Adaptive Galerkin Methods
for Parametric and Stochastic
Operator Equations

A dissertation submitted to
ETH Zürich

for the degree of
Doctor of Sciences

presented by
CLAUDE JEFFREY GITTELSON
MSc ETH Mathematics, ETH Zürich
born May 1, 1985
citizen of Villmergen AG, Switzerland
and the United States of America

accepted on the recommendation of
Prof. Dr. Christoph Schwab, ETH Zürich, examiner
Prof. Dr. Helmut Harbrecht, Universität Stuttgart, co-examiner

2011
I have experienced the Seminar for Applied Mathematics at ETH as providing a stimulating and supportive working atmosphere, which is certainly the achievement of all of its current and recent members.

Most of all, I wish to thank Prof. Christoph Schwab for the innumerable opportunities he offered me. These will certainly prove invaluable in the future. Among the professors, I have also had the pleasure of working with Ralf, Daniel and Siddhartha. Each of these collaborations has been a rewarding experience for me.

I am thankful to all of my fellow assistants, who have made my time at SAM very enjoyable, in particular Marcel, Bastian, Andrea, Florian and Annika, with whom I shared an office for various periods of the last three years. Also, I am grateful to Christian for making me run, to Daniel for letting me, to Roman for his collaboration and to Christine for her competition; to Holger and Carlos for giving me priority, to Oleg for always being in a good mood, and to Miro for the opposite reason.

This thesis was made possible by the Swiss National Science Foundation through funding under Pro*Doc Grant № 200021-120290/1. It has indirectly benefited from the ERC Advanced Grant № 247277 on Sparse Tensor Approximations of High-Dimensional and stochastic Partial Differential Equations to Christoph Schwab.
Abstract

We derive adaptive solvers for parametric and stochastic boundary value problems using techniques from adaptive wavelet methods to construct optimal spectral sparse tensor discretizations.

Boundary value problems depending on a random field are transformed into parametric equations depending on a sequence of scalar parameters if the random field is expanded into a series. Assuming independence of the coefficients in this series induces an infinite product probability measure on the parameter domain. This permits a discretization by tensorized polynomials, which are orthonormal with respect to the measure on the parameter domain. The original problem is thus recast as a countably infinite system of partial differential equations for the coefficients of the solution with respect to this basis.

Without any explicit truncation of the series, restricting to a finite set of polynomial basis functions on the parameter domain reduces this infinite system to a finite system of deterministic equations, which can in principle be solved by standard finite element methods.

We adaptively select such finite sets using adaptive wavelet methods, with tensor product polynomial bases in place of wavelets. This requires that the adaptive wavelet methods are extended to a vector-valued setting. We prove that some of our adaptive solvers construct a sequence of approximations which converges at the optimal rate, while scaling linearly in the number of active polynomial basis functions on the parameter domain.

Due to their modular design, these methods can be coupled with arbitrary solvers for the deterministic counterpart of the stochastic boundary value problem under consideration. They can be thought of as exterior iterations, in which solves of deterministic problems constitute individual steps. We generalize certain algorithms from adaptive wavelet methods to automatically select an optimal finite element discretization for each active polynomial coefficient.

Some variants of our solvers are proven to control the error uniformly in the parameter. They can therefore be used to solve purely parametric equations depending affinely on a sequence of scalar parameters, with no probability measure on the parameter domain, or stochastic problems for which the coefficients in the series expansion of the random field are correlated.
Zusammenfassung

Wir leiten adaptive Lös器 für parametrische und stochastische Randwertprobleme her, basierend auf adaptiven Wavelet Lösern. Diese konstruieren adaptiv optimale Tensordiskretisierungen.


Wir wählen adaptiv endliche Mengen von aktiven Basisfunktionen auf dem Parameterraum mittels adaptive Wavelet Verfahren, in denen wir die Wavelet Basen durch Polynome ersetzen. Zu diesem Zweck werden die adaptiven Wavelet Löser zu vektorwertigen Gleichungssystemen verallgemeinert. Wir beweisen, dass einige unserer adaptiven Löser eine Folge von Approximationen konstruieren, die mit der optimalen Rate konvergieren, wobei die Laufzeit sich linear bezüglich der Anzahl polynomialer Basisfunktionen auf dem Parameterraum verhält.

Aufgrund ihres modularen Aufbaus können diese Verfahren mit beliebigen Lösern für das deterministische Gegenstück des stochastischen Randwertproblems kombiniert werden. Sie können als äussere Iterationen aufgefasst werden, in denen das Lösen eines deterministischen Problems einen einzelnen Teilschritt darstellt. Wir verallgemeinern bestimmte Algorithmen aus adaptiven Wavelet Methoden, um automatisch eine optimale Diskretisierung für die Approximation jedes Koeffizienten in der polynomialen Annäherung der Lösung zu bestimmen.

# Contents

**Introduction** xi  
**Notation** xv  

## 1. Parametric and Stochastic Operator Equations 1  
1.1. Parametric Operators .................................................. 2  
   1.1.1. Continuous Parameter Dependence ............................. 2  
   1.1.2. Weak Formulation in the Parameter .......................... 4  
   1.1.3. Tensor Product Structure ..................................... 6  
1.2. Operators with Stochastic Coefficients ............................ 6  
   1.2.1. Transformation to a Parametric Problem .................... 6  
   1.2.2. Affine Parameter-Dependence ................................ 8  
1.3. Tensor Product Construction ........................................ 10  
   1.3.1. Expandable Operators .......................................... 10  
   1.3.2. Application to Parametric Operator Equations ............. 12  

## 2. Countably Infinite Tensor Product Frames 15  
2.1. Frames and Riesz Bases of Hilbert Spaces ........................ 15  
   2.1.1. Frames and Riesz Bases ........................................ 15  
   2.1.2. Vector Frames .................................................. 17  
   2.1.3. Tensor Product Frames ........................................ 20  
   2.1.4. Hierarchic Frames .............................................. 22  
2.2. Frames on Infinite Dimensional Domains ........................... 23  
   2.2.1. Product Domains ............................................... 23  
   2.2.2. Finite Tensor Product Frames ................................. 25  
   2.2.3. Compatibility of Finite Tensor Product Frames ............ 26  
   2.2.4. Infinite Tensor Product Frames .............................. 29  
   2.2.5. Countably Infinite Products of Boundedly Hierarchic Frames 32  

## 3. Transformation to a Discrete System 35  
3.1. Frame Representation .................................................. 35  
   3.1.1. Discretization of the Parameter Domain ..................... 35  
   3.1.2. Transformation to Sequence Spaces ........................... 37  
   3.1.3. Structure of the Operator Matrices .......................... 38  
3.2. Tensor Product Polynomial Bases .................................... 41  
   3.2.1. Construction of Orthonormal Polynomial Bases .............. 41  
   3.2.2. Representation of Affinely Parametric Operators .......... 44
## Contents

3.2.3. The Bi-Infinite Matrix Equation ................................................. 46
3.3. Further Examples ................................................................. 47
  3.3.1. Piecewise Polynomial Bases ................................................. 47
  3.3.2. Rescaled Uniform Bases ....................................................... 50

  4.1. Greedy Solution of a Generalized Knapsack Problem ......................... 54
    4.1.1. Problem Setting .............................................................. 54
    4.1.2. A Sequence of Optimal Solutions ........................................ 54
    4.1.3. Numerical Construction ...................................................... 57
  4.2. Adaptive Application of $s^*$-Computable Operators ......................... 59
    4.2.1. $s^*$-Compressibility and $s^*$-Computability ......................... 59
    4.2.2. An Adaptive Application Routine ......................................... 61
    4.2.3. Convergence Analysis ........................................................ 64
  4.3. Adaptive Application of Parametric Operators ................................ 69
    4.3.1. Sparse Semidiscrete Approximations .................................... 69
    4.3.2. Sparse Approximation of the Discrete Operator ......................... 72
    4.3.3. $s^*$-Compressibility and $s^*$-Computability .......................... 74
    4.3.4. Adaptive Application Routine ............................................. 78

5. Adaptive Iterative Solvers ....................................................... 81
  5.1. A General Iterative Method ..................................................... 82
    5.1.1. An Abstract Perturbed Stationary Linear Iteration ..................... 82
    5.1.2. Realization by Adaptive Subroutines ...................................... 84
    5.1.3. Uniform Convergence ......................................................... 87
  5.2. Extensions ................................................................................. 90
    5.2.1. Alternating Subspace Correction .......................................... 90
    5.2.2. Coarsening of the Approximate Solution .................................. 95
  5.3. Discretization of Deterministic Problems ..................................... 100
    5.3.1. Single Level Discretization ................................................. 100
    5.3.2. Adaptive Solvers ............................................................... 101
    5.3.3. Multiple Level Coarsening ................................................. 103

  6.1. Galerkin Projection ............................................................... 107
    6.1.1. Theoretical Foundations ..................................................... 107
    6.1.2. Algorithmic Aspects .......................................................... 109
    6.1.3. Conjugate Gradient Iteration .............................................. 112
  6.2. An Adaptive Iterative Solver .................................................. 114
    6.2.1. Successive Refinement of the Active Index Set .......................... 114
    6.2.2. Optimality Properties ....................................................... 119
  6.3. Fully Discrete Variant ........................................................... 123
    6.3.1. Computation of the Residual .............................................. 123
    6.3.2. A Discrete Iterative Solver ................................................ 125
7. Computational Examples and Applications 129
7.1. The Isotropic Diffusion Equation .................................................. 129
7.1.1. A Differential Equation with Stochastic Coefficients ................. 129
7.1.2. Series Expansion of the Stochastic Coefficient ....................... 131
7.1.3. Frame Representation ............................................................ 134
7.1.4. Finite Element Approximation ................................................. 136
7.1.5. A Posteriori Error Estimation .................................................. 136
7.1.6. Weak Formulation on Banach Spaces ................................. 139
7.2. Elements of a Matlab Implementation ........................................... 142
7.2.1. Data Structure for the Index Set ............................................. 142
7.2.2. Multilevel Finite Elements ..................................................... 144
7.2.3. Implementation of the Adaptive Solvers ............................... 145
7.3. Numerical Experiments ............................................................... 147
7.3.1. Model Problem ................................................................. 147
7.3.2. Comparison of Methods ....................................................... 147
7.3.3. Uniform Convergence .......................................................... 151
7.3.4. Index Sets ........................................................................ 155
7.3.5. Empirical Convergence Rates .............................................. 156
A. Tensor Products and Vector-Valued Integration 161
A.1. Notions of Measurability and Integration ........................................ 161
A.1.1. The Bochner Integral ............................................................ 161
A.1.2. Weak Measurability and the Pettis Integral ............................ 162
A.1.3. Equivalent Notions of Measurability ...................................... 163
A.2. Tensor Products of Banach Spaces ............................................. 164
A.2.1. The Algebraic Tensor Product ............................................. 164
A.2.2. Cross Norms ........................................................................ 165
A.2.3. The Projective Tensor Product ............................................. 166
A.2.4. The Injective Tensor Product ................................................ 167
A.2.5. The Hilbert Tensor Product ................................................ 168
A.2.6. Tensor Product Operators .................................................... 170
A.3. Tensor Product Structure of Function Spaces ............................. 172
A.3.1. Lebesgue–Bochner spaces .................................................... 172
A.3.2. Lebesgue–Pettis Spaces and the Injective Tensor Product .......... 173
A.3.3. Spaces of Continuous Functions .......................................... 174
A.3.4. Sequence Spaces ............................................................... 174

References 179
Curriculum Vitae 185
Introduction

Numerical analysis of partial differential equations classically studies the convergence of a sequence of approximations to the solution of a given boundary value problem. However, uncertainty in coefficients and in other input data of the boundary value problem prevents these approximations from converging to physically observable quantities, and causes a discrepancy between numerical simulations and experiments that is not due to an error in the modeling of physical phenomena, but simply to insufficient knowledge of the system at hand.

This can be taken into account by considering a range of coefficients and solving the boundary value problem with each of these. If unknown coefficients entering into the differential equation are modeled as depending on a sequence of scalar parameters, the dependence of the solution on these parameters is sought.

If statistics on the unknown coefficients are available, it may be sufficient to determine resulting statistics of the solution to the boundary value problem. For example, if the coefficients are modeled as random fields due to a known inherent variability, then the solution is also a random field, and numerical computations may target stochastic moments or other statistical quantities.

In the latter, stochastic setting, statistics of the solution can be computed by generating samples of the input data and solving the differential equation for each of these. This approach, known as the Monte Carlo method, makes minimal assumptions on the structure of the problem, and is simple to implement as it consists of solving independent deterministic problems. However, it is unable to take advantage of the regular dependence of the solution on the parameters. To a lesser extent, this also holds for quasi-Monte Carlo methods, in which samples are not drawn randomly, but by deterministic methods designed to minimize discrepancy, and for multilevel Monte Carlo methods, see e.g. [GKN+10, BSZ10, MS10]. All of these sampling methods target statistics of the solution directly and do not provide a parametric representation of the solution.

The stochastic Galerkin method approximates the coefficients of the solution with respect to a basis of functions on the parameter domain, such as tensor product polynomials. It is based on a weak formulation of the stochastic boundary value problem also in the parameter. Galerkin projections are computed by solving a coupled system of boundary value problems for the desired coefficients. If the solution depends sufficiently smoothly on the parameters, the stochastic Galerkin method has been seen to be substantially more efficient than the Monte Carlo method, see [DBO01, XK02, BTZ04, WK05, MK05, FST05, WK06, TS07, BS09, BAS10].

The stochastic collocation approach is an attempt to combine advantages of stochastic
Introduction

Galerkin and sampling methods, see [XH05, BNT07, NTW08b, NTW08a, Bie09a, WK09]. Instead of computing the Galerkin projection onto a given space of functions on the parameter domain, an interpolant is constructed. This requires only the solution of independent boundary value problems, and provides an explicit representation of the solution as a function of the parameters. However, this representation may be less efficient than that obtained by Galerkin projection.

In the past decade, new adaptive methods have emerged, which are set not in the continuous framework of the original boundary value problem, or other operator equation, but on the level of coefficients with respect to a hierarchic Riesz basis, such as a wavelet basis. Due to the norm equivalences constitutive of Riesz bases, errors and residuals in appropriate sequence spaces are equivalent to those in physically meaningful function spaces. This permits adaptive wavelet methods to be applied to any equation without modification, provided that a suitable Riesz basis is available.

There are two approaches to such methods. For symmetric elliptic problems, the error of the Galerkin projection onto the span of a set of coefficients can be estimated using a sufficiently accurate approximation of the residual of a previously computed approximate solution, see [CDD01, GHS07, DSS09]. This results in a sequence of finite-dimensional linear equations with successively larger sets of active coefficients.

Alternatively, the operator equation can be interpreted as a bi-infinite matrix equation for the coefficients with respect to a Riesz basis. In principle, this could be solved directly by an iterative method such as the Richardson iteration. Perturbing this iteration by adaptive approximate application of the bi-infinite matrix and suitable finite approximations of the right hand side leads to finitely supported sequences of coefficients, without ever explicitly specifying a set of active coefficients or computing a Galerkin projection. This approach is thus more widely applicable; it can handle saddle point problems and other systems of boundary value problems, see [CDD02].

Representatives of both approaches are optimal in the sense that the number of active coefficients in the approximation to any prescribed tolerance is less than that of the optimal approximation of the solution to this tolerance, up to a constant factor, and the work required to compute this approximation is proportional to this number. This generally requires a coarsening step, compressibility of the bi-infinite matrix in the sense that it can be approximated sufficiently by matrices with finitely many entries per row and column, and some further technical assumptions, such as restrictions on parameters of the adaptive algorithm.

A major obstacle in the numerical solution of parametric and stochastic boundary value problems is the construction of suitable spaces in which to compute approximate solutions. The goal of this thesis is to apply techniques from adaptive wavelet methods to derive adaptive methods for parametric and stochastic operator equations.

As in the stochastic Galerkin and stochastic collocation approaches, we compute the explicit parameter dependence of an approximate solution, expanded with respect to tensor product orthonormal polynomials on the parameter domain. However, our approximation is neither a Galerkin projection, nor is it constructed by interpolating...
between solutions at distinct values of the parameters. Rather, we solve a sequence of deterministic problems as part of an iterative method directly targeting the exact parametric equation.

Rather than focusing on a particular model problem, we keep our exposition as general as possible. Besides the obvious wider applicability of the obtained results, our abstract approach provides valuable insight into the structure of the class of parametric and stochastic operator equations we consider, identifying critical assumptions and pivotal constraints.

Chapter 1 introduces parametric and stochastic operator equations. We derive well-posed weak formulations on the parameter domain under the natural assumption that the operator depends continuously on the parameter. The tensor product structure of the function spaces arising in this manner is of particular importance to the discretization of these equations. Appendix A provides an overview of the relationship between tensor products of Banach spaces and vector-valued integration.

A key ingredient in the discretization process is an orthonormal basis of the space of square integrable functions on the parameter domain. Striving for maximal generality, we construct frames on infinite dimensional domains in Chapter 2. Considering infinitely many factors circumvents the need to truncate series representations of random or parametric coefficients, and thereby avoids premature approximations to the equation at hand.

In line with adaptive wavelet methods, we reformulate the original continuous problem as an equivalent bi-infinite matrix equation in Chapter 3. We also derive an intermediary formulation by discretizing only the parameter dependence of the parametric solution. This reduces the original uncountable set of independent equations to a countably infinite coupled system of equations. Using an orthonormal polynomial basis on the parameter domain leads to a particularly simple structure, which we study in detail.

The most direct approach to solving parametric and stochastic operator equations by adaptive wavelet methods is to employ a wavelet basis on the physical domain in conjunction with, for example, an orthonormal polynomial basis on the parameter domain. The original equation is thereby transformed into a bi-infinite scalar matrix equation for the coefficients with respect to the product of these two bases. As adaptive wavelet methods are formulated in this setting, they apply without modification.

An important prerequisite, however, is a routine which adaptively applies the bi-infinite matrix to a vector. This hinges on the ability to approximate the bi-infinite matrix by matrices with only finitely many elements per row and column. Such approximations, and the resulting adaptive application routines, are presented in Chapter 4.

A more flexible approach is based on the formulation of the parametric or stochastic operator equation as an infinite system of operator equations, derived by introducing a basis, such as orthonormal polynomials, only on the parameter domain. This system can be interpreted as a single equation with a linear operator represented by a bi-infinite operator matrix. Adaptive wavelet methods generalize to this vector setting. Furthermore, the resulting methods are modular in that they can be coupled with
Introduction

arbitrary discretizations of or solvers for the deterministic problem.

In Chapter 5, we consider methods based on the general adaptive wavelet method introduced in [CDD02]. These are applicable to parametric and stochastic operator equations on Hilbert spaces and Banach spaces. For suitable spatial discretizations, they provide upper bounds for the error in a Lebesgue space of square integrable vector-valued functions on the parameter domain.

With minor modifications, these methods are able to control the error uniformly in the parameter. This is the most meaningful type of convergence for parametric equations, in which no measure is provided on the parameter domain. Moreover, for stochastic equations in which the probability measure is not suitable for constructing an orthonormal basis on the parameter domain, for example if it is not a product measure, any other measure can be used, and uniform convergence implies convergence with respect to the original measure with no intangible absolute continuity constraints.

In Chapter 6, we apply a Galerkin-type adaptive wavelet method similar to [CDD02, GHS07, DSS09] to symmetric elliptic parametric or stochastic operator equations. This method provides an explicit error bound in the energy norm, which is equivalent to the Lebesgue–Bochner norm. It can be coupled with arbitrary finite elements on the physical domain, including adaptive finite element solvers.

Finally, Chapter 7 applies the above adaptive methods to the isotropic diffusion equation with a stochastic diffusion coefficient. We discuss some details of a Matlab implementation, and present results of numerical computations.

The discussion in Chapter 7 illustrates how our abstract operator equation framework maps to a concrete boundary value problem. Due to the generality of their derivation, our methods apply analogously to many other equations, such as anisotropic diffusion with a stochastic diffusion tensor, the Stokes equation with a stochastic kinematic viscosity and a mixed finite element discretization, or parabolic problems with either Galerkin time stepping or a space-time Galerkin discretization.
Notation

N  natural numbers excluding zero.
N₀  natural numbers including zero.
P(X) the set of all subsets of X.
Σ(X) the set of all finite subsets of X.
⊂  subset, X ⊂ Y if any element of X is also in Y.
Σ(X) Borel σ-algebra on X.
L(X, Y) space of bounded linear maps from X to Y.
L(X) space of bounded linear maps from X to X.
∥·∥_{X→Y} operator norm on L(X, Y).
L^*(X, Y) space of bounded antilinear maps from X to Y.
B_X closed unit ball of X.
X* space of all bounded antilinear functionals on a Banach space X.
⟨·, ·⟩_X duality pairing on X, linear in the first argument and antilinear in the second.
(·, ·)_X inner product on X, linear in the first argument and antilinear in the second.
⊗ tensor product of operators, vectors, or measures.
⊗ the algebraic tensor product. 164
⊗_α the topological tensor product with cross norm α. 165
⊗_η the Hilbert tensor product. 168
⊗_i the injective tensor product. 167
⊗_π the projective tensor product. 166
b_Φ, B_Φ frame bounds of the frame Φ. 16
Φ*, (ϕ_v*)_{v∈Ξ} canonical dual frame of Φ = (ϕ_v)_{v∈Ξ}. 17
S_Φ frame operator of the frame Φ. 16
Notation

\( T_\Phi \) synthesis operator of the frame \( \Phi \). 15

\( K_\Phi \) kernel of \( T_\Phi \). 16

\( \pi_\Phi \) projection onto the quotient space \( \ell^2(\Xi)/K_\Phi \). 16

\( L^p_\mu(\Omega; X) \) Lebesgue–Bochner space. 172

\( \overline{L}^p_\mu(\Omega; X) \) Lebesgue–Pettis space. 173

\( \ell^p(\Xi; X) \) Lebesgue–Bochner sequence space. 174

\( \mathcal{A}^a(\Xi) \) approximation space in \( \ell^2(\Xi) \). 64

\( \mathcal{A}^a(\Xi; X) \) approximation space in \( \ell^2(\Xi; X) \). 95

\( 1_X \) indicator function, \( 1_X(x) := 1 \) if \( x \in X \), and 0 otherwise.

\( \mathcal{N} \) set of neighboring pairs in a set of multiindices. 110

\( c_\eta, d_{\eta,s} \) compressibility constants. 59

\( e_m, e_v \) Kronecker sequence. 29, 45

\( \lambda \) average index length of a set of multiindices. 110

\( \log^+ \) positive logarithm, \( \log^+(x) := \log(\max(x, 1)) \).

\( \Pi_\Theta \) orthogonal projection in \( \ell^2(\Xi; X) \) onto \( \ell^2(\Theta; X) \). 108
Chapter 1.

Parametric and Stochastic Operator Equations

A parametric operator equation is, as the name suggests, an operator equation in which the operator and right hand side depend on a parameter. As such, it could in principle be solved independently for each value of the parameter. If the parameter has infinitely many admissible values, either a finite subset must be selected which in some sense sufficiently exhausts the parameter domain, or the solution can be approximated directly in a parametric form.

We follow the latter strategy, and interpret the original parametric operator equation as a single operator equation set in a space of functions on the parameter domain. It is important that this space has a tensor product structure; Lebesgue spaces of square integrable functions on the parameter domain are particularly useful. One may think of a reformulation of the parametric equation on a Lebesgue–Bochner space as a weak formulation of the parametric equation.

As an intermediate step, we formulate the parametric equation on spaces of continuous functions. This is of interest in its own right, and enables us to dispense with intangible measurability conditions on the parametric operator by assuming continuous parameter dependence. Parametric operator equations are discussed in Section 1.1.

Stochastic operator equations are parametric operator equations in which there is a probability measure on the parameter domain. More precisely, we assume that the operator and right hand side depend continuously on some coefficients, which are modeled as random variables. In Section 1.2, we reduce such equations to the setting of Section 1.1.

The resulting theory is quite satisfactory for separable Hilbert spaces. However, Lebesgue–Bochner spaces of square integrable Banach-valued functions lack a tensor product structure. Also, the theory does not extend to Lebesgue–Pettis spaces, which are always injective tensor products. In Section 1.3, we develop a different approach based directly on the assumed tensor product structure of the parametric operator, and derive a weak formulation of the parametric operator equation on Lebesgue–Pettis spaces.
Chapter 1. Parametric and Stochastic Operator Equations

1.1. Parametric Operators

1.1.1. Continuous Parameter Dependence

Let \( V \) and \( W \) be Banach spaces over \( \mathbb{K} \in \{ \mathbb{R}, \mathbb{C} \} \). Denote by \( W^* \) the space of bounded antilinear maps from \( W \) to \( \mathbb{K} \), and by \( \mathcal{L}(V, W^*) \) the Banach space of bounded linear maps from \( V \) to \( W^* \).\(^1\)

Let \( \Gamma \) be a nonempty topological space. A parametric linear operator from \( V \) to \( W^* \) with parameter domain \( \Gamma \) is a continuous map

\[
A : \Gamma \to \mathcal{L}(V, W^*) , \quad y \mapsto A(y) .
\]

(1.1.1)

For a given \( f : \Gamma \to W^* \), we are interested in determining \( u : \Gamma \to V \) such that

\[
A(y)u(y) = f(y) \quad \forall y \in \Gamma .
\]

(1.1.2)

**Assumption 1.1.A.** \( A(y) \) is bijective for all \( y \in \Gamma \).

By the open mapping theorem, Assumption 1.1.A implies that \( A(y) \) is boundedly invertible for all \( y \in \Gamma \).

**Theorem 1.1.1.** Equation (1.1.2) has a unique solution \( u : \Gamma \to V \). It is continuous if and only if \( f : \Gamma \to W^* \) is continuous.

**Proof.** By Assumption 1.1.A, (1.1.2) has the unique solution \( u(y) = A(y)^{-1} f(y) \).

Let \( D \in \mathcal{L}(V, W^*) \) be boundedly invertible. For example, \( D \) could be equal to \( A(y) \) for some \( y \in \Gamma \). Then \( y \mapsto D^{-1}A(y) \) is a continuous map from \( \Gamma \) into \( \mathcal{L}(V) \). By the abstract property [KR97, Prop. 3.1.6] of Banach algebras, the map \( \text{inv} : T \mapsto T^{-1} \) defined on the multiplicative group of \( \mathcal{L}(V) \) is continuous in the topology of \( \mathcal{L}(V) \). Therefore, \( y \mapsto \text{inv}(D^{-1}A(y)) = A(y)^{-1}D \) is continuous, and multiplying from the right by the constant \( D^{-1} \), it follows that \( y \mapsto A(y)^{-1} \) is a continuous map from \( \Gamma \) to \( \mathcal{L}(W^*, V) \).

Note that the application of an operator to a vector, *i.e.* the map \( \text{mult} : \mathcal{L}(W^*, V) \times W \to V \) defined by \( \text{mult}(T, z) = Tz \), is continuous. Therefore, if \( y \mapsto f(y) \) is continuous, then so is \( y \mapsto u(y) = \text{mult}(A(y)^{-1}, f(y)) \). Similarly, since \( \text{mult} : \mathcal{L}(W^*, V) \times V \to W^* \) is continuous, if \( y \mapsto u(y) \) is continuous, then so is \( y \mapsto f(y) = \text{mult}(A(y), u(y)) \). \( \square \)

**Example 1.1.2.** Assumption 1.1.A is assured to hold if \( A(y) \) is a perturbation of a boundedly invertible \( D \in \mathcal{L}(V, W^*) \), *i.e.*

\[
A(y) = D + R(y) , \quad y \in \Gamma ,
\]

(1.1.3)

with a continuous \( y \mapsto R(y) \in \mathcal{L}(V, W^*) \) satisfying

\[
\left\| D^{-1}R(y) \right\|_{W^* \to V} \leq \gamma < 1 \quad \forall y \in \Gamma .
\]

(1.1.4)

\(^1\)We follow the convention that, for any Banach space \( X \), the dual space \( X^* \) is the space of bounded antilinear functionals on \( X \), but the bidual \( X^{**} \) is the space of linear functionals on \( X^* \).
Then \( A(y) \) can be decomposed as
\[
A(y) = D(\text{id}_V + D^{-1}R(y)) , \quad y \in \Gamma ,
\]
and consequently, using a Neumann series in \( \mathcal{L}(V) \) to invert the second factor,
\[
A(y)^{-1} = \left( \sum_{n=0}^{\infty} \left( -D^{-1}R(y) \right)^n \right) D^{-1} , \quad y \in \Gamma .
\]

In this setting, due to (1.1.4), (1.1.5) and (1.1.6), the parametric operators \( A(y) \) and \( A(y)^{-1} \) are uniformly bounded,
\[
\| A(y) \|_{V \to W} \leq \| D \|_{V \to W} (1 + \gamma) \quad \forall y \in \Gamma , \quad (1.1.7)
\]
\[
\| A(y)^{-1} \|_{W \to V} \leq \| D^{-1} \|_{W \to V} \frac{1}{1 - \gamma} \quad \forall y \in \Gamma . \quad (1.1.8)
\]

A sufficient condition for (1.1.4) is
\[
\| R(y) \|_{V \to W} \leq \frac{\gamma}{\| D^{-1} \|_{W \to V}} \quad \forall y \in \Gamma \quad (1.1.9)
\]
with \( \gamma < 1 \). Equation (1.1.9) does not depend on the precise structure of the operators \( D^{-1}R(y) \). Therefore, Assumption 1.1.A is always satisfied if the parametric component \( R(y) \) of \( A(y) \) is sufficiently small.

**Assumption 1.1.B.** \( \Gamma \) is a compact Hausdorff space.

**Lemma 1.1.3.** There exist constants \( \bar{c}, \bar{\gamma} \in \mathbb{R} \) such that
\[
\| A(y) \|_{V \to W} \leq \bar{c} \quad \text{and} \quad \| A(y)^{-1} \|_{W \to V} \leq \bar{\gamma} \quad \forall y \in \Gamma . \quad (1.1.10)
\]

**Proof.** By assumption, the map \( y \mapsto A(y) \) is continuous. As shown in the proof of Theorem 1.1.1, \( y \mapsto A(y)^{-1} \) is also continuous. Consequently, the maps \( y \mapsto \| A(y) \|_{V \to W} \) and \( y \mapsto \| A(y)^{-1} \|_{W \to V} \) are continuous maps from \( \Gamma \) into \( \mathbb{R} \). Since \( \Gamma \) is compact by Assumption 1.1.B, the ranges of these maps are compact in \( \mathbb{R} \), and therefore bounded. \( \square \)

For any Banach space \( X \), let \( C(\Gamma; X) \) denote the Banach space of continuous maps from \( \Gamma \) to \( X \) with norm
\[
\| v \|_{C(\Gamma; X)} := \sup_{y \in \Gamma} \| v(y) \|_X , \quad v \in C(\Gamma; X) . \quad (1.1.11)
\]

In what follows, we abbreviate \( C(\Gamma) := C(\Gamma; \mathbb{K}) \).

**Corollary 1.1.4.** The operators
\[
\mathcal{A} : C(\Gamma; V) \to C(\Gamma; W^*), \quad v \mapsto [y \mapsto A(y)v(y)] \quad \text{and} \quad (1.1.12)
\]
\[
\mathcal{A}^{-1} : C(\Gamma; W^*) \to C(\Gamma; V), \quad g \mapsto [y \mapsto A(y)^{-1}g(y)] \quad (1.1.13)
\]
are well-defined, inverse to each other, and bounded with norms \( \| \mathcal{A} \| \leq \bar{c} \) and \( \| \mathcal{A}^{-1} \| \leq \bar{\gamma} \).

**Proof.** The assertion is a direct consequence of Theorem 1.1.1 and Lemma 1.1.3. \( \square \)
Chapter 1. Parametric and Stochastic Operator Equations

1.1.2. Weak Formulation in the Parameter

We extend Corollary 1.1.4 by density to Lebesgue spaces of vector-valued functions. Let \( \mathcal{B}(\Gamma) \) be the Borel \( \sigma \)-algebra on \( \Gamma \), and let \( \mu \) be a finite measure on \( (\Gamma, \mathcal{B}(\Gamma)) \). We note that \( \mu \) is a regular Borel measure, and in particular a Radon measure, see e.g. [Bau92, Satz 29.12].

For a Banach space \( X \), let \( C(\Gamma; X)/\mu \) denote the space of equivalence classes of \( \mu \)-a.e. identical functions in \( C(\Gamma; X) \).

**Lemma 1.1.5.** The operators

\[ \mathcal{A} : C(\Gamma; V)/\mu \rightarrow C(\Gamma; W^*)/\mu \quad \text{and} \quad \mathcal{A}^{-1} : C(\Gamma; W^*)/\mu \rightarrow C(\Gamma; V)/\mu \]

are well-defined.

**Proof.** Let \( v \in C(\Gamma; V) \) with \( v(y) = 0 \) for \( \mu \)-a.e. \( y \in \Gamma \). Then by linearity of \( A(y) \), \( A(y)v(y) = 0 \) for \( \mu \)-a.e. \( y \in \Gamma \). Therefore, if \( v = 0 \) in \( C(\Gamma; V)/\mu \), then \( \mathcal{A}v = 0 \) in \( C(\Gamma; W^*)/\mu \). The same argument applies to \( \mathcal{A}^{-1} \).

**Theorem 1.1.6.** For all \( 1 \leq p < \infty \), the operator \( \mathcal{A} \) from (1.1.12) extends uniquely to a boundedly invertible operator on the Lebesgue–Bochner spaces

\[ \mathcal{A} : L^p_\mu(\Gamma; V) \rightarrow L^p_\mu(\Gamma; W^*) \quad (1.1.14) \]

The norms of \( \mathcal{A} \) and \( \mathcal{A}^{-1} \) are bounded by \( \hat{c} \) and \( \hat{\varepsilon} \), respectively.

**Proof.** By [AE01, Thm. X.4.14], \( C(\Gamma; V)/\mu \) is dense in \( L^p_\mu(\Gamma; V) \), and \( C(\Gamma; W^*)/\mu \) is dense in \( L^p_\mu(\Gamma; W^*) \).2 For all \( v \in C(\Gamma; V) \), by Lemma 1.1.3,

\[ \| \mathcal{A}v \|^p_{L^p_\mu(\Gamma; W^*)} = \int_\Gamma \| A(y)v(y) \|^p_{W^*} \, d\mu(y) \leq \hat{c}^p \int_\Gamma \| v(y) \|^p_\mu \, d\mu(y) = \hat{c}^p \| v \|^p_{L^p_\mu(\Gamma; V)} . \]

Therefore, \( \mathcal{A} \) extends by continuity to an operator (1.1.14) with norm less than or equal to \( \hat{c} \). Similarly, \( \mathcal{A}^{-1} \) extends to an operator with norm at most \( \hat{\varepsilon} \), which is inverse to \( \mathcal{A} \) since this holds on the dense subspaces \( C(\Gamma; V)/\mu \) and \( C(\Gamma; W^*)/\mu \) by Corollary 1.1.4 and Lemma 1.1.5.

**Remark 1.1.7.** The operator \( \mathcal{A} \) is defined on \( L^p_\mu(\Gamma; V) \) only by continuous extension of the explicit definition (1.1.12). If \( v \) is a version of an element of \( L^p_\mu(\Gamma; V) \), then the map \( y \mapsto A(y)v(y) \) is a version of \( \mathcal{A}v \) with respect to equivalence classes of \( \mu \)-a.e. identical functions. This is clear for \( v \in C(\Gamma; V)/\mu \). For a general \( v \in L^p_\mu(\Gamma; V) \), let \( v_n \in C(\Gamma; V)/\mu, n \in \mathbb{N}, \) with \( v_n \rightarrow v \) in \( L^p_\mu(\Gamma; V) \). Then using \( (\mathcal{A}v_n)(y) = A(y)v_n(y) \) for \( \mu \)-a.e. \( y \in \Gamma \),

\[ \int_\Gamma \| A(y)v(y) - (\mathcal{A}v)(y) \|^p_{W^*} \, d\mu(y) \leq \lim_{n \to \infty} \hat{\varepsilon} \int_\Gamma \| v(y) - v_n(y) \|^p_{W^*} \, d\mu(y) = 0 . \]

2Note that the assumption in [AE01, Thm. X.4.14] that \( \Gamma \) is metric is unnecessary since this is only used to apply Urysohn’s lemma, which holds on all normal Hausdorff spaces, and in particular on compact Hausdorff spaces. Alternatively, density follows from applying the scalar result [Bau92, Satz 29.14] to functions of the form \( y \mapsto v_1(y) \) with \( E \in \mathcal{B}(\Gamma) \), which span the Lebesgue–Bochner spaces.
1.1. Parametric Operators

Similarly, the operator $\mathcal{A}^{-1}$ on $L^p_\mu(\Gamma; W^*)$ is equal to pointwise application of $A(y)^{-1}$ up to $\mu$-a.e. equivalence.

**Corollary 1.1.8.** Let $V$ and $W$ be separable Banach spaces, $1 \leq p < \infty$, and let $q$ be the Hölder conjugate of $p$. If $f \in L^p_\mu(\Gamma; W^*)$, then there is a unique $\tilde{\mu} \in L^q_\mu(\Gamma; V)$ such that

$$
\int_\Gamma \langle A(y)\tilde{\mu}(y), w(y) \rangle_W \, d\mu(y) = \int_\Gamma \langle f(y), w(y) \rangle_W \, d\mu(y) \quad \forall w \in L^q_\mu(\Gamma; W) .
$$

Furthermore, the solution $u$ of (1.1.2) is a version of $\tilde{\mu}$.

**Proof.** Theorem 1.1.6 implies that $\tilde{\mu} := \mathcal{A}^{-1} f$ is in $L^q_\mu(\Gamma; V)$. By Remark 1.1.7, passing to an arbitrary version of $\tilde{\mu}$, $A(y)\tilde{\mu}(y) = f(y)$ for $\mu$-a.e. $y \in \Gamma$. Consequently, for any $w \in L^q_\mu(\Gamma; W)$,

$$
\langle A(y)\tilde{\mu}(y), w(y) \rangle_W = \langle f(y), w(y) \rangle_W \quad \text{a.e. } y \in \Gamma .
$$

Both sides are integrable due to Hölder’s inequality. Equation (1.1.15) follows by integrating over $\Gamma$ with respect to $\mu$.

Suppose $\bar{u} \in L^q_\mu(\Gamma; V)$ satisfies (1.1.15). Let $\bar{w} \in W$. For all $E \in \mathcal{B}(\Gamma)$, setting $w(y) := 1_E(y)\bar{w}$, it follows that there is a null set $N(\bar{w}) \in \mathcal{B}(\Gamma)$ such that

$$
\langle A(y)\bar{u}(y), w(y) \rangle_W = \langle f(y), w(y) \rangle_W \quad \forall y \in \Gamma \setminus N(\bar{w}) .
$$

Choosing for $\bar{w}$ each element of a countable dense subset of $W$ implies $A(y)\bar{u}(y) = f(y)$ for $\mu$-a.e. $y \in \Gamma$. Consequently, the solution $u$ of (1.1.2) is a version of $\bar{u}$, and, in particular, $\bar{u}$ is unique. \hfill \square

In the following, we no longer distinguish between the solution $u$ of (1.1.2) and its equivalence class $\bar{u}$.

