

Master Thesis

# The Criticality Spectral Problem in Neutron Transport

Severin Thaler

August 20, 2012

Supervised by Prof. Christoph Schwab

### Abstract

This master thesis is about the criticality spectral problem in neutron transport. This is an eigenvalue problem used to determine whether a nuclear reactor is sub- or supercritical. More precisely, there is a certain eigenpair which contains the criticality information. However, such an eigenpair does not necessarily exist, so that certain conditions have to be formulated under which the existence can be shown. In this respect, the Krein-Rutman Theorem for positive compact operators plays a central role. Once the existence is established, a variational formulation of the problem is introduced. This variational formulation turns out to be quite intricate, since it does not seem to fit into existing frameworks for variationally posed eigenvalue problems. Moreover, when it comes to the finite element discretization, the right choice for the discrete trial and test spaces is not obvious. In particular, one must decide whether or not to deal with a non-square matrix eigenvalue problem. After these problems have been roughly discussed, a hierarchy of model problems is considered. For the (1 + 1)-dimensional case, a concrete finite element discretization is formulated, which is then generalized to the (2+2)-dimensional case. In doing so, the efficient evaluation of the bilinear forms, especially the scattering bilinear form, is addressed. Then, a generalized finite element method is sketched, when the number of cells of the reactor is large. Finally, numerical examples for the standard finite element method are provided.

# Contents

1	Introduction		2
2	<b>Prol</b> 2.1 2.2 2.3	<b>Dem Formulation</b> The Neutron Transport Problem         The Criticality Spectral Problem         Functional-Analytic Framework	<b>4</b> 4 8 9
3	Basic Existence Result 18		
	3.1	Formulation as an ordinary EVP	18
	3.2	Existence of the Eigenpair	23
4	Standard FEM for the Criticality Problem 32		
	4.1	Variational Formulation	32
	4.2	Model Problems	37
		4.2.1 1D space & 0D velocity	38
		4.2.2 2D space & 0D velocity	42
		4.2.3 0D space & 1D velocity	44
		4.2.4 1D space & 1D velocity	45
		4.2.5 Further Model Problems	47
	4.3	Choice of the FE subspace in 1D	48
	4.4	Fast Quadrature	52
	4.5	Choice of the FE subspace in 2D	53
5	Generalized FEM 60		60
	5.1	Homogenization and the Unit Cell Problem	60
	5.2	g-FEM	61
6	Numerical Examples & Outlook		64
	6.1	1+1D Global	64
	6.2	1+1D Unit Cell	66
	6.3	Outlook	68
Bi	Bibliography		

## Danke

Zuerst einmal möchte ich mich bei Herrn Schwab bedanken. Dafür, dass ich überhaupt eine Arbeit bei Ihnen schreiben durfte und für das spezielle Thema, das Sie mir gegeben haben. Dann muss ich mich ganz fest bei Holger Brandsmeier bedanken. Ohne Dich wäre ich in der Implementation garantiert nicht so weit gekommen. Aber auch in der Theorie hast Du mich auf einige Ideen gebracht. Du hast mir gezeigt, dass Mathematik nicht immer so schwierig sein muss. Dann bedanke ich mich bei Roman Andreev. Dafür, dass ich einfach in Dein Büro platzen durfte um Dir Fragen aller Art zu stellen. So ist beispielsweise unsere Diskussion über die optimale Wahl des Test- und Trialraumes sicher noch nicht beendet. Dann bedanke ich mich bei meinen Eltern. Danke, dass Ihr so viel Geduld mit mir habt. Danke, dass Ihr immer an mich glaubt. Zum Schluss möchte ich noch Ho-Yan danken. Wer weiss, vielleicht liest Du diese Arbeit ja einmal und dann wirst Du verstehen, was ich Dir mit den Vektorräumen erzählen wollte, genauso wie ich dann verstehen werde, was Du mir erzählen wolltest. Ich wünsche Dir alles Gute.

## **1** Introduction

In a nuclear reactor, neutrons are absorbed by fissile nuclei (e.g. uranium-235), which then split into new atoms, releasing binding energy and new neutrons. Some of these new neutrons are absorbed by other fissile nuclei, leading to a chain reaction. An important question is whether this chain reacting system is sub- or supercritical. If it is subcritical, it is not self-sustaining, i.e. the chain reaction will die out. If it is supercritical on the other hand, then the reactor will get out of control. An efficient reactor should be exactly critical. To answer this question, a certain eigenvalue problem, the criticality spectral problem, is studied. With the help of its first eigenvalue, the reactor's state can be determined. However, since the criticality spectral problem is posed in phase space, the computational complexity is high, especially when the number of fuel rods of the reactor is large. Hence, complexity reducing techniques are necessary. Among them is the homogenization technique (see [1]), which is based on the idea that for a large number of identical fuel rods, the eigenvalue problem posed on the whole reactor approximately reduces to an eigenvalue problem posed on one rod, and a homogenized diffusion problem posed on the whole reactor. More precisely, the (global) eigenfunction approximately decomposes into a product of two terms. One term is the first eigenfunction of the same criticality problem, but posed on one rod, and with periodic boundary conditions. The other term is the first eigenfunction of the homogenized diffusion problem, and does not depend on the velocity. But since the eigenfunction of the diffusion problem satisfies Dirichlet boundary conditions, this product will not only satisfy no-inflow, but also nooutflow boundary conditions, which is unphysical. One can therefore not expect that this product will be a good approximation to the global eigenfunction at the boundary of the reactor. The motivation for this thesis was therefore to introduce a new numerical approach to the criticality spectral problem, a generalized finite element method (see [2] for an overview), which is related to the homogenization idea (see [14]). However, for the criticality spectral problem, not even the standard FEM is well-established. Thus, the thesis focuses on the standard FEM and addresses the generalized FEM at the very end.

The thesis begins with the time-dependent neutron transport problem and then introduces the (time-independent) criticality spectral problem. We then describe the functional-analytic framework needed to formulate this eigenvalue problem in terms of function spaces and operators. In Section 3 we show the existence of the positive eigenpair using the Krein-Rutman Theorem for positive compact operators, which is a generalization of the Perron-Frobenius Theorem for positive matrices. The positivity appears naturally since the data functions are physical quantities. The compactness is established using a result about regular integral operators. In Section 4 we introduce a variational formulation of the criticality spectral problem for which an abstract framework is then sought, in the hope that this framework will provide error estimates for the finite element method. Some existing frameworks for variationally posed eigenvalue problems are examined and it is shown that none of them is quite appropriate for our problem. We then proceed with a hierarchy of model problems which serve several purposes. First, they have been used as test problems to validate our implementation along the way. Second, they make explicit the geometry of the reactor and the velocity domain, which makes it easier to imagine the situation. Third, they also improve the understanding of some theoretical aspects. Next, we concretely describe how to choose the FE-subspaces in the (1 + 1)-dimensional setting. This choice might not be optimal in terms of a discrete inf-sup condition, but at least leads to a square matrix eigenvalue problem. We then make a little excursion to the sum factorization technique, a fast quadrature for high-order tensor product polynomial shape functions. But since this technique only becomes relevant in higher dimensions, we proceed with the (2+2)-dimensional setting, for which we again describe how to choose the FE-subspaces explicitly, but then also explain how the fast quadrature can be used to efficiently evaluate the bilinear forms, even in the setting of the scattering bilinear form, which does not have tensor structure. In Section 5 we briefly introduce the homogenization idea together with the unit cell problem, and then sketch a generalized FEM. Finally, in Section 6, numerical examples are given, including the global- and the unit cell eigenfunction. The thesis ends with an outlook addressing interesting open problems.

## **2** Problem Formulation

### 2.1 The Neutron Transport Problem

We want to describe the population of free neutrons in the core  $\Omega \subset \mathbb{R}^3$  of a nuclear reactor. A neutron's state is determined by

- its position  $\boldsymbol{x} \in \Omega$ ;
- its velocity  $\boldsymbol{v} \in V \subset \mathbb{R}^3$ .

**Remark 2.1.1.** We note that in most of the literature the variables  $\mathbf{x}, \boldsymbol{\omega}, E$  are used, where  $\boldsymbol{\omega}$  is the direction of the neutron's velocity (a unit vector, i.e.  $\boldsymbol{\omega} \in S^2$ ) and E is its kinetic energy. However, the variables  $\mathbf{x}, \boldsymbol{\omega}, E$  do not contain more information than the variables  $\mathbf{x}, \mathbf{v}$ . That is why we shall use this minimal set of variables.

The neutrons are considered to have their velocity in just a subset V of  $\mathbb{R}^3$ . This is reasonable because a neutron's velocity cannot be arbitrary (e.g. it cannot exceed the speed of light). The neutron population is described by a function  $n(\boldsymbol{x}, \boldsymbol{v}, t)$ , so that  $n(\boldsymbol{x}, \boldsymbol{v}, t) \,\mathrm{d}\boldsymbol{x} \,\mathrm{d}\boldsymbol{v}$  is the expected number of neutrons at time t

- being in the volume element dx around the point x,
- having velocity in the element dv around v.

To have an equation for n(x, v, t) we need certain data about how the neutrons interact with the environment. We introduce

- $\Sigma_t(\boldsymbol{x}, \boldsymbol{v})$ : total macroscopic cross section. This is the probability per unit path length traveled that a neutron at position  $\boldsymbol{x}$  with velocity  $\boldsymbol{v}$  will undergo any kind of interaction with the environment.
- $\Sigma_s(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v})$  : macroscopic scattering cross section. This function is such that  $\Sigma_s(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \, \mathrm{d}\boldsymbol{v}$  is the probability per unit path length traveled that a neutron at position  $\boldsymbol{x}$  with velocity  $\boldsymbol{v}'$  will undergo scattering and have its velocity in  $\mathrm{d}\boldsymbol{v}$  afterwards.
- $\Sigma_f(\boldsymbol{x}, \boldsymbol{v}')$ : macroscopic fission cross section. This is the probability per unit path length traveled that a neutron at position  $\boldsymbol{x}$  with velocity  $\boldsymbol{v}'$  will induce a fission reaction.

- $\nu(x, v')$ : average number of neutrons emitted in a fission reaction which is induced by a neutron at position x with velocity v'.
- $\chi(\boldsymbol{x}, \boldsymbol{v})$ : neutron fission spectrum. This function is such that  $\chi(\boldsymbol{x}, \boldsymbol{v}) d\boldsymbol{v}$  is the probability that a neutron emitted in a fission reaction, which is induced by a neutron at position  $\boldsymbol{x}$ , will have its velocity in  $d\boldsymbol{v}$ .

**Remark 2.1.2.** For the exact meanings of these data functions, especially the cross sections, we refer the reader to [5] or [11].

We further denote

- $\Sigma(\boldsymbol{x}, \boldsymbol{v}) \coloneqq |\boldsymbol{v}| \Sigma_t(\boldsymbol{x}, \boldsymbol{v}),$
- $f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \coloneqq |\boldsymbol{v}'| \Sigma_s(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}),$
- $\sigma(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \coloneqq \chi(\boldsymbol{x}, \boldsymbol{v}) \nu(\boldsymbol{x}, \boldsymbol{v}') |\boldsymbol{v}'| \Sigma_f(\boldsymbol{x}, \boldsymbol{v}').$

Following [5], we are going to express the variation

 $\underbrace{n(\boldsymbol{x}, \boldsymbol{v}, t + \mathrm{d}t) \,\mathrm{d}\boldsymbol{x} \,\mathrm{d}\boldsymbol{v}}_{\text{no. of neutrons in }\mathrm{d}\boldsymbol{x} \,\mathrm{d}\boldsymbol{v} \,\mathrm{at time }t + \mathrm{d}t}_{t + \mathrm{d}t} - \underbrace{n(\boldsymbol{x}, \boldsymbol{v}, t) \,\mathrm{d}\boldsymbol{x} \,\mathrm{d}\boldsymbol{v}}_{\text{no. of neutrons in }\mathrm{d}\boldsymbol{x} \,\mathrm{d}\boldsymbol{v}}_{\text{at time }t} = \frac{\partial}{\partial t} n(\boldsymbol{x}, \boldsymbol{v}, t) \,\mathrm{d}\boldsymbol{x} \,\mathrm{d}\boldsymbol{v} \,\mathrm{d}t.$ 

This variation is mainly due to the following four physical processes.

- (1) Advection. Neutrons streaming into and neutrons leaking out of dx.
- (2) **Inscattering.** Neutrons in dx undergoing scattering and having their new velocity in dv.
- (3) Absorption & Outscattering. Neutrons in dx having their velocity in dv initially but undergoing any kind of interaction with the environment so that their new velocity will not be in dv anymore.
- (4) Fission. Neutrons produced by fission reactions in dx.

For each of these processes we are going to write down the gained (resp. lost) number of neutrons.

(1) Advection. Let  $d\sigma$  be a surface element limiting the volume element dx. The number of neutrons with velocity in dv traveling through  $d\sigma$  during the time dt is

$$\boldsymbol{v} n(\boldsymbol{x}, \boldsymbol{v}, t) \cdot \mathrm{d}\boldsymbol{\sigma} \mathrm{d}\boldsymbol{v} \mathrm{d}t.$$

Summing over all surface elements and using Gauss' Law as well as the fact that the volume element dx is infinitely small, we obtain

$$\boldsymbol{v} \cdot \nabla n(\boldsymbol{x}, \boldsymbol{v}, t) \, \mathrm{d} \boldsymbol{x} \, \mathrm{d} \boldsymbol{v} \, \mathrm{d} t$$

for the net number of neutrons with velocity in dv leaving or entering the element dx during dt.

② Inscattering. At time t, the number of neutrons in  $d\mathbf{x}$  with velocity in  $d\mathbf{v}'$  is  $n(\mathbf{x}, \mathbf{v}', t) d\mathbf{x} d\mathbf{v}'$ . The probability for each of these neutrons to undergo scattering and to have its new velocity in  $d\mathbf{v}$  during the time interval dt is

$$\Sigma_s(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) |\boldsymbol{v}'| \, \mathrm{d}t \, \mathrm{d}\boldsymbol{v}.$$

Thus, the number of neutrons in dx with velocity in dv' which scatter and have their new velocity in dv during the time interval dt is

$$|\boldsymbol{v}'|\Sigma_s(\boldsymbol{x},\boldsymbol{v}',\boldsymbol{v})n(\boldsymbol{x},\boldsymbol{v}',t) \,\mathrm{d}\boldsymbol{v}' \,\mathrm{d}\boldsymbol{x} \,\mathrm{d}\boldsymbol{v} \,\mathrm{d}t.$$

However, we must consider contributions from any v'. Hence, the number of neutrons in dx undergoing scattering and having their new velocity in dv during the time interval dt is

$$\left(\int_{V} |\boldsymbol{v}'| \Sigma_{s}(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) n(\boldsymbol{x}, \boldsymbol{v}', t) \, \mathrm{d}\boldsymbol{v}'\right) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v} \, \mathrm{d}t = \left(\int_{V} f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) n(\boldsymbol{x}, \boldsymbol{v}', t) \, \mathrm{d}\boldsymbol{v}'\right) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v} \, \mathrm{d}t$$

(3) Absorption & Outscattering. At time t, the number of neutrons in dx with velocity in dv is n(x, v, t) dx dv. The probability for each of these neutrons to undergo any kind of interaction with the environment during the time interval dt is given by

$$\Sigma_t(\boldsymbol{x}, \boldsymbol{v}) |\boldsymbol{v}| \, \mathrm{d}t.$$

Thus, the number of neutrons in dx undergoing any kind of interaction with the environment during the time interval dt is

$$|\boldsymbol{v}|\Sigma_t(\boldsymbol{x},\boldsymbol{v})n(\boldsymbol{x},\boldsymbol{v},t) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v} \, \mathrm{d}t = \Sigma(\boldsymbol{x},\boldsymbol{v})n(\boldsymbol{x},\boldsymbol{v},t) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v} \, \mathrm{d}t.$$

(4) **Fission**. At time t, the number of neutrons in  $d\mathbf{x}$  with velocity in  $d\mathbf{v}'$  is  $n(\mathbf{x}, \mathbf{v}', t) d\mathbf{x} d\mathbf{v}'$ . The probability for each of these neutrons to induce a fission reaction during the time interval dt is

$$\Sigma_f(\boldsymbol{x}, \boldsymbol{v}') |\boldsymbol{v}'| \, \mathrm{d}t.$$

Hence, the number of neutrons among them which actually do induce a fission reaction is

$$|\boldsymbol{v}'| \Sigma_f(\boldsymbol{x}, \boldsymbol{v}') n(\boldsymbol{x}, \boldsymbol{v}', t) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v}' \, \mathrm{d}t.$$

In each such fission reaction  $\nu(x, v')$  many neutrons are emitted. So the number of neutrons produced is

$$u(\boldsymbol{x}, \boldsymbol{v}') | \boldsymbol{v}' | \Sigma_f(\boldsymbol{x}, \boldsymbol{v}') n(\boldsymbol{x}, \boldsymbol{v}', t) \, \mathrm{d} \boldsymbol{x} \, \mathrm{d} \boldsymbol{v}' \, \mathrm{d} t.$$

However, not all these emitted neutrons will have their velocity in dv. But  $\chi(x, v) dv$  gives us the fraction of them which do have their velocity in dv. Thus, we obtain

$$\chi(\boldsymbol{x},\boldsymbol{v})\nu(\boldsymbol{x},\boldsymbol{v}')|\boldsymbol{v}'|\Sigma_f(\boldsymbol{x},\boldsymbol{v}')n(\boldsymbol{x},\boldsymbol{v}',t)\;\mathrm{d}\boldsymbol{v}'\;\mathrm{d}\boldsymbol{x}\;\mathrm{d}\boldsymbol{v}\;\mathrm{d}t$$

for the number of neutrons produced by fission in  $d\mathbf{x}$ , induced by neutrons with velocity in  $d\mathbf{v}'$ , during the time interval dt. But again we have to take into account all velocities  $\mathbf{v}'$ . So the number of neutrons produced by fission in  $d\mathbf{x}$ , during the time interval dtis

$$\left( \int_{V} \chi(\boldsymbol{x}, \boldsymbol{v}) \nu(\boldsymbol{x}, \boldsymbol{v}') |\boldsymbol{v}'| \Sigma_{f}(\boldsymbol{x}, \boldsymbol{v}') n(\boldsymbol{x}, \boldsymbol{v}', t) \, \mathrm{d} \boldsymbol{v}' \right) \, \mathrm{d} \boldsymbol{x} \, \mathrm{d} \boldsymbol{v} \, \mathrm{d} t \\ = \left( \int_{V} \sigma(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) n(\boldsymbol{x}, \boldsymbol{v}', t) \, \mathrm{d} \boldsymbol{v}' \right) \, \mathrm{d} \boldsymbol{x} \, \mathrm{d} \boldsymbol{v} \, \mathrm{d} t.$$

Now that we have all the gain and loss terms together, we obtain

$$\frac{\partial}{\partial t}n(\boldsymbol{x},\boldsymbol{v},t)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{v}\,\mathrm{d}t = n(\boldsymbol{x},\boldsymbol{v},t+\,\mathrm{d}t)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{v} - n(\boldsymbol{x},\boldsymbol{v},t)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{v} = -(1+(2)-(3)+(4))$$

Thus, we have

$$\frac{\partial}{\partial t}n(\boldsymbol{x},\boldsymbol{v},t) \,\mathrm{d}\boldsymbol{x} \,\mathrm{d}\boldsymbol{v} \,\mathrm{d}t + \boldsymbol{v} \cdot \nabla n(\boldsymbol{x},\boldsymbol{v},t) \,\mathrm{d}\boldsymbol{x} \,\mathrm{d}\boldsymbol{v} \,\mathrm{d}t + \Sigma(\boldsymbol{x},\boldsymbol{v})n(\boldsymbol{x},\boldsymbol{v},t) \,\mathrm{d}\boldsymbol{x} \,\mathrm{d}\boldsymbol{v} \,\mathrm{d}t \\ = \left(\int_{V} f(\boldsymbol{x},\boldsymbol{v}',\boldsymbol{v})n(\boldsymbol{x},\boldsymbol{v}',t) \,\mathrm{d}\boldsymbol{v}'\right) \,\mathrm{d}\boldsymbol{x} \,\mathrm{d}\boldsymbol{v} \,\mathrm{d}t + \left(\int_{V} \sigma(\boldsymbol{x},\boldsymbol{v}',\boldsymbol{v})n(\boldsymbol{x},\boldsymbol{v}',t) \,\mathrm{d}\boldsymbol{v}'\right) \,\mathrm{d}\boldsymbol{x} \,\mathrm{d}\boldsymbol{v} \,\mathrm{d}t.$$

Since  $d\boldsymbol{x}$ ,  $d\boldsymbol{v}$ ,  $d\boldsymbol{t}$  were chosen arbitrarily small, we arrive at

$$\begin{aligned} \frac{\partial}{\partial t} n(\boldsymbol{x}, \boldsymbol{v}, t) + \boldsymbol{v} \cdot \nabla n(\boldsymbol{x}, \boldsymbol{v}, t) + \Sigma(\boldsymbol{x}, \boldsymbol{v}) n(\boldsymbol{x}, \boldsymbol{v}, t) \\ &= \int_{V} f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) n(\boldsymbol{x}, \boldsymbol{v}', t) \, \mathrm{d} \boldsymbol{v}' + \int_{V} \sigma(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) n(\boldsymbol{x}, \boldsymbol{v}', t) \, \mathrm{d} \boldsymbol{v}'. \end{aligned}$$

This is the neutron transport equation. Of course we also need boundary and initial conditions. If the core  $\Omega$  is convex and surrounded by vacuum it makes sense to assume that neutrons which have left it will never enter again. Thus, we introduce the inflow boundary

$$\Gamma_{-} \coloneqq \left\{ (\boldsymbol{x}, \boldsymbol{v}) \in \partial \Omega \times V : \, \boldsymbol{v} \cdot \boldsymbol{n}(\boldsymbol{x}) < 0 \right\},\$$

where  $\boldsymbol{n}(\boldsymbol{x})$  is the unit outward normal to  $\partial \Omega$  at  $\boldsymbol{x}$ . The neutron transport problem can now be formulated as:

Find a function  $n(\boldsymbol{x}, \boldsymbol{v}, t)$  such that

$$\begin{cases} \operatorname{lhs}(\boldsymbol{x}, \boldsymbol{v}, t) = \operatorname{rhs}(\boldsymbol{x}, \boldsymbol{v}, t) & \forall (\boldsymbol{x}, \boldsymbol{v}) \in \Omega \times V, \quad \forall t > 0, \\ n(\boldsymbol{x}, \boldsymbol{v}, t) = 0 & \forall (\boldsymbol{x}, \boldsymbol{v}) \in \Gamma_{-}, \quad \forall t > 0, \\ n(\boldsymbol{x}, \boldsymbol{v}, 0) = n_0(\boldsymbol{x}, \boldsymbol{v}) & \forall (\boldsymbol{x}, \boldsymbol{v}) \in \Omega \times V. \end{cases}$$
(2.1)

The left hand side lhs(x, v, t) and the right hand side rhs(x, v, t) are given by

$$lhs(\boldsymbol{x}, \boldsymbol{v}, t) = \frac{\partial}{\partial t} n(\boldsymbol{x}, \boldsymbol{v}, t) + \boldsymbol{v} \cdot \nabla n(\boldsymbol{x}, \boldsymbol{v}, t) + \Sigma(\boldsymbol{x}, \boldsymbol{v}) n(\boldsymbol{x}, \boldsymbol{v}, t),$$
  

$$rhs(\boldsymbol{x}, \boldsymbol{v}, t) = \int_{V} f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) n(\boldsymbol{x}, \boldsymbol{v}', t) \, \mathrm{d}\boldsymbol{v}' + \int_{V} \sigma(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) n(\boldsymbol{x}, \boldsymbol{v}', t) \, \mathrm{d}\boldsymbol{v}'$$

## 2.2 The Criticality Spectral Problem

Following the argumentation of [11, p. 283] and [13, p. 46], we ask ourselves the question: How much do we need to reduce fission artificially, so that there exists a (nonnegative and nontrivial) stationary solution of (2.1)? Suppose that we reduce fission by dividing  $\sigma(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v})$  by some k > 0. Then the fission term becomes

$$\frac{1}{k} \int_V \sigma(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) n(\boldsymbol{x}, \boldsymbol{v}', t) \, \mathrm{d} \boldsymbol{v}'.$$

If k is large, there cannot exist a stationary solution because too few fission reactions occur so that any initial neutron distribution will automatically die out. Suppose on the other hand that k is small. Then there can also be no stationary solution because so many fission reactions occur that any initial distribution of neutrons will immediately blow up. Hence, there must be some largest value of k such that there exists a (nonnegative and nontrivial) stationary solution of (2.1). We are therefore led to the following criticality spectral problem:

Find the largest k > 0 such that there exists n(x, v) nonnegative and nontrivial satisfying

$$\begin{cases} \operatorname{lhs}(\boldsymbol{x},\boldsymbol{v}) = \operatorname{rhs}(k,\boldsymbol{x},\boldsymbol{v}) & \forall (\boldsymbol{x},\boldsymbol{v}) \in \Omega \times V, \\ n(\boldsymbol{x},\boldsymbol{v}) = 0 & \forall (\boldsymbol{x},\boldsymbol{v}) \in \Gamma_{-}. \end{cases}$$
(2.2)

The left hand side lhs(x, v) and the right hand side rhs(k, x, v) are given by

$$lhs(\boldsymbol{x}, \boldsymbol{v}) = \boldsymbol{v} \cdot \nabla n(\boldsymbol{x}, \boldsymbol{v}) + \Sigma(\boldsymbol{x}, \boldsymbol{v})n(\boldsymbol{x}, \boldsymbol{v}),$$
  

$$rhs(k, \boldsymbol{x}, \boldsymbol{v}) = \int_{V} f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v})n(\boldsymbol{x}, \boldsymbol{v}') \, \mathrm{d}\boldsymbol{v}' + \frac{1}{k} \int_{V} \sigma(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v})n(\boldsymbol{x}, \boldsymbol{v}') \, \mathrm{d}\boldsymbol{v}'.$$

The value of k determines the state of the core, i.e.

$$\begin{cases} k > 1, & \text{supercritical,} \\ k = 1, & \text{critical,} \\ k < 1, & \text{subcritical.} \end{cases}$$

The whole purpose of this eigenvalue problem is that we can adjust the system composition and geometry until k becomes one, so that we obtain a critical chain reacting system.

