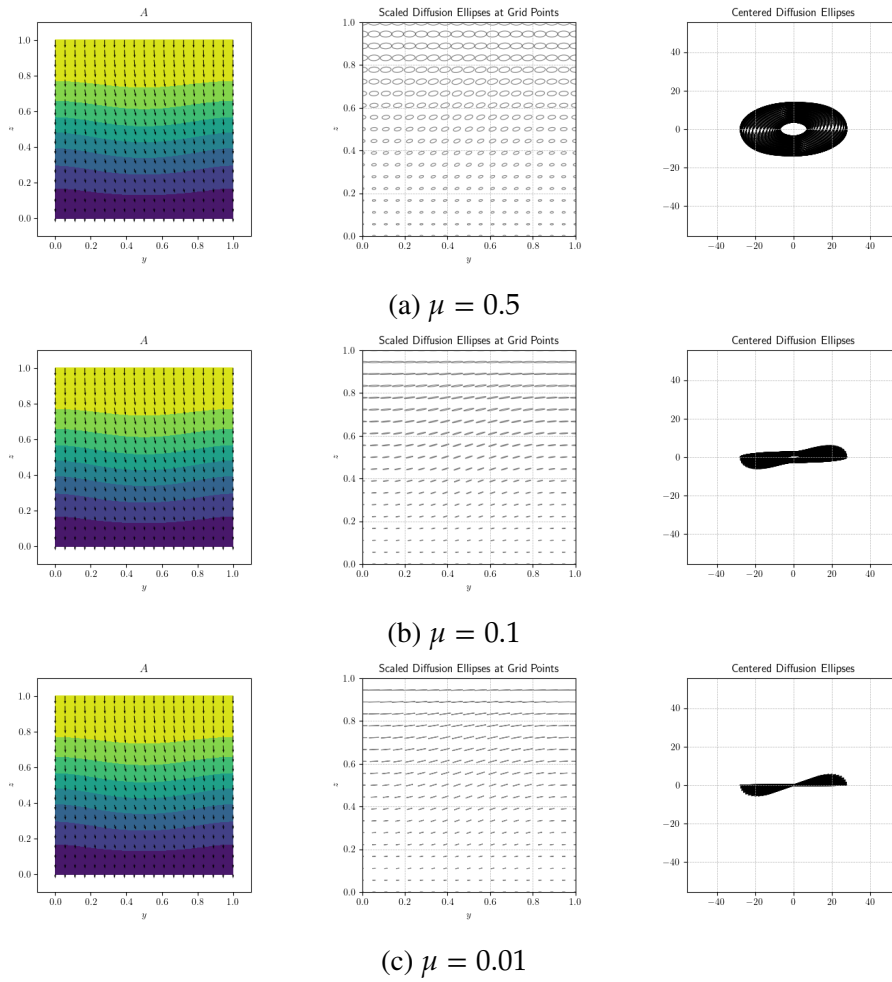


Figure 5.13: Vector field (5.8) and principal diffusion directions of the diffusion tensor  $D$  at  $\mu = 0.5, 0.1$  and  $0.01$  for the non-conservative field with  $\tau = 0.4$ .

## 5.4.2 Convergence

Since an analytic solution is not available such that  $\nabla\gamma = \vec{A}$  for the non-conservative test case empirical convergence is found using successive numerical solutions as described in Section 5.2. As was done for the conservative test cases, the grid is refined five times starting at  $i = 0$  such that  $\Delta y^{(0)} = \frac{1}{n_y - 1}$  and  $\Delta z^{(0)} = \frac{1}{n_z - 1}$  where  $n_y = n_z = 10$ . Because  $n_y = n_z$  a general step size for both directions can be denoted  $h_0 = \Delta y^{(0)} = \Delta z^{(0)}$ . This step size decreases in each refinement  $i$  such that  $h_i = h_0 2^{-i}$ . Figure

5.14 illustrates second order convergence for the non-conservative vector field (5.8) with  $\tau = 0.4$  for both natural and simple boundary conditions at  $\mu = 0.5, 0.1, 0.01$ . The observation of second order convergence can also be confirmed by looking at the convergence rates of Table 5.2.

i	BC	0	1	2	3	4
$\mu = 0.5$	Natural	2.18	2.10	2.05	2.03	2.01
	Simple	2.18	2.10	2.05	2.03	2.01
$\mu = 0.1$	Natural	2.22	2.13	2.05	2.02	2.01
	Simple	2.22	2.13	2.05	2.02	2.01
$\mu = 0.01$	Natural	2.96	2.39	2.12	2.01	1.99
	Simple	2.96	2.39	2.12	2.01	1.99

Table 5.2: Empirical convergence rates for non-conservative field (5.8) with  $\tau = 0.4$  at  $\mu = 0.5, 0.1$  and  $0.01$  where  $h_i = h_0 2^{-i}$ .

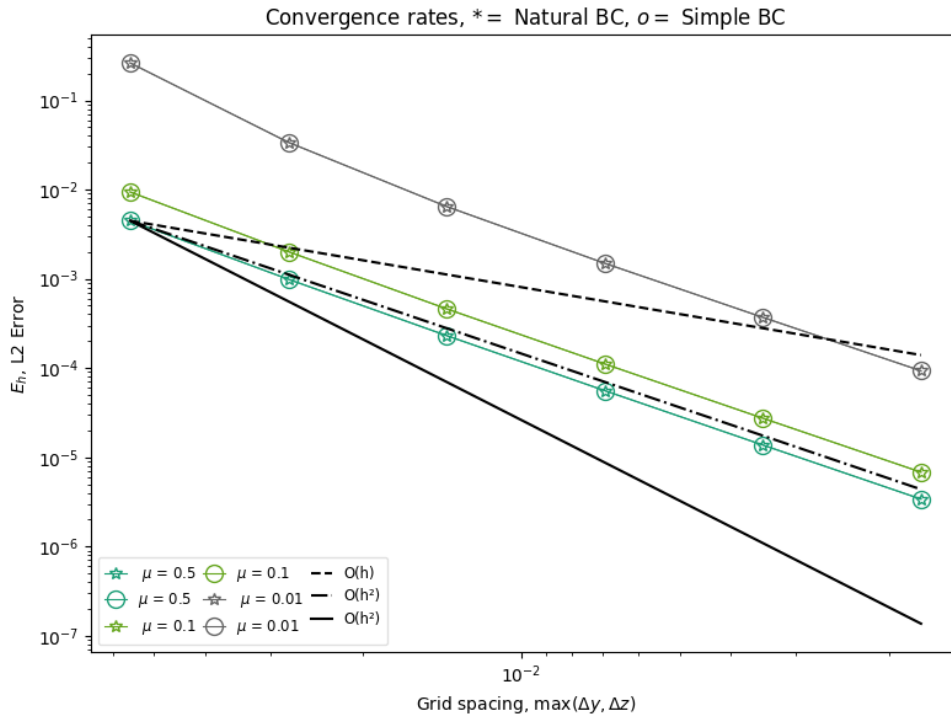


Figure 5.14: Empirical convergence of the non-conservative field (5.8) with  $\tau = 0.4$  at  $\mu = 0.5, 0.1$  and  $0.01$ .

### 5.4.3 Solution's dependence on $\mu$

The root mean square measure of  $|A \times \nabla\gamma|$  and  $|A - \nabla\gamma|$  are presented in Figure 5.15 for  $\mu \in (0, 1]$ . This is done for four refinements of the grid to analyse convergence behaviour. For each refinement we have  $h = \Delta y = \Delta z$  where  $h$  is based on the initial resolution  $h_0$  which is given above. We can see that the solution is changing with  $\mu$  significantly and the change persists when  $\Delta y, \Delta z \rightarrow 0$ .

