Bivariate Spline Solution to a Class of Reaction-Diffusion Equations

by

George Petrov Slavov

(Under the direction of Ming-Jun Lai)

Abstract

This work presents a method of solving a time dependent partial differential equation, which arises from classic models in ecology concerned with a species' population density over two dimensional domains. The species experiences population growth and diffuses over time due to overcrowding. Population growth is modeled using logistic growth with Allee effect. This work introduces the concept of discrete weak solution and establish theory for the existence, uniqueness and stability of the solution. Bivariate splines of arbitrary degree and smoothness across elements are used to approximate the discrete weak solution. More recent efforts focus on modeling the interaction of multiple species, which either compete for a common resource or one predates on the other. Some simulations of population development over some irregular domains are presented at the end.

Index words: bivariate splines, partial differential equation, nonlinear, diffusion

Bivariate Spline Solution to a Class of Reaction-Diffusion Equations

by

George Petrov Slavov

B.A., Bowdoin College, 2009

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2016

Bivariate Spline Solution to a Class of Reaction-Diffusion Equations

by

George Petrov Slavov

Approved:

Major Professor: Ming-Jun Lai

Committee: Juan Gutierrez
Jingzhi Tie
Sa'ar Hersonsky

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2016

Acknowledgments

I would like to thank my advisor Ming-Jun Lai without whose help and patience this work would not have been possible. I would also like to thank my sister Elly Slavova, my parents Nadya and Peter, and my lifelong friend Hristo Georgiev who have always been supportive of me during my struggles.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Formulation

Let $\Omega \subset \mathbb{R}^2$ be a polygonal domain. The first goal of this dissertation is to present a solution to the class of time-dependent, nonlinear partial differential equations in equation (1.1.1).

$$\begin{cases} \dfrac{\partial p(\mathbf{x},t)}{\partial t} &= \text{ div } (D(p,\mathbf{x})\nabla p(\mathbf{x},t))) + F(p(\mathbf{x},t)) \quad \mathbf{x} \in \Omega, t \in [0,T] \\ p(\mathbf{x},t) &\geq 0 \quad \mathbf{x} \in \Omega, t \in [0,T] \\ p(\mathbf{x},t) &= 0 \quad \mathbf{x} \in \partial\Omega, t \in [0,T] \\ p(\mathbf{x},0) &= p_0 \quad \mathbf{x} \in \Omega, \end{cases} \tag{1.1.1}$$

Here $D(p,\mathbf{x}) > 0$ is a known diffusive weight function, e.g. $D(p,\mathbf{x}) = D > 0$ and $F(p)$ is a growth function a.e. Lipschitz continuous and bounded above, e.g. $F(p) = A(\mathbf{x})p(1-p)$, which is a standard logistic growth function with $A(\mathbf{x})$ being a nonnegative, bounded weight function over Ω. Chapter 4, which focuses on numerical examples, will use $F(p) = Ap(1-p)(p-\sigma)$ where $\sigma \in [0,1)$. This term incorporates an Allee effect into the growth term. The significance of the Allee effect is discussed

in Section 1.2.

The PDE can be modified to satisfy a Neumann boundary condition. The remainder of this dissertation will, however, focus on studying the Dirichlet boundary condition.

The second goal of this dissertation is to present a solution to a model for population density of two species on the same domain Ω. They are modeled by the class of PDEs in equation (1.1.2), which is a natural extension of the single species system above.

Let $\Omega \subset \mathbb{R}^2$ be a polygonal domain.

$$\begin{cases} \dfrac{dp(\mathbf{x},t)}{dt} &= \text{div } (D\nabla p(\mathbf{x},t))) + F(p(\mathbf{x},t), m(\mathbf{x},t)) \quad \mathbf{x} \in \Omega, t \in [0,T] \\ \dfrac{dm(\mathbf{x},t)}{dt} &= \text{div } (E\nabla m(\mathbf{x},t))) + G(p(\mathbf{x},t), m(\mathbf{x},t)) \quad \mathbf{x} \in \Omega, t \in [0,T] \\ p(\mathbf{x},t) &\geq 0 \quad \mathbf{x} \in \Omega, t \geq 0 \\ m(\mathbf{x},t) &\geq 0 \quad \mathbf{x} \in \Omega, t \geq 0 \\ p(\mathbf{x},t) &= 0 \quad \mathbf{x} \in \partial\Omega, t \geq 0 \\ m(\mathbf{x},t) &= 0 \quad \mathbf{x} \in \partial\Omega, t \geq 0 \\ p(\mathbf{x},0) &= p_0 \quad \mathbf{x} \in \Omega \\ m(\mathbf{x},0) &= m_0 \quad \mathbf{x} \in \Omega \end{cases} \tag{1.1.2}$$

As before, $D = D(p, \mathbf{x}) > 0$ is a diffusive weight function and a corresponding function E is introduced for the second species. Growth of the two species is modeled by F and G, which could take a variety of forms. For example, $F(p,m) = Ap(1-p)(p-\sigma) - \mu pm$ would correspond to logistic growth with Allee effect of the species p and a loss of population due to the presence of species m. The term μpm is called mass action or Holling type I response. The alternative is a growth term of the form $F(p,m) = Ap(1-p)(p-\sigma) - \mu\dfrac{pm}{\xi + p}$ which features a Holling type II functional

response. We will assume that F and G are Lipschitz continuous and bounded above. The boundary conditions here are similar to the ones chosen for the single species case. Neumann boundary conditions are also possible.

This dissertation constructs a scheme to compute a numeric approximation of the PDE solution using bivariate splines on a triangulated, polygonal domain. The time variable is discretized using an implicit Euler finite difference scheme. We establish existence, uniqueness and stability of the numerical solution using techniques from convex optimization. Finally, we present a fixed-point iterative scheme for computing the approximation in a finite-dimensional spline space and prove the scheme's convergence to the discrete weak solution. Chapter 2 develops the theory for the single species case in equation (1.1.1). Chapter 3 develops the corresponding theory for the two species case in equation (1.1.2).

1.2 Literature Review

The topic of population dynamics has a lengthy history dating back to Malthus in 1798 [28] who first studied population growth. The model was overly simplistic as it gives rise to asymptotically unbounded population and so Verhulst in 1838 [34] introduced the logistic growth model $p_t = r_0 p(1 - p/k)$, which provided more realistic outcomes. p_0 here represents rate of growth and k represents carrying capacity. In 1910 Lotka [27] introduced the ODE model known today as the Lotka-Volterra equations. They were initially proposed as a model for chemical reactions between two substances with masses u and v. The equations are

$$\frac{du}{dt} = \lambda u - buv, \quad \frac{dv}{dt} = -\mu v + cuv$$

where λ, μ, b and c are positive constants. The quadratic term uv is referred to as mass action and was motivated by the tendency of chemical reactions to be faster

in the presence of a larger mass of chemicals. The equations were later adopted by Volterra in 1926 [35] as a model for predator-prey interactions, letting u and v be the number of individuals of each species. Empirical evidence for the validity of this model was brought forward in 1935 by Gause [13] and later by Huffaker in 1958 [18]. Solomon [32] and Holling [16] deemed that mass action as applied in the case of ecology was inadequate since a predator has a limited ability to consume its prey. Thus, they introduced the Holling type II functional response and so the new model became.

$$\frac{du}{dt} = \lambda u - b\frac{u}{1+mu}v, \quad \frac{dv}{dt} = -\mu v + c\frac{u}{1+mu}v$$

where m is a positive constant. With this change in place, the predator's rate of growth is not as strongly influenced by high predator density.

Cantrell and Cosner in [2] is a rich resource, which provides extensive qualitative discussion of Lotka-Volterra equations, with both mass action and Holling type II response. A central topic in the discussion is coexistence of species. In the simplest ODE formulation

$$\frac{du}{dt} = \lambda u - buv, \quad \frac{dv}{dt} = -\mu v - cuv$$

there exists no equilibrium which corresponds to the coexistence of both species. One species is guaranteed to perish in the long term. The earliest analysis of these systems gave rise to the so-called "paradox of diversity," named in the article of Hutchinson [19], which rightfully asks the question of how it is possible for our ecosystem to support so many species, which compete for a common resource. A number of modifications to the Lotka-Volterra model were proposed to address this misalignment between theory and observation.

An example of such a modification to Lotka-Volterra is agent-based models [4], which represent the environment as a discrete lattice with "particles" inhabiting the nodes of the lattice. Individual reproduction and movement can be defined as deter-

ministic or probabilistic. Such a system is well-suited to numeric simulations, which showed that it is possible for multiple species to coexist by isolating themselves in small, localized regions, thereby removing the burden of competing for a common resource. At the same time, competition from species in surrounding patches forces each patch to remain relatively stable for long periods of time. Thus, on a micro scale the species experience little competition, yet on a macro scale the species can coexist while making use of the same resource. The spatial heterogeneity of a species' population is key to its survival in this model.

Due to this emergence of spatial heterogeneity in the agent-based model, it becomes natural to choose to extend the dependence of u and v in the Lotka-Volterra equations to the spatial domain, giving rise to reaction-diffusion equations, as described in Section 1.1. They were proposed as early as 1937 by Fisher [10] who first studied the equation

$$\frac{\partial p}{\partial t} = D\frac{\partial^2 p}{\partial x^2} + kp(1-p)$$

in the one-dimensional setting and characterized the "traveling wave" solution. Since then there have been many other studies of reaction-diffusion equations, see Cantrell and Cosner in [2] for extensive bibliography and [6] for an excellent survey of publications on the topic. Lopez-Gomez in [7] formulated the problem of "competition with refuges," which explicitly defines subsets of $\mathbb{R}^2$ as more beneficial to species A than to species B and gives rise to coexistence.

A further issue with the Lotka-Volterra model is that even at very low population density, a species shows an ability to reproduce and thrive. This leads to the notorious "attofox" problem, which refers to the fact that even a fraction of an individual will eventually reproduce and its progeny will reach population capacity. In reality, low density leads to less efficient feeding, reduced effectiveness of vigilance and antipredator defenses, inbreeding depression as well as a slew of other negative outcomes; see, [36], [37], [21], [20], [14], [15], [11], [12], [33] and [30]. Lewis and

Kareiva in [26] provide a qualitative analysis of an equation with Allee effect as well as some numerical results. The addition of the Allee effect causes a marked change in the dynamics the system can exhibit as compared to the Lotka-Volterra model. Even if the initial population has density which is higher than the Allee threshold in certain regions, it is possible for the entire population to perish if the growth rate is not sufficiently large to overcome the decrease in population caused by diffusion. Using numerical experiments, the asymptotic behavior of the system is observed to depend on the shape and size of the region $\Omega \subset \mathbb{R}^2$ on which the system is defined. A complete qualitative description of this dependence remains elusive. Thus, there remains a need for robust numerical tools that shed light in specific applications of interest.

There have been a number of studies of reaction-diffusion studies in the literature which explore numerical simulations. Lewis and Kareiva [26] used finite differences, which is readily implementable in computer code and can provide good observations in synthetic tests. However, the use of finite difference methods is inadequate since realistic regions of interest could be polygonal subsets of $\mathbb{R}^2$, such counties, states or countries as derived from political maps. An example application is presented in Richter et al [31] which was solved using linear finite elements, implemented by the general purpose software suite COMSOL Multiphysics. They do not provide a convergence analysis since they relied on the proper operation of third-party software.

In light of the need to numerically solve reaction-diffusion equations, this dissertation endeavors to provide a complete implementation of a finite element scheme based on bivariate splines and convergence analysis proving the robustness of the code. The reaction-diffusion system is kept general so as to be applicable to a wide class of systems instead of focusing on a particular choice of diffusion or growth regimes. The code has proved to be powerful enough to tackle the solution for population density of a single species and of multiple interacting species. In Chapter 4 we present a

number of example population densities subject to logistic growth with or without Allee effect, predator-prey interactions with mass action or Holling type II response, and two species competing for a common resource.

1.3 Some Well-Known Theorems and Lemmas

For the sake of completeness, we list a number of lemmas used in this dissertation, which are special cases of well-known results.

Lemma 1.3.1. *Any* $a, b \geq 0$ *and* $\alpha > 0$ *satisfy*

$$ab \leq \frac{\alpha}{2}a^2 + \frac{1}{2\alpha}b^2.$$

Lemma 1.3.2 (Ladyzhenskaya's Inequality)**.** *Any* $p \in H_0^1(\Omega)$ *such that* $\Omega \subset \mathbb{R}^2$ *satisfy*

$$\|p\|_{L^4} \leq C\|p\|_{L^2}^{1/2}\|\nabla p\|_{L^2}^{1/2}.$$

Definition 1.3.1. Let X and Y be Banach spaces, $X \subset Y$. We say X is compactly embedded in Y, written $X \subset\subset Y$, if the following conditions are satisfied.

(a) $\forall x \in X$, $||p||_{L^2(\Omega)} \leq C\,||p||_{H_0^1(\Omega)}$ for some constant C.

(b) Any bounded sequence in X has a subsequence which converges in Y.

Theorem 1.3.1 (Rellich-Kondrachov [8])**.** *Suppose* $\Omega \subset \mathbb{R}^2$ *is bounded with Lipschitz boundary. Then we have the following compact embedding.*

$$H^1(\Omega) \subset\subset L^2(\Omega)$$

Theorem 1.3.2 (General Sobolev Inequality [8])**.** *If* $p \in H^2(\Omega)$*, then* $p \in C^{0,\gamma}$*, the*

space of Höder continuous functions with any exponent $0 < \gamma < 1$. *Furthermore,*

$$||p||_{C^{0,\gamma}(\Omega)} \leq C\,||p||_{H^2(\Omega)}$$

where C *is a constant independent of* p.

Some well-known theory on bivariate splines can be found in the Appendix in Chapter 6. For a more complete discussion of spline theory see [24]. For computational schemes see [1]. As the PDE of interest (1.1.1) is nonlinear, the MATLAB code used in [1] has to be extended to handle this nonlinear PDE.

Chapter 2

Modeling a Single Species

The time dependence and nonlinearity of the PDE of interest presents a challenge for a numerical scheme charged with finding an approximate solution. To tackle this challenge, this dissertation introduces a sequence of alternate formulations, which are successively more tractable.

1) Introduce a weak formulation of the PDE, which is standard for any finite element scheme.

2) Discretize the time domain thereby removing the time dependence.

3) Introduce a fixed-point iteration scheme thereby removing all nonlinearity.

4) Discretize the Hilbert space $H_0^1(\Omega)$ by using the space of bivariate splines $S_d^r(\Delta)$ of degree d and smoothness r, which leaves us with a linear, finite-dimensional problem.

2.1 Discrete Weak Formulation

Let us begin by presenting the weak formulation. Suppose $p \in H_0^1(\Omega)$ is a solution to equation (1.1.1). Then for any $q \in H_0^1(\Omega)$, p satisfies the following weak formulation

obtained by integrating by parts.

$$\int_\Omega \frac{\partial p(\mathbf{x},t)}{\partial t} q(\mathbf{x}) d\mathbf{x} = -\int_\Omega D(\mathbf{x}) \nabla p(\mathbf{x},t) \cdot \nabla q(\mathbf{x}) d\mathbf{x} + \int_\Omega F(p(\mathbf{x},t)) q(\mathbf{x}) d\mathbf{x}. \quad (2.1.1)$$

Now consider $t \in [0,T]$ and partition $0 = t_0 < t_1 < t_2 < \cdots < t_m < t_{m+1} = T$. We approximate $\frac{dp(\mathbf{x},t)}{dt}$ by its divided difference, i.e.,

$$\frac{dp(\mathbf{x},t_i)}{dt} \approx \frac{p(\mathbf{x},t_i) - p(\mathbf{x},t_{i-1})}{h}$$

with $h = t_i - t_{i-1}$. Substitute this approximation into (2.1.1) to obtain

$$\begin{aligned} &\int_\Omega p_h(\mathbf{x},t_i) q(\mathbf{x}) d\mathbf{x} + h \int_\Omega D(\mathbf{x}) \nabla p_h(\mathbf{x},t_i) \cdot \nabla q(\mathbf{x}) d\mathbf{x} \\ &- h \int_\Omega F(p_h(\mathbf{x},t_i)) q(\mathbf{x}) d\mathbf{x} = \int_\Omega p_h(\mathbf{x},t_{i-1}) q(\mathbf{x}) d\mathbf{x}. \end{aligned} \quad (2.1.2)$$

Note that the function p_h has a subscript h to indicate its dependence on the choice of h; a solution to (2.1.2) is not the same as a solution to (2.1.1) and vice-versa. In addition, both (2.1.1) and (2.1.2) obey the same boundary conditions as (1.1.1).

Definition 2.1.1. Any solution to equation (2.1.2) for fixed $h > 0$, t_{i-1} and t_i is called a discrete weak solution of (1.1.1).

The following theorem guarantees that the discrete weak solution is a good approximation of the exact solution.

Theorem 2.1.1. *Let $p(\mathbf{x},t)$ be the classical solution of (1.1.1) and $p_h(\mathbf{x},t)$ be the discrete weak solution dependent on $h > 0$. Suppose that $p(\mathbf{x},t)$ is twice differentiable with respect to t. Then*

$$\int_\Omega |p(\mathbf{x},t_i) - p_h(\mathbf{x},t_i)|^2 d\mathbf{x} \le Ch, \quad \forall i = 0, \cdots, m+1, \quad (2.1.3)$$

as $h = T/(m+1) \to 0$, where $C > 0$ is a constant independent of h.

Proof. By Taylor expansion, we have

$$\frac{dp(\mathbf{x},t_i)}{dt} = \frac{p(\mathbf{x},t_i) - p(\mathbf{x},t_{i-1})}{h} + O(h), \tag{2.1.4}$$

where $O(h)$ is a quantity bounded by Ch for a positive constant $C < \infty$. Using (2.1.1) and (2.1.2), we have

$$\begin{aligned}
&\int_\Omega \frac{dp(\mathbf{x},t_i)}{dt} q(\mathbf{x}) d\mathbf{x} - \int_\Omega \frac{p_h(\mathbf{x},t_i) - p_h(\mathbf{x},t_{i-1})}{h} q(\mathbf{x}) d\mathbf{x} \\
=& -\int_\Omega D(\mathbf{x}) \nabla (p(\mathbf{x},t_i) - p_h(\mathbf{x},t_i)) \cdot \nabla q(\mathbf{x}) d\mathbf{x} \\
&+ \int_\Omega (F(p(\mathbf{x},t_i)) - F(p_h(\mathbf{x},t_i)) q(\mathbf{x}) d\mathbf{x}.
\end{aligned}$$

Substitute (2.1.4) to obtain.

$$\begin{aligned}
&\int_\Omega \frac{p(\mathbf{x},t_i) - p(\mathbf{x},t_{i-1})}{h} q(\mathbf{x}) d\mathbf{x} - \int_\Omega \frac{p_h(\mathbf{x},t_i) - p_h(\mathbf{x},t_{i-1})}{h} q(\mathbf{x}) d\mathbf{x} \\
=& -\int_\Omega D(\mathbf{x}) \nabla (p(\mathbf{x},t_i) - p_h(\mathbf{x},t_i)) \cdot \nabla q(\mathbf{x}) d\mathbf{x} \\
&+ \int_\Omega (F(p(\mathbf{x},t_i) - F(p_h(\mathbf{x},t_i)) q(\mathbf{x}) d\mathbf{x} + O(h).
\end{aligned}$$

Letting $q = p(\mathbf{x},t_i) - p_h(\mathbf{x},t_i) \in H_0^1(\Omega)$ in the above equality, we obtain

$$\begin{aligned}
&\int_\Omega |p(\mathbf{x},t_i) - p_h(\mathbf{x},t_i)|^2 d\mathbf{x} \\
=& \int_\Omega (p(\mathbf{x},t_{i-1}) - p_h(\mathbf{x},t_{i-1}) (p(\mathbf{x},t_i) - p_h(\mathbf{x},t_i)) \\
&- h \int_\Omega D(\mathbf{x}) |\nabla (p(\mathbf{x},t_i) - p_h(\mathbf{x},t_i))|^2 d\mathbf{x} \\
&+ h \int_\Omega (F(p(\mathbf{x},t_i)) - F(p_h(\mathbf{x},t_i)) (p(\mathbf{x},t_i) - p_h(\mathbf{x},t_i)) d\mathbf{x} + O(h^2)
\end{aligned}$$

Discard the positive gradient term and use Lemma 1.3.1 with $\alpha = 1$.

$$\begin{aligned}\leq & \frac{1}{2}\int_\Omega |p(\mathbf{x},t_i) - p_h(\mathbf{x},t_i)|^2 d\mathbf{x} + \frac{1}{2}\int_\Omega |p(\mathbf{x},t_{i-1}) - p_h(\mathbf{x},t_{i-1})|^2 d\mathbf{x} \\ & + h\int_\Omega (F(p(\mathbf{x},t_i)) - F(p_h(\mathbf{x},t_i))(p(\mathbf{x},t_i) - p_h(\mathbf{x},t_i))d\mathbf{x} + O(h^2)\end{aligned}$$

Since F is Lipschitz continuous, i.e. $|F(p) - F(q)| \leq L|p - q|$, it follows that

$$\begin{aligned}\int_\Omega |p(\mathbf{x},t_i) - p_h(\mathbf{x},t_i)|^2 d\mathbf{x} \leq & \int_\Omega |p(\mathbf{x},t_{i-1}) - p_h(\mathbf{x},t_{i-1})|^2 d\mathbf{x} \\ & + 2hLC_A \int_\Omega |p(\mathbf{x},t_i) - p_h(\mathbf{x},t_i)|^2 + O(h^2),\end{aligned}$$

where $C_A = \|A\|_{L^\infty(\Omega)}$. That is, we have

$$(1 - 2hLC_A)\int_\Omega |p(\mathbf{x},t_i) - p_h(\mathbf{x},t_i)|^2 d\mathbf{x} \leq \int_\Omega |p(\mathbf{x},t_{i-1}) - p_h(\mathbf{x},t_{i-1})|^2 d\mathbf{x} + O(h^2).$$

Letting $\alpha = 1/(1 - 2hLC_A)$, we multiply α on both sides above and then repeatedly apply the inequality for $i = k, \ldots, 1$ to obtain

$$\begin{aligned}\int_\Omega |p(\mathbf{x},t_k) - p_h(\mathbf{x},t_k)|^2 d\mathbf{x} \leq & \alpha \int_\Omega |p(\mathbf{x},t_{k-1}) - p_h(\mathbf{x},t_{k-1})|^2 d\mathbf{x} + \alpha O(h^2) \leq \cdots\cdots \\ & \leq \alpha^k \int_\Omega |p(\mathbf{x},t_0) - p_h(\mathbf{x},t_0)|^2 d\mathbf{x} + O(h^2)\sum_{i=0}^{k-1}\alpha^i\end{aligned}$$

Note that $p(\mathbf{x},t_0) = p_h(\mathbf{x},t_0)$ since they obey the same boundary conditions.

$$\begin{aligned}& \leq O(h^2)\sum_{i=0}^{m}\alpha^i \leq O(h^2)\frac{\alpha^{m+1}}{\alpha - 1} \\ & \leq O(h^2)\frac{\alpha^{T/h}}{\alpha - 1} = O(h^2)(1 - 2hLC_A)^{-T/h}\frac{1 - 2hLC_A}{2hLC_A} \\ & \leq O(h^2)e^{2TL}\frac{1 - 2hLC_A}{2hLC_A} = O(h)\end{aligned}$$

The last inequality concludes the proof. □

The theorem guarantees a discrete weak solution is a close approximation of a classical solution. From this point on, we will drop the subscript h from p_h for simplicity. We will also rewrite the weak formulation slightly by altering the growth term from $F(p)$ to $pF_1(p)$. The growth functions in which epidemiologists are interested all allow such a factoring, so the model remains sufficiently general.

Let $\mathcal{A} = \{p \in H_0^1(\Omega), p(x,y) \geq 0 \text{ for a.e. } (x,y) \in \Omega\}$ be the set of admissible functions. Here $\Omega \subset \mathbb{R}^2$ is an open, bounded domain with Lipschitz boundary. By letting $p = p(\mathbf{x}, t_i)$ and $\hat{p} = p(\mathbf{x}, t_{i-1})$, we rewrite (2.1.2) into a simpler form. Thus, we look for a population density in the admissible set $p \in \mathcal{A}$ which satisfies the following equation:

$$\int_\Omega pq \, d\mathbf{x} + h \int_\Omega D(\mathbf{x}) \nabla p \cdot \nabla q \, d\mathbf{x} = \int_\Omega \hat{p} q \, d\mathbf{x} + h \int_\Omega pF_1(p) q \, d\mathbf{x} \qquad \forall q \in H_0^1(\Omega) \quad (2.1.5)$$

where $0 < K \leq D(\mathbf{x}) \leq K_2$ is a diffusive weight function and $F_1(p)$ is a growth function, which is Lipschitz continuous and bounded above by some constant M. An example of interest in epidemiology is

$$F_1(p) = A(\mathbf{x})(1-p)(p-\sigma)$$

which models population growth with an Allee effect. Here $A(\mathbf{x})$ is a given bounded, nonnegative function and $\sigma \in [0, 1)$. We can assume $p(\mathbf{x}, 0) \in \mathcal{A}$ and thus $\hat{p} \in \mathcal{A}$ as well by induction.

We would like to know that equation (2.1.5) has a unique solution. In order to do that, we note that the discrete weak formulation is the Euler-Lagrange equation

of the following energy minimization problem.

$$\min_{p\in\mathcal{A}} E(p) = \min_{H_0^1(\Omega),p\geq 0} \int_\Omega p^2 \, d\mathbf{x} + h\int_\Omega D(\mathbf{x})|\nabla p|^2 \, d\mathbf{x} - h\int_\Omega G(p) \, d\mathbf{x} - \int_\Omega \hat{p}p \, d\mathbf{x} \quad (2.1.6)$$

where

$$G(p) = \int_0^p \xi F_1(\xi) \, d\xi$$

In order to show that the functional has a minimizer, we need a lower bound for its image.

Lemma 2.1.1. *Suppose we choose* $h < \dfrac{1}{M}$. *Then for any function* $p \in \mathcal{A}$ *the energy functional given in (2.1.6) satisfies*

$$E(p) \geq C \, ||p||^2_{H_0^1(\Omega)} - ||\hat{p}||_2^2$$

for some constant $C > 0$. *In particular,* $\inf\limits_{p\in\mathcal{A}} E(p) \geq -\,||\hat{p}||_2^2 > -\infty$.