## 2.3 Functional-Analytic Framework

In this section we give the functional-analytic framework for problem (2.2). We first introduce the appropriate function spaces and operators. Then we will formulate the eigenvalue problem in this setting. The material covered in this section is an extension of the material presented in [4]. Let  $\Omega \subset \mathbb{R}^n$  be open, bounded and convex with  $\partial \Omega$  continuously differentiable. Let  $V \subset \mathbb{R}^n$  be open and bounded.<sup>1</sup> For every  $\boldsymbol{x} \in \partial \Omega$  let  $\boldsymbol{n}(\boldsymbol{x})$  denote the unit outward normal to  $\partial \Omega$  at  $\boldsymbol{x}$ . We define

$$\begin{split} & \Gamma \coloneqq \partial \Omega \times V, \\ & \Gamma_+ \coloneqq \left\{ (\boldsymbol{x}, \boldsymbol{v}) \in \partial \Omega \times V : \, \boldsymbol{v} \cdot \boldsymbol{n}(\boldsymbol{x}) > 0 \right\}, \\ & \Gamma_0 \coloneqq \left\{ (\boldsymbol{x}, \boldsymbol{v}) \in \partial \Omega \times V : \, \boldsymbol{v} \cdot \boldsymbol{n}(\boldsymbol{x}) = 0 \right\}, \\ & \Gamma_- \coloneqq \left\{ (\boldsymbol{x}, \boldsymbol{v}) \in \partial \Omega \times V : \, \boldsymbol{v} \cdot \boldsymbol{n}(\boldsymbol{x}) < 0 \right\}. \end{split}$$

For points  $(\boldsymbol{x}, \boldsymbol{v}) \in \Gamma_+$  the velocity  $\boldsymbol{v}$  points out of  $\Omega$ . That is why  $\Gamma_+$  is called the outflow boundary. Similarly, if  $(\boldsymbol{x}, \boldsymbol{v}) \in \Gamma_-$ , then the velocity  $\boldsymbol{v}$  points into  $\Omega$ . Hence, we call  $\Gamma_-$  the inflow boundary. We introduce the Hilbert space

$$W^2(\Omega \times V) := \left\{ \phi \in L^2(\Omega \times V) \text{ s.t. } \boldsymbol{v} \cdot \nabla \phi \in L^2(\Omega \times V) \right\},\$$

where the inner product is given by

$$(\phi,\psi)_{W^2(\Omega\times V)} \coloneqq (\phi,\psi)_{L^2(\Omega\times V)} + (\boldsymbol{v}\cdot\nabla\phi,\boldsymbol{v}\cdot\nabla\psi)_{L^2(\Omega\times V)}.$$

Let  $\gamma$  be the boundary measure of  $\Omega$ . We denote

$$L^{2}(\Gamma_{+}) := \left\{ \phi : \Gamma_{+} \to \mathbb{R} : \phi \text{ is measurable and } \int_{\Gamma_{+}} |\phi|^{2} \, \mathrm{d}\gamma \, \mathrm{d}\boldsymbol{v} < \infty \right\},$$
$$L^{2}(\Gamma_{-}) := \left\{ \phi : \Gamma_{-} \to \mathbb{R} : \phi \text{ is measurable and } \int_{\Gamma_{-}} |\phi|^{2} \, \mathrm{d}\gamma \, \mathrm{d}\boldsymbol{v} < \infty \right\}.$$

For  $\phi \in W^2(\Omega \times V)$  we want to define what it means for  $\phi$  to vanish on  $\Gamma_-$  (resp.  $\Gamma_+$ ), i.e. we want to define how the statement  $\phi|_{\Gamma_-} = 0$  (resp.  $\phi|_{\Gamma_+} = 0$ ) shall be understood. But here, we come up against a difficulty: For  $\phi \in W^2(\Omega \times V)$  it is not true in general, that the trace  $u|_{\Gamma_-}$  (resp.  $u|_{\Gamma_+}$ ) exists as a function in  $L^2(\Gamma_-)$  (resp.  $L^2(\Gamma_+)$ ). This is the content of the following example.

**Example 2.3.1.** Consider the open unit disk  $\Omega = B_1(0) \subset \mathbb{R}^2$ . Consider a fixed velocity  $\boldsymbol{v} = (1,0)$  coming from the left. We denote  $\Omega_-$  for the inflow boundary and  $\Omega_+$  for the outflow boundary of  $\Omega$ , i.e.  $\Omega_-$  is the left half circle (without the poles) and  $\Omega_+$  is the right half circle (without the poles).

<sup>&</sup>lt;sup>1</sup> In [7], V is assumed to be the support of a positive Radon measure  $\mu$  on  $\mathbb{R}^n$  with  $\mu(\{0\}) = 0$ . Hence, in this framework, V is allowed to be discrete, or the unit sphere  $S^{n-1} \subset \mathbb{R}^n$ . For the moment, we can work with V open and the Lebesgue measure. But then some of the model problems we will encounter, e.g. the (1+0)-dimensional one, will not be covered by this framework, but they are by the framework in [7].



Consider now the function  $u(x,y) = (1-y)^{-\frac{1}{4}}$  defined on  $\Omega$ . Then u belongs to  $L^2(\Omega)$  with

$$\begin{aligned} \|u\|_{L^{2}(\Omega)}^{2} &= \int_{\Omega} |u(x,y)|^{2} \, \mathrm{d}x \, \mathrm{d}y = \int_{-1}^{+1} \int_{-\sqrt{1-x^{2}}}^{+\sqrt{1-x^{2}}} (1-y)^{-\frac{1}{2}} \, \mathrm{d}y \, \mathrm{d}x \\ &= 2 \int_{-1}^{+1} \left(1 + \sqrt{1-x^{2}}\right)^{\frac{1}{2}} - \left(1 - \sqrt{1-x^{2}}\right)^{\frac{1}{2}} \, \mathrm{d}x \\ &= \frac{8\sqrt{2}}{3}. \end{aligned}$$

But the trace of u on  $\Omega_{-}$  is  $(1 - \sin \vartheta)^{-\frac{1}{4}}$  (in polar coordinates) which is not in  $L^2(\Omega_{-})$  (w.r.t. the boundary measure). Indeed, we have

$$||u||_{L^{2}(\Omega_{-})}^{2} = \int_{\Omega_{-}} |u|^{2} ds = \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} (1 - \sin \vartheta)^{-\frac{1}{2}} d\vartheta = \infty$$

Note that u(x, y) is constant in x, i.e. constant in the direction of  $\mathbf{v}$ . Now this was for a fixed velocity  $\mathbf{v}$ . What if  $\mathbf{v}$  ranges in  $V = B_1(0) = \Omega$ ? Then we can do the same trick for each direction. Thus, let  $\mathbf{v} \in B_1(0)$ . For this fixed velocity, we denote  $\Omega^{\mathbf{v}}_{-}$  resp.  $\Omega^{\mathbf{v}}_{+}$ for the inflow resp. outflow part of the boundary of  $\Omega$ . For this  $\mathbf{v}$  we have the function  $u_{\mathbf{v}}$  defined by  $u_{\mathbf{v}}(\hat{x}, \hat{y}) = (1 - \hat{y})^{-\frac{1}{4}}$ , where  $\hat{x}$  and  $\hat{y}$  are the coordinates w.r.t. the rotated coordinate system. The following pictures illustrate the situation.



We denote  $\mathbf{x} = (x, y)$  and define  $u(\mathbf{x}, \mathbf{v}) = u_{\mathbf{v}}(\mathbf{x})$ . This defines a function in  $L^2(\Omega \times V)$  with

$$\int_{\Omega \times V} |u(\boldsymbol{x}, \boldsymbol{v})|^2 \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v} = \int_V \int_{\Omega} |u_{\boldsymbol{v}}(\boldsymbol{x})|^2 \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v} = \int_V \frac{8\sqrt{2}}{3} \, \mathrm{d}\boldsymbol{v} = \frac{8\pi\sqrt{2}}{3},$$

since the Lebesgue integral is invariant under rotation. Moreover, u belongs to  $W^2(\Omega \times V)$ , because  $\mathbf{v} \cdot \nabla u$  exists as a function in  $L^2(\Omega \times V)$ , and is actually zero, since for each fixed  $\mathbf{v}$  the function  $u_{\mathbf{v}}(\mathbf{x})$  is constant in the direction of  $\mathbf{v}$ . However, the trace of u on all of  $\Gamma_-$  (note that  $\Gamma_-$  is the actual inflow boundary as a subset of  $\partial\Omega \times V$  and should not be confused with  $\Omega^{\mathbf{v}}_-$  which is the inflow boundary for a fixed  $\mathbf{v}$  as a subset of  $\partial\Omega$ ) does not exist as a function in  $L^2(\Gamma_-)$ . This is because for each fixed  $\mathbf{v}$  the trace of  $u_{\mathbf{v}}(\mathbf{x})$  on  $\Omega^{\mathbf{v}}_$ is not in  $L^2(\Omega^{\mathbf{v}}_-)$ . Symbolically,

$$\int_{\Gamma_{-}} |u|^2 \,\mathrm{d}s \,\mathrm{d}\boldsymbol{v} = \int_{V} \int_{\Omega_{-}^{\boldsymbol{v}}} |u|^2 \,\mathrm{d}s \,\mathrm{d}\boldsymbol{v} = \int_{V} \infty \,\mathrm{d}\boldsymbol{v} = \pi \cdot \infty = \infty,$$

since  $\Gamma_{-} = \bigcup_{v \in V} \Omega^{v}_{-} \times \{v\}$ . The function u(x, v) can be given explicitly as

$$u(\boldsymbol{x}, \boldsymbol{v}) = \left(1 - \boldsymbol{x} \cdot \frac{\boldsymbol{v}^{\perp}}{|\boldsymbol{v}|}\right)^{-\frac{1}{4}},$$

where  $\mathbf{v}^{\perp} = (v_1, v_2)^{\perp} = (-v_2, v_1)$ . Last but not least we note that the exponent  $-\frac{1}{4}$  is optimal in the sense that

$$u(\boldsymbol{x},\boldsymbol{v}) = \left(1 - \boldsymbol{x} \cdot \frac{\boldsymbol{v}^{\perp}}{|\boldsymbol{v}|}\right)^{-\kappa}$$

does not serve as a similar counterexample if  $\kappa < \frac{1}{4}$ .

Although the trace of  $\phi \in W^2(\Omega \times V)$  does not exist as an  $L^2$ -function on all of  $\Gamma_-$  (resp.  $\Gamma_+$ ) we can still hope for this to be true on compact subsets of  $\Gamma_-$  (resp.  $\Gamma_+$ ). The following result shows that this is indeed true.

**Theorem 2.3.2.** [7, Ch. XXI, §2, Sec. 2.2, Thm. 1]. Let K be a compact subset of  $\Gamma_+$  (resp.  $\Gamma_-$ ); then there exists a bounded linear map

$$W^2(\Omega \times V) \to L^2(K)$$
  
 $\phi \mapsto \phi|_K.$ 

As a consequence, we can define

$$W^2_+(\Omega \times V) := \left\{ \phi \in W^2(\Omega \times V) : \phi|_K = 0 \text{ for all } K \subset \Gamma_+ \text{ compact} \right\},$$
$$W^2_-(\Omega \times V) := \left\{ \phi \in W^2(\Omega \times V) : \phi|_K = 0 \text{ for all } K \subset \Gamma_- \text{ compact} \right\},$$

which are Hilbert spaces in their own right. We shall next derive a Green's formula for the phase space  $\Omega \times V$  based on a Green's formula for  $\Omega$ .

Lemma 2.3.3. [7, Ch. XXI, §2, Sec. 2.2, Rmk. 3]. We define the space

$$\widetilde{W}^2(\Omega \times V) \coloneqq \left\{ \phi \in W^2(\Omega \times V) : \phi|_{\Gamma_{\pm}} \in L^2(\Gamma_{\pm}, |\boldsymbol{v} \cdot \boldsymbol{n}| \, \mathrm{d}\gamma \, \mathrm{d}\boldsymbol{v}) \right\}.$$

Then for all  $\phi_1, \phi_2 \in \widetilde{W}^2(\Omega \times V)$  we have the following Green's formula

$$\int_{\Omega \times V} \phi_1(\boldsymbol{v} \cdot \nabla \phi_2) + (\boldsymbol{v} \cdot \nabla \phi_1) \phi_2 \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v} = \int_{\Gamma_+} \phi_1 \phi_2 |\boldsymbol{v} \cdot \boldsymbol{n}| \, \mathrm{d}\gamma \, \mathrm{d}\boldsymbol{v} - \int_{\Gamma_-} \phi_1 \phi_2 |\boldsymbol{v} \cdot \boldsymbol{n}| \, \mathrm{d}\gamma \, \mathrm{d}\boldsymbol{v}.$$
(2.3)

*Proof.* We have

$$\begin{split} \int_{\Omega \times V} \phi_1(\boldsymbol{v} \cdot \nabla \phi_2) + (\boldsymbol{v} \cdot \nabla \phi_1) \phi_2 \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v} \\ &= \int_V \int_\Omega \phi_1(\boldsymbol{v} \cdot \nabla \phi_2) + (\boldsymbol{v} \cdot \nabla \phi_1) \phi_2 \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v} \\ &= \underbrace{\int_V \int_\Omega \phi_1(\boldsymbol{v} \cdot \nabla \phi_2) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v}}_{(*)} + \underbrace{\int_V \int_\Omega (\boldsymbol{v} \cdot \nabla \phi_1) \phi_2 \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v}}_{(**)} . \end{split}$$

Let us consider the first term

$$\begin{aligned} (*) &= \int_{V} \int_{\Omega} \phi_{1}(\boldsymbol{v} \cdot \nabla \phi_{2}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v} = \int_{V} \int_{\Omega} \phi_{1} \nabla \cdot (\boldsymbol{v}\phi_{2}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v} \\ &= \int_{V} \left( \int_{\partial \Omega} \phi_{1}\phi_{2}(\boldsymbol{v} \cdot \boldsymbol{n}) \, \mathrm{d}\gamma - \int_{\Omega} (\nabla \phi_{1}) \cdot (\boldsymbol{v}\phi_{2}) \, \mathrm{d}\boldsymbol{x} \right) \, \mathrm{d}\boldsymbol{v} \\ &= \int_{V} \int_{\partial \Omega} \phi_{1}\phi_{2}(\boldsymbol{v} \cdot \boldsymbol{n}) \, \mathrm{d}\gamma \, \mathrm{d}\boldsymbol{v} - \int_{V} \int_{\Omega} (\boldsymbol{v} \cdot \nabla \phi_{1})\phi_{2} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v}, \end{aligned}$$

where we have used a classical Green's formula on  $\Omega$ . So we finally obtain

$$\begin{split} \int_{\Omega \times V} \phi_1(\boldsymbol{v} \cdot \nabla \phi_2) + (\boldsymbol{v} \cdot \nabla \phi_1) \phi_2 \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v} &= (*) + (**) \\ &= \int_V \int_{\partial \Omega} \phi_1 \phi_2(\boldsymbol{v} \cdot \boldsymbol{n}) \, \mathrm{d}\gamma \, \mathrm{d}\boldsymbol{v} \\ &= \int_{\Gamma_+} \phi_1 \phi_2 |\boldsymbol{v} \cdot \boldsymbol{n}| \, \mathrm{d}\gamma \, \mathrm{d}\boldsymbol{v} - \int_{\Gamma_-} \phi_1 \phi_2 |\boldsymbol{v} \cdot \boldsymbol{n}| \, \mathrm{d}\gamma \, \mathrm{d}\boldsymbol{v}. \end{split}$$

Remark 2.3.4. [7, Ch. XXI, §2, Sec. 2.2, Rmk. 3]. We actually have

$$\widetilde{W}^{2}(\Omega \times V) = \left\{ \phi \in W^{2}(\Omega \times V) : \phi|_{\Gamma_{+}} \in L^{2}(\Gamma_{+}, |\boldsymbol{v} \cdot \boldsymbol{n}| \, \mathrm{d}\gamma \, \mathrm{d}\boldsymbol{v}) \right\} \\ = \left\{ \phi \in W^{2}(\Omega \times V) : \phi|_{\Gamma_{-}} \in L^{2}(\Gamma_{-}, |\boldsymbol{v} \cdot \boldsymbol{n}| \, \mathrm{d}\gamma \, \mathrm{d}\boldsymbol{v}) \right\},\$$

which in particular implies that  $W^2_+(\Omega \times V) \subset \widetilde{W}^2(\Omega \times V)$  and  $W^2_-(\Omega \times V) \subset \widetilde{W}^2(\Omega \times V)$ .

Corollary 2.3.5. We have

$$\int_{\Omega \times V} \left( \boldsymbol{v} \cdot \nabla \phi \right) \phi \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v} \le 0 \qquad \qquad \forall \phi \in W^2_+(\Omega \times V), \tag{2.4a}$$

$$\int_{\Omega \times V} (\boldsymbol{v} \cdot \nabla \phi) \phi \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v} \ge 0 \qquad \qquad \forall \phi \in W^2_-(\Omega \times V). \tag{2.4b}$$

*Proof.* We shall prove the statement for  $W^2_+(\Omega \times V)$  only. Let  $\phi \in W^2_+(\Omega \times V)$ . According to the previous remark, we have that  $\phi \in \widetilde{W}^2(\Omega \times V)$  and formula (2.3) is valid for  $\phi_1 = \phi_2 = \phi$ . Hence,

$$2\int_{\Omega \times V} (\boldsymbol{v} \cdot \nabla \phi) \phi \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v} = \underbrace{\int_{\Gamma_{+}} |\phi|^{2} \, |\boldsymbol{v} \cdot \boldsymbol{n}| \, \mathrm{d}\gamma \, \mathrm{d}\boldsymbol{v}}_{=0} - \underbrace{\int_{\Gamma_{-}} |\phi|^{2} \, |\boldsymbol{v} \cdot \boldsymbol{n}| \, \mathrm{d}\gamma \, \mathrm{d}\boldsymbol{v}}_{\geq 0} \leq 0.$$

**Remark 2.3.6.** The result of Corollary 2.3.5 is also called the monotonicity of the advection operator. It is used in [4] during the proof of the main existence result of the positive eigenpair. However, to make the presentation clearer, we have stated this result here.

The criticality spectral equation (2.2) is an integro-differential equation. Hence, there are integral operators involved (for scattering and fission) for which we have to make sure that the induced operators are bounded. The following lemma gives us conditions for this to be true.

**Lemma 2.3.7.** [7, Ch. XXI, §2, Sec. 3.1, Lemma 1]. Let s(x, v', v) be a given function on  $\Omega \times V \times V$ . We assume that there exists a constant M > 0 such that

$$\begin{split} &\int_{V} \left| s(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \right| \ \mathrm{d} \boldsymbol{v} \leq M \quad \forall (\boldsymbol{x}, \boldsymbol{v}') \in \Omega \times V, \\ &\int_{V} \left| s(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \right| \ \mathrm{d} \boldsymbol{v}' \leq M \quad \forall (\boldsymbol{x}, \boldsymbol{v}) \in \Omega \times V. \end{split}$$

Then the operator S defined by

$$(S\phi)(\boldsymbol{x}, \boldsymbol{v}) \coloneqq \int_V s(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \phi(\boldsymbol{x}, \boldsymbol{v}') \, \mathrm{d}\boldsymbol{v}',$$

is linear and bounded from  $L^2(\Omega \times V)$  to itself.

*Proof.* We have

$$||S\phi||_{L^2(\Omega\times V)}^2 = \int_{\Omega\times V} \left| \int_V s(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \phi(\boldsymbol{x}, \boldsymbol{v}') \, \mathrm{d}\boldsymbol{v}' \right|^2 \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v}.$$

Taking the absolute value inside the integral, we obtain

$$\int_{\Omega \times V} \left| \int_{V} s(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \phi(\boldsymbol{x}, \boldsymbol{v}') \, \mathrm{d}\boldsymbol{v}' \right|^2 \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v} \le \int_{\Omega \times V} \left( \int_{V} \left| s(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \right| \left| \phi(\boldsymbol{x}, \boldsymbol{v}') \right| \, \mathrm{d}\boldsymbol{v}' \right)^2 \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v}.$$

Using Hölder's inequality for the innermost integral and our assumptions on s, we obtain

$$\begin{split} \int_{\Omega \times V} \left( \int_{V} \left| s(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \right| \left| \phi(\boldsymbol{x}, \boldsymbol{v}') \right| \, \mathrm{d}\boldsymbol{v}' \right)^{2} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v} \\ &= \int_{\Omega \times V} \left( \int_{V} \left| s(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \right|^{\frac{1}{2}} \left| s(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \right|^{\frac{1}{2}} \left| \phi(\boldsymbol{x}, \boldsymbol{v}') \right| \, \mathrm{d}\boldsymbol{v}' \right)^{2} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v} \\ &\leq \int_{\Omega \times V} \left( \int_{V} \left| s(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \right| \, \mathrm{d}\boldsymbol{v}' \right) \left( \int_{V} \left| s(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \right| \left| \phi(\boldsymbol{x}, \boldsymbol{v}') \right|^{2} \, \mathrm{d}\boldsymbol{v}' \right) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v} \\ &\leq M \int_{\Omega \times V} \left( \int_{V} \left| s(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \right| \left| \phi(\boldsymbol{x}, \boldsymbol{v}') \right|^{2} \, \mathrm{d}\boldsymbol{v}' \right) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v} \\ &= M \int_{\Omega \times V} \left| \phi(\boldsymbol{x}, \boldsymbol{v}') \right|^{2} \left( \int_{V} \left| s(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \right| \, \mathrm{d}\boldsymbol{v} \right) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v}' \\ &\leq M^{2} \int_{\Omega \times V} \left| \phi(\boldsymbol{x}, \boldsymbol{v}') \right|^{2} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v}'. \end{split}$$

Hence, we arrive at

$$||S\phi||_{L^2(\Omega \times V)} \le M ||\phi||_{L^2(\Omega \times V)}.$$

**Assumption 2.3.8.** The data functions  $f(\mathbf{x}, \mathbf{v}', \mathbf{v})$  (scattering) and  $\sigma(\mathbf{x}, \mathbf{v}', \mathbf{v})$  (fission) are nonnegative and satisfy the conditions of Lemma 2.3.7 in place of  $s(\mathbf{x}, \mathbf{v}', \mathbf{v})$ .

Due to Assumption 2.3.8 we can define the following bounded linear operators

$$\begin{split} K &: L^2(\Omega \times V) \to L^2(\Omega \times V) \\ \phi(\boldsymbol{x}, \boldsymbol{v}) &\mapsto (K\phi)(\boldsymbol{x}, \boldsymbol{v}) \coloneqq \int_V f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \phi(\boldsymbol{x}, \boldsymbol{v}') \, \mathrm{d} \boldsymbol{v}', \end{split}$$

$$F: L^{2}(\Omega \times V) \to L^{2}(\Omega \times V)$$
$$\phi(\boldsymbol{x}, \boldsymbol{v}) \mapsto (F\phi)(\boldsymbol{x}, \boldsymbol{v}) \coloneqq \int_{V} \sigma(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \phi(\boldsymbol{x}, \boldsymbol{v}') \, \mathrm{d}\boldsymbol{v}'.$$

**Assumption 2.3.9.** The data function  $\Sigma(\mathbf{x}, \mathbf{v})$  (absorption & outscattering) is nonnegative and bounded.

Under Assumption 2.3.9 we can further define the following bounded linear operator

$$\begin{split} L: \ W^2_-(\Omega \times V) &\to L^2(\Omega \times V) \\ \phi(\boldsymbol{x}, \boldsymbol{v}) &\mapsto (L\phi)(\boldsymbol{x}, \boldsymbol{v}) \coloneqq \boldsymbol{v} \cdot \nabla \, \phi(\boldsymbol{x}, \boldsymbol{v}) + \Sigma(\boldsymbol{x}, \boldsymbol{v}) \phi(\boldsymbol{x}, \boldsymbol{v}). \end{split}$$

With this functional-analytic preparation our criticality spectral problem (2.2) can now be formulated as:

$$\begin{cases} \text{Find the largest } k > 0 \text{ such that there exists} \\ \phi \in W^2_-(\Omega \times V) \text{ nonnegative and nontrivial such that} \\ L\phi = K\phi + \frac{1}{k}F\phi. \end{cases}$$
(2.5)

and

## **3** Basic Existence Result

#### 3.1 Formulation as an ordinary EVP

Our goal is to show the existence of a solution  $(k, \phi)$  to problem (2.5). For this purpose, we follow [4], which uses material from [7], to reformulate this generalized eigenvalue problem in several stages to finally arrive at an ordinary eigenvalue problem. For each reformulation we need to invert a certain operator. The first one to invert is L.

**Assumption 3.1.1.**  $\Sigma(\boldsymbol{x}, \boldsymbol{v})$  is uniformly positive, i.e. there exists  $\Sigma_0 > 0$ , such that

$$\Sigma(\boldsymbol{x}, \boldsymbol{v}) \geq \Sigma_0 \qquad \forall (\boldsymbol{x}, \boldsymbol{v}) \in \Omega \times V.$$

**Theorem 3.1.2.** [7, Ch. XXI, §2, Sec. 4.1, Thm. 4]. Under Assumption 3.1.1 we have that for each  $g \in L^2(\Omega \times V)$  there exists a unique  $\phi \in W^2_-(\Omega \times V)$ , such that  $L\phi = g$ . This allows us to consider the inverse map

$$L^{-1}: L^2(\Omega \times V) \to W^2_-(\Omega \times V) \subset L^2(\Omega \times V).$$

Seen as a map from  $L^2(\Omega \times V)$  to itself,  $L^{-1}$  is a bounded linear operator.