As  $\mu$  goes towards zero the root mean square of  $|A - \nabla\gamma|$  grows. This is expected since we are not penalizing deviations from magnitude equality between the solution gradient and  $\vec{A}$  as much for smaller  $\mu$ . Note that since the field is non-conservative we do not want to enforce such magnitude equality. Instead we have that the solution  $\gamma$  and some  $b$  exists such that  $\nabla\gamma = b\vec{A}$ . Consequently, the increase as  $\mu \rightarrow 0$  shown in the right subplot of Figure 5.15 is both anticipated and favourable. As  $\mu$  goes towards zero the root mean square of  $|A \times \nabla\gamma|$  decreases as seen in the left subplot of Figure 5.15. This is reasonable since collinearity should be satisfied when  $\gamma$  is such that  $\nabla\gamma = b\vec{A}$  for some unknown  $b$ .

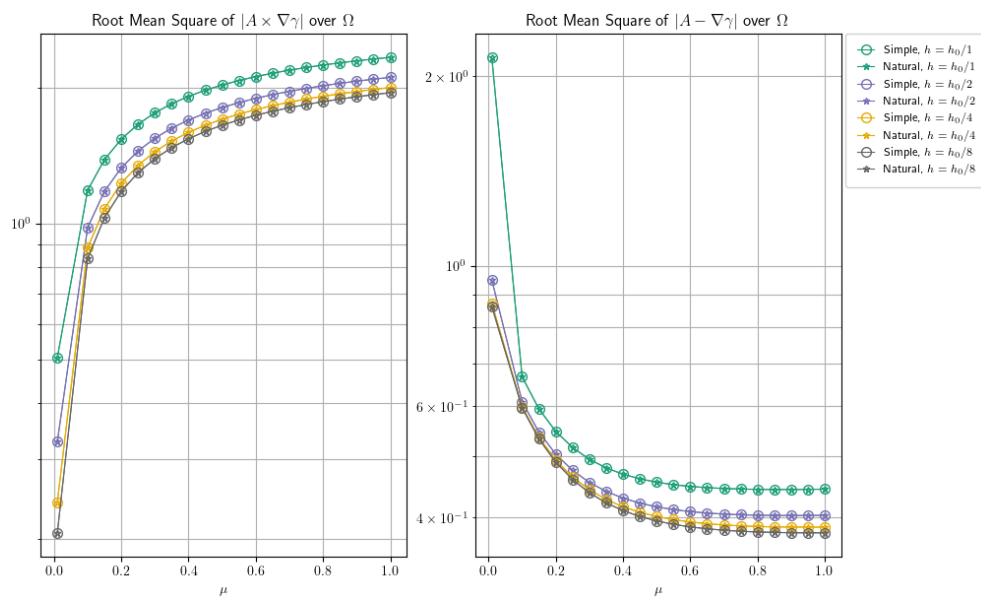


Figure 5.15: Angular and magnitude deviations between the non-conservative vector field (5.8) with  $\tau = 0.4$  and solution gradient  $\nabla\gamma$  as a function of  $\mu$ .

The behaviour seen in Figure 5.15 and discussed in the latest paragraph is illustrated further by Figure 5.16, 5.17 and 5.18. These figures include vector field  $\vec{A}$ , solution gradient, angular deviation measure  $|\vec{A} \times \nabla\gamma|$  and the factor  $b$  such that  $\nabla\gamma = b\vec{A}$  evaluated through

$$b = \frac{\vec{A}^T \nabla\gamma}{\vec{A}^T \vec{A}}. \quad (5.9)$$

By comparing the vector fields in each figure and by looking at  $|\vec{A} \times \nabla\gamma|$  in the bottom left subplot of each figure we see that  $|\vec{A} \times \nabla\gamma|$  decreases with  $\mu$ . In the bottom right subplots of Figure 5.16, 5.17 and 5.18 we look at measures of  $b$  such that  $\nabla\gamma = b\vec{A}$ . We can see that  $b$  deviates from 1 by about  $\pm 0.1$  for  $\mu = 0.5$ , about  $\pm 0.3$  for  $\mu = 0.1$  and about  $-0.9$  for  $\mu = 0.01$ . In conclusion, for smaller  $\mu$  magnitude equality between  $\vec{A}$  and  $\nabla\gamma$  is not enforced as much, which is expected, resulting in a  $b$  that deviates from 1 to a greater extent.

Returning briefly to Figure 5.15, we find that for  $\mu$  close to zero there is no apparent convergence behaviour towards a solution  $\gamma$  such that  $\nabla\gamma = \vec{A}$  as resolution is increased. The measure seems to stabilize around 0.4. Because the field is non-conservative a solution such that  $\nabla\gamma = \vec{A}$  does not exist and as resolution increases we are moving towards some solution but not one that enjoys magnitude equality between the solution gradient and  $\vec{A}$ . For  $\mu$  closer to zero, the root mean square of  $|\vec{A} - \nabla\gamma|$  stabilizes similar to when  $\mu$  was close to one however, the angular deviation measure shows indications of stable convergence as resolution increases.

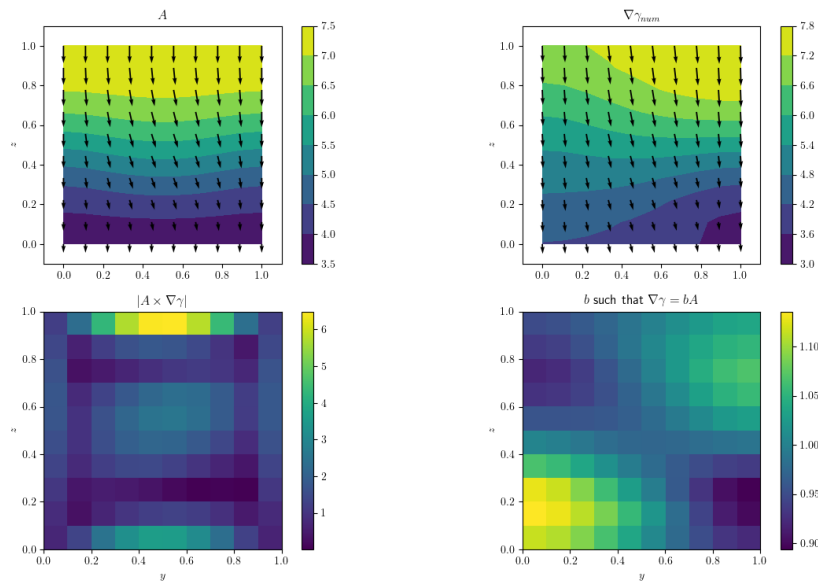


Figure 5.16: Vector field (5.8) with  $\tau = 0.4$  and solution gradient  $\nabla\gamma$  using natural boundary conditions with step size  $h_0$  at  $\mu = 0.5$ .