**Remark 1.1.9.** Let $V = W$ be a separable Hilbert space, and let $A(y) \in \mathcal{L}(V, V^*)$ be a positive operator for all $y \in \Gamma$, i.e.

$$
V \times V \ni (v, w) \mapsto \langle A(y)v, w \rangle_V
$$

is an inner product on $V$ for all $y \in \Gamma$. Then the bilinear form in (1.1.15) defines an inner product

$$
(v, w)_\mathcal{A} := \langle A(y)v, w \rangle_{L^2_\mu(\Gamma; V)} = \int_\Gamma \langle A(y)v(y), w(y) \rangle_W \, d\mu(y)
$$

on $L^2_\mu(\Gamma; V)$. It induces a norm, called the *energy norm*,

$$
\|v\|_\mathcal{A} := \sqrt{\langle A(y)v, v \rangle_{L^2_\mu(\Gamma; V)}} , \quad v \in L^2_\mu(\Gamma; V) ,
$$

which is equivalent to the standard norm on $L^2_\mu(\Gamma; V)$ with constants

$$
\frac{1}{\sqrt{c}} \|v\|_{L^2_\mu(\Gamma; V)} \leq \|v\|_\mathcal{A} \leq \sqrt{c} \|v\|_{L^2_\mu(\Gamma; V)} \quad \forall v \in L^2_\mu(\Gamma; V) .
$$

by Theorem 1.1.6. \hfill \square
1.1.3. Tensor Product Structure

Several of the spaces above have a tensor product structure. Thus the operator $\mathcal{A}$ and its inverse $\mathcal{A}^{-1}$ can be interpreted as acting on tensor products of Banach spaces. In the following, we identify the simple tensor $\varphi \otimes v$ for $\varphi \in C(\Gamma)$ and $v$ in a Banach space $X$ with the map $y \mapsto \varphi(y)v$ in $C(\Gamma; X)$, and similarly for other function spaces on $\Gamma$. In particular, $\mathcal{A}$ and $\mathcal{A}^{-1}$ act on simple tensors by

$$\mathcal{A}(\varphi \otimes v) = [y \mapsto \varphi(y)A(y)v] \quad \text{and} \quad \mathcal{A}^{-1}(\varphi \otimes g) = [y \mapsto \varphi(y)A(y)^{-1}g] \quad (1.1.19)$$

for $\varphi \in C(\Gamma)$, $v \in V$ and $g \in W^*$. 

Proposition 1.1.10. The operator $\mathcal{A}$ is a boundedly invertible operator from $C(\Gamma) \otimes_i V$ to $C(\Gamma) \otimes_i W^*$ with inverse $\mathcal{A}^{-1}$ and norms at most $\check{c}$ and $\check{c}$, respectively.

Proof. The assertion is a direct consequence of Corollary 1.1.4 since $C(\Gamma; X)$ is isometrically isomorphic to the injective tensor product $C(\Gamma) \otimes_i X$ for any Banach space $X$ by Theorem A.3.3. $\square$

A similar property holds for appropriate tensor products of Lebesgue spaces, for example for the projective tensor products $L^1_p(\Gamma) \otimes_i V$ and $L^1_p(\Gamma) \otimes_i W^*$, which are isometrically isomorphic to the Lebesgue–Bochner spaces $L^1_p(\Gamma; V)$ and $L^1_p(\Gamma; W^*)$, respectively. We are more interested in the case $p = 2$ and Hilbert tensor products.

Theorem 1.1.11. If $V$ and $W$ are separable Hilbert spaces, then $\mathcal{A}$ is a boundedly invertible operator from $L^2_p(\Gamma) \otimes \eta V$ to $L^2_p(\Gamma) \otimes \eta W^*$ with inverse $\mathcal{A}^{-1}$ and norms at most $\check{c}$ and $\check{c}$, respectively.

Proof. This is a consequence of Theorem 1.1.6 since for a separable Hilbert space $X$, the Lebesgue–Bochner space $L^1_p(\Gamma; X)$ is isometrically isomorphic to the Hilbert tensor product $L^2_p(\Gamma) \otimes \eta X$ by Theorem A.3.2. $\square$

1.2. Operators with Stochastic Coefficients

1.2.1. Transformation to a Parametric Problem

For a topological vector space $Z$, we consider the parametric operator equation

$$A_0(z)u_0(z) = f_0(z) \quad \forall z \in Z , \quad (1.2.1)$$

where $A_0$ and $f_0$ are continuous maps from $Z$ to $\mathcal{L}(V, W^*)$ and $W^*$, respectively, as in (1.1.1). Of course, abstractly, one might take $X$ to be $\mathcal{L}(V, W^*) \times W^*$, and define $A_0$ and $f_0$ simply as the projections onto the first and second coordinates; however, we think of $z$ as a coefficient, or possibly a sequence of coefficients, on which $A_0$ and $f_0$ depend.

We model this coefficient as a $Z$-valued random variable $\bar{q}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$,

$$\bar{q}: (\Omega, \mathcal{F}) \to Z . \quad (1.2.2)$$

If $A_0(\bar{q}(\omega))$ is boundedly invertible for all $\omega \in \Omega$, then by Theorem 1.1.1, the solution of (1.2.1) is a $V$-valued random variable $U(\omega) := u_0(\bar{q}(\omega))$, $\omega \in \Omega$. 

6
1.2. Operators with Stochastic Coefficients

**Assumption 1.2.A.** For a finite or countably infinite index set \( \mathcal{M} \), there is a sequence \( Y = (Y_m)_{m \in \mathcal{M}} \) of random variables \( Y_m : \Omega \to \Gamma_m \) with compact Hausdorff codomains \( \Gamma_m \) such that \( \tilde{q}(\omega) \) depends continuously on \( Y(\omega) \), i.e. for a continuous map \( q \),

\[
q : \Gamma := \prod_{m \in \mathcal{M}} \Gamma_m \to Z, \quad \tilde{q}(\omega) = q(Y(\omega)) \quad \forall \omega \in \Omega, \quad (1.2.3)
\]

and \( A_0(q(y)) \) is boundedly invertible for all \( y = (y_m)_{m \in \mathcal{M}} \in \Gamma \).

If the index set \( \mathcal{M} \) is empty, then \( \Gamma = \{0\} \) by convention, and \( q \) is constant. Thus we identify deterministic operator equations as a special case.

The product space \( \Gamma \) defined in (1.2.3) is compact and Hausdorff by Tychonoff’s theorem. Therefore, the parametric operator

\[
A : \Gamma \to \mathcal{L}(V, W^*), \quad y \mapsto A(y) := A_0(q(y)), \quad (1.2.4)
\]

satisfies Assumptions 1.1.A and 1.1.B. Furthermore, since \( A_0 \) and \( q \) are continuous by assumption, \( A(y) \) depends continuously on the parameter \( y \in \Gamma \). Similarly, \( f = f_0 \circ q \) is a continuous map from \( \Gamma \) to \( W^* \). Therefore, Theorem 1.1.1 provides a continuous solution \( u : \Gamma \to V \) of (1.1.2). It is related to \( u_0 \) and \( U \) by

\[
u(Y(\omega)) = U(\omega) = u_0(\tilde{q}(\omega)) \quad \forall \omega \in \Omega. \quad (1.2.5)
\]

More generally, \( u(y) = u_0(q(y)) \) for all \( y \in \Gamma \) by definition of \( A \) and \( f \).

Since the index set \( \mathcal{M} \) is assumed to be at most countably infinite, the Borel and product \( \sigma \)-algebras on \( \Gamma \) coincide,

\[
\mathcal{B}(\Gamma) = \bigotimes_{m \in \mathcal{M}} \mathcal{B}(\Gamma_m), \quad (1.2.6)
\]

and measurability of the map

\[
Y : (\Omega, \mathcal{F}) \to (\Gamma, \mathcal{B}(\Gamma)), \quad \omega \mapsto Y(\omega) = (Y_m(\omega))_{m \in \mathcal{M}}, \quad (1.2.7)
\]

is a consequence of the measurability of \( (Y_m)_{m \in \mathcal{M}} \) assumed in Assumption 1.2.A. Therefore, the distribution of \( Y \), i.e. the image of \( P \) under \( Y \),

\[
\rho := Y(P) \subseteq \mathcal{B}(\Gamma) \ni E \mapsto \rho(E) = P(Y^{-1}(E)), \quad (1.2.8)
\]

is a probability measure on \( (\Gamma, \mathcal{B}(\Gamma)) \).

The measure \( \rho \) could be used as \( \mu \) in the weak formulation of (1.1.2) from Section 1.1.2. However, it is useful for \( \mu \) to be a product measure on \( \Gamma \), which \( \rho \) is not unless the sequence \( (Y_m)_{m \in \mathcal{M}} \) is independent. This suggests using

\[
\mu := \bigotimes_{m \in \mathcal{M}} \rho_m \subset \mathcal{B}(\Gamma) \ni E \mapsto \rho_m(E) = P_m(Y_m^{-1}(E)), \quad (1.2.9)
\]

instead. More generally, we define

\[
\mu := \pi := \bigotimes_{m \in \mathcal{M}} \pi_m \quad (1.2.10)
\]

for arbitrary probability measures \( \pi_m \) on \( (\Gamma_m, \mathcal{B}(\Gamma_m)) \).
Chapter 1. Parametric and Stochastic Operator Equations

Remark 1.2.1. The measure \( \rho \) on \((\Gamma, \mathcal{B}(\Gamma))\) can be used to compute statistics of the solution \( U \). For example, if \( U \in L^2_{\rho}(\Omega; V) \), then the mean of \( U \) is

\[
\bar{U} := \int_{\Omega} U(\omega) \, dP(\omega) = \int_{\Gamma} u(y) \, d\rho(y) \in V,
\]

and the covariance operator of the distribution of \( U \) is given by

\[
Q_U(w, v) := \int_{\Omega} (w, U(\omega) - \bar{U})_V (U(\omega) - \bar{U}, v)_V \, dP(\omega)
\]

\[
= \int_{\Gamma} (w, u(y) - \bar{U})_V (u(y) - \bar{U}, v)_V \, d\rho(y)
\]

for \( v, w \in V \). Other statistics can be computed similarly.

Given only a sequence of approximations of \( u \), convergence to \( u \) in a space of the form \( L^p_{\rho}(\Gamma; V) \) is required for the approximations to the above statistics to converge. Numerical methods based on the weak formulation of (1.1.2) with an auxiliary measure \( \pi \) may, however, only provide convergence in \( L^q_{\pi}(\Gamma; V) \) for some \( q \). If \( \rho \ll \pi \), then by Hölder’s inequality, this is adequate if \( q \) is sufficiently large. More generally, if \( \rho \) is not absolutely continuous with respect to \( \pi \), then even convergence in \( L^\infty_{\pi}(\Gamma; V) \) is insufficient for the approximations to the above statistics to be meaningful. Instead, uniform pointwise convergence everywhere on \( \Gamma \) must be shown, for example using that \( u \) depends continuously on \( y \in \Gamma \).

1.2.2. Affine Parameter-Dependence

We assume for simplicity that \( Z \) is a real Banach space, and consider the case that the random variables \((Y_m)_{m \in \mathbb{M}}\) are coefficients in a series expansion of \( \tilde{q} \). Let \( \bar{q}, q_m \in Z, m \in \mathbb{M} \), such that

\[
\tilde{q}(\omega) = \bar{q} + \sum_{m \in \mathbb{M}} Y_m(\omega)q_m \quad \forall \omega \in \Omega
\]

with convergence in \( Z \). We assume that all of the \( Y_m \) are bounded. Then without loss of generality, \( Y_m(\Omega) \subset [-1, 1] =: \Gamma_m \) for all \( m \in \mathbb{M} \).

In this case, we define

\[
q(y) := \bar{q} + \sum_{m \in \mathbb{M}} y_m q_m, \quad y = (y_m)_{m \in \mathbb{M}} \in \Gamma = [-1, 1]^\mathbb{M}.
\]

It is clear that \( \tilde{q}(\omega) = q(Y(\omega)) \) for all \( \omega \in \Omega \). The map \( y \mapsto q(y) \) is well-defined and continuous if, for example,

\[
\sum_{m \in \mathbb{M}} \|q_m\|_Z < \infty.
\]

Then Assumption 1.2.A is satisfied if \( A(y) := A_0(q(y)) \) is bijective for all \( y \in \Gamma \).
1.2. Operators with Stochastic Coefficients

If \( \tilde{q} \) is a stochastic coefficient in a linear operator \( A_0(\tilde{q}) \), then the dependence of \( A_0(z) \) on \( z \in Z \) is often affine. We consider this case here. Let \( A_0^0 \in \mathcal{L}(V, W^*) \) and let \( A_0^1 : Z \to \mathcal{L}(V, W^*) \) be continuous and linear such that

\[
A_0(z) = A_0^0 + A_0^1(z) \quad \forall z \in Z .
\]

(1.2.16)

Inserting \( z = q(y) \) from (1.2.14) into (1.2.16) leads to

\[
A(y) = A_0^0 + A_0^1(\tilde{q}) + \sum_{m \in \mathcal{M}} y_m A_0^1(q_m) = D + \sum_{m \in \mathcal{M}} y_m R_m \quad \forall y \in \Gamma ,
\]

(1.2.17)

for \( D := A_0^0 + A_0^1(\tilde{q}) \) and \( R_m := A_0^1(q_m), m \in \mathcal{M} \). Note that

\[
R : \Gamma \to \mathcal{L}(V, W^*) , \quad y \mapsto R(y) := \sum_{m \in \mathcal{M}} y_m R_m ,
\]

(1.2.18)

is well-defined and continuous by the assumptions that \( q(y) \) depends continuously on \( y \in \Gamma \) and \( A_0^1 \) is a bounded linear map. We assume that \( D \in \mathcal{L}(V, W^*) \) is boundedly invertible, and

\[
\sum_{m \in \mathcal{M}} \|D^{-1} R_m\|_{V \to V} \leq \gamma < 1 .
\]

(1.2.19)

As in Example 1.1.2, \( A(y) \) can be inverted by a Neumann series in \( \mathcal{L}(V) \) for all \( y \in \Gamma \), since by triangle inequality and using \( |y_m| \leq 1 \),

\[
\|D^{-1} R(y)\|_{V \to V} \leq \sum_{m \in \mathcal{M}} \|D^{-1} R_m\|_{V \to V} \quad \forall y \in \Gamma .
\]

In particular, Assumption 1.2.A is satisfied.

**Remark 1.2.2.** Let \( V = W \) be a separable Hilbert space. If \( D \) is positive and \( R_m \) is symmetric for all \( m \in \mathcal{M} \), and if

\[
\sum_{m \in \mathcal{M}} \|D^{-1/2} R_m D^{-1/2}\|_{V \to V} \leq \gamma < 1 ,
\]

(1.2.20)

then \( A(y) \) is positive for all \( y \in \Gamma \) since for all \( v \in V \),

\[
\langle A(y)v, v \rangle_V = \langle Dv, v \rangle_V + \sum_{m \in \mathcal{M}} y_m \langle R_m v, v \rangle_V
\]

\[
\geq \langle Dv, v \rangle_V - \sum_{m \in \mathcal{M}} \|D^{-1/2} R_m D^{-1/2}\|_{V \to V} \langle Dv, v \rangle_V \geq (1 - \gamma) \langle Dv, v \rangle_V ,
\]

where we used \( |y_m| \leq 1 \) and

\[
\langle R_m v, v \rangle_V = \langle D^{-1/2} R_m D^{-1/2} D^{1/2} v, D^{1/2} v \rangle_V \leq \|D^{-1/2} R_m D^{-1/2}\|_{V \to V} \langle Dv, v \rangle_V
\]

for all \( m \in \mathcal{M} \). In particular, the assumptions of Remark 1.1.9 are satisfied.
Chapter 1. Parametric and Stochastic Operator Equations

1.3. Tensor Product Construction

1.3.1. Expandable Operators

Let $X_1$, $X_2$, $Y_1$, and $Y_2$ be Banach spaces. We call a linear map on the algebraic tensor product spaces
\[ A : X_1 \otimes Y_1 \to X_2 \otimes Y_2 \] (1.3.1)
expandable if there are sequences $(S_k)_{k \in \mathbb{N}} \subset \mathcal{L}(X_1, X_2)$ and $(T_k)_{k \in \mathbb{N}} \subset \mathcal{L}(Y_1, Y_2)$ satisfying
\[ \sum_{k=1}^{\infty} \|S_k\|_{X_1 \to X_2} \|T_k\|_{Y_1 \to Y_2} < \infty \] (1.3.2)
such that
\[ A = \sum_{k=1}^{\infty} S_k \otimes T_k \] (1.3.3)
i.e.
\[ A(x \otimes y) = \sum_{k=1}^{\infty} (S_k x) \otimes (T_k y) \quad \forall x \in X_1, \quad \forall y \in Y_1. \] (1.3.4)
Note that the sum in (1.3.4) converges due to (1.3.2).

**Lemma 1.3.1.** If $A$ is expandable, then for any tensor norm $\alpha$, $A$ extends by continuity to the Banach tensor product $X_1 \otimes_\alpha Y_1$ as a map into $X_2 \otimes_\alpha Y_2$ and
\[ \|A\|_{X_1 \otimes_\alpha Y_1 \to X_2 \otimes_\alpha Y_2} \leq \sum_{k=1}^{\infty} \|S_k\|_{X_1 \to X_2} \|T_k\|_{Y_1 \to Y_2} \] (1.3.5)
for any expansion (1.3.3).

**Proof.** By triangle inequality and (A.2.9),
\[ \|A\|_{X_1 \otimes_\alpha Y_1 \to X_2 \otimes_\alpha Y_2} \leq \sum_{k=1}^{\infty} \|S_k \otimes T_k\|_{X_1 \otimes_\alpha Y_1 \to X_2 \otimes_\alpha Y_2} = \sum_{k=1}^{\infty} \|S_k\|_{X_1 \to X_2} \|T_k\|_{Y_1 \to Y_2}, \]
which is finite by (1.3.2). Therefore, $A$ is a continuous linear map from $(X_1 \otimes Y_1, \alpha_{X_1,Y_1})$ to $X_2 \otimes_\alpha Y_2$, and thus it extends to the Banach tensor product space $X_1 \otimes_\alpha Y_1$. \[ \square \]

Consequently, if $A$ is invertible on the algebraic tensor product spaces and both $A$ and $A^{-1}$ are expandable, then they extend to isomorphisms between $X_1 \otimes_\alpha Y_1$ and $X_2 \otimes_\alpha Y_2$ for any tensor norm $\alpha$. We consider a case in which only conditions on $A$ are required.

Let $A$ be an expandable perturbation of a tensor product operator,
\[ A = D + R \quad \text{with} \quad D = S \otimes T \quad \text{and} \quad R = \sum_{k=1}^{\infty} S_k \otimes T_k \] (1.3.6)
1.3. Tensor Product Construction

for $S \in \mathcal{L}(X_1, X_2)$ and $T \in \mathcal{L}(Y_1, Y_2)$ boundedly invertible, and sequences $(S_k)_{k \in \mathbb{N}} \subset \mathcal{L}(X_1, X_2)$ and $(T_k)_{k \in \mathbb{N}} \subset \mathcal{L}(Y_1, Y_2)$ satisfying

$$
\sum_{k=1}^{\infty} \|S^{-1}S_k\|_{X_1 \to X_1} \|T^{-1} T_k\|_{Y_1 \to Y_1} \leq \gamma < 1.
$$

(1.3.7)

By Lemma 1.3.1, using (A.2.27),

$$
\|\mathcal{D}^{-1} \mathcal{D}\|_{X_1 \otimes_\alpha Y_1 \to X_1 \otimes_\alpha Y_1} \leq \sum_{k=1}^{\infty} \|S^{-1}S_k\|_{X_1 \to X_1} \|T^{-1} T_k\|_{Y_1 \to Y_1} \leq \gamma < 1
$$

(1.3.8)

for any tensor norm $\alpha$.

**Theorem 1.3.2.** Let $\mathcal{A}$ as in (1.3.6) satisfy (1.3.7), and let $\alpha$ be a tensor norm. Then $\mathcal{A}$ is a boundedly invertible operator from $X_1 \otimes_\alpha Y_1$ to $X_2 \otimes_\alpha Y_2$ satisfying

$$
\|\mathcal{A}\|_{X_1 \otimes_\alpha Y_1 \to X_2 \otimes_\alpha Y_2} \leq (1 + \gamma) \|S\|_{X_1 \to X_2} \|T\|_{Y_1 \to Y_2},
$$

(1.3.9)

$$
\|\mathcal{A}^{-1}\|_{X_2 \otimes_\alpha Y_2 \to X_1 \otimes_\alpha Y_1} \leq \frac{1}{1 - \gamma} \|S^{-1}\|_{X_2 \to X_1} \|T^{-1}\|_{Y_2 \to Y_1}.
$$

(1.3.10)

Furthermore, both $\mathcal{A}$ and $\mathcal{A}^{-1}$ are expandable.

**Proof.** The operator $\mathcal{A}$ has the form

$$
\mathcal{A} = \mathcal{D} (\text{id}_{X_1 \otimes_\alpha Y_1} + \mathcal{D}^{-1} \mathcal{D}) = S \otimes T \left( \text{id}_{X_1} \otimes \text{id}_{Y_1} + \sum_{k=1}^{\infty} (S^{-1} S_k) \otimes (T^{-1} T_k) \right).
$$

It is expandable by definition. Due to (1.3.8), it can be inverted using a Neumann series, and its inverse is

$$
\mathcal{A}^{-1} = \left( \sum_{n=0}^{\infty} (-\mathcal{D}^{-1} \mathcal{D})^n \right) \mathcal{D}^{-1} = \sum_{n=0}^{\infty} \sum_{k=1}^{\infty} (-1)^n ((S^{-1} S_k)^n S^{-1}) \otimes ((T^{-1} T_k)^n T^{-1}).
$$

The operator $\mathcal{A}^{-1}$ is expandable since, using (1.3.7),

$$
\sum_{n=0}^{\infty} \sum_{k=1}^{\infty} \|S^{-1} S_k\|_{X_1 \to X_1} \|T^{-1} T_k\|_{Y_1 \to Y_1}^n \leq \sum_{n=0}^{\infty} \left( \sum_{k=1}^{\infty} \|S^{-1} S_k\|_{X_1 \to X_1} \|T^{-1} T_k\|_{Y_1 \to Y_1} \right)^n = \sum_{n=0}^{\infty} \gamma^n = \frac{1}{1 - \gamma} < \infty.
$$

Estimate (1.3.9) is a consequence of (1.3.8), and (1.3.10) follows from the Neumann series representation of $\mathcal{A}^{-1}$.  

\[\square\]
Chapter 1. Parametric and Stochastic Operator Equations

1.3.2. Application to Parametric Operator Equations

We consider parametric operators that depend affinely on the parameter, as in Section 1.2.2. For a countable index set $\mathcal{M}$, the parameter domain is the set $\Gamma = [-1, 1]^{\mathcal{M}}$, which is compact by Tychonoff’s theorem. As in (1.2.17),

$$A(y) = D + R(y), \quad R(y) := \sum_{m \in \mathcal{M}} y_m R_m \quad \forall y \in \Gamma,$$

(1.3.11)

with $D \in \mathcal{L}(V, W^*)$ boundedly invertible and $R_m \in \mathcal{L}(V, W^*)$, $m \in \mathcal{M}$, such that $\|R_m\|_{V \rightarrow W^*}$ is summable over $m \in \mathcal{M}$.

Let $\mathcal{A}$ be defined as in (1.1.12). Similarly, let

$$\mathcal{D} : C(\Gamma; V) \rightarrow C(\Gamma; W^*) \quad \text{and} \quad \mathcal{R} : C(\Gamma; V) \rightarrow C(\Gamma; W^*) \quad \forall \mathcal{A} \text{ defined as in (1.1.12)}.$$  

Identifying $C(\Gamma; X)$ with $C(\Gamma) \otimes X$, see Theorem A.3.3, the operators $\mathcal{D}$ and $\mathcal{R}$ have the form

$$\mathcal{D} = \text{id}_{C(\Gamma)} \otimes D \quad \text{and} \quad \mathcal{R} = \sum_{m \in \mathcal{M}} K_m \otimes R_m,$$

(1.3.14)

where $K_m$ is the multiplication operator for $y_m$.

$$K_m : C(\Gamma) \rightarrow C(\Gamma), \quad v(\gamma) \mapsto y_m v(\gamma).$$

(1.3.15)

**Lemma 1.3.3.** The operator $K_m$ has norm one on $C(\Gamma)$. If $\mu$ is a finite measure on $\mathcal{B}(\Gamma)$, then $K_m$ extends by continuity to $L^p(\mu)$ for all $1 \leq p < \infty$, and has norm at most one. If $\mu(\{ y \in \Gamma ; |y_m| \geq \epsilon \}) > 0$ for all $\epsilon > 0$, then $K_m$ has norm one. Furthermore $K_m$ is self-adjoint on $L^2(\mu)$.

**Proof.** On any of the spaces mentioned in the assertion, $K_m$ has norm at most one since $|y_m| \leq 1$. Since $y_m = 1$ is attained, this bound is sharp on $C(\Gamma)$. Furthermore, $K_m$ extends uniquely to $L^p(\mu)$ since $C(\Gamma)$ is a dense subspace by [Bau92, Satz 29.14] for $1 \leq p < \infty$. Applying $K_m$ to $v := \epsilon^{-1/p} 1_{|y_m| \geq \epsilon}$, one easily verifies that the bound is also sharp on $L^p(\mu)$, provided that the sets $\{ y \in \Gamma ; |y_m| \geq \epsilon \}$ have positive measure. Finally, $K_m = K_m^*$ since $y_m$ is real. \qed

The operator $\mathcal{A}$ is of the form (1.3.6) with $S = \text{id}$, $T = D$, $S_m = K_m$, $T_m = R_m$, $Y_1 = V$, $Y_2 = W^*$, and $X_1 = X_2 = C(\Gamma)$. By Lemma 1.3.3, condition (1.3.7) is equivalent to (1.2.19), i.e.

$$\sum_{m \in \mathcal{M}} \|D^{-1} R_m\|_{V \rightarrow V} \leq \gamma < 1.$$

(1.3.16)

If (1.3.16) holds, Theorem 1.3.2 for the injective tensor norm $\alpha = \iota$ leads to a version of Corollary 1.1.4 for this setting with alternative bounds on the norms of $\mathcal{A}$ and $\mathcal{A}^{-1}$.
1.3. Tensor Product Construction

Proposition 1.3.4. If (1.3.16) is satisfied, then $\mathcal{A}$ is a boundedly invertible linear operator from $C(\Gamma) \otimes V$ to $C(\Gamma) \otimes W^*$ with

$$
\|\mathcal{A}\|_{C(\Gamma) \otimes V \to C(\Gamma) \otimes W^*} \leq (1 + \gamma) \|D\|_{V \to W^*},
$$

(1.3.17)

$$
\|\mathcal{A}^{-1}\|_{C(\Gamma) \otimes W^* \to C(\Gamma) \otimes V} \leq \frac{1}{1 - \gamma} \|D^{-1}\|_{W^* \to V},
$$

(1.3.18)

for any tensor norm $\alpha$.

Proof. The assertion follows from Theorem 1.3.2 using Lemma 1.3.3 and (1.3.16). □

By Lemma 1.3.3, (1.3.16) implies (1.2.19) also for $X_1 = X_2 = L^p_\mu(\Gamma)$.

Theorem 1.3.5. Let $\mu$ be a finite measure on $\mathcal{B}(\Gamma)$, $1 \leq p < \infty$, and let $\alpha$ denote a tensor norm. Then under condition (1.3.16), $\mathcal{A}$ is a boundedly invertible operator from $L^p_\mu(\Gamma) \otimes V$ to $L^p_\mu(\Gamma) \otimes W^*$ with

$$
\|\mathcal{A}\|_{L^p_\mu(\Gamma) \otimes V \to L^p_\mu(\Gamma) \otimes W^*} \leq (1 + \gamma) \|D\|_{V \to W^*},
$$

(1.3.19)

$$
\|\mathcal{A}^{-1}\|_{L^p_\mu(\Gamma) \otimes W^* \to L^p_\mu(\Gamma) \otimes V} \leq \frac{1}{1 - \gamma} \|D^{-1}\|_{W^* \to V},
$$

(1.3.20)

Proof. By Lemma 1.3.3 and (1.3.16), $\mathcal{A}$ is a continuous linear map from $C(\Gamma) \otimes V$ to $C(\Gamma) \otimes W^*$ with respect to the norms of $L^p_\mu(\Gamma) \otimes V$ and $L^p_\mu(\Gamma) \otimes W^*$, so it extends by continuity to the latter spaces. Then the assertion follows from Theorem 1.3.2. □

If $V$ and $W^*$ are separable Hilbert spaces and $p = 2$, then the norms in Theorem 1.3.5 are identical to those of Theorem 1.1.6. For Banach spaces, the Lebesgue–Bochner spaces in the latter theorem do not have a tensor product structure. However, Theorem 1.3.5 also applies to Lebesgue–Pettis spaces $\tilde{L}^p_\mu(\Gamma; X)$, with norm

$$
\|f\|_{\tilde{L}^p_\mu(\Gamma; X)} := \sup_{\varphi \in B_X} \|\varphi \circ f\|_{L^p_\mu(\Gamma)}
$$

(1.3.21)

see Appendix A.3.2.

Corollary 1.3.6. Let $\mu$ be a finite measure on $\mathcal{B}(\Gamma)$ and $1 \leq p < \infty$. Then under condition (1.3.16), $\mathcal{A}$ is a boundedly invertible operator from $\tilde{L}_\mu^p(\Gamma; V)$ to $\tilde{L}_\mu^p(\Gamma; W^*)$ with

$$
\|\mathcal{A}\|_{\tilde{L}_\mu^p(\Gamma; V) \to \tilde{L}_\mu^p(\Gamma; W^*)} \leq (1 + \gamma) \|D\|_{V \to W^*},
$$

(1.3.22)

$$
\|\mathcal{A}^{-1}\|_{\tilde{L}_\mu^p(\Gamma; W^*) \to \tilde{L}_\mu^p(\Gamma; V)} \leq \frac{1}{1 - \gamma} \|D^{-1}\|_{W^* \to V},
$$

(1.3.23)

Proof. The assertion follows from Theorem 1.3.5 since $\tilde{L}_\mu^p(\Gamma; X)$ is isometrically isomorphic to the injective tensor product space $L^p_\mu(\Gamma) \otimes X$ for any Banach space $X$, see Appendix A.3.2. □
Chapter 1. Parametric and Stochastic Operator Equations

The Lebesgue–Pettis norm (1.3.21) is weaker than the Lebesgue–Bochner norm, so Corollary 1.3.6 can be interpreted as a weaker form of Theorem 1.1.6. It is reasonable to consider the Lebesgue–Pettis norm, since it uniformly controls the error in arbitrary functionals of the solution $u$ of (1.1.2). Consequently, in the stochastic setting, this norm uniformly controls the error in statistics of any continuous linear functional. Its main advantage over Lebesgue–Bochner spaces is that Lebesgue–Pettis spaces have a tensor product structure also if $V$ and $W$ are Banach spaces.
Chapter 2.

Countably Infinite Tensor Product Frames

Given orthonormal bases of the Lebesgue spaces of square integrable functions on two domains, it is straightforward to construct an orthonormal basis on the product domain by tensorizing the original basis functions. Repeating the process, bases can be constructed on arbitrarily high dimensional product domains.

On infinite dimensional domains, one must be careful to choose the appropriate generalization of a product basis, and an additional limit argument is required to show density of such bases. Dropping the orthonormality condition, it is possible to define products of Riesz bases. However, in the infinite dimensional setting, it is necessary to impose strong decay conditions on the Riesz constants.

Generalizing further, one may even consider frames, which are redundant bases. We construct infinite tensor products of frames in Section 2.2.

This construction is facilitated by the adoption of an abstract algebraic point of view, based primarily on the synthesis operator of a frame. This formalism, introduced in Section 2.1, provides a succinct theory of frames, yielding proofs of many elementary properties with minimal effort.

2.1. Frames and Riesz Bases of Hilbert Spaces

2.1.1. Frames and Riesz Bases

Let \( H \) be a separable Hilbert space over the scalar field \( K \in \{\mathbb{R}, \mathbb{C}\} \). A frame of \( H \) is a countable sequence \( \Phi := (\phi_v)_{v \in \Xi} \subset H \) for which the synthesis operator

\[
T_\Phi : \ell^2(\Xi) \to H, \quad c \mapsto \sum_{v \in \Xi} c_v \phi_v
\]

is bounded and surjective, see e.g. [Chr03, Thm. 5.5.1] or [Hol94, Thm. 2.1]. Of course, every continuous linear map \( \ell^2(\Xi) \to H \) is of the form (2.1.1) for some countable set \( \Phi \subset H \). The adjoint of the synthesis operator, called the analysis operator, is the injective linear map

\[
T_\Phi^* : H^* \to \ell^2(\Xi), \quad f \mapsto (f(\phi_v))_{v \in \Xi}.
\]
Chapter 2. Countably Infinite Tensor Product Frames

We identify $\ell^2(\Xi)$ with its dual through the Riesz isomorphism. Let $K_{\Phi} := \ker T_\Phi \subset \ell^2(\Xi)$. By the closed range theorem, we have the orthogonal decomposition

$$\ell^2(\Xi) = \text{range } T_\Phi \oplus K_{\Phi}. \tag{2.1.3}$$

Also, the open mapping theorem implies that the induced map

$$\hat{T}_\Phi : \ell^2(\Xi)/K_{\Phi} \to H \tag{2.1.4}$$

is an isomorphism. It satisfies $T_\Phi = \hat{T}_\Phi \pi_{\Phi}$, where $\pi_{\Phi} : \ell^2(\Xi) \to \ell^2(\Xi)/K_{\Phi}$ is the projection onto the quotient space.

Note that the restriction of $\pi_{\Phi}$ to $K_{\Phi}^+ = \text{range } T_\Phi^*$ is an isometric isomorphism from $K_{\Phi}^+ \subset \ell^2(\Xi)/K_{\Phi}$. Furthermore, the identification of $\ell^2(\Xi)$ with its dual carries over to $K_{\Phi}^+ \subset \ell^2(\Xi)$ by restriction, so $(K_{\Phi}^+)^* = K_{\Phi}^+$. Therefore,

$$\pi_{\Phi}^* \circ \pi_{\Phi} : (\ell^2(\Xi)/K_{\Phi})^* \to \ell^2(\Xi)/K_{\Phi} \tag{2.1.5}$$

is an isometric isomorphism, through which we identify the spaces $(\ell^2(\Xi)/K_{\Phi})^*$ and $\ell^2(\Xi)/K_{\Phi}$. Then the adjoint $\pi_{\Phi}^*$ of $\pi_{\Phi}$ is the embedding of $\ell^2(\Xi)/K_{\Phi}$ onto $K_{\Phi}^+ = \text{range } T_\Phi^*$ in $\ell^2(\Xi)$. Also, the adjoint of $T_\Phi$ is

$$T_\Phi^* = \pi_{\Phi}^* \hat{T}_\Phi^* : H^* \to \ell^2(\Xi)/K_{\Phi}. \tag{2.1.6}$$

The two values

$$b_\Phi := \|\hat{T}_\Phi^{-1}\|_{H \to \ell^2(\Xi)/K_{\Phi}} \quad \text{and} \quad B_\Phi := \|T_\Phi\|_{\ell^2(\Xi) \to H} = \|\hat{T}_\Phi\|_{\ell^2(\Xi)/K_{\Phi} \to H} \tag{2.1.7}$$

are called the frame bounds of $\Phi$. For all $f \in H^*$,

$$b_\Phi \|f\|_{H^*} \leq \left( \sum_{v \in \Xi} |f(\varphi_v)|^2 \right)^{1/2} \leq B_\Phi \|f\|_{H^*}. \tag{2.1.8}$$

The frame $\Phi$ is called tight if $b_\Phi = B_\Phi = 1$.

The frame $\Phi$ is a Riesz basis of $H$ if $K_{\Phi} = \{0\}$, i.e. if $T_\Phi = \hat{T}_\Phi$. In this case, the frame bounds are also called Riesz constants. Any tight Riesz basis is an orthonormal basis.

The frame operator is the self-adjoint linear map

$$S_{\Phi} := T_\Phi T_\Phi^* : H^* \to H, \quad f \mapsto \sum_{v \in \Xi} f(\varphi_v)\varphi_v. \tag{2.1.9}$$

By injectivity of $T_\Phi^*$ and (2.1.3), $S_{\Phi}$ is an isomorphism. In summary, we have the following commuting diagram

$$\begin{array}{ccc}
H & \xrightarrow{T_\Phi} & \ell^2(\Xi) \\
\downarrow{\pi_{\Phi}} & & \downarrow{\pi_{\Phi}} \\
\ell^2(\Xi)/K_{\Phi} & \xrightarrow{T_\Phi^*} & H \\
\end{array} \tag{2.1.10}
$$

\begin{footnotesize}
\footnote{In order for the Riesz isomorphism to be linear in the complex case, we define $H^*$ as the space of bounded antilinear maps from $H$ to $K$.}
\end{footnotesize}
and the bottom half of (2.1.10) collapses if $\Phi$ is a Riesz basis.

The sequence $\Phi^\ast := S^{-1}_\Phi \Phi$ is a frame of $H^\ast$, called the canonical dual frame. Its synthesis operator is $T_{\Phi^\ast} = S^{-1}_\Phi T_\Phi$ and, in particular, $K_{\Phi^\ast} = K_\Phi$. Since $S^{-1}_\Phi$ is self-adjoint, the adjoint of $T_{\Phi^\ast}$ is $T_{\Phi^\ast}^\ast = T_\Phi^\ast S^{-1}_\Phi$. Therefore, the frame operator of $\Phi^\ast$ is

$$S_{\Phi^\ast} = T_{\Phi^\ast} T_{\Phi^\ast}^\ast = S^{-1}_\Phi T_\Phi T_\Phi^\ast S^{-1}_\Phi = S^{-1}_\Phi S_\Phi S^{-1}_\Phi = S^{-1}_\Phi .$$

By definition,

$$T_{\Phi^\ast} T_{\Phi^\ast} = T_\Phi T_\Phi^\ast S^{-1}_\Phi = \text{id}_H \quad \text{and} \quad T_{\Phi^\ast}^\ast T_\Phi^\ast = S^{-1}_\Phi T_\Phi^\ast T_\Phi = \text{id}_{H^\ast} .$$

Denoting the elements of $\Phi^\ast$ by $q_v^\ast := S^{-1}_\Phi q_v$, this is equivalent to

$$w = \sum_{v \in \Xi} q_v^\ast(w)q_v \quad \forall w \in H \quad \text{and} \quad f = \sum_{v \in \Xi} f(q_v)q_v^\ast \quad \forall f \in H^\ast .$$

Note that the frame operator satisfies $S_{\Phi} = \tilde{T}_{\Phi} \tilde{T}_{\Phi^\ast}$, so

$$S_{\Phi^\ast} = S^{-1}_\Phi = (\tilde{T}_{\Phi^\ast})^{-1} \tilde{T}_{\Phi}^\ast .$$

This implies

$$\tilde{T}_{\Phi^\ast} = S^{-1}_\Phi \tilde{T}_\Phi = (\tilde{T}_\Phi)^{-1} \quad \text{and} \quad \tilde{T}_{\Phi^\ast}^\ast = \tilde{T}_{\Phi} .$$

In particular, the frame bounds of $\Phi^\ast$ are $b_{\Phi} = B^{-1}$ and $B_{\Phi^\ast} = b^{-1}$. Also,

$$T_{\Phi^\ast}^\ast T_\Phi = T_{\Phi}^\ast T_{\Phi^\ast} = P_{K_\Phi^\ast} : \ell^2(\Xi) \to K_\Phi^\ast \subset \ell^2(\Xi) ,$$

where $P_{K_\Phi^\ast}$ is the orthonormal projection onto $K_\Phi^\ast \subset \ell^2(\Xi)$, since by (2.1.15),

$$T_{\Phi^\ast} T_\Phi = \pi_{\Phi^\ast} \tilde{T}_{\Phi^\ast} \pi_\Phi = \pi_{\Phi^\ast} \pi_\Phi = P_{K_\Phi^\ast} ,$$

and similarly for $\Phi$ and $\Phi^\ast$ interchanged. Note that $\pi_\Phi^\ast = \pi_\Phi$ since $K_{\Phi^\ast} = K_\Phi$.