*Proof.* For any  $t \ge 0$  we consider

$$\begin{split} G(t): L^2(\Omega \times V) &\to L^2(\Omega \times V) \\ \phi(\boldsymbol{x}, \boldsymbol{v}) &\mapsto \begin{cases} \phi(\boldsymbol{x} - t\boldsymbol{v}, \boldsymbol{v}) \mathrm{e}^{-\int_0^t \Sigma(\boldsymbol{x} + (s-t)\boldsymbol{v}, \boldsymbol{v}) \, \mathrm{d}s} & \text{if } \boldsymbol{x} - t\boldsymbol{v} \in \Omega, \\ 0 & \text{otherwise.} \end{cases} \end{split}$$

The family of operators  $\{G(t)\}_{t\geq 0}$  defines a semigroup of class  $\mathcal{C}^0$  whose infinitesimal generator is the unbounded operator -L with domain  $W^2_{-}(\Omega \times V)$  (see [7, Ch. XXI, §2, Sec. 3.2, Rmk. 10]). Using Assumption 3.1.1, we obtain

$$\begin{split} ||G(t)\phi||^{2}_{L^{2}(\Omega\times V)} &= \int_{\Omega\times V} \left| \phi(\boldsymbol{x} - t\boldsymbol{v}, \boldsymbol{v}) \mathrm{e}^{-\int_{0}^{t} \Sigma(\boldsymbol{x} + (s-t)\boldsymbol{v}, \boldsymbol{v}) \, \mathrm{d}s} \chi_{\Omega}(\boldsymbol{x} - t\boldsymbol{v}) \right|^{2} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v} \\ &= \int_{\Omega\times V} |\phi(\boldsymbol{x} - t\boldsymbol{v}, \boldsymbol{v})|^{2} \, \mathrm{e}^{-2\int_{0}^{t} \Sigma(\boldsymbol{x} + (s-t)\boldsymbol{v}, \boldsymbol{v}) \, \mathrm{d}s} \chi_{\Omega}(\boldsymbol{x} - t\boldsymbol{v}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v} \\ &\leq \mathrm{e}^{-2t\Sigma_{0}} \int_{\Omega\times V} |\phi(\boldsymbol{x} - t\boldsymbol{v}, \boldsymbol{v})|^{2} \, \chi_{\Omega}(\boldsymbol{x} - t\boldsymbol{v}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v} \\ &\leq \mathrm{e}^{-2t\Sigma_{0}} \int_{\Omega\times V} |\phi(\boldsymbol{x}, \boldsymbol{v})|^{2} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v} \\ &= \mathrm{e}^{-2t\Sigma_{0}} ||\phi||^{2}_{L^{2}(\Omega\times V)}. \end{split}$$

Hence, we have shown  $||G(t)||_{\mathcal{L}(L^2(\Omega \times V))} \leq e^{-t\Sigma_0}$  with  $\Sigma_0 > 0$  and the result is a consequence of [6, Ch. XVII, §3, Sec. 1, Prop. 1].

**Remark 3.1.3.** The previous theorem actually shows that the solution  $\phi$  of  $L\phi = g$  is given by the following formula (for a.e.  $(\boldsymbol{x}, \boldsymbol{v}) \in \Omega \times V$ )

$$\phi(\boldsymbol{x},\boldsymbol{v}) = \int_0^{d(\boldsymbol{x},\boldsymbol{v})} e^{-\int_0^t \Sigma(\boldsymbol{x}-s\boldsymbol{v},\boldsymbol{v}) \, \mathrm{d}s} g(\boldsymbol{x}-t\boldsymbol{v},\boldsymbol{v}) \, \mathrm{d}t, \qquad (3.1)$$

where  $d(\boldsymbol{x}, \boldsymbol{v})$  is the travel time defined by

$$d(\boldsymbol{x}, \boldsymbol{v}) \coloneqq \sup \left\{ \tau : \boldsymbol{x} - s \boldsymbol{v} \in \Omega \text{ for all } 0 \leq s < \tau \right\}.$$

**Example 3.1.4.** Assumption 3.1.1 is important. Indeed, let for example  $\Omega = V = (0, 1)$ . Let  $g(x, v) = v^{-\frac{1}{3}} \in L^2(\Omega \times V)$  and  $\Sigma = 0$ . Then  $\phi$  given by (3.1) does not define a function in  $W^2_{-}(\Omega \times V)$ . Indeed,

$$\phi(x,v) = \int_0^{d(x,v)} v^{-\frac{1}{3}} \, \mathrm{d}t = \int_0^{\frac{x}{v}} v^{-\frac{1}{3}} \, \mathrm{d}t = xv^{-\frac{4}{3}} \notin L^2(\Omega \times V).$$

However, if  $\Sigma = 1$ , we obtain

$$\phi(x,v) = \int_0^{d(x,v)} e^{-t} v^{-\frac{1}{3}} dt = \int_0^{\frac{x}{v}} e^{-t} v^{-\frac{1}{3}} dt = v^{-\frac{1}{3}} \left(1 - e^{-\frac{x}{v}}\right) \in W^2_-(\Omega \times V),$$
  
and  $v \cdot \frac{\partial \phi}{\partial x} + \Sigma \phi = g.$ 

Theorem 3.1.2 allows us to reformulate problem (2.5) as

 $\begin{cases} \text{Find the largest } k > 0 \text{ such that there exists} \\ \phi \in L^2(\Omega \times V) \text{ nonnegative and nontrivial such that} \\ \phi = L^{-1}K\phi + \frac{1}{k}L^{-1}F\phi, \end{cases}$ 

or equivalently as

 $\begin{cases} \text{Find the largest } k > 0 \text{ such that there exists} \\ \phi \in L^2(\Omega \times V) \text{ nonnegative and nontrivial such that} \\ (I - L^{-1}K) \phi = \frac{1}{k} L^{-1} F \phi. \end{cases}$ (3.2)

The next step is to invert the operator  $(I - L^{-1}K)$ .

**Assumption 3.1.5.** There exists  $0 < \eta < 1$ , such that

$$\int_{V} f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \, \mathrm{d}\boldsymbol{v} \leq \eta \, \Sigma(\boldsymbol{x}, \boldsymbol{v}') \qquad \forall (\boldsymbol{x}, \boldsymbol{v}') \in \Omega \times V, \qquad (3.3a)$$
$$\int_{V} f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \, \mathrm{d}\boldsymbol{v}' \leq \eta \, \Sigma(\boldsymbol{x}, \boldsymbol{v}) \qquad \forall (\boldsymbol{x}, \boldsymbol{v}) \in \Omega \times V. \qquad (3.3b)$$

**Proposition 3.1.6.** Under Assumption 3.1.5 we have that the operator  $(I - L^{-1}K)$  is boundedly invertible from  $L^2(\Omega \times V)$  to itself.

*Proof.* We define the space

$$L^{2}_{\Sigma}(\Omega \times V) \coloneqq \left\{ \phi \, : \, \Omega \times V \to \mathbb{R} \text{ measurable } : \, ||\phi||_{L^{2}_{\Sigma}(\Omega \times V)} < \infty \right\},$$

where

$$||\phi||_{L^2_{\Sigma}(\Omega \times V)} \coloneqq ||\Sigma^{\frac{1}{2}}\phi||_{L^2(\Omega \times V)}.$$

Since  $\Sigma$  is bounded and uniformly positive, we have that the spaces  $L^2(\Omega \times V)$  and  $L^2_{\Sigma}(\Omega \times V)$  coincide and that the norms are equivalent. If we can show that

$$||L^{-1}K||_{\mathcal{L}\left(L^2_{\Sigma}(\Omega \times V)\right)} < 1, \tag{3.4}$$

we obtain that  $(I - L^{-1}K)$  is boundedly invertible in  $\mathcal{L}(L^2_{\Sigma}(\Omega \times V))$  and hence also in  $\mathcal{L}(L^2(\Omega \times V))$ . It therefore remains to show (3.4). To this end, let  $g \in L^2(\Omega \times V)$  and  $\phi \in W^2_{-}(\Omega \times V)$  with  $L\phi = Kg$ . Using (2.4b), we obtain

$$\begin{split} ||\phi||_{L_{\Sigma}^{2}}^{2} &= (\Sigma\phi,\phi)_{L^{2}} = (L\phi - \boldsymbol{v}\cdot\nabla\phi,\phi)_{L^{2}} = (L\phi,\phi)_{L^{2}} - (\boldsymbol{v}\cdot\nabla\phi,\phi)_{L^{2}} \\ &\leq (L\phi,\phi)_{L^{2}} = (Kg,\phi)_{L^{2}} = \int_{\Omega} \int_{V} (Kg)(\boldsymbol{x},\boldsymbol{v})\phi(\boldsymbol{x},\boldsymbol{v}) \, \mathrm{d}\boldsymbol{v} \, \mathrm{d}\boldsymbol{x} \\ &= \int_{\Omega} \int_{V} \phi(\boldsymbol{x},\boldsymbol{v}) \int_{V} f(\boldsymbol{x},\boldsymbol{v}',\boldsymbol{v})g(\boldsymbol{x},\boldsymbol{v}') \, \mathrm{d}\boldsymbol{v}' \, \mathrm{d}\boldsymbol{v} \, \mathrm{d}\boldsymbol{x} \\ &\leq \int_{\Omega} \int_{V} |\phi(\boldsymbol{x},\boldsymbol{v})| \int_{V} f(\boldsymbol{x},\boldsymbol{v}',\boldsymbol{v})|g(\boldsymbol{x},\boldsymbol{v}')| \, \mathrm{d}\boldsymbol{v}' \, \mathrm{d}\boldsymbol{v} \, \mathrm{d}\boldsymbol{x} \\ &\leq \int_{\Omega} \int_{V} |\phi(\boldsymbol{x},\boldsymbol{v})| \int_{V} f(\boldsymbol{x},\boldsymbol{v}',\boldsymbol{v})^{\frac{1}{2}} f(\boldsymbol{x},\boldsymbol{v}',\boldsymbol{v})^{\frac{1}{2}} |g(\boldsymbol{x},\boldsymbol{v}')| \, \mathrm{d}\boldsymbol{v}' \, \mathrm{d}\boldsymbol{v} \, \mathrm{d}\boldsymbol{x}. \end{split}$$

Hence, we have

$$||\phi||_{L_{\Sigma}^{2}}^{2} \leq \int_{\Omega} \int_{V} |\phi(\boldsymbol{x}, \boldsymbol{v})| \int_{V} f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v})^{\frac{1}{2}} f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v})^{\frac{1}{2}} |g(\boldsymbol{x}, \boldsymbol{v}')| \, \mathrm{d}\boldsymbol{v}' \, \mathrm{d}\boldsymbol{v} \, \mathrm{d}\boldsymbol{x}.$$
(3.5)

In the innermost integral we use Hölder's inequality to obtain

$$\int_{V} f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v})^{\frac{1}{2}} f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v})^{\frac{1}{2}} |g(\boldsymbol{x}, \boldsymbol{v}')| \, \mathrm{d}\boldsymbol{v}' \\
\leq \left( \int_{V} f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \, \mathrm{d}\boldsymbol{v}' \right)^{\frac{1}{2}} \left( \int_{V} f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) |g(\boldsymbol{x}, \boldsymbol{v}')|^{2} \, \mathrm{d}\boldsymbol{v}' \right)^{\frac{1}{2}}.$$
(3.6)

Inserting (3.6) in (3.5), we obtain

$$||\phi||_{L^2_{\Sigma}}^2 \leq \int_{\Omega} \int_{V} |\phi(\boldsymbol{x}, \boldsymbol{v})| \left( \int_{V} f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \, \mathrm{d}\boldsymbol{v}' \right)^{\frac{1}{2}} \left( \int_{V} f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) |g(\boldsymbol{x}, \boldsymbol{v}')|^2 \, \mathrm{d}\boldsymbol{v}' \right)^{\frac{1}{2}} \, \mathrm{d}\boldsymbol{v} \, \mathrm{d}\boldsymbol{x}.$$

Using Hölder's inequality once again, but this time over  $\Omega \times V$ , we obtain

$$\begin{split} ||\phi||_{L_{\Sigma}^{2}}^{2} &\leq \left(\int_{\Omega}\int_{V}|\phi(\boldsymbol{x},\boldsymbol{v})|^{2}\int_{V}f(\boldsymbol{x},\boldsymbol{v}',\boldsymbol{v}) \,\mathrm{d}\boldsymbol{v}' \,\mathrm{d}\boldsymbol{v} \,\mathrm{d}\boldsymbol{x}\right)^{\frac{1}{2}} \\ &\cdot \left(\int_{\Omega}\int_{V}\int_{V}f(\boldsymbol{x},\boldsymbol{v}',\boldsymbol{v})|g(\boldsymbol{x},\boldsymbol{v}')|^{2} \,\mathrm{d}\boldsymbol{v}' \,\mathrm{d}\boldsymbol{v} \,\mathrm{d}\boldsymbol{x}\right)^{\frac{1}{2}} \end{split}$$

Now we can use (3.3b) of Assumption 3.1.5 to bound the first term

$$\begin{split} \left( \int_{\Omega} \int_{V} |\phi(\boldsymbol{x}, \boldsymbol{v})|^{2} \int_{V} f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \, \mathrm{d}\boldsymbol{v}' \, \mathrm{d}\boldsymbol{v} \, \mathrm{d}\boldsymbol{x} \right)^{\frac{1}{2}} &\leq \eta^{\frac{1}{2}} \left( \int_{\Omega} \int_{V} \Sigma(\boldsymbol{x}, \boldsymbol{v}) |\phi(\boldsymbol{x}, \boldsymbol{v})|^{2} \, \mathrm{d}\boldsymbol{v} \, \mathrm{d}\boldsymbol{x} \right)^{\frac{1}{2}} \\ &= \eta^{\frac{1}{2}} ||\phi||_{L^{2}_{\Sigma}}. \end{split}$$

And we can use (3.3a) of Assumption 3.1.5 to bound the second term after having used Fubini

$$\begin{split} \left(\int_{\Omega} \int_{V} \int_{V} f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) |g(\boldsymbol{x}, \boldsymbol{v}')|^{2} \, \mathrm{d}\boldsymbol{v}' \, \mathrm{d}\boldsymbol{v} \, \mathrm{d}\boldsymbol{x}\right)^{\frac{1}{2}} \\ &= \left(\int_{\Omega} \int_{V} \int_{V} f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) |g(\boldsymbol{x}, \boldsymbol{v}')|^{2} \, \mathrm{d}\boldsymbol{v} \, \mathrm{d}\boldsymbol{v}' \, \mathrm{d}\boldsymbol{x}\right)^{\frac{1}{2}} \\ &= \left(\int_{\Omega} \int_{V} |g(\boldsymbol{x}, \boldsymbol{v}')|^{2} \int_{V} f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \, \mathrm{d}\boldsymbol{v} \, \mathrm{d}\boldsymbol{v}' \, \mathrm{d}\boldsymbol{x}\right)^{\frac{1}{2}} \\ &\leq \eta^{\frac{1}{2}} \left(\int_{\Omega} \int_{V} \Sigma(\boldsymbol{x}, \boldsymbol{v}') |g(\boldsymbol{x}, \boldsymbol{v}')|^{2} \, \mathrm{d}\boldsymbol{v}' \, \mathrm{d}\boldsymbol{x}\right)^{\frac{1}{2}} \\ &= \eta^{\frac{1}{2}} ||g||_{L^{2}_{\Sigma}}. \end{split}$$

We arrive at

$$||\phi||_{L^2_{\Sigma}}^2 \le \eta ||\phi||_{L^2_{\Sigma}} ||g||_{L^2_{\Sigma}}.$$

Hence, we have  $||\phi||_{L^2_{\Sigma}} \leq \eta ||g||_{L^2_{\Sigma}}$  meaning that  $||L^{-1}K||_{\mathcal{L}(L^2_{\Sigma}(\Omega \times V))} \leq \eta < 1$ , which was to be shown.

Proposition 3.1.6 allows us to rewrite problem (3.2) as

$$\begin{cases} \text{Find the largest } k > 0 \text{ such that there exists} \\ \phi \in L^2(\Omega \times V) \text{ nonnegative and nontrivial such that} \\ A\phi = k\phi, \end{cases}$$
(3.7)

where  $A := (I - L^{-1}K)^{-1}L^{-1}F = (L - K)^{-1}F$ . So we have transformed our generalized eigenvalue problem into an ordinary one for the operator A.

## 3.2 Existence of the Eigenpair

We know that  $A = (L - K)^{-1}F$  is a bounded linear operator from  $L^2(\Omega \times V)$  to itself. However, A might not be compact. That is why we shall introduce the notion of socalled regular integral operators. To this end, let  $s(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v})$  be a function satisfying the conditions of Lemma 2.3.7. We can see the corresponding integral operator S as a map

$$\Omega \to \mathcal{L}(L^2(V))$$
$$\boldsymbol{x} \mapsto S(\boldsymbol{x}),$$

defined by

$$(S(\boldsymbol{x})\phi)(\boldsymbol{v})\coloneqq\int_V s(\boldsymbol{x},\boldsymbol{v}',\boldsymbol{v})\phi(\boldsymbol{v}')\,\mathrm{d}\boldsymbol{v}'.$$

**Definition 3.2.1.** [16, Def. 4.1]. Let  $\mathcal{K} \subset \mathcal{L}(L^2(V))$  be the subspace of compact operators. Then the integral operator S is called regular if  $S(\mathbf{x}) \in \mathcal{K}$  a.e., the mapping  $\Omega \ni \mathbf{x} \mapsto S(\mathbf{x}) \in \mathcal{K}$  is measurable and

$$\{S(\boldsymbol{x}) : \boldsymbol{x} \in \Omega\}$$

is relatively compact in  $\mathcal{L}(L^2(V))$ .

The following compactness result about regular integral operators is crucial for our eigenvalue problem.

**Theorem 3.2.2.** [16, Thm. 4.1]. Assume that the integral operator S is regular. Then  $L^{-1}S$  and  $SL^{-1}$  are compact operators from  $L^2(\Omega \times V)$  to itself.

**Assumption 3.2.3.** The integral operator F induced by  $\sigma(\mathbf{x}, \mathbf{v}', \mathbf{v})$  is regular.

**Corollary 3.2.4.** Under Assumption 3.2.3 we have that the operator A is compact. In particular, problem (3.7) (and hence problem (2.5)) can have at most a countable number of eigenvalues and of associated eigenvectors in  $W^2_{-}(\Omega \times V)$ .

*Proof.* Because of Assumption 3.2.3, we can apply Theorem 3.2.2, which gives us the compactness of  $L^{-1}F$ . Since the composition of a bounded operator with a compact operator is again a compact operator, it follows that the operator  $A = (I - L^{-1}K)^{-1}L^{-1}F$  is compact from  $L^2(\Omega \times V)$  to itself.

However, the eigenpair we are interested in, has a physical meaning. The eigenvalue must be positive (since it is a factor which reduces fission) and the eigenfunction as well (since it makes no sense to have a negative neutron distribution). That is why we have to show the existence of such a special eigenpair. For this we need to make precise the notions of nonnegativity and positivity.

**Definition 3.2.5.** Let  $g \in L^2(\Omega \times V)$ . Then g is called

- nonnegative, if  $g(\boldsymbol{x}, \boldsymbol{v}) \geq 0$ , for a.e.  $(\boldsymbol{x}, \boldsymbol{v}) \in \Omega \times V$ ,
- positive, if  $g(\boldsymbol{x}, \boldsymbol{v}) > 0$ , for a.e.  $(\boldsymbol{x}, \boldsymbol{v}) \in \Omega \times V$ .

Let  $T : L^2(\Omega \times V) \to L^2(\Omega \times V)$  be an operator. Then T is called

- nonnegative, if for all g nonnegative, we have that Tg is nonnegative,
- positive, if for all g positive, we have that Tg is positive.

Let  $s(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v})$  be a function on  $\Omega \times V \times V$ . Then s is called

- nonnegative, if  $s(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \ge 0$ , for a.e.  $(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \in \Omega \times V \times V$ ,
- positive, if  $s(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) > 0$ , for a.e.  $(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \in \Omega \times V \times V$ .

**Notation 3.2.6.** For nonnegativity we use the symbol  $\geq$  and for positivity > (e.g. if  $g \in L^2(\Omega \times V)$  is nonnegative/positive, we write  $g \geq 0/g > 0$ ). We also use these symbols to compare two functions, i.e. for  $g, h \in L^2(\Omega \times V)$ , we write  $g \geq h$  to say that  $g(\boldsymbol{x}, \boldsymbol{v}) \geq h(\boldsymbol{x}, \boldsymbol{v})$  for a.e.  $(\boldsymbol{x}, \boldsymbol{v}) \in \Omega \times V$ . Similarly for g > h.

**Definition 3.2.7.** Let  $g \in L^2(\Omega \times V)$ . Then we say that g is not of constant sign if both sets

$$\{(\boldsymbol{x}, \boldsymbol{v}) \in \Omega \times V : g(\boldsymbol{x}, \boldsymbol{v}) > 0\}, \\ \{(\boldsymbol{x}, \boldsymbol{v}) \in \Omega \times V : g(\boldsymbol{x}, \boldsymbol{v}) < 0\},$$

have nonzero measure. If one of these sets has zero measure, then we say that g is of constant sign.

**Assumption 3.2.8.** The spectral radius  $\rho$  of A is nonzero.

**Remark 3.2.9.** In [4] it is not proven that the spectral radius of A is nonzero. However, this detail becomes important when we want to apply the Krein-Rutman Theorem (see [7, Ch. XXI, §3, Sec. 4, p. 286] for the  $L^2$ -setting and [10, p. 2, Thm. 1.1] for the abstract setting. One might show Assumption 3.2.8 with a result of de Pagter about irreducible compact operators (see [8]).

**Theorem 3.2.10.** Under Assumption 3.2.8 we have that the spectral radius  $\rho$  of A is an eigenvalue. Moreover, there exists a corresponding eigenfunction  $u_{\rho}$  which is nonnegative.

*Proof.* Because of Corollary 3.2.4 we have that the operator

$$A: L^2(\Omega \times V) \to L^2(\Omega \times V)$$

given by  $A = (I - L^{-1}K)^{-1}L^{-1}F$  is compact. We can write A as

$$A = \sum_{k=0}^{\infty} \left( L^{-1} K \right)^k L^{-1} F.$$

Since  $f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v})$  and  $\sigma(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v})$  are nonnegative, it follows that the induced operators K and F are nonnegative. The operator  $L^{-1}$  is also nonnegative, even positive (see formula (3.1)). So all the operators in this series are nonnegative from which it follows that A is nonnegative. Taking into account Assumption 3.2.8, we can therefore apply the Krein-Rutman Theorem [7, Ch. XXI, §3, Sec. 4, p. 286], which says that the spectral radius  $\rho$  of A is an eigenvalue, and that there exists a corresponding nonnegative eigenfunction  $u_{\rho}$ .

Assumption 3.2.11. The data function  $\sigma(\mathbf{x}, \mathbf{v}', \mathbf{v})$  is positive (i.e.  $\sigma > 0$ ).

**Assumption 3.2.12.** The velocity space V contains a torus, i.e. there exist constants  $0 < a_0 < a_1$  such that

$$\{\boldsymbol{v}\in\mathbb{R}^n:a_0\leq|\boldsymbol{v}|\leq a_1\}\subset V.$$

**Lemma 3.2.13.** [7, Ch. XXI, §3, Sec. 4, Thm. 7]. Under Assumption 3.2.11 and Assumption 3.2.12 we have that the operator  $FL^{-1}F$  has the following property: For all g nonnegative and nontrivial, we have that  $FL^{-1}Fg$  is positive, i.e.  $\forall g \geq 0$  with  $g \neq 0$ , we have that  $FL^{-1}Fg > 0$ .