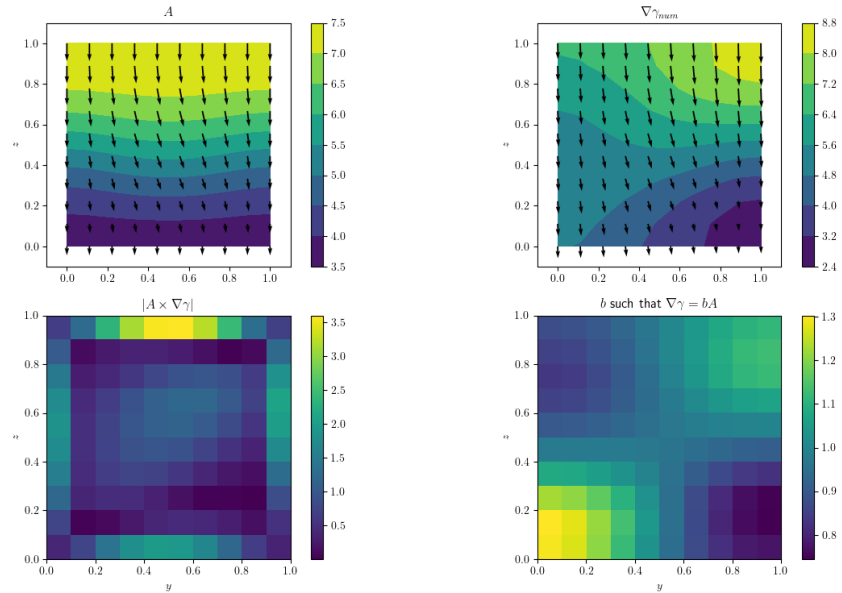


Figure 5.17: Vector field (5.8) with  $\tau = 0.4$  and solution gradient  $\nabla\gamma$  using natural boundary conditions with step size  $h_0$  at  $\mu = 0.1$ .

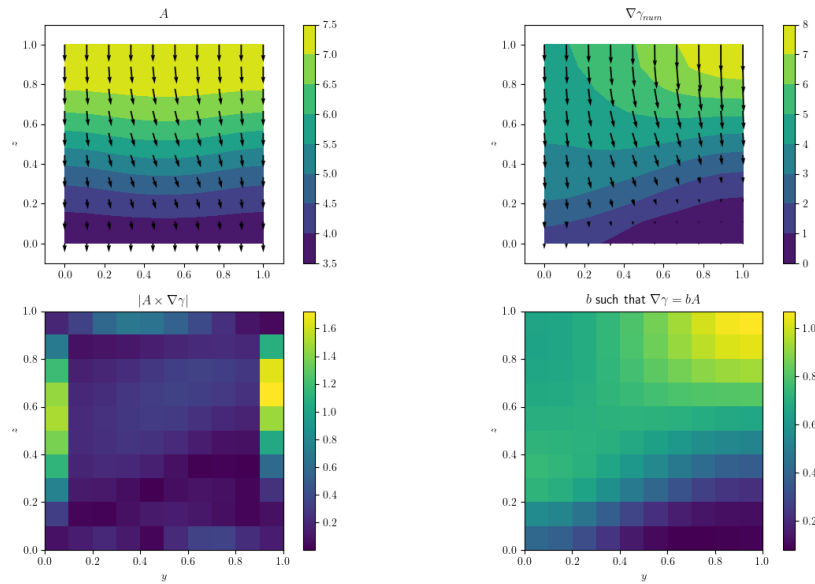


Figure 5.18: Vector field (5.8) with  $\tau = 0.4$  and solution gradient  $\nabla\gamma$  using natural boundary conditions with step size  $h_0$  at  $\mu = 0.01$ .

## 5.5 Real diapycnal vector field

The real diapycnal vector field  $\vec{A}$  is defined through absolute salinity, conservative temperature, density as well as the haline contraction coefficient and thermal expansion coefficient, see equation (2.6). Figure 5.19 illustrates the vector field (2.6) with principal diffusion directions of the diffusion tensor  $D$  at  $\mu = 0.5, 0.1$  and  $0.01$ .

### 5.5.1 Convergence

As was the case for the non-conservative test field (5.8), an analytic solution is not available such that  $\nabla\gamma = \vec{A}$ . Empirical convergence is therefore found using successive numerical solutions as described in Section 5.2. The grid is refined five times by interpolation starting at  $i = 0$  where that  $\Delta y^{(0)} = 111111$  and  $\Delta z^{(0)} = 110.43$ . This step size decreases in each refinement  $i$  such that  $h_i = h_0 2^{-i}$  in both directions:  $h_i = \Delta y^{(i)}$  and  $h_i = \Delta z^{(i)}$ . Figure 5.20 illustrates second order convergence for the real diapycnal vector field (2.6) for both natural and simple boundary conditions at  $\mu = 0.5, 0.1$  and  $0.01$ . The observation of second order convergence can also be confirmed by looking at the convergence rates of Table 5.1.

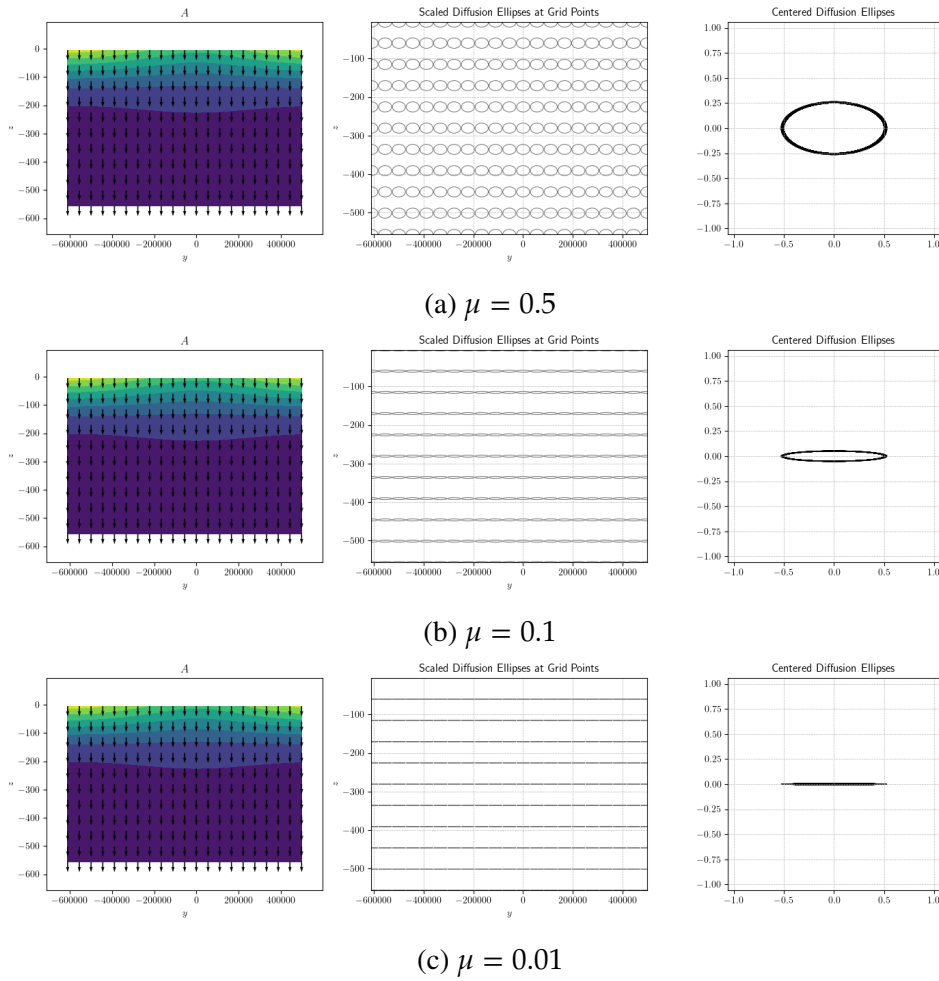


Figure 5.19: The real diapycnal vector field (2.6) and principal diffusion directions of the diffusion tensor  $D$  with refined step sizes  $\Delta y^{(1)} = \Delta y^{(0)} 2^{-1}$  and  $\Delta z^{(1)} = \Delta z^{(0)} 2^{-1}$ .

i	BC	0	1	2	3	4
$\mu = 0.5$	Natural	1.29	1.82	1.90	1.95	1.98
	Simple	1.29	1.82	1.90	1.95	1.98
$\mu = 0.1$	Natural	1.29	1.82	1.90	1.95	1.98
	Simple	1.29	1.82	1.90	1.95	1.98
$\mu = 0.01$	Natural	1.29	1.82	1.90	1.95	1.98
	Simple	1.29	1.82	1.90	1.95	1.98

Table 5.3: Empirical convergence rates for the real diapycnal vector field (2.6) at  $\mu = 0.5, 0.1$  and  $0.01$  where  $\Delta y^{(i)} = \Delta y^{(0)} 2^{-i}$  and  $\Delta z^{(i)} = \Delta z^{(0)} 2^{-i}$ .