We refer to [Chr03, Chr08] for further details and an in-depth discussion of frames.

### 2.1.2. Vector Frames

Let $\Phi$ be a frame of a separable Hilbert space $H$ over $K$ as in Section 2.1.1. Furthermore, let $V$ be a Banach space, and $\alpha$ a tensor norm. We consider the tensor product maps

$$T_\Phi^V := T_\Phi \otimes \text{id}_V : \ell^2(\Xi) \otimes_{\alpha} V \to H \otimes_{\alpha} V ,$$

$$\tilde{T}_\Phi^V := \tilde{T}_\Phi \otimes \text{id}_V : (\ell^2(\Xi) / K_\Phi) \otimes_{\alpha} V \to H \otimes_{\alpha} V .$$

By Proposition A.2.4, $T_\Phi^V$ has dense range and $\tilde{T}_\Phi^V$ is boundedly invertible. Furthermore, due to (A.2.9),

$$\| T_\Phi^V \|_{C(\Xi) \otimes_{V} V \to H \otimes_{\alpha} V} = \| \tilde{T}_\Phi^V \|_{(\ell^2(\Xi) / K_\Phi) \otimes_{\alpha} V \to H \otimes_{\alpha} V} = B_{\Phi} ,$$

$$\| (\tilde{T}_\Phi^V)^{-1} \|_{H \otimes_{\alpha} V \to (\ell^2(\Xi) / K_\Phi) \otimes_{\alpha} V} = b_{\Phi} .$$
Chapter 2. Countably Infinite Tensor Product Frames

We also define formal adjoints of \( T^V_\Phi \) and \( \check{T}^V_\Phi \):
\[
T^V_\Phi := T^{\check{\Phi}}_\Phi \otimes \text{id}_V : H^* \otimes_\alpha V^* \to \ell^2(\Xi) \otimes_\alpha V^*, \tag{2.1.21}
\]
\[
\check{T}^V_\Phi := \check{T}^{\Phi}_\Phi \otimes \text{id}_V : H^* \otimes_\alpha V^* \to (\ell^2(\Xi)/K_\Phi) \otimes_\alpha V^*. \tag{2.1.22}
\]

Proposition A.2.4 implies that \( \check{T}^V_\Phi \) is boundedly invertible, and both operators have norm \( B_\Phi \).

**Theorem 2.1.1.** For any tensor norm \( \alpha \), \( T^V_\Phi \) is surjective and \( T^V_\Phi \) is injective. Furthermore, the kernel of \( T^V_\Phi \) is the closure of \( K_\Phi \otimes V \) in \( \ell^2(\Xi) \otimes_\alpha V \), and the range of \( T^V_\Phi \) is \( K^+_\Phi \otimes_\alpha V^* \), which embeds continuously into \( \ell^2(\Xi) \otimes_\alpha V^* \).

**Proof.** Due to (2.1.12), using (A.2.27),
\[
(T^{\check{\Phi}}_\Phi \otimes \text{id}_V)(T^{\Phi}_\Phi \otimes \text{id}_V) = (T^{\check{\Phi}}_\Phi T^{\Phi}_\Phi) \otimes \text{id}_V = \text{id}_H \otimes \text{id}_V = \text{id}_{H \otimes_\alpha V},
\]
which is surjective, so in particular \( T^V_\Phi = T^{\check{\Phi}}_\Phi \otimes \text{id}_V \) is surjective. Similarly, interchanging \( \Phi \) and \( \check{\Phi} \),
\[
(T^{\check{\Phi}}_\Phi \otimes \text{id}_V^*)T^{\Phi}_\Phi \otimes \text{id}_V^*) = (T^{\check{\Phi}}_\Phi T^{\Phi}_\Phi) \otimes \text{id}_V^* = \text{id}_H^* \otimes \text{id}_V^* = \text{id}_{H^* \otimes_\alpha V},
\]
which is injective. Thus also \( T^V_\Phi = T^{\check{\Phi}}_\Phi \otimes \text{id}_V^* \) is injective.

Since \( T^{\Phi}_\Phi \) is an isomorphism of \( H^* \) onto \( K^+_\Phi \subset \ell^2(\Xi) \), Proposition A.2.4 implies that \( T^{\check{\Phi}}_\Phi \otimes \text{id}_V \) is an isomorphism of \( H^* \otimes_\alpha V^* \) onto \( K^+_\Phi \otimes_\alpha V^* \). As this is also an injective map into \( \ell^2(\Xi) \otimes_\alpha V^* \), \( K^+_\Phi \otimes_\alpha V^* \) embeds into \( \ell^2(\Xi) \otimes_\alpha V^* \).

Let \( K^V_\Phi \) denote the closure of \( K_\Phi \otimes V \) in \( \ell^2(\Xi) \otimes_\alpha V \). It is clear that \( K_\Phi \otimes V \subset \ker T^V_\Phi \), and since the latter is closed, it follows that \( K^V_\Phi \subset \ker T^V_\Phi \).

Due to (2.1.16),
\[
(T^{\check{\Phi}}_\Phi \otimes \text{id}_V^*)T^{\Phi}_\Phi \otimes \text{id}_V^*) = (T^{\check{\Phi}}_\Phi T^{\Phi}_\Phi) \otimes \text{id}_V^* = P_{K^V_\Phi} \otimes \text{id}_V^*,
\]
where \( P_{K^V_\Phi} \) is the orthonormal projection of \( \ell^2(\Xi) \) onto \( K^V_\Phi \subset \ell^2(\Xi) \). By injectivity of \( T^{\check{\Phi}}_\Phi \otimes \text{id}_V^* \), it follows that
\[
\ker T^V_\Phi = \ker(T^{\check{\Phi}}_\Phi \otimes \text{id}_V) = \ker(P_{K^V_\Phi} \otimes \text{id}_V). 
\]

Let \( v \in \ell^2(\Xi) \otimes_\alpha V \setminus K^V_\Phi \). Since \( K^V_\Phi \) is closed, the Hahn–Banach theorem implies that there is a continuous linear functional \( \psi_0 \) on \( \ell^2(\Xi) \otimes_\alpha V \) that vanishes on \( K^V_\Phi \), but \( \psi_0(v) \neq 0 \). For all \( c \in \ell^2(\Xi) \) and all \( v \in V \),
\[
\psi_0(c \otimes v) = \psi_0((P_{K^V_\Phi} c) \otimes v) + \psi_0((P_{K^V_\Phi} c) \otimes v) = \psi_0((P_{K^V_\Phi} c) \otimes v) = \psi_0((P_{K^V_\Phi} c) \otimes v),
\]
where \( P_{K^V_\Phi} = \text{id}_{\ell^2(\Xi)} - P_{K^V_\Phi} \) is the orthogonal projection onto \( K^V_\Phi \). It follows by linearity and continuity that
\[
\psi_0 = \psi_0 \circ (P_{K^V_\Phi} \otimes \text{id}_V). 
\]
In particular \( \psi_0(v) \neq 0 \) implies \( v \notin \ker(P_{K^V_\Phi} \otimes \text{id}_V) \) and thus \( v \notin \ker T^V_\Phi \). Since \( v \in \ell^2(\Xi) \otimes_\alpha V \setminus K^V_\Phi \) was arbitrary, \( \ker T^V_\Phi \subset K^V_\Phi \). \( \Box \)
Chapter 2. Frames and Riesz Bases of Hilbert Spaces

Theorem 2.1.1 should be contrasted with Propositions A.2.5 and A.2.6, which, for general Banach spaces and operators, require additional assumptions on the tensor norm \( a \) to deduce injectivity and surjectivity of tensor product operators.

By Theorem A.3.5, the tensor product spaces \( \ell^2(\Xi) \otimes_\alpha V \) and \( \ell^2(\Xi) \otimes_\alpha V^* \) can be interpreted as spaces of sequences in \( V \) and \( V^* \), respectively, indexed by \( \Xi \). The generalized synthesis and analysis operators behave in much the same way as in the scalar setting.

**Proposition 2.1.2.** For all \( v = (v_\nu)_{\nu \in \Xi} \in \ell^2(\Xi) \otimes_\alpha V \),

\[
T^V_{\phi} v = \sum_{\nu \in \Xi} \phi_\nu \otimes v_\nu
\]

(2.1.23)

with unconditional convergence in \( H \otimes_\alpha V \). For all \( g \in H^* \otimes_\alpha V^* \),

\[
T^V_{\phi} g = ((\phi_\nu \otimes \text{id}_{V^*})g)_{\nu \in \Xi},
\]

(2.1.24)

where \( \phi_\nu \) is interpreted in the bidual space \( H^{**} \).

**Proof.** Let \( v = c \otimes v \) for \( v \in V \) and \( c = (c_\nu)_{\nu \in \Xi} \), \( c_\nu = \delta_{\mu \nu} \) for a \( \mu \in \Xi \), i.e. \( v_\mu = v \) and \( v_\nu = 0 \) for all \( \nu \neq \mu \). Then by (2.1.17) and (2.1.1),

\[
T^V_{\phi} v = (T^V_{\phi} c) \otimes v = \sum_{\nu \in \Xi} \phi_\nu \otimes v_\nu.
\]

It follows by linearity that (2.1.23) holds if \( v_\nu = 0 \) for all but finitely many \( \nu \in \Xi \). For general \( v \in \ell^2(\Xi) \otimes_\alpha V \), let \( \Xi_n \subset \Xi \) for all \( n \in \mathbb{N} \) such that \( \Xi_n \uparrow \Xi \) and \( \Xi_n \) is finite. Furthermore, let \( v^n = (v^n_\nu)_{\nu \in \Xi} \) be given by \( v^n_\nu := v_\nu \) if \( \nu \in \Xi_n \) and \( v^n_\nu := 0 \) if \( \nu \in \Xi \setminus \Xi_n \). Then \( v^n \in \ell^2(\Xi) \otimes V \) and by Theorem A.3.6, \( v^n \to v \) in \( \ell^2(\Xi) \otimes_\alpha V \). Therefore, by continuity of \( T^V_{\phi} \),

\[
T^V_{\phi} v = \lim_{n \to \infty} T^V_{\phi} v^n = \lim_{n \to \infty} \sum_{\nu \in \Xi_n} \phi_\nu \otimes v_\nu
\]

with convergence in \( H \otimes_\alpha V \).

Let \( g = f \otimes g \) for \( f \in H^* \) and \( h \in V^* \). Then by (2.1.21) and (2.1.2),

\[
T^V_{\phi} g = (T^V_{\phi} f) \otimes h = (f(\phi_\nu))_{\nu \in \Xi} \otimes h = (f(\phi_\nu)h)_{\nu \in \Xi},
\]

where the last equality is due to the embedding of \( \ell^2(\Xi) \otimes_\alpha V^* \) into \( c_0(\Xi; V^*) \). Since \( f(\phi_\nu)h = (\phi_\nu \otimes \text{id}_{V^*})(f \otimes h) \), (2.1.24) holds for simple tensors, and it follows by linearity that it holds on the algebraic tensor product \( H^* \otimes V^* \). For every \( v \in \Xi \), \( \phi_\nu \otimes \text{id}_{V^*} \) is a continuous linear map from \( H^* \otimes_\alpha V^* \) to \( V^* \). Therefore, (2.1.24) extends by continuity to \( H^* \otimes_\alpha V^* \). \( \square \)

\footnote{We follow the convention that the dual space \( H^* \) is the space of bounded antilinear functionals on \( H \), but the bidual \( H^{**} \) is the space of linear functionals on \( H \).}
Chapter 2. Countably Infinite Tensor Product Frames

It follows from the respective definitions that

\[ T^V_{\Phi} = \hat{T}_{\Phi} V^V \quad \text{and} \quad T^V_{\Phi} = \pi^V_{\Phi} \hat{T}^V_{\Phi} \tag{2.1.25} \]

for \( \pi^V_{\Phi} := \pi_{\Phi} \otimes \text{id}_V \) and \( \pi^V_{\Phi} := \pi_{\Phi}^* \otimes \text{id}_{V^*} \). Due to Theorem 2.1.1, \( \pi^V_{\Phi} \) is an embedding of \( (\ell^2(\mathbb{Z})/K_{\Phi}) \otimes_{\alpha} V^* \) into \( \ell^2(\mathbb{Z}) \otimes_{a} V^* \), with range \( K_{\Phi} \otimes_{a} V^* \). If \( \alpha \) is an injective tensor norm, then \( K_{\Phi} \otimes_{a} V^* \) is a closed subspace of \( \ell^2(\mathbb{Z}) \otimes_{a} V^* \). Furthermore, injectivity of \( \alpha \) implies that \( \ker T^V_{\Phi} = K_{\Phi} \otimes_{a} V \), seen as a closed subspace of \( \ell^2(\mathbb{Z}) \otimes_{a} V \). If \( \alpha \) is a projective tensor norm, then \( \pi^V_{\Phi} \) is a quotient map from \( \ell^2(\mathbb{Z}) \otimes_{a} V \) to \( (\ell^2(\mathbb{Z})/K_{\Phi}) \otimes_{a} V \).

We assume that \( V \) is a separable Hilbert space, and \( \alpha \) is the Hilbert tensor norm, see Appendix A.2.5. We identify \( \ell^2(\mathbb{Z}) \otimes_{\eta} V \) with \( \ell^2(\mathbb{Z}; V) \), and its dual \( \ell^2(\mathbb{Z}; V)^* = (\ell^2(\mathbb{Z}) \otimes_{\eta} V)^* \) with the spaces \( \ell^2(\mathbb{Z}; V^*) \) and \( \ell^2(\mathbb{Z}) \otimes_{\eta} V^* \). Then the generalized synthesis and analysis operators have the form

\[ T^V_{\Phi} = T_{\Phi} \otimes \text{id}_V : \ell^2(\mathbb{Z}; V) \to H \otimes_{\eta} V , \tag{2.1.26} \]

\[ T^V_{\Phi} = T_{\Phi}^* \otimes \text{id}_{V^*} : (H \otimes_{\eta} V)^* \to \ell^2(\mathbb{Z}; V^*) . \tag{2.1.27} \]

By Theorem 2.1.1, the kernel of \( T^V_{\Phi} \) is \( \ker T^V_{\Phi} = K_{\Phi} \otimes_{\eta} V \) and \( T^V_{\Phi} \) induces an isomorphism

\[ \hat{T}^V_{\Phi} = \hat{T}_{\Phi} \otimes \text{id}_V : \ell^2(\mathbb{Z}; V)/(K_{\Phi} \otimes_{\eta} V) \to H \otimes_{\eta} V , \tag{2.1.28} \]

which coincides with (2.1.18) due to (2.1.25). Due to its tensor product structure, \( (\hat{T}^V_{\Phi})^* = \hat{T}^{V^*}_{\Phi} \).

**Lemma 2.1.3.** Under the identification \( V^{**} = V \),

\[ (\hat{T}^V_{\Phi})^{-1} = (\hat{T}^{V^*}_{\Phi})^* \quad \text{and} \quad ((\hat{T}^{V^*}_{\Phi})^*)^{-1} = \hat{T}^V_{\Phi} . \tag{2.1.29} \]

**Proof:**

\[ (\hat{T}^V_{\Phi})^{-1} = (\hat{T}_{\Phi} \otimes \text{id}_V)^{-1} = \hat{T}_{\Phi}^* \otimes \text{id}_{V^*} = (\hat{T}_{\Phi}^* \otimes \text{id}_{V^*})^* = (\hat{T}^{V^*}_{\Phi})^* , \]

\[ ((\hat{T}^{V^*}_{\Phi})^*)^{-1} = (\hat{T}_{\Phi}^* \otimes \text{id}_{V^*})^{-1} = (\hat{T}_{\Phi}^*)^{-1} \otimes \text{id}_{V^*} = \hat{T}_{\Phi} \otimes \text{id}_V = \hat{T}^V_{\Phi} . \]

\[ \square \]

### 2.1.3. Tensor Product Frames

In the setting of Section 2.1.2, let \( V \) be a separable Hilbert space and let \( \Psi = (\psi_{\lambda})_{\lambda \in \Theta} \) be a frame of \( V \). Define the countable sequence \( \Phi \times \Psi := (\varphi_{v} \otimes \psi_{\lambda})_{(v, \lambda) \in \Xi \times \Theta} \) and the corresponding synthesis operator

\[ T_{\Phi \times \Psi} : \ell^2(\Xi \times \Theta) \to H \otimes \text{V} , \quad c \mapsto \sum_{v \in \Xi, \lambda \in \Theta} c_{v, \lambda} \varphi_{v} \otimes \psi_{\lambda} . \tag{2.1.30} \]

Let \( K_{\Phi \times \Psi} := \ker T_{\Phi \times \Psi} \); then \( T_{\Phi \times \Psi} \) induces a map \( \hat{T}^V_{\Phi \times \Psi} \) on \( \ell^2(\Xi \times \Theta)/K_{\Phi \times \Psi} \) as in Section 2.1.1.
2.1. Frames and Riesz Bases of Hilbert Spaces

**Theorem 2.1.4.** The map \( T_{\Phi,\Psi} \) is surjective and has the form \( T_{\Phi,\Psi} = T_{\Phi} \otimes T_{\Psi} \). It can be decomposed as

\[
T_{\Phi,\Psi} = T_{\Phi}^{\Psi} T_{\Psi}^{\Psi} = T_{\Phi}^H T_{\Psi}^H
\]

(2.1.31)

and its kernel is

\[
K_{\Phi,\Psi} = (K_{\Phi} \otimes \eta \ell^2(\Theta)) + (\ell^2(\Xi) \otimes \eta K_{\Psi}),
\]

(2.1.32)

where \( K_{\Phi} = \ker T_{\Phi} \subset \ell^2(\Xi) \) and \( K_{\Psi} = \ker T_{\Psi} \subset \ell^2(\Theta) \). The induced map on the quotient space has the form \( \tilde{T}_{\Phi,\Psi} = \tilde{T}_{\Phi} \otimes \tilde{T}_{\Psi} \). Furthermore,

\[
B_{\Phi,\Psi} = B_{\Phi} B_{\Psi}, \quad (2.1.33)
\]

(2.1.34)

**Proof.** By continuity of the tensor product, for all \( c \in \ell^2(\Xi) \) and \( d \in \ell^2(\Theta) \),

\[
(T_{\Phi} \otimes T_{\Psi})(c \otimes d) = \left( \sum_{\nu \in \Xi} c_{\nu} \varphi_{\nu} \right) \otimes \left( \sum_{\lambda \in \Theta} d_{\lambda} \psi_{\lambda} \right) = \sum_{\nu, \lambda \in \Xi, \Theta} c_{\nu} d_{\lambda} \varphi_{\nu} \otimes \psi_{\lambda}.
\]

The last term corresponds to (2.1.30). This implies surjectivity of \( T_{\Phi,\Psi} \) since surjectivity of \( T_{\Phi} \otimes T_{\Psi} \) follows from the surjectivity of \( T_{\Phi} \) and \( T_{\Psi} \) by Proposition A.2.6. Also,

\[
T_{\Phi,\Psi} = (T_{\Phi} \otimes \text{id}_V)(\text{id}_{\ell^2(\Xi)} \otimes T_{\Psi}) = T_{\Phi}^V T_{\Psi}
\]

and similarly for the other decomposition in (2.1.31). Since \( \tilde{T}_{\Phi} \) and \( \tilde{T}_{\Psi} \) are isomorphisms, their tensor product

\[
\tilde{T}_{\Phi} \otimes \tilde{T}_{\Psi} : \left( \ell^2(\Xi) / K_{\Phi} \right) \otimes \eta \left( \ell^2(\Theta) / K_{\Psi} \right) \rightarrow H \otimes \eta V
\]

is an isomorphism by Proposition A.2.4. Therefore, \( (\tilde{T}_{\Phi} \otimes \tilde{T}_{\Psi})^{-1} \circ T_{\Phi,\Psi} \) induces an isomorphism

\[
\ell^2(\Xi \times \Theta) / K_{\Phi,\Psi} \rightarrow \left( \ell^2(\Xi) / K_{\Phi} \right) \otimes \eta \left( \ell^2(\Theta) / K_{\Psi} \right),
\]

through which we identify these two spaces. In particular, \( \tilde{T}_{\Phi,\Psi} = \tilde{T}_{\Phi} \otimes \tilde{T}_{\Psi} \), and properties (2.1.33) and (2.1.34) follow from \( (\tilde{T}_{\Phi} \otimes \tilde{T}_{\Psi})^{-1} = \tilde{T}_{\Phi}^{-1} \otimes \tilde{T}_{\Psi}^{-1} \) since the operator norm of the tensor product of two operators is the product of the norms of these operators.

Finally, by (2.1.3),

\[
K_{\Phi,\Psi} = \text{range} T_{\Phi,\Psi} = \text{range} T_{\Phi} \otimes T_{\Psi} = K_{\Phi} \otimes \eta K_{\Psi}
\]

\[
= \left( K_{\Phi} \otimes \eta \ell^2(\Theta) \right) ^\perp \cap \left( \ell^2(\Xi) \otimes \eta K_{\Psi} \right) ^\perp.
\]

Equation (2.1.32) follows since \( (K_{\Phi} \otimes \eta \ell^2(\Theta)) + (\ell^2(\Xi) \otimes \eta K_{\Psi}) \) is closed by [SST01]. \( \square \)
Chapter 2. Countably Infinite Tensor Product Frames

Theorem 2.1.4 states that $\Phi \times \Psi$ is a frame of $H \otimes \eta V$. By (2.1.32), $K_{\Phi \times \Psi} = [0]$ if and only if $K_\Phi = [0]$ and $K_\Psi = [0]$, so $\Phi \times \Psi$ is a Riesz basis of $H \otimes \eta V$ if and only if both $\Phi$ and $\Psi$ are Riesz bases.

It is useful to note that the dual of a product frame is the product of the dual frames. The precise formulation of this statement requires the identification of $(H \otimes \eta V)^*$ with $H^* \otimes \eta V^*$.

Corollary 2.1.5. The canonical dual frame of $\Phi \times \Psi$ is

$$ (\Phi \times \Psi)^* = \Phi^* \times \Psi^* = (\varphi_{\eta}^* \otimes \psi_{\lambda}^*)(\nu, \lambda) \in \Xi \times \Theta . $$

Proof. The frame operator of $\Phi \times \Psi$ is

$$ S_{\Phi \times \Psi} = (T_{\Phi} \otimes T_\Psi)(T_{\Phi} \otimes T_\Psi)^* = (T_{\Phi} T_{\Phi}^*) \otimes (T_{\Psi} T_{\Psi}^*) = S_\Phi \otimes S_\Psi . $$

Therefore, $S_{\Phi \times \Psi}^{-1} = S_\Phi^{-1} \otimes S_\Psi^{-1}$ and

$$ (\Phi \times \Psi)^* = S_{\Phi \times \Psi}^{-1}(\Phi \times \Psi) = (S_\Phi^{-1} \varphi_{\nu} \otimes S_\Psi^{-1} \psi_\lambda)(\nu, \lambda) \in \Xi \times \Theta . $$

2.1.4. Hierarchic Frames

Given a frame $\Phi = (\varphi_{\nu})_{\nu \in \Xi}$ of $H$, it is often useful to introduce a hierarchic structure on $\Xi$ as a way of encoding which indices are, in some sense, more important than which others. If $H$ is a function space, an index $\nu' \in \Xi$ may be considered more important than $\nu \in \Xi$ e.g. if $\varphi_{\nu'}$ has a coarser scale than $\varphi_{\nu}$, or if, in addition, the supports of $\varphi_{\nu'}$ and $\varphi_{\nu}$ overlap. Such a relation defines a partial order on $\Xi$.

A boundedly hierarchic frame of $H$ is a tuple $(\Phi, \prec)$, where $\Phi$ is a frame of $H$ with index set $\Xi$, and $\prec$ is a strict partial order on $\Xi$ with finitely many minimal elements and such that

$$ \# \{ \nu' \in \Xi ; \nu' \prec \nu \} < \infty \quad \forall \nu \in \Xi , $$

i.e. each index has only finitely many predecessors. We use the notation $\nu' \preceq \nu$ for $\nu' \prec \nu$ or $\nu' = \nu$.

In numerical approximations, elements of $H$ are approximated by expansions in $\varphi_{\nu}$ for $\nu$ in a finite subset of $\Xi$. Since $\nu'$ should be included in such a finite subset before $\nu$, (2.1.37) ensures that each $\nu \in \Xi$ can be reached by a finite refinement of any initial subset of $\Xi$.

Example 2.1.6. A graded frame of $H$ is a frame $\Phi$ with index set $\Xi$ and a map $\ell : \Xi \to \mathbb{N}_0$ assigning to each index $\nu \in \Xi$ a level $\ell(\nu) \in \mathbb{N}_0$. Such a grading function defines a partial order on $\Xi$ by

$$ \nu' \prec \nu \quad \Leftrightarrow \quad \ell(\nu') < \ell(\nu) , \quad \nu, \nu' \in \Xi . $$

If only finitely many indices are mapped to each $n \in \mathbb{N}_0$, then $(\Phi, \prec)$ is a boundedly hierarchic frame. In this case, we call $\Phi$ finitely graded. \[ \Box \]
2.2. Frames on Infinite Dimensional Domains

**Proposition 2.1.7.** Let \( \Phi = (\varphi_\nu)_{\nu \in \Xi} \) and \( \Psi = (\psi_\lambda)_{\lambda \in \Theta} \) be boundedly hierarchic frames of \( H \) and \( V \), respectively, and define a partial order on \( \Xi \times \Theta \) by

\[(v', \lambda') \preceq (v, \lambda) :\iff v' \preceq v \land \lambda' \preceq \lambda, \quad (2.1.39)\]

and \((v', \lambda') < (v, \lambda)\) if \((v', \lambda') \preceq (v, \lambda)\) and \((v', \lambda') \neq (v, \lambda)\). Then \((\Phi \times \Psi, <)\) is a boundedly hierarchic frame of \( H \otimes V \).

**Proof.** By Theorem 2.1.4, \( \Phi \times \Psi \) is a frame of \( H \otimes V \) with index set \( \Xi \times \Theta \). It is clear that \(<\) is a strict partial order on \( \Xi \times \Theta \). An index \((v, \lambda) \in \Xi \times \Theta\) is minimal if and only if \( v \) is minimal in \( \Xi \) and \( \lambda \) is minimal in \( \Theta \). By definition, \((v', \lambda') < (v, \lambda)\) only if \( v' \preceq v \) and \( \lambda' \preceq \lambda \). Therefore, the finiteness conditions are satisfied. \( \square \)

**Remark 2.1.8.** The partial order \(<\) given by \((2.1.39)\) is the unique partial order on \( \Xi \times \Theta \) that is compatible with traces, i.e. for any \( \lambda \in \Theta \), \((v', \lambda) < (v, \lambda)\) if and only if \( v' < v \), and similarly with reversed roles.

**Remark 2.1.9.** If \( \Phi \) and \( \Psi \) are graded frames, the product \( \Phi \times \Psi \) is not naturally graded. The product of the grading functions is a map

\[\Xi \times \Theta \to \mathbb{N}_0^2, \quad (v, \lambda) \mapsto (\ell(v), \ell(\lambda)).\quad (2.1.40)\]

To define a grading function for \( \Phi \times \Psi \), one needs to compose \((2.1.40)\) with a suitable map \( \mathbb{N}_0^2 \to \mathbb{N}_0 \). For example, one might define the grading function as

\[\ell(v) + \ell(\lambda), \quad \ell(v)^2 + \ell(\lambda)^2 \quad \text{or} \quad \max(\ell(v), \ell(\lambda)), \quad (v, \lambda) \in \Xi \times \Theta. \quad (2.1.41)\]

However, each of the choices \((2.1.41)\) induces a different partial order \((2.1.38)\) on \( \Xi \times \Theta \).

### 2.2. Frames on Infinite Dimensional Domains

#### 2.2.1. Product Domains

Let \( \mathcal{M} \) be an arbitrary index set. For all \( m \in \mathcal{M} \), let \((\Gamma_m, \Sigma_m)\) be a measurable space, and let \( \pi_m \) be a probability measure on \((\Gamma_m, \Sigma_m)\). Define the product space

\[(\Gamma, \Sigma, \pi) := \bigotimes_{m \in \mathcal{M}} (\Gamma_m, \Sigma_m, \pi_m), \quad (2.2.1)\]

which is again a probability space.\(^3\)

**Example 2.2.1.** In Section 1.2, we consider products of topological spaces \( \Gamma_m \) with Borel \( \sigma \)-algebras \( \Sigma_m := \mathcal{B}(\Gamma_m) \). If \( \mathcal{M} \) is countable, then \( \Sigma = \mathcal{B}(\Gamma) \) is the Borel \( \sigma \)-algebra on the product domain \( \Gamma \).

\(^3\)See e.g. [Bau02, Section 9] for a general construction of arbitrary products of probability spaces.
Chapter 2. Countably Infinite Tensor Product Frames

Let $\mathcal{P}(\mathcal{M})$ denote the power set of $\mathcal{M}$, and $\mathcal{F}(\mathcal{M}) \subset \mathcal{P}(\mathcal{M})$ the lattice of finite subsets of $\mathcal{M}$. Define the finite products

\[(\Gamma_I, \Sigma_I, \pi_I) := \bigotimes_{n \in I}(\Gamma_n, \Sigma_n, \pi_n), \quad I \in \mathcal{F}(\mathcal{M}). \tag{2.2.2}\]

The coordinate maps

\[y_I: (\Gamma, \Sigma) \to (\Gamma_I, \Sigma_I), \quad I \in \mathcal{F}(\mathcal{M}), \tag{2.2.3}\]

are measurable. They generate $\sigma$-algebras on $\Gamma$, which, abusing notation, we denote by $\Sigma_I := \sigma(y_I)$. Note that, by definition,

\[\Sigma = \sigma(y_I; I \in \mathcal{F}(\mathcal{M})) = \sigma(\Sigma_I; I \in \mathcal{F}(\mathcal{M})). \tag{2.2.4}\]

We denote by $L^2_{\pi|\Sigma_I}(\Gamma)$ the space of $\Sigma_I$-measurable elements of $L^2_{\pi}(\Gamma)$.

**Lemma 2.2.2.** For all $I \in \mathcal{F}(\mathcal{M})$, the map

\[L^2_{\pi_I}(\Gamma_I) \to L^2_{\pi}(\Gamma), \quad v \mapsto v \circ y_I, \tag{2.2.5}\]

is an isometry with range $L^2_{\pi|\Sigma_I}(\Gamma)$.

**Proof.** Let $I \in \mathcal{F}(\mathcal{M})$ and $v \in L^2_{\pi_I}(\Gamma_I)$. Then $v \circ y_I \in L^2_{\pi}(\Gamma)$ and $v \circ y_I$ is $\Sigma_I = \sigma(y_I)$-measurable, so $v \circ y_I \in L^2_{\pi|\Sigma_I}(\Gamma)$. Conversely, let $w \in L^2_{\pi|\Sigma_I}(\Gamma)$. By the Doob–Dynkin lemma, there is a measurable function $v$ on $(\Gamma_I, \Sigma_I)$ such that $w = v \circ y_I$. Furthermore, since $\pi_I = y_I(\pi)$,

\[\int_{\Gamma_I} |v|^2 \, d\pi_I = \int_{\Gamma} |v \circ y_I|^2 \, d\pi = \int_{\Gamma} |w|^2 \, d\pi.\]

This shows $v \in L^2_{\pi_I}(\Gamma_I)$ and that the map is an isometry. \(\square\)

In particular, Lemma 2.2.2 embeds the spaces $L^2_{\pi_m}(\Gamma_m)$ into $L^2_{\pi}(\Gamma)$ for all $m \in \mathcal{M}$. We abbreviate $y_m := y_{[m]}$ and $\Sigma_m := \Sigma_{[m]}$ for $m \in \mathcal{M}$.

We recall the monotone class theorem, see for example [Pro05, Theorem I.8]. A set $\mathcal{M}$ of real-valued functions on $\Gamma$ is multiplicative if $v, w \in \mathcal{M}$ implies $vw \in \mathcal{M}$. A monotone vector space over $\Gamma$ is a real vector space $\mathcal{S}$ of bounded, real-valued functions on $\Gamma$ such that all constants are in $\mathcal{S}$ and if $(v_n)_{n \in \mathbb{N}}$ is a sequence in $\mathcal{S}$ with $0 \leq v_n \leq v_{n+1}$ for all $n \in \mathbb{N}$ and $v := \sup_n v_n$ is a bounded function on $\Gamma$, then $v \in \mathcal{S}$.

**Theorem 2.2.3 (Monotone Class Theorem).** Let $\mathcal{M}$ be a multiplicative class of bounded, real-valued functions on $\Gamma$, and let $\mathcal{S}$ be a monotone vector space containing $\mathcal{M}$. Then $\mathcal{S}$ contains all bounded $\sigma(\mathcal{M})$-measurable functions.

**Proposition 2.2.4.** $\bigcup_{I \in \mathcal{F}(\mathcal{M})} L^2_{\pi|\Sigma_I}(\Gamma)$ is a dense subspace of $L^2_{\pi}(\Gamma)$.
2.2. Frames on Infinite Dimensional Domains

Proof. Let $\mathcal{B} := \bigcup_{I \in \mathcal{F}(\mathcal{M})} L^2_{\text{in}}(I) \subset L^2_{\text{in}}(\Gamma)$ and define $\mathcal{S} := \mathcal{B} \cap L^\infty_{\text{in}}(\Gamma)$ as the vector space of bounded functions in $\mathcal{B}$. Let $\mathcal{W} := \{1; S \in \bigcup_{I \in \mathcal{F}(\mathcal{M})} \Sigma_I\}$ be the set of indicator functions that are in $L^\infty_{\text{in}}(I)$ for some $I \in \mathcal{F}(\mathcal{M})$. Then $\mathcal{W} \subset \mathcal{S}$, $1 \in \mathcal{S}$, and $\mathcal{W}$ is closed under multiplication. Let $0 \leq v_1 \leq v_2 \leq \cdots$ be a pointwise monotonic sequence in $\mathcal{S}$ and $v := \sup_n v_n$ its supremum. If $v \in L^\infty_{\text{in}}(\Gamma) \subset L^2_{\text{in}}(\Gamma)$, then $(v_n)_n$ converges to $v$ in $L^2_{\text{in}}(\Gamma)$ by dominated convergence. Since $\mathcal{B}$ is closed in $L^2_{\text{in}}(\Gamma)$, $v \in \mathcal{B}$ and therefore $v \in \mathcal{S}$. Thus $\mathcal{S}$ is a monotone vector space and, using $\Sigma = \sigma(\mathcal{W})$, the monotone class theorem implies $\mathcal{S} = L^\infty_{\text{in}}(\Gamma)$.

If $v \in L^2_{\text{in}}(\Gamma)$, then for any $N \in \mathbb{N}$, $v_1_{\lVert I \leq N} \in L^\infty_{\text{in}}(\Gamma) = \mathcal{S} \subset \mathcal{B}$ and $v \in \mathcal{B}$ by dominated convergence. \hfill \Box

Remark 2.2.5. Note that the proof of Proposition 2.2.4 does not use the product structure of the measure $\pi$. Therefore, the assertion holds for any finite measure $\pi$ on the product measurable space $(\Gamma, \Sigma)$.

2.2.2. Finite Tensor Product Frames

For all $m \in \mathcal{M}$, let $\Phi_m := (\psi^m_k)_{k \in \Lambda_m}$ be a frame of $L^2_{\text{in}}(\Gamma_m)$. For any $I \in \mathcal{F}(\mathcal{M})$, define

$$\Phi_I := \prod_{m \in I} \Phi_m = (\psi_I^m)_{(m) \in \Lambda_I}, \quad \psi_I^m := \bigotimes_{m \in I} \psi^m_{v_m}, \quad \Lambda_I := \prod_{m \in I} \Lambda_m.$$  \hfill (2.2.6)

Repeated application of Theorem 2.1.4 implies that $\Phi_I$ is a frame of $L^2_{\text{in}}(\Gamma_I)$ with bounds

$$B_{\Phi_I} = \prod_{m \in I} B_{\Phi_m} \quad \text{and} \quad b_{\Phi_I} = \prod_{m \in I} b_{\Phi_m}. \hfill (2.2.7)$$

Assumption 2.2.A. The frame constants of $\Phi_m$ satisfy $b_{\Phi_m} \leq 1 \leq B_{\Phi_m}$ for all $m \in \mathcal{M}$ and

$$(\log(B_{\Phi_m}))_{m \in \mathcal{M}} \quad \text{and} \quad (\log(b_{\Phi_m}))_{m \in \mathcal{M}} \in \ell^1(\mathcal{M}). \hfill (2.2.8)$$

As in Lemma 2.2.2, we identify $\psi_I^m \in \Phi_I$ with $\psi_I \circ y_I$ and interpret $\Phi_I$ as a subset of $L^2_{\text{in}}(I)$ for all $I \in \mathcal{F}(\mathcal{M})$.

Proposition 2.2.6. $\Phi_I$ is a frame of $L^2_{\text{in}}(\Gamma_I)$ with bounds satisfying $B_{\Phi} \leq b_{\Phi} \leq 1 \leq B_{\Phi} \leq B_{\Phi}$ for

$$B_{\Phi} := \prod_{m \in \mathcal{M}} B_{\Phi_m} \quad \text{and} \quad b_{\Phi} := \prod_{m \in \mathcal{M}} b_{\Phi_m}. \hfill (2.2.9)$$

The kernel of the synthesis operator $T_{\Phi_I}$ is

$$K_{\Phi_I} = \sum_{m \in I} K_{\Phi_m} \bigotimes_{n \in I \setminus m} \ell^2(\Lambda_n) \subset \ell^2(\Lambda_I). \hfill (2.2.10)$$

In particular, if $\Phi_m$ is a Riesz basis for all $m \in I$, then $\Phi_I$ is also a Riesz basis.