*Proof.* Let  $g \in L^2(\Omega \times V)$  with  $g \ge 0$  and  $g \ne 0$ . We have

$$(FL^{-1}Fg)(\boldsymbol{x}, \boldsymbol{v})$$

$$= \int_{V} \sigma(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \left( L^{-1}Fg \right)(\boldsymbol{x}, \boldsymbol{v}') \, \mathrm{d}\boldsymbol{v}'$$

$$= \int_{V} \sigma(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \int_{0}^{d(\boldsymbol{x}, \boldsymbol{v}')} \mathrm{e}^{-\int_{0}^{t} \Sigma(\boldsymbol{x} - s\boldsymbol{v}', \boldsymbol{v}') \, \mathrm{d}s} (Fg)(\boldsymbol{x} - t\boldsymbol{v}', \boldsymbol{v}') \, \mathrm{d}t \, \mathrm{d}\boldsymbol{v}'$$

$$= \int_{V} \sigma(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \int_{0}^{d(\boldsymbol{x}, \boldsymbol{v}')} \mathrm{e}^{-\int_{0}^{t} \Sigma(\boldsymbol{x} - s\boldsymbol{v}', \boldsymbol{v}') \, \mathrm{d}s}$$

$$\cdot \int_{V} \sigma(\boldsymbol{x} - t\boldsymbol{v}', \boldsymbol{v}'', \boldsymbol{v}')g(\boldsymbol{x} - t\boldsymbol{v}', \boldsymbol{v}'') \, \mathrm{d}\boldsymbol{v}'' \, \mathrm{d}t \, \mathrm{d}\boldsymbol{v}'$$

$$= \int_{V} \sigma(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \int_{0}^{\infty} \mathrm{e}^{-\int_{0}^{t} \Sigma(\boldsymbol{x} - s\boldsymbol{v}', \boldsymbol{v}') \, \mathrm{d}s} \chi_{\Omega}(\boldsymbol{x} - t\boldsymbol{v}')$$

$$\cdot \int_{V} \sigma(\boldsymbol{x} - t\boldsymbol{v}', \boldsymbol{v}'', \boldsymbol{v}')g(\boldsymbol{x} - t\boldsymbol{v}', \boldsymbol{v}'') \, \mathrm{d}\boldsymbol{v}'' \, \mathrm{d}t \, \mathrm{d}\boldsymbol{v}'$$

$$= \int_{V} \int_{0}^{\infty} \int_{V} \sigma(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \mathrm{e}^{-\int_{0}^{t} \Sigma(\boldsymbol{x} - s\boldsymbol{v}', \boldsymbol{v}') \, \mathrm{d}s} \chi_{\Omega}(\boldsymbol{x} - t\boldsymbol{v}')$$

$$\cdot \sigma(\boldsymbol{x} - t\boldsymbol{v}', \boldsymbol{v}'', \boldsymbol{v}')g(\boldsymbol{x} - t\boldsymbol{v}', \boldsymbol{v}'') \, \mathrm{d}\boldsymbol{v}'' \, \mathrm{d}t \, \mathrm{d}\boldsymbol{v}'$$

$$= \int_{V} \int_{0}^{\infty} \int_{V} \sigma(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \mathrm{e}^{-\int_{0}^{t} \Sigma(\boldsymbol{x} - s\boldsymbol{v}', \boldsymbol{v}') \, \mathrm{d}s} \chi_{\Omega}(\boldsymbol{x} - t\boldsymbol{v}')$$

$$\cdot \sigma(\boldsymbol{x} - t\boldsymbol{v}', \boldsymbol{v}'', \boldsymbol{v}')g(\boldsymbol{x} - t\boldsymbol{v}', \boldsymbol{v}'') \, \mathrm{d}\boldsymbol{v}'' \, \mathrm{d}t \, \mathrm{d}\boldsymbol{v}'$$

$$= \int_{V} \int_{0}^{\infty} \int_{V} \sigma(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \mathrm{e}^{-\int_{0}^{t} \Sigma(\boldsymbol{x} - s\boldsymbol{v}', \boldsymbol{v}') \, \mathrm{d}s} \chi_{\Omega}(\boldsymbol{x} - t\boldsymbol{v}')$$

$$\cdot \sigma(\boldsymbol{x} - t\boldsymbol{v}', \boldsymbol{v}'', \boldsymbol{v}')g(\boldsymbol{x} - t\boldsymbol{v}', \boldsymbol{v}'') \, \mathrm{d}\boldsymbol{v}' \, \mathrm{d}t \, \mathrm{d}\boldsymbol{v}' .$$

In the innermost integral we substitute  $\boldsymbol{v}'$  with  $\boldsymbol{x}'$  via  $\boldsymbol{x}'\coloneqq\boldsymbol{x}-t\boldsymbol{v}'$  to obtain

$$\begin{split} &\int_{V} \int_{0}^{\infty} \int_{\Omega} \sigma\left(\boldsymbol{x}, \frac{\boldsymbol{x} - \boldsymbol{x}'}{t}, \boldsymbol{v}\right) e^{-\int_{0}^{t} \Sigma\left(\boldsymbol{x} - s\frac{\boldsymbol{x} - \boldsymbol{x}'}{t}, \frac{\boldsymbol{x} - \boldsymbol{x}'}{t}\right) \, \mathrm{d}s} \chi_{V}\left(\frac{\boldsymbol{x} - \boldsymbol{x}'}{t}\right) \\ &\cdot \sigma\left(\boldsymbol{x}', \boldsymbol{v}'', \frac{\boldsymbol{x} - \boldsymbol{x}'}{t}\right) g\left(\boldsymbol{x}', \boldsymbol{v}''\right) \frac{1}{t^{n}} \, \mathrm{d}\boldsymbol{x}' \, \mathrm{d}t \, \mathrm{d}\boldsymbol{v}''. \end{split}$$

This can be written as

$$\int_{\Omega \times V} L\left(\boldsymbol{x}, \boldsymbol{v}; \boldsymbol{x}', \boldsymbol{v}''\right) g\left(\boldsymbol{x}', \boldsymbol{v}''\right) \, \mathrm{d}\boldsymbol{x}' \, \mathrm{d}\boldsymbol{v}'',$$

where the kernel  ${\cal L}$  is given by

$$L\left(\boldsymbol{x}, \boldsymbol{v}; \boldsymbol{x}', \boldsymbol{v}''\right) = \int_{0}^{\infty} \sigma\left(\boldsymbol{x}, \frac{\boldsymbol{x} - \boldsymbol{x}'}{t}, \boldsymbol{v}\right) e^{-\int_{0}^{t} \Sigma\left(\boldsymbol{x} - s\frac{\boldsymbol{x} - \boldsymbol{x}'}{t}, \frac{\boldsymbol{x} - \boldsymbol{x}'}{t}\right) ds} \\ \cdot \chi_{V}\left(\frac{\boldsymbol{x} - \boldsymbol{x}'}{t}\right) \sigma\left(\boldsymbol{x}', \boldsymbol{v}'', \frac{\boldsymbol{x} - \boldsymbol{x}'}{t}\right) \frac{1}{t^{n}} dt.$$

Because of Assumption 3.2.11 and Assumption 3.2.12, we have that L > 0. Since we also have that  $g \ge 0$  and  $g \ne 0$ , we obtain that  $FL^{-1}Fg > 0$ .

**Theorem 3.2.14.** [7, Ch. XXI, §3, Sec. 4, Thm. 7]. Under Assumption 3.2.11 and Assumption 3.2.12 we have that the eigenpair  $(\rho, u_{\rho})$  has the following additional properties.

- (i)  $u_{\rho}$  is positive,
- (ii) there is no other eigenfunction associated with another eigenvalue which is of constant sign,
- (iii)  $\rho$  is simple.

*Proof.* (i) Let us first show that  $u_{\rho}$  is positive. We have

$$\rho u_{\rho} = A u_{\rho} = (I - L^{-1}K)^{-1}L^{-1}F u_{\rho},$$

which implies that

$$L^{-1}Fu_{\rho} = \rho(I - L^{-1}K)u_{\rho} = \rho u_{\rho} - \rho L^{-1}Ku_{\rho} \le \rho u_{\rho}$$

since the operator  $L^{-1}K$  is nonnegative and  $u_{\rho}$  is nonnegative. We therefore have

$$\rho u_{\rho} \ge L^{-1} F u_{\rho}.$$

Since the operator  $L^{-1}$  is positive, it is enough to show that  $Fu_{\rho}$  is positive in order to deduce that  $u_{\rho}$  is positive. So we need to show that  $Fu_{\rho} > 0$ . The operator F being linear and nonnegative we obtain

$$\rho F u_{\rho} \ge F L^{-1} F u_{\rho}.$$

But  $u_{\rho}$  is nonnegative and nontrivial so that Lemma 3.2.13 implies

$$FL^{-1}Fu_{\rho} > 0.$$

This in turn implies that  $Fu_{\rho} > 0$  and therefore  $u_{\rho} > 0$  which was to be shown.

(ii) We now show that there is no other eigenfunction associated with another eigenvalue which is of constant sign. To this end, we consider the so-called adjoint problem

$$L^*\psi = K^*\psi + \frac{1}{k}F^*\psi,$$

with the boundary condition  $\psi|_{\Gamma_+} = 0$ . The operators  $L^*$ ,  $K^*$  and  $F^*$  are the adjoints of L, K and F. We have

$$\begin{split} (L^*\psi)\left(\boldsymbol{x},\boldsymbol{v}\right) &= -\boldsymbol{v}\cdot\nabla\psi(\boldsymbol{x},\boldsymbol{v}) + \Sigma(\boldsymbol{x},\boldsymbol{v})\psi(\boldsymbol{x},\boldsymbol{v}),\\ (K^*\psi)\left(\boldsymbol{x},\boldsymbol{v}\right) &= \int_V f(\boldsymbol{x},\boldsymbol{v},\boldsymbol{v}')\psi(\boldsymbol{x},\boldsymbol{v}') \,\mathrm{d}\boldsymbol{v}',\\ (F^*\psi)\left(\boldsymbol{x},\boldsymbol{v}\right) &= \int_V \sigma(\boldsymbol{x},\boldsymbol{v},\boldsymbol{v}')\psi(\boldsymbol{x},\boldsymbol{v}') \,\mathrm{d}\boldsymbol{v}'. \end{split}$$

If  $(\cdot, \cdot)$  denotes the inner product on  $L^2(\Omega \times V)$  we have that

$$\begin{split} (L\phi,\psi) &= (\phi,L^*\psi) \quad \forall \phi \in W^2_+(\Omega \times V), \quad \forall \psi \in W^2_-(\Omega \times V), \\ (K\phi,\psi) &= (\phi,K^*\psi) \quad \forall \phi \in L^2(\Omega \times V), \quad \forall \psi \in L^2(\Omega \times V), \\ (F\phi,\psi) &= (\phi,F^*\psi) \quad \forall \phi \in L^2(\Omega \times V), \quad \forall \psi \in L^2(\Omega \times V). \end{split}$$

Using the fact that for two bounded linear operators  $T_1$  and  $T_2$  we have that  $\rho(T_1T_2) = \rho(T_2T_1)$  (i.e. the spectral radius is invariant under commutation), we obtain

$$\rho((L-K)^{-1}F) = \rho(F(L-K)^{-1}) = \rho((L^*-K^*)^{-1}F^*).$$

But this allows us to apply the same argumentation so far used for problem (2.5) also for the adjoint problem. We therefore obtain that  $\rho$  is also an eigenvalue of the adjoint problem and that there exists a corresponding positive eigenfunction  $u_{\rho}^*$ . So let us now assume that we have a (real) eigenpair  $(k, \psi)$  of our original problem (2.5) with  $k \neq \rho$ . We begin with

$$L\psi = K\psi + \frac{1}{k}F\psi,$$

and take the  $L^2$ -inner product of both sides with  $u_{\rho}^*$ . Then we obtain

$$\left(L\psi, u_{\rho}^{*}\right) = \left(K\psi + \frac{1}{k}F\psi, u_{\rho}^{*}\right).$$
(3.8)

We also take the equation

$$L^* u_{\rho}^* = K^* u_{\rho}^* + \frac{1}{\rho} F^* u_{\rho}^*$$

and take the  $L^2$ -inner product of both sides with  $\psi$ . Then we obtain

$$\left(L^* u_{\rho}^*, \psi\right) = \left(K^* u_{\rho}^* + \frac{1}{\rho} F^* u_{\rho}^*, \psi\right).$$
(3.9)

Subtracting (3.9) from (3.8) we arrive at

$$(L\psi, u_{\rho}^{*}) - (L^{*}u_{\rho}^{*}, \psi) = \left(K\psi + \frac{1}{k}F\psi, u_{\rho}^{*}\right) - \left(K^{*}u_{\rho}^{*} + \frac{1}{\rho}F^{*}u_{\rho}^{*}, \psi\right).$$

Since both  $u_{\rho}$  and  $\psi$  are real-valued, we obtain for the left hand side

$$(L\psi, u_{\rho}^{*}) - (L^{*}u_{\rho}^{*}, \psi) = (\psi, L^{*}u_{\rho}^{*}) - (L^{*}u_{\rho}^{*}, \psi) = (L^{*}u_{\rho}^{*}, \psi) - (L^{*}u_{\rho}^{*}, \psi) = 0,$$

and similarly for the right hand side

$$\left(K\psi + \frac{1}{k}F\psi, u_{\rho}^{*}\right) - \left(K^{*}u_{\rho}^{*} + \frac{1}{\rho}F^{*}u_{\rho}^{*}, \psi\right) = \left(\frac{1}{k} - \frac{1}{\rho}\right)\left(F\psi, u_{\rho}^{*}\right).$$

But this means that

$$\left(\frac{1}{k} - \frac{1}{\rho}\right) \left(F\psi, u_{\rho}^{*}\right) = 0.$$

And hence  $(F\psi, u_{\rho}^*) = 0$ . Since  $u_{\rho}^*$  is positive and F is a positive operator (since the kernel  $\sigma$  is positive) we conclude that  $\psi$  cannot be of constant sign.

(iii) We now show that  $\rho$  is simple. Assume the contrary, i.e. there exists  $w_{\rho}$  eigenvector of A such that  $u_{\rho}$  and  $w_{\rho}$  are linearly independent. Then it is possible to choose a  $\gamma \in \mathbb{R}$ such that  $v = u_{\rho} + \gamma w_{\rho}$  is not of constant sign. We denote  $(\cdot, \cdot)$  for the inner product on  $L^{2}(\Omega \times V)$  and consider

$$E_{1} \coloneqq \left( \left| (L-K)^{-1}F(L-K)^{-1}Fv \right|, F^{*}u_{\rho}^{*} \right) = \left( \left| A^{2}v \right|, F^{*}u_{\rho}^{*} \right) \\ = \left( \left| \rho^{2}v \right|, F^{*}u_{\rho}^{*} \right) = \rho^{2} \left( \left| v \right|, F^{*}u_{\rho}^{*} \right),$$

and

$$E_{2} \coloneqq \left( (L-K)^{-1}F(L-K)^{-1}F|v|, F^{*}u_{\rho}^{*} \right) = \left( F(L-K)^{-1}F|v|, (L^{*}-K^{*})^{-1}F^{*}u_{\rho}^{*} \right)$$
$$= \left( F(L-K)^{-1}F|v|, \rho u_{\rho}^{*} \right)$$
$$= \left( F|v|, \rho (L^{*}-K^{*})^{-1}F^{*}u_{\rho}^{*} \right)$$
$$= \left( F|v|, \rho^{2}u_{\rho}^{*} \right)$$
$$= \left( |v|, \rho^{2}F^{*}u_{\rho}^{*} \right)$$
$$= \rho^{2} \left( |v|, F^{*}u_{\rho}^{*} \right)$$
$$= E_{1}.$$

Hence we have  $E_1 = E_2$ . Let us now consider the functions  $|v| \pm v$ . They are first of all nonnegative. But since v is not of constant sign they are also nontrivial. We have

$$(L-K)^{-1} = (L(I-L^{-1}K))^{-1} = (I-L^{-1}K)^{-1}L^{-1} = \sum_{m=0}^{\infty} (L^{-1}K)^m L^{-1} \ge L^{-1},$$

since the operators  $L^{-1}$  and K are nonnegative (we note that for two operators  $T_1$  and  $T_2$  on  $L^2(\Omega \times V)$  the notation  $T_1 \geq T_2$  is understood as  $T_1g \geq T_2g$  for all  $g \in L^2(\Omega \times V)$  where the latter notion of inequality has already been introduced). We therefore obtain

$$(L-K)^{-1}F(L-K)^{-1}F(|v|\pm v) \ge (L-K)^{-1}FL^{-1}F(|v|\pm v) \ge L^{-1}FL^{-1}F(|v|\pm v).$$

The functions  $|v| \pm v$  being nonnegative and nontrivial, Lemma 3.2.13 implies that  $FL^{-1}F(|v|\pm v) > 0$ . Since  $L^{-1}$  is a positive operator, we obtain  $L^{-1}FL^{-1}F(|v|\pm v) > 0$ . Thus

$$(L-K)^{-1}F(L-K)^{-1}F(|v|\pm v) > 0,$$

which implies that

$$(L-K)^{-1}F(L-K)^{-1}F|v| > |(L-K)^{-1}F(L-K)^{-1}Fv|$$

But this in turn leads to the conclusion that  $E_2 > E_1$  which is a contradiction. We conclude that  $\rho$  must be simple<sup>2</sup>.

<sup>&</sup>lt;sup>2</sup>More precisely, we showed that  $\rho$  is geometrically simple. To deduce that  $\rho$  is algebraically simple, one might use a similar argumentation as in [7, Ch. XXI, §3, Sec. 4, Rmk. 12].

Now that  $\rho$  is simple, we can talk about the unique positive eigenfunction associated with it (up to a scaling). Because of Theorem 3.2.14 (ii) this is the only eigenfunction of A which has a physical meaning.

## 4 Standard FEM for the Criticality Problem

### 4.1 Variational Formulation

We recall that Theorem 3.2.14 gives us the existence of a positive eigenpair  $(\rho, u_{\rho})$ . However, so far, we have been concerned with the operator formulation. But for the numerical treatment of the criticality problem, we need a variational formulation. If we then use a finite element method, we will arrive at a generalized matrix eigenvalue problem whose principal eigenvalue might not be real although the one of the continuous problem is. That is why we need to complexify the Hilbert spaces considered so far to allow complex eigenpairs. So we introduce  $H_1 := {}^{\mathbb{C}}W^2_{-}(\Omega \times V)$  and  $H_2 := {}^{\mathbb{C}}L^2(\Omega \times V)$  the complexifications of the corresponding real Hilbert spaces. We denote  $(\cdot, \cdot)_1$  respectively  $(\cdot, \cdot)_2$  for the corresponding (complex) inner products. Thus, so far, we have been interested in eigenpairs  $(\lambda, u) \in \mathbb{C} \times H_1 \setminus \{0\}$  satisfying

$$(L-K)u = \lambda F u. \tag{4.1}$$

And actually, we were not interested in all the eigenpairs, but in the one special eigenpair  $(\lambda, u)$  with  $\lambda$  being real, positive, of smallest modulus, and simple, and u being real and positive. We now pick a test function  $v \in H_2$  and take the complex  $L^2$ -inner product of both sides of (4.1) with v (i.e. we multiply (4.1) with the complex conjugate of  $v \in H_2$  and integrate over phase space  $\Omega \times V$ ) to obtain

$$\underbrace{((L-K)u,v)_2}_{=a(u,v)} = \lambda \underbrace{(Fu,v)_2}_{=b(u,v)}.$$

We have denoted a for the sesquilinear form on the left hand side, and b for the sesquilinear form on the right hand side. Hence, (4.1) can be reformulated as: Find  $(\lambda, u) \in \mathbb{C} \times H_1 \setminus \{0\}$  satisfying

$$a(u,v) = \lambda b(u,v) \qquad \forall v \in H_2, \tag{4.2}$$

and then extract the special eigenpair of interest. We refer to (4.2) as the variational formulation of (4.1). Next, we try to find an abstract framework for this variational formulation that will give us error estimates for the finite element method. A first candidate for such a framework is [3, Ch. 2, Sec. 8], respectively [12]. Hence, we need to

check the main assumptions on the sesquilinear forms a and b made there. First of all, both a and b are continuous on  $H_1 \times H_2$ . Second, we need to check the inf-sup condition for a. The form  $a(\cdot, \cdot)$  satisfies the *inf-sup condition* if

$$\inf_{u \in H_1 \setminus \{0\}} \sup_{v \in H_2 \setminus \{0\}} \frac{|a(u,v)|}{\|u\|_{H_1} \|v\|_{H_2}} \ge \gamma > 0,$$
(4.3a)

$$\forall v \in H_2 \setminus \{0\} : \sup_{u \in H_1 \setminus \{0\}} |a(u, v)| > 0.$$
 (4.3b)

If  $A : H_1 \to H_2$  is the operator representing the sesquilinear form a, i.e.  $a(u, v) = (Au, v)_2$  for all  $u \in H_1$  and for all  $v \in H_2$ , i.e., A is nothing else than the operator L - K, then we have that a satisfies the inf-sup condition if and only if A is boundedly invertible. And because of the bounded inverse Theorem, it is enough to show that A is bijective, because the inverse map will then automatically be bounded. The bijectivity of A is established in the following Theorem.

**Theorem 4.1.1.** [7, Ch. XXI, §2, Sec. 4.1, Thm. 4]. Under Assumption 3.1.5 we have that the operator  $A : W^2_{-}(\Omega \times V) \to L^2(\Omega \times V)$  is bijective.

Note that this Theorem establishes the bijectivity of A in the real setting, i.e. as a map from  $W_{-}^{2}(\Omega \times V)$  to  $L^{2}(\Omega \times V)$  (if we do not use a superscript  $\mathbb{C}$ , it means that the spaces are real). However, the bijectivity of A in the complex setting, i.e. as a map from  $H_{1}$  to  $H_{2}$ , then immediately follows. We conclude that a satisfies the infsup condition. The last assertion left to check so that (4.2) fits into the setting of [3, Ch. 2, Sec. 8], respectively [12], is, that the map  $A^{-1}F : H_{1} \to H_{1}$  is compact. We know that  $A^{-1}F : H_{2} \to H_{2}$  is compact (see the proof of Corollary 3.2.4), but this is not enough to deduce that  $A^{-1}F : H_{1} \to H_{1}$  is compact. If  $H_{1}$  was compactly embedded in  $H_{2}$ , i.e. if  $W_{-}^{2}(\Omega \times V)$  was compactly embedded in  $L^{2}(\Omega \times V)$ , then we could follow the argumentation in [3, Ch. 2, Sec. 8], to obtain the compactness of  $A^{-1}F : H_{1} \to H_{1}$ . However,  $W_{-}^{2}(\Omega \times V)$  is not compactly embedded in  $L^{2}(\Omega \times V)$ , as shown in the following two remarks.

**Remark 4.1.2.** We first show that  $W^2(\Omega \times V)$  is not compactly embedded in  $L^2(\Omega \times V)$ . To see this, consider  $\Omega = (-1, 1) = V$ . We define the following sequence of functions

$$f_n(x,v) = \begin{cases} C(n) & \text{if } |v| < 2^{-n}, \\ 0 & \text{otherwise.} \end{cases} \quad n \ge 1.$$

Each function  $f_n$  is supported in a velocity slice as illustrated in the following picture.


First of all,  $f_n$  belongs to  $L^2(\Omega \times V)$ . But since it does not depend on x, its derivative w.r.t. x exists and is zero. Hence,  $f_n$  belongs to  $W^2(\Omega \times V)$  and  $||f_n||_{W^2(\Omega \times V)} = ||f_n||_{L^2(\Omega \times V)}$ . It remains to examine the  $L^2$ -norm of  $f_n$ . We have

$$||f_n||_{L^2(\Omega \times V)}^2 = C(n)^2 2^{-n+2}.$$

Hence, if we choose  $C(n) := 2^{\frac{n}{2}-1}$ , then we achieve  $||f_n||_{L^2(\Omega \times V)} = 1$  for all n. Then, the sequence  $f_n$  is a bounded sequence in  $W^2(\Omega \times V)$ . However, there does not exist a subsequence which converges in  $L^2(\Omega \times V)$ . To see this, let n > m, and consider

$$||f_n - f_m||^2_{L^2(\Omega \times V)} \ge 4 C(m)^2 \left(2^{-m} - 2^{-n}\right) = 2^m \left(2^{-m} - 2^{-n}\right) = 1 - 2^{m-n} \ge \frac{1}{2}.$$

But this means that there cannot exist a subsequence of  $\{f_n\}_{n\geq 1}$  which is Cauchy in  $L^2(\Omega \times V)$ , i.e. there cannot exist a subsequence of  $\{f_n\}_{n\geq 1}$  which converges in  $L^2(\Omega \times V)$ .

**Remark 4.1.3.** One might hope that  $W^2_{-}(\Omega \times V)$  is compactly embedded in  $L^2(\Omega \times V)$ . However, this is also not the case. To show this though, we cannot use the functions  $f_n$  from before, since they do not belong to  $W^2_{-}(\Omega \times V)$ , i.e. they do not satisfy the no-inflow boundary conditions. Instead, we let  $\epsilon > 0$  and consider the little square  $(-\epsilon, \epsilon)^2 \subset \Omega \times V = (-1, 1)^2$  centered at the origin. We then define the "tent" function  $f_{\epsilon}$  supported on this square. Formally,  $f_{\epsilon}$  is defined as

$$f_{\epsilon}(x,v) \coloneqq \begin{cases} f_{\epsilon}(x) & \text{if } (x,v) \in (-\epsilon,\epsilon)^2, \\ 0 & \text{if } (x,v) \notin (-\epsilon,\epsilon)^2, \end{cases}$$

where  $f_{\epsilon}(x)$  is defined as

$$f_{\epsilon}(x) \coloneqq \begin{cases} \frac{C(\epsilon)}{\epsilon} \left( x + \epsilon \right) & \text{if } -\epsilon < x \le 0, \\ \frac{C(\epsilon)}{\epsilon} \left( \epsilon - x \right) & \text{if } 0 < x < \epsilon, \end{cases}$$

where  $C(\epsilon) > 0$  is the height of the tent and depends on  $\epsilon$ . Since  $f_{\epsilon}(x, v)$  is continuous in x for fixed v, the partial derivative  $\frac{\partial f_{\epsilon}}{\partial x}$  exists as a function in  $L^2(\Omega \times V)$ . Hence,  $f_{\epsilon}$  belongs to  $W^2(\Omega \times V)$ , and since it surely satisfies the no-inflow boundary conditions, it even belongs to  $W^2_{-}(\Omega \times V)$ . Let us now determine the norms  $||f_{\epsilon}||_{L^2(\Omega \times V)}$  and  $||v \frac{\partial f_{\epsilon}}{\partial x}||_{L^2(\Omega \times V)}$ . We have

$$||f_{\epsilon}||_{L^{2}(\Omega \times V)}^{2} = \int_{\Omega \times V} f_{\epsilon}^{2} \, \mathrm{d}x \, \mathrm{d}v = \int_{(-\epsilon,\epsilon)^{2}} f_{\epsilon}^{2} \, \mathrm{d}x \, \mathrm{d}v = 2\left(\frac{1}{3}C(\epsilon)^{2}\epsilon\right)2\epsilon = \frac{4}{3}C(\epsilon)^{2}\epsilon^{2},$$

since this integral is nothing but the volume under the graph of  $f_{\epsilon}^2$ . If we now set  $C(\epsilon) \coloneqq \frac{1}{\epsilon}$ , then we obtain  $||f_{\epsilon}||_{L^2(\Omega \times V)}^2 = \frac{4}{3}$ , independent of  $\epsilon > 0$ . We further have

$$\left| \left| v \frac{\partial f_{\epsilon}}{\partial x} \right| \right|_{L^2(\Omega \times V)}^2 = \int_{\Omega \times V} v^2 \left( \frac{\partial f_{\epsilon}}{\partial x} \right)^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{(-\epsilon,\epsilon)^2} v^2 \frac{C(\epsilon)^2}{\epsilon^2} \, \mathrm{d}x \, \mathrm{d}v = \frac{4}{3} C(\epsilon)^2 \epsilon^2 = \frac{4}{3},$$

since we have chosen  $C(\epsilon) = \frac{1}{\epsilon}$ . We therefore have that  $||f_{\epsilon}||^2_{W^2(\Omega \times V)} = \frac{4}{3} + \frac{4}{3} = \frac{8}{3}$ , independent of  $\epsilon > 0$ . We now consider the sequence  $\{g_n\}_{n\geq 1}$ , where  $g_n \coloneqq f_{2^{-n}}$ . This is then a sequence in  $W^2_{-}(\Omega \times V)$ , which is bounded in  $W^2_{-}(\Omega \times V)$  (i.e. bounded w.r.t. the  $||\cdot||_{W^2(\Omega \times V)}$ -norm). In what follows, we will show that this sequence cannot have a subsequence which converges in  $L^2(\Omega \times V)$ . Let n > m. We have that  $g_m$  is supported on  $(-\epsilon, \epsilon)^2$ , where  $\epsilon = 2^{-m}$ , and that  $g_n$  is supported on  $(-\delta, \delta)^2$ , where  $\delta = 2^{-n}$ . The following picture illustrates the situation.