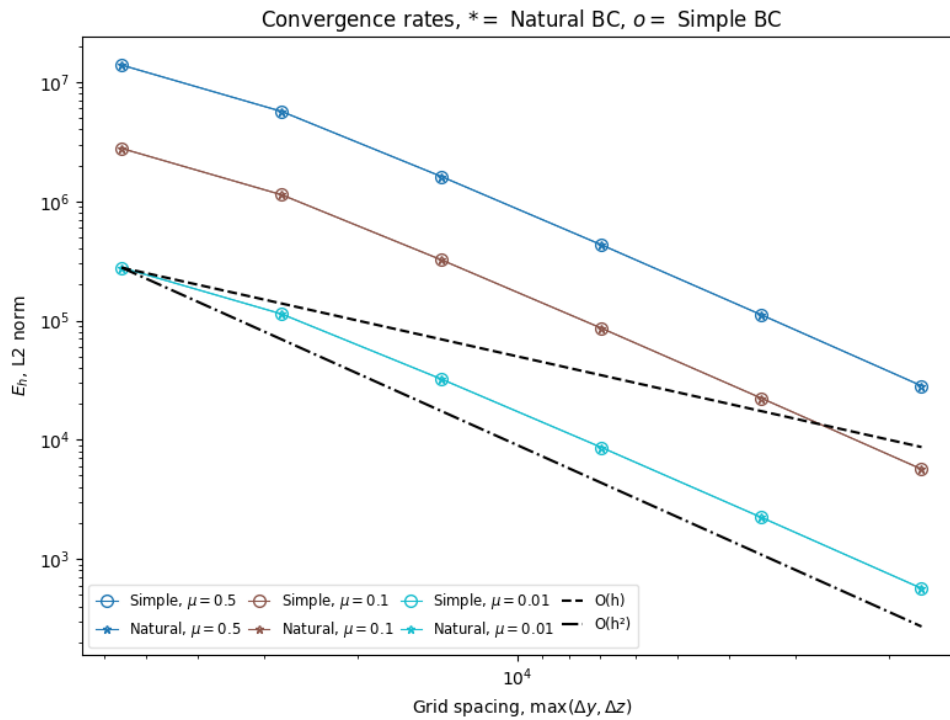


Figure 5.20: Empirical convergence of the real diapycnal vector field (2.6) at  $\mu = 0.5, 0.1$  and  $0.01$ .

In Figure 5.21, the solution gradient as well as the accuracy measure  $|A - \nabla\gamma|$  can be found. We can see that the difference  $|A - \nabla\gamma|$  is around  $10^{-3} - 10^{-4}$  for all test cases. This difference is noteworthy however, as indicated by the convergence rates the accuracy can be improved by increasing resolution. This can be seen by comparing the accuracy measure  $|A - \nabla\gamma|$  between Figure 5.21 and 5.22.

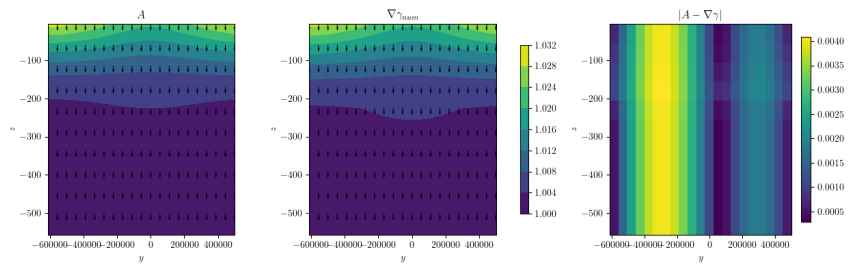
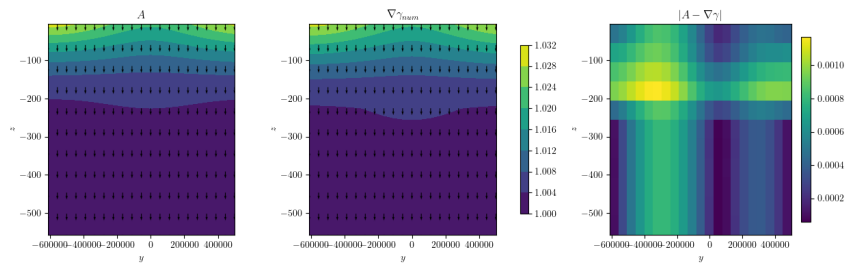
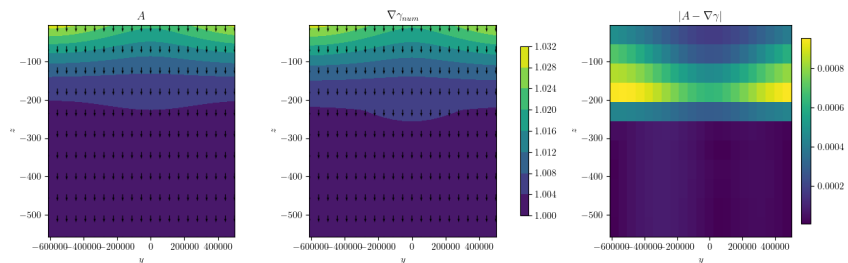
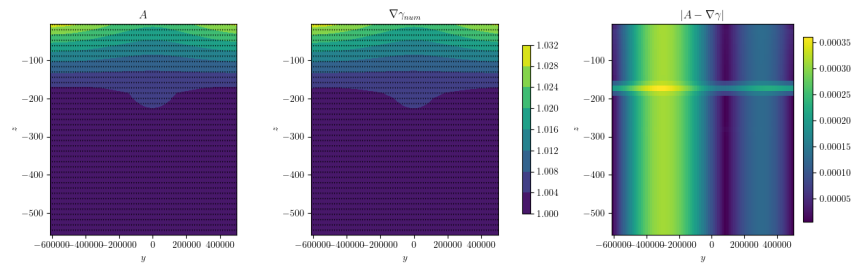
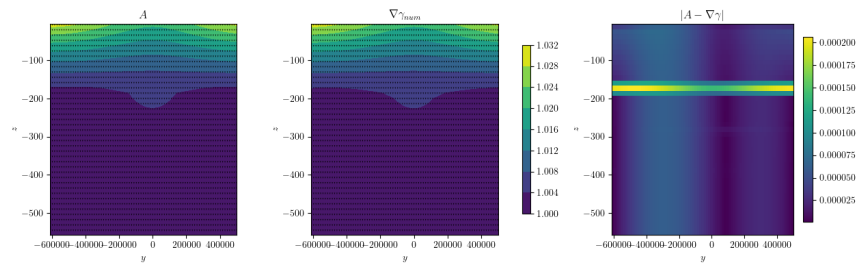
(a)  $\mu = 0.5$ (b)  $\mu = 0.1$ (c)  $\mu = 0.01$ 

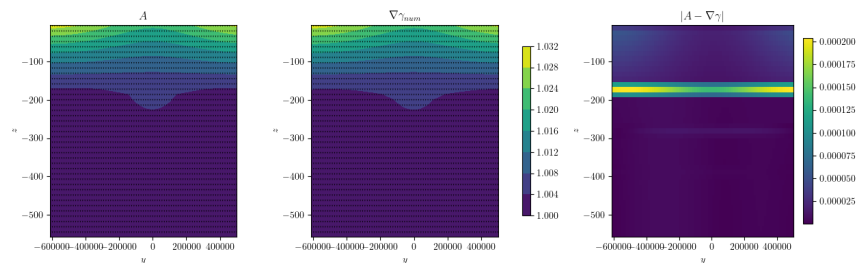
Figure 5.21: Vector field  $\vec{A}$  and solution gradient  $\nabla\gamma$  using natural boundary conditions with refined step sizes  $\Delta y^{(1)} = \Delta y^{(0)}2^{-1}$  and  $\Delta z^{(1)} = \Delta z^{(0)}2^{-1}$ .