Proof. The assertion follows from Lemma 2.2.2 and recursive application of Theorem 2.1.4. By Assumption 2.2.A, the frame bounds $B_{\Phi_I}$ and $b_{\Phi_I}$ given in (2.2.7) are increasing and decreasing in $I$, respectively, and have limits (2.2.9). \hfill \Box
Chapter 2. Countably Infinite Tensor Product Frames

2.2.3. Compatibility of Finite Tensor Product Frames

Assumption 2.2.B. For each $m \in \mathcal{M}$, there is an index $0 \in \Lambda_m$ such that $\phi_0^m = 1$.

Define the increment index sets $\Delta_m := \Lambda_m \setminus \{0\}$ for $m \in \mathcal{M}$, and their products

\[
\Delta_\emptyset := \{0\} \quad \text{and} \quad \Delta_I := \prod_{m \in I} \Delta_m \ , \quad \emptyset \neq I \in \mathcal{F}(\mathcal{M}) . \tag{2.2.11}
\]

The disjoint union of these sets is

\[
\Lambda := \bigsqcup_{I \in \mathcal{F}(\mathcal{M})} \Delta_I . \tag{2.2.12}
\]

This is the index set of the sequence of functions

\[
\Phi := (\phi_\nu)_{\nu \in \Lambda} \subset L^2_\mathcal{M}(\Gamma) . \tag{2.2.13}
\]

We define the subsequences

\[
\Phi^j_I := \Phi|_{\Delta_I} = (\phi_\nu)_{\nu \in \Delta_I} \subset L^2_\mathcal{M}(\Gamma) \ , \quad I \in \mathcal{F}(\mathcal{M}) . \tag{2.2.14}
\]

Let $I \in \mathcal{F}(\mathcal{M})$. The support of $\nu \in \Lambda_I$ is

\[
\text{supp } \nu := \{m \in I \mid \nu_m \neq 0 \in \Lambda_m \} \in \mathcal{P}(I) \subset \mathcal{F}(\mathcal{M}) ; \tag{2.2.15}
\]

we denote the restriction of $\nu$ to its support by

\[
\hat{\nu} := \nu|_{\text{supp } \nu} = (\nu_m)_{m \in \text{supp } \nu} \in \Delta_{\text{supp } \nu} \subset \Lambda . \tag{2.2.16}
\]

For all $\nu \in \Lambda_I$, $\hat{\nu} \in \Delta_I$ for $J = \text{supp } \nu \in \mathcal{P}(I)$, and conversely, if $\mu \in \Delta_I$ with $J \in \mathcal{P}(I)$, then there is exactly one $\nu \in \Lambda_I$ with $\mu = \hat{\nu}$. Assumption 2.2.B implies that $\phi_\nu = \phi_{\hat{\nu}}$ for all $\nu \in I$ since

\[
\phi_\nu(y) = \left( \prod_{m \in \text{supp } \nu} \phi_\nu^m(y_m) \right) \left( \prod_{m \notin \text{supp } \nu} \phi_0^m(y_m) \right) = \phi_{\hat{\nu}}(y) , \quad y \in \Gamma .
\]

Thus,

\[
\Phi_I = (\phi_\nu)_{\nu \in \Lambda_I} = (\phi_{\hat{\nu}})_{\mu \in \Delta_I, J \in \mathcal{P}(I)} = (\Phi^j_I)_{J \in \mathcal{P}(I)} , \tag{2.2.17}
\]

and we identify

\[
\Lambda_I = \bigsqcup_{J \in \mathcal{P}(I)} \Delta_I \subset \Lambda . \tag{2.2.18}
\]

By (2.2.12), $\ell^2(\Lambda)$ can be written as an orthogonal direct sum,

\[
\ell^2(\Lambda) = \bigoplus_{I \in \mathcal{F}(\mathcal{M})} \ell^2(\Delta_I) . \tag{2.2.19}
\]
2.2. Frames on Infinite Dimensional Domains

Furthermore, (2.2.18) implies
\[
\ell^2(\Lambda_I) = \bigoplus_{J \in \mathcal{P}(I)} \ell^2(\Lambda_J) \subset \ell^2(\Lambda), \quad I \in \mathcal{F} (\mathcal{M}) .
\] (2.2.20)

In particular, since \( \Lambda_m = \Delta_{\phi} \cup \Delta_m \) for all \( m \in \mathcal{M} \), we have the orthogonal decomposition
\[
\ell^2(\Lambda_m) = \ell^2(\Delta_{\phi}) \oplus \ell^2(\Delta_m), \quad m \in \mathcal{M} .
\] (2.2.21)

**Assumption 2.2.C.** For all \( m \in \mathcal{M} \), \( K_{\Phi_m} \subset \ell^2(\Lambda_m) \) is orthogonal to \( \ell^2(\Delta_{\phi}) \).

Of course, Assumption 2.2.C is trivially satisfied if \( \Phi_m \) is a Riesz basis of \( L^2_{\text{loc}}(I_m) \) for all \( m \in \mathcal{M} \).

If \( I, J \in \mathcal{F} (\mathcal{M}) \) with \( J \subset I \), then (2.2.20) implies \( \ell^2(\Lambda_I) \subset \ell^2(\Lambda_J) \), and by (2.2.17),
\[
T_{\Phi_I} |_{\ell^2(\Lambda_I)} = T_{\Phi_J} .
\] (2.2.22)

Define the synthesis operators
\[
T_{\Phi_I} := T_{\Phi_J} |_{\ell^2(\Lambda_I)} = T_{\Phi_I} |_{\ell^2(\Lambda_I)}, \quad J \subset I \in \mathcal{F} (\mathcal{M}) ,
\] (2.2.23)
and their kernels
\[
K_{\Phi_I} := \ker T_{\Phi_I} = K_{\Phi_J} \cap \ell^2(\Lambda_I) = K_{\Phi_J} \cap \ell^2(\Lambda_J), \quad J \subset I \in \mathcal{F} (\mathcal{M}) .
\] (2.2.24)

By (2.2.21) and (2.2.24), Assumption 2.2.C states that \( K_{\Phi_m} = K_{\Phi_m} \) for all \( m \in \mathcal{M} \).

**Lemma 2.2.7.**
\[
K_{\Phi_I} = \bigoplus_{J \in \mathcal{P}(I)} K_{\Phi_J}, \quad I \in \mathcal{F} (\mathcal{M}) .
\] (2.2.25)

**Proof.** If \( I = \emptyset \), then \( K_{\Phi_{\emptyset}} = \{ 0 \} \subset \ell^2(\Delta_{\phi}) \) since \( \varphi_0 = 1 \). We proceed by induction over \#I. Let \( I \in \mathcal{F} (\mathcal{M}) \) be nonempty. For an \( m \in I \), define \( I' := I \setminus \{ m \} \). Then Proposition 2.2.6 implies
\[
K_{\Phi_I} = K_{\Phi_m} \otimes \ell^2(\Lambda_{I'}) + \ell^2(\Lambda_m) \otimes K_{\Phi_{I'}} .
\]

By (2.2.21) and since \( \ell^2(\Delta_{\phi}) \) is one-dimensional,
\[
\ell^2(\Lambda_m) \otimes K_{\Phi_{I'}} = \ell^2(\Lambda_m) \otimes K_{\Phi_{I'}} + K_{\Phi_{I'}} .
\]

Using \( K_{\Phi_m} = K_{\Phi_m} \) by Assumption 2.2.C, (2.2.20) and (2.2.25) for \( I' \),
\[
K_{\Phi_I} = K_{\Phi_m} \otimes \bigoplus_{J \in \mathcal{P}(I')} \ell^2(\Lambda_J) + \ell^2(\Lambda_m) \otimes \bigoplus_{J \in \mathcal{P}(I')} K_{\Phi_J} + \bigoplus_{J \in \mathcal{P}(I')} K_{\Phi_J} \overset{\text{(2.2.24)}}{=} \bigoplus_{J \in \mathcal{P}(I')} \left( K_{\Phi_m} \otimes \ell^2(\Lambda_J) + \ell^2(\Lambda_m) \otimes K_{\Phi_J} \right) + \bigoplus_{J \in \mathcal{P}(I')} K_{\Phi_J} .
\]

27
By Proposition 2.2.6 and (2.2.24), for all $J \in \mathcal{P}(l')$,
\[
K_{\Phi^J_I} = K_{\Phi^J_I} \cap \ell^2(\Lambda_I) \\
= \left( K_{\Phi^J_I} \otimes \ell^2(\Lambda_I) + \ell^2(\Lambda_I) \otimes K_{\Phi^J_I} \right) \cap \left( \ell^2(\Lambda_I) \otimes \ell^2(\Lambda_I) \right) \\
= K_{\Phi^J_I} \otimes \ell^2(\Lambda_I) + \ell^2(\Lambda_I) \otimes K_{\Phi^J_I}.
\]

Therefore,
\[
K_{\Phi^J_I} = \bigoplus_{J \in \mathcal{P}(l')} K_{\Phi^J_I} + \bigoplus_{J \in \mathcal{P}(l')} K_{\Phi^J_I}.
\]

The sum remaining in this expansion is an orthogonal direct sum by (2.2.20) since $K_{\Phi^J_I} \subset \ell^2(\Lambda_I)$ for all $J \in \mathcal{P}(l)$. \hfill \qed

As a direct consequence of Lemma 2.2.7, the quotient spaces satisfy
\[
\ell^2(\Lambda_I)/K_{\Phi^J_I} = \bigoplus_{J \in \mathcal{P}(l')} \ell^2(\Lambda_I)/K_{\Phi^J_I}, \quad I \in \mathcal{F}(\mathcal{M}) , \tag{2.2.26}
\]
through the identification of $c + K_{\Phi^J_I}$ with $(c_J + K_{\Phi^J_I})_{J \in \mathcal{P}(l)}$ if $c = (c_J)_{J \in \mathcal{P}(l)}$. In particular, $\ell^2(\Lambda_I)/K_{\Phi^J_I}$ is a subspace of $\ell^2(\Lambda_I)/K_{\Phi^J_I}$ if $J \subset I$. Together with (2.2.22), this implies
\[
\hat{T}_{\Phi^J_I}|_{\ell^2(\Lambda_I)/K_{\Phi^J_I}} = \hat{T}_{\Phi^J_I}, \quad J \subset I \in \mathcal{F}(\mathcal{M}) . \tag{2.2.27}
\]

A similar decomposition to (2.2.26) holds for the orthogonal complements $K_{\Phi^J_I}^\perp \subset \ell^2(\Lambda_I)$ and $K_{\Phi^J_I}^\perp \subset \ell^2(\Lambda_I)$ in place of the quotient spaces.

The operators $T_{\Phi^J_I}$ induce maps on the quotient spaces,
\[
\hat{T}_{\Phi^J_I} : \ell^2(\Lambda_I)/K_{\Phi^J_I} \to c + K_{\Phi^J_I} \mapsto T_{\Phi^J_I}c . \tag{2.2.28}
\]

By (2.2.23),
\[
\hat{T}_{\Phi^J_I}|_{\ell^2(\Lambda_I)/K_{\Phi^J_I}} = \hat{T}_{\Phi^J_I}, \quad J \subset I \in \mathcal{F}(\mathcal{M}) . \tag{2.2.29}
\]

Define the spaces
\[
R_{\Phi^J_I} := \text{range } T_{\Phi^J_I} = \text{range } \hat{T}_{\Phi^J_I}, \quad L^2_{\mathcal{M}}(\Gamma) , \tag{2.2.30}
\]

\[\text{Lemma 2.2.8.} \quad \text{For all } I \in \mathcal{F}(\mathcal{M}), \text{ } R_{\Phi^J_I} \text{ is a closed subspace of } L^2_{\mathcal{M}}(\Gamma). \text{ Furthermore, if } I \neq J \in \mathcal{F}(\mathcal{M}), \text{ then } R_{\Phi^J_I} \cap R_{\Phi^J_I} = \{0\}. \]

\[\text{Proof.} \quad \text{For all } I \in \mathcal{F}(\mathcal{M}), \]
\[
R_{\Phi^J_I} = \text{range } \hat{T}_{\Phi^J_I} = \hat{T}_{\Phi^J_I} \left( \ell^2(\Lambda_I)/K_{\Phi^J_I} \right) = S_{\Phi^J_I} \left( \hat{T}_{\Phi^J_I}^{-1}(\ell^2(\Lambda_I)/K_{\Phi^J_I}) \right),
\]

which is a closed subspace of $L^2_{\mathcal{M}}(\Gamma)$ by Proposition 2.2.6. Let $J \in \mathcal{F}(\mathcal{M})$. Suppose $c \in \ell^2(\Lambda_I)$ and $d \in \ell^2(\Lambda_J)$ with $T_{\Phi^J_I}c = T_{\Phi^J_I}d = v$ for $v \in L^2_{\mathcal{M}}(\Gamma)$. Then $c - d \in K_{\Phi^J_I}$. If $I \neq J$, Lemma 2.2.7 implies $c \in K_{\Phi^J_I}$ and $d \in K_{\Phi^J_I}$. In particular, $v = 0$. \hfill \qed

28
In particular, Lemma 2.2.8 states that $\Phi^4_i$ is a frame of $R_{\Phi^4_i} \subset L^2_{\nu_i}(\Gamma)$ for all $I \in \mathcal{F}(\mathcal{M})$. By Proposition 2.2.6 and (2.2.17),

$$L^2_{m|\Sigma^j}(\Gamma) = \sum_{j \in \mathcal{P}(I)} R_{\Phi^4_i}, \quad I \in \mathcal{F}(\mathcal{M}), \quad (2.2.31)$$

and the sum in (2.2.31) is direct but not necessarily orthogonal. Note that

$$\sum_{j \in \mathcal{P}(I)} L^2_{m|\Sigma^j}(\Gamma) \subset L^2_{m|\Sigma^j}(\Gamma), \quad I \in \mathcal{F}(\mathcal{M}), \quad (2.2.32)$$

and $R_{\Phi^4_i}$ extends this space to $L^2_{m|\Sigma^j}(\Gamma)$.

**Lemma 2.2.9.** Let $I \in \mathcal{F}(\mathcal{M})$ such that, for all $m \in I$, $\Phi^i_m$ is a tight frame of $L^2_{m_{\nu_i}}(\Gamma_m)$. Then

$$R_{\Phi^i} = H_I := L^2_{m|\Sigma^j}(\Gamma) \oplus \sum_{j \in \mathcal{P}(I) \setminus I} L^2_{m|\Sigma^j}(\Gamma). \quad (2.2.33)$$

**Proof.** By Proposition 2.2.6, $\Phi^i_j$ is a tight frame of $L^2_{m|\Sigma^j}(\Gamma)$. Therefore, $\hat{T}_{\Phi^i_j}$ is an isometric isomorphism. The assertion follows since $R_{\Phi^i_j} = \hat{T}_{\Phi^i_j}(L^2(\Delta_I)/(K_{\Phi^i_j}))$ and by (2.2.26),

$$L^2(\Delta_I)/(K_{\Phi^i_j}) = L^2(\Lambda_I)/(K_{\Phi^i_j}) \oplus \sum_{j \in \mathcal{P}(I) \setminus I} L^2(\Delta_I)/(K_{\Phi^i_j}). \quad \square$$

In particular, under the condition that $\Phi^i_m$ is tight for all $m \in \mathcal{M}$, the spaces $R_{\Phi^i} \subset L^2_{\nu_i}(\Gamma)$ are independent of $\Phi^i_m$ and mutually orthogonal. By Lemma 2.2.9 and (2.2.31), we have the orthogonal decompositions

$$L^2_{m|\Sigma^j}(\Gamma) = \bigoplus_{j \in \mathcal{P}(I)} H_I, \quad I \in \mathcal{F}(\mathcal{M}). \quad (2.2.34)$$

Proposition 2.2.4 implies

$$L^2_{\nu_i}(\Gamma) = \bigoplus_{I \in \mathcal{F}(\mathcal{M})} H_I. \quad (2.2.35)$$

The subspace $H_I \subset L^2_{\nu_i}(\Gamma)$ can be interpreted as the space of square integrable functions on $\Gamma$ depending only on the dimensions $m \in I$ for $I \in \mathcal{F}(\mathcal{M})$, and such that the integral over any of these dimensions is zero.

### 2.2.4. Infinite Tensor Product Frames

We show that $\Phi^i$ defined in (2.2.13) is a frame of $L^2_{\nu_i}(\Gamma)$. Note that, even though $\Gamma$ may be infinite dimensional, each frame element $\varphi_v$ for $v \in \Lambda$ depends on only finitely many dimensions.

The synthesis operator $T_{\Phi^i}$ must map the Kronecker sequence $e_v \in L^2(\Lambda)$ that is one at $v \in \Lambda$ and zero otherwise onto $\varphi_v \in L^2_{\nu_i}(\Gamma)$ for all $v \in \Lambda$. Equivalently,

$$T_{\Phi^i} e_v = T_{\Phi^i} \quad \forall I \in \mathcal{F}(\mathcal{M}). \quad (2.2.36)$$

Furthermore, $T_{\Phi^i}$ is uniquely determined by (2.2.36) since $L^2(\Lambda)$ is spanned by $(e_v)_{v \in \Lambda}$. 

29
Chapter 2. Countably Infinite Tensor Product Frames

**Theorem 2.2.10.** The sequence $\Phi$ from (2.2.13) is a frame of $L^2_\mathcal{M}(\Gamma)$ with frame bounds (2.2.9). The kernel of its synthesis operator $T_\Phi$ is

$$\ker T_\Phi = K_\Phi = \bigoplus_{I \in \mathcal{F}(\mathcal{M})} K_{\Phi_I^\perp}.$$  \hspace{1cm} (2.2.37)

In particular, $\Phi$ is a Riesz basis of $L^2_\mathcal{M}(\Gamma)$ if and only if $\Phi_m$ is a Riesz basis of $L^2_{\mathcal{M}_m}(\Gamma_m)$ for all $m \in \mathcal{M}$.

**Proof.** Define $K_\Phi \subset \ell^2(\Lambda)$ by (2.2.37), and let $\pi_\Phi$ be the canonical projection of $\ell^2(\Lambda)$ onto $\ell^2(\Lambda)/K_\Phi$. By (2.2.19), as in (2.2.26),

$$\ell^2(\Lambda)/K_\Phi = \bigoplus_{I \in \mathcal{F}(\mathcal{M})} \ell^2(\Lambda_I)/K_{\Phi^\perp_I}.$$ \hspace{1cm} (2.2.38)

We identify $\ell^2(\Lambda)/K_\Phi$, with its embedding in $\ell^2(\Lambda)/K_\Phi$ through (2.2.26) and (2.2.38). Since finitely supported sequences are dense in $\ell^2(\Lambda)$, and any finitely supported sequence is in $\ell^2(\Lambda_I)$ for some $I \in \mathcal{F}(\mathcal{M})$,

$$\bigcup_{I \in \mathcal{F}(\mathcal{M})} \ell^2(\Lambda_I)/K_{\Phi_I} \subset \ell^2(\Lambda)/K_\Phi$$

is dense. We construct the synthesis operator on this domain. Condition (2.2.36) dictates the definition

$$\hat{T} : \bigcup_{I \in \mathcal{F}(\mathcal{M})} \ell^2(\Lambda_I)/K_{\Phi_I} \to L^2_\mathcal{M}(\Gamma), \quad \hat{T}_{\ell^2(\Lambda_I)/K_{\Phi_I}} := \hat{T}_{\Phi_I},$$ \hspace{1cm} (2.2.39)

which is consistent by (2.2.27). Using that the spaces $\ell^2(\Lambda)/K_{\Phi_I}$ are nested,

$$\|\hat{T}\| = \sup_{I \in \mathcal{F}(\mathcal{M})} \left\|\hat{T}_{\ell^2(\Lambda_I)/K_{\Phi_I}}\right\| = \sup_{I \in \mathcal{F}(\mathcal{M})} B_{\Phi_I} = \prod_{m \in \mathcal{M}} B_{\Phi_m} = B_{\Phi}$$ \hspace{1cm} (2.2.40)

since $B_{\Phi_m} \geq 1$ for all $m \in \mathcal{M}$. By surjectivity of $\hat{T}_{\Phi_I}$, for all $I \in \mathcal{F}(\mathcal{M})$,

$$\text{range } \hat{T} = \bigcup_{I \in \mathcal{F}(\mathcal{M})} \text{range } \hat{T}_{\Phi_I} = \bigcup_{I \in \mathcal{F}(\mathcal{M})} L^2_{\mathcal{M}_I(\Gamma)}.$$ \hspace{1cm} (2.2.41)

Since $\hat{T}_{\Phi_I}$ is injective for all $I \in \mathcal{F}(\mathcal{M})$, $\hat{T}$ is invertible on its range with inverse

$$\hat{T}^{-1} : \bigcup_{I \in \mathcal{F}(\mathcal{M})} L^2_{\mathcal{M}_I(\Gamma)} \to \bigcup_{I \in \mathcal{F}(\mathcal{M})} \ell^2(\Lambda_I)/K_{\Phi_I}, \quad \hat{T}^{-1}_{L^2_{\mathcal{M}_I(\Gamma)}} := \hat{T}_{\Phi_I}^{-1}.$$ \hspace{1cm} (2.2.42)

Proposition 2.2.4 states that the domain of $\hat{T}^{-1}$ is dense in $L^2_\mathcal{M}(\Gamma)$. As in (2.2.40),

$$\|\hat{T}^{-1}\| = \sup_{I \in \mathcal{F}(\mathcal{M})} \left\|\hat{T}^{-1}_{L^2_{\mathcal{M}_I(\Gamma)}}\right\| = \sup_{I \in \mathcal{F}(\mathcal{M})} B_{\Phi_I}^{-1} = \prod_{m \in \mathcal{M}} B_{\Phi_m}^{-1} = B_{\Phi}^{-1}.$$ \hspace{1cm} (2.2.43)
2.2. Frames on Infinite Dimensional Domains

The operators \( \hat{T} \) and \( \hat{T}^{-1} \) extend by continuity and density to
\[
\hat{T}_\Phi : \ell^2(\Lambda) / K_\Phi \to L^2_\pi(\Gamma) \quad \text{and} \quad \hat{T}^{-1}_\Phi : L^2_\pi(\Gamma) \to \ell^2(\Lambda) / K_\Phi ,
\]
which, as the notation suggests, are inverse to each other. By (2.2.27) and since \( \pi_\Phi|_{\ell^2(\Lambda_I)} = \pi_\Phi_I \) for all \( I \in \mathcal{F}(\mathcal{M}) \), the map
\[
T_\Phi = \hat{T}_\Phi \pi_\Phi
\]
(2.2.45)
satisfies (2.2.36). Therefore, (2.2.45) is the synthesis operator of \( \Phi \). By construction, \( T_\Phi \) is bounded with norm \( B_\Phi \), and surjectivity of \( \hat{T}_\Phi \) implies
\[
\text{range } T_\Phi = \text{range } \hat{T}_\Phi = L^2_\pi(\Gamma) .
\]
(2.2.46)
Equation (2.2.37) follows from (2.2.45), Lemma 2.2.7 and injectivity of \( \hat{T}_\Phi \).

\[\square\]

Remark 2.2.11. If \( \mathcal{M} \) is countable, then as a countable union of countable sets, \( \Lambda \) is also countable. If \( \mathcal{M} \) is uncountable, \( \Lambda \) is also an uncountable set, and \( L^2_\pi(\Gamma) \) is not separable. Although our definition of frames does not cover this setting, everything in Section 2.1.1 remains true.

Corollary 2.2.12. If \( \Phi_m \) is a tight frame of \( L^2_\pi(\Gamma_m) \) for all \( m \in \mathcal{M} \), then \( \Phi \) is a tight frame of \( L^2_\pi(\Gamma) \). The synthesis operator has the form
\[
T_\Phi = \bigoplus_{l \in \mathcal{F}(\mathcal{M})} T_{\Phi_I} ,
\]
(2.2.47)
where \( T_{\Phi_I} : \ell^2(\Lambda_I) \to H_I \) is given by (2.2.23) for all \( l \in \mathcal{F}(\mathcal{M}) \).

Proof. Theorem 2.2.10 implies that \( \Phi \) is a tight frame of \( L^2_\pi(\Gamma) \) since \( b_\Phi = B_\Phi = 1 \) by (2.2.9). Equation (2.2.47) follows from (2.2.36), (2.2.23) and Lemma 2.2.9 using the decompositions (2.2.19) and (2.2.35).

\[\square\]

Corollary 2.2.13. If \( \Phi_m \) is an orthonormal basis of \( L^2_\pi(\Gamma_m) \) for all \( m \in \mathcal{M} \), then \( \Phi \) is an orthonormal basis of \( L^2_\pi(\Gamma) \) and its synthesis operator is of the form (2.2.47).

Proof. By Theorem 2.2.10 and Corollary 2.2.12, \( \Phi \) is a tight Riesz basis \( L^2_\pi(\Gamma) \).

In the setting of Corollary 2.2.12, the canonical dual frame \( \Phi^*_J \) is equal to \( \Phi \). Since Assumption 2.2.A is trivially satisfied in this setting, we will consider examples of tight frames below.

Remark 2.2.14. If \( \Phi \) is not tight, then the canonical dual frames \( \Phi^*_J \) are not necessarily nested for \( J \in \mathcal{F}(\mathcal{M}) \). By Corollary 2.1.5,
\[
\Phi^*_J = \prod_{m \in J} \Phi^*_m = \left( \bigotimes_{m \in J} (\phi^*_m) \right)_{v \in \Lambda_J} .
\]
(2.2.48)
Chapter 2. Countably Infinite Tensor Product Frames

By the Riesz representation theorem, the dual element \( \bigotimes_{m \in I} \langle q^m \rangle^* \) can be represented on \( L^2_{\mathcal{M}}(\mathcal{I}) \) as the inner product with a function \( \psi^I \in L^2_{\mathcal{M}}(\mathcal{I}) \) for any \( I \in \mathcal{F}(\mathcal{M}) \) with \( j \subset I \). However, in general these functions depend on \( I \) and are related to each other only as conditional expectations. In particular, the Riesz representation theorem on \( L^2_{\mathcal{M}}(\mathcal{I}) \) may identify \( \bigotimes_{m \in I} \langle q^m \rangle^* \) with a function that depends on infinitely many coordinates of \( \mathcal{I} \).

Remark 2.2.15. Corollary 2.2.12 suggests an alternative approach to constructing a frame of \( L^2_{\mathcal{M}}(\mathcal{I}) \). If \( \Phi^I = (\varphi_m)_{m \in I} \) is a frame of \( H_I \) for all \( I \in \mathcal{F}(\mathcal{M}) \) with uniform bounds

\[
B_\Phi := \sup_{I \in \mathcal{F}(\mathcal{M})} B_{\Phi^I} < \infty \quad \text{and} \quad b_\Phi := \inf_{I \in \mathcal{F}(\mathcal{M})} b_{\Phi^I} > 0 ,
\]

then equation (2.2.47) defines the synthesis operator of a frame \( \Phi \) of \( L^2(\mathcal{I}) \) with bounds \( B_\Phi \) and \( b_\Phi \). In particular, if \( \Phi^I \) is a Riesz basis for all \( I \in \mathcal{F}(\mathcal{M}) \), then \( \Phi \) is also a Riesz basis. These and similar abstract results follow from elementary properties of direct sums, see e.g. [KR97, Sec. 2.6]. The main difficulty associated with this approach is the direct construction of frames or bases of the orthogonal increment spaces \( H_I, I \in \mathcal{F}(\mathcal{M}) \).

If \( \Phi^\mathcal{I}_m = (\varphi^m_k)_{k \in \Delta_m} \) is a frame of the space \( L^2_{\mathcal{M}}(\Gamma_m) \) of square integrable functions on \( \Gamma_m \) with zero mean for all \( m \in \mathcal{M} \), and the frame bounds \( b_{\Phi^\mathcal{I}_m} := b_{\Phi^I} \) and \( B_{\Phi^\mathcal{I}_m} := B_{\Phi^I} \) satisfy Assumption 2.2.1A, then the products

\[
\Phi^I_m := \prod_{m \in I} \Phi^\mathcal{I}_m = \left( \bigotimes_{m \in I} \langle q^m \rangle \right)_{m \in \Delta_I} , \quad \Delta_I := \prod_{m \in I} \Delta_m ,
\]

constitute a frame of \( H_I \) for each \( I \in \mathcal{F}(\mathcal{M}) \). The frame bounds (2.2.49) of the resulting frame \( \Phi \) of \( L^2_{\mathcal{M}}(\mathcal{I}) \) are given by (2.2.9). More generally, one might construct frames of \( H_I \) with no tensor product structure.

2.2.5. Countably Infinite Products of Boundedly Hierarchic Frames

We show that Proposition 2.1.7 generalizes to infinite products of boundedly hierarchic frames.

The frames considered in Theorem 2.2.10 and Remark 2.2.15 have some hierarchic structure. It seems natural to define \( v' \preceq v \) if \( \text{supp } v' \subseteq \text{supp } v \). However, \( \Phi \) with this partial order on \( \Lambda \) does not satisfy the conditions of a boundedly hierarchic frame since \( \Delta_I \) can be infinite for \( I \in \mathcal{F}(\mathcal{M}) \).

Proposition 2.2.16. If, for all \( m \in \mathcal{M} \), \( \Phi_m \) is a boundedly hierarchic frame of \( L^2_{\mathcal{M}}(\Gamma_m) \) such that \( 0 \preceq k \) for all \( k \in \Delta_m \), then \( \Phi \) is a boundedly hierarchic frame when \( \Lambda \) is endowed with the partial order

\[
v' \preceq v :\iff v'_m \preceq v_m \quad \forall m \in \mathcal{M} ,
\]

and \( v' \preceq v \) if \( v' \preceq v \land v' \neq v \).
Proof. By definition, the only minimal element in $\Lambda$ is 0. Due to (2.2.12), for any $\nu \in \Lambda$, $\nu_m$ is minimal in $\Lambda_m$ for all but finitely many $m \in \mathcal{M}$. Then it follows by repeated application of Proposition 2.1.7 that $\nu$ has only finitely many predecessors. □

Remark 2.2.17. Note that the partial order defined in Proposition 2.2.16 is compatible with the supports of indices $\nu \in \Lambda$ in that if $\nu' \prec \nu$, then $\text{supp } \nu' \subset \text{supp } \nu$. Furthermore, it is compatible with traces. Let $I \in \mathcal{F}(\mathcal{M})$ and $\mu \in \Lambda$ with $\text{supp } \mu \subset \mathcal{M} \setminus I$. Then for $\nu, \nu' \in \Lambda_I$, $(\nu', \mu) \prec (\nu, \mu)$ if and only if $\nu' \prec \nu$, where $(\nu, \mu) \in \Lambda$ is given by $(\nu, \mu)_m := \nu_m$ for $m \in I$ and $(\nu, \mu)_m := \mu_m$ for $m \in \mathcal{M} \setminus I$, and $\prec$ is defined on $\Lambda_I$ by restriction of (2.2.51) or, equivalently, by recursive application of (2.1.39). This property uniquely characterizes $\prec$ on $\Lambda$ since any two indices $\nu, \nu' \in \Lambda$ are in $\Lambda_I$ for some $I \in \mathcal{F}(\mathcal{M})$. □
Chapter 3.
Transformation to a Discrete System

Using the weak formulations derived in Chapter 1, parametric operator equations can be represented as bi-infinite operator matrix equations by passing to coefficients with respect to the infinite product frames constructed in Chapter 2. This reduces the original uncountable set of independent equations to a countably infinite coupled system of equations. We derive these equations in an abstract setting for general frames in Section 3.1.

The infinite system of equations has a particularly simple structure for parametric operators that depend affinely on the parameter, if these are represented with respect to tensorized polynomials, which are orthonormal with respect to the measure on the parameter domain. In this case, all integrals over the parameter domain can be evaluated explicitly using the three term recursion formulas for the one-dimensional orthonormal polynomials. This is discussed in Section 3.2; we present some alternative tensor product bases in Section 3.3.

3.1. Frame Representation

3.1.1. Discretization of the Parameter Domain

Let $(\Gamma, \Sigma, \pi)$ be a measure space, and let $\Phi = (\varphi_v)_{v \in \Lambda}$ be a frame of $L^2(\Gamma)$. Let $\alpha$ and $\beta$ be tensor norms, and let

$$\mathcal{A} : L^2_{\pi}(\Gamma) \otimes_{\alpha} V \rightarrow L^2_{\pi}(\Gamma) \otimes_{\beta} W^*$$

be a bounded linear operator with norm less than or equal to $\hat{c}$. We define a frame representation $\mathcal{A}$ of $\mathcal{A}$ as

$$\mathcal{A} := T_{\Phi} W \mathcal{A} T_{\Phi}^V : \ell^2(\Lambda) \otimes_{\alpha} V \rightarrow \ell^2(\Lambda) \otimes_{\beta} W^* \quad$$

(3.1.1)

Then both $\mathcal{A}$ and $\hat{\mathcal{A}}$ are bounded with norm at most $B^2_{\Phi, \hat{c}}$. Equation (2.1.25) implies

$$\mathcal{A} = \pi_{\Phi}^V W \pi_{\Phi}^V \mathcal{A} \pi_{\Phi}^V \pi_{\Phi}^V$$

(3.1.4)

for $\pi_{\Phi}^V = \pi_{\Phi} \otimes \text{id}_V$ and $\pi_{\Phi}^W = \pi_{\Phi} \otimes \text{id}_{W^*}$. 

35
Chapter 3. Transformation to a Discrete System

We assume that $\mathcal{A}$ is boundedly invertible and that the norm of $\mathcal{A}^{-1}$ is less than or equal to $\tilde{c}$. Then also $\mathcal{H}$ is boundedly invertible, and the norm of $\mathcal{H}^{-1}$ is at most $b^{-2}\tilde{c}$, see Section 2.1.2. For an $f \in L^2_\pi(\Gamma) \otimes \beta W^*$, we are interested in solving

$$\mathcal{A} u = f.$$  \hfill (3.1.5)

In the settings of Theorem 1.1.11 or Theorem 1.3.5 with $p = 2$, (3.1.5) is a weak form of (1.1.2). We define frame representations of the right hand side,

$${\tilde{f}} := T^W_\Phi f \in \ell^2(\Lambda) \otimes \beta W^*,$$  \hfill (3.1.6)

$${\tilde{\mathcal{H}} u} := T^V_\Phi f \in (\ell^2(\Lambda)/K_\Phi) \otimes \beta W^*.$$  \hfill (3.1.7)

Thus we have the transformed equations

$$\mathcal{H} u = {\tilde{f}},$$  \hfill (3.1.8)

$$\mathcal{H} \hat{u} = {\tilde{\mathcal{H}} u},$$  \hfill (3.1.9)

where $u \in \ell^2(\Lambda) \otimes V$ and $\hat{u} \in (\ell^2(\Lambda)/K_\Phi) \otimes V$.

**Theorem 3.1.1.** Let $f \in L^2_\pi(\Gamma) \otimes \beta W^*$. The operator $\mathcal{H}$ in (3.1.3) is boundedly invertible, and the solutions of (3.1.5) and (3.1.9) are related by

$$u = T^V_\Phi \hat{u}.$$  \hfill (3.1.10)

Furthermore, there is a $u \in \ell^2(\Lambda) \otimes V$ satisfying (3.1.8), and all such $u$ are characterized by

$$u = T^V_\Phi u.$$  \hfill (3.1.11)

If $\Phi$ is a Riesz basis of $L^2_\pi(\Gamma)$, then $u$ is unique.

**Proof.** Since $T^W_\Phi$ and $T^V_\Phi$ are boundedly invertible by Proposition A.2.4, see Section 2.1.2, $\mathcal{H}$ is boundedly invertible, and thus (3.1.9) has a unique solution $\hat{u} \in (\ell^2(\Lambda)/K_\Phi) \otimes V$ for any $\tilde{f} \in (\ell^2(\Lambda)/K_\Phi) \otimes \beta W^*$. Applying $(T^W_\Phi)^{-1}$ to (3.1.9), it follows that $T^V_\Phi \hat{u}$ solves (3.1.5), which shows (3.1.10) by uniqueness of $u$.

Applying $T^W_\Phi$ to (3.1.11), it follows that if $u$ satisfies (3.1.11), it also satisfies (3.1.8). Conversely, let $u$ be a solution of (3.1.8). Applying $T^W_\Phi$ to (3.1.8), we get

$$(S_\Phi \otimes \text{id}_{W^*}) \mathcal{A} u = T^W_\Phi u = T^W_\Phi \tilde{f} = (S_\Phi \otimes \text{id}_{W^*}) f.$$  

Since $S_\Phi \otimes \text{id}_{W^*}$ is boundedly invertible by Proposition A.2.4, $T^W_\Phi u$ satisfies (3.1.5), and (3.1.11) follows by uniqueness of $u$.

Due to Theorem 2.1.1, $T^V_\Phi$ is surjective, and thus there is a $u \in \ell^2(\Lambda) \otimes V$ satisfying (3.1.11), which by the above is equivalent to (3.1.8). Furthermore, this $u$ is unique up to an element of $\ker T^V_\Phi$, which is the closure of $K_\Phi \otimes V$ in $\ell^2(\Lambda) \otimes V$. In particular, if $\Phi$ is a Riesz basis of $L^2_\pi(\Gamma)$, then $K_\Phi = \{0\}$, and therefore the solution $u$ of (3.1.11) is unique. □
3.1. Frame Representation

Theorem 3.1.1 applies in particular to the Hilbert tensor norm. We identify $\ell^2(A) \otimes_{\eta} X$ with $\ell^2(A; X)$ and $L^2_\eta(I) \otimes_{\eta} X$ with $L^2_\eta(I; X)$ for any separable Hilbert space $X$.

**Corollary 3.1.2.** If $V$ and $W$ are separable Hilbert spaces, then $u \in \ell^2(A; V)$ satisfies (3.1.8) if and only if it satisfies (3.1.11), and such $u$ exist for any $f \in L^2_\eta(I; W^*)$.

**Proof.** In Theorem 3.1.1, we set $\alpha$ and $\beta$ equal to the Hilbert tensor norm. Then the assertion follows using the above identifications. $\square$

### 3.1.2. Transformation to Sequence Spaces

Let $V$ and $W$ be separable Hilbert spaces. Furthermore, let $\Psi = (\psi_i)_{i \in \Xi}$ be a frame of $V$ and $\Theta = (\theta_k)_{k \in \Gamma}$ a frame of $W$. By Theorem 2.1.4, the products $\Phi \times \Psi = (\varphi_k \otimes \psi_i)_{(k,i) \in \Lambda \times \Xi}$ and $\Phi \times \Theta = (\varphi_k \otimes \theta_k)_{(k,k) \in \Lambda \times \Gamma}$ are frames of $L^2_\eta(I; V)$ and $L^2_\eta(I; W)$, respectively.

Similarly to Section 3.1.1, we define the bi-infinite matrices

\[
A := T_{\Phi \times \Psi} \circ T_{\Phi \times \Psi} : \ell^2(\Lambda \times \Xi) \to \ell^2(\Lambda \times \Gamma),
\]

\[
\hat{A} := \hat{T}_{\Phi \times \Psi} \circ \hat{T}_{\Phi \times \Psi} : \ell^2(\Lambda \times \Xi) / K_{\Phi \times \Psi} \to \ell^2(\Lambda \times \Gamma) / K_{\Phi \times \Theta}.
\]

Both $A$ and $\hat{A}$ are bounded with norms at most $b^{2}_{\Phi} b_{\Psi} b_{\Theta} \hat{c}$. Furthermore, $\hat{A}$ is invertible, and the norm of $\hat{A}^{-1}$ is bounded by $b^{-2}_{\Phi} b^{-1}_{\Psi} b^{-1}_{\Theta} \hat{c}$.