Denoting  $A_1 := (-\epsilon, \epsilon) \times (\delta, \epsilon)$  and  $A_2 := (-\epsilon, \epsilon) \times (-\epsilon, -\delta)$ , we obtain

$$||g_n - g_m||_{L^2(\Omega \times V)}^2 = \int_{\Omega \times V} |g_n - g_m|^2 \, \mathrm{d}x \, \mathrm{d}v \ge \int_{A_1 \cup A_2} |g_n - g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}v \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}v \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}v \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}v \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}v \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}v \, \mathrm{d}v = \int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}v \, \mathrm{d}v = \int_{A_$$

since  $g_n$  vanishes on  $A_1 \cup A_2$ . We proceed to obtain

$$\int_{A_1 \cup A_2} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = 2 \int_{A_1} |g_m|^2 \, \mathrm{d}x \, \mathrm{d}v = 2 \left( 2 \cdot \frac{1}{3} C(\epsilon)^2 \epsilon \right) (\epsilon - \delta) = \frac{4}{3} \left( 1 - \frac{\delta}{\epsilon} \right),$$

since  $C(\epsilon) = \frac{1}{\epsilon}$ . Inserting  $\epsilon = 2^{-m}$  and  $\delta = 2^{-n}$ , we obtain

$$||g_n - g_m||^2_{L^2(\Omega \times V)} \ge \frac{4}{3} (1 - 2^{m-n}) \ge \frac{2}{3}$$

since n > m. Hence, the sequence  $\{g_n\}_{n \ge 1}$  cannot have a subsequence which is Cauchy in  $L^2(\Omega \times V)$ , and therefore it cannot have a subsequence which converges in  $L^2(\Omega \times V)$ .

We conclude that the framework presented in [3, Ch. 2, Sec. 8], respectively [12], is not quite appropriate for our variational problem. We can then consider another framework, e.g. [9]. But there the (continuous) trial and test spaces coincide. However, in our variational formulation the test space is clearly larger than the trial space. One might only consider test functions in  $H_1 \subset H_2$ , but then it is not clear whether *a* still satisfies the inf-sup condition. Matters are complicated further by the fact that the transport operator T = L - K :  $W^2_-(\Omega \times V) \rightarrow L^2(\Omega \times V)$  is not self-adjoint (w.r.t. the  $L^2$ -inner product). This is definitely due to the advection term, since the absorption term is clearly symmetric, and the scattering term is, if we assume that f(x, v', v) = f(x, v, v'). Consequently, the bilinear form *a* (defined on the real Hilbert spaces) is not symmetric. But much of the literature deals with the situation where *a* is symmetric. Hence the right framework still needs to be found. Nevertheless, we shall now proceed with the Galerkin disretization.

The Galerkin discretization is obtained by restricting the variational problem (4.2) to finite-dimensional subspaces  $\mathcal{V}_1$  of  $H_1$ , and  $\mathcal{V}_2$  of  $H_2$ , i.e. we want to find eigenpairs  $(\lambda, \phi) \in \mathbb{C} \times \mathcal{V}_1 \setminus \{0\}$  satisfying

$$a(\phi,\psi) = \lambda b(\phi,\psi), \quad \text{for all } \psi \in \mathcal{V}_2.$$
 (4.4)

Note that the subspaces  $\mathcal{V}_1$  and  $\mathcal{V}_2$  are allowed to be complex here. But in practice, they are chosen to be real. Moreover, we are not interested in all the eigenvalues of (4.4), but in the one with smallest modulus. Let now  $\phi_1, \ldots, \phi_N$  be a basis of  $\mathcal{V}_1$  and  $\psi_1, \ldots, \psi_M$  be a basis of  $\mathcal{V}_2$ . Since we are looking for  $\phi$  in  $\mathcal{V}_1$ , we can write  $\phi$  in terms of the basis  $\phi_1, \ldots, \phi_N$ , i.e.

$$\phi = \sum_{j=1}^{N} x_j \phi_j.$$

Since a and b are sesquilinear, it suffices that (4.4) is satisfied for the basis functions  $\psi_1, \ldots, \psi_M$ . But this leads to a generalized matrix eigenvalue problem

$$Ax = \lambda Bx,\tag{4.5}$$

where  $A \in \mathbb{R}^{M \times N}$  is given by  $A = (A_{ij})$ , where  $A_{ij} = a(\phi_j, \psi_i)$ , and  $B \in \mathbb{R}^{M \times N}$  is given by  $B = (B_{ij})$ , where  $B_{ij} = b(\phi_j, \psi_i)$ . The vector x contains the degrees of freedom  $x_1, \ldots, x_N$  of  $\phi$ . But now we encounter a problem: The matrices A and B are in general non-square, since the dimensions of  $\mathcal{V}_1$  and  $\mathcal{V}_2$  are not necessarily the same. To find out whether this really causes problems we need to know how one computes the eigenvalue  $\lambda$  of (4.5) which is of smallest modulus in practice. This is done by the power method. In each step this method takes an iterate  $x_k$  and updates it to  $x_{k+1}$  in the following way. It solves the system

$$Ay = Bx_k,\tag{4.6}$$

and then sets  $x_{k+1} \coloneqq \frac{y}{||y||}$ . However, if the matrix A is non-square, solving the system (4.6) has to be understood in a least-squares sense. Anyway, such non-square eigenvalue problems are not well established in the literature, and besides, all the eigenvalue solvers in e.g. Python or Matlab just take square matrices as arguments. Of course we could manipulate (4.5) to arrive at a square eigenvalue problem (by e.g. taking only the upper square blocks of A and B). But it is not clear how this manipulation will affect the spectrum (spurious eigenvalues!). An easier way to obtain a square eigenvalue problem is to choose  $\mathcal{V}_1$  and  $\mathcal{V}_2$  to be of the same dimension. Or even stronger, we can choose these spaces to be the same. This is the approach we will choose.

### 4.2 Model Problems

Before we describe how to choose the subspace  $\mathcal{V} = \mathcal{V}_1 = \mathcal{V}_2$  explicitly, we will consider a hierarchy of model problems. Hierarchical in the sense that the dimensions of the spatial and the velocity domain are increased. These model problems will help us to get a better understanding of the criticality problem both theoretically and practically. Theoretically, because we will consider special cases of problems which were so far considered in an abstract way. Practically, because we will make the reactor composition explicit, so that it becomes easier to imagine the situation. Moreover, these model problems have been used to test our implementation.

#### 4.2.1 1D space & 0D velocity

For some L > 0 we consider the one-dimensional reactor  $\Omega = (0, L) \subset \mathbb{R}$ . We let  $\mu > 0$ and assume that a neutron has either velocity  $v_1 = +\mu$  or  $v_2 = -\mu$ .<sup>3</sup> The neutron population is then described by two functions  $\phi_1$  and  $\phi_2$  for the respective velocities  $v_1$ and  $v_2$ . The criticality spectral equation then becomes the following system of ordinary differential equations

$$\mu \phi_1'(x) + \Sigma_1(x)\phi_1(x) = f_{21}(x)\phi_2(x) + \frac{1}{k} \left(\sigma_{11}(x)\phi_1(x) + \sigma_{21}(x)\phi_2(x)\right), -\mu \phi_2'(x) + \Sigma_2(x)\phi_2(x) = f_{12}(x)\phi_1(x) + \frac{1}{k} \left(\sigma_{12}(x)\phi_1(x) + \sigma_{22}(x)\phi_2(x)\right).$$

$$(4.7)$$

 $\Sigma_i(x)$  models the probability for a neutron to be absorbed if it is located at x and has velocity  $v_i$ .  $f_{ij}(x)$  models how probable it is for a neutron at x with velocity  $v_i$  to scatter and have  $v_j$  as its new velocity. Similarly,  $\sigma_{ij}(x)$  models how probable it is that a neutron at x with velocity  $v_i$  will induce a fission reaction and that a neutron produced in this fission reaction will have  $v_j$  as its velocity. We point out that these parameters are not probabilities in the sense that they are numbers between zero and one. In reality, they can be larger than one. However, to get an intuitive understanding of (4.7), one can still think of them as probabilities. Since  $\phi_1$  is the spatial density for the neutrons moving to the right, and  $\phi_2$  the spatial density for the neutrons moving to the left, the no-inflow boundary conditions become

$$\phi_1(0) = 0,$$
  
 $\phi_2(L) = 0.$ 

**Example 4.2.1.** Let us consider the system of equations (4.7). Setting all the data L,  $\mu$ ,  $\Sigma$ , f,  $\sigma$  to one, we obtain

$$\phi_1' + \phi_1 = \phi_2 + \frac{1}{k} (\phi_1 + \phi_2),$$
  
$$-\phi_2' + \phi_2 = \phi_1 + \frac{1}{k} (\phi_1 + \phi_2),$$

together with the boundary conditions  $\phi_1(0) = 0$  and  $\phi_2(1) = 0$ . We bring this system of ordinary differential equations to matrix form

<sup>&</sup>lt;sup>3</sup>Hence the velocity space is  $V = \{-\mu, +\mu\}$ , i.e. discrete. In particular, V is not open, and therefore this model problem is not covered by the framework used so far. But as already mentioned in footnote 1, the framework of [7] covers this testcase. The corresponding Radon measure would be a Dirac measure at  $-\mu$  plus a Dirac measure at  $+\mu$ .

$$\begin{pmatrix} \phi_1' \\ \phi_2' \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{1}{k} - 1 & \frac{1}{k} + 1 \\ -1 - \frac{1}{k} & 1 - \frac{1}{k} \end{pmatrix}}_{=A} \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}.$$

The solution is given by

$$\begin{pmatrix} \phi_1(x)\\ \phi_2(x) \end{pmatrix} = e^{Ax} \begin{pmatrix} \phi_1(0)\\ \phi_2(0) \end{pmatrix} = e^{Ax} \begin{pmatrix} 0\\ \phi_2(0) \end{pmatrix}.$$

Assuming that k > 0 we have

$$e^{Ax} = \begin{pmatrix} \frac{2\sqrt{k}\cos\left(\frac{2x}{\sqrt{k}}\right) + \sin\left(\frac{2x}{\sqrt{k}}\right) - k\sin\left(\frac{2x}{\sqrt{k}}\right)}{2\sqrt{k}} & \frac{(1+k)\sin\left(\frac{2x}{\sqrt{k}}\right)}{2\sqrt{k}} \\ \frac{-\sin\left(\frac{2x}{\sqrt{k}}\right) - k\sin\left(\frac{2x}{\sqrt{k}}\right)}{2\sqrt{k}} & \frac{2\sqrt{k}\cos\left(\frac{2x}{\sqrt{k}}\right) - \sin\left(\frac{2x}{\sqrt{k}}\right) + k\sin\left(\frac{2x}{\sqrt{k}}\right)}{2\sqrt{k}} \end{pmatrix}.$$

Hence,

$$\phi_2(1) = \frac{2\sqrt{k}\cos\left(\frac{2}{\sqrt{k}}\right) - \sin\left(\frac{2}{\sqrt{k}}\right) + k\sin\left(\frac{2}{\sqrt{k}}\right)}{2\sqrt{k}}\phi_2(0).$$

We are interested in finding the largest k > 0, such that  $\phi_2(1) = 0$ . Thus, we need



Figure 4.1: The graph of M

There is indeed a largest k > 0 which is a zero of M(k) (compare to Theorem 3.2.10). This k is approximately 1.351, i.e. we are supercritical. If we set  $\phi_2(0) = 1$ , the corresponding eigenfunction is



Figure 4.2: The two components of the eigenfunction (4.8)

Note that both components of the eigenfunction are positive (compare to Theorem 3.2.14 (i)). Note also that there is at least a second positive real eigenvalue which is approximately 0.243. However, the corresponding eigenfunction is

$$\begin{pmatrix} \phi_1(x) \\ \phi_2(x) \end{pmatrix} = \begin{pmatrix} 1.261 \sin(4.058x) \\ 1.014 (0.986 \cos(4.058x) - 0.757 \sin(4.058x)) \end{pmatrix},$$
(4.9)

whose both components are not of constant sign, and hence not physically meaningful (compare to Theorem 3.2.14 (ii)).



Figure 4.3: The two components of the nonphysical eigenfunction (4.9)

**Example 4.2.2.** This time we change the physical parameters in such a way that we obtain

$$\phi_1' + 5\phi_1 = 2\phi_2 + \frac{1}{k} \left( \phi_1 + \frac{1}{2}\phi_2 \right),$$
  
$$-\phi_2' + \phi_2 = \phi_1 + \frac{1}{k} \left( 3\phi_1 + \phi_2 \right),$$
  
(4.10)

where we again have the boundary conditions  $\phi_1(0) = 0$  and  $\phi_2(1) = 0$ . First, since  $\Sigma_1 = 5$  and  $\Sigma_2 = 1$ , it is much more probable for neutrons with velocity +1 to be absorbed than for neutrons with velocity -1. Thus, concerning absorption, there is a tendency towards neutrons with velocity -1, since they are less affected by absorption. On the other hand, since  $f_{21} = 2 > 1 = f_{12}$ , it is more probable for a neutron to scatter from -1 to +1 than from +1 to -1. Thus, concerning scattering, there is a tendency towards neutrons with velocity +1. But since  $\sigma_{12} + \sigma_{22} = 4 > \frac{3}{2} = \sigma_{11} + \sigma_{21}$ , we have that more neutrons with velocity -1 are produced in fission reactions than neutrons with velocity +1. Hence, concerning the production of neutrons in fission reactions, there is again a tendency towards neutrons with velocity -1. To sum up, it is not clear whether the overall tendency is towards neutrons with velocity +1 or -1. It might therefore be possible to balance neutron losses and gains, so that we can hope for the existence of a positive eigenpair. We start with the matrix form of (4.10) which is

$$\begin{pmatrix} \phi_1'\\ \phi_2' \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{1}{k} - 5 & \frac{1}{2k} + 2\\ -1 - \frac{3}{k} & 1 - \frac{1}{k} \end{pmatrix}}_{=A} \begin{pmatrix} \phi_1\\ \phi_2 \end{pmatrix}.$$

Again we must consider the bottom right entry of  $e^A$ . There is indeed a largest k > 0 for which this entry is zero and this k is approximately 1.056. Again we are slightly supercritical. However, this time, the two components of the eigenfunction are more interesting, because they are not symmetric to each other.



Figure 4.4: The two components of the eigenfunction for problem (4.10)

Next we ask ourselves what happens when the material parameters are not constant in the whole reactor, but piecewise constant. To this end, let  $x_0, \ldots, x_N$  be an increasing sequence of points in [0, L] with  $x_0 = 0$  and  $x_N = L$ . We assume that in each interval  $I_i = (x_i, x_{i+1})$  the cross sections are constant. We therefore obtain

$$\mu \phi_1' + \Sigma_1^i \phi_1 = f_{21}^i \phi_2 + \frac{1}{k} \left( \sigma_{11}^i \phi_1 + \sigma_{21}^i \phi_2 \right), -\mu \phi_2' + \Sigma_2^i \phi_2 = f_{12}^i \phi_1 + \frac{1}{k} \left( \sigma_{12}^i \phi_1 + \sigma_{22}^i \phi_2 \right),$$
(4.11)

on  $I_i$ . As usual we have the boundary conditions  $\phi_1(0) = 0$  and  $\phi_2(L) = 0$ . On each subinterval  $I_i$  we can solve (4.11) as before, since the data is constant. Hence, we denote  $h_i = x_{i+1} - x_i$ , and

$$A_{i} = \frac{1}{\mu} \begin{pmatrix} \frac{\sigma_{11}^{i}}{k} - \Sigma_{1}^{i} & f_{21}^{i} + \frac{\sigma_{21}^{i}}{k} \\ -f_{12}^{i} - \frac{\sigma_{12}^{i}}{k} & \Sigma_{2}^{i} - \frac{\sigma_{22}^{i}}{k} \end{pmatrix},$$

and then we obtain

$$\begin{pmatrix} \phi_1(L) \\ \phi_2(L) \end{pmatrix} = e^{A_{N-1}h_{N-1}} e^{A_{N-2}h_{N-2}} \cdots e^{A_1h_1} e^{A_0h_0} \begin{pmatrix} 0 \\ \phi_2(0) \end{pmatrix} \cdot \frac{F M}{\sum 2.2 2.5} \frac{1}{f 0.4 2.3} \frac{1}{\sigma 1.8 0.2}$$

Table 4.1: Explicit values of the material parameters for Example 4.2.3

**Example 4.2.3.** We set L and  $\mu$  to one and consider a slice of fissile material (F) that has width  $a \leq 1$ . The slice is centered at  $\frac{1}{2}$ . A moderator (M) is surrounding the fissile material. We assume that the material parameters are independent of the velocities and are given in Table 4.1. If a = 0.25, then we obtain k = 0.862, i.e. we are subcritical. However, if a = 0.75, then k = 1.141, i.e. we are supercritical. The following picture shows the two components of the eigenfunction in the case where a = 0.25.



Figure 4.5: The two components of the eigenfunction for a = 0.25

#### 4.2.2 2D space & 0D velocity

This time we consider a two-dimensional reactor  $\Omega = (0, L) \times (0, L) \subset \mathbb{R}^2$ . We let  $\boldsymbol{\mu} = (\mu_1, \mu_2) \in \mathbb{R}^2$  and assume that a neutron either has velocity  $\boldsymbol{v}_1 = \boldsymbol{\mu}$  or  $\boldsymbol{v}_2 = -\boldsymbol{\mu}$ .

The neutron population is therefore described by two spatial density functions  $\phi_1(x_1, x_2)$ (for the neutrons having velocity  $v_1 = \mu$ ) and  $\phi_2(x_1, x_2)$  (for the neutrons having velocity  $v_2 = -\mu$ ). Denoting  $\boldsymbol{x} = (x_1, x_2)$ , the criticality spectral equation then becomes the following system of equations

$$\boldsymbol{\mu} \cdot \nabla \phi_{1}(\boldsymbol{x}) + \Sigma_{1}(\boldsymbol{x})\phi_{1}(\boldsymbol{x}) = f_{21}(\boldsymbol{x})\phi_{2}(\boldsymbol{x}) + \frac{1}{k} \left(\sigma_{11}(\boldsymbol{x})\phi_{1}(\boldsymbol{x}) + \sigma_{21}(\boldsymbol{x})\phi_{2}(\boldsymbol{x})\right), - \boldsymbol{\mu} \cdot \nabla \phi_{2}(\boldsymbol{x}) + \Sigma_{2}(\boldsymbol{x})\phi_{2}(\boldsymbol{x}) = f_{12}(\boldsymbol{x})\phi_{1}(\boldsymbol{x}) + \frac{1}{k} \left(\sigma_{12}(\boldsymbol{x})\phi_{1}(\boldsymbol{x}) + \sigma_{22}(\boldsymbol{x})\phi_{2}(\boldsymbol{x})\right).$$
(4.12)

The no-inflow boundary conditions depend on the direction of  $\mu$ . Let us assume for simplicity that the  $x_1$ -component  $\mu_1$  of  $\mu$  is positive, so that  $\mu$  always points to the right. The following pictures illustrate the no-inflow boundary conditions depending on the sign of  $\mu_2$ . The red respectively green part of  $\partial\Omega$  is where  $\phi_1$  respectively  $\phi_2$  must vanish.



We can further specify the material geometry in the reactor. To this end, we divide  $(0, L) \times (0, L)$  uniformly into a grid of  $N^2$  cells. In each cell we take the midpoint m. Then we consider the disk centered at m with diameter  $a \leq \frac{L}{N}$ . In this manner we obtain  $N^2$  disks (see the illustration for N = 3).



These disks contain the fissile material and the rest of  $\Omega$  is occupied by the moderator. We note that as in the (1 + 0)-dimensional case the material parameters  $\Sigma$ , f,  $\sigma$  are independent of  $\boldsymbol{x}$  within the fissile material or the moderator. In particular, we can use

the same explicit values as given in Table 4.1 for testing purposes. However, there is a problem. Consider the system of equations (4.12). We assume that  $\mu_2 = 0$ , so that concerning the boundary conditions, we are in the situation of the middle picture of the three ones given before. Let  $\boldsymbol{x}_0$  be a point on the red part of  $\partial\Omega$ . Consider now the path  $\gamma(t) = \boldsymbol{x}_0 + t\boldsymbol{\mu}$  starting at  $\boldsymbol{x}_0$  and passing  $\Omega$  in the direction of  $\boldsymbol{\mu}$ . Evaluating (4.12) at  $\boldsymbol{x} = \gamma(t)$ , we obtain

$$\widetilde{\phi}_{1}'(t) + \widetilde{\Sigma}_{1}(t)\widetilde{\phi}_{1}(t) = \widetilde{f}_{21}(t)\widetilde{\phi}_{2}(t) + \frac{1}{k}\left(\widetilde{\sigma}_{11}(t)\widetilde{\phi}_{1}(t) + \widetilde{\sigma}_{21}(t)\widetilde{\phi}_{2}(t)\right),$$

$$-\widetilde{\phi}_{2}'(t) + \widetilde{\Sigma}_{2}(t)\widetilde{\phi}_{2}(t) = \widetilde{f}_{12}(t)\widetilde{\phi}_{1}(t) + \frac{1}{k}\left(\widetilde{\sigma}_{12}(t)\widetilde{\phi}_{1}(t) + \widetilde{\sigma}_{22}(t)\widetilde{\phi}_{2}(t)\right),$$

$$(4.13)$$

where  $\tilde{g}(t) = g(\gamma(t))$  for a corresponding function  $g = g(\boldsymbol{x})$ . But this means that on  $\gamma$ , the (2+0)-dimensional problem (4.12) is just the (1+0)-dimensional problem (4.7) we have already encountered. This in turn implies that the largest eigenvalue k > 0 of (4.12) is the one coming from (4.13) when the path  $\gamma$  is the one passing the most fissile material (i.e. the one that exactly bisects a row of disks). In particular, this eigenvalue is not simple. Its multiplicity is N, since there are N such paths.

#### 4.2.3 0D space & 1D velocity

We consider the neutrons to be concentrated in one point. A neutron cannot change its position, but its velocity. Moreover, we assume that this velocity can only change along a fixed axis. Hence, the velocity domain V is a subset of  $\mathbb{R}$  (e.g. V = (-1, 1) or  $V = (-1, -\frac{1}{2}) \cup (\frac{1}{2}, 1)$ ). The neutron population is thus described by  $\phi(v)$ , a density w.r.t. velocity. The criticality spectral equation becomes

$$\Sigma(v)\phi(v) = \int_V f(v',v)\phi(v') \,\mathrm{d}v' + \frac{1}{k} \int_V \sigma(v',v)\phi(v') \,\mathrm{d}v'.$$

We note that this time there is no advection term and hence no boundary condition.

**Example 4.2.4.** We consider the following example

$$\phi(v) = \frac{c}{k} \int_{-1}^{1} |v - v'| \phi(v') \, \mathrm{d}v', \quad v \in (-1, 1),$$

where c is a parameter to control the fission, and we assumed that scattering does not occur. The larger c, the more fission reactions. This is a Fredholm integral equation of the second kind whose solution (see [17, Sec. 4.1, Ex. 6]) is

$$\phi(v) = C_1 \cosh\left(mv\right) + C_2 \sinh\left(mv\right),$$

where  $m = \sqrt{\frac{2c}{k}}$ , and the constants  $C_1$ ,  $C_2$  are determined by the conditions

$$\phi'(-1) + \phi'(1) = 0,$$
  
$$\phi(-1) + \phi(1) + 2\phi'(-1) = 0.$$

This leads to the following system

$$\underbrace{\begin{pmatrix} 0 & 2m\cosh(m) \\ 2\cosh(m) - 2m\sinh(m) & 2m\cosh(m) \end{pmatrix}}_{=M} \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$
 (4.14)

Since we are interested in nontrivial solutions of (4.14), we have to make sure that the determinant of M, which is

$$4m\cosh\left(m\right)\left(m\sinh\left(m\right) - \cosh\left(m\right)\right),\tag{4.15}$$

vanishes. The smallest m > 0 for which (4.15) is zero is 1.19968. The corresponding largest k > 0 is then given by  $k = \frac{2c}{m^2} = 1.38963c$ . If the fission parameter c is set to one, then we are supercritical with k = 1.38963. If it is set to 0.7, then we are subcritical with k = 0.972743. The corresponding eigenfunction is always given by  $\cosh(mv)$ .