(a)  $\mu = 0.5$



(b)  $\mu = 0.1$



(c)  $\mu = 0.01$

Figure 5.22: Vector field  $\vec{A}$  and solution gradient  $\nabla\gamma$  using natural boundary conditions with refined step sizes  $\Delta y^{(3)} = \Delta y^{(0)}2^{-3}$  and  $\Delta z^{(3)} = \Delta z^{(0)}2^{-3}$ .

## 5.5.2 Solutions dependence on $\mu$

Figure 5.23 shows the root mean square measure of  $|\vec{A} \times \nabla\gamma|$  and  $|\vec{A} - \nabla\gamma|$  as a function of  $\mu$  for successively finer grids. The grid is refined in both directions by a factor of two. Therefore,  $h$  of Figure 5.23 represents changes in both  $\Delta y$  and  $\Delta z$ . We find that both measures decrease as  $\mu \rightarrow 0$ . In the leftmost subplot of Figure 5.23, we find that  $|\vec{A} \times \nabla\gamma|$  moves towards zero as  $\mu \rightarrow 0$ . However, the rightmost subplot of Figure 5.23 as well as Figure 5.22 indicates that  $|\vec{A} - \nabla\gamma|$  stabilizes as  $\mu \rightarrow 0$ . Both measures move towards zero when the grid is refined;  $\Delta y, \Delta z \rightarrow 0$ . This convergence behaviour indicates that the field is conservative. However, a conclusion regarding the characteristics of the field should not be drawn without further investigation.

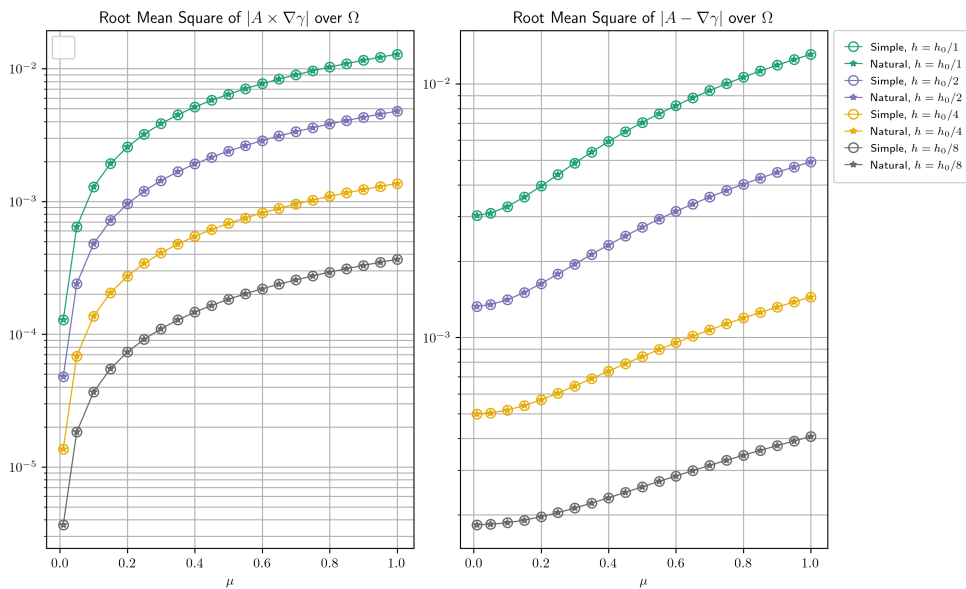


Figure 5.23: Angular and magnitude deviations between the real diapycnal vector field (2.6) and solution gradient  $\nabla\gamma$  as a function of  $\mu$ .

# Chapter 6

## Conclusion

In this chapter, the results of Chapter 5 are first summarized and then limitations of the method are discussed. In this discussion we give suggestions on what would be good to focus on when continuing the development of this method to find an approximate neutral density variable for the world's oceans.

### 6.1 Summary

In this study the neutral density variable is modelled by an optimization problem and then solved for through finite differences of the corresponding Euler-Lagrange equation and boundary conditions.

The results in Chapter 5 begins with a discussion of four conservative fields. The first two fields are grid aligned and hence result in an isotropic diffusion equation. The latter two fields deviate from the grid causing anisotropic diffusion for  $\mu \neq 1$ . When the isotropic diffusion equations are solved the solution gradient have no angular deviation to  $\vec{A}$  for all  $\mu$  and the root mean square of  $|\vec{A} - \nabla\gamma|$  is seemingly independent of  $\mu$ . Conversely, for the latter two conservative fields, where an anisotropic diffusion equation is solved, the root mean square of  $|A - \nabla\gamma|$  is increased as  $\mu \rightarrow 0$ . This is especially prominent for vector field (5.7). While the solution improves as  $\Delta y, \Delta z \rightarrow 0$ , this result might point at some numerical instability as  $\mu \rightarrow 0$  for anisotropic diffusion. Two observations could be possible causes of this behaviour. Firstly, when  $\mu$  is zero, wellposedness is not guaranteed and while we have stable convergence behaviour for all  $\mu$  we are not guaranteed to converge to a solution such that  $\nabla\gamma = \vec{A}$  when  $\mu \rightarrow 0$ . Secondly, as  $\mu \rightarrow 0$  the condition number of the diffusion tensor will increase causing an increased condition number of the

matrix of the linear system of equations which makes it harder to solve.

A non-conservative test case was constructed to investigate how the method would react when a solution  $\gamma$  such that  $\vec{A} = \nabla\gamma$  does not exist. When looking at the root mean square of the angular deviation as well as the difference between  $\vec{A}$  and the solution gradient we found that these measures changes gradually with  $\mu$ . The measure on angular deviation decreases as  $\mu \rightarrow 0$ . This result is in line with the upper bound presented in the second point of Section 3.3. Conversely, the general difference measure, which looks at magnitude and angular deviation, increase with  $\mu \rightarrow 0$  as expected.

In a further effort to investigate the behaviour of the solver we also consider the real diapycnal vector field (2.6). Since the scale of the horizontal direction is much bigger than that of the vertical direction, the vector field  $\vec{A}$  is approximately aligned with the z-direction. When looking at the angular and magnitude deviations between the solution gradient and the diapycnal vector field  $\vec{A}$  we find that the error decreases gradually as  $\mu \rightarrow 0$ . This indicates non-conservativeness of the vector field. However, convergence behaviour as resolution increases is consistent for all  $\mu$  indicating a conservative field. A conclusion regarding the characteristics of the field can not be drawn without further investigation.

In this study, the simple and natural Neumann boundary conditions were used in the numerical experiments. As pointed out in the last remark of Section 3.1.3 these will be equivalent if the vector field  $\vec{A}$  is perpendicular and/or parallel to the normal vector at the boundary. The solutions produced by the method in our numerical experiments are consistent with this theoretical result. Using the simple boundary condition is advantageous as it removes complexity and some numerical instability in implementation. The real diapycnal vector field is, for the domain considered in this study, perpendicular or parallel to the normal vector at the boundaries. Hence, when this method is applied to physical oceanography it might be sufficient to reduce the natural Neumann boundary condition to its simpler form and use that in implementation.