Proof. First we will present an upper bound for one of the terms. Recall our assumption that $F_1(p)$ is bounded above. Thus, $F_1(p) \leq M$ for some constant M.

$$G(p) = \int_0^p \xi F_1(\xi) \, d\xi \leq M\int_0^p \xi \, d\xi = \frac{M}{2}p^2$$
$$\int_\Omega G(p) \, d\mathbf{x} \leq \frac{M}{2} \, ||p||_2^2$$

Now we prove the lower bound for the entire functional. We use the Cauchy-Schwarz inequality, the upper bound for $G(p)$ we just established and $D(\mathbf{x}) \geq K$.

$$\begin{aligned} E(p) &\geq ||p||_2^2 + hK \, ||\nabla p||_2^2 - \frac{hM}{2} \, ||p||_2^2 - ||\hat{p}||_2 \, ||p||_2 \\ &= \left(1 - \frac{hM}{2}\right) ||p||_2^2 + hK \, ||\nabla p||_2^2 - ||\hat{p}||_2 \, ||p||_2 \end{aligned}$$

Use our assumption for h in this Lemma and Lemma 1.3.1 on the last term with $\alpha = 2$.

$$\begin{aligned} &\geq \frac{1}{2}\left|\left|p\right|\right|_2^2 + hK\left|\left|\nabla p\right|\right|_2^2 - \left|\left|\hat{p}\right|\right|_2^2 - \frac{1}{4}\left|\left|p\right|\right|_2^2 \\ &\geq \min\left\{\frac{1}{4}, hK\right\}\left|\left|p\right|\right|_{H_0^1(\Omega)}^2 - \left|\left|\hat{p}\right|\right|_2^2 \end{aligned}$$

□

Lemma 2.1.2. *If $h < 1/M$, the energy functional in (2.1.6) is weakly lower semi-continuous on $H^1(\Omega)$. That is, if $p_k \to p^*$ weakly in $H^1(\Omega)$, then*

$$E(p^*) \leq \liminf_{k\to\infty} E(p_k)$$

Proof. Set $m := \liminf_{k\to\infty} E(p_k)$. By passing to a subsequence we can assume that $E(p_k) - m < 1/k$. That is, $\lim_{k\to\infty} E(p_k) = m$. Any weakly convergent sequence is bounded in $H^1(\Omega)$ norm, so by the Rellich-Kondrachov theorem (Theorem 1.3.1), we can pass to another subsequence which converges strongly in $L^2(\Omega)$. Taking one last subsequence, we can assume that $p_k \to p^*$ a.e. in Ω.

Fix $\epsilon > 0$. By Egoroff's theorem there exists a measurable set U_ϵ such that $p_k \to p^*$ uniformly on U_ϵ and $|\Omega - U_\epsilon| < \epsilon$. Also write

$$V_\epsilon = \left\{ x \in \Omega \middle| |p^*(\mathbf{x})| + |\nabla p^*(\mathbf{x})| < \frac{1}{\epsilon} \right\} \tag{2.1.7}$$

Then $|\Omega - V_\epsilon| \to 0$ as $\epsilon \to 0$. Let $O_\epsilon = U_\epsilon \cap V_\epsilon$ and note that

$$|\Omega - O_\epsilon| = |(\Omega - U_\epsilon) \cup (\Omega - V_\epsilon)| \leq |\Omega - U_\epsilon| + |\Omega - V_\epsilon| \to 0 \quad \text{as } \epsilon \to 0$$

Now

$$E(p_k) + \int_\Omega \hat{p} p_k \, d\mathbf{x} = \int_\Omega p_k^2 + hD(\mathbf{x})|\nabla p_k|^2 - hG(p_k) \, d\mathbf{x}$$

From the proof of Lemma 2.1.1 we know that the right-hand side is nonnegative.

$$\geq \int_{O_\epsilon} p_k^2 + hD(\mathbf{x})|\nabla p_k|^2 - hG(p_k) \, d\mathbf{x}$$

Since the function $\eta : \mathbb{R}^n \to \mathbb{R}$ given by $\eta(\mathbf{x}) = |x|^2$ is convex, it follows that

$$\geq \int_{O_\epsilon} p_k^2 + hD(\mathbf{x}) \left(|\nabla p^*|^2 + 2\nabla p^* \cdot (\nabla p_k - \nabla p^*) \right) \tag{2.1.8}$$

$$- hG(p_k) \, d\mathbf{x} \tag{2.1.9}$$

$$E(p_k) + \int_\Omega \hat{p} p_k \, d\mathbf{x} \geq \int_{O_\epsilon} p_k^2 + hD(\mathbf{x})|\nabla p^*|^2 - hG(p_k) \, d\mathbf{x}$$

$$+ \int_{O_\epsilon} 2hD(\mathbf{x}) \nabla p^* \cdot (\nabla p_k - \nabla p^*) \, d\mathbf{x} \tag{2.1.10}$$

Recall equation (2.1.7) and note that in the first integral every term is bounded above. In addition, $p_k \to p^*$ uniformly on O_ϵ and G is an absolutely continuous function, so $G(p_k) \to G(p^*)$ uniformly on O_ϵ. Thus,

$$\lim_{k\to\infty} \int_{O_\epsilon} p_k^2 + hD(\mathbf{x})|\nabla p^*|^2 - hG(p_k) \, d\mathbf{x} = \int_{O_\epsilon} (p^*)^2 + hD(\mathbf{x})|\nabla p^*|^2 - hG(p^*) \, d\mathbf{x} \tag{2.1.11}$$

As for the second integral, note that $\nabla p_k \to \nabla p^*$ weakly in $L^2(\Omega; \mathbb{R}^n)$. Since $hD(\mathbf{x})\nabla p^* \in L^2(\Omega; \mathbb{R}^n)$ it follows that

$$\lim_{k\to\infty} \int_{O_\epsilon} 2hD(\mathbf{x}) \nabla p^* \cdot (\nabla p_k - \nabla p^*) \, d\mathbf{x} = 0 \tag{2.1.12}$$

We then take limits as $k \to \infty$ on both sides of (2.1.10) and as a result of (2.1.11)

and (2.1.12), we have

$$m + \int_\Omega \hat{p}p^* \, d\mathbf{x} \geq \int_{O_\epsilon} (p^*)^2 + hD(\mathbf{x})|\nabla p^*|^2 - hG(p^*) \, d\mathbf{x}$$
$$m \geq \int_{O_\epsilon} (p^*)^2 + hD(\mathbf{x})|\nabla p^*|^2 - hG(p^*) \, d\mathbf{x} - \int_\Omega \hat{p}p^* \, d\mathbf{x}$$

Now we take the limit as $\epsilon \to 0$. Since the integrand is nonnegative and $O_\epsilon \uparrow \Omega$, the monotone convergence theorem guarantees that

$$m \geq \int_\Omega (p^*)^2 + hD(\mathbf{x})|\nabla p^*|^2 - hG(p^*) - \hat{p}p^* \, d\mathbf{x}$$
$$m \geq E(p^*)$$

□

Theorem 2.1.2. *There exists a function $p^* \in \mathcal{A}$ which minimizes the energy functional $E(p)$ defined in (2.1.6).*

Proof. Set $m := \inf_{p \in \mathcal{A}} E(p)$ and choose a minimizing sequence $\{p_k\}$. Then $E(p_k) \to m$. As a result of Lemma 2.1.1 we know that

$$||p_k||_{H_0^1(\Omega)} \leq E(p_k) + ||\hat{p}||_2^2$$

$E(p_k) \to m$, so $\sup_k E(p_k) < \infty$. Thus, the minimizing sequence is bounded in $H_0^1(\Omega)$. Since $H_0^1(\Omega)$ is weakly compact, there exists a subsequence p_k which converges weakly to some function $p^* \in H_0^1(\Omega)$. We'd like to know that p^* is also in the admissible set $\mathcal{A}$. By the Rellich-Kondrachov theorem (Theorem 1.3.1), we can pass to a subsequence which converges strongly in $L^2(\Omega)$. By taking another subsequence, we can assume that $p_k \to p^*$ a.e., so we conclude that $p^* \geq 0$ a.e. That is, p^* is in the admissible set $\mathcal{A}$.

It remains to show that p^* is a minimizer of $E(p)$. Lemma 2.1.2 assures us that

$$E(p^*) \leq \liminf_{k\to\infty} E(p_k) = m \tag{2.1.13}$$

Since $p^* \in \mathcal{A}$, we have $m \leq E(p)$. Together with (2.1.13), this implies that $E(p^*) = m = \min_{p\in\mathcal{A}} E(p)$. □

Lemma 2.1.3. *Recall that* $M = \max_{x\in\Omega, p\geq 0} F_1(p)$. *Let* $M' = \max_{x\in\Omega,p\geq 0} F_1'(p)$. *If* h *is small enough so that*

$$2 - hM - hM' p_{max} > 0$$

then the functional $E(p)$ *defined in (2.1.6) is* μ*-strongly convex. That is,* $\exists \mu > 0$ *such that*

$$E(y) \geq E(\mathbf{x}) + \langle \nabla E(\mathbf{x}), x - y\rangle + \frac{\mu}{2} ||x - y||_2^2$$

where $\langle \nabla E(\mathbf{x}), x - y\rangle$ *is the Gâteaux derivative of* E *at the point* x *in the direction* $x - y$.

Proof. We use an equivalent formulation of μ-strong convexity. It is enough to show that $\forall q \in O$ we have

$$\partial^2 E(p, q) \geq \mu ||q||_2^2$$

We compute the second Gâteaux derivative. Let $q \in H_0^1(\Omega)$. Then the second derivative is given by $\mathcal{F}''(0)$.

$$\begin{aligned}
\mathcal{F}(t) =& E(p + tq) \\
\mathcal{F}'(t) =& \int_\Omega 2(p + tq)q \, d\mathbf{x} + 2h \int_\Omega D(\mathbf{x})\nabla(p + tq) \cdot \nabla q \, d\mathbf{x} \\
& - h \int_\Omega (p + tq)F_1(p + tq)q \, d\mathbf{x} - \int_\Omega \hat{p}q \\
\mathcal{F}''(t) =& 2 \int_\Omega q^2 \, d\mathbf{x} + 2h \int_\Omega D(\mathbf{x})|\nabla q|^2 \, d\mathbf{x} - h \int_\Omega F_1(p + qt)q^2 \\
& + (p + tq)F_1'(p + tq)q^2 \, d\mathbf{x}
\end{aligned}$$

$$\begin{aligned}\partial^2 E(p,q) = \mathcal{F}''(0) =& 2\,||q||_2^2 + 2h\int_\Omega D(\mathbf{x})|\nabla q|^2\,d\mathbf{x} - \\ & h\int_\Omega F_1(p)q^2 - h\int_\Omega pF_1'(p)q^2\,d\mathbf{x} \\ \geq& 2\,||q||_2^2 - hM\,||q||_2^2 - hM'p_{\max}\,||q||_2^2 \\ =& (2 - hM - hM'p_{\max})\,||q||_2^2\end{aligned}$$

as desired. □

Theorem 2.1.3. *The energy functional in (2.1.6) has a unique minimizer.*

Proof. Suppose p and $\tilde{p}$ are both minimizers of $E(p)$. Then for any $q \in H_0^1(\Omega)$ we have

$$\langle \nabla E(p,q)\rangle = \langle \nabla E(\tilde{p},q)\rangle = 0$$

By Lemma 2.1.3 the following two inequalities hold.

$$\begin{aligned}E(p) &\geq E(\tilde{p}) + \frac{\mu}{2}\,||p-\tilde{p}||_2^2 \\ E(\tilde{p}) &\geq E(p) + \frac{\mu}{2}\,||p-\tilde{p}||_2^2\end{aligned}$$

Add the two inequalities.

$$0 \geq \mu\,||p-\tilde{p}||_2^2$$

Thus, $p = \tilde{p}$ a.e. □

Definition 2.1.2 (Sobolev gradient)**.** Fix $p \in H_0^1$ and let $L : H_0^1(\Omega) \to \mathbb{R}$ be given by $L(q) = \langle \nabla E(p), q\rangle$. Since L is a bounded linear functional, by the Riesz representation theorem, there exists a unique element $v \in H_0^1(\Omega)$ such that $L(q) = \langle v, q\rangle_{H^1(\Omega)}$. We call v the Sobolev gradient of E at p and will denote it $\nabla E(p)$.

The Sobolev gradient is a natural extension of the concept of gradient of a dif-

ferentiable function on $\mathbb{R}^n$ and similarly functions as the direction of steepest ascent for the functional E at the point p. For a discussion of the properties of Sobolev gradients see Neuberger [29].

Definition 2.1.3. Let $\mathcal{P}_+ : \mathbb{R} \to \mathbb{R}$.

$$\mathcal{P}_+(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

and $\mathcal{P}_-(x) = -\mathcal{P}_+(-x)$.

The function $\mathcal{P}_+$ extends to an operator on absolutely continuous functions by composition. That is, $\mathcal{P}_+(f)(x) = \mathcal{P}_+ \circ f(x)$. The result is absolutely continuous. Ultimately, we need to extend this operator to functions in Sobolev space.

Lemma 2.1.4. *The operator $\mathcal{P}_+$ extends to an operator $\mathcal{P}_+ : H_0^1(\Omega) \to H_0^1(\Omega)$ such that for any $f \in H_0^1(\Omega)$, $\mathcal{P}_+(f)$ is nonnegative a.e. and*

$$||\mathcal{P}_+(f)||_{H_0^1(\Omega)} \leq ||f||_{H_0^1(\Omega)}.$$

Proof. Note that $\mathcal{P}_+ : \mathbb{R} \to \mathbb{R}$ is a nonexpansive function. That is,

$$|\mathcal{P}_+(x) - \mathcal{P}_+(y)| \leq |x - y|.$$

Take any sequence $f_n \in C_c^\infty$ such that $||f_n - f||_{H_0^1(\Omega)} \to 0$. All f_n are absolutely continuous and so, $\mathcal{P}_+(f_n)$ is defined. Let $\omega_n = f_n^{-1}((-\infty, 0))$. Note that $\mathcal{P}_+(f_n)\big|_{\omega_n} \equiv 0$ and $\nabla\mathcal{P}_+(f_n)\big|_{\omega_n} \equiv 0$. Furthermore,

$$\begin{aligned} ||\mathcal{P}_+(f_n)||^2_{H_0^1(\Omega)} &= \int_{\Omega\setminus\omega} |f_n|^2 \, d\mathbf{x} + \int_{\Omega\setminus\omega} |\nabla f_n|^2 \, d\mathbf{x} \\ &\leq \int_\Omega |f_n|^2 \, d\mathbf{x} + \int_\Omega |\nabla f_n|^2 \, d\mathbf{x} = ||f_n||^2_{H_0^1(\Omega)}. \end{aligned}$$

The inequality above shows $\mathcal{P}_+$ is a nonexpansive map in the Sobolev norm. In addition, $\mathcal{P}_+(f_n)$ is also a Cauchy sequence since

$$||\mathcal{P}_+(f_n) - \mathcal{P}_+(f_m)||_{H_0^1(\Omega)} \leq ||f_n - f_m||_{H_0^1(\Omega)} \to 0.$$

Then let $\mathcal{P}_+(f) := \lim_{n\to\infty} \mathcal{P}_+(f_n)$. Since $\mathcal{P}_+$ is a nonexpansive map in Sobolev norm, we are assured that $||\mathcal{P}_+(f)|| \leq ||f||_{H_0^1(\Omega)}$. Note that $\mathcal{P}_+(f)$ is nonnegative a.e.

All that remains is to show this construction is well defined. To that end, choose a different sequence $g_n \in C_c^\infty(\Omega)$ such that $||g_n - f||_{H_0^1(\Omega)} \to 0$. Suppose $\mathcal{P}_+(g_n) \to g$.

$$\begin{aligned} ||\mathcal{P}_+(f) - g||_{H_0^1(\Omega)} &= ||\mathcal{P}_+(f) - g + \mathcal{P}_+(f_n) - \mathcal{P}_+(f_n) + \mathcal{P}_+(g_n) - \mathcal{P}_+(g_n)||_{H_0^1(\Omega)} \\ &\leq ||\mathcal{P}_+(f) - \mathcal{P}_+(f_n)||_{H_0^1(\Omega)} + ||g - \mathcal{P}_+(g_n)||_{H_0^1(\Omega)} \\ &+ ||\mathcal{P}_+(f_n) - \mathcal{P}_+(g_n)||_{H_0^1(\Omega)} \end{aligned}$$

By definition, the first two norms on the right hand side can be made arbitrarily small by choosing a large n. The last term can be bounded as well by

$$||\mathcal{P}_+(f_n) - \mathcal{P}_+(g_n)||_{H_0^1(\Omega)} \leq ||f_n - g_n||_{H_0^1(\Omega)} \leq ||f_n - f||_{H_0^1(\Omega)} + ||f - g_n||_{H_0^1(\Omega)}$$

which can also be made arbitrarily small. Thus, $g = \mathcal{P}_+(f)$. □

The motivation for the construction of $\mathcal{P}_+$ is to be able to "project" any Sobolev function into the admissible set of solutions $\mathcal{A}$, but note that this is not a true projection of a vector space onto a subspace since $\mathcal{P}_+$ is not linear nor is $\mathcal{A}$ a subspace.

Lemma 2.1.5. *Suppose $p^* \in \mathcal{A}$ is the minimizer of (2.1.6). Let $v = -\dfrac{\nabla E(p^*)}{||\nabla E(p^*)||}$ be the direction of steepest descent. Then $\mathcal{P}_+(v) = 0$. In addition, if $h \leq \dfrac{2}{M}$, then*

$$p^* = \mathcal{P}_+(p^* - \tau \nabla E(p^*)).$$

for small enough $\tau > 0$.

Proof. Suppose $v_1 = \mathcal{P}_+(v) \in \mathcal{A}$ is nonzero. $\langle \nabla E(p^*), v_1 \rangle$ is either positive, negative or zero.

If $\langle \nabla E(p^*), v_1 \rangle < 0$, that would contradict the fact that p^* is the minimizer since $E(p^* + \tau v_1) < E(p^*)$ for small enough τ.

If $\langle \nabla E(p^*), v_1 \rangle > 0$, let $v_2 = \mathcal{P}_-(v)$ so that $v = v_1 + v_2$. Then

$$\langle \nabla E(p^*), v \rangle = \langle \nabla E(p^*), v_1 \rangle + \langle \nabla E(p^*), v_2 \rangle > \langle \nabla E(p^*), v_2 \rangle.$$

According to Lemma 2.1.4, $||v_2||_{H_0^1(\Omega)} \leq ||v||_{H_0^1(\Omega)}$, making v_2 a vector of unit length or smaller, which would contradict the fact that v is the direction of steepest descent.

We conclude that $\langle \nabla E(p^*), v_1 \rangle = 0$, making v_2 the direction of steepest descent which is positive nowhere and thus $\mathcal{P}_+(v_2) = 0$.

To prove the second conclusion, we claim that $E(\mathcal{P}_+(p)) \leq E(p)$ for all $p \in H_0^1(\Omega)$. Consequently, $E(\mathcal{P}_+(p^* - \tau \nabla E(p^*))) \leq E(p^* - \tau \nabla E(p^*))$. But if τ is small enough, moving in the direction of steepest descent will cause the energy to decrease. Hence,

$$E(\mathcal{P}_+(p^* - \tau \nabla E(p^*))) \leq E(p^* - \tau \nabla E(p^*)) \leq E(p^*).$$

If the inequality is strict, that would contradict that p^* is the minimizer. If we have equality, then by Theorem 2.1.3 the conclusion follows.

Now we prove the claim. Let $\omega = p^{-1}(-\infty, 0) = \{x \in \Omega : p(x) \leq 0\}$ and split each energy functional into integrals over ω and over $\Omega \backslash \omega$. The integrals over $\Omega \backslash \omega$ are identical, so we focus on ω. Note that $G(0) = 0$ by definition. We will also need

the following estimate.

$$G(p) = \int_0^p \xi F_1(\xi)\, d\xi \le M \int_0^p \xi\, d\xi = \frac{M}{2} p^2$$
$$\int_\omega G(p)\, d\mathbf{x} \le \frac{M}{2} \int_\omega p^2\, d\mathbf{x}$$

Recall that we assumed $\hat{p} \ge 0$. Then

$$\int_\omega p^2\, d\mathbf{x} + h \int_\omega D(x)|\nabla p|^2\, d\mathbf{x} - h \int_\omega G(p)\, d\mathbf{x} - \int_\omega \hat{p} p$$
$$\ge \int_\omega p^2\, d\mathbf{x} - \frac{hM}{2} \int_\omega p^2\, d\mathbf{x} = \left(1 - \frac{hM}{2}\right) \int_\omega p^2\, d\mathbf{x} \ge 0$$

while

$$\int_\omega \mathcal{P}_+(p)^2\, d\mathbf{x} + h \int_\omega D(x)|\nabla \mathcal{P}_+(p)|^2\, d\mathbf{x} - h \int_\omega G(\mathcal{P}_+(p))\, d\mathbf{x} - \int_\omega \hat{p} \mathcal{P}_+(p) = 0$$

since $\mathcal{P}_+(p) \equiv 0$ on ω. Thus, $E(\mathcal{P}_+(p)) \le E(p)$.

□

Theorem 2.1.4. *A function $p \in \mathcal{A}$ is the minimizer of (2.1.6) if and only if p is a discrete weak solution to (2.1.5).*

Proof. It is clear that when $p \in \mathcal{A}$ is a discrete weak solution (2.1.5), we have $\langle \nabla E(p), q \rangle = 0$. Hence, we have

$$\langle \nabla E(p), q - p \rangle \ge 0, \quad \forall q \in \mathcal{A}. \tag{2.1.14}$$

Thus, p is a minimizer of (2.1.6).

On the other hand, as seen from Theorem 2.1.3, there is a unique minimizer p^* of (2.1.6). We can use the standard projected gradient method to find p^* which is given as follows.

Algorithm 2.1.1. *Starting with $P^1 = \hat{p} \in \mathcal{A}$, we iteratively compute $\tilde{P}^{k+1}$*

$$\tilde{P}^{k+1} = P^k - \tau \nabla E(P^k) \tag{2.1.15}$$

and find $P^{k+1} = \mathcal{P}_+(\tilde{P}^{k+1})$ for $k = 1, \cdots$, until the consecutive error $\|P^{k+1} - P^k\|_2$ is within a tolerance, where $\tau > 0$ is a step size.

Using Lemma 2.1.5 and the inequality below

$$\int_\Omega |\mathcal{P}_+(p) - \mathcal{P}_+(q)|^2 d\mathbf{x} \leq \int_\Omega |p - q|^2 d\mathbf{x}, \quad \forall p, q \in L^2(\Omega). \tag{2.1.16}$$

we conclude that

$$\begin{aligned}
&\|P^{k+1} - p^*\|^2_{L^2(\Omega)} = \int_\Omega |\mathcal{P}_+(\tilde{P}^{k+1}) - \mathcal{P}_+(p^*)|^2 d\mathbf{x} \\
&\leq \int_\Omega |\tilde{P}^{k+1} - (p^* - \tau \nabla E(p^*))|^2 d\mathbf{x} \\
&= \int_\Omega |P^k - p^* - \tau(\nabla E(P^k) - \nabla E(p^*))|^2 d\mathbf{x} \\
&= \|P^k - p^*\|^2_{L^2(\Omega)} - 2\tau \int_\Omega (P^k - p^*)(\nabla E(P^k) - \nabla E(p^*)) d\mathbf{x} \\
&+ \tau^2 \|\nabla E(P^k) - \nabla E(p^*)\|^2_{L^2(\Omega)} \\
&\leq \|P^k - p^*\|^2_{L^2(\Omega)} (1 - 2\tau\mu + \tau^2 L^2)
\end{aligned}$$

where μ is the constant in Lemma 2.1.3 and L is the Lipschitz constant of ∇E. As long as $\tau < 2\mu/(L^2)$, we have $\rho = 1 - 2\tau\mu + \tau^2 L^2 < 1$. For example, $\tau = \mu/L^2$ is a good choice. Thus, it follows that $P^k, k \geq 1$ are a Cauchy sequence and converge to p^* in $L^2(\Omega)$ norm.

Furthermore, we can consider $\|\tilde{P}^{k+1} - \tilde{P}^{\ell+1}\|^2_{L^2(\Omega)}$ and use the above analysis to conclude that $\tilde{P}^k, k \geq 1$ are a Cauchy sequence in $L^2(\Omega)$ norm and hence, converge

to a limit by $\tilde{P}^*$. It follows that $p^* = \mathcal{P}_+(\tilde{P}^*)$ almost every where in Ω since

$$\begin{aligned}\|p^* - \mathcal{P}_+(\tilde{P}^*)\|^2_{L^2(\Omega)} &\leq \|p^* - P^k\|^2_{L^2(\Omega)} + \|P^k - \mathcal{P}_+(\tilde{P}^*)\|^2_{L^2(\Omega)} \\ &= \|p^* - P^k\|^2_{L^2(\Omega)} + \|\mathcal{P}_+(\tilde{P}^k) - \mathcal{P}_+(\tilde{P}^*)\|^2_{L^2(\Omega)} \\ &\leq \|p^* - P^k\|^2_{L^2(\Omega)} + \|\tilde{P}^k - \tilde{P}^*\|^2_{L^2(\Omega)} \to 0\end{aligned}$$

when $k \to \infty$.

We now claim that $\tilde{P}^* \in \mathcal{A}$. Otherwise, let $\omega = \{(x, y) \in \Omega, \tilde{P}^* < 0\}$. If ω is an open set with a positive measure, we can choose a function $q \in H^1_0(\Omega)$ such that $q = 1$ in an interior of ω and 0 outside of ω such that

$$\begin{aligned}\int_\Omega \frac{p^* - \tilde{P}^*}{\tau} q d\mathbf{x} &= \langle \nabla E(p^*), q\rangle \qquad (2.1.17)\\ &= \int_\Omega p^* q\mathbf{x} + h\int_\Omega D(\mathbf{x})\nabla p^* \cdot \nabla q d\mathbf{x} - h\int_\Omega p^* F_1(p^*) q d\mathbf{x} - \int_\Omega \hat{p} q d\mathbf{x} \\ &= \int_\omega p^* q\mathbf{x} + h\int_\omega D(\mathbf{x})\nabla p^* \cdot \nabla q d\mathbf{x} - h\int_\omega p^* F_1(p^*) q d\mathbf{x} - \int_\omega \hat{p} q d\mathbf{x}\end{aligned}$$

It follows that

$$0 < \int_\omega \frac{-\tilde{P}^*}{\tau} d\mathbf{x} = -\int_\omega \hat{p} q d\mathbf{x} \leq 0 \qquad (2.1.18)$$

which is a contradiction as $\hat{p} \geq 0$. If ω is not an open set, we can choose an open set $\widetilde{\omega}$ containing ω such that the measure of $\widetilde{\omega}\backslash\omega$ is arbitrarily close to zero. We shall have an equality similar to (2.1.17) with $\widetilde{\omega}$ replacing ω and use a $q \in H^1_0(\Omega)$ which is 1 on an interior of $\widetilde{\omega}$ while zero out of $\widetilde{\omega}$. As $p^* \in H^1_0(\Omega)$, the terms on the right-hand side of the modified version of (2.1.17) can be arbitrarily small except for the last term, i.e. $-\int_{\widetilde{\omega}} \hat{p} q d\mathbf{x}$ while the left-hand side term is $\int_{\widetilde{\omega}} \frac{-\tilde{P}^*}{\tau} q d\mathbf{x} > 0$. These show that ω has to be of zero measure. Hence, $\tilde{P}^* \geq 0$ almost everywhere in Ω. That is, $p^* = \mathcal{P}_+(\tilde{P}^*) = \tilde{P}^*$. From (2.1.15), it follows that $\langle \nabla E(p^*), q\rangle = 0, \quad \forall q \in H^1_0(\Omega)$. $\square$

Remark 2.1.1. *Theorem 2.1.4 implies that there exists a unique discrete weak solu-*

tion to (2.1.5).