The operators $A$, $\hat{A}$, $\Psi$ and $\hat{\Psi}$ are related by

\[
A = (T^\ell_{\Theta})^* \Psi T^\ell_{\Psi} = \pi_{\Phi \times \Theta} \hat{A} \pi_{\Phi \times \Psi},
\]

\[
\hat{A} = (T^\ell_{\Theta})^* \hat{\Psi} \hat{T}^\ell_{\Psi},
\]

due to Theorem 2.1.4.

To discretize equation (3.1.5), we represent $f$ by

\[
f := T_{\Phi \times \Theta} f \in \ell^2(\Lambda \times \Gamma),
\]

\[
\hat{f} := \hat{T}_{\Phi \times \Theta} f \in \ell^2(\Lambda \times \Gamma) / K_{\Phi \times \Theta}.
\]

Then $f = (T^\ell_{\Theta})^* \hat{f}$, $\hat{f} = (\hat{T}^\ell_{\Theta})^* f$ and $f = \pi_{\Phi \times \Theta}^* \hat{f}$ by Theorem 2.1.4.

**Theorem 3.1.3.** $u \in \ell^2(\Lambda \times \Xi)$ solves the bi-infinite matrix equation

\[
Au = f
\]

if and only if it is related to the solution $u$ of (3.1.5) by

\[
u = T_{\Phi \times \Psi} u.
\]

In particular, a solution $u$ of (3.1.18) exists, and it is unique up to an element of $K_{\Phi \times \Psi}$.
Chapter 3. Transformation to a Discrete System

Proof. Applying $T_{\Phi \times \Theta}^*$ to (3.1.19) and inserting (3.1.5) and (3.1.16), it is evident that (3.1.18) follows from (3.1.19).

If $u \in \ell^2(\Lambda \times \Xi)$ satisfies (3.1.18), then applying $T_{\Phi \times \Theta}$ leads to

$$S_{\Phi \times \Theta} \mathcal{A} T_{\Phi \times \Theta} u = T_{\Phi \times \Theta} A u = T_{\Phi \times \Theta} f^\prime = S_{\Phi \times \Theta} f .$$

Since $S_{\Phi \times \Theta}$ is boundedly invertible, it follows that $T_{\Phi \times \Psi} u$ satisfies (3.1.5), and therefore (3.1.19) holds.

Consequently, $u$ is characterized by (3.1.19). A solution of (3.1.19) exists since $T_{\Phi \times \Psi}$ is surjective, and it is unique up to an element of $\ker T_{\Phi \times \Psi} = K_{\Phi \times \Theta}$. □

Similarly, (3.1.5) is equivalent to $\hat{A} u = f^\prime$ via $u = T_{\Phi \times \Theta} \hat{u}$. This amounts to considering (3.1.18) in the quotient spaces $\ell^2(\Lambda \times \Xi)/K_{\Phi \times \Psi}$ and $\ell^2(\Lambda \times \Xi)/K_{\Phi \times \Theta}$.

If $A$ is not symmetric positive definite, it is useful to consider the discrete normal equations

$$A^* A u = A^* f .$$

(3.1.20)

Here, $A^*$ is the adjoint bi-infinite matrix

$$A^* = T_{\Phi \times \Psi} \mathcal{A}^* T_{\Phi \times \Theta} : \ell^2(\Lambda \times \Xi) \rightarrow \ell^2(\Lambda \times \Xi) .$$

(3.1.21)

Note that $A^*$ is only injective if $\Phi$ and $\Theta$ are Riesz bases of $L^2(\Gamma)$ and $W$, respectively.

**Theorem 3.1.4.** $A u \in \ell^2(\Lambda \times \Xi)$ solves (3.1.20) if and only if it solves (3.1.18).

**Proof.** If $u \in \ell^2(\Lambda \times \Xi)$ satisfies (3.1.18), then (3.1.20) follows by applying $A^*$. Let $u \in \ell^2(\Lambda \times \Xi)$ solve (3.1.20). Note that by (2.1.9),

$$A^* A = T_{\Phi \times \Psi} \mathcal{A}^* T_{\Phi \times \Theta} \quad \text{and} \quad A^* f = T_{\Phi \times \Psi} \mathcal{A}^* f .$$

Therefore, applying $T_{\Phi \times \Psi}$ to (3.1.20) leads to

$$S_{\Phi \times \Psi} \mathcal{A}^* T_{\Phi \times \Theta} S_{\Phi \times \Psi} u = S_{\Phi \times \Psi} \mathcal{A}^* S_{\Phi \times \Theta} f .$$

Since $S_{\Phi \times \Psi}$, $\mathcal{A}^*$ and $S_{\Phi \times \Theta}$ are all invertible, it follows that $T_{\Phi \times \Psi} u$ satisfies (3.1.5), and the assertion follows using Theorem 3.1.3. □

### 3.1.3. Structure of the Operator Matrices

In the setting of Section 3.1.1, for all $\nu, \mu \in \Lambda$, we define the linear operator $A_{\nu \mu} \in \mathcal{L}(V, W^*)$ by

$$A_{\nu \mu} v := (q_{\nu} \otimes \text{id}_{W^*}) \mathcal{A} (q_{\mu} \otimes v) , \quad v \in V .$$

(3.1.22)

Here, $q_{\nu} \in L^2(\Gamma)$ is interpreted as an element of the bidual space, i.e. it is a linear functional on $L^2(\Gamma)$ defined by integration against the complex conjugate $\overline{q}_{\nu}$. We interpret spaces of the form $\ell^2(\Lambda) \otimes_\alpha X$ as spaces of sequences in $X$, see Theorem A.3.5.
Remark 3.1.6. If by integration against the complex conjugate $\bar{\nu} f = L_{\ell^2(\Lambda)} \otimes_{\beta} W^*$ in the sense of the Bochner integral, due to the isomorphisms of $L^2(\Gamma) \otimes_{\eta} W^*$ with $L^2(\Gamma; W^*)$ and $L^2(\Lambda) \otimes_{\eta} W^*$ with $L^2(\Lambda; W^*)$, then (3.1.25) holds in the sense of the Pettis integral since for any $\psi \in W^*$, $\psi(\varphi_v \otimes \id_{W^*}) f = (\varphi_v \otimes \psi) f = \varphi_v (\id_{L^2(\Gamma)} \otimes \psi) f = \varphi_v f = \int_{\Gamma} \psi(\varphi_v) \, d\tau$.
Chapter 3. Transformation to a Discrete System

The assertion follows using (A.1.10). Similarly, \( A_{\nu} v \) for \( v \in V \) can be interpreted as

\[
A_{\nu} v = \int \phi_\nu \sigma_i(\varphi_\nu \otimes v) \, d\pi \in \mathbb{W}^*
\]  
(3.1.26)

in the sense of the Bochner integral if \( V \) and \( W \) are separable Hilbert spaces, and in the sense of the Pettis integral if the integrand is Pettis integrable.

**Corollary 3.1.7.** The operator equation (3.1.8) is equivalent to the countably infinite system

\[
\sum_{\mu \in \Lambda} A_{\nu \mu} u_\mu = f_\nu \quad \forall \nu \in \Lambda
\]  
(3.1.27)

for \( A_{\nu \mu} \) from (3.1.22) and \( f_\nu \) from (3.1.24) in the sense that \( u = (u_\mu)_{\mu \in \Lambda} \in \ell^2(\Lambda) \otimes_\alpha V \) solves (3.1.8) if and only if its coefficients \( u_\mu \) satisfy (3.1.27).

**Proof.** The assertion follows from Theorem 3.1.5 since \( \hat{f} = (f_\nu)_{\nu \in \Lambda} \).

Let \( V \) and \( W \) be separable Hilbert spaces. Representing \( (A_{\nu \mu})_{\nu, \mu \in \Lambda} \) using the frames \( \Psi \) and \( \Theta \) of \( V \) and \( W \), respectively, leads to the operator \( A \) from (3.1.12). Define

\[
A_{\nu \mu} := T^*_{\Theta} A_{\nu \mu} T_{\Psi} : \ell^2(\Xi) \to \ell^2(\gamma) , \quad \nu, \mu \in \Lambda .
\]  
(3.1.28)

**Theorem 3.1.8.** For all \( v = (v_{\mu})_{(\mu, \nu) \in \Lambda} \in \ell^2(\Lambda \times \Xi) \)

\[
Av = \left( \sum_{\mu \in \Lambda} A_{\nu \mu} v_{\mu} \right)_{\nu \in \Lambda} \in \ell^2(\Lambda \times \gamma) ,
\]  
(3.1.29)

where \( v_{\mu} = (v_{\mu})_{\nu \in \Lambda} \).

**Proof.** We first note that for \( v = (v_{\mu})_{\mu \in \Lambda} \in \ell^2(\Lambda \times \Xi) \) and \( g = (g_{\nu})_{\nu \in \Lambda} \in \ell^2(\Lambda \times \gamma) \),

\[
\left( T^\ell_\Psi v_{\mu} \right)_{\nu} = T_{\Psi}v_{\mu} \in V \quad \text{ and } \quad \left( (T^\ell_\Theta)^* g_{\nu} \right)_{\nu} = T^*_{\Theta}g_{\nu} \in W^*
\]

for all \( \nu, \mu \in \Lambda \), since \( T^\ell_\Psi = \text{id}_{\ell^2(\Lambda)} \otimes T_{\Psi} \) and \( (T^\ell_\Theta)^* = \text{id}_{\ell^2(\Lambda)} \otimes T_{\Theta} \) by (2.1.26) and (2.1.27).

Consequently, (3.1.14) and (3.1.23) imply

\[
(Av)_{\nu} = \left( (T^\ell_{\Theta})^* T^\ell_\Psi \right) v_{\nu} = T^*_{\Theta} \sum_{\mu \in \Lambda} A_{\nu \mu} T_{\Psi} v_{\mu} = \sum_{\mu \in \Lambda} A_{\nu \mu} v_{\mu}
\]

for all \( \nu \in \Lambda \).

**Corollary 3.1.9.** The bi-infinite matrix equation (3.1.18) is equivalent to the countably infinite system

\[
\sum_{\mu \in \Lambda} A_{\nu \mu} u_\mu = f_\nu \quad \forall \nu \in \Lambda ,
\]  
(3.1.30)

where \( u_\mu = (u_{\mu})_{\nu \in \Xi} \in \ell^2(\Xi) \) and \( f_\nu = (f_{\nu \xi})_{\xi \in \gamma} \in \ell^2(\gamma) \).

**Proof.** Since \( T^\ell_{\Theta} f_{\nu} = \text{id}_{\ell^2(\Xi)} \otimes T_{\Theta} f_{\nu} \), we have \( f_{\nu} = T_{\Theta} f_{\nu} \) for \( f = (T^\ell_{\Theta})^* \hat{f} \) and \( \hat{f} = (f_\nu)_{\nu \in \Lambda} \). The rest of the assertion follows from Theorem 3.1.8.
3.2. Tensor Product Polynomial Bases

3.2.1. Construction of Orthonormal Polynomial Bases

We consider the case that, for all \( m \in \mathcal{M} \), \( \Gamma_m \subset \mathbb{R} \) and the \( \sigma \)-algebra \( \Sigma_m \) contains the Borel \( \sigma \)-algebra \( \mathcal{B}(\Gamma_m) \). If the moments

\[
M^m_n := \int_{\Gamma_m} \xi^n \, d\pi_m(\xi), \quad n \in \mathbb{N}_0,
\]

are finite, we can construct a sequence of orthonormal polynomials in \( L^2_{\pi_m}(\Gamma_m) \).

Let \( \Lambda_m := \{0, 1, \ldots, N - 1\} \) if the support of \( \pi_m \) has cardinality \( N \in \mathbb{N} \), and \( \Lambda_m := \mathbb{N}_0 \) otherwise.\(^1\) Let \( P_{-1} := 0 \), \( P_0^n := 1 \), and

\[
\beta^n_m P^n_m(\xi) := (\xi - \alpha^n_n)P^n_{n-1}(\xi) - \beta^{m-1}_n P^n_{n-2}(\xi), \quad n \in \Lambda_m \setminus \{0\},
\]

with

\[
\alpha^n_n := \int_{\Gamma_m} \xi P^n_m(\xi)^2 \, d\pi_m(\xi) \quad \text{and} \quad \beta^n_n := \frac{e^n_{m-1}}{c^n_n},
\]

where \( e^n_m \) is the leading coefficient of \( P^n_m \), \( P^n_0 := 1 \), and \( P^n_m \) is chosen as normalized in \( L^2_{\pi_m}(\Gamma_m) \). The values \((\alpha^n_m)_{n \in \Lambda_m} \) and \((\beta^n_m)_{n \in \Lambda_m} \) are conveniently tabulated for many common distributions \( \pi_m \); see Table 3.1 for the coefficients of a few classical polynomials on the interval \([-1, 1]\), or e.g. [Gau04], which tabulates \((\beta^n_m)^2 \) in place of \( \beta^n_m \).

**Lemma 3.2.1.** The sequence \( P_m := (P^n_m)_{n \in \Lambda_m} \) is orthonormal in \( L^2_{\pi_m}(\Gamma_m) \). Furthermore, \( P^n_m \) is a polynomial of degree \( n \) for all \( n \in \Lambda_m \).

**Proof.** This is shown e.g. in [Gau04, Sec. 1.3.2]. By the Gram–Schmidt orthogonalization process applied to the linearly independent monomials \((\xi^n)_{n \in \Lambda_m} \) in \( L^2_{\pi_m}(\Gamma_m) \), orthonormal polynomials \((P_n)_{n \in \Lambda_m} \) in \( L^2_{\pi_m}(\Gamma_m) \) exist. In particular, \( P_n \) is a polynomial of degree \( n \) for all \( n \in \Lambda_m \). We show that they satisfy the three term recursion (3.2.2). Note that \( P_0 = 1 \) since \( \pi_m \) is a probability measure. Let \( \alpha_n \) and \( \beta_n \) be defined as in (3.2.3) for \((P_n)_{n \in \Lambda_m} \) in place of \((P^n_m)_{n \in \Lambda_m} \).

For all \( n \in \Lambda_m \setminus \{0\} \), \( \beta_n P_n(\xi) - \xi P_{n-1}(\xi) \) is a polynomial of degree at most \( n - 1 \). Therefore, using the orthonormality of \((P^n_k)_{k=0}^n\),

\[
\beta_n P_n(\xi) - \xi P_{n-1}(\xi) = \gamma_{n-1} P_{n-1}(\xi) + \gamma_{n-2} P_{n-2}(\xi) + \cdots + \gamma_0
\]

with

\[
\gamma_k = \int_{\Gamma_m} (\beta_k P_k(\xi) - \xi P_k(\xi)) P_k d\pi_m(\xi) = -\int_{\Gamma_m} \xi P_{n-1}(\xi) P_k d\pi_m(\xi)
\]

for \( k = 0, 1, \ldots, n - 1 \). In particular, \( \gamma_{n-1} = -\alpha_{n-1} \), and \( \gamma_k = 0 \) for \( k \leq n - 3 \) since \( \xi P_k(\xi) \) is a polynomial of degree at most \( n - 2 \). Note that \( \xi P_{n-2}(\xi) = \beta_{n-1} P_{n-1}(\xi) + q(\xi) \) for a polynomial \( q \) of degree at most \( n - 3 \), so \( \gamma_{n-2} = -\beta_{n-1} \). Using that both \((P^n_m)_{n \in \Lambda_m} \) and \((P_n)_{n \in \Lambda_m} \) are normalized in \( L^2_{\pi_m}(\Gamma_m) \), it follows by induction that \( P^n_m = P^n_m \) up to a scalar factor of absolute value one for all \( n \in \Lambda_m \). \( \square \)

\(^1\)The support of \( \pi_m \) is defined as the set of \( \xi \in \Gamma_m \) such that \( \pi_m(U) > 0 \) for all open neighborhoods \( U \) of \( \xi \).
Chapter 3. Transformation to a Discrete System

The measure \( \pi_m \) is **determinate** if it is uniquely characterized by its moments \( (M_m^n)_{n \in \mathbb{N}_0} \).

We note that \( \pi_m \) is always determinate if its support is bounded.

**Proposition 3.2.2.** If \( \pi_m \) is determinate, then \( P_m = (P_{m,n})_{n \in \mathbb{N}_0} \) is an orthonormal basis of \( L^2_{\pi_m}(\Gamma_m) \).

**Proof.** Orthonormality is shown in Lemma 3.2.1. Completeness is due to F. Riesz, see e.g. [Fre71, Theorem 4.3], [Ber96, Theorem 2.1] and [Rie23]. \( \square \)

We assume that \( \pi_m \) is determinate for all \( m \in \mathcal{M} \). Then \( P_m \) satisfies Assumptions 2.2.A and 2.2.B since \( P_m^0 = 1 \) and \( b_{P_m} = b_{P_m} = 1 \) for all \( m \in \mathcal{M} \). Furthermore, it is graded with the polynomial degree as the grading function. Since there is only a single basis function of each degree, \( P_m \) is boundedly hierarchic.

We define the countable tensor product polynomials

\[
P := (P_v)_{v \in \Lambda}, \quad P_v := \bigotimes_{m \in \mathcal{M}} P_{v_m}^m, \quad v \in \Lambda,
\]

with \( \Lambda \) as in (2.2.12). Note that each of these functions depends on only finitely many dimensions,

\[
P_v(y) = \prod_{m \in \mathcal{M}} P_{v_m}^m(y_m) = \prod_{n \in \supp v} P_{v_n}^m(y_m), \quad v \in \Lambda,
\]

since \( P_0^m = 1 \) for all \( m \in \mathcal{M} \), where the support of \( v \in \Lambda \) is defined by (2.2.15).

**Corollary 3.2.3.** \( P \) is an orthonormal basis of \( L^2_{\pi_m}(\Gamma) \).

**Proof.** The assertion follows from Corollary 2.2.13 and Proposition 3.2.2. \( \square \)

The tensor product polynomial basis \( P \) is sometimes referred to as the generalized polynomial chaos basis and goes back to [XK02, Wie38]. We mention a couple of important examples of orthonormal polynomials. See e.g. [Sze75, Gau04] for further examples and an in-depth discussion.

**Example 3.2.4 (Legendre Polynomials).** If \( \Gamma_m = [-1, 1] \) for all \( m \in \mathcal{M} \) and \( \pi_m \) is the uniform probability measure on \( \Gamma_m \), i.e. \( d\pi_m = \frac{d\xi}{2} \), on the Borel \( \sigma \)-algebra \( \mathcal{B}([-1, 1]) \), then \( P_m \) consists of the normalized Legendre polynomials defined by Rodrigues’ formula

\[
P_m^m(\xi) = L_n(\xi) := \frac{\sqrt{2n + 1}}{2^n n!} \frac{d^n}{d\xi^n} (\xi^2 - 1)^n, \quad n \in \mathbb{N}_0.
\]

Normalized Legendre polynomials satisfy the three term recursion

\[
\frac{n + 1}{\sqrt{2n + 3} \sqrt{2n + 1}} L_{n+1}(\xi) = \xi L_n(\xi) - \frac{n}{\sqrt{2n + 1} \sqrt{2n - 1}} L_{n-1}(\xi), \quad n \in \mathbb{N}_0,
\]

with \( L_{-1} := 0 \). In particular, \( \alpha_n = 0 \) for all \( n \in \mathbb{N}_0 \) and

\[
\beta_n = \frac{n}{\sqrt{2n + 1} \sqrt{2n - 1}} = \frac{1}{\sqrt{4 - n^2}} \in \left( \frac{1}{2}, \frac{1}{\sqrt{3}} \right), \quad n \in \mathbb{N}.
\]
3.2. Tensor Product Polynomial Bases

Table 3.1: Recursion coefficients of orthonormal polynomials on \([-1, 1]\) w.r.t. \(w(\xi)\, d\xi\).

<table>
<thead>
<tr>
<th>Name</th>
<th>(w(\xi))</th>
<th>(\alpha_n)</th>
<th>(\beta_n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Legendre</td>
<td>(\frac{1}{\sqrt{\pi}}(1 - \xi^2)^{-1/2})</td>
<td>0</td>
<td>(\frac{\sqrt{4 - n^2}}{2}, n = 1)</td>
</tr>
<tr>
<td>Chebyshev #1</td>
<td>(\frac{2}{\pi}(1 - \xi^2)^{1/2})</td>
<td>0</td>
<td>(\frac{1}{2}, n = 2)</td>
</tr>
<tr>
<td>Chebyshev #2</td>
<td>(\frac{1}{\pi}(1 - \xi^{-1/2}(1 + \xi)^{1/2})</td>
<td>(\frac{1}{2}, n = 0)</td>
<td>(\frac{1}{2}, n = 2)</td>
</tr>
<tr>
<td>Chebyshev #3</td>
<td>(\frac{1}{\pi}(1 - \xi^{-1/2}(1 + \xi)^{-1/2})</td>
<td>(\frac{1}{2}, n = 0)</td>
<td>(\frac{1}{2}, n = 2)</td>
</tr>
<tr>
<td>Chebyshev #4</td>
<td>(\frac{1}{\pi}(1 - \xi^{-1/2}(1 + \xi)^{-1/2})</td>
<td>(\frac{1}{2}, n = 0)</td>
<td>(\frac{1}{2}, n = 2)</td>
</tr>
<tr>
<td>Gegenbauer, (\lambda &gt; -\frac{1}{2})</td>
<td>(\frac{\Gamma((1+1)\nu)}{\sqrt{\pi\Gamma(((1+2)\nu)}}(1 - \xi^2)^{\lambda-1/2})</td>
<td>0</td>
<td>(\frac{1}{\sqrt{2\nu+1}}, n = 1)</td>
</tr>
</tbody>
</table>

The first few Legendre polynomials are

\[ L_0(\xi) = 1, \quad L_1(\xi) = \sqrt{3} \xi, \quad L_2(\xi) = \frac{\sqrt{5}}{2} (3\xi^2 - 1). \] (3.2.9)

Corollary 3.2.3 implies that the tensor product Legendre polynomials \((L_\nu)_{\nu\in\Lambda}\) defined by (3.2.4) and (3.2.5) with (3.2.6) are an orthonormal basis of \(L^2_\pi([-1, 1]^\#)\), where \(\pi\) is the product measure from (2.2.1).

**Example 3.2.5 (Jacobi Polynomials).** Jacobi polynomials generalize Legendre polynomials to certain nonuniform distributions on \([-1, 1]\). For parameters \(a > -1\) and \(b > -1\), we consider the probability measure \(w(\xi)\, d\xi\) for the weight function

\[ w(\xi) = 2^{-(a+b+1)} \frac{\Gamma(a + b + 1)}{\Gamma(a + 1)\Gamma(b + 1)} (1 - \xi)^a (1 + \xi)^b, \quad \xi \in [-1, 1]. \] (3.2.10)

The Jacobi polynomials can be constructed through the recursion (3.2.2) with the coefficients

\[ \alpha_0 = \frac{b - a}{a + b + 2}, \quad \alpha_n = \frac{b^2 - a^2}{(2n + 1 + b)(2n + a + b + 2)}, \quad n \geq 1, \] (3.2.11)

and

\[ \beta_n = \begin{cases} \sqrt{\frac{4(a + 1)(b + 1)}{(a + b + 2)^3(a + b + 3)}} & \text{if } n = 1, \\ \sqrt{\frac{4n(n + a)(n + b)(n + a + b)}{(2n + a + b)^3(2n + a + b + 1)(2n + a + b - 1)}} & \text{if } n \geq 2. \end{cases} \] (3.2.12)

See Table 3.1 for the coefficients of a few particular cases of Jacobi polynomials.\(^2\)
Chapter 3. Transformation to a Discrete System

Example 3.2.6 (Hermite Polynomials). Let $\Gamma_m = \mathbb{R}$ for all $m \in \mathcal{M}$ and let $\pi_m$ be the standard Gaussian measure

$$d\pi_m(\xi) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\xi^2}{2}\right) \, d\xi, \quad m \in \mathbb{N}, \quad (3.2.13)$$

on the Borel $\sigma$-algebra $\mathcal{B}(\mathbb{R})$. Then the orthonormal polynomials are the normalized Hermite polynomials given by Rodrigues’ formula

$$p_n^m(\xi) = H_n(\xi) := \frac{(-1)^n}{\sqrt{n!}} \exp\left(\frac{\xi^2}{2}\right) \frac{d^n}{d\xi^n} \exp\left(-\frac{\xi^2}{2}\right), \quad n \in \mathbb{N}_0. \quad (3.2.14)$$

Hermite polynomials satisfy the three term recursion

$$\sqrt{n+1} H_{n+1}(\xi) = \xi H_n(\xi) - \sqrt{n} H_{n-1}(\xi) \quad n \in \mathbb{N}_0, \quad (3.2.15)$$

with $H_{-1} := 0$. The first few Hermite polynomials are

$$H_0(\xi) = 1, \quad H_1(\xi) = \xi, \quad H_2(\xi) = \frac{1}{\sqrt{2}} (\xi^2 - 1). \quad (3.2.16)$$

The tensorized Hermite polynomials $(H_{\nu})_{\nu \in \mathcal{I}}$ defined by (3.2.4) and (3.2.5) for (3.2.14) are an orthonormal basis of $L^2_{\nu}(\mathbb{R}^d)$ by Corollary 3.2.3. Here, the product measure $\pi$ from (2.2.1) is the centered Gaussian measure on $\mathbb{R}^d$ with Cameron–Martin space $\ell^2(\mathcal{M})$. ∎

3.2.2. Representation of Affinely Parametric Operators

We consider the setting of Sections 1.2.2 and 1.3.2, i.e. $\Gamma_m = [-1, 1]$ for all $m \in \mathcal{M}$ and

$$\mathcal{A} = \text{id}_{L^2_\nu(\Gamma)} \otimes D + \sum_{m \in \mathcal{M}} K_m \otimes R_m : L^2_{\nu}(\Gamma) \otimes_\alpha V \to L^2_{\nu}(\Gamma) \otimes_\alpha W^*, \quad (3.2.17)$$

where $\sigma$ is a tensor norm. We assume that condition (1.3.16) is satisfied, which implies that $\mathcal{A}$ is boundedly invertible, see Theorem 1.3.5.

Proposition 3.2.7. Let $\Phi = (\varphi_\nu)_{\nu \in \mathcal{I}}$ be an orthonormal basis of $L^2_{\nu}(\Gamma)$. Then

$$\mathcal{A} = T^W_\Phi \mathcal{A} T^V_\Phi = I \otimes D + \sum_{m \in \mathcal{M}} K_m \otimes R_m \quad (3.2.18)$$

for $K_m := T^W_\Phi K_m T^V_\Phi$ and $I := \text{id}_{\mathcal{L}(\mathcal{Y})}$, with convergence in $\mathcal{L}(L^2(\Lambda) \otimes_\alpha V, L^2(\Lambda) \otimes_\alpha W^*)$.

Proof. Since $\mathcal{A}$ is expandable due to (1.3.16), the sum in (3.2.17) converges in $\mathcal{L}(L^2_\nu(\Gamma) \otimes_\alpha V, L^2_\nu(\Gamma) \otimes_\alpha W^*)$. Therefore, using the definitions (2.1.17) and (2.1.21),

$$\mathcal{A} = (T^V_\Phi \otimes \text{id}_W)(T^\mathcal{A} \otimes D)(T^V_\Phi \otimes \text{id}_V) + \sum_{m \in \mathcal{M}} (T^V_\Phi \otimes \text{id}_W)(K_m \otimes R_m)(T^V_\Phi \otimes \text{id}_V)$$

$$= (T^V_\Phi T^V_\Phi) \otimes D + \sum_{m \in \mathcal{M}} (T^V_\Phi K_m T^V_\Phi) \otimes R_m,$$

with convergence in $\mathcal{L}(L^2(\Lambda) \otimes_\alpha V, L^2(\Lambda) \otimes_\alpha W^*)$. By assumption, $\Phi$ is an orthonormal basis of $L^2_\nu(\Gamma)$, and thus $T^V_\Phi$ is unitary, i.e. $T^V_\Phi = T^{-1}_\Phi$. □

44
Lemma 3.2.8. For any orthonormal basis $\Phi$ of $L^2_\pi(\Gamma)$, $K_m = T_\Phi K_m T_\Phi \in \mathcal{L}(\ell^2(\Lambda))$ with $K_m^* = K_m$ and
\[
\|K_m\|_{\ell^2(\Lambda)\to\ell^2(\Lambda)} \leq 1 \quad \forall m \in \mathcal{M}.
\]
Equality holds in (3.2.19) if $\pi_m([\xi \in [-1, 1] ; |\xi| \geq \epsilon]) > 0$ for all $\epsilon > 0$.

Proof. By Lemma 1.3.3, since $T_\Phi$ is unitary,
\[
\|K_m\|_{\ell^2(\Lambda)\to\ell^2(\Lambda)} = \|K_m\|_{L^2_\pi(\Gamma)\to L^2_\pi(\Gamma)} \leq 1,
\]
with equality if $\pi_m([\xi \in [-1, 1] ; |\xi| \geq \epsilon]) > 0$ for all $\epsilon > 0$. Furthermore, $K_m$ is self-adjoint since $K_m$ is self-adjoint. \qed

We consider the case of tensor product polynomial bases $\Phi = P$ from (3.2.4). The factors $P_m = (P^m_n)_{n \in \Lambda_m}$ form orthonormal bases of $L^2_{\pi_m}(\Gamma_m)$ with $\Gamma_m = [-1, 1]$ by Proposition 3.2.2. By the recursion formula (3.2.2),
\[
\xi P^m_n(\xi) = \rho^m_{n+1} P^m_{n+1}(\xi) + c^m_n P^m_n(\xi) + \alpha^m_{n+1} P^m_{n-1}(\xi), \quad n \in \Lambda_m, \quad \xi \in [-1, 1],
\]
where $P^m_n := 0$ for $n \in \mathbb{Z} \setminus \Lambda_m$.

Lemma 3.2.9. For all $m \in \mathcal{M}$, the operator $K_m = T_p K_m T_p$ has the form
\[
(K_m c)_\mu = \beta^m_{\mu+1} c_{\mu+\epsilon_m} + \alpha^m_{\mu} c_{\mu} + \beta^m_{\mu-1} c_{\mu-\epsilon_m}, \quad \mu \in \Lambda,
\]
for $c = (c_\mu)_{\mu \in \Lambda} \in \ell^2(\Lambda)$, where $\epsilon_m$ is the Kronecker sequence $(\epsilon_m)_n = \delta_{nm}$, and $c_\mu := 0$ if $\mu < 0$ for any $m \in \mathcal{M}$.

Proof. Let $c = (c_\mu)_{\mu \in \Lambda} \in \ell^2(\Lambda)$ and $m \in \mathcal{M}$. Then by (3.2.21),
\[
K_m T_p c = \sum_{\mu \in \Lambda} c_\mu K_m P_\mu = \sum_{\mu \in \Lambda} c_\mu (\beta^m_{\mu+1} P_{\mu+\epsilon_m} + \alpha^m_{\mu} P_{\mu} + \beta^m_{\mu-1} P_{\mu-\epsilon_m})
\]
\[
= \sum_{\mu \in \Lambda} (\rho^m_{\mu+1} c_{\mu+\epsilon_m} + \alpha^m_{\mu} c_{\mu} + \beta^m_{\mu-1} c_{\mu-\epsilon_m}) P_\mu.
\]
Equation (3.2.22) follows since $T_p^* = T_p^{-1}$. \qed

Remark 3.2.10. If $\pi_m$ is a symmetric measure on $[-1, 1]$, then $c^m_n = 0$ for all $n \in \Lambda_m$ by symmetry of the integral (3.2.3). In this case, the action of $K_m$ from Lemma 3.2.9 is simplified.
Chapter 3. Transformation to a Discrete System

Proposition 3.2.11. For all \( v = (v_\mu)_{\mu \in \Lambda} \in \ell^2(\Lambda) \otimes_\alpha V \),

\[
\forall v = \left( Dv_\mu + \sum_{m \in \mathcal{M}} R_m (\beta^m_{\mu m+1} v_{\mu+\epsilon_m} + \alpha^m_{\mu m} v_\mu + \beta^m_{\mu m} v_{\mu-\epsilon_m}) \right)_{\mu \in \Lambda}
\]  

(3.2.23)
in \( \ell^2(\Lambda) \otimes_\alpha W^* \), where \( v_\mu := 0 \) if \( \mu_m < 0 \) for any \( m \in \mathcal{M} \).

**Proof.** By Theorem A.3.5, \( \ell^2(\Lambda) \otimes_\alpha V \) and \( \ell^2(\Lambda) \otimes_\alpha W^* \) can be interpreted as spaces of sequences in \( V \) and \( W^* \), respectively. By continuity and linearity of \( \forall \), it suffices to show (3.2.23) for simple tensors \( v = c \otimes \nu \) with \( c = (c_\mu)_{\mu \in \Lambda} \in \ell^2(\Lambda) \) and \( \nu \in V \). In this case,

\[
\forall (c \otimes \nu) = c \otimes (D\nu) + \sum_{m \in \mathcal{M}} (K_m c) \otimes (R_m \nu)
\]

By Lemma 3.2.9, for all \( \mu \in \Lambda \),

\[
(\forall (c \otimes \nu))_\mu = c_\mu D\nu + \sum_{m \in \mathcal{M}} (\beta^m_{\mu m+1} c_{\mu+\epsilon_m} + \alpha^m_{\mu m} c_\mu + \beta^m_{\mu m} c_{\mu-\epsilon_m}) R_m \nu,
\]

which coincides with (3.2.23). \( \square \)

Let \( f \in L^2_\pi(\Gamma) \otimes_\alpha W^* \). Then by Theorem A.3.5, \( \check{f} := T^W f \in \ell^2(\Lambda) \otimes_\alpha W^* \) can be identified with a sequence \( \check{f} = (f_\nu)_{\nu \in \Lambda} \) in \( W^* \). By (3.2.24), \( f_\nu = (P_\nu \otimes \text{id}_{W^*}) f \), where \( P_\nu \in L^2_\pi(\Gamma) \) is interpreted as a continuous linear functional on \( L^2_\pi(\Gamma) \) by integration against its complex conjugate \( \bar{P}_\nu \). As in Remark 3.1.6,

\[
f_\nu = \int f \bar{P}_\nu \, d\pi \in W^*, \quad \nu \in \Lambda.
\]  

(3.2.24)

for a suitable interpretation of the integral.

**Theorem 3.2.12.** For tensor product polynomial bases \( \Phi = P \), the operator equation (3.1.8) is equivalent to the countably infinite system of equations

\[
Du_\nu + \sum_{m \in \mathcal{M}} R_m (\beta^m_{\nu m+1} u_{\nu+\epsilon_m} + \alpha^m_{\nu m} u_\nu + \beta^m_{\nu m} u_{\nu-\epsilon_m}) = f_\nu, \quad \nu \in \Lambda,
\]  

(3.2.25)

with \( u = (u_\nu)_{\nu \in \Lambda} \in \ell^2(\Lambda) \otimes_\alpha V \).

**Proof.** The assertion follows from Proposition 3.2.11 and \( \check{f} = (f_\nu)_{\nu \in \Lambda} \). \( \square \)

### 3.2.3. The Bi-Infinite Matrix Equation

In the setting of Section 3.2.2, we assume that \( V \) and \( W \) are separable Hilbert spaces, and that \( \alpha \) is the Hilbert tensor norm. Furthermore, let \( \Psi = (\psi_\kappa)_{\kappa \in \Xi} \) be a frame of \( V \) and \( \Theta = (\theta_\kappa)_{\kappa \in \Upsilon} \) a frame of \( W \). We define the frame representations of \( D \) and \( R_m \),

\[
D := T^\Theta DT^\Psi \quad \text{and} \quad R_m := T^\Theta R_m T^\Psi, \quad m \in \mathcal{M},
\]

(3.2.26)

which are bounded linear operators from \( \ell^2(\Xi) \) to \( \ell^2(\Upsilon) \), and as such can be interpreted as bi-infinite matrices.
3.3. Further Examples

**Proposition 3.2.13.** For any orthonormal basis $\Phi = (\varphi_{\nu})_{\nu \in \Lambda}$ of $L^2_\mu(\Gamma)$,

$$A = T^*_{\Phi \leq \Theta} \circ T_{\Phi \leq \Psi} = I \otimes D + \sum_{m \in \mathcal{M}} K_m \otimes R_m,$$  

(3.2.27)

with convergence in $\mathcal{L}(\ell^2(\Lambda \times \Xi), \ell^2(\Lambda \times \gamma))$. Furthermore,

$$A^* = T^*_{\Phi \leq \Theta} \circ T^*_{\Phi \leq \Psi} = I \otimes D^* + \sum_{m \in \mathcal{M}} K_m \otimes R_m^*,$$  

(3.2.28)

with convergence in $\mathcal{L}(\ell^2(\Lambda \times \gamma), \ell^2(\Lambda \times \Xi))$.

**Proof.** The first part of the assertion is a consequence of Proposition 3.2.7 and (3.1.14) due to Theorem 2.1.4. Equation (3.2.28) follows since $I^* = I$ and $K^*_m = K_m$, see Lemma 3.2.8.

In the case of tensor product polynomial bases $\Phi = P$ as in (3.2.4), the application of $A$ simplifies to

$$Av = \left(Dv_\mu + \sum_{\nu \in \Lambda} R_m (\beta^m_{\mu+1} v_{\mu+\epsilon_m} + \alpha^m_{\mu} v_\mu + \beta^{m\nu}_m v_{\mu-\epsilon_m})\right)_{\mu \in \Lambda},$$  

(3.2.29)

for $v = (v_\mu)_{\mu \in \Lambda} = (v_{\mu\nu})_{(\mu,\nu) \in \Lambda \times \Xi} \in \ell^2(\Lambda \times \Xi)$, and similarly for the adjoint $A^*$. This follows from Proposition 3.2.11 with $V$ and $W$ replaced by $\ell^2(\Xi)$ and $\ell^2(\gamma)$, respectively.

**Theorem 3.2.14.** For tensor product polynomial bases $\Phi = P$, the operator equation (3.1.18) is equivalent to the countably infinite system of equations

$$Du_\nu + \sum_{m \in \Lambda} R_m (\beta^m_{\nu+1} u_{\nu+\epsilon_m} + \alpha^m_{\nu} u_\nu + \beta^{m\nu}_m u_{\nu-\epsilon_m}) = f_\nu, \quad \nu \in \Lambda,$$  

(3.2.30)

with $u_\mu = (u_{\mu\nu})_{\nu \in \Xi} \in \ell^2(\Xi)$ and $f_\nu = (f_{\nu\kappa})_{\kappa \in \gamma} = T_{\Theta}^* f_\nu \in \ell^2(\gamma)$.