While we have presented here a proof-of-concept, there are several ways the method needs to be developed and improved to make it applicable to real physical oceanography cases.

## 6.2 Limitations and future work

We now give suggestions on what would be good to focus on when continuing the development of this method to find an approximate neutral density variable for the world's oceans. One of the first things to look at is the generalization of the finite difference method to three dimensional domains. A natural part of this will be increased complexity in the finite differencing of the boundary value problem.

The finite difference method of this study assumes regularity of the domain which includes uniform steps in vertical direction  $\Delta z$ . This is an idealization of how most hydrographic datasets are constructed as  $\Delta z$  often increases the further away from the surface you go. Before looking at the neutrality of the neutral density variable this idealization needs to be eliminated. Such an effort requires modification of the finite difference discretization whose current form can be found in Section 4.1 and appendix A.

Another idealization of the domain considered in this study when compared to realistic ocean basins is that the boundary have simple outer unit normal vectors;  $\vec{n} = [\pm 1, 0]^T$  and  $\vec{n} = [0, \pm 1]^T$ . In reality the boundary values in hydrographic datasets follows the bathymetry of the ocean basin. Figure 5.1 and 5.2 illustrates how such bathymetries might look but even this is a simple example and the ocean basin can have more variation at the bottom and sides. It is important to investigate the consequence of a complex bathymetry for two reasons. Firstly, to understand the accuracy of the finite difference approximations near the boundary when it varies drastically. Secondly, it is important to understand to what extent increased complexity of the bathymetry leads to increased complexity in implementation of the method.

Currently, there are still some questionmarks with regards to the optimization problem and if these terms are sufficient in finding the most appropriate neutral density variable. Issues include physical contribution to the solution in areas where the magnitude of the diapycnal vector  $\vec{A}$  is small. One possible improvement is to include normalization of terms by the magnitude of  $\vec{A}$  however, it might also be advantageous to make terms non-linear in  $\gamma$  or  $\nabla\gamma$ . In the latter case, it is important to note that non-linearities in  $\gamma$  or  $\nabla\gamma$  results in more involved derivation of the Euler Lagrange equation. It is also probable that the boundary value problem becomes more difficult to discretise when including non-linearities in  $\gamma$  or  $\nabla\gamma$ . To get a more rigorous understanding of

the method, future work could also focus on firstly, creating a solid theoretical understanding of the solution in the limit  $\mu \rightarrow 0$  and secondly, formalizing equivalence and other commonalities between the optimization problem and the Euler-Lagrange equation.

After eliminating idealizations of the domain and improving the physical accuracy of the optimization problem it is of high interest to evaluate the neutrality of the neutral surfaces produced by this method. This will make it possible to diagnose the method of this study and compare it to previous approaches that constructs neutral surfaces.

## References

- [1] Y. You, “Review of global ocean intermediate water masses: 1. part a, the neutral density surface (the ‘mcdougall surface’) as a study frame for water-mass analysis,” *Journal of Ocean University of China*, vol. 5, no. 3, pp. 187–199, 2006. [Page 1.]
- [2] R. H. Stewart, *Introduction to Physical Oceanography*, 2008. [Page 1.]
- [3] T. J. McDougall, “Neutral surfaces,” *Journal of physical oceanography*, vol. 17, no. 11, pp. 1950–1964, 1987. [Pages 2, 7, and 8.]
- [4] F. Roquet, G. Madec, T. J. McDougall, and P. M. Barker, “Accurate polynomial expressions for the density and specific volume of seawater using the TEOS-10 standard,” *Ocean Modelling*, vol. 90, no. 0, pp. 29–43, 2015. doi: <http://dx.doi.org/10.1016/j.ocemod.2015.04.002> ISBN: 1463-5003. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1463500315000566> [Pages 2 and 5.]
- [5] H. B. Phillips, *Vector analysis*. New York: Wiley, 1933. ISBN 991-931647-4 [Pages 2 and 9.]
- [6] T. J. McDougall and D. R. Jackett, “On the helical nature of neutral trajectories in the ocean,” *Progress in oceanography*, vol. 20, no. 3, pp. 153–183, 1988. [Pages 2 and 10.]
- [7] D. R. Jackett and T. J. McDougall, “A neutral density variable for the world’s oceans,” *Journal of physical oceanography*, vol. 27, no. 2, pp. 237–263, 1997. [Pages 2 and 9.]
- [8] Y. Lang, G. J. Stanley, T. J. McDougall, and P. M. Barker, “A pressure-invariant neutral density variable for the world’s oceans,” *Journal of physical oceanography*, vol. 50, no. 12, pp. 3585–3604, 2020. [Page 2.]

- [9] C. Eden and J. Willebrand, “Neutral density revisited,” *Deep-sea research. Part II, Topical studies in oceanography*, vol. 46, no. 1, pp. 33–54, 1999. [Pages 2, 3, and 12.]
- [10] G. J. Stanley, T. J. McDougall, and P. M. Barker, “Algorithmic improvements to finding approximately neutral surfaces,” *Journal of advances in modeling earth systems*, vol. 13, no. 5, 2021. [Pages 2 and 3.]
- [11] A. Klocker, T. J. McDougall, and D. R. Jackett, “A new method for forming approximately neutral surfaces,” *Ocean science*, vol. 5, no. 2, pp. 155–172, 2009. [Page 3.]
- [12] T. J. McDougall and P. M. Barker, “Getting started with teos-10 and the gibbs seawater (gsw) oceanographic toolbox,” SCOR/IAPSO WG127, Tech. Rep., 2011. [Pages 5, 6, and 7.]
- [13] F. Peeters, G. Piepke, R. Kipfer, R. Hohmann, and D. M. Imboden, “Description of stability and neutrally buoyant transport in freshwater lakes,” *Limnology and oceanography*, vol. 41, no. 8, pp. 1711–1724, 1996. [Page 8.]
- [14] A. Persson and L.-C. Böiers, *Analys i flera variabler*, 3rd ed. Lund: Studentlitteratur, 2005. ISBN 91-44-03869-0 [Page 9.]
- [15] H. H. Goldstine, *A history of the calculus of variations from the 17th through the 19th century*, 1st ed., ser. Studies in the History of Mathematics and Physical Sciences ; Volume 5. New York: Springer-Verlag, 1980. ISBN 1-4613-8106-1 [Page 14.]
- [16] P. de Fermat, *Methodus ad disquirendam maximam et minimam*, 1636. [Page 15.]
- [17] F. Rindler, *Calculus of Variations*, 1st ed., ser. Universitext. Cham: Springer International Publishing, 2018. ISBN 3-319-77637-1 [Page 16.]
- [18] J. Hadamard, “Sur un problème mixte aux dérivées partielles,” *Bulletin de la Société mathématique de France*, vol. 2, pp. 208–224, 1903. [Page 23.]
- [19] L. C. Evans, *Partial differential equations*, 2nd ed., ser. Graduate studies in mathematics, 19. Providence, R.I: American Mathematical Society, 2010. ISBN 0-8218-4974-3 [Page 23.]

- [20] B. van Es, B. Koren, and H. J. de Blank, “Finite-difference schemes for anisotropic diffusion,” *Journal of computational physics*, vol. 272, pp. 526–549, 2014. [Pages vi, 33, 36, and 37.]