Lemma 2.1.6. *The minimizer p^* of the energy functional (2.1.6), hereby denoted by $E_{\hat{p}}$, is stable with respect to perturbations in $\hat{p}$. In particular, if we let q^* be the minimizer associated with the energy functional*

$$E_{\hat{q}}(q) = \int_\Omega q^2 \, d\mathbf{x} + h \int_\Omega D(\mathbf{x}) |\nabla q|^2 \, d\mathbf{x} - h \int_\Omega G(q) \, d\mathbf{x} - \int_\Omega \hat{q} q \, d\mathbf{x}$$

then we are assured that

$$||p^* - q^*||_2 \leq \frac{1}{\mu} ||\hat{p} - \hat{q}||_2$$

Proof. Since p^* is the minimizer, we know $\partial E_{\hat{p}}(p^*, \nu) = 0$ for all ν. Similarly, $\partial E_{\hat{q}}(q^*, \nu) = 0$ for all ν. As a result of Lemma 2.1.3 we get the following two inequalities.

$$E_{\hat{p}}(q^*) \geq E_{\hat{p}}(p^*) + \frac{\mu}{2} \|p^* - q^*\|_2^2$$
$$E_{\hat{q}}(p^*) \geq E_{\hat{q}}(q^*) + \frac{\mu}{2} \|p^* - q^*\|_2^2$$

We add the two inequalities. After some cancellation we obtain the following inequality.

$$-\langle \hat{p}, q^* \rangle - \langle \hat{q}, p^* \rangle \geq -\langle \hat{p}, p^* \rangle - \langle \hat{q}, q^* \rangle + \mu \|p^* - q^*\|_2^2$$
$$\langle \hat{p}, p^* - q^* \rangle - \langle \hat{q}, p^* - q^* \rangle \geq \mu \|p^* - q^*\|_2^2$$
$$\langle \hat{p} - \hat{q}, p^* - q^* \rangle \geq \mu \|p^* - q^*\|_2^2$$

We use the Cauchy-Schwarz's inequality to conclude

$$\|p^* - q^*\|_2 \leq \frac{1}{\mu} \|\hat{p} - \hat{q}\|_2$$

which is the desired inequality. □

2.2 Bivariate Spline Approximation of the Discrete Weak Solution

2.2.1 The Discrete Weak Solution in Finite Dimensional Space

So far we have established that there exists a unique discrete weak solution to the problem posed in (2.1.5). Our next goal is to find an approximate solution in a finite-dimensional spline space. That is, we will approximate p and $\hat{p}$ by using the spline space $S_d^r(\triangle)$ defined as follows.

Definition 2.2.1 (Spline Space)**.** Let $\triangle$ be a given triangulation of a domain Ω. Then we define the spline space of smoothness r and degree d over $\triangle$ by,

$$S_d^r(\triangle) = \{s \in C^r(\Omega) \mid s|_T \in \mathcal{P}_d,\ \forall\ T \in \triangle\},$$

where $\mathcal{P}_d$ is the space of polynomials of degree at most d.

We shall denote the basis of this space as $\{\phi_j\}_{1\leq j\leq n}$. We now set out to find $p^* \in S_d^r(\triangle)$ which satisfies the following equation.

$$\int_\Omega pq\ d\mathbf{x} + h\int_\Omega D(\mathbf{x})\nabla p\cdot\nabla q\ d\mathbf{x} = \int_\Omega \hat{p}q\ d\mathbf{x} + h\int_\Omega pF_1(p)q\ d\mathbf{x} \qquad \forall q \in S_d^r(\triangle) \quad (2.2.1)$$

Theorem 2.2.1. *If h is small enough, then there exists $p^* \in S_d^r(\triangle)$ which satisfies equation (2.2.1).*

Proof. The proof of this theorem is constructive and we only give an overview of the construction here. The detail is contained in the rest of this subsection and the next subsection. We first devise an iterative computational scheme. Each iteration

requires solving a simple linear equation, for which we can guarantee the existence of such iterative solution. We then show that this sequence of iterative solutions actually forms a Cauchy sequence. Thus, the sequence converges to a spline in $S_d^r(\triangle)$ which is a finite dimensional, and hence a complete space. Finally, by simply taking limits as the number of iteration goes to infinity, we demonstrate that we get a discrete weak spline solution satisfying (2.2.1). □

Theorem 2.2.2. *The weak solution of (2.2.1) is unique.*

Proof. The proof is analogous to the one in Theorem 2.1.3. Detail is omitted here. □

2.2.2 The Computational Scheme

At each time step t_i, we have to solve the nonlinear problem (2.2.1). Our approach is to linearize the equation using a fixed-point method.

Algorithm 2.2.1. *Writing $\hat{p} = p(\mathbf{x}, i-1)$ or $\hat{p} = p_0(\mathbf{x})$, the initial value, find $p^{(k)}, k \geq 1$ such that*

$$\int_\Omega p^{(k)} q + hD \int_\Omega \nabla p^{(k)} \cdot \nabla q = \langle \hat{p}, q \rangle + h \int_\Omega p^{(k)} F_1\left(p^{(k-1)}\right) q \, d\mathbf{x} \quad \forall q \in S_d^r(\triangle) \quad (2.2.2)$$

for $k = 1, 2, \ldots,$ until a given accuracy for $\|p^{(k)} - p^{(k-1)}\|$ is met.

Remark 2.2.1. *We stated in the outline of the proof for Theorem 2.2.1 that we will show the sequence of $p^{(k)}$ is Cauchy and hence converges to a limit $p^* \in S_d^r(\triangle)$. Note that in (2.2.2), we can take the limit as $k \to \infty$ of both sides and obtain precisely (2.2.1). This requires the use of the Dominated Convergence Theorem and so we prove boundedness of all the iterates in Theorem 2.2.3.*

Lemma 2.2.1. *Given splines $p^{(k-1)}$ and $\hat{p}$, there exists a unique spline solution for $p^{(k)}$ in equation (2.2.2).*

Proof. Let ϕ_j be any spline basis function. Any spline function in $S_d^r(\triangle)$ can be written as $\sum_{i=1}^n c_i\phi_i$. Let ϕ_j be any spline basis function. Let $\vec{c}$ be the vector of coefficients for $p^{(k)}$ and $\vec{p}$ be the vector of coefficients for $\hat{p}$. Define the following matrices.

$$\begin{aligned} M(i,j) &:= \int_\Omega \phi_i\phi_j \, d\mathbf{x} \\ K_D(i,j) &:= \int_\Omega D(\mathbf{x})\nabla\phi_i \cdot \nabla\phi_j \, d\mathbf{x} \\ M_{F_1(p^{(k-1)})}(i,j) &:= \int_\Omega F_1(p^{(k-1)})\phi_i\phi_j \, d\mathbf{x} \end{aligned}$$

Note that all these matrices are symmetric. In addition, M is positive-definite.

We have to solve (2.2.2) for each $q \in S_d^r(\triangle)$, but it's sufficient to solve for each basis spline ϕ_j. Thus, we have n equations and n unknowns in the coefficient vector, which is equivalent to the following linear system.

$$\begin{aligned} M\vec{c} + hK_D\vec{c} &= M\vec{p} + hM_{F_1(p^{(k-1)})}\vec{c} \\ \left(M + hK_D - hM_{F_1(p^{(k-1)})}\right)\vec{c} &= M\vec{p} \end{aligned}$$

Let $L = M + hK_D - hM_{F_1(p^{(k-1)})}$. M is positive-definite and invertible. If h is small enough, L is also invertible. Thus, we can solve for $\vec{c}$, the spline coefficients of $p^{(k)}$. □

Theorem 2.2.3. *If $h < 1/M$, then the successive solutions $p^{(k)}$ of the equation (2.2.2) satisfy*

$$\left|\left|p^{(k)}\right|\right|_2 \leq \frac{1}{1-hM}\left|\left|\hat{p}\right|\right|_2 \tag{2.2.3}$$

$$\left|\left|\nabla p^{(k)}\right|\right|_2 \leq \frac{1}{\sqrt{hK}}\sqrt{\left|\left|p^{(k)}\right|\right|_2\left(\left|\left|\hat{p}\right|\right|_2 - (1-hM)\left|\left|p^{(k)}\right|\right|_2\right)} \tag{2.2.4}$$

If we substitute the estimate from (2.2.3) into (2.2.4), we obtain a bound which is less sharp but is independent of k.

$$\left|\left|\nabla p^{(k)}\right|\right|_2 \leq \frac{1}{\sqrt{hK}}\sqrt{||p^{(k)}||_2\,||\hat{p}||_2} \leq \frac{1}{\sqrt{hK(1-hM)}}\,||\hat{p}||_2$$

Proof. Substitute $q = p$ into (2.2.2). Then

$$\left|\left|p^{(k)}\right|\right|_2^2 + h\underbrace{\int_\Omega D(\mathbf{x})|\nabla p^{(k)}|^2\,d\mathbf{x}}_{\geq 0} = \langle \hat{p}, p^{(k)}\rangle + h\int_\Omega F_1(p^{(k-1)})(p^{(k)})^2\,d\mathbf{x}$$

Use the Cauchy-Schwarz inequality and the fact that $F_1(p) \leq M$ for any p.

$$\begin{aligned}
\left|\left|p^{(k)}\right|\right|_2^2 &\leq ||\hat{p}||_2\left|\left|p^{(k)}\right|\right|_2 + hM\left|\left|p^{(k)}\right|\right|_2^2\\
\left|\left|p^{(k)}\right|\right|_2 &\leq ||\hat{p}||_2 + hM\left|\left|p^{(k)}\right|\right|_2\\
\left|\left|p^{(k)}\right|\right|_2 &\leq \frac{1}{1-hM}\,||\hat{p}||_2
\end{aligned}$$

Now we prove the bound for $\nabla p^{(k)}$ by substituting $q = p$ once more into (2.2.2).

$$\begin{aligned}
\left|\left|p^{(k)}\right|\right|_2^2 + h\int_\Omega D(\mathbf{x})|\nabla p^{(k)}|^2\,d\mathbf{x} &= \langle \hat{p}, p^{(k)}\rangle + h\int_\Omega F_1(p^{(k-1)})(p^{(k)})^2\,d\mathbf{x}\\
\left|\left|p^{(k)}\right|\right|_2^2 + hK\left|\left|\nabla p^{(k)}\right|\right|_2^2 &\leq ||\hat{p}||_2\left|\left|p^{(k)}\right|\right|_2 + hM\left|\left|p^{(k)}\right|\right|_2^2\\
hK\left|\left|\nabla p^{(k)}\right|\right|_2^2 &\leq ||\hat{p}||_2\left|\left|p^{(k)}\right|\right|_2 - \left|\left|p^{(k)}\right|\right|_2^2 + hM\left|\left|p^{(k)}\right|\right|_2^2\\
hK\left|\left|\nabla p^{(k)}\right|\right|_2^2 &\leq \left|\left|p^{(k)}\right|\right|_2\left(||\hat{p}||_2 - (1-hM)\left|\left|p^{(k)}\right|\right|_2\right)\\
\left|\left|\nabla p^{(k)}\right|\right|_2 &\leq \frac{1}{\sqrt{hK}}\sqrt{||p^{(k)}||_2\left(||\hat{p}||_2 - (1-hM)\,||p^{(k)}||_2\right)}
\end{aligned}$$

□

Remark 2.2.2. *The constant in the bound for $\nabla p^{(k)}$, which can be found under the square root, is non-negative as a result of the bound for $p^{(k)}$. In fact, it can be very close to zero.*

Remark 2.2.3. *Since we are now working within a finite-dimensional space, all norms are equivalent. As a result, we have just established that p and its derivatives are bounded functions. That is,*

$$\left|\left|p^{(k)}\right|\right|_\infty \leq \frac{C}{1-hM} \left|\left|\hat{p}\right|\right|_2$$

Theorem 2.2.4. *If h is small enough so that*

$$hL\frac{C}{(1-hM)^2} \left|\left|\hat{p}\right|\right|_2 < 1$$

where C is the constant from Remark 2.2.3, then successive iterates in Algorithm (2.2.1) are Cauchy in $L^2(\Omega)$. That is,

$$\left|\left|p^{(k)} - p^{(k-1)}\right|\right|_2 \leq \alpha \left|\left|p^{(k-1)} - p^{(k-2)}\right|\right|_2$$

where $0 < \alpha < 1$.

Proof. Take two successive solutions which satisfy the following equations.

$$\begin{aligned}
\int_\Omega p^{(k)} q \, d\mathbf{x} + h \int_\Omega D(\mathbf{x}) \nabla p^{(k)} \cdot \nabla q \, d\mathbf{x} &= \int_\Omega \hat{p} q \, d\mathbf{x} \\
&\quad + h \int_\Omega p^{(k)} F_1(p^{(k-1)}) q \, d\mathbf{x} \\
\int_\Omega p^{(k-1)} q \, d\mathbf{x} + h \int_\Omega D(\mathbf{x}) \nabla p^{(k-1)} \cdot \nabla q \, d\mathbf{x} &= \int_\Omega \hat{p} q \, d\mathbf{x} \\
&\quad + h \int_\Omega p^{(k-1)} F_1(p^{(k-2)}) q \, d\mathbf{x}
\end{aligned}$$

Subtract the two equations and substitute $q = p^{(k)} - p^{(k-1)}$.

$$\left|\left|p^{(k)} - p^{(k-1)}\right|\right|_2^2 + h \underbrace{\int_\Omega D(\mathbf{x}) |\nabla p^{(k)} - \nabla p^{(k-1)}|^2 \, d\mathbf{x}}_{\geq 0}$$
$$= h \int_\Omega \left(F_1(p^{(k-1)})p^{(k)} - F_1(p^{(k-2)})p^{(k-1)}\right)\left(p^{(k)} - p^{(k-1)}\right) d\mathbf{x}$$

Add and subtract $F_1(p^{(k-1)})$ and rearrange.

$$\left|\left|p^{(k)} - p^{(k-1)}\right|\right|_2^2 \leq h \int_\Omega F_1(p^{(k-1)}) \left(p^{(k)} - p^{(k-1)}\right)^2$$
$$+ h \int_\Omega \left(F_1(p^{(k-1)}) - F_1(p^{(k-2)})\right) p^{(k-1)}(p^{(k)} - p^{(k-1)}) \, d\mathbf{x}$$

Use remark 2.2.3 to bound $|p^{(k-1)}|$.

$$\left|\left|p^{(k)} - p^{(k-1)}\right|\right|_2^2 \leq hM \left|\left|p^{(k)} - p^{(k-1)}\right|\right|_2^2$$
$$+ h\frac{C}{1-hM} \left|\left|\hat{p}\right|\right|_2 \int_\Omega \left|F_1(p^{(k-1)}) - F_1(p^{(k-2)})\right| \left|p^{(k)} - p^{(k-1)}\right| \, d\mathbf{x}$$

Group like terms.

$$(1-hM) \left|\left|p^{(k)} - p^{(k-1)}\right|\right|_2^2$$
$$\leq h\frac{C}{1-hM} \left|\left|\hat{p}\right|\right|_2 \int_\Omega \left|F_1(p^{(k-1)}) - F_1(p^{(k-2)})\right| \left|p^{(k)} - p^{(k-1)}\right| \, d\mathbf{x}$$

$F_1(p)$ is assumed to be Lipschitz continuous with constant L_F.

$$\leq hL_F\frac{C}{1-hM} \left|\left|\hat{p}\right|\right|_2 \int_\Omega \left|p^{(k-1)} - p^{(k-2)}\right| \left|p^{(k)} - p^{(k-1)}\right| \, d\mathbf{x}$$

Apply the Cauchy-Schwartz inequality.

$$(1-hM)\left|\left|p^{(k)}-p^{(k-1)}\right|\right|_2^2 \leq hL_F\frac{C}{1-hM}\left|\left|\hat{p}\right|\right|_2\left|\left|p^{(k-1)}-p^{(k-2)}\right|\right|_2\left|\left|p^{(k)}-p^{(k-1)}\right|\right|_2$$

$$\left|\left|p^{(k)}-p^{(k-1)}\right|\right|_2 \leq hL_F\frac{C}{(1-hM)^2}\left|\left|\hat{p}\right|\right|\left|\left|p^{(k-1)}-p^{(k-2)}\right|\right|_2$$

We can choose a small enough h so that $\alpha = hL\frac{C}{(1-hM)^2}\left|\left|\hat{p}\right|\right|$ satisfies $0<\alpha<1$.

□

2.2.3 Bivariate Spline Approximation to the Discrete Weak Solution in Sobolev Space

In this subsection, we show that the spline solutions obtained above are a good approximation to the weak solution in (2.1.5). Let p^* be the weak solution of (2.1.5) and let S^* be the spline solution which is the limit of the iterative solutions from Algorithm 2.2.1. By using Lemma 2.1.3 and noting that $\nabla E(p^*,q)=0$ for any $q\in H_0^1(\Omega)$, we have

$$E(S^*)-E(p^*)\geq\frac{\mu}{2}\left|\left|S^*-p^*\right|\right|_2^2 \tag{2.2.5}$$

Let S_{p^*} be the quasi-interpolant of p^* in the spline space $S_d^r(\triangle)$ as in the Appendix. Since S^* is the minimizer of (2.1.6) with respect to all $q\in S_d^r(\triangle)$, we conclude that $E(S_{p^*})>E(S^*)$. Together with (2.2.5) we can write

$$\frac{\mu}{2}\left|\left|S^*-p^*\right|\right|_2^2\leq E(S_{p^*})-E(p^*) \tag{2.2.6}$$

Theorem 2.2.5. *Suppose that $h>0$ is small enough and p^*, the weak solution of (2.1.5), is in $H^{m+1}(\Omega)$ with $m\geq 1$. Then S^*, the limit of the iterative solutions from Algorithm 2.2.1, approximates p^* in the following sense:*

$$\left|\left|S^*-p^*\right|\right|_2\leq C|\triangle|^m|p^*|_{m+1,2,\Omega} \tag{2.2.7}$$

where C is a constant.

Proof. We rewrite equation (2.2.6)

$$\begin{aligned}\frac{\mu}{2}\left|\left|S^*-p^*\right|\right|_2^2 \leq & \int_\Omega S_{p^*}^2-(p^*)^2\,d\mathbf{x}+h\int_\Omega D(\mathbf{x})\left(|\nabla S_{p^*}|^2-|\nabla p^*|^2\right)\,d\mathbf{x}\\ &+h\int_\Omega G(p^*)-G(S_{p^*})\,d\mathbf{x}\\ =&\int_\Omega (S_{p^*}-p^*)(S_{p^*}+p^*)\,d\mathbf{x}\\ &+h\int_\Omega D(\mathbf{x})(\nabla S_{p^*}-\nabla p^*)\cdot(\nabla S_{p^*}+\nabla p^*)\,d\mathbf{x}\\ &+h\int_\Omega G(p^*)-G(S_{p^*})\,d\mathbf{x}\end{aligned}$$

G is a differentiable function by construction. Since $p^* \in H^2(\Omega)$, by Theorem 1.3.2 we conclude that p^* is Hölder continuous and hence it has some maximal value M^* on the compact set $\overline{\Omega}$. Analogously, we can conclude the same for S_{p^*}. As a result, $G'(p)$ has a maximum value on the compact set $[0, M^*]$ and so G is Lipschitz continuous with some constant L_G. Continuing where we left off above, we use the Cauchy-Schwarz inequality and L_G:

$$\begin{aligned}\leq&\left|\left|S_{p^*}-p^*\right|\right|_2\left|\left|S_{p^*}+p^*\right|\right|_2\\ &+hK_2\left|\left|\nabla S_{p^*}-\nabla p^*\right|\right|_2\left|\left|\nabla S_{p^*}+\nabla p^*\right|\right|_2\\ &+hL_G\int_\Omega |p^*-S_{p^*}|\,d\mathbf{x}\\ \leq&C_1\left|\left|S_{p^*}-p^*\right|\right|_2+hK_2C_2\left|\left|\nabla S_{p^*}-\nabla p^*\right|\right|_2\\ &+hL_G|\Omega|^{1/2}\left|\left|p^*-S_{p^*}\right|\right|_2\end{aligned}$$

where $C_1 = \left|\left|S_{p^*}\right|\right|_2 + \left|\left|p^*\right|\right|_2$, $C_2 = \left|\left|\nabla S_{p^*}\right|\right|_2 + \left|\left|\nabla p^*\right|\right|_2$.

By the approximation property of nonnegative preserving interpolatory splines, Theorem 2.3 in [22] and the standard approximation property of spline spaces, i.e. Theorem 6.0.1 together with the Markov inequality, i.e. Theorem 6.0.2 as in Ap-

pendix, we then write

$$||S_{p^*} - p^*||_2 \leq C_3|\triangle|^2|p^*|_{2,2,\Omega}$$
$$||\nabla S_{p^*} - \nabla p^*||_2 \leq C_4|\triangle||p^*|_{2,2,\Omega}$$

where $|\triangle|$ is the length of the longest edge in the triangulation and C_3 and C_4 are constants independent of p^*. □

As a corollary, we have that $E(S_{p^*}) - E(p^*) \to 0$ as $|\triangle| \to 0$.

Chapter 3

Multiple Interacting Species

We now shift our attention to finding solutions to the system modeling two interacting population densities as presented in equation (1.1.2).

In order to define a weak formulation of the PDE, we define the following set of admissible solutions

$$\mathcal{A} = \{(p, m) \in H_0^1(\Omega) \times H_0^1(\Omega) \mid p(\mathbf{x}) \geq 0, m(\mathbf{x}) \geq 0 \text{ for a.e. } \mathbf{x} \in \Omega\}.$$

The set consists of the subset of functions in the standard Sobolev space with trace zero which satisfy the nonnegative condition almost everywhere. Let $\Omega \subset \mathbb{R}^2$ be an open, bounded domain with Lipschitz boundary.

3.1 Discrete Weak Formulation

Suppose p, m are classical solutions to equation (1.1.2). Then for any $q \in H_0^1(\Omega)$, p and m satisfy the following weak formulation obtained by integrating by parts.

$$\int_\Omega \frac{\partial p(\mathbf{x},t)}{\partial t} q(\mathbf{x}) d\mathbf{x} = -\int_\Omega D(\mathbf{x}) \nabla p(\mathbf{x},t) \cdot \nabla q(\mathbf{x}) d\mathbf{x} + \int_\Omega F(p,m) q(\mathbf{x}) d\mathbf{x} \tag{3.1.1}$$

$$\int_\Omega \frac{\partial m(\mathbf{x},t)}{\partial t} q(\mathbf{x}) d\mathbf{x} = -\int_\Omega E(\mathbf{x}) \nabla m(\mathbf{x},t) \cdot \nabla q(\mathbf{x}) d\mathbf{x} + \int_\Omega G(p,m) q(\mathbf{x}) d\mathbf{x} \tag{3.1.2}$$

Consider $t \in [0,T]$ and partition $0 = t_0 < t_1 < t_2 < \cdots < t_m < t_{m+1} = T$. We approximate $\frac{dp(\mathbf{x},t)}{dt}$ and $\frac{dm(\mathbf{x},t)}{dt}$ by its divided difference, i.e.,

$$\frac{dp(\mathbf{x},t_i)}{dt} \approx \frac{p(\mathbf{x},t_i) - p(\mathbf{x},t_{i-1})}{h}$$

with $h = t_i - t_{i-1}$. Substitute this approximation into (3.1.1) and (3.1.2) to obtain

$$\begin{aligned} &\int_\Omega p_h(\mathbf{x},t_i) q(\mathbf{x}) d\mathbf{x} + h \int_\Omega D(\mathbf{x}) \nabla p_h(\mathbf{x},t_i) \cdot \nabla q(\mathbf{x}) d\mathbf{x} \\ &- h \int_\Omega F(p_h(\mathbf{x},t_i), m_h(\mathbf{x},t_i)) q(\mathbf{x}) d\mathbf{x} = \int_\Omega p_h(\mathbf{x},t_{i-1}) q(\mathbf{x}) d\mathbf{x} \end{aligned} \tag{3.1.3}$$

$$\begin{aligned} &\int_\Omega m_h(\mathbf{x},t_i) q(\mathbf{x}) d\mathbf{x} + h \int_\Omega E(\mathbf{x}) \nabla m_h(\mathbf{x},t_i) \cdot \nabla q(\mathbf{x}) d\mathbf{x} \\ &- h \int_\Omega G(p_h(\mathbf{x},t_i), m_h(\mathbf{x},t_i)) q(\mathbf{x}) d\mathbf{x} = \int_\Omega m(\mathbf{x},t_{i-1}) q(\mathbf{x}) d\mathbf{x}. \end{aligned} \tag{3.1.4}$$

Note that the functions p_h and m_h have the subscript h to indicate their dependence on the choice of h; a solution to (3.1.3) is not a solution to (3.1.1) and vice-versa.

Definition 3.1.1. Any pair of functions $(p(\mathbf{x},t_i), m(\mathbf{x},t_i)) \in \mathcal{A}$, which satisfy equations (3.1.3) and (3.1.4) for fixed $h > 0$, t_i and t_{i-1}, are called discrete weak solutions of (1.1.2).

Similar to Theorem 2.1.1, we can guarantee that the discrete weak solutions are good approximations to the exact solutions.

Theorem 3.1.1. *Let $(p(\mathbf{x},t), m(\mathbf{x},t))$ be classical solutions of (1.1.2). Suppose that F and G are Lipschitz continuous. Let $p_h(\mathbf{x},t_i)$ and $m_h(\mathbf{x},t_i)$ for $i = 1,\ldots,m+1$ be the discrete weak solutions with $p_h(\mathbf{x},t_0) = p(\mathbf{x},t_0)$ and $m_h(\mathbf{x},t_0) = m(\mathbf{x},t_0)$. Suppose that $p(\mathbf{x},t)$ and $m(\mathbf{x},t)$ are twice differentiable with respect to t. Then*

$$\int_\Omega |p(\mathbf{x},t_i) - p_h(\mathbf{x},t_i)|^2 d\mathbf{x} + \int_\Omega |m(\mathbf{x},t_i) - m_h(\mathbf{x},t_i)|^2 d\mathbf{x} \le Ch, \quad \forall i = 0,\cdots,m+1, \tag{3.1.5}$$

as $h = T/(m+1) \to 0$, where $C > 0$ is a constant independent of h.