#### 4.2.4 1D space & 1D velocity

We consider again the one-dimensional reactor  $\Omega = (0, L) \subset \mathbb{R}$ . A neutron has its velocity in  $V \subset \mathbb{R}$ , where e.g. V = (-1, 1) or  $V = (-1, -\frac{1}{2}) \cup (\frac{1}{2}, 1)$ . The neutron distribution  $\phi(x, v)$  is a density in phase space and satisfies

$$v\frac{\partial\phi}{\partial x}(x,v) + \Sigma(x,v)\phi(x,v)$$

$$= \int_{V} f(x,v',v)\phi(x,v') \,\mathrm{d}v' + \frac{1}{k} \int_{V} \sigma(x,v',v)\phi(x,v') \,\mathrm{d}v'.$$
(4.16)

The no-inflow boundary conditions become

$$\begin{split} \phi(0,v) &= 0 \qquad \forall v > 0, \\ \phi(L,v) &= 0 \qquad \forall v < 0. \end{split}$$

It is important to first understand the source problem

$$v\frac{\partial\phi}{\partial x}(x,v) + \Sigma(x,v)\phi(x,v) = g(x,v), \qquad (4.17)$$

where  $g \in L^2(\Omega \times V)$ . Important for both theoretical and practical reasons. Theoretically, because this problem is nothing but the operator equation  $L\phi = g$  and studying it will help us to understand the operator  $L^{-1}$ . Practically, because (4.17) can be used as a testcase to validate our implementation. Concretely, if we discretize (4.16) with a finite element method, we will arrive at a generalized matrix eigenvalue problem. Before we can solve this generalized matrix eigenvalue problem, we have to be sure that the matrices appearing were set up correctly, e.g. the one discretizing the operator L. But since this matrix is basically the same as the one appearing in a source problem for L, we can test this matrix with problem (4.17). For these reasons, we shall now determine the exact solution of this source problem for general enough data.

To this end, let  $x_0, \ldots, x_N$  be an increasing sequence of points in [0, L] with  $x_0 = 0$ and  $x_N = L$ . In each interval  $I_i = (x_i, x_{i+1}), i = 0, \ldots, N-1$ , there is a material  $M_i$ . We assume that in each  $M_i$  the data  $\Sigma(x, v)$  and g(x, v) reduce to functions  $\Sigma_i(v)$  respectively  $g_i(v)$ .

$$\begin{array}{cccc} \Sigma_{0}(v) & \Sigma_{i}(v) & \Sigma_{N-1}(v) \\ g_{0}(v) & g_{i}(v) & g_{N-1}(v) \\ & & M_{0} & M_{i} & M_{N-1} \\ 0 = x_{0} & x_{1} & x_{i} & x_{i+1} & x_{N-1} & x_{N} = L \end{array}$$

For  $(x, v) \in \Omega \times V$  let N(x, v) be the number of materials a neutron visits, when it starts at x with velocity -v. Let  $M_{i_1}, \ldots, M_{i_{N(x,v)}}$  be the ordered sequence of materials this neutron visits, i.e. it first visits  $M_{i_1}$ , then  $M_{i_2}$ , and finally  $M_{i_{N(x,v)}}$ . Let  $d_{i_1}(x, v), \ldots, d_{i_{N(x,v)}}(x, v)$  be the times which are needed for this neutron to pass  $M_{i_1}, \ldots, M_{i_{N(x,v)}}$ . Then the solution of (4.17) at (x, v) is given by

$$\begin{split} \phi(x,v) &= \frac{g_{i_1}(v)}{\Sigma_{i_1}(v)} \left(1 - e^{-\Sigma_{i_1}(v)d_{i_1}(x,v)}\right) \\ &+ e^{-\Sigma_{i_1}(v)d_{i_1}(x,v)} \frac{g_{i_2}(v)}{\Sigma_{i_2}(v)} \left(1 - e^{-\Sigma_{i_2}(v)d_{i_2}(x,v)}\right) \\ &+ e^{-\Sigma_{i_1}(v)d_{i_1}(x,v)} e^{-\Sigma_{i_2}(v)d_{i_2}(x,v)} \frac{g_{i_3}(v)}{\Sigma_{i_3}(v)} \left(1 - e^{-\Sigma_{i_3}(v)d_{i_3}(x,v)}\right) \\ &+ \cdots \\ &+ e^{-\Sigma_{i_1}(v)d_{i_1}(x,v)} \cdots e^{-\Sigma_{i_N(x,v)-1}(v)d_{i_N(x,v)-1}(x,v)} \frac{g_{i_N(x,v)}(v)}{\Sigma_{i_N(x,v)}(v)} \left(1 - e^{-\Sigma_{i_N(x,v)}(v)d_{i_N(x,v)}(x,v)}\right). \end{split}$$

This can be shown by using formula (3.1). To validate our implementation we shall also need a source problem for the operator L - K, i.e. we consider the following problem

$$v\frac{\partial\phi}{\partial x}(x,v) + \Sigma(x,v)\phi(x,v) - \int_{V} f(x,v',v)\phi(x,v') \,\mathrm{d}v' = g(x,v). \tag{4.18}$$

To obtain a testproblem for (4.18) one can take the testproblem described for (4.17) and then adapt the right hand side accordingly, so that the solution remains the same. This is done in the following example.

**Example 4.2.5.** Let  $\Omega = (0,1)$  and  $V = (-1, -\frac{1}{2}) \cup (\frac{1}{2}, 1)$ . Let  $\Sigma \equiv \Sigma_0$  and  $g \equiv g_0$ . Then the solution of (4.17) is given by

$$\phi(x,v) = \frac{g_0}{\Sigma_0} \begin{cases} 1 - e^{-\Sigma_0 \frac{x}{v}} & \text{if } v > 0, \\ 1 - e^{-\Sigma_0 \frac{x-1}{v}} & \text{if } v < 0. \end{cases}$$

So we know that  $L\phi = g$ . But this implies  $(L - K)\phi = g - K\phi$ , i.e.  $\phi$  solves (4.18), where the right hand side is  $g - K\phi$ . It remains to determine  $K\phi$ . To this end, we choose  $f(x, v', v) = \frac{1}{(v')^2}$ . Then we obtain

$$\begin{split} (K\phi)(x,v) &= \int_{V} f(x,v',v)\phi(x,v') \, \mathrm{d}v' \\ &= \frac{g_{0}}{\Sigma_{0}} \left( \int_{-1}^{-\frac{1}{2}} \frac{1}{(v')^{2}} \left( 1 - e^{-\Sigma_{0}\frac{x-1}{v}} \right) \, \mathrm{d}v' + \int_{\frac{1}{2}}^{1} \frac{1}{(v')^{2}} \left( 1 - e^{-\Sigma_{0}\frac{x}{v}} \right) \, \mathrm{d}v' \right) \\ &= \frac{g_{0}}{\Sigma_{0}} \left( 2 - \frac{e^{2\alpha} - e^{\alpha}}{\alpha} - \frac{e^{-\beta} - e^{-2\beta}}{\beta} \right), \end{split}$$

where  $\alpha = \Sigma_0(x-1)$  and  $\beta = \Sigma_0 x$ .

#### 4.2.5 Further Model Problems

Let us first consider the (2 + 1)-dimensional problem. There, the reactor  $\Omega$  is  $(0, L) \times (0, L) \subset \mathbb{R}^2$  and we assume that the neutrons can move in any direction, but their speed is fixed. The velocity space V can then be chosen as  $S^1 \subset \mathbb{R}^{2,4}$  The neutron density  $\phi$  is then a density w.r.t. position and angle and satisfies

$$\boldsymbol{v} \cdot \nabla \phi(\boldsymbol{x}, \boldsymbol{v}) + \Sigma(\boldsymbol{x}, \boldsymbol{v}) \phi(\boldsymbol{x}, \boldsymbol{v}) = \int_{V} f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \phi(\boldsymbol{x}, \boldsymbol{v}') \, \mathrm{d}\boldsymbol{v}' + \frac{1}{k} \int_{V} \sigma(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \phi(\boldsymbol{x}, \boldsymbol{v}') \, \mathrm{d}\boldsymbol{v}'.$$

<sup>&</sup>lt;sup>4</sup>Again V is not open. However, this model problem still fits into the framework of [7].

Note that if  $V = S^1$ , the integrals are taken w.r.t. the surface measure on  $S^{1.5}$ 

Let us now consider the **quasi** (2 + 2)-dimensional problem. Again we start from the two-dimensional reactor  $\Omega = (0, L) \times (0, L)$ . The neutrons can still move in any direction, but this time, their speed does not need to be fix, but is in a certain range. Hence, the velocity space is  $V = \{ v \in \mathbb{R}^2 : a \leq |v| \leq b \}$ . The neutron density is a density w.r.t. the Lebesgue measure on  $\Omega$  and w.r.t. the Lebesgue measure on V. The equation is still

$$\boldsymbol{v} \cdot \nabla \phi(\boldsymbol{x}, \boldsymbol{v}) + \Sigma(\boldsymbol{x}, \boldsymbol{v}) \phi(\boldsymbol{x}, \boldsymbol{v}) = \int_{V} f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \phi(\boldsymbol{x}, \boldsymbol{v}') \, \mathrm{d}\boldsymbol{v}' + \frac{1}{k} \int_{V} \sigma(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \phi(\boldsymbol{x}, \boldsymbol{v}') \, \mathrm{d}\boldsymbol{v}',$$

but this time the boundary conditions are mixed, i.e. denoting

$$\begin{split} & \Gamma_1 = (0,L) \times \{0\}, \\ & \Gamma_2 = \{1\} \times (0,L), \\ & \Gamma_3 = (0,L) \times \{1\}, \\ & \Gamma_4 = \{0\} \times (0,L), \end{split}$$

we have no-inflow boundary conditions on  $\Gamma_2$  and  $\Gamma_4$  and periodic boundary conditions on  $\Gamma_1$  and  $\Gamma_3$ . Having periodic boundary conditions amounts to saying that a neutron leaving  $\Gamma_1$  will enter  $\Gamma_3$  at exactly the opposite point, and the same for a neutron leaving  $\Gamma_3$ .

In the actual (2 + 2)-dimensional problem the situation is the same as in the previous one except that now we also have no-inflow boundary conditions on  $\Gamma_1$  and  $\Gamma_3$ .

### 4.3 Choice of the FE subspace in 1D

Recall the (1 + 1)-dimensional problem: The reactor is  $\Omega = (0, L)$ , and the velocity domain is  $V = (-1, -\frac{1}{2}) \cup (\frac{1}{2}, 1)$ . The equation is

$$v \frac{\partial \phi}{\partial x}(x,v) + \Sigma(x,v)\phi(x,v)$$

$$= \int_{V} f(x,v',v)\phi(x,v') \,\mathrm{d}v' + \frac{1}{k} \int_{V} \sigma(x,v',v)\phi(x,v') \,\mathrm{d}v',$$
(4.19)

<sup>&</sup>lt;sup>5</sup>Again we see the advantage of the framework presented in [7]. Because there the velocity space is defined as the support of a positive Radon measure  $\mu$  and then all the integrals are taken w.r.t. that measure  $\mu$ .

and the boundary conditions are

$$\begin{split} \phi(0,v) &= 0 \qquad \forall v > 0, \\ \phi(L,v) &= 0 \qquad \forall v < 0. \end{split}$$

We are looking for  $\phi \in W^2_{-}(\Omega \times V)$ . We then multiplied (4.19) with a test function  $\psi \in L^2(\Omega \times V)$ , and integrated over phase space to obtain the variational formulation

$$\begin{split} \int_{\Omega \times V} v \frac{\partial \phi}{\partial x}(x, v) \psi(x, v) \, \mathrm{d}x \, \mathrm{d}v + \int_{\Omega \times V} \Sigma(x, v) \phi(x, v) \psi(x, v) \, \mathrm{d}x \, \mathrm{d}v \\ &= \int_{\Omega \times V} \int_{V} f(x, v', v) \phi(x, v') \, \mathrm{d}v' \psi(x, v) \, \mathrm{d}x \, \mathrm{d}v \\ &+ \frac{1}{k} \int_{\Omega \times V} \int_{V} \sigma(x, v', v) \phi(x, v') \, \mathrm{d}v' \psi(x, v) \, \mathrm{d}x \, \mathrm{d}v. \end{split}$$

We then obtained the Galerkin formulation by restricting the solution space to  $\mathcal{V}_1 \subset W^2_-(\Omega \times V)$  and the test space to  $\mathcal{V}_2 \subset L^2(\Omega \times V)$ . This leads to a generalized matrix eigenvalue problem in m rows and n columns, where n is the dimension of  $\mathcal{V}_1$ , and m is the dimension of  $\mathcal{V}_2$ . To guarantee a square matrix eigenvalue problem, we can choose the subspaces  $\mathcal{V}_1$  and  $\mathcal{V}_2$  to be the same. Hence, we are left to specify the space  $\mathcal{V} = \mathcal{V}_1 = \mathcal{V}_2$ . To this end, let us now specify the meshes for the spatial and the velocity domain. The spatial domain shall be composed of N copies of a unit cell. The unit cell  $\hat{\Omega}$  is (0, 1), where in the first third  $(0, \frac{1}{3})$  we have the moderator (M), in the second third  $(\frac{1}{3}, \frac{2}{3})$  we have the fissile material (F), and in the last third  $(\frac{2}{3}, 1)$  we again have the moderator.

$$\widehat{\Omega} = \begin{array}{cccc} M & F & M \\ 0 & \frac{1}{3} & \frac{2}{3} & 1 \end{array}$$

$$\Omega = \begin{array}{cccc} \widehat{\Omega} & \widehat{\Omega} & & & \\ 0 & 1 & 2 & & N-1 & N \end{array}$$

The mesh or grid for  $\Omega$  is now chosen so that the grid points coincide with the material discontinuities, except that between two unit cells we also choose a grid point (although there is no material discontinuity there, since two moderator materials match together). More precisely, the grid points are  $0, \frac{1}{3}, \frac{2}{3}, 1, \frac{4}{3}, \frac{5}{3}, 2, \frac{7}{3}, \ldots, N - \frac{1}{3}, N$ . Let us denote  $\mathcal{T}_{\Omega}$  for this mesh. We use the notation  $K_x$  to denote an element of  $\mathcal{T}_{\Omega}$ . The subscript x is used to emphasize that we consider a spatial element. Given a polynomial degree p we consider the space

$$S^{p,1}(\Omega,\mathcal{T}_{\Omega}) = \left\{ u \in H^1(\Omega) : u|_{K_x} \circ F_{K_x} \in \mathcal{P}_p(0,1), \, \forall \, K_x \in \mathcal{T}_{\Omega} \right\},\$$

where  $F_{K_x}$  denotes the reference element mapping from  $\widehat{K}_x = (0,1)$  to  $K_x$ .<sup>6</sup>  $\mathcal{P}_p(0,1) = \mathcal{P}_p$  is the space of polynomials on the unit interval of degree at most p. A basis  $\mathcal{B}(\mathcal{P}_p)$  of  $\mathcal{P}_p$  is given by

$$\mathcal{B}(\mathcal{P}_p) = \{\xi, 1-\xi\} \cup \{\xi(1-\xi)l_k(\xi) : k = 0, \dots, p-2\},\$$

where  $l_k$  is the k-th Legendre polynomial (on (0, 1)). We have that  $\{\xi, 1 - \xi\}$  are external, since they do not vanish on both boundary points 0 and 1. The other polynomials  $\{\xi(1-\xi)l_k(\xi) : k = 0, \ldots, p-2\}$  are internal. A basis  $\mathcal{B}(S^{p,1}(\Omega, \mathcal{T}_{\Omega}))$  of  $S^{p,1}(\Omega, \mathcal{T}_{\Omega})$  is now given by

$$\mathcal{B}\left(S^{p,1}(\Omega,\mathcal{T}_{\Omega})\right) = \left\{f \circ F_{K_{\tau}}^{-1} \text{ extended to a hat function } : K_{x} \in \mathcal{T}_{\Omega} \text{ and } f \in \mathcal{B}\left(\mathcal{P}_{p}\right)\right\}.$$

So we just take an element  $K_x \in \mathcal{T}_{\Omega}$ . Via the reference element mapping we can map a polynomial in  $\mathcal{B}(\mathcal{P}_p)$  to a function on  $K_x$  (which actually is again a polynomial of the same degree since here  $F_{K_x}$  is affine). If the polynomial taken is internal, we can just extend it by zero without creating a discontinuity. If it is external, we can extend it by zero only on one side. On the other side we have to connect it with the corresponding external polynomial. Let us now consider the velocity domain V. We can partition both  $\left(-1, -\frac{1}{2}\right)$  and  $\left(\frac{1}{2}, 1\right)$  uniformly. Let us denote  $\mathcal{T}_V$  for this mesh. For an element of  $\mathcal{T}_V$ we use the notation  $K_v$ . Let q be a polynomial degree and consider the space

$$S^{q,0}(V, \mathcal{T}_V) = \left\{ u \in L^2(V) : u|_{K_v} \circ F_{K_v} \in \mathcal{P}_q(0, 1), \, \forall \, K_v \in \mathcal{T}_V \right\},\$$

where  $F_{K_v}$  denotes the reference element mapping from  $\widehat{K}_v = (0,1)$  to  $K_v$ . A basis  $\mathcal{B}(S^{q,0}(V,\mathcal{T}_V))$  of  $S^{q,0}(V,\mathcal{T}_V)$  is now given by

$$\mathcal{B}\left(S^{q,0}(V,\mathcal{T}_V)\right) = \left\{f \circ F_{K_v}^{-1} : K_v \in \mathcal{T}_V \text{ and } f \in \mathcal{B}\left(\mathcal{P}_q\right)\right\},\$$

i.e. we consider a velocity element  $K_v$  and a polynomial  $f \in \mathcal{B}(\mathcal{P}_q)$ . Via  $F_{K_v}^{-1}$  we can map f to a polynomial on  $K_v$ . This polynomial  $f \circ F_{K_v}^{-1}$  is then automatically a basis

$$S^{p,\ell}(\Omega,\mathcal{T}_{\Omega}) = \left\{ u \in H^{\ell}(\Omega) : u|_{K_x} \circ F_{K_x} \in \mathcal{P}_p(0,1), \, \forall \, K_x \in \mathcal{T}_{\Omega} \right\}.$$

<sup>&</sup>lt;sup>6</sup>This notation comes from the more general one

Also note that we do not have to use the same polynomial degree on each element  $K_x$ . This will become relevant in the generalized FEM (see Chapter 5) where we will use a small polynomial degree in the interior of the reactor and a higher polynomial degree the closer we get to the boundary.

function of  $S^{q,0}(V, \mathcal{T}_V)$ . So the only thing different this time is that we do not have to extend to a hat function since we do not need to be continuous. We can finally define the subspace  $\mathcal{V}$  as

$$\mathcal{V} \coloneqq \operatorname{Span} \left\{ \phi : \phi(x, v) = \phi_{\Omega}(x) \cdot \phi_{V}(v), \text{ where } \phi_{\Omega} \in \mathcal{B}\left(S^{p, 1}(\Omega, \mathcal{T}_{\Omega})\right) \\ \text{and } \phi_{V} \in \mathcal{B}\left(S^{q, 0}(V, \mathcal{T}_{V})\right) \text{ and } \phi|_{\Gamma_{-}} = 0 \right\}.$$

In words this means that we take the product of a function in  $\mathcal{B}(S^{p,1}(\Omega, \mathcal{T}_{\Omega}))$  (which is a function of the spatial variable x) and a function in  $\mathcal{B}(S^{q,0}(V, \mathcal{T}_V))$  (which is a function of the velocity variable v). Not all the functions obtained in this way (which are now functions in both variables x and v) satisfy the no-inflow boundary conditions. We therefore delete the ones which do not satisfy them. The space  $\mathcal{V}$  is then defined as the span of the remaining ones.

**Example 4.3.1.** Let us consider the case where N = 2 (i.e. two cells). Moreover, the velocity mesh  $\mathcal{T}_V$  shall consist of the two elements  $\left(-1, -\frac{1}{2}\right)$  and  $\left(\frac{1}{2}, 1\right)$  only. The polynomial degrees shall be p = 1 = q. The following pictures show the functions of  $\mathcal{B}\left(S^{p,1}(\Omega, \mathcal{T}_{\Omega})\right)$  and  $\mathcal{B}\left(S^{q,0}(V, \mathcal{T}_V)\right)$ .



The product of the *i*-th function in  $\mathcal{B}(S^{p,1}(\Omega,\mathcal{T}_{\Omega}))$  with the *j*-th function in  $\mathcal{B}(S^{q,0}(V,\mathcal{T}_{V}))$ shall be denoted with the pairing (i, j). We see now that not all the products (i, j) satisfy the no-inflow boundary conditions. The ones that do not are (0,2), (0,3) and (6,0), (6,1). These have to be deleted. The space  $\mathcal{V}$  is defined as the span of the remaining products. In the code, the no-inflow boundary conditions were implemented as follows. The 0-th function of  $\mathcal{B}(S^{p,1}(\Omega,\mathcal{T}_{\Omega}))$  gets the attribute "-1", and the 6-th the attribute "+1". Similarly for the functions of  $\mathcal{B}(S^{q,0}(V,\mathcal{T}_V))$ , where the ones supported on  $(-1,-\frac{1}{2})$ get the attribute "-1", and the ones supported on  $(\frac{1}{2}, 1)$  get the attribute "+1". The products (i, j) which do not satisfy the no-inflow boundary conditions are then exactly the ones where the attribute corresponding to i and the attribute corresponding to j are not of the same sign, i.e. (-1,+1) and (+1,-1). By deleting these products we mean the following. First we set up the Galerkin matrices A and B for both operators L-Kand F, but with all the degrees of freedom, i.e. with all products (i, j). Then we delete the degrees of freedom corresponding to products (i, j) which do not satisfy the no-inflow boundary conditions (and this is done using the attributes  $\pm 1$ ). In this way we obtain two smaller matrices A and B which are still square, because we delete these degrees of

freedom for both the trial and the test space (since we chose  $\mathcal{V} = \mathcal{V}_1 = \mathcal{V}_2$ ). The reduced generalized matrix eigenvalue problem  $\widetilde{A}x = \lambda \widetilde{B}x$  is then solved.

**Remark 4.3.2.** In the numerical experiments (see Chapter 6) we actually also used continuous elements w.r.t. the velocity variable, although we are not forced to do so. This is because the code we used just uses continuous elements. Hence, we used the subspace

$$\mathcal{V} = Span\left\{\phi : \phi(x,v) = \phi_{\Omega}(x) \cdot \phi_{V}(v), \text{ where } \phi_{\Omega} \in \mathcal{B}\left(S^{p,1}(\Omega,\mathcal{T}_{\Omega})\right) \\ and \ \phi_{V} \in \mathcal{B}\left(S^{q,1}(V,\mathcal{T}_{V})\right) \text{ and } \phi|_{\Gamma_{-}} = 0\right\}.$$

We could have modified the code to use discontinuous elements in the velocity variable though. If the eigenfunction is discontinuous in the velocity variable, then using continuous elements might lead to oscillations when they try to resolve the discontinuity.

## 4.4 Fast Quadrature

Since we now know the space  $\mathcal{V}$  and also the basis functions (in fact we first defined the basis functions and then defined  $\mathcal{V}$  as the span of them), the next step is to say how we can set up the Galerkin matrices for L - K and F efficiently by exploring the tensor structure of the basis functions. We shall see in the next section, that modulo an assembly and a transformation, this boils down to the following situation. We have a reference element  $(0,1)^d \subset \mathbb{R}^d$  and a function  $g(\xi_1,\ldots,\xi_d)$  :  $(0,1)^d \to \mathbb{R}$ . The function g will play the role of the material data transformed back to the reference element together with the corresponding Jacobian. In each direction  $k = 1,\ldots,d$ , we have  $I_k$ many functions  $u_{i_k}(\xi_k)$ ,  $i_k = 1,\ldots,I_k$ , for the trial space, and  $J_k$  many functions  $v_{j_k}(\xi_k)$ ,  $j_k = 1,\ldots,J_k$ , for the test space. The  $u_{i_k}$ 's and the  $v_{j_k}$ 's are just the polynomial shape functions we choose in each direction for the trial space respectively for the test space. Our problem consists in computing the integrals

$$\int_{(0,1)^d} g(\xi_1, \dots, \xi_d) u_{i_1}(\xi_1) \cdots u_{i_d}(\xi_d) v_{j_1}(\xi_1) \cdots v_{j_d}(\xi_d) \,\mathrm{d}\boldsymbol{\xi}.$$
(4.20)

In each direction k we choose  $Q_k$  many quadrature points  $\xi_{k,q_k}$  and quadrature weights  $w_{k,q_k}$ ,  $q_k = 1, \ldots, Q_k$ . Then (4.20) is approximated as

$$\sum_{q_1=1}^{Q_1} \cdots \sum_{q_d=1}^{Q_d} w_{1,q_1} \cdots w_{d,q_d} g(\xi_{1,q_1}, \dots, \xi_{d,q_d}) u_{i_1}(\xi_{1,q_1}) \cdots u_{i_d}(\xi_{d,q_d}) v_{j_1}(\xi_{1,q_1}) \cdots v_{j_d}(\xi_{d,q_d}).$$
(4.21)

We are interested in computing the element matrix  $E = E(i_1, j_1, \ldots, i_d, j_d)$  (or better the element tensor or element array), whose entries are given by (4.21). These sums can now be factorized as

$$\sum_{q_1=1}^{Q_1} w_{1,q_1} u_{i_1}(\xi_{1,q_1}) v_{j_1}(\xi_{1,q_1}) \cdots \sum_{q_d=1}^{Q_d} w_{d,q_d} u_{i_d}(\xi_{d,q_d}) v_{j_d}(\xi_{d,q_d}) g(\xi_{1,q_1},\dots,\xi_{d,q_d}).$$

Hence, we define

$$E^{(0)}(q_1,\ldots,q_d) \coloneqq g(\xi_{1,q_1},\ldots,\xi_{d,q_d})$$

and

$$E^{(1)}(i_d, j_d; q_1, \dots, q_{d-1}) \coloneqq \sum_{q_d=1}^{Q_d} w_{d,q_d} u_{i_d}(\xi_{d,q_d}) v_{j_d}(\xi_{d,q_d}) E^{(0)}(q_1, \dots, q_d),$$

and

$$E^{(2)}(i_{d-1}, j_{d-1}, i_d, j_d; q_1, \dots, q_{d-2})$$
  
$$\coloneqq \sum_{q_{d-1}=1}^{Q_{d-1}} w_{d-1, q_{d-1}} u_{i_{d-1}}(\xi_{d-1, q_{d-1}}) v_{j_{d-1}}(\xi_{d-1, q_{d-1}}) E^{(1)}(i_d, j_d; q_1, \dots, q_{d-1}),$$

and so on and so forth. In this manner we recursively obtain arrays  $E^{(0)}, E^{(1)}, E^{(2)}, \ldots$ , until we arrive at  $E^{(d)} = E$ , which is the desired element array. This is the fast quadrature technique described in [15] and is used in our implementation. If  $I_1 = \cdots = I_d = O(n)$ ,  $J_1 = \cdots = J_d = O(n)$ , and  $Q_1 = \cdots = Q_d = O(n)$  for some parameter n, then the complexity of the fast quadrature is  $O(n^{2d+1})$  compared to  $O(n^{3d})$  for the naive quadrature.