# Appendix A

## Finite difference discretization

The general form of the Euler-Lagrange equation with Neumann boundary conditions of this study is specified in  $(BVP')$ . Suppose  $F = K\nabla\gamma$  and

$$F := \begin{bmatrix} F^{(y)} \\ F^{(z)} \end{bmatrix}, \quad K := \begin{bmatrix} K^{(yy)} & K^{(yz)} \\ K^{(zy)} & K^{(zz)} \end{bmatrix}.$$

The flux terms,  $F^{(y)}$  and  $F^{(z)}$ , are given by

$$F^{(y)} = K^{(yy)} \frac{\partial\gamma}{\partial y} + K^{(yz)} \frac{\partial\gamma}{\partial z},$$

$$F^{(z)} = K^{(zy)} \frac{\partial\gamma}{\partial y} + K^{(zz)} \frac{\partial\gamma}{\partial z}.$$

The Euler Lagrange equation is discretized by finite difference schemes beginning with the diffusion term and then moving onto the source term.

The divergence is evaluated on cell centres making the flux terms evaluated on cell faces such that

$$(\nabla \cdot F)(y_i, z_j) \approx \frac{F_{i+\frac{1}{2},j}^{(y)} - F_{i-\frac{1}{2},j}^{(y)}}{\Delta y} + \frac{F_{i,j+\frac{1}{2}}^{(z)} - F_{i,j-\frac{1}{2}}^{(z)}}{\Delta z}.$$

For all  $(i, j) \in \mathcal{I}_y \times \mathcal{I}_z$  we have,

$$\begin{aligned}
 F_{i+\frac{1}{2},j}^{(y)} &= \bar{K}_{i+\frac{1}{2},j}^{(yy)} d_y \gamma_{i+\frac{1}{2},j} + \bar{K}_{i+\frac{1}{2},j}^{(yz)} d_z \gamma_{i+\frac{1}{2},j} \\
 &= \bar{K}_{i+\frac{1}{2},j}^{(yy)} \frac{\gamma_{i+1,j} - \gamma_{i,j}}{\Delta y} + \bar{K}_{i+\frac{1}{2},j}^{(yz)} \frac{\gamma_{i+1,j+1} + \gamma_{i,j+1} - \gamma_{i,j-1} - \gamma_{i+1,j-1}}{4\Delta z} \\
 &= \frac{\bar{K}_{i+\frac{1}{2},j}^{(yy)}}{\Delta y} (\gamma_{i+1,j} - \gamma_{i,j}) + \frac{\bar{K}_{i+\frac{1}{2},j}^{(yz)}}{4\Delta z} (\gamma_{i+1,j+1} + \gamma_{i,j+1} - \gamma_{i,j-1} - \gamma_{i+1,j-1}),
 \end{aligned}$$

$$\begin{aligned}
 F_{i,j+\frac{1}{2}}^{(z)} &= \bar{K}_{i,j+\frac{1}{2}}^{(zy)} d_y \gamma_{i,j+\frac{1}{2}} + \bar{K}_{i,j+\frac{1}{2}}^{(zz)} d_z \gamma_{i,j+\frac{1}{2}} \\
 &= \bar{K}_{i,j+\frac{1}{2}}^{(zy)} \frac{\gamma_{i+1,j+1} + \gamma_{i+1,j} - \gamma_{i-1,j+1} - \gamma_{i-1,j}}{4\Delta y} + \bar{K}_{i,j+\frac{1}{2}}^{(zz)} \frac{\gamma_{i,j+1} - \gamma_{i,j}}{\Delta z} \\
 &= \frac{\bar{K}_{i,j+\frac{1}{2}}^{(zy)}}{4\Delta y} (\gamma_{i+1,j+1} + \gamma_{i+1,j} - \gamma_{i-1,j+1} - \gamma_{i-1,j}) + \frac{\bar{K}_{i,j+\frac{1}{2}}^{(zz)}}{\Delta z} (\gamma_{i,j+1} - \gamma_{i,j}).
 \end{aligned}$$

To simplify matters, it is sufficient to derive the flux expressions on the face in the positive direction as  $F_{i-\frac{1}{2},j}^{(y)}$  and  $F_{i,j-\frac{1}{2}}^{(z)}$  can be defined through the formulas for  $F_{i+\frac{1}{2},j}^{(y)}$  and  $F_{i,j+\frac{1}{2}}^{(z)}$  such that

$$\begin{aligned}
 F_{i-\frac{1}{2},j}^{(y)} &= F_{(i-1)+\frac{1}{2},j}^{(y)} \\
 \Leftrightarrow F_{i-\frac{1}{2},j}^{(y)} &= \frac{\bar{K}_{i-\frac{1}{2},j}^{(yy)}}{\Delta y} (\gamma_{i,j} - \gamma_{i-1,j}) + \frac{\bar{K}_{i-\frac{1}{2},j}^{(yz)}}{4\Delta z} (\gamma_{i,j+1} + \gamma_{i-1,j+1} - \gamma_{i-1,j-1} - \gamma_{i,j-1}),
 \end{aligned}$$

$$\begin{aligned}
 F_{i,j-\frac{1}{2}}^{(z)} &= F_{i,(j-1)+\frac{1}{2}}^{(z)} \\
 \Leftrightarrow F_{i,j-\frac{1}{2}}^{(z)} &= \frac{\bar{K}_{i,j-\frac{1}{2}}^{(zy)}}{4\Delta y} (\gamma_{i+1,j} + \gamma_{i+1,j-1} - \gamma_{i-1,j} - \gamma_{i-1,j-1}) + \frac{\bar{K}_{i,j-\frac{1}{2}}^{(zz)}}{\Delta z} (\gamma_{i,j} - \gamma_{i,j-1}).
 \end{aligned}$$

These four expressions for the flux components at cell faces are put into the divergence term such that

$$\begin{aligned}
\frac{F_{i+\frac{1}{2},j}^{(y)} - F_{i-\frac{1}{2},j}^{(y)}}{\Delta y} &= \left( \frac{\bar{K}_{i+\frac{1}{2},j}^{(yy)}}{\Delta y \Delta y} (\gamma_{i+1,j} - \gamma_{i,j}) \right. \\
&+ \frac{\bar{K}_{i+\frac{1}{2},j}^{(yz)}}{4\Delta z \Delta y} (\gamma_{i+1,j+1} + \gamma_{i,j+1} - \gamma_{i,j-1} - \gamma_{i+1,j-1}) \\
&- \left( \frac{\bar{K}_{i-\frac{1}{2},j}^{(yy)}}{\Delta y \Delta y} (\gamma_{i,j} - \gamma_{i-1,j}) \right. \\
&+ \left. \frac{\bar{K}_{i-\frac{1}{2},j}^{(yz)}}{4\Delta z \Delta y} (\gamma_{i,j+1} + \gamma_{i-1,j+1} - \gamma_{i-1,j-1} - \gamma_{i,j-1}) \right) \\
&= \frac{\bar{K}_{i+\frac{1}{2},j}^{(yy)}}{\Delta y \Delta y} \gamma_{i+1,j} \\
&+ \left( -\frac{\bar{K}_{i+\frac{1}{2},j}^{(yy)}}{\Delta y \Delta y} - \frac{\bar{K}_{i-\frac{1}{2},j}^{(yy)}}{\Delta y \Delta y} \right) \gamma_{i,j} \\
&+ \frac{\bar{K}_{i+\frac{1}{2},j}^{(yz)}}{4\Delta z \Delta y} \gamma_{i+1,j+1} \\
&+ \left( \frac{\bar{K}_{i+\frac{1}{2},j}^{(yz)}}{4\Delta z \Delta y} - \frac{\bar{K}_{i-\frac{1}{2},j}^{(yz)}}{4\Delta z \Delta y} \right) \gamma_{i,j+1} \\
&+ \left( -\frac{\bar{K}_{i+\frac{1}{2},j}^{(yz)}}{4\Delta z \Delta y} + \frac{\bar{K}_{i-\frac{1}{2},j}^{(yz)}}{4\Delta z \Delta y} \right) \gamma_{i,j-1} \\
&- \frac{\bar{K}_{i+\frac{1}{2},j}^{(yz)}}{4\Delta z \Delta y} \gamma_{i+1,j-1} + \frac{\bar{K}_{i-\frac{1}{2},j}^{(yy)}}{\Delta y \Delta y} \gamma_{i-1,j} \\
&- \frac{\bar{K}_{i-\frac{1}{2},j}^{(yz)}}{4\Delta z \Delta y} \gamma_{i-1,j+1} + \frac{\bar{K}_{i-\frac{1}{2},j}^{(yz)}}{4\Delta z \Delta y} \gamma_{i-1,j-1}
\end{aligned}$$