Proof. By Taylor expansion, we have

$$\frac{dp(\mathbf{x},t_i)}{dt} = \frac{p(\mathbf{x},t_i) - p(\mathbf{x},t_{i-1})}{h} + O(h), \tag{3.1.6}$$

where $O(h)$ is a quantity bounded by Ch for a positive constant $C < \infty$. Using (3.1.1) and (3.1.3), we have

$$\begin{aligned}
&\int_\Omega \frac{dp(\mathbf{x},t_i)}{dt} q(\mathbf{x}) d\mathbf{x} - \int_\Omega \frac{p_h(\mathbf{x},t_i) - p_h(\mathbf{x},t_{i-1})}{h} q(\mathbf{x}) d\mathbf{x} \\
=& -\int_\Omega D(\mathbf{x}) \nabla (p(\mathbf{x},t_i) - p_h(\mathbf{x},t_i)) \cdot \nabla q(\mathbf{x}) d\mathbf{x} \\
&+ \int_\Omega (F(p(\mathbf{x},t_i), m(\mathbf{x},t_i)) - F(p_h(\mathbf{x},t_i), m_h(\mathbf{x},t_i)) q(\mathbf{x}) d\mathbf{x}.
\end{aligned}$$

Substitute (3.1.6) to obtain

$$\begin{aligned}
&\int_\Omega \frac{p(\mathbf{x},t_i) - p(\mathbf{x},t_{i-1})}{h} q(\mathbf{x}) d\mathbf{x} - \int_\Omega \frac{p_h(\mathbf{x},t_i) - p_h(\mathbf{x},t_{i-1})}{h} q(\mathbf{x}) d\mathbf{x} \\
=& O(h) - \int_\Omega D(\mathbf{x}) \nabla (p(\mathbf{x},t_i) - p_h(\mathbf{x},t_i)) \cdot \nabla q(\mathbf{x}) d\mathbf{x} \\
&+ \int_\Omega (F(p(\mathbf{x},t_i), m(\mathbf{x},t_i)) - F(p_h(\mathbf{x},t_i), m_h(\mathbf{x},t_i)) q(\mathbf{x}) d\mathbf{x}.
\end{aligned}$$

Letting $q = p(\mathbf{x}, t_i) - p_h(\mathbf{x}, t_i)$ in the above inequality, we obtain

$$\begin{aligned}
&\int_\Omega |p(\mathbf{x}, t_i) - p_h(\mathbf{x}, t_i)|^2 d\mathbf{x} \\
=&\int_\Omega (p(\mathbf{x}, t_{i-1}) - p_h(\mathbf{x}, t_{i-1})(p(\mathbf{x}, t_i) - p_h(\mathbf{x}, t_i)) + O(h^2) \\
&- h\int_\Omega D(\mathbf{x})|\nabla(p(\mathbf{x}, t_i) - p_h(\mathbf{x}, t_i))|^2 d\mathbf{x} \\
&+ h\int_\Omega (F(p(\mathbf{x}, t_i), m(\mathbf{x}, t_i)) - F(p_h(\mathbf{x}, t_i), m_h(\mathbf{x}, t_i))(p(\mathbf{x}, t_i) - p_h(\mathbf{x}, t_i))d\mathbf{x}
\end{aligned}$$

Discard the positive gradient term and use Lemma 1.3.1 with $\alpha = 1$.

$$\begin{aligned}
\leq&\frac{1}{2}\int_\Omega |p(\mathbf{x}, t_i) - p_h(\mathbf{x}, t_i)|^2 d\mathbf{x} + \frac{1}{2}\int_\Omega |p(\mathbf{x}, t_{i-1}) - p_h(\mathbf{x}, t_{i-1})|^2 d\mathbf{x} + O(h^2) \\
&+ h\int_\Omega (F(p(\mathbf{x}, t_i), m(\mathbf{x}, t_i)) - F(p_h(\mathbf{x}, t_i), m_h(\mathbf{x}, t_i))(p(\mathbf{x}, t_i) - p_h(\mathbf{x}, t_i))d\mathbf{x}
\end{aligned}$$

Since F is Lipschitz continuous, i.e. $|F(p, m) - F(q, n)| \leq L\sqrt{|p - q|^2 + |m - n|^2}$, it follows that

$$\begin{aligned}
&\int_\Omega |p(\mathbf{x}, t_i) - p_h(\mathbf{x}, t_i)|^2 d\mathbf{x} \\
&\leq \int_\Omega |p(\mathbf{x}, t_{i-1}) - p_h(\mathbf{x}, t_{i-1})|^2 d\mathbf{x} + O(2h^2) \\
&+ 2hL\int_\Omega \sqrt{|p(\mathbf{x}, t_i) - p_h(\mathbf{x}, t_i)|^2 + |m(\mathbf{x}, t_i) - m_h(\mathbf{x}, t_i)|^2}(p(\mathbf{x}, t_i) - p_h(\mathbf{x}, t_i))d\mathbf{x}
\end{aligned}$$

Use Lemma 1.3.1 with $\alpha = 1$ again.

$$\begin{aligned}
&\leq \int_\Omega |p(\mathbf{x}, t_{i-1}) - p_h(\mathbf{x}, t_{i-1})|^2 d\mathbf{x} + O(2h^2) \\
&+ 2hL\int_\Omega |p(\mathbf{x}, t_i) - p_h(\mathbf{x}, t_i)|^2 + \frac{1}{2}|m(\mathbf{x}, t_i) - m_h(\mathbf{x}, t_i)|^2 d\mathbf{x}
\end{aligned}$$

Rearrange the inequality to obtain

$$(1-2hL)\int_\Omega |p(\mathbf{x},t_i)-p_h(\mathbf{x},t_i)|^2 d\mathbf{x} - hL\int_\Omega |m(\mathbf{x},t_i)-m_h(\mathbf{x},t_i)|^2 d\mathbf{x}$$
$$\leq \int_\Omega |p(\mathbf{x},t_{i-1})-p_h(\mathbf{x},t_{i-1})|^2 d\mathbf{x} + O(h^2).$$

Similarly, we can derive

$$(1-2hL)\int_\Omega |m(\mathbf{x},t_i)-m_h(\mathbf{x},t_i)|^2 d\mathbf{x} - hL\int_\Omega |p(\mathbf{x},t_i)-p_h(\mathbf{x},t_i)|^2 d\mathbf{x}$$
$$\leq \int_\Omega |m(\mathbf{x},t_{i-1})-m_h(\mathbf{x},t_{i-1})|^2 d\mathbf{x} + O(h^2).$$

Adding the two inequalities together, we obtain

$$(1-3hL)\left[\int_\Omega |p(\mathbf{x},t_i)-p_h(\mathbf{x},t_i)|^2 d\mathbf{x} + \int_\Omega |m(\mathbf{x},t_i)-m_h(\mathbf{x},t_i)|^2 d\mathbf{x}\right] \tag{3.1.7}$$
$$\leq \int_\Omega |p(\mathbf{x},t_{i-1})-p_h(\mathbf{x},t_{i-1})|^2 d\mathbf{x} + \int_\Omega |m(\mathbf{x},t_{i-1})-m_h(\mathbf{x},t_{i-1})|^2 d\mathbf{x} + O(h^2).$$

Letting $\alpha = 1/(1-3hL)$ and

$$e_i = \int_\Omega |p(\mathbf{x},t_i)-p_h(\mathbf{x},t_i)|^2 d\mathbf{x} + \int_\Omega |m(\mathbf{x},t_i)-m_h(\mathbf{x},t_i)|^2 d\mathbf{x},$$

we multiply α on the both sides above and then repeatedly apply the inequality for $i = k, \ldots, 1$ to obtain

$$e_k \leq \alpha e_{k-1} + O(h^2) \leq \cdots\cdots$$
$$\leq \alpha^k e_0 + O(h^2)\sum_{i=0}^{k-1}\alpha^i$$

Note that $e_0 = 0$ since the same boundary conditions apply to p and p_h and to m and m_h.

$$\leq O(h^2)\frac{\alpha^m}{\alpha - 1} = O(h)$$

The final step in the inequality is analogous to the one proved in Theorem 2.1.1. $\square$

Now that we've defined the discrete weak solution and seen guarantees that it is a good approximation to a classical solution, we seek an approach to finding such a discrete weak solution. The approach uses many of the techniques from Chapter 2 for the case of a single species.

In order to simplify the notation, we will use the shortened notation $p = p_h(\mathbf{x}, t_i)$ and $\hat{p} = p_h(\mathbf{x}, t_{i-1})$. Similarly, $m = m_h(\mathbf{x}, t_i)$ and $\hat{m} = p_h(\mathbf{x}, t_{i-1})$. We suppress the time step size h, the spatial variable and all times t_i other than the current and previous ones, unless there is a specific need to pay attention to them. Thus, a much more concise description of the discrete weak formulation is.

$$\int_\Omega pq \, d\mathbf{x} + h\int_\Omega D(\mathbf{x})\nabla p \cdot \nabla q \, d\mathbf{x} - h\int_\Omega F(p,m)q \, d\mathbf{x} = \int_\Omega \hat{p}q \, d\mathbf{x} \tag{3.1.8}$$

$$\int_\Omega mq \, d\mathbf{x} + h\int_\Omega E(\mathbf{x})\nabla m \cdot \nabla q \, d\mathbf{x} - h\int_\Omega G(p,m)q \, d\mathbf{x} = \int_\Omega \hat{m}q d\mathbf{x}. \tag{3.1.9}$$

Since $p_h(\mathbf{x}, t_0)$ and $m_h(\mathbf{x}, t_0)$ are given initial conditions of the PDE, we can assume that $\hat{p}$ and $\hat{m}$ are known and can focus on finding a solution for p, the next time step.

Recall from the introductory chapter that we assume F and G are Lipschitz continuous and bounded above. We must now impose the additional condition that we can write

$$F(p,m) = pF_1(p,m) \text{ and } G(p,m) = mG_1(p,m). \tag{3.1.10}$$

and the factors F_1 and G_1 are also Lipschitz continuous and bounded above. Let $f(p,m) = \int_0^p F(\xi, m)d\xi = \int_0^p \xi F_1(\xi, m)d\xi$. Similar for $g(m,p) = \int_0^m \xi G_1(\xi, p)d\xi$. As

the system of diffusive PDE (1.1.2) is nonlinear, we build a fixed point iteration scheme which linearizes the problem of finding discrete weak solutions p and m. We start with an initial guess $(p^k, m^k) \in \mathcal{A}$ and find $(p^{k+1}, m^{k+1}) \in \mathcal{A}$. We shall show that the sequences $\{p^k, k \geq 1\}$ and $\{m^k, k \geq 1\}$ are Cauchy. The limits will form the discrete weak solution at t_i. Let

$$\mathcal{E}_1(p) := \int_\Omega p^2 \, d\mathbf{x} + h \int_\Omega D(\mathbf{x}) |\nabla p|^2 \, d\mathbf{x} - h \int_\Omega f(p, m_i^k) \, d\mathbf{x} - \int_\Omega \hat{p} p \, d\mathbf{x} \tag{3.1.11}$$

$$\mathcal{E}_2(m) := \int_\Omega m^2 \, d\mathbf{x} + h \int_\Omega E(\mathbf{x}) |\nabla m|^2 \, d\mathbf{x} - h \int_\Omega g(p_i^k, m) \, d\mathbf{x} - \int_\Omega \hat{m} m \, d\mathbf{x} \tag{3.1.12}$$

Initially, we let $p^1 = \hat{p}$ and $m^1 = \hat{m}$ and define a minimization problem.

$$\min_{(p,m) \in \mathcal{A}} \mathcal{E}(p, m) = \min_{(p,m) \in \mathcal{A}} \mathcal{E}_1(p) + \mathcal{E}_2(m). \tag{3.1.13}$$

Then we have the following existence and uniqueness result.

Theorem 3.1.2. *There exists a unique pair $(p^{k+1}, m^{k+1}) \in \mathcal{A}$ which minimizes the energy functional $\mathcal{E}(p, m)$ in (3.1.13).*

Proof. The proof is analogous to the one of Theorem 2.1.2. □

The motivation behind this definition of the energy functionals in 3.1.11 and 3.1.12 is that their Euler-Lagrange equations, computed using Gâteaux derivatives, are given by

$$\int_\Omega p^{k+1} q \, d\mathbf{x} + h \int_\Omega D(\mathbf{x}) \nabla p^{k+1} \cdot \nabla q \, d\mathbf{x} = \int_\Omega \hat{p} q \, d\mathbf{x} + h \int_\Omega F(p^{k+1}, m^k) q \, d\mathbf{x} \tag{3.1.14}$$

$$\int_\Omega m^{k+1} q \, d\mathbf{x} + h \int_\Omega E(\mathbf{x}) \nabla m^{k+1} \cdot \nabla q \, d\mathbf{x} = \int_\Omega \hat{m} q \, d\mathbf{x} + h \int_\Omega G(p^k, m^{k+1}) q \, d\mathbf{x}. \tag{3.1.15}$$

where $q \in H_0^1(\Omega)$ is an arbitrary element, possibly different in each equation. These are analogous to the discrete weak problem from 3.1.8 and 3.1.9 but modified into a fixed-point iteration scheme whose limit solves the discrete weak problem. What we

are not assured, however, is that the minimizer is necessarily a critical point of $\mathcal{E}$, i.e. that it is a solution to the Euler-Lagrange equations, since $\mathcal{A}$ is not an open set. We later prove that the minimizer is indeed a critical point.

Theorem 3.1.3. *Suppose that $(p^{k+1}, m^{k+1}) \in \mathcal{A}$ satisfy equations 3.1.14 and 3.1.15. Then (p^{k+1}, m^{k+1}) is the minimizer of (3.1.13).*

Proof. Consider the constrained minimization problem:

$$\min_{\mathbf{x} \in C} \mathcal{E}(\mathbf{x}), \tag{3.1.16}$$

where $\mathcal{E}(\mathbf{x})$ is a convex function over the convex set $C \subset H$ and H is a Hilbert space. Suppose $\mathcal{E}$ is differentiable. Then any minimizer $\mathbf{w}^*$ of (3.1.16) satisfies

$$\langle \nabla \mathcal{E}(\mathbf{w}^*), \mathbf{x} - \mathbf{w}^* \rangle \geq 0, \quad \forall \mathbf{x} \in C. \tag{3.1.17}$$

On the other hand, if $\mathbf{w}^* \in C$ satisfies (3.1.17), then $\mathbf{w}^*$ is a minimizer of (3.1.16).

We know $(p^{k+1}, m^{k+1}) \in \mathcal{A}$ which is a convex subset of the Hilbert space $H_0^1(\Omega) \times H_0^1(\Omega)$. Then $\mathcal{E}(p, m) = \mathcal{E}_1(p) + \mathcal{E}_2(m)$ is a convex and differentiable function, a fact whose proof is omitted here but can be readily reproduced by consulting the proof of Lemma 2.1.3. Equations 3.1.14 and (3.1.15) imply that $\langle \nabla \mathcal{E}(\mathbf{w}^*), \mathbf{x} \rangle = 0$ for all $\mathbf{x} \in \{(q_1, q_2) \mid q_1, q_1 \in H_0^1(\Omega)\}$, where $\mathbf{w}^* = (p^{k+1}, m^{k+1})$. This in turn implies that the inequality in 3.1.17 holds, and thus (p^{k+1}, m^{k+1}) is a minimizer of (3.1.13). □

Theorem 3.1.4. *There exists at most one pair (p^{k+1}, m^{k+1}) satisfying (3.1.15) if $h > 0$ is small enough.*

Proof. Suppose that there is another pair $(\tilde{p}, \tilde{m})$ satisfying

$$\int_\Omega \tilde{p} q \, d\mathbf{x} + h \int_\Omega D(\mathbf{x}) \nabla \tilde{p} \cdot \nabla q \, d\mathbf{x} = \int_\Omega \hat{p} q \, d\mathbf{x} + h \int_\Omega F(\tilde{p}, m^k) q \, d\mathbf{x} \tag{3.1.18}$$

$$\int_\Omega \tilde{m} q \, d\mathbf{x} + h \int_\Omega E(\mathbf{x}) \nabla \tilde{m} \cdot \nabla q \, d\mathbf{x} = \int_\Omega \hat{m} q \, d\mathbf{x} + h \int_\Omega G(p^k, \tilde{m}) q \, d\mathbf{x} \tag{3.1.19}$$

for all $q \in H_0^1(\Omega)$. Subtract equations 3.1.18 and 3.1.14 to obtain

$$\int_\Omega (p^{k+1} - \tilde{p})q \, d\mathbf{x} + h \int_\Omega D(\mathbf{x}) \nabla (p^{k+1} - \tilde{p}) \cdot \nabla q \, d\mathbf{x} = h \int_\Omega (F(p^{k+1}, m^k) - F(\tilde{p}, m^k))q \, d\mathbf{x}.$$

Letting $q = p^{k+1} - \tilde{p}$, we have

$$\|p^{k+1} - \tilde{p}\|^2 + h\|\sqrt{D}\nabla(p^{k+1} - \tilde{p})\|^2 \leq hL\|p^{k+1} - \tilde{p}\|^2$$
$$(1 - hL)\|p^{k+1} - \tilde{p}\|^2 \leq 0$$

As long as $hL < 1$, we conclude that $p^{k+1} - \tilde{p} \equiv 0$.

An analogous argument for $m^{k+1} - \tilde{m} \equiv 0$ can be made. □

Finally we show

Theorem 3.1.5. *Let $(p^{k+1}, m^{k+1}) \in \mathcal{A}$ be the minimizer of (3.1.13). Then the pair are discrete weak solutions to (3.1.15).*

Proof. As we know, there is a unique minimizer (p^*, m^*) of (3.1.13). We can use the standard projected gradient method to find (p^*, m^*). For convenience, we only discuss how to compute p^* which is given as follows.

Algorithm 3.1.1. *Starting with $P^1 = \hat{p} \in \mathcal{A}$, we iteratively compute $\tilde{P}^{k+1}$*

$$\tilde{P}^{k+1} = P^k - \tau \nabla E(P^k) \tag{3.1.20}$$

and find $P^{k+1} = \mathcal{P}_+(\tilde{P}^{k+1})$ for $k = 1, \cdots,$ until the consecutive error $\|P^{k+1} - P^k\|_2$ is within a tolerance, where $\tau > 0$ is a step size.

Recall Lemma 2.1.5 and

$$\int_\Omega |\mathcal{P}_+(p) - \mathcal{P}_+(q)|^2 d\mathbf{x} \leq \int_\Omega |p - q|^2 d\mathbf{x}, \quad \forall p, q \in L^2(\Omega). \tag{3.1.21}$$

Thus, we have

$$\begin{aligned}
\|P^{k+1}-p^*\|^2_{L^2(\Omega)} &= \int_\Omega |\mathcal{P}_+(\tilde{P}^{k+1})-\mathcal{P}_+(p^*)|^2 d\mathbf{x} \\
&\leq \int_\Omega |\tilde{P}^{k+1}-(p^*-\tau\nabla\mathcal{E}_1(p^*))|^2 d\mathbf{x} \\
&= \int_\Omega |P^k-p^*-\tau(\nabla\mathcal{E}_1(P^k)-\nabla\mathcal{E}_1(p^*))|^2 d\mathbf{x} \\
&= \|P^k-p^*\|^2_{L^2(\Omega)} - 2\tau\int_\Omega (P^k-p^*)(\nabla\mathcal{E}_1(P^k)-\nabla\mathcal{E}_1(p^*))d\mathbf{x} \\
&+ \tau^2\|\nabla\mathcal{E}_1(P^k)-\nabla\mathcal{E}_1(p^*)\|^2_{L^2(\Omega)} \\
&\leq \|P^k-p^*\|^2_{L^2(\Omega)}(1-2\tau\mu+\tau^2L^2)
\end{aligned}$$

where μ is the strong convexity constant and L is the Lipschitz constant of $\nabla\mathcal{E}_1$. As long as $\tau < 2\mu/(L^2)$, we have $\rho = 1-2\tau\mu+\tau^2L^2 < 1$. For example, $\tau = \mu/L^2$ is a good choice. Thus, it follows that $P^k, k \geq 1$ are a Cauchy sequence and converge to p^* in $L^2(\Omega)$ norm.

Furthermore, we can consider $\|\tilde{P}^{k+1}-\tilde{P}^{\ell+1}\|^2_{L^2(\Omega)}$ and use the above analysis to conclude that $\tilde{P}^k, k \geq 1$ are a Cauchy sequence in $L^2(\Omega)$ norm and hence, converge to a limit by $\tilde{P}^*$. It follows that $p^* = \mathcal{P}_+(\tilde{P}^*)$ almost everywhere in Ω since

$$\begin{aligned}
\|p^*-\mathcal{P}_+(\tilde{P}^*)\|^2_{L^2(\Omega)} &\leq \|p^*-P^k\|^2_{L^2(\Omega)} + \|P^k-\mathcal{P}_+(\tilde{P}^*)\|^2_{L^2(\Omega)} \\
&= \|p^*-P^k\|^2_{L^2(\Omega)} + \|\mathcal{P}_+(\tilde{P}^k)-\mathcal{P}_+(\tilde{P}^*)\|^2_{L^2(\Omega)} \\
&\leq \|p^*-P^k\|^2_{L^2(\Omega)} + \|\tilde{P}^k-\tilde{P}^*\|^2_{L^2(\Omega)} \to 0
\end{aligned}$$

when $k \to \infty$. Thus, $p^* = \tilde{P}^* + \tau\nabla\mathcal{E}(p^*)$ for some τ.

We now claim that $\tilde{P}^* \in \mathcal{A}$. Otherwise, let $\omega = \{(x,y) \in \Omega, \tilde{P}^* < 0\}$. If ω is an open set with a positive measure, we can choose a function $q \in H^1_0(\Omega)$ such that

$q = 1$ in an interior of ω and 0 outside of ω such that

$$\begin{aligned}\int_\Omega \frac{p^* - \tilde{P}^*}{\tau} q d\mathbf{x} &= \langle \nabla \mathcal{E}_1(p^*), q\rangle \\ &= \int_\Omega p^* q\mathbf{x} + h\int_\Omega D(\mathbf{x})\nabla p^* \cdot \nabla q d\mathbf{x} - h\int_\Omega p^* F_1(p^*) q d\mathbf{x} - \int_\Omega \hat{p} q d\mathbf{x} \\ &= \int_\omega p^* q\mathbf{x} + h\int_\omega D(\mathbf{x})\nabla p^* \cdot \nabla q d\mathbf{x} - h\int_\omega p^* F_1(p^*) q d\mathbf{x} - \int_\omega \hat{p} q d\mathbf{x}\end{aligned} \tag{3.1.22}$$

It follows that

$$0 < \int_\omega \frac{-\tilde{P}^*}{\tau} d\mathbf{x} = -\int_\omega \hat{p} q d\mathbf{x} \leq 0 \tag{3.1.23}$$

which is a contradiction. If ω is not an open set, we can choose an open set $\widetilde{\omega}$ containing ω such that the measure of $\widetilde{\omega}\backslash\omega$ is arbitrarily close to zero. We shall have an equality similar to (3.1.22) with $\widetilde{\omega}$ replacing ω and use a $q \in H_0^1(\Omega)$ which is 1 on an interior of $\widetilde{\omega}$ while zero out of $\widetilde{\omega}$. As $p^* \in H_0^1(\Omega)$, the terms on the right-hand side of the modified version of (3.1.22) can be arbitrarily small except for the last term, i.e. $-\int_{\widetilde{\omega}} \hat{p} q d\mathbf{x}$ while the left-hand side term is $\int_{\widetilde{\omega}} \frac{-\tilde{P}^*}{\tau} q d\mathbf{x} > 0$. These show that ω has to be of zero measure. Hence, $\tilde{P}^* \geq 0$ almost everywhere in Ω. That is, $p^* = \mathcal{P}_+(\tilde{P}^*) = \tilde{P}^*$. From (3.1.20), it follows that $\langle \nabla E(p^*), q\rangle = 0, \quad \forall q \in H_0^1(\Omega)$. $\square$

Let us design another algorithm to compute (p^{k+1}, m^{k+1}) to ensure the convergence is in $H^1(\Omega)$.

Algorithm 3.1.2. *Let $p^1 = \hat{p}$ and $m^1 = \hat{m}$. For each $n \geq 1$ use Algorithm 3.1.1 to find $\tilde{p}^{n+1}, \tilde{m}^{n+1} \in H_0^1(\Omega)$ such that $\forall q \in H_0^1(\Omega)$ the following equations are satisfied.*

$$\int_\Omega \tilde{p}^{n+1} q \, d\mathbf{x} + h\int_\Omega D(\mathbf{x})\nabla \tilde{p}^{n+1} \cdot \nabla q \, d\mathbf{x} = \int_\Omega \hat{p} q \, d\mathbf{x} + h\int_\Omega F(p^n, m^k) q \, d\mathbf{x} \tag{3.1.24}$$

$$\int_\Omega \tilde{m}^{n+1} q \, d\mathbf{x} + h\int_\Omega E(\mathbf{x})\nabla \tilde{m}^{n+1} \cdot \nabla q \, d\mathbf{x} = \int_\Omega \hat{m} q \, d\mathbf{x} + h\int_\Omega G(p^k, m^n) q \, d\mathbf{x} \tag{3.1.25}$$

Now let $p^{n+1} = \mathcal{P}_{[0,1]}(\tilde{p}^{n+1})$ and $m^{n+1} = \mathcal{P}_{[0,1]}(\tilde{m}^{n+1})$, where $\mathcal{P}_{[0,1]}$ stands for a projec-

tion defined by

$$\mathcal{P}_{[0,1]}(p)(\mathbf{x}) = \begin{cases} 1, & \text{if } p(\mathbf{x}) \geq 1, \mathbf{x} \in \Omega \\ p(\mathbf{x}) & \text{if } 0 < p(\mathbf{x}) < 1, \mathbf{x} \in \Omega \\ 0 & \text{if } p(\mathbf{x}) \leq 0, \mathbf{x} \in \Omega. \end{cases} \tag{3.1.26}$$

We now show that the new sequence $\{p^n, m^n, n \geq 1\}$ converges in $H^1(\Omega)$ norm. Consider the difference of of the first equation in (3.1.24) involving $\tilde{p}^{m+1}$ and $\tilde{p}^m$ and then let $q = \tilde{p}^{m+1} - \tilde{p}^m$. We have

$$\begin{aligned} &\int_\Omega |\tilde{p}^{m+1} - \tilde{p}^m|^2 \, d\mathbf{x} + h \int_\Omega D(\mathbf{x}) |\nabla \tilde{p}^{m+1} - \nabla \tilde{p}^m|^2 \, d\mathbf{x} \\ =& h \int_\Omega (F(p^m, m^k) - F(p^{m-1}, m^k)(\tilde{p}^{m+1} - \tilde{p}^m) \, d\mathbf{x} \\ \leq& hL \int_\Omega |p^m - p^{m-1}||\tilde{p}^{m+1} - \tilde{p}^m| d\mathbf{x} \\ \leq& \frac{hL}{2} \|p^m - p^{m-1}\|^2 + \frac{hL}{2} \|\tilde{p}^{m+1} - \tilde{p}^m\|^2. \end{aligned} \tag{3.1.27}$$

It follows that

$$(1 - hL/2)\|\tilde{p}^{m+1} - \tilde{p}^m\|^2 \leq \frac{hL}{2} \|p^m - p^{m-1}\|^2.$$

We notice that $\|p^{m+1} - p^m\|^2 \leq \|\tilde{p}^{m+1} - \tilde{p}^m\|^2$. Thus, we have

$$(1 - hL/2)\|p^{m+1} - p^m\|^2 \leq \frac{hL}{2} \|p^m - p^{m-1}\|^2.$$

Letting $\alpha = hL/(2 - hL)$, we have

$$\|p^{m+1} - p^m\|^2 \leq \alpha \|p^m - p^{m-1}\|^2 \leq \cdots \leq \alpha^m \|p^2 - p^1\|^2.$$

Thus, p^m is a Cauchy sequence in $L^2(\Omega)$. Furthermore, from (3.1.27, we have

$$(1 - hL/2)\|\tilde{p}^{m+1} - \tilde{p}^m\|^2 + h \int_\Omega D(\mathbf{x}) |\nabla \tilde{p}^{m+1} - \nabla \tilde{p}^m|^2 \leq \frac{hL}{2} \|p^m - p^{m-1}\|^2 \leq C\alpha^m.$$

for a positive constant $C = hL\|p^2 - p^1\|^2$. That is, when $D(\mathbf{x}) \geq K > 0$,

$$(1 - hL/2)\|\tilde{p}^{m+1} - \tilde{p}^m\|^2 + K\|\nabla\tilde{p}^{m+1} - \nabla\tilde{p}^m|^2 \leq \frac{L}{2}\|p^m - p^{m-1}\|^2 \leq C\alpha^m$$

for all $m \geq 1$ and hence, $\tilde{p}^m, m \geq 1$ are a Cauchy sequence in $H^1(\Omega)$ norm. Let p^* be the limit of $p^m, m \geq 1$ and $\tilde{p}^*$ be the limit of $\tilde{p}^m, m \geq 1$. It is easy to see that $\mathcal{P}_{[0,1]}(\tilde{p}^*) = p^*$. Now we let $m \to \infty$ in (3.1.24) to have

$$\int_\Omega \tilde{p}^* q \, d\mathbf{x} + h\int_\Omega D(\mathbf{x})\nabla\tilde{p}^* \cdot \nabla q \, d\mathbf{x} = \int_\Omega \hat{p} q \, d\mathbf{x} + h\int_\Omega F(p^*, m^k) q \, d\mathbf{x} \tag{3.1.28}$$

for all $q \in H_0^1(\Omega)$. If $\tilde{p}^* \in [0,1]$, then $p^* = \tilde{p}^*$ and hence, (3.1.28) shows p^* is a minimizer, i.e, $\langle \nabla\mathcal{E}_1(p^*), q\rangle = 0$ for all $q \in H_0^1(\Omega)$. On the other hand, if p^* is the minimizer, we have

$$\langle \nabla\mathcal{E}_1(p^*), q - p^*\rangle \geq 0, \quad \forall q \in H_0^1(\Omega).$$

In particular, letting $q = \tilde{p}^*$, we have

$$\begin{aligned}&\int_\Omega p^*(\tilde{p}^* - p^* \, d\mathbf{x} + h\int_\Omega D(\mathbf{x})\nabla p^* \cdot \nabla(\tilde{p}^* - p^*) \, d\mathbf{x} \\ &- \int_\Omega \hat{p}(\tilde{p^*} - p^*) \, d\mathbf{x} - h\int_\Omega F(p^*, m^k)(\tilde{p}^* - p^*) \, d\mathbf{x} \geq 0.\end{aligned}$$

Using (3.1.28), we have

$$\int_\Omega (p^* - \tilde{p}^*)(\tilde{p}^* - p^* \, d\mathbf{x} + h\int_\Omega D(\mathbf{x})\nabla(p^* - \tilde{p}^*) \cdot \nabla(\tilde{p}^* - p^*) \, d\mathbf{x} \geq 0.$$

It follows that $\tilde{p}^* \equiv p^*$.