#### 4.5 Choice of the FE subspace in 2D

In what follows, we want to specify the finite-dimensional subspace  $\mathcal{V}$  in the (2+2)-dimensional setting. To this end, we consider a partition of the spatial domain  $\Omega = (0, L)^2$ . First of all,  $\Omega$ , which is the whole core, is composed of cells  $\widehat{\Omega}$ . Each such cell is partitioned into nine elements  $K_x$  as depicted in the following picture.



 $\mathcal{T}_{\Omega}$  shall be the obtained mesh. For each spatial element  $K_{\boldsymbol{x}} \in \mathcal{T}_{\Omega}$ , we denote  $F_{K_{\boldsymbol{x}}}$  for the reference element mapping, i.e.

$$F_{K_{\boldsymbol{x}}} : \widehat{K}_{\boldsymbol{x}} \to K_{\boldsymbol{x}}$$
$$\widehat{\boldsymbol{x}} \mapsto \boldsymbol{x}.$$

The reference element  $\widehat{K}_{\boldsymbol{x}}$  is always the unit square, i.e.  $(0,1)^2$ . We denote  $J_{F_{K_{\boldsymbol{x}}}} = J_{F_{K_{\boldsymbol{x}}}}(\widehat{\boldsymbol{x}})$  for the Jacobian, which in general is not a constant (except for the square element in the middle of  $\widehat{\Omega}$ ). We now discretize the velocity domain. We take V to be a torus, i.e.  $V = \{\boldsymbol{v} \in \mathbb{R}^2 : a \leq |\boldsymbol{v}| \leq b\}$ . We partition V uniformly w.r.t. angle and modulus. In this manner we obtain a mesh  $\mathcal{T}_V$  which looks like a spider net.



We denote  $F_{K_{\boldsymbol{v}}}$  for the reference element mapping from  $\widehat{K}_{\boldsymbol{v}}$  (= (0,1)<sup>2</sup>) to  $K_{\boldsymbol{v}}$ , i.e.

$$F_{K_{\boldsymbol{v}}} : \widehat{K}_{\boldsymbol{v}} \to K_{\boldsymbol{v}}$$
$$\widehat{\boldsymbol{v}} \mapsto \boldsymbol{v}.$$

And as before we denote  $J_{F_{K_v}}$  for the Jacobian of  $F_{K_v}$ . Consider now a tuple  $K_x \times K_v$  of a spatial element and a velocity element. The mapping  $F_{K_x \times K_v}$  defined by

$$F_{K_{\boldsymbol{x}} \times K_{\boldsymbol{v}}} : \widehat{K}_{\boldsymbol{x}} \times \widehat{K}_{\boldsymbol{v}} \to K_{\boldsymbol{x}} \times K_{\boldsymbol{v}}$$
$$(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{v}}) \mapsto (F_{K_{\boldsymbol{x}}}(\widehat{\boldsymbol{x}}), F_{K_{\boldsymbol{v}}}(\widehat{\boldsymbol{v}})),$$

is then a reference element mapping for this tuple. We note that the Jacobian of  $F_{K_{\boldsymbol{x}} \times K_{\boldsymbol{v}}}$ is the product of the Jacobian of  $F_{K_{\boldsymbol{x}}}$  and the Jacobian of  $F_{K_{\boldsymbol{v}}}$ , i.e.  $J_{F_{K_{\boldsymbol{x}} \times K_{\boldsymbol{v}}}} = J_{F_{K_{\boldsymbol{x}}}} J_{F_{K_{\boldsymbol{v}}}}$ . Next, let  $\boldsymbol{p} = (p_1, p_2) \in \mathbb{N}_0^2$  and consider  $\mathcal{P}_{\boldsymbol{p}} = \mathcal{P}_{\boldsymbol{p}}(0, 1)^2$ , the space of polynomials on the unit square of degree at most  $p_1$  in the first variable and of degree at most  $p_2$  in the second variable. A basis  $\mathcal{B}(\mathcal{P}_{\boldsymbol{p}})$  of  $\mathcal{P}_{\boldsymbol{p}}$  is given by

$$\mathcal{B}(\mathcal{P}_{p}) = \{f_1 \cdot f_2 : f_1 \in \mathcal{B}(\mathcal{P}_{p_1}) \text{ and } f_2 \in \mathcal{B}(\mathcal{P}_{p_2})\}.$$

Let now  $\boldsymbol{q} = (q_1, q_2) \in \mathbb{N}_0^2$  be another tuple of polynomial degrees. We introduce the spaces

$$S^{\boldsymbol{p},1}(\Omega,\mathcal{T}_{\Omega}) = \left\{ u \in H^{1}(\Omega) : u|_{K_{\boldsymbol{x}}} \circ F_{K_{\boldsymbol{x}}} \in \mathcal{P}_{\boldsymbol{p}}, \forall K_{\boldsymbol{x}} \in \mathcal{T}_{\Omega} \right\}, S^{\boldsymbol{q},0}(V,\mathcal{T}_{V}) = \left\{ u \in L^{2}(V) : u|_{K_{\boldsymbol{v}}} \circ F_{K_{\boldsymbol{v}}} \in \mathcal{P}_{\boldsymbol{q}}, \forall K_{\boldsymbol{v}} \in \mathcal{T}_{V} \right\},$$

which have the bases

$$\mathcal{B}\left(S^{\boldsymbol{p},1}(\Omega,\mathcal{T}_{\Omega})\right) = \left\{f \circ F_{K_{\boldsymbol{x}}}^{-1} \text{ extended to a hat function } : K_{\boldsymbol{x}} \in \mathcal{T}_{\Omega} \text{ and } f \in \mathcal{B}\left(\mathcal{P}_{\boldsymbol{p}}\right)\right\},\$$
$$\mathcal{B}\left(S^{\boldsymbol{q},0}(V,\mathcal{T}_{V})\right) = \left\{f \circ F_{K_{\boldsymbol{x}}}^{-1} : K_{\boldsymbol{v}} \in \mathcal{T}_{V} \text{ and } f \in \mathcal{B}\left(\mathcal{P}_{\boldsymbol{q}}\right)\right\}.$$

We can finally define the Galerkin subspace

$$\mathcal{V} \coloneqq \operatorname{Span} \left\{ \phi : \phi(\boldsymbol{x}, \boldsymbol{v}) = \phi_{\Omega}(\boldsymbol{x}) \cdot \phi_{V}(\boldsymbol{v}), \text{ where } \phi_{\Omega} \in \mathcal{B}\left(S^{\boldsymbol{p}, 1}(\Omega, \mathcal{T}_{\Omega})\right) \\ \text{and } \phi_{V} \in \mathcal{B}\left(S^{\boldsymbol{q}, 0}(V, \mathcal{T}_{V})\right) \text{ and } \phi|_{\Gamma_{-}} = 0 \right\}.$$

First, we have to say what we mean by "extended to a hat function". To this end, let  $K_{\boldsymbol{x}} \in \mathcal{T}_{\Omega}$  and  $f \in \mathcal{B}(\mathcal{P}_{\boldsymbol{p}})$ . We consider  $f \circ F_{K_{\boldsymbol{x}}}^{-1}$ . We have  $f = f_1 f_2$ , where  $f_1 \in \mathcal{B}(\mathcal{P}_{p_1})$  and  $f_2 \in \mathcal{B}(\mathcal{P}_{p_2})$ . Recall that

$$\mathcal{B}(\mathcal{P}_p) = \{\xi, 1-\xi\} \cup \{\xi(1-\xi)l_k(\xi) : k = 0, \dots, p-2\},\$$

where  $l_k$  is the k-th Legendre polynomial.  $\{\xi, 1-\xi\}$  are external, and  $\{\xi(1-\xi)l_k(\xi) : k = 0, \dots, p-2\}$  are internal. Depending on whether  $f_1$  and  $f_2$  are internal or external, we can now explain how we have to extend  $f \circ F_{K_x}^{-1}$  to a hat function.

**Case 1.**  $(f_1, f_2) = (internal, internal)$ 

We then have that f vanishes on  $\partial \widehat{K}_{\boldsymbol{x}} = \partial(0,1)^2$ , and hence  $f \circ F_{K_{\boldsymbol{x}}}^{-1}$  vanishes on  $\partial K_{\boldsymbol{x}}$ . We can then extend  $f \circ F_{K_{\boldsymbol{x}}}^{-1}$  by zero outside of  $K_{\boldsymbol{x}}$  without creating a discontinuity. The obtained basis function can be called **quad associated**.

#### Case 2. $(f_1, f_2) = (internal, external)$

If  $f_1$  is internal and  $f_2$  is external (or vice versa) there is a unique edge e of  $K_x$  where  $f \circ F_{K_x}^{-1}$  does not vanish. If  $e \in \partial \Omega$ , then we can extend  $f \circ F_{K_x}^{-1}$  by zero outside of  $K_x$ . If  $e \notin \partial \Omega$ , then we must extend  $f \circ F_{K_x}^{-1}$  to the neighbor element of  $K_x$  (i.e. the one that shares the edge e with  $K_x$ ) by choosing the appropriate shape function on it. More precisely, if  $K'_x$  denotes the neighbor element, we extend  $f \circ F_{K_x}^{-1}$  by  $f \circ F_{K_x}^{-1}$ , where we assume that  $F_{K_x}$  is such that  $f \circ F_{K_x}^{-1}$  also does not vanish on e. After all we obtain a basis function which is edge associated.

#### **Case 3.** $(f_1, f_2) = (external, external)$

In this case there is a unique vertex v of  $K_x$  where  $f \circ F_{K_x}^{-1}$  does not vanish. We then have to extend  $f \circ F_{K_x}^{-1}$  to all the elements incident to v by taking the appropriate shape functions on them. Again this means that on another element  $K'_x$  incident to v, we extend by  $f \circ F_{K'_x}^{-1}$ , where  $F_{K'_x}$  is chosen so that  $f \circ F_{K'_x}^{-1}$  does not vanish on v. After this extension we obtain a basis function which is **vertex associated**.

Having now understood the functions of  $\mathcal{B}(S^{p,1}(\Omega, \mathcal{T}_{\Omega}))$ , we ask ourselves how to implement the no-inflow boundary conditions. First, we can associate to each edge on the boundary of  $\Omega$  a number between 0 and 3 depending on which part of the boundary the edge is lying.



We denote  $\operatorname{attr}(e)$  for this number associated to the boundary edge e. Similarly, we can associate to each velocity element  $K_v$  a number between 0 and 3, denoted as  $\operatorname{attr}(K_v)$ .

![](_page_58_Figure_9.jpeg)

For  $g \in \mathcal{B}\left(S^{\boldsymbol{q},0}(V,\mathcal{T}_V)\right)$  we define  $\operatorname{attr}(g) = \operatorname{attr}(K_{\boldsymbol{v}})$ , where  $K_{\boldsymbol{v}}$  is the element on which g is supported. Let now  $\mathcal{N}_{\text{delete}} \coloneqq \{(0,1), (0,2), (1,2), (1,3), (2,0), (2,3), (3,0), (3,1)\}$ . With the help of the set  $\mathcal{N}_{\text{delete}}$  we can now determine which products  $\phi_{\Omega}\phi_V$  do not satisfy the no-inflow boundary conditions and hence have to be deleted. So let  $\phi(\boldsymbol{x}, \boldsymbol{v}) = \phi_{\Omega}(\boldsymbol{x})\phi_V(\boldsymbol{v})$  with  $\phi_{\Omega} \in \mathcal{B}\left(S^{\boldsymbol{p},1}(\Omega,\mathcal{T}_{\Omega})\right)$  and  $\phi_V \in \mathcal{B}\left(S^{\boldsymbol{q},0}(V,\mathcal{T}_V)\right)$ .

**Case 1.**  $\phi_{\Omega}$  is quad associated.

We then have that  $\phi_{\Omega}$  vanishes on  $\partial\Omega$  and hence  $\phi = \phi_{\Omega}\phi_V$  automatically satisfies the no-inflow boundary conditions. Thus, this  $\phi$  will not be deleted.

#### **Case 2.** $\phi_{\Omega}$ is edge associated.

Let e be the corresponding edge. If  $e \notin \partial \Omega$ , then  $\phi_{\Omega}$  vanishes on  $\partial \Omega$ , and thus  $\phi$  will not be deleted. If  $e \in \partial \Omega$ , then we have to consider attr(e) and attr( $\phi_V$ ). If  $(attr(e), attr(\phi_V)) \in \mathcal{N}_{delete}$ , then  $\phi$  does not satisfy the no-inflow boundary conditions and has to be deleted.

#### **Case 3.** $\phi_{\Omega}$ is vertex associated.

Let v be this vertex. If v is in the interior of  $\Omega$ , then  $\phi$  automatically satisfies the no-inflow boundary conditions and will thus not be deleted. So we now assume that v is on the boundary of  $\Omega$ . If v is not a corner of  $\Omega$ , then there are two edges  $e_1$  and  $e_2$  on  $\partial\Omega$  on which  $\phi_{\Omega}$  does not vanish, and these edges have the same attribute, i.e.  $attr(e_1) = attr(e_2)$ . If  $(attr(e_1), attr(\phi_V)) \in \mathcal{N}_{delete}$ , then  $\phi$  has to be deleted. If v happens to be one of the four corners of  $\Omega$ , then there are still two edges  $e_1$  and  $e_2$ on  $\partial\Omega$  on which  $\phi_{\Omega}$  does not vanish. But this time, these edges do not have the same attribute. If now  $(attr(e_1), attr(\phi_V)) \in \mathcal{N}_{delete}$  or  $(attr(e_2), attr(\phi_V)) \in \mathcal{N}_{delete}$ , then  $\phi$ has to be deleted.

Now that we have described how to implement the no-inflow boundary conditions, we consider the assembly of the Galerkin matrices for the operators L - K and F. In this respect, we want to show how the fast quadrature (see Section 4.4) comes in. We will show this for the absorption and the scattering part of the bilinear form a. So let  $\phi$  and  $\psi$  be two basis functions of  $\mathcal{V}$ . We consider the integral

$$\int_{\Omega \times V} \Sigma(\boldsymbol{x}, \boldsymbol{v}) \phi(\boldsymbol{x}, \boldsymbol{v}) \, \mathrm{d} \boldsymbol{x} \, \mathrm{d} \boldsymbol{v}.$$

Splitting this integral into a sum of integrals over the elements  $K_{\boldsymbol{x}} \times K_{\boldsymbol{v}}$ , we obtain

$$\sum_{K_{\boldsymbol{x}} \times K_{\boldsymbol{v}}} \int_{K_{\boldsymbol{x}} \times K_{\boldsymbol{v}}} \Sigma(\boldsymbol{x}, \boldsymbol{v}) \phi(\boldsymbol{x}, \boldsymbol{v}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v}.$$

Hence, we are left with

$$\int_{K_{\boldsymbol{x}}\times K_{\boldsymbol{v}}} \Sigma(\boldsymbol{x},\boldsymbol{v})\phi(\boldsymbol{x},\boldsymbol{v})\psi(\boldsymbol{x},\boldsymbol{v}) \,\mathrm{d}\boldsymbol{x} \,\mathrm{d}\boldsymbol{v}.$$

But since  $\phi$  and  $\psi$  are basis functions of  $\mathcal{V}$ , they split on  $K_{\boldsymbol{x}} \times K_{\boldsymbol{v}}$ , i.e.  $\phi(\boldsymbol{x}, \boldsymbol{v}) = \phi_{K_{\boldsymbol{x}}}(\boldsymbol{x})\phi_{K_{\boldsymbol{v}}}(\boldsymbol{v})$  and  $\psi(\boldsymbol{x}, \boldsymbol{v}) = \psi_{K_{\boldsymbol{x}}}(\boldsymbol{x})\psi_{K_{\boldsymbol{v}}}(\boldsymbol{v})$  on  $K_{\boldsymbol{x}} \times K_{\boldsymbol{v}}$  (note that they actually split on all of  $\Omega \times V$ ). Thus, the above integral becomes

$$\int_{K_{\boldsymbol{x}}\times K_{\boldsymbol{v}}} \Sigma(\boldsymbol{x},\boldsymbol{v}) \phi_{K_{\boldsymbol{x}}}(\boldsymbol{x}) \phi_{K_{\boldsymbol{v}}}(\boldsymbol{v}) \psi_{K_{\boldsymbol{x}}}(\boldsymbol{x}) \psi_{K_{\boldsymbol{v}}}(\boldsymbol{v}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{v}.$$

But now comes a crucial point. We have that  $\Sigma$  only depends on  $\boldsymbol{v}$  on a given spatial element  $K_{\boldsymbol{x}}$ . This is due to the fact that the elements  $K_{\boldsymbol{x}}$  were chosen so that within each  $K_{\boldsymbol{x}}$  the material is homogeneous, i.e. either the fissile material or the moderator. Hence, on  $K_{\boldsymbol{x}}$ , we can write  $\Sigma(\boldsymbol{x}, \boldsymbol{v}) = \Sigma_{K_{\boldsymbol{x}}}(\boldsymbol{v})$ . But then the above integral splits as

$$\int_{K_{\boldsymbol{x}}} \phi_{K_{\boldsymbol{x}}}(\boldsymbol{x}) \psi_{K_{\boldsymbol{x}}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \int_{K_{\boldsymbol{v}}} \Sigma_{K_{\boldsymbol{x}}}(\boldsymbol{v}) \phi_{K_{\boldsymbol{v}}}(\boldsymbol{v}) \psi_{K_{\boldsymbol{v}}}(\boldsymbol{v}) \, \mathrm{d}\boldsymbol{v}.$$

Again we shall emphasize that this splitting only occurs if the mesh is chosen so that in each element  $K_x$  the material is homogeneous.<sup>7</sup> We now focus on the integral over  $K_v$ . Going back to the reference element, we can write this integral as

$$\int_{\widehat{K}_{\boldsymbol{v}}}\widehat{\Sigma}_{K_{\boldsymbol{v}}}(\widehat{\boldsymbol{v}})\phi_{\widehat{K}_{\boldsymbol{v}}}(\widehat{\boldsymbol{v}})\psi_{\widehat{K}_{\boldsymbol{v}}}(\widehat{\boldsymbol{v}})J_{F_{K_{\boldsymbol{v}}}}(\widehat{\boldsymbol{v}})\,\mathrm{d}\widehat{\boldsymbol{v}},$$

where  $\widehat{\Sigma}_{K_{\boldsymbol{x}}} = \Sigma_{K_{\boldsymbol{x}}} \circ F_{K_{\boldsymbol{v}}}, \ \phi_{\widehat{K}_{\boldsymbol{v}}} = \phi_{K_{\boldsymbol{v}}} \circ F_{K_{\boldsymbol{v}}}, \ \text{and} \ \psi_{\widehat{K}_{\boldsymbol{v}}} = \psi_{K_{\boldsymbol{v}}} \circ F_{K_{\boldsymbol{v}}}.$  We use now the variable  $\boldsymbol{\xi}$  instead of  $\widehat{\boldsymbol{v}}$  to make the connection to the fast quadrature more visible. Since  $\phi_{\widehat{K}_{\boldsymbol{v}}} \in \mathcal{B}(\mathcal{P}_{\boldsymbol{q}}), \ \text{we have} \ \phi_{\widehat{K}_{\boldsymbol{v}}}(\boldsymbol{\xi}) = u_{i_1}(\xi_1)u_{i_2}(\xi_2), \ \text{where} \ u_{i_1} \in \mathcal{B}(\mathcal{P}_{q_1}), \ i_1 = 1, \ldots, q_1 + 1, \ \text{and} \ u_{i_2} \in \mathcal{B}(\mathcal{P}_{q_2}), \ i_2 = 1, \ldots, q_2 + 1 \ \text{(we use an enumeration of the polynomials in } \mathcal{B}(\mathcal{P}_{q_1}), \ \text{respectively} \ \mathcal{B}(\mathcal{P}_{q_2})). \ \text{Similarly, since} \ \psi_{\widehat{K}_{\boldsymbol{v}}} \in \mathcal{B}(\mathcal{P}_{\boldsymbol{q}}), \ \text{we have} \ \psi_{\widehat{K}_{\boldsymbol{v}}}(\boldsymbol{\xi}) = v_{j_1}(\xi_1)v_{j_2}(\xi_2), \ \text{where} \ v_{j_1} \in \mathcal{B}(\mathcal{P}_{q_1}), \ j_1 = 1, \ldots, q_1 + 1, \ \text{and} \ v_{j_2} \in \mathcal{B}(\mathcal{P}_{q_2}), \ j_2 = 1, \ldots, q_2 + 1. \ \text{So denoting} \ g(\boldsymbol{\xi}) = \widehat{\Sigma}_{K_{\boldsymbol{x}}}(\boldsymbol{\xi}) J_{F_{K_{\boldsymbol{v}}}}(\boldsymbol{\xi}), \ \text{we obtain}$ 

$$\int_{(0,1)^2} g(\xi_1,\xi_2) u_{i_1}(\xi_1) u_{i_2}(\xi_2) v_{j_1}(\xi_1) v_{j_2}(\xi_2) \,\mathrm{d}\boldsymbol{\xi}$$

But this leads to a fast quadrature on  $(0,1)^2$ ! Let us now do a similar consideration for the scattering part of the bilinear form a,

$$\int_{\Omega \times V} \int_V f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \phi(\boldsymbol{x}, \boldsymbol{v}') \, \mathrm{d} \boldsymbol{v}' \psi(\boldsymbol{x}, \boldsymbol{v}) \, \mathrm{d} \boldsymbol{x} \, \mathrm{d} \boldsymbol{v},$$

where again  $\phi$  and  $\psi$  are basis functions of  $\mathcal{V}$ . Again we start with decomposing the integral over the elements. We obtain

<sup>&</sup>lt;sup>7</sup> The splitting which results in less computational effort is not the only advantage of such a mesh. The second one is that we can capture the points where the eigenfunction is not differentiable w.r.t.  $\boldsymbol{x}$ .

$$\sum_{K_{\boldsymbol{x}} \times K_{\boldsymbol{v}}} \sum_{K_{\boldsymbol{v}}'} \int_{K_{\boldsymbol{x}} \times K_{\boldsymbol{v}}} \int_{K_{\boldsymbol{v}}'} f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \phi(\boldsymbol{x}, \boldsymbol{v}') \, \mathrm{d} \boldsymbol{v}' \psi(\boldsymbol{x}, \boldsymbol{v}) \, \mathrm{d} \boldsymbol{x} \, \mathrm{d} \boldsymbol{v}.$$

Thus, it remains to compute the term

$$\int_{K_{\boldsymbol{x}}\times K_{\boldsymbol{v}}}\int_{K_{\boldsymbol{v}}'}f(\boldsymbol{x},\boldsymbol{v}',\boldsymbol{v})\phi(\boldsymbol{x},\boldsymbol{v}')\,\mathrm{d}\boldsymbol{v}'\psi(\boldsymbol{x},\boldsymbol{v})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{v}.$$

But again we have that on  $K_x$ , the data f does not depend on x anymore, i.e.  $f(x, v', v) = f_{K_x}(v', v)$  on  $K_x$ . Moreover, we have that  $\phi$  and  $\psi$  split. We therefore obtain

$$\int_{K_{\boldsymbol{x}}} \phi_{K_{\boldsymbol{x}}}(\boldsymbol{x}) \psi_{K_{\boldsymbol{x}}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \int_{K_{\boldsymbol{v}}' \times K_{\boldsymbol{v}}} f_{K_{\boldsymbol{x}}}(\boldsymbol{v}', \boldsymbol{v}) \phi_{K_{\boldsymbol{v}}'}(\boldsymbol{v}') \psi_{K_{\boldsymbol{v}}}(\boldsymbol{v}) \, \mathrm{d}\boldsymbol{v}' \, \mathrm{d}\boldsymbol{v}.$$

Now the first term leads to a fast quadrature on  $(0,1)^2$  (after transformation). But the second term is slightly different. We now focus on it. We pull this integral back to the reference element to obtain

$$\int_{\widehat{K}'_{\boldsymbol{v}}\times\widehat{K}_{\boldsymbol{v}}}\widehat{f}_{K_{\boldsymbol{x}}}(\widehat{\boldsymbol{v}}',\widehat{\boldsymbol{v}})\phi_{\widehat{K}'_{\boldsymbol{v}}}(\widehat{\boldsymbol{v}}')\psi_{\widehat{K}_{\boldsymbol{v}}}(\widehat{\boldsymbol{v}})J_{F_{K'_{\boldsymbol{v}}}}(\widehat{\boldsymbol{v}}')J_{F_{K_{\boldsymbol{v}}}}(\widehat{\boldsymbol{v}})\,\mathrm{d}\widehat{\boldsymbol{v}}'\,\mathrm{d}\widehat{\boldsymbol{v}}$$

But note that both  $\phi_{\widehat{K}'_{v}} = \phi_{K'_{v}} \circ F_{K'_{v}}$  and  $\psi_{\widehat{K}_{v}} = \psi_{K_{v}} \circ F_{K_{v}}$  belong to  $\mathcal{B}(\mathcal{P}_{q})$ . So using the notation  $\boldsymbol{\xi} = (\widehat{\boldsymbol{v}}', \widehat{\boldsymbol{v}})$ , we have  $\phi_{\widehat{K}'_{v}}(\xi_{1}, \xi_{2}) = u_{i_{1}}(\xi_{1})u_{i_{2}}(\xi_{2})$  with  $u_{i_{1}} \in \mathcal{B}(\mathcal{P}_{q_{1}}), i_{1} =$  $1, \ldots, q_{1} + 1$ , and  $u_{i_{2}} \in \mathcal{B}(\mathcal{P}_{q_{2}}), i_{2} = 1, \ldots, q_{2} + 1$ . Similarly,  $\psi_{\widehat{K}_{v}}(\xi_{3}, \xi_{4}) = v_{j_{3}}(\xi_{3})v_{j_{4}}(\xi_{4})$ with  $v_{j_{3}} \in \mathcal{B}(\mathcal{P}_{q_{1}}), j_{3} = 1, \ldots, q_{1} + 1$ , and  $v_{j_{4}} \in \mathcal{B}(\mathcal{P}_{q_{2}}), j_{4} = 1, \ldots, q_{2} + 1$ . Thus, if we denote  $g(\boldsymbol{\xi}) = \widehat{f}_{K_{x}}(\boldsymbol{\xi})J_{F_{K'_{v}}}(\xi_{1}, \xi_{2})J_{F_{K_{v}}}(\xi_{3}, \xi_{4})$ , we obtain

$$\int_{(0,1)^4} g(\xi_1,\xi_2,\xi_3,\xi_4) u_{i_1}(\xi_1) u_{i_2}(\xi_2) v_{j_3}(\xi_3) v_{j_4}(\xi_4) \,\mathrm{d}\boldsymbol{\xi}.$$

So we see that this is almost a fast quadrature on  $(0,1)^4$ . But if we introduce "dummy" functions  $u_{i_3}, u_{i_4}, v_{j_1}, v_{j_2}$ , which are all constant one functions (and their indices range from 1 to 1), then we can write this integral as

$$\int_{(0,1)^4} g(\xi_1,\xi_2,\xi_3,\xi_4) u_{i_1}(\xi_1) u_{i_2}(\xi_2) u_{i_3}(\xi_3) u_{i_4}(\xi_4) v_{j_1}(\xi_1) v_{j_2}(\xi_2) v_{j_3}(\xi_3) v_{j_4}(\xi_4) \,\mathrm{d}\boldsymbol{\xi}$$

But this leads to a fast quadrature on  $(0, 1)^4$ !