**Proof.** Since $(T^W_{\Theta})^* = id_{\Psi} \otimes T^*_{\Theta}$, we have $f_\nu = T_{\Theta}^* f_\nu$ for $f = (T^W_{\Theta})^* f$ and $f = (f_\nu)_{\nu \in \Lambda}$. The rest of the assertion follows from Proposition 3.2.13.

3.3. Further Examples

3.3.1. Piecewise Polynomial Bases

The tensor product polynomial basis $P$ defined in Section 3.2 can be generalized by dividing each domain $\Gamma_m$ into subdomains and considering orthonormal polynomials on each of these. In order to enforce Assumption 2.2.B, we need to treat piecewise constants separately.

As in Section 3.2, we assume $\Gamma_m \subset \mathbb{R}$ for all $m \in \mathcal{M}$, and the $\sigma$-algebra $\Sigma_m$ contains the Borel $\sigma$-algebra $\mathcal{B}(\Gamma_m)$. Furthermore, if $\pi_m$ has unbounded support, then it possesses moments of any order and is determinate.
Chapter 3. Transformation to a Discrete System

Let \( q_0^m := P_0^m \) be the constant function 1 on \( \Gamma_n \), and \( P_{n,0}^m := P_n^m \) the orthonormal polynomial of degree \( n \) for \( n \in \Lambda_{m,0} := \Lambda_m \setminus \{0\} \subset \mathbb{N} \). Then the orthonormal polynomials from Section 3.2 are

\[
P_m = (P_n^m)_{n \in \Lambda_m} = \{q_0^m\} \cup (P_n^m,0)_{n \in \Lambda_{m,0}}.
\]

In particular, \((P_n^m,0)_{n \in \Lambda_{m,0}}\) is an orthonormal basis of the space \( L^2_{\pi_m}(\Gamma_m) \) of square integrable functions on \( \Gamma_m \) with zero mean.

For any set \( I \subset \Gamma_m \), let \( P'(I) \) denote the space of equivalence classes in \( L^2_{\pi_m}(\Gamma_m) \) of polynomials on \( I \) with zero mean, and define \( I_{m,0} := \Gamma_m \). Let \( I_{m,(0,0)}, I_{m,(0,1)} \subset I_{m,0} \) be Borel sets with positive measure with respect to \( \pi_m \) such that \( I_{m,0} = I_{m,(0,0)} \cup I_{m,(0,1)} \). Define the Haar wavelet on this decomposition as the piecewise constant function

\[
\psi_0^m := \begin{cases} 
\sqrt{\frac{\partial \Theta_m}{\pi_m}} & \text{on } I_{m,(0,0)}, \\
-\sqrt{\frac{\partial \Theta_m}{\pi_m}} & \text{on } I_{m,(0,1)}.
\end{cases}
\]

Then \( \psi_0^m \) has mean zero and norm one in \( L^2_{\pi_m}(\Gamma_m) \). Furthermore, let \((P_n^{m,(0,0)})_{n \in \Lambda_{m,0}}\) and \((P_n^{m,(0,1)})_{n \in \Lambda_{m,0}}\) be hierarchical orthonormal polynomial bases of \( P'(I_{m,(0,0)}) \) and \( P'(I_{m,(0,1)}) \), respectively, in \( L^2_{\pi_m}(\Gamma_m) \), extended to \( \Gamma_m \) by zero. Note that the first element of each of these bases is a polynomial of degree one, not zero. Combined with the constant function \( q_0^m \) and the Haar wavelet \( \psi_0^m \), we have the set of functions

\[
\{q_0^m\} \cup \{\psi_0^m\} \cup (P_n^{m,(0,0)})_{n \in \Lambda_{m,0}} \cup (P_n^{m,(0,1)})_{n \in \Lambda_{m,0}}.
\]

We show below that this is an orthonormal basis of \( L^2_{\pi_m}(\Gamma_m) \). Further refinements of the domain \( I_{m,0} = I_{m,(0,0)} \cup I_{m,(0,1)} \) lead to further bases similar to (3.3.3).

In a general refinement step, we begin with a basis of \( L^2_{\pi_m}(\Gamma_m) \) of the form

\[
\{q_0^m\} \cup \{\psi_0^m\} \cup \bigcup_{\delta \in \partial \Theta_m} (P_n^{m,\delta})_{n \in \Lambda_{m,\delta}}.
\]

Clearly, (3.3.3) is of the form (3.3.4) with \( \Theta_m = \emptyset \) and \( \partial \Theta_m = \{(0,0),(0,1)\} \), and (3.3.1) has the same form with \( \Theta_m = \emptyset \) and \( \partial \Theta_m = \emptyset \). The general meaning of (3.3.4) will become clear from the following recursive construction.

We first mention that \( \Theta_m = \Theta_m^j \cup \partial \Theta_m \) is a nonempty set of finite sequences in \([0,1]\). Consider the partial order of such sequences given by \( \delta \leq \eta \) if there is a \( j \in \mathbb{N}_0 \) such that \( \delta \) is equal to the first \( j \) elements of \( \eta \). The set \( \Theta_m \) has the property that if \( \eta \in \Theta_m \) and \( \delta \leq \eta \), then \( \delta \in \Theta_m \). The set \( \partial \Theta_m \) consists of all the maximal elements of \( \Theta_m \) with respect to the partial order \( \leq \), and \( \Theta_m^i := \Theta_m \setminus \partial \Theta_m \) contains all non-maximal elements of \( \Theta_m \).

For each \( \delta \in \Theta_m \), we associate a set \( I_{m,\delta} \subset \Gamma_m \). These sets are hierarchical in the sense that if \( \delta \in \Theta_m^j \), then \( (\delta,0) \) and \( (\delta,1) \) are elements of \( \Theta_m \) and \( I_{m,\delta} = I_{m,(\delta,0)} \cup I_{m,(\delta,1)} \). In particular, using \( \emptyset \in \Theta_m \), we have the disjoint decomposition

\[
\Gamma_m = I_{m,0} = \bigsqcup_{\delta \in \partial \Theta_m} I_{m,\delta}.
\]
3.3. Further Examples

The set \((P^m_\gamma)_{\gamma \in \Lambda_m, \delta}\) is a hierarchical orthonormal polynomial basis of \(P^*(I_m, \delta)\) in \(L^2_{\pi_m}(\Gamma_m)\). Let \(\gamma_m \subset \partial \Theta_m\) be an arbitrary subset, and define

\[
\partial Y_m := \{(\delta, i) ; \; \delta \in Y_m, \; i \in \{0, 1\}\}.
\]  

For all \(\delta \in Y_m\), let \(I_{m,\delta,0}, I_{m,\delta,1} \subset I_{m,\delta}\) be Borel sets with positive measure with respect to \(\pi_m\) such that \(I_{m,\delta} = I_{m,\delta,0} \cup I_{m,\delta,1}\). Define the Haar wavelet on this decomposition of \(I_{m,\delta}\) as the piecewise constant function

\[
\psi^m_\delta := \begin{cases} 
\sqrt{\frac{\pi_m(I_{m,\delta,0})}{\pi_m(I_{m,\delta,1})\pi_m(I_{m,\delta})}} \text{ on } I_{m,\delta,0}, \\
-\sqrt{\frac{\pi_m(I_{m,\delta,1})}{\pi_m(I_{m,\delta,0})\pi_m(I_{m,\delta})}} \text{ on } I_{m,\delta,1}, \\
0 \text{ otherwise.}
\end{cases}
\]  

By construction, \(\psi^m_\delta\) has mean zero and norm one in \(L^2_{\pi_m}(\Gamma_m)\). For all \(\delta \in \partial Y_m\), let \((P^m_n)_{n \in \Lambda_m, \delta}\) be a hierarchical orthonormal polynomial basis of \(P^*(I_m, \delta)\) in \(L^2_{\pi_m}(\Gamma_m)\). Then the refined basis

\[
\{\varphi^m_0\} \cup (\psi^m_\delta)_{\delta \in \Theta_m \cup \partial Y_m} \cup \bigsqcup_{\delta \in \partial Y_m \cup \partial \Theta_m \setminus Y_m} (P^m_n)_{n \in \Lambda_m, \delta}
\]  

is of the form (3.3.4).

**Proposition 3.3.1.** All sets of the form (3.3.4) constructed by the above recursion are orthonormal bases of \(L^2_{\pi_m}(\Gamma_m)\).

**Proof.** All elements of (3.3.4) are normalized in \(L^2_{\pi_m}(\Gamma_m)\) by construction. The Haar wavelets defined in (3.3.7) have mean zero, and are therefore orthogonal to the constant \(\varphi^m_0\). Furthermore, for \(\delta, \eta \in \Theta_m\), either \(I_{m,\delta}\) and \(I_{m,\eta}\) are disjoint, or they are nested. In the first case, \(\psi^m_\delta\) and \(\psi^m_\eta\) are trivially orthogonal. In the second case, if \(I_{m,\delta} \subset I_{m,\eta}\), then \(\psi^m_\delta\) is constant on \(I_{m,\eta}\) by construction, so orthogonality follows since \(\psi^m_\eta\) has zero mean. Finally, the functions \(\varphi^m_0\) and \(\psi^m_\eta\) for \(\eta \in \Theta_m\) are constant on \(I_{m,\delta}\) for all \(\delta \in \partial \Theta_m\). Therefore, they are orthogonal to the polynomials \((P^m_n)_{n \in \Lambda_m, \delta}\), and these polynomials are orthogonal to each other by definition.

By construction, the set (3.3.4) spans the space \(P(I_m, \delta)\) of polynomial functions on \(I_m, \delta\) for \(\delta \in \partial \Theta_m\). The sum of these spaces contains \(P(\Gamma_m)\), which is dense in \(L^2_{\pi_m}(\Gamma_m)\) since, by assumption, \(\pi_m\) possesses moments of any order and is determinate if it has infinite support. □

**Corollary 3.3.2.** If \(\Phi_m\) is a basis of the form (3.3.4) for all \(m \in \mathcal{M}\), then \(\Phi\) defined in (2.2.13) is an orthonormal basis of \(L^2_{\pi}(\Gamma)\).

**Proof.** The assertion follows from Proposition 3.3.1 and Corollary 2.2.13. □

**Remark 3.3.3.** We note that (3.3.4) is an example of a boundedly hierarchic frame that is not graded if a basis function is considered a predecessor of another basis function if the former has a strictly larger support, or the same support and a strictly lower polynomial degree.
Chapter 3. Transformation to a Discrete System

Remark 3.3.4. By a similar argument as in Proposition 3.3.1, if \( \Phi_m \) consists of orthonormal polynomial bases on each set \( I_m, \delta \) for \( \delta \in \partial \Theta_m \), then \( \Phi_m \) is an orthonormal basis of \( L^2_{\pi_m}(\Gamma_m) \), and Proposition 2.2.6 implies that \( \Phi_I \) is an orthonormal basis of \( L^2_{\pi_m}(I) \) for all \( I \in \mathcal{F} \). However, since Assumption 2.2.B is not satisfied, these bases lack the hierarchic structure described in Section 2.2.3 that is used in Theorem 2.2.10. Bases of this form are used in [WK05, WK06] under the name multi-element generalized polynomial chaos.

3.3.2. Rescaled Uniform Bases

Consider the case \( \Gamma_m \subset \mathbb{R} \) for all \( m \in \mathcal{M} \), and the \( \sigma \)-algebra \( \Sigma_m \) contains the Borel \( \sigma \)-algebra \( \mathcal{B}(\Gamma_m) \). Without loss of generality, we assume \( \Gamma_m = \mathbb{R} \) by extending \( \pi_m \) to \( \mathbb{R} \). Furthermore, we assume that \( \Sigma_m \) is contained in the \( \sigma \)-algebra of Lebesgue-measurable sets and \( \pi_m \) is absolutely continuous with respect to the Lebesgue measure \( \lambda \).

Let \( F_m \) be the distribution function of \( \pi_m \), i.e. \( F_m(z) := \pi_m((0, z]) \), \( z \in \mathbb{R} \). Then \( F_m \) maps \( \mathbb{R} \) onto \([0, 1]\) monotonically and continuously. Its \( \pi_m \)-a.e. inverse is given by the quantile function \( G_m : [0, 1] \to \mathbb{R} \),

\[
G_m(t) := \sup \{ z \in \mathbb{R} ; F_m(z) < t \} \quad , \quad t \in [0, 1] .
\]  

(3.3.9)

Lemma 3.3.5. \( F_m(G_m(t)) = t \) for all \( t \in [0, 1] \), and \( G_m(F_m(z)) = z \) for \( \pi_m \)-a.e. \( z \in \mathbb{R} \).

Proof. Since \( \pi_m \ll \lambda \), there is a function \( f_m \in L^1_\lambda(\mathbb{R}) \) such that \( d\pi_m = f_m d\lambda \), and \( f_m \geq 0 \). Equivalently, \( F_m \) is absolutely continuous and \( F_m' = f_m \lambda \)-almost everywhere. By definition, \( f_m > 0 \) \( \pi_m \)-almost everywhere, so consequently, \( F_m'(z) > 0 \) for \( \pi_m \)-a.e. \( z \in \mathbb{R} \). For all such \( z \), using monotonicity and continuity of \( F_m \),

\[
G_m(F_m(z)) = \sup \{ w \in \mathbb{R} ; F_m(w) < F_m(z) \} = z .
\]

The first part of the assertion follows from the monotonicity and surjectivity of \( F_m \) since for all \( t \in [0, 1] \),

\[
F_m(G_m(t)) = F_m(\sup \{ z \in \mathbb{R} ; F_m(z) < t \}) = \sup \{ F_m(z) ; F_m(z) < t \} , \quad z \in \mathbb{R} \} = t .
\]

\[ \square \]

For all \( m \in \mathcal{M} \), let \( \Psi_m := (\varphi^m_k)_{k \in A_m} \) be a tight frame of the Lebesgue space \( L^2_\lambda([0, 1]) \). For example, \( \Psi_m \) could be an orthonormal basis of \( L^2_\lambda([0, 1]) \). Define the functions

\[
\varphi^m_k(z) := \psi^m_k(F_m(z)) , \quad z \in \mathbb{R} , \quad k \in A_m ,
\]  

(3.3.10)

and \( \Phi_m := (\varphi^m_k)_{k \in A_m} \).

Proposition 3.3.6. \( \Phi_m \) is a tight frame of \( L^2_{\pi_m}(\mathbb{R}) \) for all \( m \in \mathcal{M} \). If \( \Psi_m \) is an orthonormal basis of \( L^2_\lambda([0, 1]) \), then \( \Phi_m \) is an orthonormal basis of \( L^2_{\pi_m}(\mathbb{R}) \).

Proof. Note that the image measure \( F_m(\pi_m) \) is the Lebesgue measure \( \lambda \) on \([0, 1] \) since for all \( t \in [0, 1] \),

\[
\pi_m(F_m^{-1}([0, t])) = \pi_m((0, t)) = \lambda([0, t]) = F_m(G_m(t)) = F_m(t) .
\]
Therefore, for any \( v \in L^1_\Lambda([0, 1]) \), \( v \circ F_m \in L^1_{\pi_m}(\mathbb{R}) \) and

\[
\int_{\mathbb{R}} v(F_m(z)) \, d\pi_m(z) = \int_{[0,1]} v(t) \, dF_m(\pi_m)(t) = \int_0^1 v(t) \, dt.
\]

Applying this to \( v = w^2 \) for \( w \in L^2_\Lambda([0, 1]) \) leads to an isometry

\[
\zeta : L^2_\Lambda([0, 1]) \to L^2_{\pi_m}(\mathbb{R}), \quad w \mapsto w \circ F_m.
\]

Since the inverse of \( \zeta \) is \( u \mapsto u \circ G_m \) by Lemma 3.3.5, \( \zeta \) is surjective and thus \( T\Phi_m = \zeta T\Psi_m \) is also surjective, which shows that \( \Phi_m \) is a frame, and it is a Riesz basis if \( \Psi_m \) is a Riesz basis. The rest of the assertion follows since \( \zeta \) and \( \zeta^{-1} \) are isometries. \( \square \)

**Corollary 3.3.7.** For all \( m \in \mathcal{M} \), let \( \Psi_m = (\psi^m_k)_{k \in \Lambda_m} \) be a tight frame of \( L^2_\Lambda([0, 1]) \) with \( 0 \in \Lambda_m \) and \( \psi^m_0 = 1 \) satisfying Assumption 2.2.C. If \( \Phi_m \) is induced by \( \Psi_m \) through (3.3.10) for all \( m \in \mathcal{M} \), then \( \Phi \) given by (2.2.13) is a tight frame of \( L^2_\Lambda(\mathcal{T}) \). If \( \Psi_m \) are orthonormal bases, then so is \( \Phi \).

**Proof.** Note that \( \Phi_m \) satisfies Assumptions 2.2.B and 2.2.C since \( \phi^m_0 = 1 \circ F_m = 1 \) and \( K_{\Phi_m} = K_{\Psi_m} \subset L^2(\Lambda_m) \) for all \( m \in \mathcal{M} \). Then the assertion follows from Proposition 3.3.6 and Theorem 2.2.10. \( \square \)

**Example 3.3.8 (Legendre Polynomials).** The Legendre polynomials

\[
L_n(t) := \frac{\sqrt{2n + 1}}{n!} \frac{d^n}{dt^n}(t^2 - 1)^n, \quad t \in [0, 1], \quad n \in \mathbb{N}_0,
\]

form an orthonormal basis of \( L^2_\Lambda([0, 1]) \). Of course, the resulting basis functions \( \phi^m_k = L_n \circ F_m \) are generally not polynomials.

**Example 3.3.9 (Haar Wavelets).** Define the scaling function and mother wavelet on \([0, 1]\) by

\[
\psi(t) := 1, \quad \varphi(t) := \begin{cases} 
1, & 0 \leq t < 1/2, \\
-1, & 1/2 \leq t < 1,
\end{cases}
\]

and extend these functions to \( \mathbb{R} \) by zero. Then the Haar wavelets on \([0, 1]\) are

\[
\psi_n,k(t) := 2^n t^k \psi(2^{-n-1}t - k), \quad n \in \mathbb{N}, \quad k = 0, \ldots, 2^n - 1,
\]

and \( \varphi_0 := \varphi = 1 \). The set \( \Psi_m = (\psi_n,k)_{n,k} \) is an orthonormal basis of \( L^2_\Lambda([0, 1]) \), consisting of the piecewise constant functions \( \phi^m_n = \psi_n,k \circ F_m \) on \( \mathbb{R} \). This construction can be generalized to orthonormal piecewise polynomial (multi)wavelets of arbitrary order.

**Example 3.3.10 (Fourier Basis).** The functions

\[
\psi_0(t) := 1, \quad \psi_{2k}(t) := \sqrt{2} \cos(2\pi kt), \quad \psi_{2k+1}(t) := \sqrt{2} \sin(2\pi kt), \quad k \in \mathbb{N},
\]

form an orthonormal basis of \( L^2_\Lambda([0, 1]) \). By Corollary 3.3.7, the tensor products of the compositions \( \phi_n \circ F_m \) are an orthonormal basis of the product space \( L^2_\Lambda(\mathcal{T}) \).
Chapter 3. Transformation to a Discrete System

For constructions of tight frames of $L_\lambda^2([0, 1])$ that are not orthonormal bases, we refer to [CS08, CHS04, LN06, Rei08]. For our purposes, we would need to adapt these constructions to ensure Assumptions 2.2.B and 2.2.C.
Chapter 4.

Adaptive Application of the Discrete Operator

Adaptive wavelet methods are based on the ability to efficiently approximate the action of a linear operator on a vector, both discretized in a wavelet basis, up to an arbitrary prescribed tolerance. Such approximate application routines are used directly as a perturbed application of the operator in an iterative method, and also to estimate residuals in the Galerkin methods. They control the refinement of the adaptive discretization and permit an error bound for verifying the termination criterion.

Generally, these adaptive application routines make use of a sequence of sparse approximations to the bi-infinite matrix representing an operator in a wavelet basis. To apply the operator to a vector, the latter is partitioned according to the magnitude of its elements. More accurate approximations of the operator are then used for more significant parts of the vector. Details are presented in Section 4.2.

Wavelets are relevant to adaptive wavelet methods only in that they provide sparse approximations for many differential and integral operators. The adaptive methods themselves are abstract procedures, acting only on numerical vectors of coefficients. They extend to any linear operator equation discretized by a Riesz basis, provided that an adaptive application routine is available.

We develop just such a routine for discretized parametric operators in Section 4.3. Therefore, adaptive wavelet methods such as [CDD01, CDD02, GHS07, DSS09] can be applied to parametric operator equations with a wavelet discretization on the physical domain. Similarly, the adaptive frame methods [Ste03, DFR07, DRW+07] can be employed in conjunction with a frame discretization. In either case, the resulting algorithm is fully adaptive, approximating each coefficient function in a polynomial expansion of the parametric solution by a different set of basis functions.

We begin with a digression. At multiple points in the adaptive method, we are faced with maximizing a sum of discrete values under a linear constraint. Greedy methods provide a simple but highly effective solution procedure, which we discuss in Section 4.1.
Chapter 4. Adaptive Application of the Discrete Operator

4.1. Greedy Solution of a Generalized Knapsack Problem

4.1.1. Problem Setting

We consider a discrete optimization problem in which both the objective and the constraints are given by sums over an arbitrary set \( \mathcal{M} \subset \mathbb{N}_0 \). For each \( m \in \mathcal{M} \), we have two increasing sequences \((c_j^m)_{j \in \mathbb{N}_0}\) and \((\omega_j^m)_{j \in \mathbb{N}_0}\), which we interpret as costs and values. We define the total cost of a \( j = (j_m)_{m \in \mathcal{M}} \in \mathbb{N}^\mathcal{M}_0 \) as

\[
c_j := \sum_{m \in \mathcal{M}} c_j^m, \tag{4.1.1}
\]

and the total value of \( j \) as

\[
\omega_j := \sum_{m \in \mathcal{M}} \omega_j^m. \tag{4.1.2}
\]

Our goal is to maximize \( \omega_j \) under a constraint on \( c_j \), or to minimize \( c_j \) under a constraint on \( \omega_j \).

Remark 4.1.1. The above two goals are essentially equivalent. If \( j \in \mathbb{N}_0^\mathcal{M} \) such that for all \( i \in \mathbb{N}_0^\mathcal{M} \), \( c_i \leq c_j \) implies \( \omega_i \leq \omega_j \), then by contraposition, \( \omega_i > \omega_j \) implies \( c_i > c_j \). Similarly, if \( c_i < c_j \) implies \( \omega_i < \omega_j \), then also \( \omega_i \geq \omega_j \) implies \( c_i \geq c_j \). In both cases, the two statements are equivalent.

Remark 4.1.2. The classical knapsack problem is equivalent to the above optimization problem in the case that \( \mathcal{M} \) is finite, and for all \( m \in \mathcal{M} \), \( c_0^m = 0 \) and \( \omega_1^m = \omega_j^m \) for all \( j \geq 1 \). Then without loss of generality, we can set \( c_0^m := 0 \) for all \( m \in \mathcal{M} \), and the values \( c_j^m \) for \( j \geq 2 \) are irrelevant due to the assumption that \( (c_j^m)_{j \in \mathbb{N}_0} \) is increasing. Optimal sequences \( j \in \mathbb{N}_0^\mathcal{M} \) will only take the values 0 and 1, and can thus be interpreted as subsets of \( \mathcal{M} \).

Remark 4.1.3. We are interested in particular in minimizing an error under constraints on the computational cost of an approximation with this error tolerance. Given sequences \((e_j^m)_{j \in \mathbb{N}_0}\) and \((\epsilon_j^m)_{j \in \mathbb{N}_0}\) of errors and corresponding costs, we define a sequence of values by \( \omega_j^m \) := \( -\epsilon_j^m \). If \((e_j^m)_{j \in \mathbb{N}_0}\) is decreasing, then \((\omega_j^m)_{j \in \mathbb{N}_0}\) is increasing. Typically, as \( j \to \infty \), we have \( e_j^m \to 0 \) and \( c_j^m \to \infty \). Then, although it is increasing, \((\omega_j^m)_{j \in \mathbb{N}_0}\) remains bounded. In particular, it is reasonable to assume that \((\omega_j^m)_{j \in \mathbb{N}_0}\) increases more slowly than \((e_j^m)_{j \in \mathbb{N}_0}\), in a sense that is made precise below.

4.1.2. A Sequence of Optimal Solutions

We iteratively construct a sequence \((f^k)_{k \in \mathbb{N}_0}\) in \( \mathbb{N}_0^\mathcal{M} \) such that, under some assumptions, each \( f^k \) is optimal in the sense of Remark 4.1.1. For all \( m \in \mathcal{M} \) and all \( j \in \mathbb{N}_0 \), let

\[
\Delta c_j^m := c_{j+1}^m - c_j^m \quad \text{and} \quad \Delta \omega_j^m := \omega_{j+1}^m - \omega_j^m. \tag{4.1.3}
\]
4.1. Greedy Solution of a Generalized Knapsack Problem

Furthermore, let $q^m_j$ denote the quotient of these two increments,

$$q^m_j := \frac{\Delta \omega^m_j}{\Delta c^m_j}, \quad j \in \mathbb{N}_0,$$

(4.1.4)

which can be interpreted as the value to cost ratio of passing from $j$ to $j+1$ in the index $m \in \mathcal{M}$.

Let $j^0 := 0 \in \mathbb{N}_0$. For all $k \in \mathbb{N}_0$, we construct $j^{k+1}$ from $j^k$ as follows. Let $m_k = m \in \mathbb{N}_0$ maximize $q^m_{j^k}$. Existence of such maxima is ensured by the last statement in Assumption 4.1.A. If the maximum is not unique, select $m_k$ to be minimal among all maxima. Then define $j^{k+1}_m := j^k_m + 1$, and set $j^{k+1} := j^{k+1}_m$ for all $m \neq m_k$. For this sequence, we abbreviate $c^k := c^k_j$ and $\omega^k := \omega^k_j$.

Assumption 4.1.A. For all $m \in \mathcal{M}$,

$$c^m_0 = 0 \quad \text{and} \quad \Delta c^m_j > 0 \quad \forall j \in \mathbb{N}_0,$$

(4.1.5)

i.e. $(c^m_j)_{j \in \mathbb{N}_0}$ is strictly increasing. Also, $(\omega^m_{j^k})_{j^k \in \mathcal{M}} \in \ell^1(\mathcal{M})$ and $(\omega^m_{j^k})_{j^k \in \mathbb{N}_0}$ is nondecreasing for all $m \in \mathcal{M}$, i.e. $\Delta \omega^m_j \geq 0$ for all $j \in \mathbb{N}_0$. Furthermore, for each $m \in \mathcal{M}$, the sequence $(q^m_j)_{j \in \mathbb{N}_0}$ is nonincreasing, i.e. if $i \geq j$, then $q^m_i \leq q^m_j$. Finally, for any $\varepsilon > 0$, there are only finitely many $m \in \mathcal{M}$ for which $q^m_0 \geq \varepsilon$.

The assumption that $(q^m_j)_{j \in \mathbb{N}_0}$ is nonincreasing is equivalent to

$$\frac{\Delta \omega^m_i}{\Delta \omega^m_j} \leq \frac{\Delta c^m_i}{\Delta c^m_j} \quad \text{if} \quad i \geq j$$

(4.1.6)

if $\Delta \omega^m_j > 0$. In this sense, $(\omega^m_j)_{j \in \mathbb{N}_0}$ increases more slowly than $(c^m_j)_{j \in \mathbb{N}_0}$. Also, this assumption implies that if $\Delta \omega^m_j = 0$, then $\omega^m_i = \omega^m_j$ for all $i \geq j$.

We define a total order on $\mathcal{M} \times \mathbb{N}_0$ by

$$(m, j) < (n, i) \quad \text{if} \quad \begin{cases} q^m_j > q^n_i \quad \text{or} \\ q^m_j = q^n_i \quad \text{and} \quad m < n \quad \text{or} \\ q^m_j = q^n_i \quad \text{and} \quad m = n \quad \text{and} \quad j < i. \end{cases}$$

(4.1.7)

To any sequence $j = (j_m)_{m \in \mathcal{M}}$ in $\mathbb{N}_0$, we associate the set

$$\|j\| := \{(m, j) \in \mathcal{M} \times \mathbb{N}_0 : j < j_m\}.$$  

(4.1.8)

Lemma 4.1.4. For all $k \in \mathbb{N}_0$, $\|k\| := \|j^k\|$ consists of the first $k$ terms of $\mathcal{M} \times \mathbb{N}_0$ with respect to the order $\prec$.  

55
Chapter 4. Adaptive Application of the Discrete Operator

Proof. The assertion is trivial for \( k = 0 \). Assume it holds for some \( k \in \mathbb{N}_0 \). By definition,
\[
\|k + 1\| = \|k\| \cup \{(m_k, j_m)\},
\]
and \((m_k, j_m)\) is the \(<\)-minimal element of the set \( \{(m, j_m) ; m \in \mathcal{M}\} \). For each \( m \in \mathcal{M} \), Assumption 4.1.A implies \( q_i^m \leq q_j^m \) for all \( i \geq j_m + 1 \). Therefore, \((m, j_m) < (m, i)\) for all \( i \geq j_m + 1 \), and consequently \((m_k, j_m)\) is the \(<\)-minimal element of \( \mathcal{M} \times \mathbb{N}_0 \setminus \|k\| \). □

Theorem 4.1.5. For all \( k \in \mathbb{N}_0 \), the sequence \( j^k \) maximizes \( \omega_j \) among all finitely supported sequence \( j = (j_m)_{m \in \mathcal{M}} \) in \( \mathbb{N}_0 \) with \( c_j \leq c_k \). Furthermore, if \( c_j < c_k \) and there exist \( k \) pairs \((m, i) \in \mathcal{M} \times \mathbb{N}_0 \) with \( \Delta \omega_i^m > 0 \), then \( \omega_j < \omega_k \).

Proof. Let \( k \in \mathbb{N} \) and let \( j = (j_m)_{m \in \mathcal{M}} \) be a finitely supported sequence in \( \mathbb{N}_0 \) with \( c_j \leq c_k \). By definition,
\[
\omega_j = \sum_{m \in \mathcal{M}} \omega_0^m + \sum_{m \in \mathcal{M}} \sum_{i=0}^{j_m-1} q_i^m \Delta c_i^m = \omega_p + \sum_{(m, j) \in \mathcal{J}} q_i^m \Delta c_i^m.
\]

Therefore, the assertion reduces to
\[
\sum_{(m, j) \in \mathcal{J} \setminus \|k\|} q_i^m \Delta c_i^m \leq \sum_{(m, j) \in \mathcal{J} \setminus \|k\|} q_i^m \Delta c_i^m.
\]

Note that by (4.1.1) and (4.1.3),
\[
\sum_{(m, j) \in \mathcal{J} \setminus \|k\|} \Delta c_i^m = c_j - c' \quad \text{for} \quad c' := \sum_{(m, j) \in \mathcal{J} \setminus \|k\|} \Delta c_i^m.
\]

By Lemma 4.1.4 and (4.1.7), \( q := \sum_{(m, j) \in \mathcal{J} \setminus \|k\|} q_i^m \Delta c_i^m \) satisfies \( q \leq q_i^m \) for all \((m, i) \in \|k\|\), and \( q_i^m \leq q \) for all \((m, i) \in \mathcal{M} \times \mathbb{N}_0 \setminus \|k\|\). In particular, \( q > 0 \) if there exist \( k \) pairs \((m, i) \in \mathcal{M} \times \mathbb{N}_0 \) with \( q_i^m > 0 \) since \( \#\|k\| = k \). Consequently,
\[
\sum_{(m, j) \in \mathcal{J} \setminus \|k\|} q_i^m \Delta c_i^m \leq q \sum_{(m, j) \in \mathcal{J} \setminus \|k\|} \Delta c_i^m = q(c_j - c') \leq q(c_k - c') \leq \sum_{(m, j) \in \mathcal{J} \setminus \|k\|} q_i^m \Delta c_i^m,
\]

and this inequality is strict if \( q > 0 \) and \( c_k > c_j \). □

The optimality property in Theorem 4.1.5 can be reinterpreted as in Remark 4.1.1, i.e. \( j^k \) also minimizes \( c_j \) among \( j \) with \( \omega_j \geq \omega_k \).
4.1. Greedy Solution of a Generalized Knapsack Problem

4.1.3. Numerical Construction

We consider numerical methods for constructing the sequence \((j^k)_{k \in \mathbb{N}_0}\) from Section 4.1.2. We assume that, for each \(m \in \mathcal{M}\), the sequences \((c^m_j)_{j \in \mathbb{N}_0}\) and \((\omega^m_j)_{j \in \mathbb{N}_0}\) are stored as linked lists.

Initially, we consider the case that \(\mathcal{M}\) is finite with \(\# \mathcal{M} =: M\). To construct \((j^k)_{k \in \mathbb{N}_0}\), we use a list \(\mathcal{N}\) of the triples \((m, j^k_m, q^m_{j^k_m})\), sorted in ascending order with respect to \(\prec\). This list may be realized as a linked list or as a tree. The data structure must provide functions for removing the minimal element from the list, and for inserting new elements.

\[
\text{NextOpt}[j, \mathcal{N}] \mapsto [j_m, \mathcal{N}]
\]

\[
m \leftarrow \text{PopMin}(\mathcal{N})
\]

\[
j_m \leftarrow j_m + 1
\]

\[
q \leftarrow (\omega^m_{j_m+1} - \omega^m_{j_m})/(c^m_{j_m+1} - c^m_{j_m})
\]

\[
\mathcal{N} \leftarrow \text{Insert}(\mathcal{N}, (m, j_m, q))
\]

**Proposition 4.1.6.** Let \(\mathcal{N}_0\) be initialized as \(\{(m, 0, q^m_0) ; m \in \mathcal{M}\}\) and \(j^0 := 0 \in \mathbb{N}_0\). Then the recursive application of

\[
\text{NextOpt}[j^k, \mathcal{N}_k] \mapsto [j^{k+1}, m_k, \mathcal{N}_{k+1}]
\]

constructs the sequence \((j^k)_{k \in \mathbb{N}_0}\) as defined above. Initialization of the data structure \(\mathcal{N}_0\) requires \(O(M \log M)\) operations and \(O(M)\) memory. One step of (4.1.9) requires \(O(M)\) operations if \(\mathcal{N}\) is realized as a linked list, and \(O(\log M)\) operations if \(\mathcal{N}\) is realized as a tree. The total number of operations required by the first \(k\) steps is \(O(kM)\) in the former case and \(O(k \log M)\) in the latter. In both cases, the total memory requirement for the first \(k\) steps is \(O(M + k)\).

**Proof.** Recursive application of NextOpt as in (4.1.9) constructs the sequence \((j^k)_{k \in \mathbb{N}_0}\) by Lemma 4.1.4 and the definition of \(\prec\). In the \(k\)-th step, the element \(m_k\) is removed from \(\mathcal{N}\) and reinserted in a new position. Therefore, the size of \(\mathcal{N}\) remains constant at \(M\). The computational cost of (4.1.9) is dominated by the insert operation on \(\mathcal{N}\), which has the complexity stated above. \(\square\)

We turn to the case that \(\mathcal{M}\) is countably infinite. By enumerating the elements of \(\mathcal{M}\), it suffices to consider the case \(\mathcal{M} = \mathbb{N}\). We assume in this case that the sequence \((q^m_{j^k_m})_{m \in \mathcal{M}}\) is nonincreasing.

As above, we use a list \(\mathcal{N}\) of triples \((m, j^k_m, q^m_{j^k_m})\) to construct the sequence \((j^k)_{k \in \mathbb{N}_0}\). However, \(\mathcal{N}\) should only store triples for which \(m\) is a candidate for the next value of \(m_k\), i.e. all \(m\) with \(j^k_m \neq 0\) and the smallest \(m\) with \(j^k_m = 0\). As in the finite case, \(\mathcal{N}\) can be realized as a linked list or a tree. The data structure should provide functions for removing the smallest element with respect to the ordering \(\prec\), and for inserting a new element.
Chapter 4. Adaptive Application of the Discrete Operator

\[
\text{NextOptInf}[j, N, M] \mapsto [j, m, N, M]
\]

- \( m \leftarrow \text{PopMin}(N) \)
- \( j_m \leftarrow j_m + 1 \)
- \( q \leftarrow (\omega^m_{j_m+1} - \omega^m_{j_m})/(c^m_{j_m+1} - c^m_{j_m}) \)
- \( N \leftarrow \text{Insert}(N, (m, j_m, q)) \)
  
  if \( m = M \) then
  
    \( M \leftarrow M + 1 \)
    \( q \leftarrow (\omega^M_1 - \omega^M_0)/c^M_1 \)
    \( N \leftarrow \text{Insert}(N, (M, 1, q)) \)
  
end

Proposition 4.1.7. Let \( N_0 \) be initialized as \( \{(1, 0, q^1)\} \), \( M_0 := 1 \) and \( f^0 := 0 \in \mathbb{N}_0^\infty \). Then the recursion

\[
\text{NextOptInf}[f^k, N_k, M_k] \mapsto [f^{k+1}, m_k, N_{k+1}, M_{k+1}]
\]  \hspace{1cm} (4.1.10)

constructs the sequence \((f^k)_{k \in \mathbb{N}_0}\) as defined above. For all \( k \in \mathbb{N}_0 \), the ordered set \( N_k \) contains exactly \( M_k \) elements, and \( M_k \leq k \). The \( k \)-th step of (4.1.10) requires \( \mathcal{O}(k) \) operations if \( N \) is realized as a linked list, and \( \mathcal{O}(\log k) \) operations if \( N \) is realized as a tree. The total number of operations required by the first \( k \) steps is \( \mathcal{O}(k^2) \) in the former case and \( \mathcal{O}(k \log k) \) in the latter. In both cases, the total memory requirement for the first \( k \) steps is \( \mathcal{O}(k) \).