Next we need to show that both sequences $p^k, k \geq 1$ and $m^k, k \geq 1$ converge in $H^1(\Omega)$. Indeed, if the pair $(p^{k+1}, m^{k+1}) \to (p^*, m^*)$ in $H^1(\Omega)$, then we can show (p^*, m^*) is the discrete weak solution of (1.1.2). To this end, we use the first equation

in (3.1.15) for $k+1$ and for k and compute the difference:

$$\begin{aligned}&\int_\Omega (p^{k+1}-p^k)q\,d\mathbf{x} + h\int_\Omega D(\mathbf{x})\nabla(p^{k+1}-p^k)\cdot\nabla q\,d\mathbf{x}\\ &\quad = h\int_\Omega (F(p^{k+1},m^k)-F(p^k,m^{k-1}))q\,d\mathbf{x}\end{aligned}$$

for all $q \in H_0^1(\Omega)$. In particular, we choose $q = p^{k+1}-p^k$ in the equation above to have

$$\begin{aligned}&\int_\Omega |p^{k+1}-p^k|^2 d\mathbf{x} + h\int_\Omega D(\mathbf{x})|\nabla(p^{k+1}-p^k)|^2\\ &\quad \le hL\frac{3}{2}\int_\Omega\int_\Omega |p^{k+1}-p^k|^2 d\mathbf{x} + hL\frac{1}{2}\int_\Omega |m^k-m^{k-1}|^2 d\mathbf{x}. \qquad (3.1.29)\end{aligned}$$

Similarly for the density function m we have

$$\begin{aligned}&\int_\Omega |m^{k+1}-m^k|^2 d\mathbf{x} + h\int_\Omega D(\mathbf{x})|\nabla(m^{k+1}-m^k)|^2\\ &\quad \le hL\frac{3}{2}\int_\Omega\int_\Omega |m^{k+1}-m^k|^2 d\mathbf{x} + hL\frac{1}{2}\int_\Omega |p^k-p^{k-1}|^2 d\mathbf{x}. \qquad (3.1.30)\end{aligned}$$

Combining the above two inequalities yields

$$\begin{aligned}&(1-3hL/2)\left(\int_\Omega |p^{k+1}-p^k|^2 d\mathbf{x} + \int_\Omega |m^{k+1}-m^k|^2 d\mathbf{x}\right)\\ &\quad \le \frac{hL}{2}\left(\int_\Omega |m^k-m^{k-1}|^2 d\mathbf{x} + \int_\Omega |p^k-p^{k-1}|^2 d\mathbf{x}\right). \qquad (3.1.31)\end{aligned}$$

Letting $\alpha = hL/(2-3hL) < 1$ if $h > 0$ small enough, we see

$$\int_\Omega |p^{k+1}-p^k|^2 d\mathbf{x} + \int_\Omega |m^{k+1}-m^k|^2 d\mathbf{x} \le \alpha\left(\int_\Omega |p^k-p^{k-1}|^2 d\mathbf{x} + \int_\Omega |m^k-m^{k-1}|^2 d\mathbf{x}\right)$$

for all $k \ge 1$. It follows that $\int_\Omega |p^k-p^{k-1}|^2 d\mathbf{x} + \int_\Omega |m^k-m^{k-1}|^2 d\mathbf{x}, k \ge 1$ is a Cauchy sequence in $\mathbb{R}$ and hence, p^k and m^k are convergent strongly in $L^2(\Omega)$. Furthermore,

from (3.1.29) we have

$$h\int_\Omega D(\mathbf{x})|\nabla(p^{k+1}-p^k)|^2 \le hL\frac{1}{2}\int_\Omega |m^k-m^{k-1}|^2 d\mathbf{x}$$

if $1-3hL/2>0$. Thus, we conclude

$$\int_\Omega D(\mathbf{x})|\nabla(p^{k+1}-p^k)|^2 \le \frac{L}{2}\int_\Omega |m^k-m^{k-1}|^2 d\mathbf{x} \le \frac{L}{2}\alpha^{k-1}.$$

Hence, if $D(\mathbf{x}) \ge K > 0$, we know $\nabla p^k, k \ge 1$ are a Cauchy sequence and hence, $\nabla p^k, k \ge 1$ converge in $L^2(\Omega)$ strongly. In summary, $p^k, k \ge 1$ converge in $H^1(\Omega)$. Similar for m^k. These complete the proof of the following theorem.

Theorem 3.1.6. *Suppose that F and G are in the form of (2.1) and Lipschitz continuous functions over $[0,1]\times[0,1]$. Suppose that $D(\mathbf{x}) \ge K > 0$ and $E(\mathbf{x}) \ge K > 0$. For any known $p_h(\mathbf{x}, t_{i-1})$ and $m_h(\mathbf{x}, t_{i-1})$, we start with $p^1 = \hat{p} = p_h(, \mathbf{t_{i-1}})$ and $m^1 = \hat{m} = m_h(\mathbf{x}, t_{i-1})$ and compute p^{k+1}, m^{k+1} from (2.1.5) for $k \ge 1$. Then p^k, m^k converge strongly in $H^1(\Omega)$ to $p_i^*, m_i^* \in \mathcal{A}$ which the discrete weak solution at t_i.*

Proof. Based on the discussion above, we can see p_i^*, m_i^* satisfy (3.1.8) and (3.1.9). We shall denote them by $p_h(\mathbf{x}, t_i) = p_i^*$ and $m_h(\mathbf{x}, t_i) = m_i^*$. This computational procedure generates the discrete weak solutions of the PDE (1.1.2). □

3.2 The Computational Scheme

We use bivariate spline functions to implement the algorithm described in Theorem 3.1.6. For details on the use bivariate spline functions we direct the reader to [1] and to the Appendix in Chapter 6.

We reuse the definition for spline space from Definition 2.2.1. We shall denote the basis of this space as $\{\phi_j\}_{1\le j\le n}$. For convenience, we let $\mathcal{S}(\triangle) = S_d^r(\triangle) \cap H_0^1(\Omega)$. Our computational algorithm is given as follows:

Algorithm 3.2.1. *Assuming we have* $p_h(\mathbf{x}, t_{i-1}), m_h(\mathbf{x}, t_{i-1}) \in \mathcal{S}(\triangle)$, *we set out to find* $p_h(\mathbf{x}, t_i), m_h(\mathbf{x}, t_i) \in \mathcal{S}(\triangle)$ *by iteratively solving the following equations starting with* $p^1 = p_h(\mathbf{x}, t_{i-1})$ *and* $m^1 = m_h(\mathbf{x}, t_{i-1})$ *and for* $k = 1, 2, \cdots$, *do the computations in*

$$\begin{aligned} &\int_\Omega p^{k+1} q(\mathbf{x}) d\mathbf{x} + h \int_\Omega D(\mathbf{x}) \nabla p^{k+1} \cdot \nabla q(\mathbf{x}) d\mathbf{x} \\ &- h \int_\Omega F(p^{k+1}, m^k) q(\mathbf{x}) d\mathbf{x} = \int_\Omega p_h(\mathbf{x}, t_{i-1}) q(\mathbf{x}) d\mathbf{x}, \quad \forall q \in \mathcal{S}(\triangle) \end{aligned} \tag{3.2.1}$$

and

$$\begin{aligned} &\int_\Omega m^{k+1} q(\mathbf{x}) d\mathbf{x} + h \int_\Omega E(\mathbf{x}) \nabla m^{k+1} \cdot \nabla q(\mathbf{x}) d\mathbf{x} \\ &- h \int_\Omega G(p^k, m^{k+1}) q(\mathbf{x}) d\mathbf{x} = \int_\Omega m_h(\mathbf{x}, t_{i-1}) q(\mathbf{x}) d\mathbf{x}, \quad \forall q \in \mathcal{S}(\triangle) \end{aligned} \tag{3.2.2}$$

until $p^{k+1} - p^k$ *and* $m^{k+1} - m^k$ *are within a tolerance in* $H^1(\Omega)$ *norm. Note that the computation of* p^{k+1} *and* m^{k+1} *requires an iterative algorithm, which is adapted from the case of a single species in Chapter 2, since* F *and* G *are nonlinear.*

Let $S_p(\mathbf{x}, t_i)$ and $S_m(\mathbf{x}, t_i)$ be the limit of the iterative solutions $p^k, m^k, k \geq 1$ produced in Algorithm 3.2.1. That is, $S_p(\mathbf{x}, t_i)$ and $S_m(\mathbf{x}, t_i)$ are spline solutions for (1.1.2). It is interesting to see if they approximate the discrete weak solution $p_h(\mathbf{x}, t_i), m_h(\mathbf{x}, t_i)$ of (1.1.2). Let $S_p^*(\cdot, t_i)$ be the best spline approximation of $p_h(\mathbf{x}, t_i)$ in $\mathcal{S}(\triangle)$. As a result of Theorem 10.4 in [24], it follows that

$$\begin{aligned} \left|\left| S_{p_h}^* - p_h \right|\right|_2 &\leq C_3 |\triangle|^2 |p_h|_{2,2,\Omega} \\ \left|\left| \nabla S_p^* - \nabla p_h \right|\right|_2 &\leq C_4 |\triangle| |p_h|_{2,2,\Omega}, \end{aligned}$$

where $|\triangle|$ is the length of the longest edge in the triangulation and C_3 and C_4 are constants independent of p. Similar for $S_m^*(\cdot, t_i)$.

Chapter 4

Numerical Simulations

In this chapter we present results from the numerical solver which implements the algorithm presented in this report. The code is written in C++ and Octave, a free and libre clone of Matlab. My contribution builds on a substantial codebase written by Dr. Ming-Jun Lai and Dr. Paul Wenston who implemented the myriad algorithms necessary for the creation and manipulation of bivariate splines.

In order to be confident in the accuracy of the solver, we present synthetic tests in which an exact solution of the PDE is compared to the solver's solution. Unfortunately, there are no known exact solutions to the PDEs presented in this report other than the constant steady-states $p(\mathbf{x}, t) = 0$, $p(\mathbf{x}, t) = 1$ and $p(\mathbf{x}, t) = \sigma$. Thus, it is useful to add a forcing term to the PDE which is specifically chosen such that a desired function $p(\mathbf{x}, t)$ is an exact solution to the PDE.

$$\frac{\partial p}{\partial t} = \text{ div } (D\nabla p)) + Ap(1-p)(p-\sigma) + f(\mathbf{x}, t), \tag{4.0.1}$$

For example, let $p(\mathbf{x}, t) = txy$, $D(\mathbf{x}) = 1$, $A(\mathbf{x}) = 1$ and $\sigma = 0.1$. Then choosing

$$f(\mathbf{x}, t) = xy - txy(1 - txy)\left(txy - \frac{1}{10}\right)$$

makes $p(\mathbf{x}, t)$ a solution of (4.0.1). The boundary condition is set to fit the known exact solution. Adding the forcing term to the numerical solver then allows me to attempt to recover $p(\mathbf{x}, t)$ and compare the numerical solution to the exact solution.

Similarly, the predator-prey system is modified with a pair of forcing functions.

$$\begin{aligned} \frac{\partial p}{\partial t} &= \operatorname{div}(D_1 \nabla p) + Ap(1-p)(1-\sigma) - \alpha pm + f(\mathbf{x}, t) \\ \frac{\partial m}{\partial t} &= \operatorname{div}(D_2 \nabla m) + Bm(1-m)(1-\gamma) + \beta pm + g(\mathbf{x}, t) \end{aligned} \tag{4.0.2}$$

Once the accuracy of the solver is confirmed, we present some visualizations of solutions in the form of surfaces and plots of total population over Ω as a function of time.

4.1 Accuracy of Single Species Numerical Solution

In all the tests below, we solve the system in (4.0.1) for $t \in [0, 1]$ using spline degree 5, for various time steps h and various triangulation sizes N_T. We then measure the error $||p(\mathbf{x}, 1) - S_p(\mathbf{x}, 1)||_\infty$, where S_p is the spline numerical solution, and tabulate the results.

Example 4.1.1. This test function is a polynomial of degree 4 and decays over time. The error decreases roughly like $O(h)$. Even modestly small values of N_T give good errors because the test function is exactly representable as a spline. The domain is $\mathbf{x} \in [0, 1] \times [0, 1]$.

$p(\mathbf{x}, t)$	$D(\mathbf{x})$	$A(\mathbf{x})$	σ
$\dfrac{13x(x-1)y(y-1)}{1+t}$	0.005	1	0.1

Example 4.1.2. This example builds on the previous one but complicates the model by introducing a nonlinear diffusion term. The error decreases roughly like $O(h)$.

h \ N_T	2	8	32	128	512
5×10^{-2}	3.94×10^{-2}	3.30×10^{-2}	3.44×10^{-2}	3.44×10^{-2}	3.44×10^{-2}
5×10^{-3}	5.40×10^{-2}	4.14×10^{-3}	3.34×10^{-3}	3.35×10^{-3}	3.35×10^{-3}
5×10^{-4}	5.57×10^{-2}	5.59×10^{-3}	3.31×10^{-4}	3.33×10^{-4}	3.34×10^{-4}
5×10^{-5}	5.59×10^{-2}	5.75×10^{-3}	3.10×10^{-5}	3.33×10^{-5}	3.33×10^{-5}

Table 4.1: Error measurement $||p(\mathbf{x}, 1) - S_p(\mathbf{x}, 1)||_\infty$ in Example 4.1.1.

Once again we see that small N_T are sufficient to achieve the optimal rate. The domain is $\mathbf{x} \in [0, 1] \times [0, 1]$.

The nonconstant diffusion term places a substantial burden on the solver, and thus computing a solution for $h = 5 \times 10^{-5}$ proved too slow.

$p(\mathbf{x}, t)$	$D(\mathbf{x})$	$A(\mathbf{x})$	σ
$\dfrac{13x(x-1)y(y-1)}{1+t}$	$0.005e^{-(x-.5)^2-(y-.5)^2}$	1	0.1

h \ N_T	2	8	32	128
5×10^{-2}	1.92×10^{-2}	1.69×10^{-2}	1.66×10^{-2}	1.66×10^{-2}
5×10^{-3}	4.65×10^{-3}	1.97×10^{-3}	1.68×10^{-3}	1.68×10^{-3}
5×10^{-4}	4.35×10^{-3}	4.75×10^{-4}	1.69×10^{-4}	1.69×10^{-4}

Table 4.2: Error measurement $||p(\mathbf{x}, 1) - S_p(\mathbf{x}, 1)||_\infty$ in Example 4.1.2.

Example 4.1.3. This test function is not a polynomial, so it is not exactly representable as a spline. We see that it does not present a serious challenge for the solver even for small N_T. The domain is $\mathbf{x} \in [0, 1] \times [0, 1]$.

$p(\mathbf{x}, t)$	$D(\mathbf{x})$	$A(\mathbf{x})$	σ
$\dfrac{\sin(\pi x)\sin(\pi y)}{1+t}$	$0.005e^{-(x-.5)^2-(y-.5)^2}$	1	0.1

h \ N_T	2	8	32	128
5×10^{-2}	1.98×10^{-2}	1.92×10^{-2}	1.84×10^{-2}	1.84×10^{-2}
5×10^{-3}	5.99×10^{-3}	2.65×10^{-3}	1.88×10^{-3}	1.87×10^{-3}
5×10^{-4}	5.61×10^{-3}	9.64×10^{-4}	1.96×10^{-4}	1.87×10^{-4}

Table 4.3: Error measurement $||p(\mathbf{x},1) - S_p(\mathbf{x},1)||_\infty$ in Example 4.1.3.

4.2 Accuracy of Predator-Prey Numerical Solution

In all the tests below, we solve the system in (4.0.2) for $t \in [0,1]$ using spline degree 6, for various time steps h and various triangulation sizes N_T. We then measure the sum of the errors $||p(\mathbf{x},1) - S_p(\mathbf{x},1)||_\infty + ||m(\mathbf{x},1) - S_m(\mathbf{x},1)||_\infty$, where S_p and S_m are the numerical solutions, and tabulate the results.

Example 4.2.1. These test functions do not depend on the spatial variable $\mathbf{x}$ and as a result the PDEs are reduced to ODEs. The size of the triangulation has no effect on accuracy. The domain is $\mathbf{x} \in [0,1] \times [0,1]$.

$p(\mathbf{x},t)$	$m(\mathbf{x},t)$	$D_1(\mathbf{x})$	$D_2(\mathbf{x})$	$A(\mathbf{x})$	$B(\mathbf{x})$	σ	γ	α	β
$\frac{e^t}{1+e^t}$	$\frac{e^t}{1+3e^t}$	0.005	0.005	1	1	0.1	0.15	1	-1

h \ N_T	2
1×10^{-1}	3.04×10^{-3}
1×10^{-2}	3.07×10^{-4}
1×10^{-3}	3.08×10^{-5}
1×10^{-4}	3.08×10^{-6}

Table 4.4: $||p(\mathbf{x},1) - S_p(\mathbf{x},1)||_\infty + ||m(\mathbf{x},1) - S_m(\mathbf{x},1)||_\infty$ in Example 4.2.1.

Example 4.2.2. These test functions do not depend on the time variable t and as a result the PDE is time-independent. The step size h has negligible effect on accuracy, which is the result of floating-point errors. The domain is $\mathbf{x} \in [0,1] \times [0,1]$.

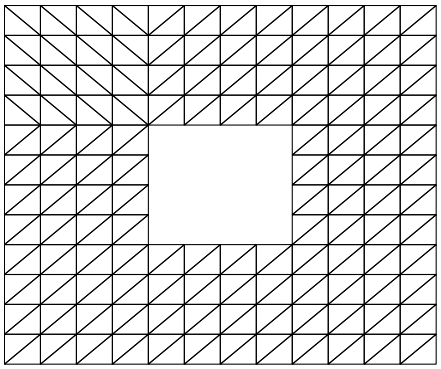
$p(\mathbf{x},t)$	$m(\mathbf{x},t)$	$D_1(\mathbf{x})$	$D_2(\mathbf{x})$	$A(\mathbf{x})$	$B(\mathbf{x})$	σ	γ	α	β
$\frac{1}{2}\cos(2\pi x)$	$\cos(2\pi x)$	0.005	0.005	1	1	0.1	0.15	1	-1

h \ N_T	2	8	32	128	512
1×10^{-1}	1.15	2.37×10^{-2}	1.57×10^{-4}	8.01×10^{-7}	7.32×10^{-8}
1×10^{-2}	1.16	2.42×10^{-2}	1.59×10^{-4}	8.17×10^{-7}	7.50×10^{-8}
1×10^{-3}	1.16	2.43×10^{-2}	1.60×10^{-4}	8.18×10^{-7}	7.52×10^{-8}

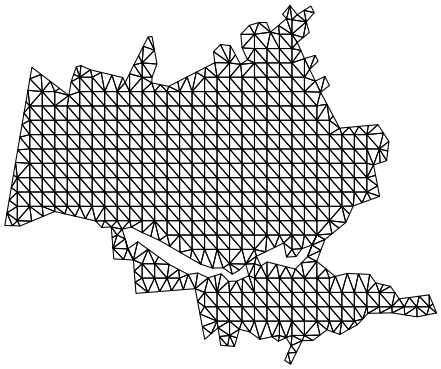
Table 4.5: $||p(\mathbf{x},1) - S_p(\mathbf{x},1)||_\infty + ||m(\mathbf{x},1) - S_m(\mathbf{x},1)||_\infty$ in Example 4.2.2.

Example 4.2.3. These test functions depend on both t and $\mathbf{x}$, testing the full generality of the solver. This example has a small time derivative, so the solver is expected to work well. The domain of $\mathbf{x}$ is shown in Figure 4.1a.

$p(\mathbf{x},t)$	$m(\mathbf{x},t)$	$D_1(\mathbf{x})$	$D_2(\mathbf{x})$	$A(\mathbf{x})$	$B(\mathbf{x})$	σ	γ	α	β
$\sin\left(\frac{\pi t}{5} + 2\pi x\right)$	$\cos\left(\frac{\pi t}{5} + 2\pi x\right)$	0.005	0.005	1	1	0.1	0.15	1	-1



(a) Square domain with hole.



(b) City of Bandiagara, Mali.

Figure 4.1: Triangulations commonly used in numerical simulations.

h \ N_T	16	64	256	1024
1×10^{-1}	2.22	4.20×10^{-2}	3.49×10^{-2}	3.49×10^{-2}
1×10^{-2}	2.25	4.29×10^{-2}	3.47×10^{-3}	3.46×10^{-3}
1×10^{-3}	2.25	4.38×10^{-2}	4.79×10^{-4}	3.46×10^{-4}
1×10^{-4}	2.25	4.39×10^{-2}	3.29×10^{-4}	3.42×10^{-5}

Table 4.6: $||p(\mathbf{x},1) - S_p(\mathbf{x},1)||_\infty + ||m(\mathbf{x},1) - S_m(\mathbf{x},1)||_\infty$ in Example 4.2.3.

Example 4.2.4. In contrast to Example 4.2.3, these test functions have a substantial time derivative, which provides the solver with a more significant challenge. The rest of the settings are identical.

$p(\mathbf{x},t)$	$m(\mathbf{x},t)$	$D_1(\mathbf{x})$	$D_2(\mathbf{x})$	$A(\mathbf{x})$	$B(\mathbf{x})$	σ	γ	α	β
$\sin(2\pi(t+x))$	$\cos(2\pi(t+x))$	0.005	0.005	1	1	0.1	0.15	1	-1

h \ N_T	16	64	256	1024
1×10^{-1}	6.09×10^{-1}	5.69×10^{-1}	5.86×10^{-1}	5.91×10^{-1}
1×10^{-2}	2.68×10^{-1}	7.19×10^{-2}	6.92×10^{-2}	6.97×10^{-2}
1×10^{-3}	2.62×10^{-1}	1.04×10^{-2}	7.01×10^{-3}	7.08×10^{-3}
1×10^{-4}	2.61×10^{-1}	6.71×10^{-3}	6.91×10^{-4}	7.09×10^{-4}

Table 4.7: $||p(\mathbf{x},1) - S_p(\mathbf{x},1)||_\infty + ||m(\mathbf{x},1) - S_m(\mathbf{x},1)||_\infty$ in Example 4.2.4.

Example 4.2.5. The examples presented thus far all make use of a numerical solver based on the backward Euler method for differential equations. That scheme is stable when faced with stiff systems, but it suffers from somewhat poor numerical accuracy unless a very small time step is chosen. The Backward Differentiation Formula (BDF) is a well-known linear multistep method, which generalizes the backward Euler scheme. The scheme is readily adaptable to the predator-prey system presented in this chapter.

To illustrate the benefit of this method, we run the solver with the same initial conditions and parameters as Example 4.2.4, which presents the solver with the most difficult circumstances. Thus, a direct comparison can be made. There is clear improvement for small time steps h, and yet $h = 0.001$ produces identical results to $h = 0.01$ with this new scheme. We can only surmise that decreasing the time step too far causes too much floating-point truncation error and presents no further improvement. It is also clear that a very fine triangulation is needed to achieve an accuracy on the order of 1×10^{-7}.

In practice, using a modestly small time step such as $h = 0.01$ substantially improves the running time of the numerical solver compared to using the backward Euler method with $h = 1 \times 10^{-4}$, since it achieves the same error rates at two orders of magnitude less time.

$p(\mathbf{x},t)$	$m(\mathbf{x},t)$	$D_1(\mathbf{x})$	$D_2(\mathbf{x})$	$A(\mathbf{x})$	$B(\mathbf{x})$	σ	γ	α	β
$\sin(2\pi(t+x))$	$\cos(2\pi(t+x))$	0.005	0.005	1	1	0.1	0.15	1	-1

h \ N_T	16	64	256	1024
1×10^{-1}	2.04×10^{-1}	3.43×10^{-2}	3.15×10^{-2}	3.15×10^{-2}
1×10^{-2}	2.61×10^{-1}	6.29×10^{-3}	7.52×10^{-5}	4.03×10^{-7}
1×10^{-3}	2.61×10^{-1}	6.29×10^{-3}	7.53×10^{-5}	3.85×10^{-7}

Table 4.8: $||p(\mathbf{x},1) - S_p(\mathbf{x},1)||_\infty + ||m(\mathbf{x},1) - S_m(\mathbf{x},1)||_\infty$ in Example 4.2.5, using BDF of order 2.