and

$$\begin{aligned}
 \frac{F_{i,j+\frac{1}{2}}^{(z)} - F_{i,j-\frac{1}{2}}^{(z)}}{\Delta z} &= \left( \frac{\bar{K}_{i,j+\frac{1}{2}}^{(zy)}}{4\Delta y\Delta z} (\gamma_{i+1,j+1} + \gamma_{i+1,j} - \gamma_{i-1,j+1} - \gamma_{i-1,j}) \right. \\
 &\quad \left. + \frac{\bar{K}_{i,j+\frac{1}{2}}^{(zz)}}{\Delta z\Delta z} (\gamma_{i,j+1} - \gamma_{i,j}) \right) \\
 &\quad - \left( \frac{\bar{K}_{i,j-\frac{1}{2}}^{(zy)}}{4\Delta y\Delta z} (\gamma_{i+1,j} + \gamma_{i+1,j-1} - \gamma_{i-1,j} - \gamma_{i-1,j-1}) \right. \\
 &\quad \left. + \frac{\bar{K}_{i,j-\frac{1}{2}}^{(zz)}}{\Delta z\Delta z} (\gamma_{i,j} - \gamma_{i,j-1}) \right) \\
 &= \frac{\bar{K}_{i,j+\frac{1}{2}}^{(zy)}}{4\Delta y\Delta z} \gamma_{i+1,j+1} + \left( \frac{\bar{K}_{i,j+\frac{1}{2}}^{(zy)}}{4\Delta y\Delta z} - \frac{\bar{K}_{i,j-\frac{1}{2}}^{(zy)}}{4\Delta y\Delta z} \right) \gamma_{i+1,j} \\
 &\quad - \frac{\bar{K}_{i,j+\frac{1}{2}}^{(zy)}}{4\Delta y\Delta z} \gamma_{i-1,j+1} + \left( -\frac{\bar{K}_{i,j+\frac{1}{2}}^{(zy)}}{4\Delta y\Delta z} + \frac{\bar{K}_{i,j-\frac{1}{2}}^{(zy)}}{4\Delta y\Delta z} \right) \gamma_{i-1,j} \\
 &\quad + \frac{\bar{K}_{i,j+\frac{1}{2}}^{(zz)}}{\Delta z\Delta z} \gamma_{i,j+1} + \left( -\frac{\bar{K}_{i,j+\frac{1}{2}}^{(zz)}}{\Delta z\Delta z} - \frac{\bar{K}_{i,j-\frac{1}{2}}^{(zz)}}{\Delta z\Delta z} \right) \gamma_{i,j} \\
 &\quad - \frac{\bar{K}_{i,j-\frac{1}{2}}^{(zy)}}{4\Delta y\Delta z} \gamma_{i+1,j-1} \\
 &\quad + \frac{\bar{K}_{i,j-\frac{1}{2}}^{(zy)}}{4\Delta y\Delta z} \gamma_{i-1,j-1} + \frac{\bar{K}_{i,j-\frac{1}{2}}^{(zz)}}{\Delta z\Delta z} \gamma_{i,j-1}.
 \end{aligned}$$

To improve readability as well as ease implementation, consider the following shift in notation for the coefficients:

$$\begin{aligned}
 k^{(yy)} &:= \frac{K^{(yy)}}{\Delta y^2}, & k^{(yz)} &:= \frac{K^{(yz)}}{4\Delta y\Delta z}, \\
 k^{(zy)} &:= \frac{K^{(zy)}}{4\Delta y\Delta z}, & k^{(zz)} &:= \frac{K^{(zz)}}{\Delta z^2},
 \end{aligned}$$

$$\begin{aligned}
 \bar{k}_+^{(*)} &:= \bar{k}_{i+\frac{1}{2},j}^{(*)}, & \bar{k}_-^{(*)} &:= \bar{k}_{i-\frac{1}{2},j}^{(*)} \text{ for } * \in \{yy, yz\}, \\
 \bar{k}_+^{(*)} &:= \bar{k}_{i,j+\frac{1}{2}}^{(*)}, & \bar{k}_-^{(*)} &:= \bar{k}_{i,j-\frac{1}{2}}^{(*)} \text{ for } * \in \{zy, zz\}.
 \end{aligned}$$

The full divergence term evaluated at  $(i, j) \in \mathcal{I}_y \times \mathcal{I}_z$ , defined through these new coefficients, is

$$\begin{aligned}
(\nabla \cdot F)(y_i, z_j) &\approx \left( -\bar{k}_-^{(yz)} - \bar{k}_+^{(zy)} \right) \gamma_{i-1, j+1} \\
&+ \left( \bar{k}_+^{(yz)} - \bar{k}_-^{(yz)} + \bar{k}_+^{(zz)} \right) \gamma_{i, j+1} \\
&+ \left( \bar{k}_+^{(zy)} + \bar{k}_+^{(yz)} \right) \gamma_{i+1, j+1} \\
&+ \left( -\bar{k}_+^{(zy)} + \bar{k}_-^{(zy)} + \bar{k}_-^{(yy)} \right) \gamma_{i-1, j} \\
&+ \left( -\bar{k}_+^{(zz)} - \bar{k}_-^{(zz)} - \bar{k}_+^{(yy)} - \bar{k}_-^{(yy)} \right) \gamma_{i, j} \\
&+ \left( \bar{k}_+^{(zy)} - \bar{k}_-^{(zy)} + \bar{k}_+^{(yy)} \right) \gamma_{i+1, j} \\
&+ \left( \bar{k}_-^{(yz)} + \bar{k}_-^{(zy)} \right) \gamma_{i-1, j-1} \\
&+ \left( -\bar{k}_+^{(yz)} + \bar{k}_-^{(yz)} + \bar{k}_-^{(zz)} \right) \gamma_{i, j-1} \\
&+ \left( -\bar{k}_+^{(yz)} - \bar{k}_-^{(zy)} \right) \gamma_{i+1, j-1}.
\end{aligned}$$

The source term used in  $(BVP')$  is analogously evaluated on cell centres and hence  $f$  is evaluated on cell faces such that for

$$f := \begin{bmatrix} f^{(y)} \\ f^{(z)} \end{bmatrix}$$

we have,

$$\begin{aligned}
\nabla \cdot f(y_i, z_j) &= \frac{\partial f^{(y)}}{\partial y}(y_i, z_j) + \frac{\partial f^{(z)}}{\partial z}(y_i, z_j) \\
&\approx \frac{\bar{f}_{i+\frac{1}{2}, j}^{(y)} - \bar{f}_{i-\frac{1}{2}, j}^{(y)}}{\Delta y} + \frac{\bar{f}_{i, j+\frac{1}{2}}^{(z)} - \bar{f}_{i, j-\frac{1}{2}}^{(z)}}{\Delta z} \quad \forall (i, j) \in \mathcal{I}_y \times \mathcal{I}_z.
\end{aligned}$$