Proof. It follows from the definitions that recursive application of \( \text{NextOptInf} \) as in (4.1.10) constructs the sequence \((f^k)_{k \in \mathbb{N}_0}\). In the \( k \)-th step, the element \( m_k \) is removed from \( N \) and reinserted in a new position. If \( m_k = M \), an additional element is inserted, and \( M \) is incremented. Therefore, the number of elements in \( N_k \) is \( M_k \), and \( M_k \leq k \). The computational cost of (4.1.10) is dominated by the insert operation on \( N_k \), which has the complexity stated above, see e.g. [CLRS09]. \( \square \)

Remark 4.1.8. As mentioned above, \((c^m_j)_{j \in \mathbb{N}_0} \) and \((\omega^m_j)_{j \in \mathbb{N}_0} \) are assumed to be stored in a linked list for each \( m \in \mathcal{M} \). By removing the first element from the \( \mathcal{M}_k \)-th list in the \( k \)-th step of (4.1.9) or (4.1.10), \( \text{NextOpt} \) and \( \text{NextOptInf} \) only ever access the first two elements of one of these lists, which takes \( \mathcal{O}(1) \) time. The memory locations of the lists can be stored in a hash table for efficient access. \( \square \)

Remark 4.1.9. An appropriate way to store \((f^k)_{k \in \mathbb{N}_0}\) is to collect \((m_k)_{k \in \mathbb{N}_0}\) in a linked list. Then \( f^k \) can be reconstructed by reading the first \( k \) elements of this list, which takes \( \mathcal{O}(k) \) time independently of the size of the list. Also, the total memory requirement is \( \mathcal{O}(\bar{k}) \) if the first \( \bar{k} \) elements are stored. \( \square \)
4.2. Adaptive Application of $s^*$-Computable Operators

4.2.1. $s^*$-Compressibility and $s^*$-Computability

Let $X$ and $Y$ be Banach spaces, and let $\Xi$ and $\Theta$ be countable sets. For tensor norms $\alpha$ and $\beta$, we consider a bounded linear operator

$$\mathcal{A} : L^2(\Xi) \otimes_\alpha X \to L^2(\Theta) \otimes_\beta Y.$$  \hspace{1cm} (4.2.1)

Any such operator $\mathcal{A}$ can be identified with a bi-infinite operator matrix $[A_{\nu\mu}]_{\mu \in \Xi, \nu \in \Theta}$ with $A_{\nu\mu} \in \mathcal{L}(X, Y)$ via

$$\mathcal{A} x = \left( \sum_{\mu \in \Xi} A_{\nu\mu} x_{\mu} \right)_{\nu \in \Theta}$$ \hspace{1cm} (4.2.2)

for $x = (x_{\mu})_{\mu \in \Xi} \in L^2(\Xi) \otimes_\alpha X$, with unconditional convergence in $L^2(\Theta) \otimes_\beta Y$. This follows as in Theorem 3.1.5 and is due to the sequence space structure of $L^2(\Xi) \otimes_\alpha X$ and $L^2(\Theta) \otimes_\beta Y$; see Theorem A.3.5. The operators $A_{\nu\mu}$ are given by

$$A_{\nu\mu} x = (p_{\nu} \otimes \text{id}_Y) \mathcal{A}(\epsilon_{\mu} \otimes x), \quad x \in X,$$ \hspace{1cm} (4.2.3)

where $\epsilon_{\mu}$ is the Kronecker sequence on the set $\Xi$ for the index $\mu \in \Xi$ and $p_{\nu}$ is the projection in $L^2(\Theta)$ onto the coordinate $\nu \in \Theta$.

We call the operator $\mathcal{A}$ $n$-sparse if for each $\mu \in \Xi$, at most $n$ of the maps $A_{\nu\mu}$, $\nu \in \Theta$, are nontrivial, and sparse if it is $n$-sparse for some $n \in \mathbb{N}_0$.

**Definition 4.2.1.** An operator $\mathcal{A} \in \mathcal{L}(L^2(\Xi) \otimes_\alpha X, L^2(\Theta) \otimes_\beta Y)$ is $s^*$-compressible for an $s^* \in (0, \infty]$ if there exists a sequence $(\mathcal{A}_j)_{j \in \mathbb{N}}$ in $\mathcal{L}(L^2(\Xi) \otimes_\alpha X, L^2(\Theta) \otimes_\beta Y)$ such that $\mathcal{A}_j$ is $n_j$-sparse with $(n_j)_{j \in \mathbb{N}} \in \mathbb{N}^\mathbb{N}$ satisfying

$$c_{\mathcal{A}} := \sup_{j \in \mathbb{N}} \frac{n_{j+1}}{n_j} < \infty$$ \hspace{1cm} (4.2.4)

and for every $s \in (0, s^*)$,

$$d_{\mathcal{A}, s} := \sup_{j \in \mathbb{N}} n_j^s \left\| \mathcal{A} - \mathcal{A}_j \right\|_{L^2(\Xi) \otimes_\alpha X \to L^2(\Theta) \otimes_\beta Y} < \infty.\hspace{1cm} (4.2.5)$$

The operator $\mathcal{A}$ is strictly $s^*$-compressible if, in addition,

$$\sup_{s \in (0, s^*)} d_{\mathcal{A}, s} < \infty.$$ \hspace{1cm} (4.2.6)

**Remark 4.2.2.** Equation (4.2.5) states that for all $s \in (0, s^*)$, the approximation errors satisfy

$$e_{\mathcal{A}, j} := \left\| \mathcal{A} - \mathcal{A}_j \right\|_{L^2(\Xi) \otimes_\alpha X \to L^2(\Theta) \otimes_\beta Y} \leq d_{\mathcal{A}, s} n_j^{-s}, \quad j \in \mathbb{N}.\hspace{1cm} (4.2.7)$$

59
Chapter 4. Adaptive Application of the Discrete Operator

If $s' < \infty$, this is equivalent to the condition that $(n^s_j e_{\mathfrak{A},j})_{j \in \mathbb{N}}$ grows subalgebraically in $n_j$, i.e.

$$n^s_j e_{\mathfrak{A},j} \leq \inf_{r > 0} d_{\mathfrak{A},s'} - n^s_j, \quad j \in \mathbb{N}.$$  \hfill (4.2.8)

Strict $s'$-compressibility states that the right hand side of (4.2.8) is bounded in $j$, i.e.

$$d_{\mathfrak{A},s'} = \sup_{j \in \mathbb{N}} n^s_j e_{\mathfrak{A},j} = \sup_{j \in \mathbb{N}} n^s_j e_{\mathfrak{A},j} = \sup_{s \in (0,s')} d_{\mathfrak{A},s} < \infty.$$  \hfill (4.2.9)

Of course, $s'$-compressibility implies strict $s$-compressibility for all $s \in (0,s')$.

**Proposition 4.2.3.** Let $\mathfrak{A} \in \mathcal{L}(\ell^2(\Xi) \otimes_{\alpha} X, \ell^2(\Theta) \otimes_{\beta} Y)$ be $s'$-compressible with an approximating sequence $(\mathfrak{A}_j)_{j \in \mathbb{N}}$ as in Definition 4.2.1, and set $\mathfrak{A}_0 := 0$. There is a map $j : [0, \infty) \to \mathbb{N}_0$ such that $\mathfrak{A}_{j(r)}$ is $r$-sparse for all $r \in [0, \infty)$ and for all $s \in (0,s')$,

$$e_{\mathfrak{A},j(r)} = \|\mathfrak{A} - \mathfrak{A}_{j(r)}\|_{\ell^2(\Xi) \otimes_{\alpha} X \to \ell^2(\Theta) \otimes_{\beta} Y} \leq \max(c^s_{\mathfrak{A}} d_{\mathfrak{A},s}, n^s_1 e_{\mathfrak{A},0}) r^{-s}$$  \hfill (4.2.10)

for $r > 0$, where $e_{\mathfrak{A},0} = \|\mathfrak{A}\|_{\ell^2(\Xi) \otimes_{\alpha} X \to \ell^2(\Theta) \otimes_{\beta} Y}$.

**Proof.** Set $n_0 := 0$ and define

$$j(r) := \max\{j \in \mathbb{N}_0 ; n_j \leq r\}, \quad r \in [0, \infty).$$  \hfill (4.2.11)

Then $\mathfrak{A}_{j(r)}$ is $r$-sparse, and if $j(r) \geq 1$,

$$e_{\mathfrak{A},j(r)} \leq d_{\mathfrak{A},s} n^{-s}_{j(r)} \leq d_{\mathfrak{A},s} c^s_{\mathfrak{A}} n^{-s}_{j(r)+1} \leq d_{\mathfrak{A},s} c^s_{\mathfrak{A}} r^{-s}$$

by (4.2.7) and (4.2.4). If $j(r) = 0$, then $r < n_1$, and

$$e_{\mathfrak{A},j(r)} = e_{\mathfrak{A},0} \leq e_{\mathfrak{A},0} n^s_1 r^{-s}.$$  \hfill \Box

In particular, Proposition 4.2.3 implies that Definition 4.2.1 coincides with the notion of $s'$-compressibility for example in [GHS07, SS09], i.e. one can assume $n_j = j$ in the definition of $s'$-compressibility at the cost of increasing the constants (4.2.5) and obscuring the discrete structure of the sparse approximating sequence. We denote the resulting compressibility constants by

$$\bar{d}_{s} := \sup_{r \in (0,\infty)} r^s \|\mathfrak{A} - \mathfrak{A}_{j(r)}\|_{\ell^2(\Xi) \otimes_{\alpha} X \to \ell^2(\Theta) \otimes_{\beta} Y} \leq \max(c^s_{\mathfrak{A}} d_{\mathfrak{A},s}, n^s_1 e_{\mathfrak{A},0}) < \infty$$  \hfill (4.2.12)

for $s \in (0,s')$, where $j(r)$ is given by (4.2.11). Also, it follows using Proposition 4.2.3 that $s'$-compressible operators are in the class $\mathcal{B}_s$ defined in [CDD01] for all $s \in [0,s')$.

In the scalar case $X = Y = \mathbb{K}$, it is conceivable to construct the sparse approximations to an operator. In this setting, we use the notation $A$ for $\mathfrak{A}$ in $\mathcal{L}(\ell^2(\Xi), \ell^2(\Theta))$, and $a_{\nu \mu}$ for the coefficient operators $A_{\nu \mu}$ from (4.2.3), which are just scalars.
4.2. Adaptive Application of \( s^* \)-Computable Operators

**Definition 4.2.4.** An operator \( A \in \mathcal{L}(\ell^2(\Xi), \ell^2(\Theta)) \) is \( s^* \)-computable for an \( s^* \in (0, \infty) \) if it is \( s^* \)-compressible with an approximating sequence \((A_j)_{j \in \mathbb{N}}\) as in Definition 4.2.1 such that \( A_j \) is \( n_j \)-sparse and there exists a routine

\[
\text{Build}_{A}[j, \mu] \mapsto \left((v_i)_{i=1}^{n_j}, (a_i)_{i=1}^{n_j}\right)
\]

such that the \( \mu \)-th column of \( A_j \) is equal to

\[
\sum_{i=1}^{n_j} a_i e_{v_i} \quad (4.2.14)
\]

and there is a constant \( b_A \) such that the number of arithmetic operations and storage locations used by a call of \( \text{Build}_{A}[j, \mu] \) is less than \( b_A n_j \) for any \( j \in \mathbb{N} \) and \( \mu \in \Xi \).

Note that the indices \( v_i \) in (4.2.13) are not assumed to be distinct, so a single entry of \( A_j \) may be given by a sum of values \( a_i \). However, the total number of \( a_i \) computed by \( \text{Build}_{A}[j, \mu] \) is at most \( n_j \).

4.2.2. An Adaptive Application Routine

It was shown in [CDD01, CDD02] that \( s^* \)-computable operators can be applied efficiently to finitely supported vectors. A routine with computational advantages was presented in [DSS09]. We extend this method by using a greedy algorithm to solve the optimization problem at the heart of the routine.

Let \( A \in \mathcal{L}(\ell^2(\Xi), \ell^2(\Theta)) \) and for all \( k \in \mathbb{N}_0 \), let \( A_k \) be \( N_k \)-sparse with

\[
||A - A_k||_{\ell^2(\Xi) \to \ell^2(\Theta)} \leq \varepsilon_{A,k} \quad (4.2.15)
\]

We consider a partitioning of a vector \( v \in \ell^2(\Xi) \) into \( v|_{[p]} := v|_{\Xi_p}, p = 1, \ldots, P \), for disjoint index sets \( \Xi_p \subseteq \Xi \). This can be approximate in that \( v|_{[1]} + \cdots + v|_{[P]} \) only approximates \( v \) in \( \ell^2(\Xi) \). We think of \( v|_{[1]} \) as containing the largest elements of \( v \), \( v|_{[2]} \) the next largest, and so on.

Such a partitioning can be constructed by the approximate sorting algorithm

\[
\text{BucketSort}[v, \epsilon] \mapsto \left((v|_{[p]})_{p=1}^{P}, (\Xi_p)_{p=1}^{P}\right) \quad (4.2.16)
\]

which, given a finitely supported \( v \in \ell^2(\Xi) \) and a threshold \( \epsilon > 0 \), returns index sets

\[
\Xi_p := \mu \in \Xi \mid |v|_P \in (2^{-p/2} ||v|_{\ell^\infty}, 2^{-(p-1)/2} ||v|_{\ell^\infty})
\]

and \( v|_{[p]} := v|_{\Xi_p} \), see [Met02, Bar05, GHS07, DSS09]. The integer \( P \) is minimal with

\[
2^{-p/2} ||v|_{\ell^\infty} \sqrt{\# \text{supp } v} \leq \epsilon \quad (4.2.18)
\]

By [GHS07, Rem. 2.3] or [DSS09, Prop. 4.4], the number of operations and storage locations required by a call of \( \text{BucketSort}[v, \epsilon] \) is bounded by

\[
\# \text{supp } v + \max(1, |\log(||v|_{\ell^\infty} \sqrt{\# \text{supp } v/\epsilon}|)) \quad (4.2.19)
\]
Chapter 4. Adaptive Application of the Discrete Operator

This analysis uses that every $v_\mu, \mu \in \Xi$, can be mapped to $p$ with $\mu \in \Xi_p$ in constant time by evaluating

$$p := 1 + 2 \log_2 \left( \frac{||v||_{\ell^2(\Xi)}}{|v_\mu|} \right).$$

(4.2.20)

Alternatively, any standard comparison-based sorting algorithm can be used to construct the partitioning of $v$, albeit with an additional logarithmic factor in the complexity.

For any $k = (k_p)_{p=1}^\ell \in \mathbb{N}_0^\ell$, with $\ell \in \mathbb{N}_0$ determined as in $\text{Apply}_A[v, \epsilon]$, define

$$\zeta_k := \sum_{p=1}^\ell \bar{e}_{A_k_p} \|v[p]\|_{\ell^2(\Xi_p)}$$

and

$$\sigma_k := \sum_{p=1}^\ell N_{k_p}(\# \text{supp } v[p]).$$

(4.2.21)

$\text{Apply}_A[v, \epsilon] \mapsto z$

$$(v[p])_{p=1}^\ell \leftarrow \text{BucketSort} \left[ v, \frac{\epsilon}{2\bar{e}_{A,0}} \right]$$

compute the minimal $\ell \in \{0, 1, \ldots, P\}$ s.t.

$$\delta := \bar{e}_{A,0} \left\| v - \sum_{p=1}^\ell v[p] \right\|_{\ell^2(\Xi)} \leq \frac{\epsilon}{2}$$

$k = (k_p)_{p=1}^\ell \leftarrow (0)_{p=1}^\ell$

while $\zeta_k > \epsilon - \delta$ do

$k \leftarrow \text{NextOpt}[k]$ with objective $-\zeta_k$ and cost $\sigma_k$

end

$z \leftarrow \sum_{p=1}^\ell A_{k_p} v[p]$.

The algorithm $\text{Apply}_A[v, \epsilon]$ has three distinct parts. First, the elements of $v$ are divided into buckets according to their magnitude. Elements smaller than a certain tolerance are neglected. This truncation of the vector $v$ produces an error of at most $\delta \leq \epsilon/2$.

Next, a greedy algorithm is used to assign to each segment $v[p]$ of $v$ a sparse approximation $A_{k_p}$ of $A$. Starting with $A_{k_p} = 0$ for all $p = 1, \ldots, \ell$, these approximations are refined iteratively until an estimate for the error resulting from the approximation of $A$ by $A_{k_p}$ for all $p = 1, \ldots, \ell$ is bounded by $\zeta \leq \epsilon - \delta$.

Finally, the multiplications determined by the previous two steps are performed. A few basic properties of this method are summarized in the following proposition.

**Proposition 4.2.5.** For any finitely supported $v \in \ell^2(\Xi)$ and any $\epsilon > 0$, if $\text{Apply}_A[v, \epsilon]$ terminates, its output is a finitely supported $z \in \ell^2(\Theta)$ with

$$\# \text{supp } z \leq \sum_{p=1}^\ell N_{k_p}(\# \text{supp } v[p])$$

(4.2.22)
4.2. Adaptive Application of \( s^* \)-Computable Operators

and

\[
\|A\mathbf{v} - \mathbf{z}\|_{\ell^2(\Theta)} \leq \delta + \zeta_k \leq \epsilon , \tag{4.2.23}
\]

where \( k = (k_p)_{p=1}^\ell \) is the vector constructed by the greedy algorithm in \( \text{Apply}_A[v, \epsilon] \). Furthermore, the number of arithmetic operations required by the final step of \( \text{Apply}_A[v, \epsilon] \) is bounded by

\[
\sum_{p=1}^\ell N_{k_p}(\#\supp\mathbf{v}_p) \tag{4.2.24}
\]

if the relevant entries of \( A \) are precomputed.

**Proof.** We show (4.2.23). Since \( \|A\|_{\ell^2(\Xi) \rightarrow \ell^2(\Theta)} \leq \tilde{\epsilon}_{A,0} \),

\[
\left\| A\mathbf{v} - A \sum_{p=1}^\ell \mathbf{v}_p \right\|_{\ell^2(\Theta)} \leq \tilde{\epsilon}_{A,0} \left\| \mathbf{v} - \sum_{p=1}^\ell \mathbf{v}_p \right\|_{\ell^2(\Xi)} = \delta \leq \frac{\epsilon}{2} .
\]

By (4.2.15), if \( k = (k_p)_{p=1}^\ell \) is the final value of \( k \),

\[
\sum_{p=1}^\ell \left\| A\mathbf{v}_p - A_{k_p} \mathbf{v}_p \right\|_{\ell^2(\Theta)} \leq \sum_{p=1}^\ell \tilde{\epsilon}_{A,k_p} \left\| \mathbf{v}_p \right\|_{\ell^2(\Xi_p)} = \zeta_k \leq \epsilon - \delta .
\]

Let \( \mathbf{v} \in \ell^2(\Xi) \) be finitely supported and \( \epsilon > 0 \). Note that by (4.2.17) and (4.2.18),

\[
\left\| \mathbf{v} - \sum_{p=1}^P \mathbf{v}_p \right\|_{\ell^2(\Xi)} \leq 2^{-P/2} \left\| \mathbf{v} \right\|_{\ell^2(\Xi)} \sqrt{\#\supp\mathbf{v}} \leq \frac{\epsilon}{2\tilde{\epsilon}_{A,0}} ,
\]

so \( \ell \) is well defined. It is not immediately clear, however, that the greedy algorithm in \( \text{Apply}_A[v, \epsilon] \) terminates. This requires some additional assumptions. For all \( k \in \mathbb{N}_0 \), define

\[
\eta_k := \frac{\tilde{\epsilon}_{A,k} - \tilde{\epsilon}_{A,k+1}}{N_{k+1} - N_k} . \tag{4.2.25}
\]

**Assumption 4.2.A.** \( (\tilde{\epsilon}_{A,k})_{k \in \mathbb{N}_0} \) is nonincreasing and converges to 0; \( N_0 = 0 \) and \( (N_k)_{k \in \mathbb{N}_0} \) is strictly increasing. Furthermore, the sequence \( (\eta_k)_{k \in \mathbb{N}_0} \) is nonincreasing.

Note that Assumption 4.2.A implies Assumption 4.1.A. Let \( \mathcal{M} \) denote the set of \( p \in \{0, \ldots, P\} \) for which \( \supp\mathbf{v}_p \neq \emptyset \). For all \( p \in \mathcal{M} \), the sequences of costs and values from Section 4.1 are given by

\[
c_k^p := N_k(\#\supp\mathbf{v}_p) \quad \text{and} \quad \omega_k^p := -\tilde{\epsilon}_{A,k} \left\| \mathbf{v}_p \right\|_{\ell^2(\Xi)} . \tag{4.2.26}
\]
Chapter 4. Adaptive Application of the Discrete Operator

By Assumption 4.2.A, \( c''_0 = 0 \), \((c''_k)_{k \in \mathbb{N}_0}\) is strictly increasing and \((\sigma''_k)_{k \in \mathbb{N}_0}\) is nondecreasing for all \( p \in \mathcal{M} \). Also,

\[
\bar{q}^p_k = \frac{\Delta\sigma''_k}{\Delta^p} = \eta_k \frac{\|v[p]\|_{L^p(\Xi)}}{\# \text{supp } v[p]}
\]  

(4.2.27)

is nonincreasing in \( k \) for all \( p \in \mathcal{M} \).

**Proposition 4.2.6.** For any \( k \) generated in \( \text{Apply}_A[v, \epsilon] \), if \( j \in \mathbb{N}_0^\epsilon \) with \( \sigma_j \leq \sigma_k \), then \( \bar{\zeta}_j \geq \bar{\zeta}_k \). If \( j \in \mathbb{N}_0^\epsilon \) with \( \sigma_j \leq \sigma_k \), then \( \sigma_j \geq \sigma_k \).

**Proof.** The assertion follows from Theorem 4.1.5 with (4.2.26) by Assumption 4.2.A. Note that \( \sigma_j \geq 0 \) for all \( j \in \mathbb{N}_0^\epsilon \), and if \( \sigma_k > 0 \), the second statement in Theorem 4.1.5 applies.

Let \((k_i)_{i \in \mathbb{N}_0}\) denote the sequence of \( k \) generated in \( \text{Apply}_A[v, \epsilon] \) if the loop is not terminated. We abbreviate \( \bar{\zeta}_i := \bar{\zeta}_{k_i} \) and \( \sigma_i := \sigma_{k_i} \).

**Remark 4.2.7.** In particular, Proposition 4.2.6 implies convergence of the greedy subroutine in \( \text{Apply}_A[v, \epsilon] \). Since \( N_{k+1} \geq N_k + 1 \) for all \( k \in \mathbb{N}_0 \) and \( k_{i,p} = 0 \) for all \( i \in \mathbb{N}_0 \) if \( \# \text{supp } v[p] = 0 \), \( \sigma_i \) goes to infinity as \( i \to \infty \). Since \( \bar{\zeta}_j \) can be made arbitrarily small for suitable \( j \in \mathbb{N}_0^\epsilon \), it follows that \( \bar{\zeta}_i \to 0 \).

### 4.2.3. Convergence Analysis

For \( v \in \ell^2(\Xi) \) and \( N \in \mathbb{N}_0 \), let \( P_N(v) \) be a best \( N \)-term approximation of \( v \), that is, \( P_N(v) \) is an element of \( \ell^2(\Xi) \) that minimizes \( \|v - P_N(v)\|_{L^2(\Xi)} \) over \( v_N \in \ell^2(\Xi) \) with \( \# \text{supp } v_N \leq N \). Following [DSS09], for \( s \in (0, \infty) \), we define

\[
\|v\|_{\mathcal{A}^s(\Xi)} := \sup_{\epsilon > 0} \epsilon \left( \min \left\{ N \in \mathbb{N}_0 : \|v - P_N(v)\|_{L^2(\Xi)} \leq \epsilon \right\} \right)^s
\]  

(4.2.28)

and

\[
\mathcal{A}^s(\Xi) := \left\{ v \in \ell^2(\Xi) : \|v\|_{\mathcal{A}^s(\Xi)} < \infty \right\}.
\]  

(4.2.29)

Note that \( \|v\|_{\mathcal{A}^s(\Xi)} \) defines a quasinorm on \( \mathcal{A}^s(\Xi) \) that is equivalent to the usual quasinorm on this space. Setting \( \epsilon = \|v - P_N(v)\|_{L^2(\Xi)} - \eta \) with \( \eta \geq 0 \), it follows that

\[
\sup_{N \in \mathbb{N}_0} N^s \|v - P_N(v)\|_{L^2(\Xi)} \leq \|v\|_{\mathcal{A}^s(\Xi)} = \sup_{N \in \mathbb{N}_0} (N + 1)^s \|v - P_N(v)\|_{L^2(\Xi)}.
\]  

(4.2.30)

**Assumption 4.2.B.** \( \bar{p}_A := \sup_{k \in \mathbb{N}_0} \frac{\bar{\sigma}_{A,k}}{\bar{\sigma}_{A,k+1}} < \infty. \)

Assumption 4.2.B states that the values \( \bar{\sigma}_{A,k} \) are spaced sufficiently regularly, with at most geometric convergence to 0. In particular, \( \bar{\sigma}_{A,k} > 0 \) for all \( k \in \mathbb{N}_0 \), i.e., if \( A \) is sparse, this is not reflected in the bounds \( \bar{\sigma}_{A,k} \). An admissible value is \( \bar{\sigma}_{A,k} = d^s \eta \) since for all \( k \in \mathbb{N}_0 \),

\[
\frac{\bar{\sigma}_{A,k}}{\bar{\sigma}_{A,k+1}} \leq \left( \frac{N_{k+1}}{N_k} \right)^s \leq c_A^s < \infty.
\]
4.2. Adaptive Application of $s^*$-Computable Operators

**Lemma 4.2.8.** For all $i \in \mathbb{N}_0$, $\zeta_i \leq \bar{r}_A \zeta_{i+1}$.

**Proof.** Let $i \in \mathbb{N}_0$. Note that $\zeta_i - \zeta_{i+1} = (\bar{e}_{A,k_i} - \bar{e}_{A,k_{i+1}}) \left\| v_{[q_i]} \right\|_{\ell^2}$ and $\zeta_{i+1} \geq \bar{e}_{A,k_{i+1}} \left\| v_{[q_i]} \right\|_{\ell^2}$. Therefore,

$$\frac{\zeta_i}{\zeta_{i+1}} = 1 + \frac{\zeta_i - \zeta_{i+1}}{\zeta_{i+1}} \leq 1 + \frac{\bar{e}_{A,k_i} - \bar{e}_{A,k_{i+1}}}{\bar{e}_{A,k_{i+1}}} = \frac{\bar{e}_{A,k_i}}{\bar{e}_{A,k_{i+1}}} \leq \bar{r}_A.$$  

□

The following is adapted from [DSS09, Thm. 4.6]. We emphasize in advance that knowledge of $s$ and $s^*$ is not required in $\text{Apply}_A[v, c]$. The algorithm satisfies Theorem 4.2.9 with any $s^*$ for which $A$ is $s^*$-compressible, provided that the bounds $\bar{r}_{A,k}$ from (4.2.15) decay at the rate implied by $s^*$-compressibility.

**Theorem 4.2.9.** Let $v \in \ell^2(\Xi)$ be finitely supported and $e > 0$. A call of $\text{Apply}_A[v, c]$ produces a finitely supported $z \in \ell^2(\Theta)$ with

$$\left\| A v - z \right\|_{\ell^2(\Theta)} \leq \delta + \zeta_k \leq e.$$  

(4.2.31)

If $A$ is $s^*$-compressible for an $s^* \in (0, \infty]$ and $\sup_{k \in \mathbb{N}} \bar{e}_{A,k} N_k^{s^*} < \infty$ for all $s \in (0, s^*)$, then for any $s \in (0, s^*)$,

$$\# \text{supp } z \leq \sigma_k \leq e^{-1/s} \left\| v \right\|_{\ell^2(\Xi)}^{1/s}$$  

(4.2.32)

with a constant depending only on $s$, $\bar{e}_{A,0}, c_A$, $N_1$, $(d_{A,s})_{s \in (0,s^*)}$ and $\bar{r}_A$.

**Proof.** Convergence of $\text{Apply}_A[v, c]$ follows from Proposition 4.2.6, see Remark 4.2.7. Then (4.2.31) is shown in Proposition 4.2.5.

Let $k = (k_p)_{p=1}^\ell$ be the final value of $k$ in $\text{Apply}_A[v, c]$, and $s \in (0, s^*)$. By Proposition 4.2.5, to prove (4.2.32) it suffices to show that there is a $j \in \mathbb{N}_0$ with $\zeta_j \leq \zeta_k \leq \zeta$ and $\sigma_j \leq e^{-1/s} \left\| v \right\|_{\ell^2(\Xi)}^{1/s}$, Then Proposition 4.2.6 implies

$$\# \text{supp } z \leq \sigma_k \leq e^{-1/s} \left\| v \right\|_{\ell^2(\Xi)}^{1/s}.$$  

The construction of such a $j$ is analogous to the proof of [DSS09, Thm. 4.6] with $\zeta$ in place of $e - \delta$. We provide it here for completeness.

Let $\tau \in (0, 2)$ be defined by $\tau^{-1} = s + \frac{1}{2}$, and let $s < \tau < s^* < s^*$. Then

$$\# \text{supp } v_{[p]} \leq \# \left\{ \mu \in \Xi \mid v_{[\mu]} > 2^{-p/2} \left\| v \right\|_{e^0} \right\} \leq 2^{p/2} \left\| v \right\|_{e^0} \left\| v \right\|_{e^{s^*}}^\tau,$$  

see e.g. [DeV98]. In particular,

$$\left\| v_{[p]} \right\|_{\ell^2} \leq 2^{-p/2} \left\| v \right\|_{e^0} \sqrt{\# \text{supp } v_{[p]} \leq 2^{-p/2} \left\| v \right\|_{e^0}^{1/2} \left\| v \right\|_{e^{s^*}}^\tau}.$$  

Let $j \geq \ell$ be the smallest integer with $\sum_{p=1}^\ell 2^{-(l-p)\tau/2} \left\| v_{[p]} \right\|_{\ell^2} \leq \zeta$ and let $j = (j_p)_{p=1}^\ell \in \mathbb{N}_0^\ell$ with $j_p$ minimal such that $\bar{e}_{A,j_p} \leq 2^{-(l-p)\tau/2}$. Then

$$\zeta_j = \sum_{p=1}^\ell \bar{e}_{A,j_p} \left\| v_{[p]} \right\|_{\ell^2} \leq \sum_{p=1}^\ell 2^{-(l-p)\tau/2} \left\| v_{[p]} \right\|_{\ell^2} \leq \zeta.$$  

65
Chapter 4. Adaptive Application of the Discrete Operator

It remains to be shown that \( \sigma_j \leq \epsilon^{-1/s} \|v\|_{\sigma^s}^{1/s} \).

If \( Jp \geq 2 \), since \( \tilde{e}_{A,lp-1}N_{lp-1}^{s_2} \leq 1 \),

\[
N_{lp} \leq N_{lp-1} \leq \tilde{e}_{A,lp-1}^{-1/s_2} \leq 2^{(j-p)(s_1/s_2)\tau/2}.
\]

This estimate extends to \( Jp \in [0, 1] \) since \( p \leq J \). Therefore, using \( s_1 < s_2 \),

\[
\sigma_j = \sum_{p=1}^{\ell} N_{lp}(\# \text{supp } v_p) \leq \sum_{p=1}^{\ell} 2^{(j-p)(s_1/s_2)\tau/2} 2^{-pst/2} \|v\|_{\ell^\infty}^{-1} \|v\|_{\sigma^s}^{s_2} \\
\leq 2^{2j(\ell)(s_1/s_2)\tau/2 - pst/2} \|v\|_{\ell^\infty}^{-1} \|v\|_{\sigma^s}^{s_2} \leq 2^{\ell\tau/2} \|v\|_{\ell^\infty}^{-1} \|v\|_{\sigma^s}^{s_2}.
\]

Thus, the assertion reduces to \( 2^{\ell\tau/2} \|v\|_{\ell^\infty}^{-1} \|v\|_{\sigma^s}^{s_2} \leq \epsilon^{-1/s} \|v\|_{\sigma^s}^{1/s} \).

If \( J = \ell \), by minimality of \( \ell \),

\[
\varepsilon < \tilde{e}_{A,0} \left\| v - \sum_{p=1}^{\ell-1} v_p \right\|_2 = \tilde{e}_{A,0} \sqrt{\sum_{p=\ell}^{\infty} \|v_p\|_2^2} \leq \tilde{e}_{A,0} 2^{-tst/2} \|v\|_{\ell^\infty}^{1-t/2} \|v\|_{\sigma^s}^{t/2}.
\]

If \( J > \ell \), then by minimality of \( \ell \), using \( s < s_1 \),

\[
\zeta < \sum_{p=1}^{\ell} 2^{-t(1-p)s_1/2} \|v_p\|_2 \leq \sum_{p=1}^{\ell} 2^{-t(1-p)s_1/2} 2^{-ps/2} \|v_p\|_{\ell^\infty}^{-1} \|v_p\|_2 \\
\leq 2^{-t(1-p)s_1/2 - pst/2} \|v_p\|_{\ell^\infty}^{-1} \|v_p\|_2 \leq 2^{-t(1-s_2)/2} \|v\|_{\sigma^s}^{1-s_2/2} \|v\|_{\sigma^s}^{t/2}.
\]

Lemma 4.2.8 implies \( \epsilon \leq \tilde{r}_A \zeta \). Therefore, in both cases,

\[
\epsilon \leq 2^{-tst/2} \|v\|_{\ell^\infty}^{-1} \|v\|_{\sigma^s}^{t/2},
\]

or equivalently,

\[
2^{\ell\tau/2} \|v\|_{\ell^\infty}^{-1} \|v\|_{\sigma^s}^{s_2} \leq \epsilon^{-1/s} \|v\|_{\sigma^s}^{1/s},
\]

which completes the proof. \( \square \)

We note that the constant in (4.2.32) may degenerate as \( s \to s^* \).

It is known that \( s^* \)-compressible operators \( A \) map \( \sigma^s(\Xi) \) boundedly into \( \sigma^s(\Theta) \), see [CDD01, Proposition 3.8]. Theorem 4.2.9 implies that this carries over to the approximate multiplication routine \( \text{App}1y_{A} \).

**Corollary 4.2.10.** Let \( A \) be \( s^* \)-compressible for some \( s^* \in (0, \infty) \), and assume that for all \( s \in (0, s^*) \), \( \sup_{k \in \mathbb{N}} e_{A,k}N_k^s < \infty \). Then for any \( s \in (0, s^*) \) there is a constant \( C \) depending only on \( s, e_{A,0}, c_{A}, N_1, (d_{A,s})_{s \in (0, s^*)} \) and \( t_A \) such that for all \( v \in \sigma^s(\Xi) \) and all \( \epsilon > 0 \), the output \( z \) of \( \text{App}1y_{A}[v, \epsilon] \) satisfies

\[
\|z\|_{\sigma^{s'}(\Xi)} \leq C \|v\|_{\sigma^s(\Xi)}.
\]  
(4.2.33)

66
4.2. Adaptive Application of $s^*$-Computable Operators

Proof. Let $z$ be the output of $\text{AppLy}_A[v, \epsilon]$ for some $v \in \mathcal{A}^s(\Xi)$ and some $\epsilon > 0$, and define $w := Av$. By [CDD01, Proposition 3.8], $w \in \mathcal{A}^s(\Theta)$, and $\|w\|_{\mathcal{A}^s(\Theta)} \leq \|v\|_{\mathcal{A}^s(\Xi)}$. Therefore, it suffices to show $\|z\|_{\mathcal{A}^s(\Theta)} \leq \|w\|_{\mathcal{A}^s(\Theta)}$. Since $z$ is finitely supported, $z \in \mathcal{A}^s(\Theta)$. Let $N := \#\text{supp } z$. Theorem 4.2.9 implies

$$\|w - z\|_{\mathcal{A}^s(\Theta)} \leq \|w\|_{\mathcal{A}^s(\Theta)} N^{-s}.$$ 

For any $n \geq N$, $P_n(z) = z$, and thus $(n + 1)^s \|z - P_n(z)\|_{\mathcal{A}^s(\Theta)} = 0$. Let $n \leq N - 1$ and $z_n \in \ell^2(\Theta)$ with $\#\text{supp } z_n \leq n$. Then

$$(n + 1)^s \|z - z_n\|_{\mathcal{A}^s(\Theta)} \leq (n + 1)^s \|w - z\|_{\mathcal{A}^s(\Theta)} + (n + 1)^s \|w - z_n\|_{\mathcal{A}^s(\Theta)}.$$ 

The first term is bounded by

$$(n + 1)^s \|w - z\|_{\mathcal{A}^s(\Theta)} \leq (n + 1)^s N^{-s} \|w\|_{\mathcal{A}^s(\Theta)} \leq \|w\|_{\mathcal{A}^s(\Theta)}.$$ 

Taking the infimum over $z_n$ with $\#\text{supp } z_n \leq n$, we have

$$(n + 1)^s \|z - P_n(z)\|_{\mathcal{A}^s(\Theta)} \leq \|w\|_{\mathcal{A}^s(\Theta)} + (n + 1)^s \inf_{z_n} \|w - z_n\|_{\mathcal{A}^s(\Theta)} \leq \|w\|_{\mathcal{A}^s(\Theta)}.$$ 

The assertion follows by taking the supremum over $n \in \mathbb{N}_0$. \hfill \Box

By (4.2.19), the number of operations and storage locations required by $\text{BucketSort}$ in a call of $\text{AppLy}_A[v, \epsilon]$ is bounded by

$$\#\text{supp } v + \max(1, \lceil \log(2\bar{\epsilon}_{A, \Theta} \|v\|_{\ell^\infty}) \sqrt{\#\text{supp } v/\epsilon} \rceil) \leq 1 + \#\text{supp } v + \log(e^{-1} \|v\|_{\ell^\infty}). \quad (4.2.34)$$

The value of $\ell$ can be determined with at most $\#\text{supp } v$ operations. We assume that the values of $\|v_p\|_{\mathcal{A}^s(\Xi)}$ are known from the computation of $\ell$. Then by Proposition 4.1.6, initialization of the greedy subroutine requires $O(\ell \log \ell)$ operations, and each iteration requires $O(1 + \log \ell)$ operations if a tree data structure is used for $\mathcal{N}$ from Section 4.1.3. As $\|k\|_{\ell^1}$ iterations are performed if $k = (k_p)_{p=1}^\ell$ is the final value of $k$ in $\text{AppLy}_A[v, \epsilon]$, the total cost of determining $\ell$ and $k$ is on the order of

$$\#\text{supp } v + \ell \log^+ \ell + (1 + \log^+ \ell) \sum_{p=1}^\ell k_p, \quad (4.2.35)$$

where $\log^+ x := \log(\max(x, 1))$. Since $\ell \leq P$, (4.2.18) implies

$$\ell \leq 1 + \log^+(\#\text{supp } v) + \log^+(e^{-1} \|v\|_{\ell^\infty}). \quad (4.2.36)$$

Finally, the number of arithmetic operations required by the last step of $\text{AppLy}_A[v, \epsilon]$ is bounded by

$$\sigma_k = \sum_{p=1}^\ell N_{k_p}(\#\text{supp } v_p), \quad (4.2.37)$$

and this value is optimal in the sense of Proposition 4.2.6. If $A$ is $s^*$-computable for any $s^* \in (0, \infty)$, then (4.2.37) includes the assembly costs of $A_{k_p}$.
Chapter 4. Adaptive Application of the Discrete Operator

**Theorem 4.2.11.** Let \( v \in \ell^2(\mathbb{Z}) \) be finitely supported and \( \epsilon > 0 \). If \( A \) is \( s^* \)-computable for an \( s^* \in (0, \infty) \) and \( \sup_{k \in \mathbb{N}} \tilde{e}_A N_k^{s^*} < \infty \) for all \( s \in (0, s^*) \), then for any \( s \in (0, s^*) \), the number of operations and storage locations required by \( \text{Apply}_A[v, \epsilon] \) is less than a multiple of

\[
1 + \#\text{supp } v + e^{-1/s} \|v\|_{1/s}^{1/s} \left( 1 + \log^+ \left( \#\text{supp } v + e^{-1} \|v\|_{1/s} \right) \right) \tag{4.2.38}
\]

with a constant depending only on \( s, \tilde{e}_A, c_A, N_1, (d_{A,s})_{s \in (0, s^*)}, r_A \) and \( b_A \). The double logarithmic term in (4.2.38) is due only to the greedy subroutine and does not apply to the storage requirements.\(^1\)

**Proof.** We first note that

\[
\log(e^{-1} \|v\|_{1/s}) \leq e^{-1/s} \|v\|_{1/s}^{1/s} \leq e^{-1/s} \|v\|_{1/s}^{1/s}.
\]

Therefore and by (4.2.34), the cost of BucketSort is less than

\[
1 + \#\text{supp } v + \log(e^{-1} \|v\|_{1/s}) \leq 1 + \#\text{supp } v + e^{-1/s} \|v\|_{1/s}^{1/s}.
\]

The cost of the last step of \( \text{Apply}_A[v, \epsilon] \) is \( \alpha_k \), which in Theorem 4.2.9 is bounded by

\[
\alpha_k \leq e^{-1/s} \|v\|_{1/s}^{1/s}.
\]

The cost of the rest of \( \text{Apply}_A[v, \epsilon] \) is given in (4.2.35). By (4.2.36), for \( \chi > 1 \),

\[
\ell \log \ell \leq \ell^\chi \leq 1 + \log(\#\text{supp } v)^\chi + \log(e^{-1} \|v\|_{1/s})^\chi
\]

\[
\leq 1 + \#\text{supp } v + e^{-1/s} \|v\|_{1/s}^{1/s} \leq 1 + \#\text{supp } v + e^{-1/s} \|v\|_{1/s}^{1/s}.
\]

Since

\[
\ell \leq 1 + \log(\#\text{supp } v) + \log(e^{-1} \|v\|_{1/s}) \leq 1 + \log(\#\text{supp } v + e^{-1} \|v\|_{1/s}),
\]

we have

\[
\log \ell \leq C + \log(1 + \log(\#\text{supp } v + e^{-1} \|v\|_{1/s})) \leq 1 + \log \log(\#\text{supp } v + e^{-1} \|v\|_{1/s}).
\]

Finally, since \( k \leq N_k \) for all \( k \in \mathbb{N}_0 \) and \( k_p = 0 \) if \( \#\text{supp } v_{[p]} = 0 \),

\[
\sum_{p=1}^\ell k_p \leq \sum_{p=1}^\ell N_{k_p}(\#\text{supp } v_{[p]}) = \alpha_k \leq e^{-1/s} \|v\|_{1/s}^{1/s}.
\]

\( \Box \)

**Remark 4.2.12.** The double logarithmic term in (4.2.38) can be dropped under mild conditions. If \( N_k \geq k^\alpha \) for an \( \alpha > 1 \), then by Hölder’s inequality,

\[
\sum_{p=1}^\ell k_p \leq \sum_{p=1}^\ell N_{k_p}^{1/\alpha} \leq \left( \sum_{p=1}^\ell N_{k_p} \right)^{1/\alpha} \ell^{\alpha/\alpha - 1}.
\]

\(^1\)As above, \( \log^+ x \coloneqq \log(\max(x, 1)) \).