4.3 Simulations of One Species

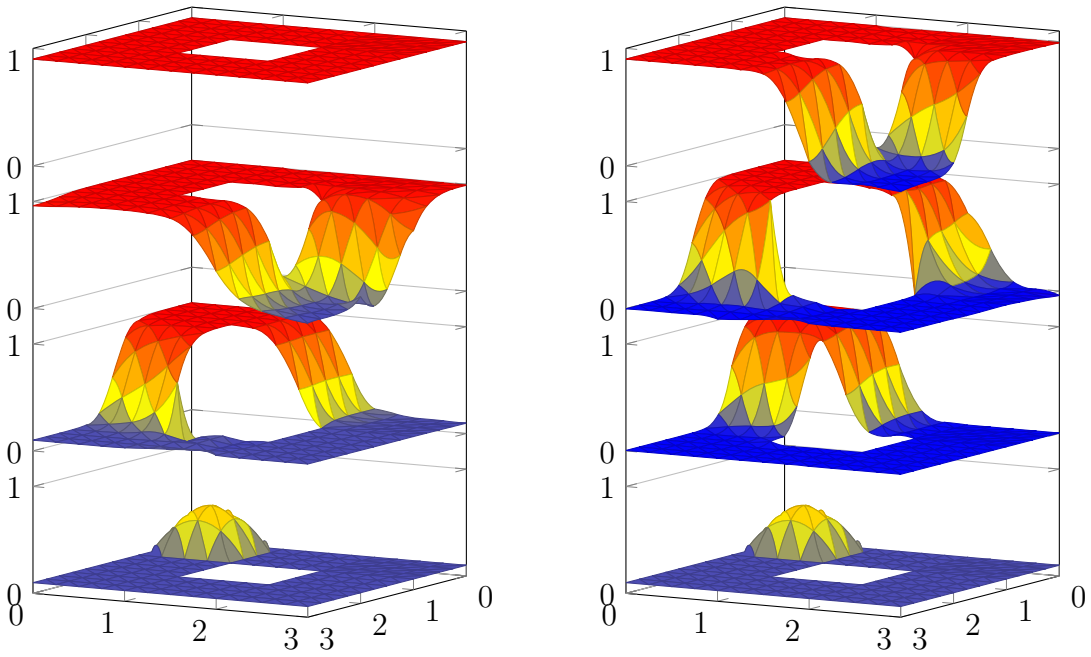
We run simulations to find a solution of (1.1.1) for various initial conditions and parameters. We shall use the two triangulated domains shown in Figures 4.1a and 4.1b.

We provide several examples to show how various growth functions affect the rate at which the solution reaches the asymptotically stable constant solution of $p(x, y) = 1$ or $p(x, y) = 0$.

Figures 4.2 through 4.5 show several 3D renders of how solutions grow over time over two domains indicated in Fig 4.1b. Each subfigure shows four equally-spaced time slices, plotted on the same xy-axes, one on top of each other, allowing the reader to observe how the solution grows over time. Initial time slice appears as the bottommost surface and the final state is the topmost surface. In addition, each figure shows the effect of varying the Allee threshold σ. With low σ we see a very quick spread since any amount of infection will expand to infect all individuals. Higher σ corresponds to a need for a critical mass before infection can permanently establish itself in a region. A high value for σ causes the average population to grow more slowly as seen in Figure 4.7. It can introduce sharp rises in population density between regions where $p(x) < \sigma$ and regions where $p(x) > \sigma$ as seen in Figure 4.6e. In order to make the difference in the behavior of the solution clearer, the value of t for each time slice is indicated in the caption of each figure.

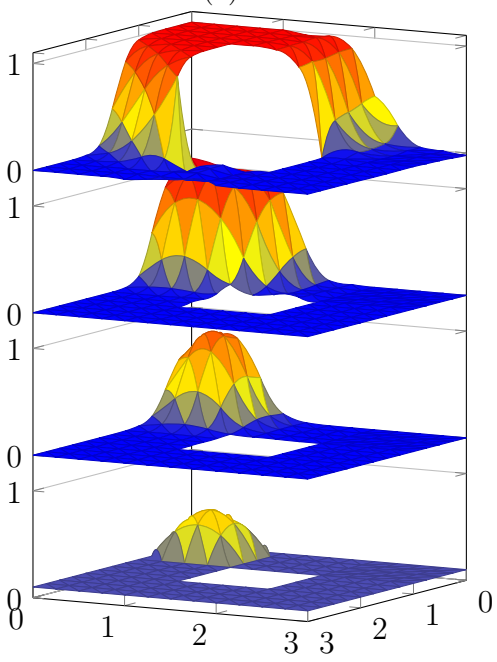
Figures 4.7 through show average population over time over the city of Bandiagara, Mali. Each subfigure corresponds to a certain set of initial conditions for the PDE, while separating the cases by the choice for σ, emphasizing the effect σ has on the rate at which the population reaches an asymptotically stable solution.

We can observe some expected behavior from the solutions presented in Figure 4.2. The initial condition is uniformly $p = 0.1$ on a large portion of Ω with an isolated bump function in one corner. In Figure 4.2b the second time slice shows the population has become extinct on the area where $p = 0.1$. At the same time the bump grows to population capacity and eventually spreads life into formerly dead areas. We observe similar results in Figure 4.2c, but the rate at which the population grows has been severely diminished. In Figure 4.2d, the threshold σ is so high that

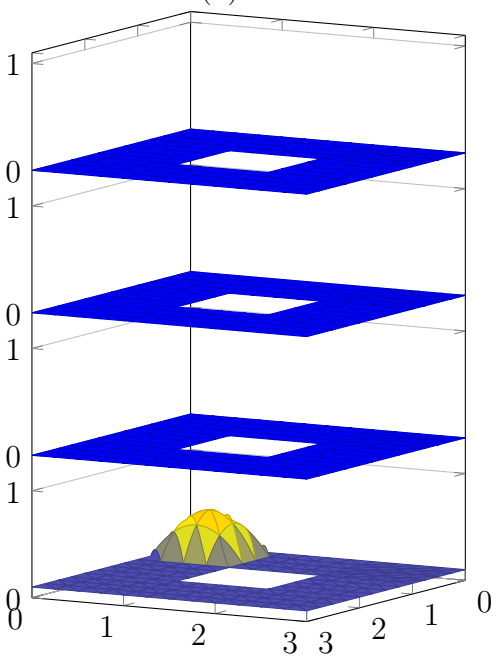


(a) $\sigma = 0.1$

(b) $\sigma = 0.2$

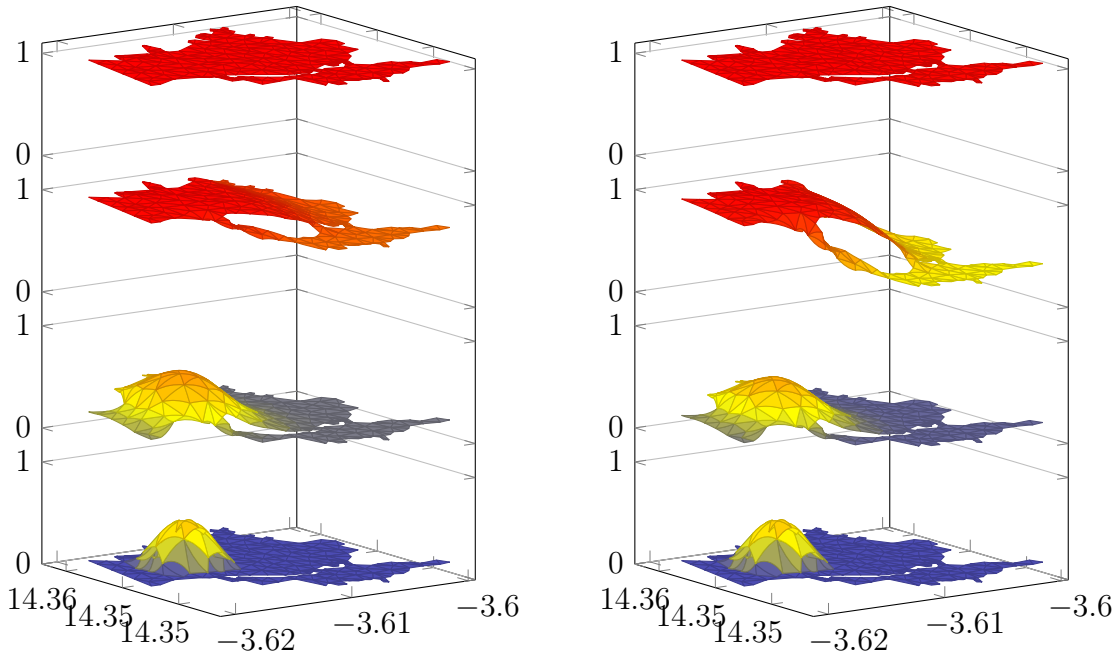


(c) $\sigma = 0.3$



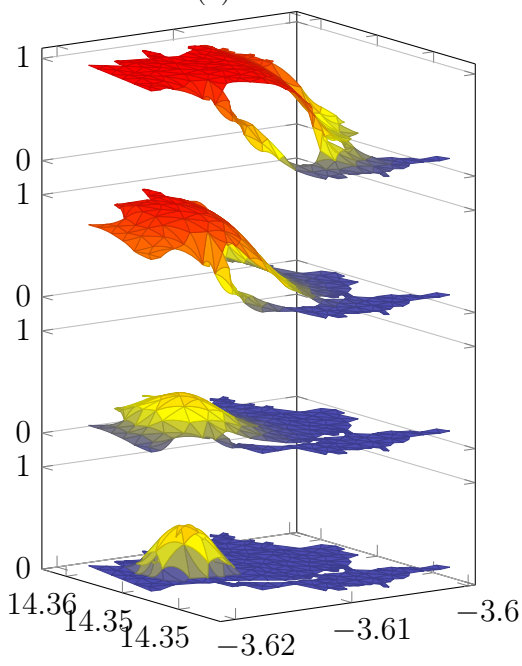
(d) $\sigma = 0.5$

Figure 4.2: Donut-shape domain. Constant growth and diffusion. Various Allee effect thresholds σ. The vertical axis shows population density $p \in [0, 1]$ at 4 points in time: $t \in \{0, 20, 45, 90\}$, where the bottom manifold represents $t = 0$, and the top manifold represents $t = 90$ in each case.

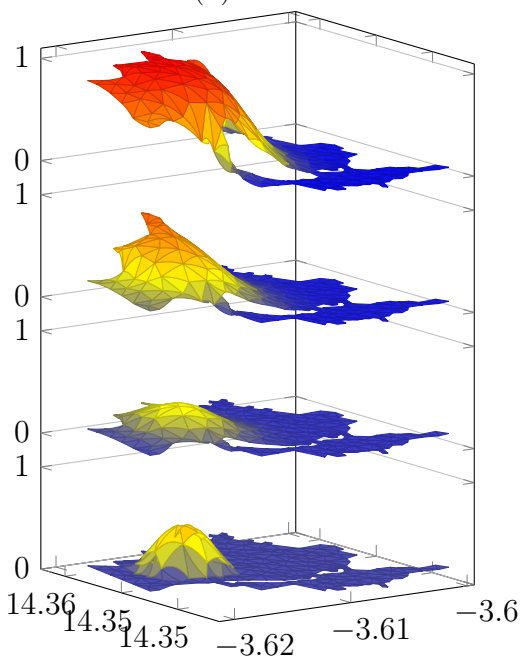


(a) $\sigma = 0.02$

(b) $\sigma = 0.05$

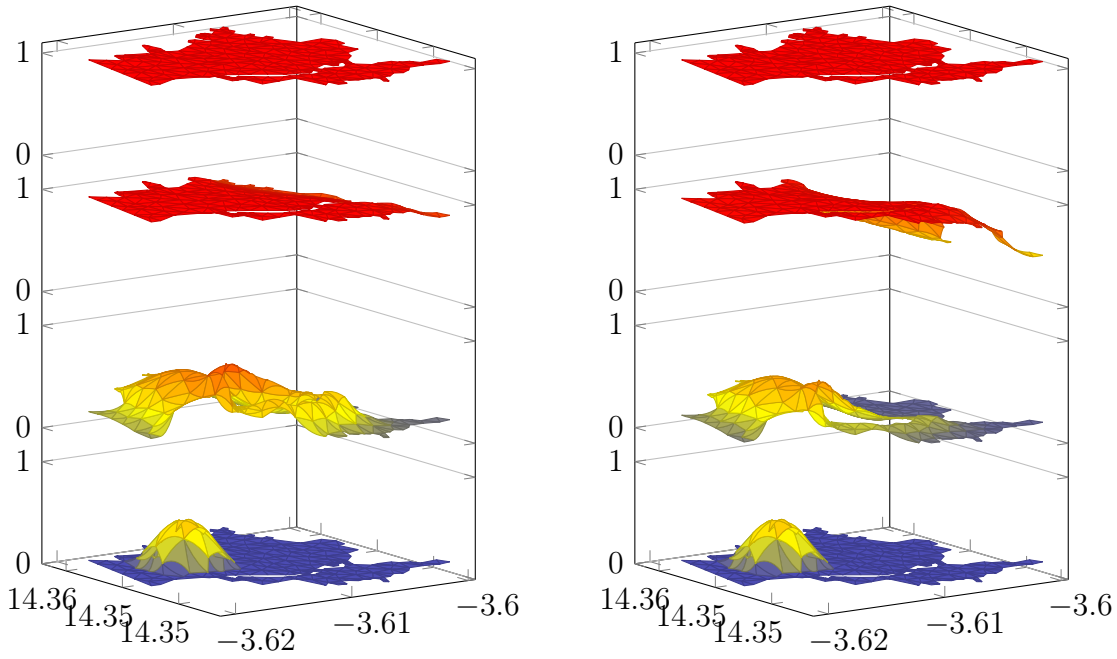


(c) $\sigma = 0.1$



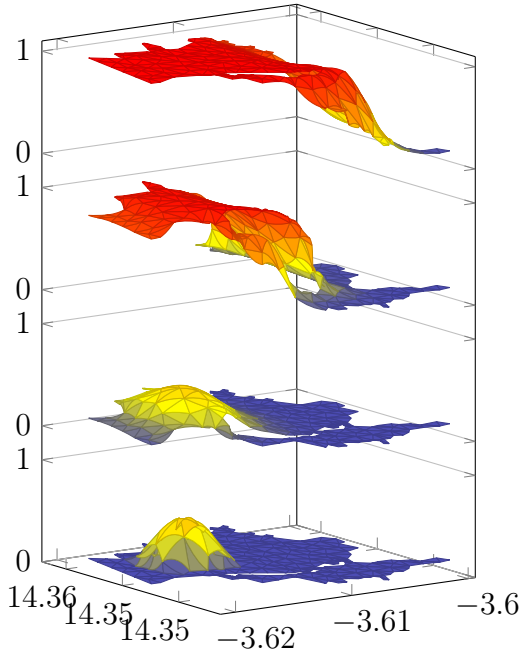
(d) $\sigma = 0.15$

Figure 4.3: City of Bandiagara, Mali. Constant growth and diffusion. Various Allee effect thresholds σ. The vertical axis shows population density $p \in [0, 1]$ at 4 points in time: $t \in \{0, 5, 13, 20\}$, where the bottom manifold represents $t = 0$, and the top manifold represents $t = 20$ in each case.

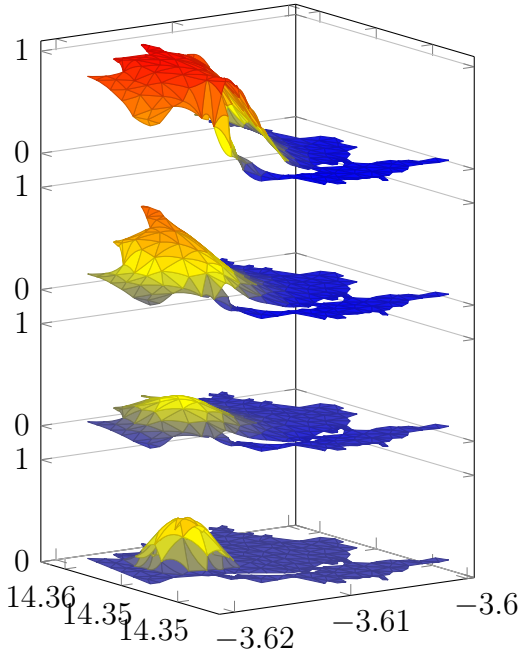


(a) $\sigma = 0.02$

(b) $\sigma = 0.05$

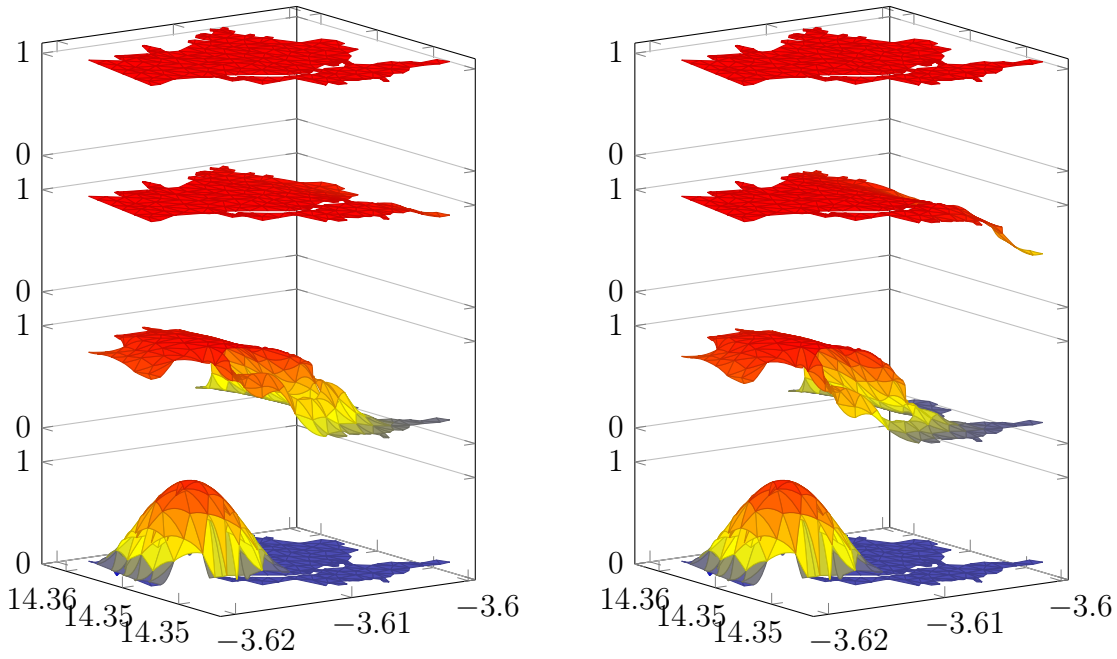


(c) $\sigma = 0.1$



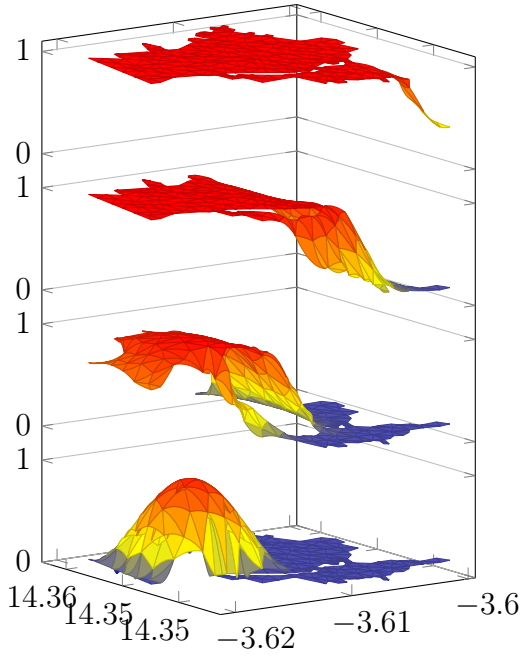
(d) $\sigma = 0.15$

Figure 4.4: City of Bandiagara, Mali. Constant diffusion. Various Allee effect thresholds. Growth function is piecewise-constant with triple magnitude for patches near the city's river. The vertical axis shows population density $p \in [0, 1]$ at 4 points in time: $t \in \{0, 5, 13, 20\}$, where the bottom manifold represents $t = 0$, and the top manifold represents $t = 20$ in each case.

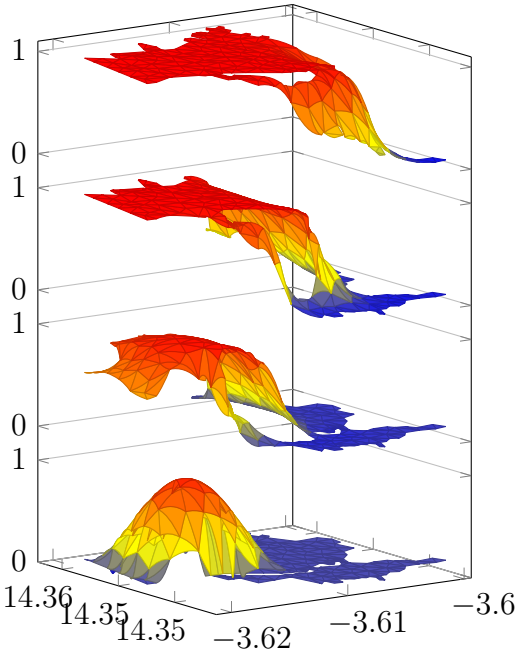


(a) $\sigma = 0.02$

(b) $\sigma = 0.05$

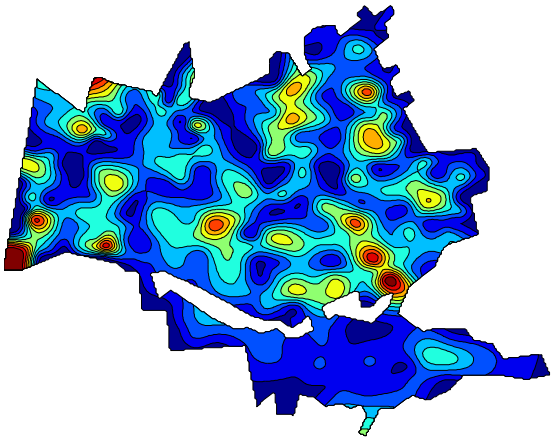


(c) $\sigma = 0.1$

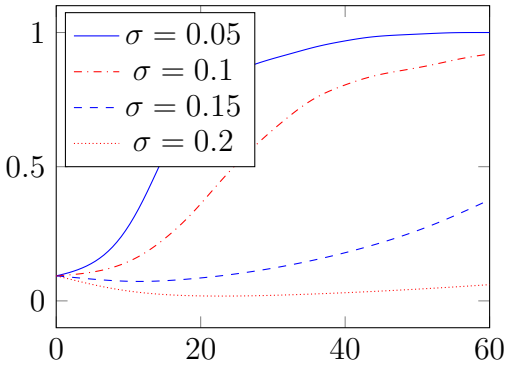


(d) $\sigma = 0.15$

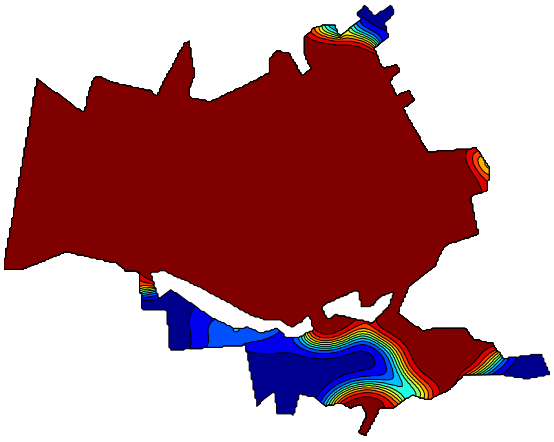
Figure 4.5: City of Bandiagara, Mali. Same as Figure 4.4 but the initial condition has a much higher total population. The vertical axis shows population density $p \in [0, 1]$ at 4 points in time: $t \in \{0, 5, 13, 20\}$, where the bottom manifold represents $t = 0$, and the top manifold represents $t = 20$ in each case.



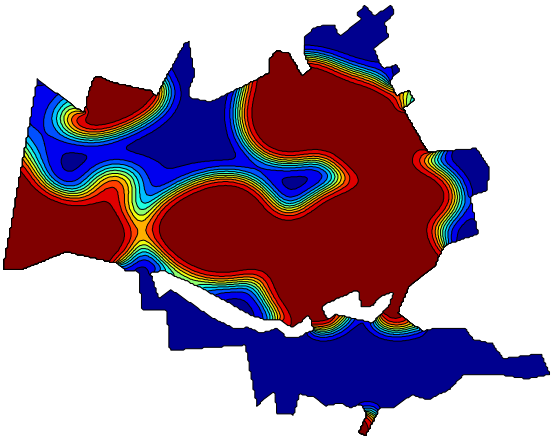
(a) Initial population density



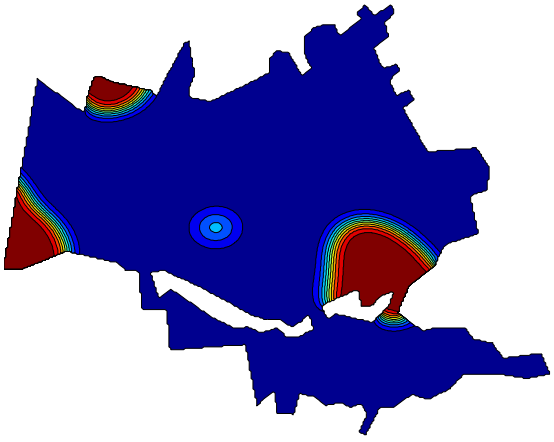
(b) Average population over time for various σ.



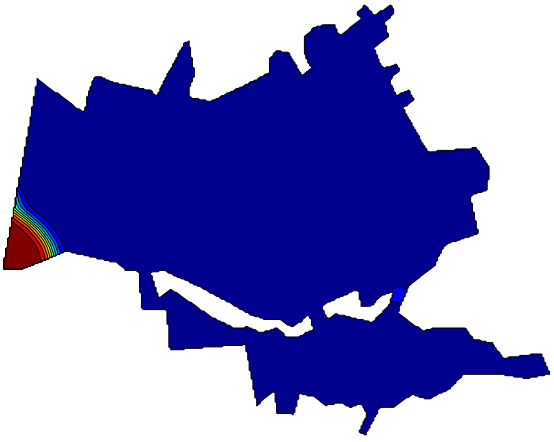
(c) $\sigma = 0.05$



(d) $\sigma = 0.1$

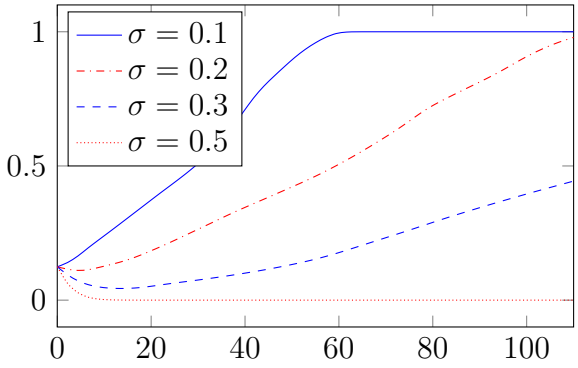


(e) $\sigma = 0.15$

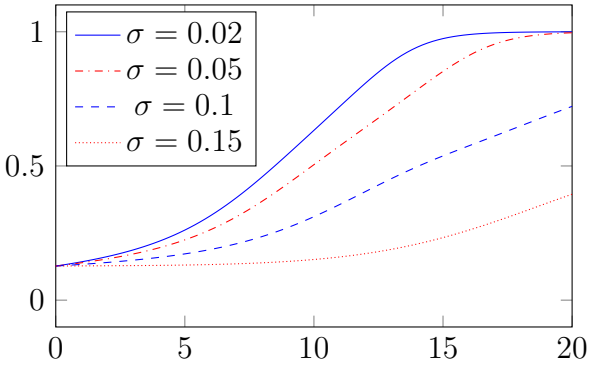


(f) $\sigma = 0.2$

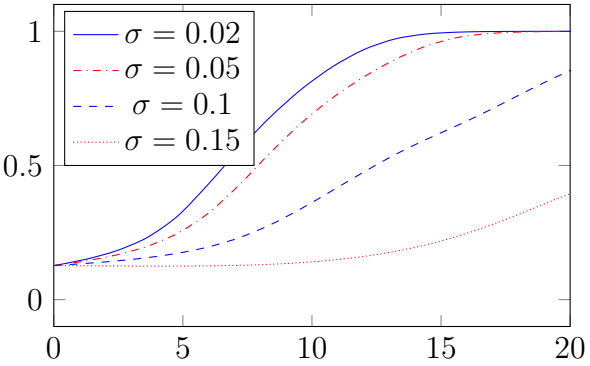
Figure 4.6: City of Bandiagara, Mali. We used spline data fitting on data of infected population density as presented in [3] and applied our model to examine future development. Figures 4.6c through 4.6f correspond to the same time $t = 27$.