## **5** Generalized FEM

## 5.1 Homogenization and the Unit Cell Problem

We consider a reactor  $\Omega_N = (0, N)^2$  where in each direction we have N cells, so  $N^2$  cells in total. The cell is  $\widehat{\Omega} = (0, 1)^2$ . We want to understand the behavior of the eigenfunction as  $N \to \infty$ . Note that we keep the size of the cells fixed, but increase their number, so that the size of the whole reactor increases.

![](_page_62_Figure_3.jpeg)

As described in [4, Part IV, Sec. 1.1.1], we are thus facing a sequence of criticality problems indexed by N

$$\boldsymbol{v} \cdot \nabla \phi + \Sigma(\boldsymbol{x}, \boldsymbol{v}) \phi = \int_{V} f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \phi(\boldsymbol{x}, \boldsymbol{v}') \, \mathrm{d}\boldsymbol{v}' + \lambda \int_{V} \sigma(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \phi(\boldsymbol{x}, \boldsymbol{v}') \, \mathrm{d}\boldsymbol{v}' \quad \text{in } \Omega_{N} \times V_{N} \\ \phi = 0 \qquad \qquad \text{on } \Gamma_{-},$$

where the data  $\Sigma$ ,  $f, \sigma$  are  $\widehat{\Omega}$ -periodic. We know that for each N there exists the positive eigenpair which we denote from now on as  $(\lambda_N, \phi_N)$ . We now introduce the unit cell problem, which is basically the same eigenvalue problem but posed on the unit cell  $\widehat{\Omega} = (0, 1)^2$  and with periodic boundary conditions (instead of no-inflow). Thus, we consider

$$\boldsymbol{v} \cdot \nabla \phi + \Sigma(\boldsymbol{x}, \boldsymbol{v}) \phi = \int_{V} f(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \phi(\boldsymbol{x}, \boldsymbol{v}') \, \mathrm{d}\boldsymbol{v}' + \lambda \int_{V} \sigma(\boldsymbol{x}, \boldsymbol{v}', \boldsymbol{v}) \phi(\boldsymbol{x}, \boldsymbol{v}') \, \mathrm{d}\boldsymbol{v}' \quad \text{in } \widehat{\Omega} \times V,$$
$$\boldsymbol{x} \mapsto \phi(\boldsymbol{x}, \boldsymbol{v}) \ \widehat{\Omega}\text{-periodic.}$$

As for the eigenvalue problem posed on the whole core, one can show that the unit cell problem has an eigenvalue which is real positive, of smallest modulus, with multiplicity one, and such that its associated eigenfunction is positive (see [4, Part II, Sec. 2.2.1, Thm. 2.2.4]). Let us denote  $(\lambda_{uc}, \phi_{uc})$  for this eigenpair. By extending  $\phi_{uc}$  periodically, we can see it as a function defined on all of  $\Omega \times V$ . The whole point now is that for large N we have that (see [4, Part IV, Sec. 1.1.2])

$$\phi_N(\boldsymbol{x}, \boldsymbol{v}) \approx u(\boldsymbol{x}) \phi_{\rm uc}(\boldsymbol{x}, \boldsymbol{v}), \tag{5.1}$$

where  $u(\boldsymbol{x})$  is the first eigenfunction of a homogenized diffusion problem posed on the whole core  $\Omega$ . What can we learn from expression (5.1)? Well, we see that for large N, the eigenfunction  $\phi_N$  approximately decomposes into a product of a function just depending on  $\boldsymbol{x}$  (and modeling the macroscopic behavior of  $\phi_N$ ) and the solution of the unit cell problem (which models the microscopic behavior of  $\phi_N$ ). This observation will lead us to the construction of a suitable finite element subspace as described in the next section.

## 5.2 g-FEM

We now introduce the idea of g-FEM, at least for our particular problem. To sum up, the generalized finite element method does not use the standard basis functions which are composed of polynomial shape functions. So let us start with the unit cell problem posed on  $\hat{\Omega} \times V$ , where  $\hat{\Omega} = (0, 1)^2$ . We have the usual meshes on  $\hat{\Omega}$  and V as depicted in the figures.

![](_page_63_Figure_6.jpeg)

We can use *p*-FEM with a high polynomial degree to solve the unit cell problem with these two meshes. The high polynomial degree is not a problem since we are just on one cell and thus there will not be too many degrees of freedom. Note that this is not the case for *p*-FEM on the whole reactor, where we have to use a high polynomial degree on each cell. Assume now that we have computed an approximation  $\phi_{uc}^{approx}$  to the eigenfunction  $\phi_{uc}$  of the unit cell problem. Note that by extending  $\phi_{uc}^{approx}$  periodically, we can see it as a function defined on all of  $\Omega \times V$ . We now consider the whole reactor again together with the velocity domain. But this time we group the cells of the reactor to obtain a macro mesh  $\mathcal{T}_{\Omega}^{\text{macro}}$ . The macro mesh  $\mathcal{T}_{V}^{\text{macro}}$  for the velocity domain will consist of only four elements. The following pictures illustrate these two macro meshes where we chose 16 cells on  $\Omega$ .

![](_page_64_Figure_1.jpeg)

In order to treat our actual eigenvalue problem, which is posed on the whole reactor, we will choose a Galerkin subspace  $\mathcal{V}$  different from the one used so far. Since we know that the global eigenfunction has a microscopic structure (given by  $\phi_{uc}$ , respectively  $\phi_{uc}^{approx}$ ) and a macroscopic one, it makes sense to choose a  $\mathcal{V}$  which has exactly this kind of structure. Note that this structure is only valid if the number of cells of the reactor is large (which is not the case in the above illustration with just 16 cells). As in Section 4.5, we introduce

$$S^{\boldsymbol{p},1}\left(\Omega,\mathcal{T}_{\Omega}^{\mathrm{macro}}\right) = \left\{ u \in H^{1}(\Omega) : u|_{K_{\boldsymbol{x}}} \circ F_{K_{\boldsymbol{x}}} \in \mathcal{P}_{\boldsymbol{p}}, \forall K_{\boldsymbol{x}} \in \mathcal{T}_{\Omega}^{\mathrm{macro}} \right\}, \\ S^{\boldsymbol{q},0}\left(V,\mathcal{T}_{V}^{\mathrm{macro}}\right) = \left\{ u \in L^{2}(V) : u|_{K_{\boldsymbol{x}}} \circ F_{K_{\boldsymbol{x}}} \in \mathcal{P}_{\boldsymbol{q}}, \forall K_{\boldsymbol{v}} \in \mathcal{T}_{V}^{\mathrm{macro}} \right\},$$

where the bases for the two spaces are

$$\mathcal{B}\left(S^{\boldsymbol{p},1}(\Omega,\mathcal{T}_{\Omega}^{\mathrm{macro}})\right) = \left\{f \circ F_{K_{\boldsymbol{x}}}^{-1} \text{ extended to a hat function } : K_{\boldsymbol{x}} \in \mathcal{T}_{\Omega}^{\mathrm{macro}} \text{ and } f \in \mathcal{B}\left(\mathcal{P}_{\boldsymbol{p}}\right)\right\},\$$
$$\mathcal{B}\left(S^{\boldsymbol{q},0}(V,\mathcal{T}_{V}^{\mathrm{macro}})\right) = \left\{f \circ F_{K_{\boldsymbol{v}}}^{-1} : K_{\boldsymbol{v}} \in \mathcal{T}_{V}^{\mathrm{macro}} \text{ and } f \in \mathcal{B}\left(\mathcal{P}_{\boldsymbol{q}}\right)\right\}.$$

But this time we consider different polynomial degrees on each  $K_x$ . Namely in the following way. The further away  $K_x$  is from the center of the reactor w.r.t. the coordinate

 $x_1$ , the higher we want the polynomial degree on that element w.r.t.  $x_1$ . Analogously for  $x_2$ . As we will see in a moment, the idea behind this is that in the interior of the reactor the global eigenfunction is basically given by the unit cell solution, whereas towards the boundary we multiply the unit cell solution with macroscopic polynomials to model the behavior of the global eigenfunction close to  $\partial \Omega$ . Consider now

$$\mathcal{M} \coloneqq \left\{ u : u(\boldsymbol{x}, \boldsymbol{v}) = u_{\Omega}(\boldsymbol{x}) \cdot u_{V}(\boldsymbol{v}), \text{ where } u_{\Omega} \in \mathcal{B}\left(S^{\boldsymbol{p},1}(\Omega, \mathcal{T}_{\Omega}^{\text{macro}})\right) \\ \text{and } u_{V} \in \mathcal{B}\left(S^{\boldsymbol{q},0}(V, \mathcal{T}_{V}^{\text{macro}})\right) \text{ and } u|_{\Gamma_{-}} = 0 \right\}.$$

This set is not new, since in Section 4.5 we have defined  $\mathcal{V}$  to be the span of it (but for the finer meshes  $\mathcal{T}_{\Omega}$  and  $\mathcal{T}_{V}$ ). But here comes now the difference: We take  $\mathcal{V}^{\text{g-FEM}}$ to be the span of functions which are products of a function in  $\mathcal{M}$  and the unit cell eigenfunction  $\phi_{\text{uc}}^{\text{approx}}$ . We therefore set

$$\mathcal{V}^{ ext{g-FEM}} \coloneqq ext{ Span} \left\{ \phi \, : \, \phi(\boldsymbol{x}, \boldsymbol{v}) = u(\boldsymbol{x}, \boldsymbol{v}) \cdot \phi_{ ext{uc}}^{ ext{approx}}(\boldsymbol{x}, \boldsymbol{v}), ext{ where } u \in \mathcal{M} 
ight\}.$$

We see that this is not a standard finite element subspace. The space  $\mathcal{V}^{\text{g-FEM}}$  depends on the unit cell eigenfunction, which itself results from a finite element method. This is the idea: First solve the unit cell problem and then compute the global eigenfunction using a second FEM with  $\mathcal{V}^{\text{g-FEM}}$  as the trial (and test) space.

#### Remarks 5.2.1.

- The macroscopic functions  $u(\boldsymbol{x}, \boldsymbol{v}) \in \mathcal{M}$  depend on  $\boldsymbol{x}$  and on  $\boldsymbol{v}$ , in contrast to (5.1), where u just depends on  $\boldsymbol{x}$ . But this is necessary to achieve that  $\phi(\boldsymbol{x}, \boldsymbol{v}) = u(\boldsymbol{x}, \boldsymbol{v}) \cdot \phi_{uc}^{approx}(\boldsymbol{x}, \boldsymbol{v})$  satisfies no-inflow boundary conditions, but at the same time can have non-zero outflow.
- We should have that the functions spanning  $\mathcal{V}^{g\text{-}FEM}$  are linearly independent. But the functions in  $\mathcal{B}\left(S^{\mathbf{p},1}(\Omega,\mathcal{T}_{\Omega}^{macro})\right)$  are, and the functions in  $\mathcal{B}\left(S^{\mathbf{q},0}(V,\mathcal{T}_{V}^{macro})\right)$ , too, and so the ones in  $\mathcal{M}$ . Thus, it remains to show that the products of functions in  $\mathcal{M}$  with  $\phi_{uc}^{approx}$  are again linearly independent. But since  $\phi_{uc}^{approx} > 0$ (hopefully!), this is the case.
- We have to ask ourselves a general question here. To determine whether the reactor is super- or subcritical, we actually just need the eigenvalue. So why are we putting so much effort into the approximation of the eigenfunction? First, the eigenfunction also contains relevant physical information. Second, in practice, the eigenvalue is computed together with the eigenfunction (→ power method). This also indicates that if we want to have a good approximation to the eigenvalue, we better make sure to have a good approximation to the eigenfunction, since the error for the eigenvalue and the error for the eigenfunction are usually closely related. But we note that since an abstract framework for the FEM could not be found, error estimates for the eigenpair are not available.

# 6 Numerical Examples & Outlook

## 6.1 1+1D Global

We consider the (1 + 1)-dimensional case with the following data

- $\Omega = (0, 10)$ , i.e. 10 cells
- $V = (-1, -\frac{1}{2}) \cup (\frac{1}{2}, 1)$

$$f(x, v', v) = \begin{cases} f_{\mathcal{F}}(v', v) & \text{if } x \in \mathcal{F}, \\ f_{\mathcal{M}}(v', v) & \text{if } x \in \mathcal{M}. \end{cases}$$

• 
$$f_{\rm F}(v',v) = \overline{f}_{\rm F} (v')^2 e^{-(v'-v)^2}$$
 with  $\overline{f}_{\rm F} = 0.25$ 

•  $f_{\rm M}(v',v) = \overline{f}_{\rm M} (v')^2 e^{-(v'-v)^2}$  with  $\overline{f}_{\rm M} = 0.5$ 

•

•

$$\sigma(x, v', v) = \begin{cases} \sigma_{\mathrm{F}}(v', v) & \text{if } x \in \mathrm{F}, \\ \sigma_{\mathrm{M}}(v', v) & \text{if } x \in \mathrm{M}. \end{cases}$$

$$\Sigma(x,v) = \begin{cases} \Sigma_{\rm F}(v) & \text{if } x \in {\rm F}, \\ \Sigma_{\rm M}(v) & \text{if } x \in {\rm M}. \end{cases}$$

• 
$$\Sigma_{\rm F}(v) = \frac{1}{\eta}\overline{f}_{\rm F} \left(2v^2 + b(v)\right)$$
  
•  $\Sigma_{\rm M}(v) = \frac{1}{\eta}\overline{f}_{\rm M} \left(2v^2 + b(v)\right)$   
•  $\eta = 0.5$ 

$$b(v) = 0.25 \left( 4 + 2e^{-(v+1)^2} (v-1) - 2e^{-(v+0.5)^2} (v-0.5) + 8v^2 + 2e^{-(v-0.5)^2} (v+0.5) - 2e^{-(v-1)^2} (v+1) \right)$$

The reason why  $\Sigma$  is chosen like that is to ensure that Assumption 3.1.5 holds, i.e. to ensure that the operator L - K is invertible. For the finite element discretization (see Section 4.3) we choose

- the spatial mesh that coincides with the material discontinuities, i.e. we have  $3 \times 10$  elements, since there are 10 cells and each consists of 3 elements.
- the velocity mesh with just two elements, i.e.  $(-1, -\frac{1}{2})$  and  $(\frac{1}{2}, 1)$ .
- the polynomial degree in x to be p = 6, and the polynomial degree in v to be q = 6.
- continuous elements on V, although we just need continuous elements in the spatial variable (see Remark 4.3.2).

For the eigenvalue we obtain k = 0.666007, and the eigenfunction  $\phi$  with  $\|\phi\|_{L^2(\Omega \times V)} = 1$ is given in Figure 6.1. We observe that  $\phi$  is symmetric in the sense that  $\phi(2m - x, -v) = \phi(x, v)$ , where m is the center of the reactor (here m = 5). This is because the material parameters have exactly this symmetry. Indeed, one can check that  $\tilde{\phi}(x, v) := \phi(2m - x, -v)$  satisfies the same eigenvalue equation. But since the first eigenvalue is simple (see Theorem 3.2.14 (iii)), this implies that  $\tilde{\phi}$  and  $\phi$  must be identical up to a scaling. The scaling factor is either +1 or -1. If it was -1, then  $\phi$  would not be of constant sign. But since we know it is (see Theorem 3.2.14 (ii)), the scaling factor must be +1, and the observed symmetry follows.

![](_page_67_Figure_6.jpeg)

Figure 6.1: The normalized eigenfunction of the whole reactor

We let  $(k^{p,q}, \phi^{p,q})$  be the eigenpair computed with the polynomial degrees p and q. We take  $(k^{6,6}, \phi^{6,6})$  as our reference solution  $(k^{\text{ref}}, \phi^{\text{ref}})$  and compare it to  $(k^{r,r}, \phi^{r,r})$  for  $r = 1, 2, \ldots, 5$ . Figure 6.2 shows  $|k^{\text{ref}} - k^{r,r}|$  for  $r = 1, 2, \ldots, 5$ . Figure 6.3 shows  $\|\phi^{\text{ref}} - \phi^{r,r}\|_{L^2(\Omega \times V)}$  for  $r = 1, 2, \ldots, 5$ .

![](_page_68_Figure_0.jpeg)

Figure 6.2: The eigenvalue error  $|k^{\text{ref}} - k^{r,r}|$  for  $r = 1, 2, \dots, 5$ 

![](_page_68_Figure_2.jpeg)

Figure 6.3: The eigenfunction error  $\|\phi^{\text{ref}} - \phi^{r,r}\|_{L^2(\Omega \times V)}$  for  $r = 1, 2, \dots, 5$ 

## 6.2 1+1D Unit Cell

We now consider the unit cell problem with the same material data as before. The spatial domain is now one cell, i.e. (0, 1). The velocity domain remains the same. Instead of noinflow boundary conditions we have periodic boundary conditions, i.e.  $\phi(0, v) = \phi(1, v)$  for all  $v \in V$ , so that the FE-subspace  $\mathcal{V}$  described in Section 4.3 has to be modified accordingly. For the discretization we choose

- the spatial mesh consisting of the three elements  $(0, \frac{1}{3}), (\frac{1}{3}, \frac{2}{3}), (\frac{2}{3}, 1)$ .
- again the velocity mesh consisting of the two elements  $(-1, -\frac{1}{2})$  and  $(\frac{1}{2}, 1)$ .
- the polynomial degree in x to be p = 10, and the one in v to be q = 10.
- again continuous elements on V, although we do not have to.

For the eigenvalue of the unit cell problem we obtain  $k_{\rm uc} = 0.670647$ . The corresponding eigenfunction  $\phi_{\rm uc}$  with  $\|\phi_{\rm uc}\|_{L^2(\widehat{\Omega} \times V)} = 1$  is given in Figure 6.4.

![](_page_69_Figure_1.jpeg)

Figure 6.4: The normalized eigenfunction of the unit cell

We want to plot the error of the eigenpair depending on the polynomial degree. For polynomial degrees p and q, let  $(k_{\rm uc}^{p,q}, \phi_{\rm uc}^{p,q})$  be the eigenpair computed with them. As our reference solution we take  $(k_{\rm uc}^{\rm ref}, \phi_{\rm uc}^{\rm ref}) = (k_{\rm uc}^{10,10}, \phi_{\rm uc}^{10,10})$ . Figure 6.5 shows  $|k_{\rm uc}^{\rm ref} - k_{\rm uc}^{r,r}|$  for  $r = 1, 2, \ldots, 9$ . Figure 6.6 shows  $\|\phi_{\rm uc}^{\rm ref} - \phi_{\rm uc}^{r,r}\|_{L^2(\widehat{\Omega} \times V)}$  for  $r = 1, 2, \ldots, 9$ .

![](_page_69_Figure_4.jpeg)

Figure 6.5: The eigenvalue error  $|k_{\rm uc}^{\rm ref} - k_{\rm uc}^{r,r}|$  for  $r = 1, 2, \ldots, 9$ 

Figure 6.6 clearly shows the exponential convergence rate. This is due to the fact that the data functions were chosen nicely. In the velocity variable, they are smooth. In the spatial variable they are not smooth globally, but on each of the three elements, which are exactly the elements of the spatial mesh.<sup>8</sup>

<sup>&</sup>lt;sup>8</sup>Recall footnote 7.

![](_page_70_Figure_0.jpeg)

Figure 6.6: The eigenfunction error  $\|\phi_{uc}^{ref} - \phi_{uc}^{r,r}\|_{L^2(\widehat{\Omega} \times V)}$  for r = 1, 2, ..., 9

## 6.3 Outlook

Several important problems are not vet solved. First of all, an abstract framework for the variational formulation still needs to be found. Second, it is not clear how to choose the discrete trial and test spaces optimally. Since in the continuous variational formulation the trial space is strictly contained in the test space, it might be advisable to choose discrete subspaces which also have this property. Then one might obtain a good pair of trial and test spaces in terms of an inf-sup condition, but the resulting matrix eigenvalue problem will be non-square. Concerning the generalized FEM, the next step would be the  $L^2$ -best approximation in 1 + 1D with which one can find out how to choose the macro meshes. Besides, if the abstract framework for the variational formulation has been found, it should be possible to do a thorough error analysis of g-FEM in 1 + 1D. Once we are familiar with g-FEM in 1 + 1D, we can proceed to the 2 + 2D setting. However, it has to be pointed out, that one first needs to understand the standard FEM for the criticality problem. From a computational point of view, we need a more powerful eigenvalue solver than the one available in Python. For this it shall be noted, that the matrices A and B are not completely sparse, because of the scattering respectively the fission part of the bilinear forms a and b. They are sparse w.r.t. the spatial variable, but not w.r.t. the velocity variable. Also, interesting from a theoretical and a practical point of view, is the case where the operator L-K is singular or close to. Because so far, we assumed that absorption is strong enough compared to scattering (recall the important Assumption 3.1.5). But it might be very well the case that there are regions in the reactor where this condition is not satisfied. The theoretical problem will then be to relax the conditions needed to prove the existence of the positive eigenpair. And last but not least, we have to be honest and go back to the very beginning of the thesis, where we introduced the criticality spectral problem. This formulation of the criticality problem is the so-called k-formulation. However, there is also another formulation of the criticality problem, called the  $\alpha$ -formulation (see [7,

Ch. XXI, §3, Sec. 1]). The connection between these two formulations is not obvious. The  $\alpha$ -formulation can be justified more rigorously, whereas the k-formulation seems to be better suitable for computations. Hence, it would be interesting to study the connection between these two formulations both theoretically, and numerically.
## **Bibliography**

- G. Allaire and G. Bal. Homogenization of the criticality spectral equation in neutron transport. ESAIM: Mathematical Modelling and Numerical Analysis, 33(04):721– 746, 1999.
- [2] I. Babuska, U. Banerjee, and J.E. Osborn. Generalized finite element methods: main ideas, results, and perspective. Technical report, DTIC Document, 2004.
- [3] I. Babuška and J. Osborn. Eigenvalue problems. Handbook of numerical analysis, 2:641–787, 1991.
- [4] G. Bal. Couplage d'équations et homogénéisation en transport neutronique. PhD thesis, 1997.
- [5] J. Bussac and P. Reuss. Traité de neutronique: Physique et calcul des réacteurs nucléaires avec application aux réacteurs à eau pressurisée et aux réacteurs à neutrons rapides. 1978.
- [6] R. Dautray, J.L. Lions, M. Artola, and M. Cessenat. Mathematical Analysis and Numerical Methods for Science and Technology: Evolution Problems I, volume 5. Springer Verlag, 1992.
- [7] R. Dautray, J.L. Lions, M. Artola, and M. Cessenat. Mathematical Analysis and Numerical Methods for Science and Technology: Evolution Problems II, volume 6. Springer Verlag, 1993.
- [8] B. de Pagter. Irreducible compact operators. Mathematische Zeitschrift, 192(1):149– 153, 1986.
- [9] J. Descloux. On spectral approximation. part 2. error estimates for the galerkin method. rairo. *Analyse Numerique*, 12:113–119, 1978.
- [10] Y. Du. Order structure and topological methods in nonlinear partial differential equations: Maximum principles and applications, volume 1. World Scientific Pub Co Inc, 2006.
- [11] J.J. Duderstadt and W.R. Martin. Transport theory. Transport theory., by Duderstadt, JJ; Martin, WR. Chichester (UK): John Wiley & Sons, 10+ 613 p., 1, 1979.
- [12] W.G. Kolata. Approximation in variationally posed eigenvalue problems. Numerische Mathematik, 29(2):159–171, 1978.

- [13] E.E. Lewis and WF Miller. Computational methods of neutron transport. 1984.
- [14] AM Matache, I. Babuška, and C. Schwab. Generalized p-FEM in homogenization. Numer. Math., 86(2):319–375, 2000.
- [15] JM Melenk, K. Gerdes, and C. Schwab. Fully discrete hp-finite elements: Fast quadrature. Computer Methods in Applied Mechanics and Engineering, 190(32-33):4339–4364, 2001.
- [16] M. Mokhtar-Kharroubi. Mathematical Topics in Neutron Transport Theory: New Aspects, volume 46. World Scientific Pub Co Inc, 1997.
- [17] A.D. Polianin and A.V. Manzhirov. Handbook of integral equations. CRC press, 2008.