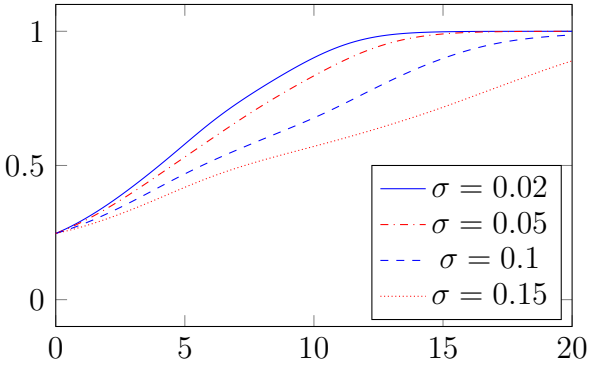
(a) Average population plot for simulations in Figure 4.2.



(b) Average population plot for simulations in Figure 4.3.



(c) Average population plot for simulations in Figure 4.4.



(d) Average population plot for simulations in Figure 4.5.

Figure 4.7: Average population density in Ω plotted over time for each of the four preceding figures.

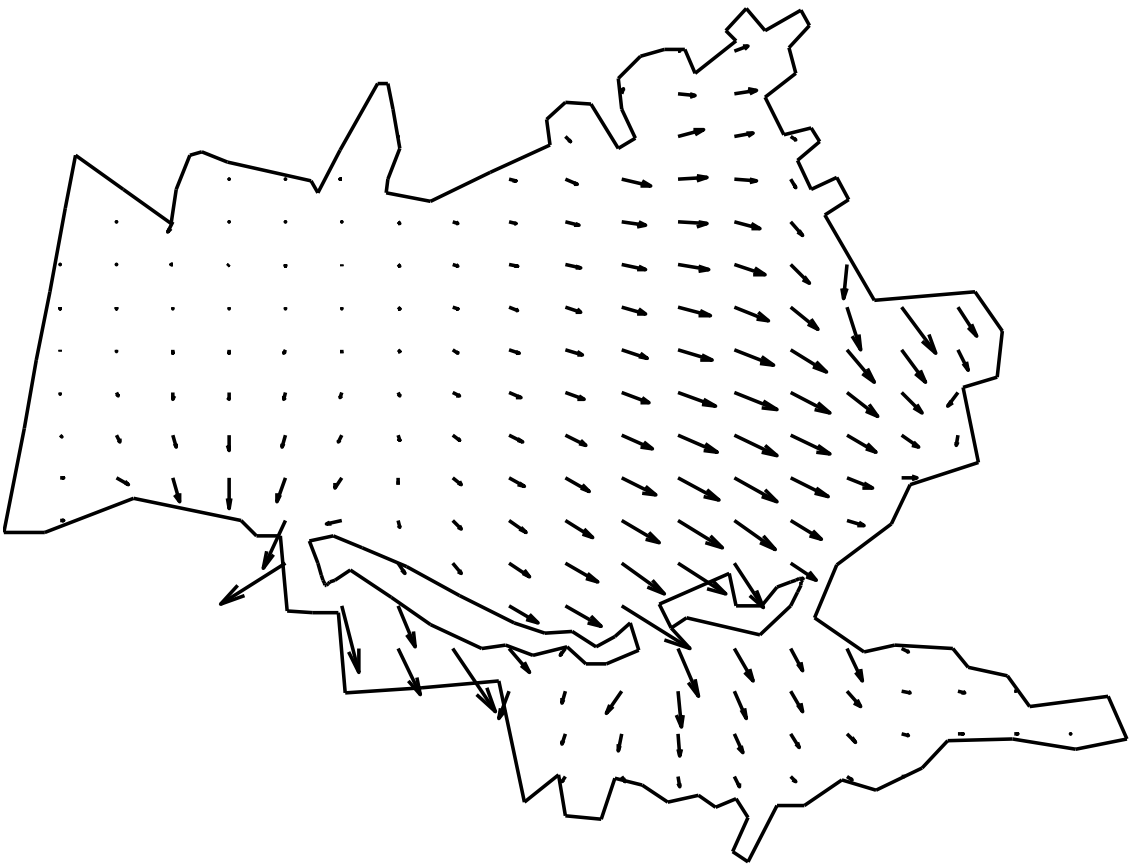


Figure 4.8: A contour plot of ∇p which corresponds to Figure 4.3a at T=15, indicating the direction in which infection spreads.

the population becomes extinct everywhere and very quickly.

4.3.1 Simulations of Malaria Study

It is well-known that malaria is one of the leading causes of mortality in the world and an estimated 3.3 billion people are at risk of malaria (cf. [40]). The World Health Organization is interested in spatial models which can identify high-risk zones of infection on a fine geographical scale as indicated in the 18th and 20th WHO reports (cf. [38] and [39]). An example of such a study can be found in [5] and in [3] where Coulibaly et al. provide data samples of individuals infected with malaria in

Bandiagara, Mali [3]. We mimic the data values presented in Figure 2 of Coulibaly et al. [3] to form an initial value for our PDE model using a bivariate spline data fitting technique (cf. [1]). Then we use our MATLAB program to simulate the development of malarial infection over a period of time using various Allee parameters. Our results are presented in Figure 4.6. σ plays a vital role in the growth rate of the infected region. When σ is small, the initial population of infected individuals is sufficiently large to cover a majority of the region by the end of the simulation. When σ is a bit larger we see that some regions become free from infection for a while since the local population density is less than σ. In Figure 4.6b we see that with a high enough σ it is possible for average infected population to decrease at first yet ultimately return to growth. This kind of phenomenon would be difficult to capture with a traditional SIR model with no spatial considerations.

An appropriate calibration of this constant based on real data would be an important achievement, as high-risk zones can be identified using our model by examining a time-horizon of one year and analyzing regions where infection has taken hold.

An additional benefit to our method is that the use of splines allows us to produce smooth population density surfaces. Fisher's [10] traveling waves travel in the direction of steepest-descent on the surface and thus can visualized quite well by a contour graph of ∇p, which is helpful in identifying the direction of the spread of infection. Figure 4.8 illustrates the pattern of disease transmission.

4.4 Simulations of Two Species

We showcase some examples of numerical simulations of multiple interacting species including predator-prey and competition mode. Unless noted otherwise, all examples feature Neumann boundary conditions and restrict population density to nonnegative numbers. Simulation is run with step size h over the time span $[0, T]$.

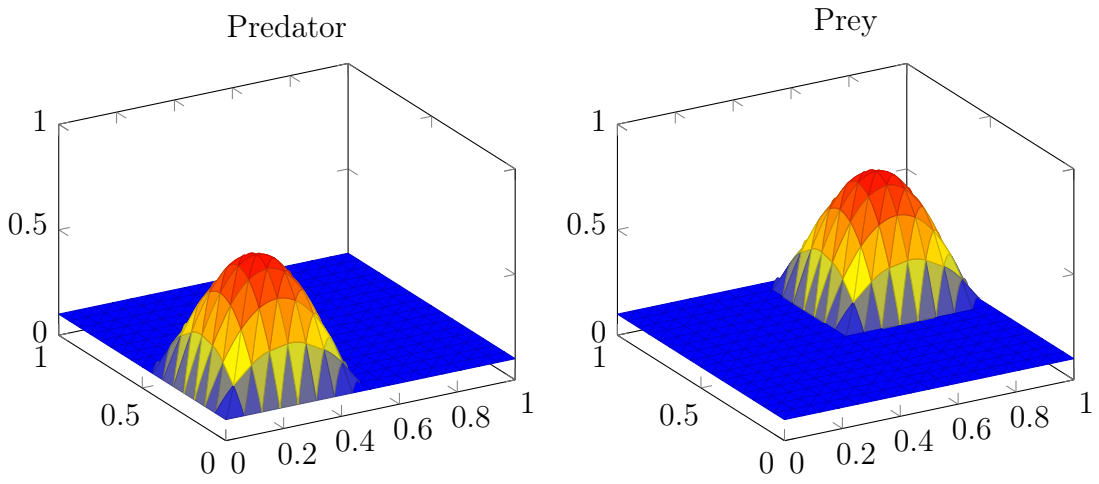


Figure 4.9: Initial conditions for Example 4.4.1.

4.4.1 Predator-Prey Interaction

In this section we focus on examples involving predator-prey interactions.

Example 4.4.1. We examine the following system, which models classical predator-prey interaction with an added diffusive term.

$$\frac{\partial p}{\partial t} = \nabla(D\nabla p) + \alpha p - \beta pm$$
$$\frac{\partial m}{\partial t} = \nabla(D\nabla m) + \delta pm - \gamma m$$

The parameters used are as follows.

D	α	β	γ	δ	h	T
0.005	8	10	6	20	0.001	20

For the PDE case, the initial condition can be seen in Figure 4.9. For the ODE case, the initial condition is predator = 0.165 and prey = 0.165, which is the total population of each species in the PDE case. We compare the behavior of the PDE to the corresponding ODE with no diffusive term by examining their phase diagrams

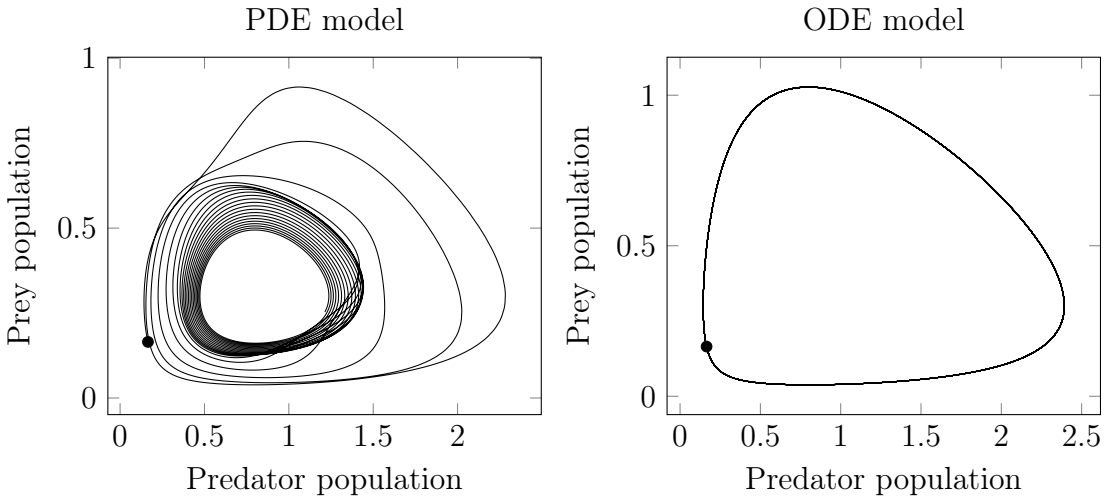


Figure 4.10: Phase portraits comparing the behavior of the PDE solution and the ODE solution. The emphasized point on the curve is the initial condition.

in Figure 4.10. We can see the PDE model exhibits quite different dynamics. A notable difference is that the PDE phase diagram exhibits a self-intersecting curve, which is impossible for a homogeneous ODE according to existence and uniqueness theory. We also see that the PDE phase diagram is asymptotically a limit cycle, which looks much like the ODE phase diagram, albeit with much lower populations of each species. The convergence to a limit cycle is the expected result of adding diffusion to the system, making populations density identical everywhere given infinite time.

Example 4.4.2. Looking back to Figure 4.9, note that the initial conditions of predator and prey are essentially bump functions, but they also have a baseline density of 0.1 outside of the bump. This guarantees that interaction between the species will occur immediately. If each population density is uniformly reduced by 0.07, this baseline would be reduced to 0.03. We can repeat the experiment with identical parameters and observe the effect on the phase diagram in Figure 4.11. The PDE system now has more unpredictable dynamics since the populations of predator and prey are more independent until diffusion acts to bring them together.

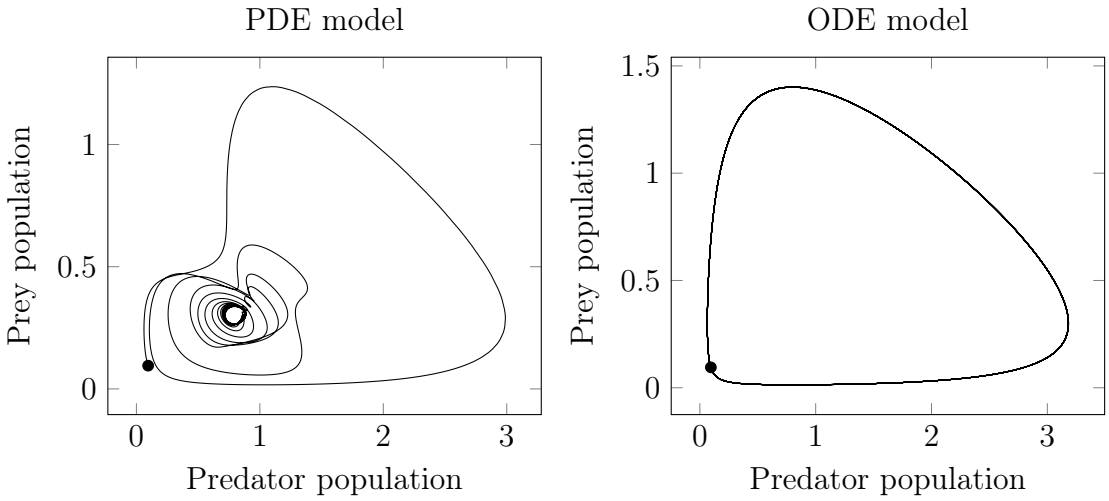


Figure 4.11: Similar to Figure 4.10 but with much smaller baseline density for each population.

Example 4.4.3. In this example we present a different predator-prey system, which features an Allee effect and a Holling type II response.

$$\frac{\partial p}{\partial t} = \nabla(D\nabla p) + \nu p(1-p)(p-\sigma) - \mu pm$$
$$\frac{\partial m}{\partial t} = \nabla(E\nabla m) + \frac{pm}{\xi + p} - \eta m$$

The parameters used are as follows.

D	E	ν	σ	μ	ξ	η	h	T
5×10^{-3}	5×10^{-7}	4	0.15	1	0.3	0.4	0.001	60

The phase diagrams in Figure 4.12 display rather different outcomes. While the ODE system led to the immediate extinction of both species, the populations in the PDE system survive much longer. In fact, the PDE system shows that initially both populations thrive with no sign of future extinction. Figure 4.13 illustrates the spatial distribution of both species at four different $t \in [0, 12]$ in order to show the coexistence of predator and prey. We see that the predator chases the prey along the edges of the domain. Ultimately, the density of prey falls under the Allee threshold and the

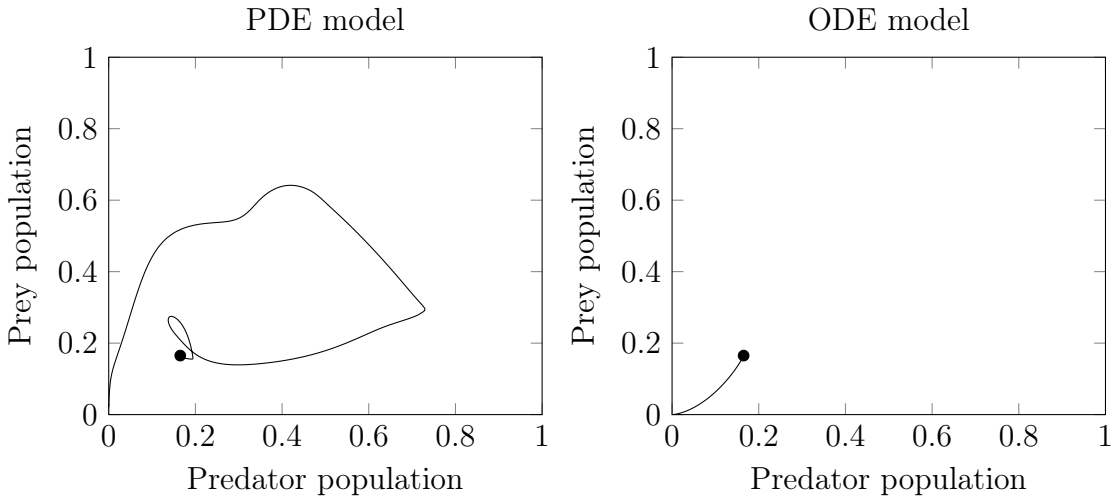


Figure 4.12: Phase diagrams for the PDE and ODE models in Example 4.4.3. The emphasized point on the curve is the initial condition.

species goes extinct.

4.4.2 Resource Competition

In this section we focus on species which are in competition for a common resource. The system is modeled by the following equation.

$$\frac{\partial p}{\partial t} = \nabla(D\nabla p) + Ap(1-p)(p-\sigma) - \alpha pm$$
$$\frac{\partial m}{\partial t} = \nabla(E\nabla m) + Bp(1-p)(p-\gamma) - \beta pm$$

Both populations are subject to an Allee effect and a high concentration of one species in a certain area causes the other species to decline. A great number of tests done during this study, even ones not presented in this dissertation, show that asymptotically the solutions for p and m tend to constant surfaces with at least one species' extinction. The precise circumstances leading to one species' domination over the other are elusive since a change in any of the parameters can lead to a change in the

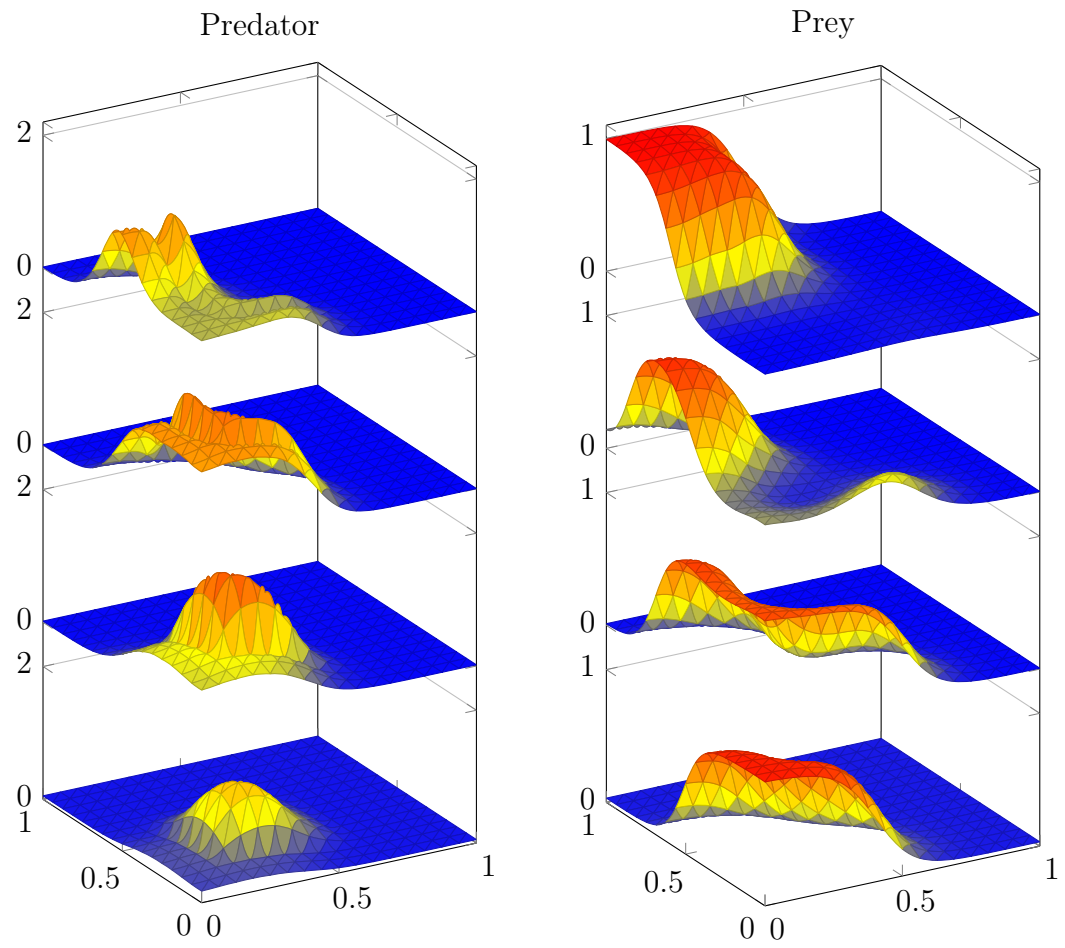


Figure 4.13: Population density over time of predator and prey in Example 4.4.3. The vertical axis shows population density as a percentage of population capacity at four points in time: $t \in \{3, 6, 8, 12\}$, where the bottom surface represents $t = 3$, and the top surface represents $t = 12$. The initial population distributions are $p = 0.1$ almost everywhere but with a localized bump.

long-term survivor. If a certain patch of $\Omega_1 \subset \Omega$ is more favorable to the growth rate of one species and the complement $\Omega_2 = \Omega \backslash \Omega_1$ is more favorable to the other species, then the two will coexist.

We present a few examples showing the effect of different diffusion rates, different choice of Allee threshold, and mildly different initial conditions. Our tests show that changes to any of these variables, while keeping all others the same, can cause one species to outlive the other. We also present an example of heterogeneous growth

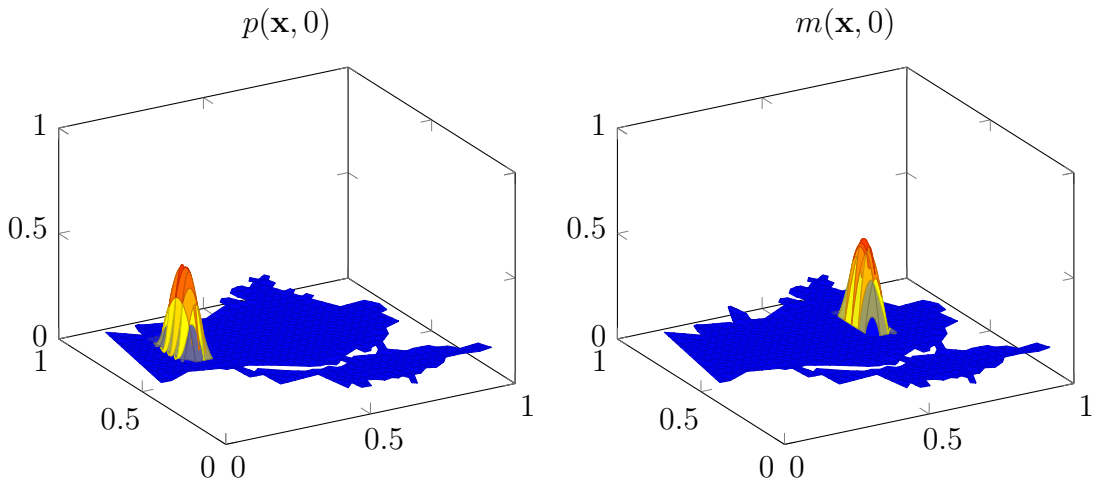


Figure 4.14: Some initial population densities for species p and m used for the competition model. Apart from the bumps in each density function, both populations are constant with density 0.1.

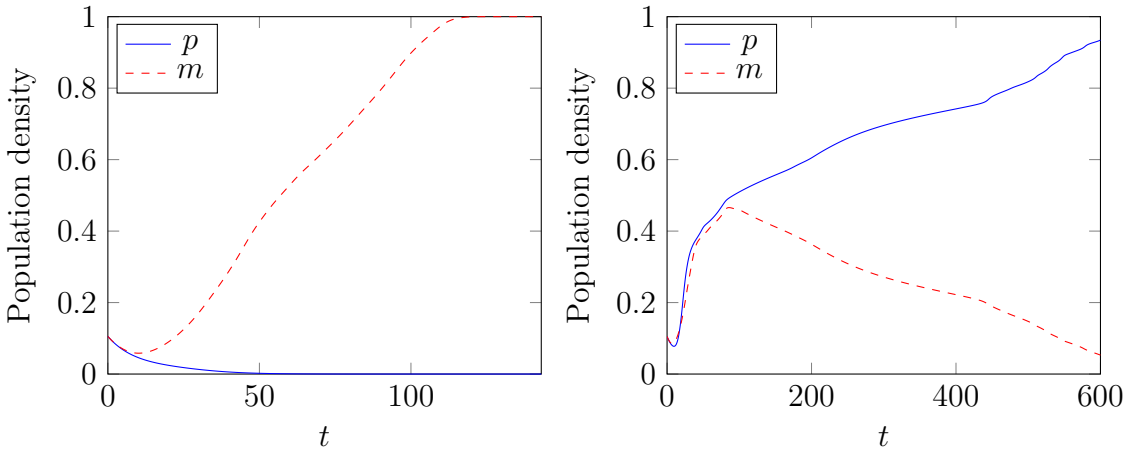
rates which led to coexistence.

Example 4.4.4. In this example all parameters are kept equal except for a difference in rate of diffusion for the two species.

D	E	A	B	α	β	σ	γ	h	T
3×10^{-4}	1×10^{-4}	1	1	1	1	0.1	0.1	0.01	600

The initial conditions are shown in Figure 4.14. The results of this example are presented in Figure 4.15a as average populations over the domain Ω since the precise evolution of the surface is not very interesting in this example. We can see that the slower diffuser prevails while the faster diffuser becomes extinct. Slower diffusion seems to provides an advantage, but as we shall see in subsequent examples, it is not a guarantee for the survival of a species.

Example 4.4.5. We now make a modification to the parameters from Example 4.4.4 to show that the difference in diffusion is not sufficient to guarantee the survival of



(a) Results of Example 4.4.4.

(b) Results of Example 4.4.5.

Figure 4.15: Average population over time for a pair of species competing for a common resource. Only one species survives in the long run.

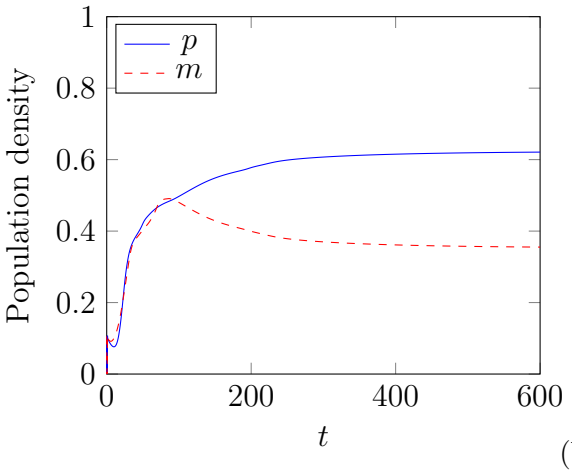
one species over the other. In this example we decrease the Allee threshold for both species to 0.05 and observe a reversal of the long-term survivor.

D	E	A	B	α	β	σ	γ	h	T
3×10^{-4}	1×10^{-4}	1	1	1	1	0.05	0.05	0.01	600

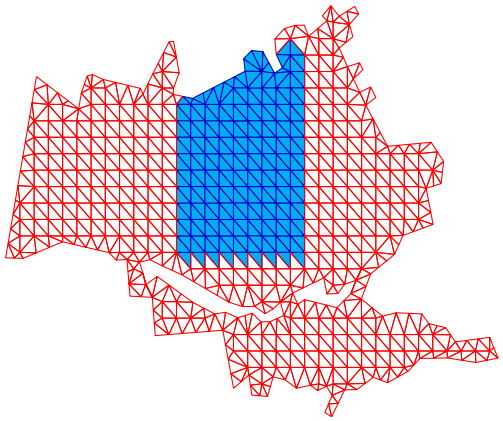
The initial conditions are the same as in Example 4.4.4, see Figure 4.14. The results of this example are presented in Figure 4.15b as average populations over the domain Ω. We can see that the faster diffuser prevails while the slower diffuser becomes extinct.

Example 4.4.6. We now show an example in which make full use of the spatial heterogeneity of the model by splitting Ω into two patches, each of which offers a more favorable growth rate to one species compared to the other. See Figure 4.16b for a description of the patches. The favorable growth rate is 20% higher than the baseline.

D	E	A	B	α	β	σ	γ	h	T
3×10^{-4}	1×10^{-4}	1 or 1.2	1 or 1.2	1	1	0.05	0.05	0.01	600



(a) Results of Example 4.4.6.



(b) Triangulation with two designated regions for Example 4.4.6.

Figure 4.16: Average population over time for a pair of species competing for a common resource, showing coexistence is possible. The blue, shaded region is favorable to species p and the red, unshaded region is favorable to species m.

The results can be seen in Figure 4.16a, which show that the two species coexist. It is interesting to examine the exact evolution of the population density functions in Figure 4.17. We see that a front forms along one of the edges where the terrain becomes more favorable to one species compared to the other. The majority of Ω is more favorable for the growth of species m, yet species p manages to maintain a presence in most of the western region.

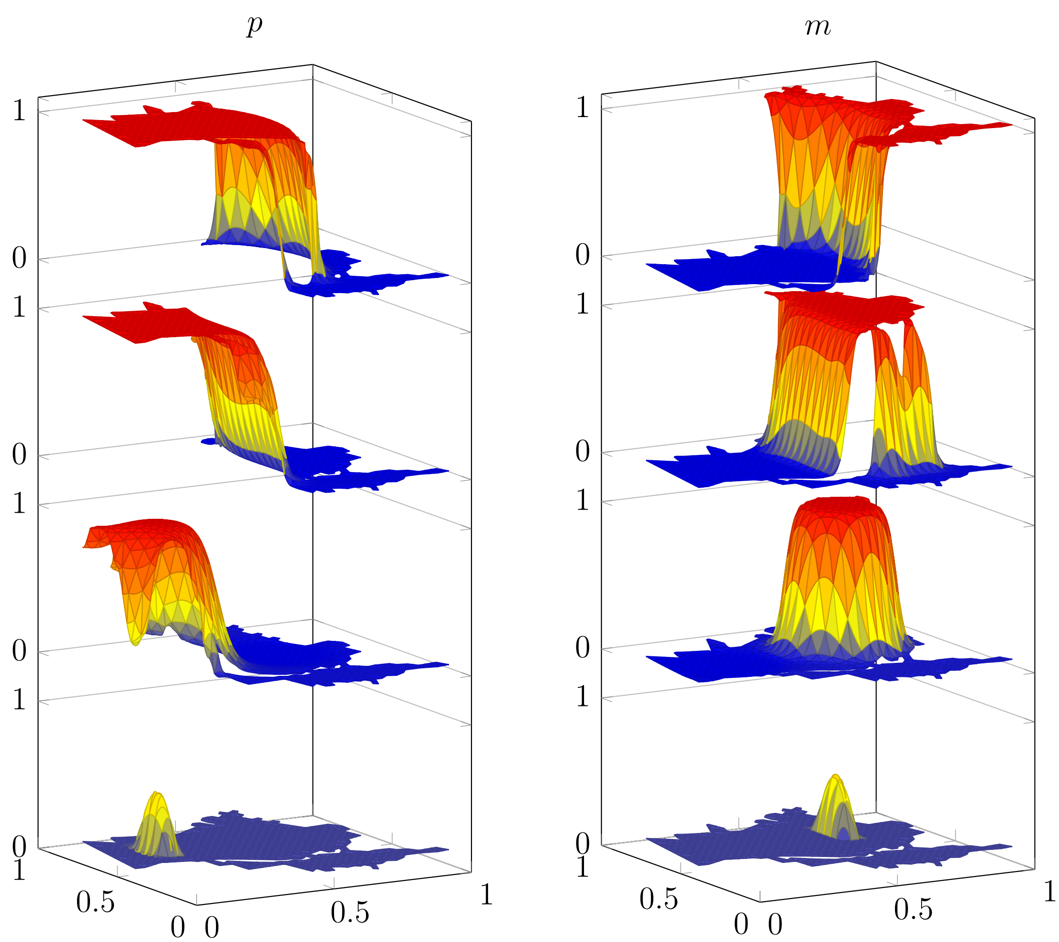


Figure 4.17: Population density over time of two species competing for a common resource in Example 4.4.6. The vertical axis shows population density as a percentage of population capacity at four points in time: $t \in \{0, 100, 200, 600\}$, where the bottom surface represents $t = 0$, and the top surface represents $t = 600$.

Chapter 5

Remarks and Future Research Problems

5.1 Higher Order Approximation of Time Derivative

In Chapter 4, Example 4.2.5 I presented evidence that an $O(h^2)$ approximation of the time derivative yields a substantial improvement in accuracy and thus it is strongly recommended. The scheme I used is based on the backward differentiation formula (BDF) of order two. I adapted BDF of order two to come up with the following modified discrete weak formulation.

$$\int_\Omega pq \, d\mathbf{x} = \int_\Omega \left[\frac{4}{3}\hat{p} - \frac{1}{3}\tilde{p}\right] q \, d\mathbf{x} - \frac{2}{3}h \int_\Omega D(\mathbf{x})\nabla p \cdot \nabla q \, d\mathbf{x} + \frac{2}{3}h \int_\Omega pF_1(p)q \, d\mathbf{x}$$

where $\hat{p} = p(\mathbf{x}, t_{i-1})$ as before and $\tilde{p} = p(\mathbf{x}, t_{i-2})$. Note that the equation fits the old framework by interpreting $\left[\frac{4}{3}\hat{p} - \frac{1}{3}\tilde{p}\right]$ as $\hat{p}$ in the old scheme and choosing a step size $\frac{2}{3}h$. A smaller step size can only help with the convergence results. In addition, $\left[\frac{4}{3}\hat{p} - \frac{1}{3}\tilde{p}\right]$ is also an element of $H_0^1(\Omega)$.

In order to invoke the theory, which we have already established for the existence, uniqueness and stability of the discrete weak solution, we need to guarantee that

$\left[\frac{4}{3}\hat{p} - \frac{1}{3}\tilde{p}\right]$ is in the admissible set $\mathcal{A}$. Since $\mathcal{A}$ is not a vector space, how can we guarantee that such a linear combination is an element in $\mathcal{A}$? Intuitively, the values of $\hat{p}$ and $\tilde{p}$ are comparable when h is small and so the resulting difference should be positive. However, it is unlikely that estimates obtained with H^1 norms will allow us to make such a conclusion. This question is interesting and warrants further investigation. In practice, it does not cause problems for the numerical implementation.

The algorithm can be initialized with $\tilde{p} = p(\mathbf{x}, t_0)$, which is the known initial data. In order to compute $\hat{p}$, one could naively take a single step using the tried and true $O(h)$ backward Euler scheme and then proceed using the second order. For my numerical experiments I elected to compute $\hat{p}$ using backward Euler, but in order not to lose crucial accuracy in that first step, I instead take several smaller time steps. For example, say $h = 0.1$. I would like to know $p(\mathbf{x}, 0.1)$ with error order $O(h^2)$, so that I can proceed to compute $p(\mathbf{x}, 0.2)$ with the same error order. I can accomplish that by taking ten steps using $h = 0.01$, thus taking me to $p(\mathbf{x}, 0.1)$ with $O(h^2)$ accuracy.

Theorem 2.1.1 can be modified to mimic the proof of the $O(h^2)$ convergence of BFD of order two used in ordinary differential equations. Such a proof need only make use of the various bounds already established on the growth term $F(p)$ and would further require that the classical solution $p(\mathbf{x}, t)$ is thrice differentiable.

5.2 Three or More Species

Theoretically, a general model for three or more species does not seem to present a serious roadblock. The optimization approach used to provide the justification for existence, uniqueness and stability of the discrete weak solution can be naturally extended to include a longer sum of energy terms, one for each species. Admittedly, I have not yet attempted to trace the details to confirm the viability of this approach, so this provides a direction for further research.

The numerical implementation of a system of reaction-diffusion equations incorporating three or more species would present a challenge. The algorithm might not scale as well as one might hope because of the nature of the implicit algorithm used to discretize the time variable, namely the need to run an iterative algorithm to solve for each time discrete time slice $p(\mathbf{x}, t_i)$. It would be worth exploring exactly how serious the computational complexity really is. There should be ample opportunities for parallelization.

5.3 Finding Appropriate Parameters

A much more daunting prospect is to solve the inverse problem of finding parameters for the reaction-diffusion equation, which would produce the best model to fit a particular set of empirical measurements. That is, given samples of population density $\{p_{ij}\}_{i\in[1\ldots N],j\in[1\ldots M]}$ at some discrete times t_i and discrete locations $\mathbf{x}_j$, find a diffusive factor $D(\mathbf{x})$, a growth function $A(\mathbf{x})$, and an Allee threshold σ which minimize the l^2 error

$$\sum_{i=1}^{N}\sum_{j=1}^{M}|p(\mathbf{x}_j, t_i) - p_{ij}|^2$$

This is a nontrivial problem, as any inverse problem tends to be. I have attempted to solve it using sequential quadratic programming, which is a standard numerical minimization scheme, but the problem proved too large for the algorithm to produce a solution in reasonable time. Some exploitation of the particular structure of this problem will be necessary if it is to be solved with a practical running time.

Chapter 6

Appendix A: Preliminary on Bivariate Splines

In this section, we explain bivariate spline functions of any degree d and smoothness $r \geq 1$ over arbitrary triangulation $\triangle$. Most of the following discussion can be found in [24]. We outline these functions here just for convenience. Let Ω be a polygonal domain in $\mathbb{R}^2$ and $\triangle$ a triangulation of Ω. That is, $\triangle$ is a finite collection of triangles $T \subset \Omega$ such that $\cup_{T \in \triangle} T = \Omega$ and the intersection of any two triangles is either the empty set, a common edge, or a common vertex. For each $T \in \triangle$, let $|T|$ denote the length of the longest edge of T, and let ρ_T be the radius of the inscribed circle of T. The longest edge length in the triangulation $\triangle$ is denoted by $|\triangle|$ and is referred to as the size of the triangulation. For any triangulation $\triangle$ we define its shape parameter by

$$\kappa_{\triangle} := \frac{|\triangle|}{\rho_{\triangle}}, \tag{6.0.1}$$

where $\rho_{\triangle}$ is the minimum of the radii of the in-circles of the triangles of $\triangle$. The shape parameter for a single triangle, κ_T, satisfies

$$\kappa_T := \frac{|T|}{\rho_T} \leq \frac{2}{\tan(\theta_T/2)} \leq \frac{2}{\sin(\theta_T/2)}, \tag{6.0.2}$$

where θ_T is the smallest angle in the triangle T. The shape of a given triangulation affects how well we can approximate a function over the triangulation. Hence we have the following definition of a β-quasi-uniform triangulation.

Definition 6.0.1 (β-Quasi-Uniform Triangulation)**.** Let $0 < \beta < \infty$. A triangulation $\triangle$ is a *β-quasi-uniform* triangulation provided that

$$\frac{|\triangle|}{\rho_\triangle} \leq \beta.$$

Once we have a triangulation, we define the spline space of degree d and smoothness r over that triangulation as follows:

Definition 6.0.2 (Spline Space)**.** Let $\triangle$ be a given triangulation of a domain Ω. Then we define the spline space of smoothness r and degree d over $\triangle$ by,

$$S_d^r(\triangle) = \{s \in C^r(\Omega) \mid s|_T \in \mathcal{P}_d, \ \forall\, T \in \triangle\},$$

where $\mathcal{P}_d$ is the space of polynomials of degree at most d.

We next explain how to represent a spline function in $S_d^r(\triangle)$. Let

$$T = \langle (x_1, y_1), (x_2, y_2), (x_3, y_3) \rangle.$$

For any point (x, y), let b_1, b_2, b_3 be the solution of

$$\begin{aligned} x &= b_1 x_1 + b_2 x_2 + b_3 x_3 \\ y &= b_1 y_1 + b_2 y_2 + b_3 y_3 \\ 1 &= b_1 + b_2 + b_3. \end{aligned}$$

(b_1, b_2, b_3) are the so-called barycentric coordinates of (x, y) with respect to T. Note that b_i is a linear polynomial of (x, y) for $i = 1, 2, 3$. Fix a degree $d > 0$. For

$i+j+k=d$, let

$$B_{ijk}^T(x,y) = \frac{d!}{i!j!k!} b_1^i b_2^j b_3^k$$

which is called Bernstein-Bézier polynomial. Let

$$S|_T = \sum_{i+j+k=d} c_{ijk}^T B_{ijk}^T(x,y).$$

We use $\mathbf{s} = (c_{ijk}^T, i+j+k=d, T \in \triangle)$ to represent the coefficient vector for spline function $S \in S_d^{-1}(\triangle)$. In order to make $S \in S_d^0(\triangle)$, we have to construct a smoothness matrix H such that $H\mathbf{s} = 0$ ensure that S is a continuous function. Such a smoothness matrix is known and in fact it is known for any smoothness $r \geq 0$ (cf. [9]).

Note that Bernstein-Bézier representation of spline functions is very convenient for basic evaluation, derivatives and integration. We use the de Casteljau algorithm to evaluate a Bernstein-Bézier polynomial at any point inside the triangle. It is a simple and stable computation. See [24]. Let $T = \langle \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3 \rangle$ and $S|_T = \sum_{i+j+k=d} c_{ijk} B_{ijk}(x,y)$. Then the directional derivative $D_{\mathbf{v}_2-\mathbf{v}_1} S|_T$ is

$$D_{\mathbf{v}_2-\mathbf{v}_1} S|_T = d \sum_{i+j+k=d-1} (c_{i,j+1,k} - c_{i+1,j,k}) B_{ijk}(x,y).$$

Similar for $D_{\mathbf{v}_3-\mathbf{v}_1} S|_T$. D_x and D_y are linear combinations of these two directional derivatives. Let s be a spline with $s|_T = \sum_{i+j+k=d} c_{ijk}^T B_i jk(x,y), T \in \triangle$ in $S_d^r(\triangle)$. Then

$$\int_\Omega s(x,y)dxdy = \sum_{T\in\triangle} \frac{A_T}{\binom{d+2}{2}} \sum_{i+j+k=d} c_{ijk}^T.$$

If $p = \sum_{i+j+k=d} a_{ijk} B_{ijk}(x,y)$ and $q = \sum_{i+j+k=d} b_{ijk} B_{ijk}(x,y)$ over a triangle T, then

$$\int_T p(x,y)q(x,y)dxdy = \mathbf{a}^\top M_d \mathbf{b},$$

where $\mathbf{a} = (a_{ijk}, i+j+k=d)^\top$, $\mathbf{b} = (b_{ijk}, i+j+k=d)^\top$, M_d is a symmetric

matrix with known entries (a formula for these entries is known (cf. [24]). These elementary operations have been implemented in MATLAB. See [1]. Many different linear and nonlinear partial differential equations have been solved by using these bivariate spline functions. See [25], [1], [17].

When $d \geq 3r+2$ the spline space $S_d^r(\triangle)$ possesses an optimal approximation order which is achieved by the use of a quasi-interpolation operator. Let $\|f\|_{L_p(\Omega)}$ denote the usual L_p norm of f over Ω, $|f|_{m,p,\Omega}$ denotes the L_p norm of the m^{th} derivatives of f over Ω, and $W_p^{m+1}(\Omega)$ stands for the usual Sobolev space over Ω.

To define the quasi-interpolation operator, we need a set of linear functionals

$$\{\lambda_{ijk,T} | i+j+k=d, T \in \triangle\},$$

which are based on values of f at the set of domain points over triangles in $\triangle$, that is

$$\lambda_{ijk,T}(f) = \sum_{|\nu|=d} a_\nu^{ijk} f(\xi_\nu^T), \tag{6.0.3}$$

where $\xi_\nu^T = (i\mathbf{v}_1^T + j\mathbf{v}_2^T + k\mathbf{v}_3^T)/d$ for $\nu = (i,j,k)$ with $i+j+k=d$ and $\mathbf{v}_i, i=1,2,3$ are vertexes of triangle T.

A quasi-interpolation operator of f is defined by

$$Qf := \sum_{T\in\triangle} \sum_{i+j+k=d} \lambda_{ijk,T}(f) B_{ijk}^T. \tag{6.0.4}$$

Now, we are ready to state a theorem on optimal approximation order (cf. [23] and [24]).

Theorem 6.0.1 (Optimal Approximation Order)**.** *Assume $d \geq 3r+2$ and let $\triangle$ be a triangulation of Ω. Then there exists a quasi-interpolatory operator $Qf \in S_d^r(\triangle)$ mapping $f \in L_1(\Omega)$ into $S_d^r(\triangle)$ such that Qf achieves the optimal approximation*

order: if $f \in W_p^{m+1}(\Omega)$,

$$\|D_x^\alpha D_y^\beta (Qf - f)\|_{L_p(\Omega)} \leq C|\triangle|^{m+1-\alpha-\beta}|f|_{m+1,p,\Omega} \tag{6.0.5}$$

for all $\alpha + \beta \leq m + 1$ *with* $0 \leq m \leq d$, *where* D_x *and* D_y *denote the derivatives with respect to the first and second variables and the constant* C *depends only on the degree* d *and the smallest angle* $\theta_\triangle$ *and may be dependent on the Lipschitz condition on the boundary of* Ω.

We sometimes need to use the so-called Markov inequality to compare the size of the derivative of a polynomial with the size of the polynomial itself on a given triangle t. As a spline function is a piecewise polynomial function, this inequality can be also applied to any spline function. See [24] for a proof.

Theorem 6.0.2. *Let* $t := \langle v_1, v_2, v_3 \rangle$ *be a triangle, and fix* $1 \leq q \leq \infty$. *Then there exists a constant* K *depending only on* d *such that for every polynomial* $p \in \mathcal{P}_d$, *and any nonnegative integers* α *and* β *with* $0 \leq \alpha + \beta \leq d$,

$$\|D_1^\alpha D_2^\beta p\|_{q,t} \leq \frac{K}{\rho_t^{\alpha+\beta}} \|p\|_{q,t}, \quad 0 \leq \alpha + \beta \leq d, \tag{6.0.6}$$

where ρ_t *denotes the radius of the largest circle inscribed in* t.

More detail on the theory of bivariate splines can be found in [24] and their computational schemes in [1].

Bibliography

[1] G. Awanou, M.-J. Lai, and P. Wenston. The multivariate spline method for scattered data fitting and numerical solutions of partial differential equations. *Wavelets and splines: Athens*, pages 24–74, 2005.

[2] R. S. Cantrell and C. Cosner. *Spatial ecology via reaction-diffusion equations*. John Wiley & Sons, 2004.

[3] D. Coulibaly, S. Rebaudet, M. Travassos, Y. Tolo, M. Laurens, A. Kone, K. Traore, A. Guindo, I. Diarra, A. Niangaly, M. Daou, A. Dembele, M. Sissoko, B. Kouriba, N. Dessay, J. Gaudart, R. Piarroux, M. Thera, C. Plowe, and O. Doumbo. Spatio-temporal analysis of malaria within a transmission season in Bandiagara, Mali. *Malaria Journal*, 12(1):82, 2013.

[4] D. DeAngelis and L. Gross. *Individual-Based Models and Approaches in Ecology: Populations, Communities and Ecosystems*. Chapman & Hall, New York, 1992.

[5] A. Dicko, C. Mantel, B. Kouriba, I. Sagara, M. A. Thera, S. Doumbia, M. Diallo, B. Poudiougou, M. Diakite, and O. K. Doumbo. Season, fever prevalence and pyrogenic threshold for malaria disease definition in an endemic area of mali. *Tropical Medicine & International Health*, 10(6):550–556, 2005.

[6] Y. Du and J. Shi. Some recent results on diffusive predator-prey models in spatially heterogeneous environment, in: Nonlinear dynamics and evolution equations. *Fields Inst. Commun*, pages 95–135, 2006.

[7] J. Eilbeck, J. Furter, and J. Lopez-Gomez. Coexistence in the competition model with diffusion. *Journal of differential equations*, 107(1):96–139, 1994.

[8] L. C. Evans. *Partial Differential Equations (Volume 19 of Graduate studies in mathematics)*. American Mathematical Society, 1998.

[9] G. Farin. Triangular bernstein-bézier patches. *Computer Aided Geometric Design*, 3(2):83–127, 1986.

[10] R. A. Fisher. The wave of advance of advantageous genes. *Annals of Eugenics*, 7(4):355–369, 1937.

[11] C. Folt. Predation: Direct and indirect impacts on aquatic communities. chapter An experimental analysis of costs and benefits of zooplankton aggregation, pages 300–314. Univ. Press of New England, New Hampshire, 1987.

[12] W. Foster and J. Treherne. Evidence for the dilution effect in the selfish herd from fish predation on a marine insect. 1981.

[13] G. Gause. Experimental demonstration of volterra's periodic oscillations in the numbers of animals. *Journal of Experimental Biology*, 12(1):44–48, 1935.

[14] M. Gilpin and M. Soule. Conservation biology: The science of scarcity and diversity. chapter Inbreeding in natural populations of birds and mammals, pages 35–56. Sinauer Associates, Massachusetts, 1986.

[15] M. Gilpin and M. Soule. Conservation biology: The science of scarcity and diversity. chapter Minimum viable populations: Processes of species extinctions, pages 19–34. Sinauer Associates, Massachusetts, 1986.

[16] C. S. Holling. Some characteristics of simple types of predation and parasitism. *The Canadian Entomologist*, 91(07):385–398, 1959.

[17] X.-L. Hu, D.-F. Han, and M.-J. Lai. Bivariate splines of various degrees for numerical solution of partial differential equations. *SIAM Journal on Scientific Computing*, 29(3):1338–1354, 2007.

[18] C. Huffaker. Biological control of weeds with insects. *Annual review of entomology*, 4(1):251–276, 1959.

[19] G. E. Hutchinson. Homage to santa rosalia or why are there so many kinds of animals? *The American Naturalist*, 93(870):145–159, 1959.

[20] R. Kenward. Hawks and doves: factors affecting success and selection in goshawk attacks on woodpigeons. *The Journal of Animal Ecology*, pages 449–460, 1978.

[21] H. Kruuk. Predators and anti-predator behaviour of the black-headed gull (larus ridibundus l.). *Behaviour. Supplement*, pages III–129, 1964.

[22] M.-J. Lai and C. Meile. Scattered data interpolation with nonnegative preservation using bivariate splines. *Computer Aided Geometric Design*, 34:37–49, 2015.

[23] M.-J. Lai and L. L. Schumaker. On the approximation power of bivariate splines. *Advances in Computational Mathematics*, 9(3-4):251–279, 1998.

[24] M.-J. Lai and L. L. Schumaker. *Spline functions on triangulations*. Number 110 in Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2007.

[25] M.-J. Lai and P. Wenston. Bivariate splines for fluid flows. *Computers & fluids*, 33(8):1047–1073, 2004.

[26] M. Lewis and P. Kareiva. Allee dynamics and the spread of invading organisms. *Theoretical Population Biology*, 43(2):141–158, 1993.

[27] A. Lotka. Contribution to the theory of periodic reaction. *J. Phys. Chem.*, 14(3):271–274, 1910.

[28] T. R. Malthus. *An Essay on the Principle of Population Or a View of Its Past and Present Effects on Human Happiness, an Inquiry Into Our Prospects Respecting the Future Removal Or Mitigation of the Evils which it Occasions by Rev. TR Malthus*. Reeves and Turner, 1872.

[29] J. Neuberger. *Sobolev gradients and differential equations*. Springer Science & Business Media, 2009.

[30] R. Pulliam and T. Caraco. Behavioral ecology: An evolutionary approach. chapter Living in groups: Is there an optimal group size? Sinauer Associates, Massachusetts, 1984.

[31] O. Richter, S. Moenickes, and F. Suhling. Modelling the effect of temperature on the range expansion of species by reaction–diffusion equations. *Mathematical biosciences*, 235(2):171–181, 2012.

[32] M. Solomon. The natural control of animal populations. *The Journal of Animal Ecology*, pages 1–35, 1949.

[33] P. Turchin and P. Kareiva. Aggregation in aphis varians: an effective strategy for reducing predation risk. *Ecology*, 70(4):1008–1016, 1989.

[34] P.-F. Verhulst. Notice sur la loi que la population poursuit dans son accroissement correspondance mathématique et physique 10: 113-121. Technical report, Retrieved 09/08, 2009.

[35] V. Volterra. Fluctuations in the abundance of a species considered mathematically. *Nature*, 118:558–560, 1926.

[36] M. Way and C. Banks. Intra-specific mechanisms in relation to the natural regulation of numbers of aphis fabae. *An. Appl. Biol.*, 59:529–565, 1967.

[37] M. Way, M. Cammell, A. Watson, et al. Aggregation behaviour in relation to food utilization by aphids. In *Animal populations in relation to their food resources. A symposium of the British Ecological Society, Aberdeen 24-28 March 1969.*, pages 229–247. Oxford and Edinburgh, Blackwell., 1970.

[38] World Health Organization Expert Committee on Malaria. 18th report. Technical report, Geneva: WHO, 1986.

[39] World Health Organization Expert Committee on Malaria. 20th report. Technical report, Geneva: WHO, 2000.

[40] World Health Organization Expert Committee on Malaria. World malaria report. Technical report, Geneva: WHO, 2011.