68
4.3. Adaptive Application of Parametric Operators

Furthermore, for a $\chi > 1$, as in the proof of Theorem 4.2.11,
\[
\ell \log \ell \leq (\ell^\alpha) \leq \left(1 + \# \text{supp } v + \epsilon^{-1/\alpha} \|v\|^{1/s}_{1/s}\right)^{\frac{\alpha}{\alpha}}.
\]
It follows that
\[
\log \ell \sum_{p=1}^{\ell} k_p \leq \alpha^{1/\alpha} \left(1 + \# \text{supp } v + \epsilon^{-1/\alpha} \|v\|^{1/s}_{1/s}\right)^{\frac{\alpha}{\alpha}} \leq 1 + \# \text{supp } v + \epsilon^{-1/s} \|v\|^{1/s}_{1/s},
\]
and (4.2.38) can be replaced by
\[
1 + \# \text{supp } v + \epsilon^{-1/s} \|v\|^{1/s}_{1/s}
\]
in Theorem 4.2.11, with a constant that also depends on $\alpha$. The assumption $N_k \geq k^\alpha$ is generally not restrictive, since by (4.2.4), $N_k$ may grow exponentially for a $s^*$-compressible operator.

4.3. Adaptive Application of Parametric Operators

4.3.1. Sparse Semidiscrete Approximations

We first cite a result due to Stechkin connecting the order of summability of a sequence to the convergence of best $N$-term approximations in a weaker sequence norm, see e.g. [CDS10b, DeV98]. Note that, although it is formulated only for nonnegative sequences, Lemma 4.3.1 applies directly to e.g. Lebesgue–Bochner spaces of Banach space valued sequences by passing to the norms of the elements of such sequences.

Lemma 4.3.1. Let $0 < p \leq q$ and let $(c_v)_{v \in \Xi} \in \ell^2(\Xi)$ with $c_v \geq 0$ for all $v \in \Xi$. For all $N \in \mathbb{N}_0$, let $\Xi_N$ be the set of the first $N$ indices in a decreasing rearrangement of $(c_v)_{v \in \Xi}$. Then
\[
\left(\sum_{v \in \Xi_N \subseteq \Xi} c_v^q\right)^{1/q} \leq (N + 1)^{-r} \|c_v\|_{\ell^p(\Xi)}^q, \quad r := \frac{1}{p} - \frac{1}{q} \geq 0
\]
for all $N \in \mathbb{N}_0$.

Proof. Selecting a decreasing rearrangement of $(c_v)_{v \in \Xi}$, we assume without loss of generality that $\Xi = \mathbb{N}$ and $c := (c_n)_{n \in \mathbb{N}}$ is nonincreasing. Due to the elementary estimate
\[
\|c\|_{\ell^p}^p = \sum_{i=1}^{\infty} c_i^p \geq \sum_{i=1}^{n} c_i^p \geq \sum_{i=1}^{n} c_n^p = nc_n^p,
\]
we have $c_n \leq n^{-1/p} \|c\|_{\ell^p}$ for all $n \in \mathbb{N}$. Therefore, using $q - p \geq 0$,
\[
\sum_{n=N+1}^{\infty} c_n^q \leq \sum_{n=N+1}^{\infty} c_n^{q-\rho-p} \leq \|c\|_{\ell^p}^p (N + 1)^{-(q-p)/p} \|c\|_{\ell^p}^{q-p} = (N + 1)^{-\rho} \|c\|_{\ell^p}^q
\]
for all $N \in \mathbb{N}_0$, with $r$ as in (4.3.1).
Chapter 4. Adaptive Application of the Discrete Operator

We turn to the setting of Sections 1.2.2 and 1.3.2. For Banach spaces $V$ and $W$, and a tensor norm $\alpha$, the operator $A \in \mathcal{L}(L^2(\Gamma) \otimes V, L^2(\Gamma) \otimes W^*)$ has the form

$$A = \text{id}_{L^2(\Gamma)} \otimes D + \sum_{m \in \mathcal{M}} K_m \otimes R_m. \quad (4.3.2)$$

We assume that condition (1.3.16) is satisfied, which ensures invertibility of $\mathcal{A}$, see Theorem 1.3.5.

Let $P = (P_{\nu})_{\nu \in \Lambda}$ be a tensor product orthonormal basis of $L^2(\Gamma)$ as in (3.2.4). By Proposition 3.2.7, discretization of $\mathcal{A}$ by $P$ leads to

$$A = T^W_P \mathcal{A} T^V_P = I \otimes D + \sum_{m \in \mathcal{M}} K_m \otimes R_m \quad (4.3.3)$$

for $K_m := T^p K_m T^p$ and $I := \text{id}_{\ell^2(\Lambda)}$.

Selecting a decreasing rearrangement, we assume $\mathcal{M} = \mathbb{N}$ and the sequence of norms $(\|R_m\|_{V \rightarrow W^*})_{m \in \mathbb{N}}$ is nonincreasing. For all $M \in \mathbb{N}_0$, define the operator

$$\mathcal{A}_M := I \otimes D + \sum_{m=1}^{M} K_m \otimes R_m \in \mathcal{L}(\ell^2(\Lambda) \otimes V, \ell^2(\Lambda) \otimes W^*). \quad (4.3.4)$$

**Theorem 4.3.2.** Let $s > 0$. If either

$$\|R_m\|_{V \rightarrow W^*} \leq s \delta_{\mathcal{A}} (m + 1)^{-s-1} \quad \forall m \in \mathbb{N} \quad (4.3.5)$$

or

$$\left( \sum_{m=1}^{\infty} \|R_m\|_{V \rightarrow W^*}^{s+1} \right)^{1/(s+1)} \leq \delta_{\mathcal{A}}, \quad (4.3.6)$$

then

$$\|\mathcal{A} - \mathcal{A}_M\|_{\mathcal{L}(\ell^2(\Lambda) \otimes V, \ell^2(\Lambda) \otimes W^*)} \leq \delta_{\mathcal{A}} (M + 1)^{-s} \quad \forall M \in \mathbb{N}_0. \quad (4.3.7)$$

**Proof.** By Proposition 3.2.7 and Lemma 3.2.8, using (4.3.4),

$$\|\mathcal{A} - \mathcal{A}_M\|_{\mathcal{L}(\ell^2(\Lambda) \otimes V, \ell^2(\Lambda) \otimes W^*)} \leq \sum_{m=M+1}^{\infty} \|R_m\|_{V \rightarrow W^*}. \quad (4.3.8)$$

If (4.3.5) holds, then (4.3.7) follows using

$$\sum_{m=M+1}^{\infty} (m + 1)^{-s-1} \leq \int_{M+1}^{\infty} t^{-s-1} \, dt = \frac{1}{s} (M + 1)^{-s}. \quad (4.3.9)$$

If (4.3.6) is satisfied, then

$$\sum_{m=M+1}^{\infty} \|R_m\|_{V \rightarrow W^*} \leq \left( \sum_{m=1}^{\infty} \|R_m\|_{V \rightarrow W^*}^{s+1} \right)^{1/(s+1)} (M + 1)^{-s}$$

by Lemma 4.3.1. \qed

70
4.3. Adaptive Application of Parametric Operators

Corollary 4.3.3. If the assumptions of Theorem 4.3.2 hold for all \( s \in (0, s^*) \), then \( \mathcal{A} \) is \( s^* \)-compressible with \( c_{\mathcal{A}} = 4 \) and \( d_{\mathcal{A},s} \leq 3^s \delta_{\mathcal{A},s} \).

Proof. By Lemma 3.2.9 and Remark 3.2.10, \( K_m \) is 3-sparse in general and 2-sparse if \( \pi_m \) is symmetric. We set \( n := 2 \) if \( \pi_m \) is symmetric for all \( m \in \mathbb{N} \), and \( n := 3 \) otherwise. Since \( \mathcal{I} \) is 1-sparse, \( \mathcal{A}[M] \) is \((nM + 1)\)-sparse and (4.2.4) reads

\[
\begin{align*}
\mathcal{A}[M] := \mathcal{I} \otimes D + \sum_{m=1}^{M} K_m \otimes R_m \in \mathcal{L}(\ell^2(\Lambda \times \Xi), \ell^2(\Lambda \times Y)).
\end{align*}
\]

Corollary 4.3.4. Under the conditions of Theorem 4.3.2,

\[
\|A - A[M]\|_{\mathcal{L}(\ell^2(\Lambda \times \Xi) \to \ell^2(\Lambda \times Y))} \leq B_\Psi B_\Phi \delta_{\mathcal{A},s}(M + 1)^{-s}
\] (4.3.9)

for all \( M \in \mathbb{N}_0 \).

Proof. The assertion follows from Theorem 4.3.2 since \( A[M] = (T^{\ell^2(\Lambda)})^* \Psi[M] T^{\ell^2(\Lambda)} \).

Remark 4.3.5. Let

\[
\begin{align*}
\|D\|_{\mathcal{L}(\ell^2(\Xi) \to \ell^2(\Upsilon))} & \leq \bar{\epsilon}_{0,0} \quad \text{and} \quad \|R_m\|_{\mathcal{L}(\ell^2(\Xi) \to \ell^2(\Upsilon))} \leq \bar{\epsilon}_{m,0}, \quad m \in \mathbb{N}.
\end{align*}
\] (4.3.10)

Then by (3.2.27) and (4.3.8),

\[
\|A - A[M]\|_{\mathcal{L}(\ell^2(\Lambda \times \Xi) \to \ell^2(\Lambda \times Y))} \leq \sum_{m=M+1}^{\infty} \bar{\epsilon}_{m,0}.
\] (4.3.11)

For any \( s > 0 \), if either

\[
\bar{\epsilon}_{m,0} \leq \frac{1}{m+1} \delta_{\mathcal{A},s}(m + 1)^{-s} \quad \forall m \in \mathbb{N}
\] (4.3.12)

or

\[
\left( \sum_{m=1}^{\infty} \frac{1}{m+1} \right)^{s+1} \leq \delta_{\mathcal{A},s},
\] (4.3.13)

then it follows as in Theorem 4.3.2 that

\[
\sum_{m=M+1}^{\infty} \bar{\epsilon}_{m,0} \leq \delta_{\mathcal{A},s}(M + 1)^{-s} \quad \forall M \in \mathbb{N}_0.
\] (4.3.14)

\( \square \)
Chapter 4. Adaptive Application of the Discrete Operator

4.3.2. Sparse Approximation of the Discrete Operator

We consider the setting of Section 3.2.3. Let \((D_i)_{i∈\mathbb{N}_0}\) and \((R_{m,j})_{j∈\mathbb{N}_0}\) be approximating sequences of \(D\) and \(R_m\), respectively, such that \(D_i\) is \(n_{0,i}\)-sparse and \(R_{m,j}\) is \(n_{m,j}\)-sparse, \(m ∈ \mathbb{N}\). We assume \(n_{m,0} = 0\) and \(n_{m,j}\) is strictly increasing in \(j\) for all \(m ∈ \mathbb{N}_0\). Furthermore, let

\[
\|D - D_i\|_{ℓ^2(\mathbb{Z})→ℓ^2(\Omega)} ≤ \varepsilon_{0,i} \quad \text{and} \quad \|R_m - R_{m,j}\|_{ℓ^2(\mathbb{Z})→ℓ^2(\Omega)} ≤ \varepsilon_{m,j}
\]  \hspace{1cm} (4.3.15)

for all \(m ∈ \mathbb{N}\).

Remark 4.3.6. Estimates of the form (4.3.15) are shown for example in [BDD06, Ste04] for a large class of operators in wavelet bases. We note that these a priori compression error estimates are asymptotic, i.e. they hold only up to an unspecified constant. For our purposes, it is important to know the dependence of this constant on \(m ∈ \mathbb{N}_0\). In particular, \(\varepsilon_{m,j}\) should be chosen such that (4.3.15) holds at least up to a uniform constant. We assume for simplicity that such bounds are known exactly.

For all finitely supported sequences \(j := (j_m)_{m∈\mathbb{N}_0}\) in \(\mathbb{N}_0\), define the operator

\[
A_j := I ⊗ D_0 + \sum_{m=1}^{∞} K_m ⊗ R_{m,j} .
\]  \hspace{1cm} (4.3.16)

Let \(σ_m := 2\) if the distribution \(π_m\) is symmetric, and \(σ_m := 3\) otherwise. We set \(σ_0 := 1\) and define \(n_{m,j} := σ_m n_{m,j}\) for \(m ∈ \mathbb{N}_0\). Then for all \(j ∈ \mathbb{N}_0\), \(I ⊗ D_j\) is \(n_{0,j}\)-sparse and \(K_m ⊗ R_{m,j}\) is \(n_{m,j}\)-sparse, \(m ∈ \mathbb{N}\).

Lemma 4.3.7. For any finitely supported sequence \(j = (j_m)_{m∈\mathbb{N}_0}\) in \(\mathbb{N}_0\), \(A_j\) is \(N_j\)-sparse for

\[
N_j := \sum_{m=0}^{∞} n_{m,j} ,
\]  \hspace{1cm} (4.3.17)

and

\[
\|A - A_j\|_{ℓ^2(\mathbb{Z})→ℓ^2(\mathbb{Z}×\Omega)} ≤ \sum_{m=0}^{∞} \varepsilon_{m,j} =: \varepsilon_{A,j} .
\]  \hspace{1cm} (4.3.18)

Proof. The first part of the assertion follows by construction since \(I\) is 1-sparse and \(K_m\) is \(σ_m\)-sparse for all \(m ∈ \mathbb{N}\). Equation (4.3.18) is a consequence of Proposition 3.2.13 and Lemma 3.2.8. □

We use the greedy algorithm from Section 4.1 to select specific \(j\) in (4.3.16). The cost \(c_j\) and objective \(ω_j\) are given by

\[
c_j := N_j = \sum_{m=0}^{∞} n_{m,j} \quad \text{and} \quad ω_j := -\varepsilon_{A,j} = \sum_{m=0}^{∞} -\varepsilon_{m,j} .
\]  \hspace{1cm} (4.3.19)
4.3. Adaptive Application of Parametric Operators

We initialize \( j_0 := 0 \in \mathbb{N}_0^{N_0} \) and construct \((j_k)_{k \in \mathbb{N}_0}\) recursively by

\[
j_{k+1} := \text{NextOptInf}[j_k], \quad k \in \mathbb{N}_0,
\]

using (4.3.19). Then

\[
A_k := A_{j_k}, \quad k \in \mathbb{N}_0,
\]

defines a sequence of approximations of \( A \). By Lemma 4.3.7, \( A_k \) is \( N_k := N_j \)-sparse and its distance to \( A \) is bounded by \( \bar{e}_{A,k} := \bar{e}_{A,j} \).

Under mild assumptions, (4.3.21) defines the optimal \( N_k \)-sparse approximation of \( A \) given the bounds (4.3.15) and the estimates in Lemma 4.3.7.

**Assumption 4.3.A.** For all \( m \in \mathbb{N}, n_{m,0} = 0 \) and the \((n_{m,j})_{j \in \mathbb{N}_0}\) is strictly increasing. The sequence \((\bar{e}_{m,j})_{m \in \mathbb{N}}\) is in \( \ell^1 \), and \((e_{m,j})_{j \in \mathbb{N}_0}\) is nonincreasing. Furthermore, if \( i \geq j \), then

\[
-\left(\frac{(\bar{e}_{m,j+1} - \bar{e}_{m,j})}{n_{m,j+1} - n_{m,j}}\right) \leq \frac{(\bar{e}_{m,j+1} - \bar{e}_{m,j})}{n_{m,j+1} - n_{m,j}},
\]

and \( n_{m,1}(\bar{e}_{m,1} - \bar{e}_{m,0}) \) is nonincreasing in \( m \).

**Corollary 4.3.8.** For all \( k \in \mathbb{N}_0, j_k \) minimizes the error bound \( \bar{e}_{A,j} \) among all finitely supported sequences \( j \) in \( \mathbb{N}_0 \) with sparsity bound \( N_j \leq N_k \). Furthermore, if \( \bar{e}_{A,k} \neq 0 \), then \( j_k \) minimizes \( N_j \) among all \( j \) with \( \bar{e}_{A,j} \leq \bar{e}_{A,k} \).

**Proof.** The assertion follows from Theorem 4.1.5, see Remark 4.1.1, since Assumption 4.3.A implies Assumption 4.1.A for (4.3.19).

We consider the complexity of a routine \( \text{Build}_A \) as in Def. 4.2.4 for constructing columns of \( A_k \), interpreted as bi-infinite matrices. To this end, we assume that such assembly routines are available for \( D \) and \( R_m, m \in \mathbb{N} \). More specifically, the routines

\[
\text{Build}_0[j, i] \mapsto \left[(\lambda_i)_{i=1}^{n_0}, (d_i)_{i=1}^{n_0}\right],
\]

\[
\text{Build}_m[j, i] \mapsto \left[(\lambda_i)_{i=1}^{n_m}, (\mu_i)_{i=1}^{n_m}\right], \quad m \in \mathbb{N},
\]

construct all nonzero elements of the \( i \)-th column of \( D_j \) and \( R_{m,j} \), respectively, using no more than \( b_m n_{m,j} \) arithmetic operations and storage locations for a constant \( b_m \) independent of \( j \) and \( i \).

**Lemma 4.3.9.** The number of arithmetic operations and storage locations required by a call of \( \text{Build}_A[k, (\mu, i)] \) is bounded uniformly in \( k \) by

\[
N_k + \sum_{m=0}^{\infty} b_m n_{m,k,i}.
\]

**Proof.** This is a direct consequence of the assumptions on \( \text{Build}_m, m \in \mathbb{N}_0 \).
Remark 4.3.10. It is often necessary to construct $j_k$ before calling $\text{Build}_A[k, \cdot]$, for example to determine $N_k$ and $\tilde{e}_{A,k}$. In this case, we can assume $j_k$ to be readily available in $\text{Build}_A[k, \cdot]$. Otherwise, $\text{NextOptInf}$ can be used to compute $j_k$ in the first call of $\text{Build}_A[k, \cdot]$. If this is done directly for an arbitrary $k \in \mathbb{N}_0$, it adds $O(k \log(k))$ to the complexity of $\text{Build}_A[k, \cdot]$ even if $\mathcal{N}$ is realized by a tree data structure, which may dominate e.g. if $N_k \leq k$. However, if $\text{Build}_A[k, \cdot]$ is called successively for $k \in \mathbb{N}$ and the values $j_k, \mathcal{N}$ and $M$ are cached, then the cost of $\text{NextOptInf}$ is negligible even if $\mathcal{N}$ is realized by a simple linked list.

4.3.3. $s'$-Compressibility and $s'$-Computability

We establish $s'$-compressibility of $A$ in Theorem 4.3.14, and $s'$-computability in Corollary 4.3.16. To this end, we first derive some preliminary estimates.

For an $s > 0$, assume for the moment that $D$ and $R_m, m \in \mathbb{N}$, are strictly $s$-compressible. By Proposition 4.2.3, there is a map $j_0: [0, \infty) \to \mathbb{N}_0$ such that the sparse approximation $D_{j_0(r)}$ is $r$-sparse and

$$
\|D - D_{j_0(r)}\|_{\mathcal{E}(\mathbb{S}) \rightarrow \mathcal{E}(\mathbb{T})} \leq \tilde{e}_{0,j_0(r)} \leq \tilde{d}_{0,3} r^{-s}, \quad r > 0,
$$

(4.3.23)
4.3. Adaptive Application of Parametric Operators

with \( d_{0,s} := d_{D,s} \).\(^2\) Similarly, for all \( m \in \mathbb{N} \) there is a map \( j_m : [0, \infty) \to \mathbb{N}_0 \) such that the sparse approximation \( R_{m,j_m(r)} \) is \( m^{-1} \)-sparse and

\[
\left\| R_m - R_{m,j_m(r)} \right\|_{\ell^2(\mathbb{Z}) \to \ell^2(\mathbb{Z})} \leq \bar{e}_{m,j_m(r)} \leq d_{m,s} r^{-s}, \quad r > 0,
\]

(4.3.24)

with \( d_{m,s} := \sigma_m d_{R,s} \).

**Lemma 4.3.11.** If \( (d_{m,s})_m \in \ell^{\frac{1}{m^s}}(\mathbb{N}_0) \), then for all \( r > 0 \) there is a finitely supported sequence \( j(r) \) in \( \mathbb{N}_0 \) such that \( N_{j(r)} \leq r \) and

\[
\bar{e}_{A,j(r)} \leq \left( \sum_{n=0}^{\infty} d_{m,s} \right)^{s+1} r^{-s}.
\]

(4.3.25)

**Proof.** Let \( t > 0 \) and define \( r_m := \frac{d_{m,s}}{m^{-s}} t \) for all \( m \in \mathbb{N}_0 \). Set \( j := (j_m(r_m))_{m \in \mathbb{N}_0} \). This sequence is finitely supported since \( r_m < 1 \) for all but finitely many \( m \in \mathbb{N}_0 \). By Lemma 4.3.7,

\[
N_j = \sum_{m=0}^{\infty} \bar{e}_{m,j_m(r_m)} \leq \sum_{m=0}^{\infty} d_{m,s} R_m \leq d_{m,s} t^s =: r
\]

and

\[
\bar{e}_{A,j} = \sum_{m=0}^{\infty} \bar{e}_{m,j_m(r_m)} \leq \sum_{m=0}^{\infty} \bar{e}_{m,s} R_m \leq \sum_{m=0}^{\infty} \frac{1}{m^{-s}} \left( \sum_{n=0}^{\infty} d_{m,s} \right)^{s+1} r^{-s}.
\]

\[\square\]

If \( (d_{m,s})_m \) is not in \( \ell^{\frac{1}{m^s}}(\mathbb{N}_0) \), a similar property still holds if we replace the infinite sum by a partial sum. Define

\[
\bar{e}_{A[M],j} := \sum_{m=0}^{M} \bar{e}_{m,j_m}.
\]

(4.3.26)

Then for all sequences \( j \) in \( \mathbb{N}_0 \) with support in \([0, 1, \ldots, M]\),

\[
\bar{e}_{A,j} = \bar{e}_{A[M],j} + \sum_{m=M+1}^{\infty} \bar{e}_{m,0}.
\]

(4.3.27)

**Lemma 4.3.12.** For all \( M \in \mathbb{N}_0 \) and all \( r > 0 \), there is a sequence \( j(r) \) in \( \mathbb{N}_0 \) with support in \([0, 1, \ldots, M]\) such that \( N_{j(r)} \leq r \) and

\[
\bar{e}_{A[M],j(r)} \leq \left( \sum_{m=0}^{M} d_{m,s} \right)^{s+1} r^{-s}.
\]

(4.3.28)

**Proof.** The proof is analogous to the proof of Lemma 4.3.11. \(\square\)

\(^2\)Proposition 4.2.3 initially only implies that the first term in (4.3.23) is bounded by the third. However, if (4.3.23) does not hold, we can replace \( \bar{e}_{0,j(r)} \) by \( d_{0,s} r^{-s} \) in (4.3.15).
Chapter 4. Adaptive Application of the Discrete Operator

Proposition 4.3.13. Let (4.3.12) or (4.3.13) be satisfied for an \( s_0 > 0 \) and
\[
\left( \sum_{n=0}^{M} d_{m,1} \right)^{s+1} \leq d_{s} M^t, \quad M \in \mathbb{N},
\]
(4.3.29)
with \( d_{s} > 0 \) and \( t_{s} \geq 0 \). Then for all \( r \in [1, \infty) \) there is a finitely supported sequence \( j(r) \) in \( \mathbb{N}_0 \) such that \( N_{j(r)} \leq r \) and
\[
\bar{e}_{A,j(r)} \leq (\hat{d}_{s} + \delta_{A,s_0}) r^{\frac{s}{s(1+s)}}.
\]
(4.3.30)
Equation (4.3.14) implies
\[
\sum_{m=M+1}^{\infty} \bar{e}_{m,0} \leq \delta_{A,s_0} (M + 1)^{-s_0} \leq \delta_{A,s_0} r^{rac{s_0}{s(1+s)}}.
\]
Then the assertion follows using (4.3.27). \( \square \)

The above estimates combine with Corollary 4.3.8 to show \( s'-\text{compressibility} \) of \( A \) with the approximating sequence \( (A_k)_{k \in \mathbb{N}} \) from Section 4.3.2. Define the constants
\[
\bar{c}_m := \max \left( \bar{\eta}_{m,1}, \sup_{j \in \mathbb{N}} \frac{\bar{\eta}_{m,j+1}}{\bar{\eta}_{m,j}} \right) < \infty, \quad m \in \mathbb{N}_0.
\]
(4.3.31)
Note that \( c_D \leq \bar{c} \) and \( c_{R,s} \leq \sigma_m \bar{c}_m \) for \( m \in \mathbb{N} \).

Theorem 4.3.14. Let \( s_0^*, s_0^* \in (0, \infty) \) and assume
\[
\bar{c} := \sup_{m \in \mathbb{N}_0} \bar{c}_m < \infty.
\]
(4.3.32)
1. If \( (\bar{d}_{m,s})_s \in \ell^\infty(\mathbb{N}_0) \) for all \( s \in (0, s_0^*) \), then \( A \) is \( s'-\text{compressible} \) for \( s' = s_0^* \).
2. If (4.3.12) or (4.3.13) holds for all \( s \in (0, s_0^*) \) and (4.3.29) holds for all \( s \in (0, s_0^*) \) with \( t_{s} \leq \hat{t} < \infty \), then \( A \) is \( s'-\text{compressible} \) for
\[
s' = \frac{s_0^*}{1 + \hat{t}/s_0^*}.
\]
(4.3.33)
In both cases, \( (A_k)_{k \in \mathbb{N}} \) is a valid approximating sequence with \( c_A \leq \bar{c} \),
\[
d_{A,s} \leq \|(\bar{d}_{m,s})_s\|_{\ell^\infty(\mathbb{N}_0)}, \quad s \in (0, s')
\]
(4.3.34)
in the first case and
\[
d_{A,s} \leq \inf_{s_0^* < s \leq s_0^*} \left( \bar{d}_{s(1+\hat{t}/s_0^*)} + \delta_{A,s_0} \right), \quad s \in (0, s')
\]
(4.3.35)
in the second case.
4.3. Adaptive Application of Parametric Operators

Proof. Condition (4.3.32) ensures (4.2.4) for $(A_k)_{k \in \mathbb{N}}$ since for $k \in \mathbb{N}$ and $j := k_{m_j}$, if $j \geq 1$, 
\[
\frac{N_{k+1}}{N_k} = \frac{N_k + \bar{r}_{m_k,j+1} - \bar{r}_{m_k,j}}{N_k} = \frac{n + \bar{r}_{m_k,j+1}}{n + \bar{r}_{m_k,j}} \leq \frac{\bar{r}_{m_k,j+1}}{\bar{r}_{m_k,j}} \leq \varepsilon_{n_k},
\]
where $n = N_k - \bar{r}_{m_k,j} \geq 0$, and if $j = 0$, 
\[
\frac{N_{k+1}}{N_k} = \frac{N_k + \bar{r}_{m_k,1}}{N_k} \leq \bar{r}_{m_k,1} \leq \varepsilon_{n_k}.
\]
Let $s \in (0, s^*)$. In case 1, Corollary 4.3.8 and Lemma 4.3.11 with $r = N_k$ imply 
\[
\varepsilon_{A,k} \leq \varepsilon_{A,j(N_k)} \leq \left( \sum_{m=0}^{\infty} \frac{d_m}{m} \right)^{s+1} N_k^{-s}.
\]
In case 2, select $s_0 \in (0, s^*_0)$ and $s_\sigma \in (0, s^*_\sigma)$ such that 
\[
s = \frac{s_0}{1 + t/s_\sigma}.
\]
This is possible since the right hand side is increasing in $s_0$ and $s_\sigma$. By monotonicity, (4.3.29) holds with $t_s = \hat{t}$. Then Corollary 4.3.8 and Proposition 4.3.13 with $r = N_k$ imply 
\[
\varepsilon_{A,k} \leq \varepsilon_{A,j(N_k)} \leq \left( d_{s_0} + \delta_{A,s_\sigma} \right) N_k^{-s}.
\]
Equation (4.3.35) follows since $s_0 = s(1 + t/s_\sigma).$ \hspace{1cm} \Box

Remark 4.3.15. Comparing the first case of Theorem 4.3.14 with Theorem 4.3.2, we note that the conditions on $\|R_m\|_{V \rightarrow W'}$ and $d_{m,s}$ are very similar. This can be interpreted as uniform $s^*$-compressibility of the rescaled operators $\|R_m\|_{V \rightarrow W'} R_m$. 

Under the assumption that the sequence $(j_k)_{k \in \mathbb{N}}$ is available, $s^*$-computability of $A$ follows from Theorem 4.3.14 as a corollary.

Corollary 4.3.16. In the setting of Theorem 4.3.14, if 
\[
\sup_{m \in \mathbb{N}_0} b_m < \infty \tag{4.3.36}
\]
for $b_m$ from Section 4.3.2 and the sequences $j_k$ are given as in Remark 4.1.9, then $A$ is $s^*$-computable and $\text{Build}_A$ is a valid assembly routine.

Proof. $s^*$-computability follows from Theorem 4.3.14. By Lemma 4.3.9, (4.3.36) and Remark 4.1.9, the number of arithmetic operations and storage locations required by a call of $\text{Build}_A[k, \cdot]$ is $O(N_k)$.

If $j_k$ are not readily available, Proposition 4.1.7 implies that recursive application of $\text{NextOptInf}$ can construct $j_k$ in $O(k \log(k))$ time. Thus $A$ is still $s^*$-computable if $k \log(k) \leq N_k$. As discussed in Remark 4.3.10, the cost of computing $j_k$ from $j_{k-1}$ using $\text{NextOptInf}$ is only $O(\log(k))$. Therefore, if $\text{NextOptInf}$ is used to construct $j_k$ in the first call of $\text{Build}_A[k, \cdot]$, then $\text{Build}_A[k, \cdot]$ requires $O(N_k)$ operations provided that $j_{k-1}$ is known, for example from a previous call of $\text{Build}_A[k-1, \cdot]$. \hspace{1cm} \Box

77
Chapter 4. Adaptive Application of the Discrete Operator

4.3.4. Adaptive Application Routine

In this section, we analyze the structure of the adaptive multiplication routine $\text{Apply}_A$ from Section 4.2.2 for a discretized parametric operator $A$ and the approximating sequence $(A_k)$ from Section 4.3.2.

By Assumption 4.3.A and Lemma 4.3.9, $(N_k)_{k \in \mathbb{N}}$ is strictly increasing, and $N_0 = 0$ since $j_0 = 0$. By definition, $(j_{k,m})_{k \in \mathbb{N}_0}$ is nondecreasing for all $m \in \mathbb{N}_0$. Therefore, Assumption 4.3.A implies that $(\bar{e}_{A,k})_{k \in \mathbb{N}_0}$ is nondecreasing. If $\bar{e}_{m,j} \to 0$ as $j \to \infty$ for all $m \in \mathbb{N}_0$, since $(\bar{e}_{m,0})_{m \in \mathbb{N}_0} \in \ell^1$ by Assumption 4.3.A, Corollary 4.3.8 implies that $\bar{e}_{A,k} \to 0$ as $k \to \infty$. We note that

$$\eta_k = \frac{\bar{e}_{A,k} - \bar{e}_{A,k+1}}{N_{k+1} - N_k} = \frac{\bar{e}_{mk,j_{k,m}} - \bar{e}_{mk,j_{k,m}+1}}{n_{mk,j_{k,m}+1} - n_{mk,j_{k,m}}},$$

(4.3.37)

which is nonincreasing in $k$ by construction of $(j_k)_{k \in \mathbb{N}_0}$, see Lemma 4.1.4. Consequently, Assumption 4.2.A is satisfied under the sole additional requirement that $\bar{e}_{m,j} \to 0$ as $j \to \infty$ for all $m \in \mathbb{N}_0$.

Also, since

$$\frac{\bar{e}_{A,k}}{\bar{e}_{A,k+1}} \leq \frac{\bar{e}_{A,k}}{\bar{e}_{A,k+1}} \leq \frac{\bar{e}_{mk,j_{k,m}}}{\bar{e}_{mk,j_{k,m}+1}},$$

Assumption 4.2.B is satisfied if

$$\sup_{m \in \mathbb{N}_0} \sup_{\bar{e}_{m,j+1}} < \infty .$$

(4.3.38)

Assuming the sequences $(j_k)$ and $(m_k)$ are known, the first two parts of $\text{Apply}_A[v, \varepsilon]$ can be used to partition the vector $v$ into $(v[p])_{p=1}^\ell$ and a negligible remainder term, and to assign to each of these a $A_p \in \mathbb{N}_0$.

The final step of $\text{Apply}_A[v, \varepsilon]$ performs the multiplications

$$z := \sum_{p=1}^\ell A_{j_p} v[p] .$$

(4.3.39)

Using the tensor product structure from Proposition 3.2.13, (4.3.39) can be decomposed into multiplications with the coefficient operators $D_j$ and $R_{m,j} m \in \mathbb{N}$.

Let $v[p]_{\mu}$ denote the $\mu$-th coefficient of $v[p]$, i.e., $v[p]_{\mu} = (v_{\mu})_i$ for $i \in \Xi$ such that $(\mu, i) \in \Xi_p$. Then assuming $\pi_m$ is symmetric for all $m \in \mathbb{N}$, $z = (z_{\mu})_{\mu \in \Lambda}$ with

$$z_{\mu} = \sum_{p=1}^\ell \left( D_{j_{p},\mu} v[p]_{\mu} + \sum_{m=1}^{M_{\mu}} \rho_{\mu,m+1} R_{m,j_{p,m}} v[p]_{\mu+1} + \rho_{\mu,m} R_{m,j_{p,m}} v[p]_{\mu} \right) ,$$

(4.3.40)

where $M_{\mu} := \max \{ m \in \mathbb{N}_0 \mid j_{k,m} \neq 0 \}$. This does not, however, represent an efficient way to construct $z$. It is not clear which $z_{\mu}$ are nonzero, and many multiplications with $R_{m,j}$ are done twice. The routine $\text{Multiply}_A$ does the same computation efficiently, for arbitrary $\pi_m$, by looping over $p$ and the index set of $v[p]$.  

78
4.3. Adaptive Application of Parametric Operators

\[
\text{Multiply}_A[(v[p])^\ell_{p=1}, (k[p])^\ell_{p=1}] \mapsto z
\]

\[
z \leftarrow 0
\]

\[
\text{for } p = 1, \ldots, \ell \text{ do}
\]

\[
\text{forall } \mu \in \Lambda \text{ with } v[p][\mu] \neq 0 \text{ do}
\]

\[
z_\mu \leftarrow z_\mu + D_{h[p]} v[p][\mu]
\]

\[
\text{for } m = 1, \ldots, M_p \text{ do}
\]

\[
w \leftarrow R_{m, h[p]} v[p][\mu]
\]

\[
z_{\mu + e_m} \leftarrow z_{\mu + e_m} + \beta_{\mu+1}^m w
\]

\[
\text{if } \mu_m \geq 1 \text{ then } z_{\mu - e_m} \leftarrow z_{\mu - e_m} + \beta_{\mu_m}^m w
\]

\[
\text{if } \sigma_m = 3 \text{ then } z_\mu \leftarrow z_\mu + a_{\mu_m}^m w
\]

end

end

Remark 4.3.17. In \(\text{Multiply}_A[(v[p])^\ell_{p=1}, (k[p])^\ell_{p=1}]\), each multiplication with \(R_{m,j}\) is performed only once, and copied to \(\sigma_m\) components of \(z\). This suggests defining \(\bar{n}_m \coloneqq n_{R_m}\) for \(m \in \mathbb{N}\), without the factor of \(\bar{\sigma}_m\) from the original definition.

Remark 4.3.18. By Proposition 3.2.13, the discrete adjoint \(A^*\) of a discretized parametric operator \(A\) has the same tensor product structure as \(A\). Therefore, sparse approximations of \(A^*\) can be constructed as in Section 4.3.2, with \(D\) and \(R_{m,j}, m \in \mathbb{N}\), replaced by their adjoints. Theorem 4.3.14 carries over to show \(s^*\)-compressibility of \(A^*\) under suitable assumptions, and \(s^*\)-computability follows as a corollary. In particular, \(\text{Apply}_A^*\) has the same structure as described above.

As suggested in [CDD02], \(\text{Apply}_A^*\) and \(\text{Apply}_A\) combine to an adaptive multiplication routine for \(A^*A\),

\[
\text{Apply}_A^*A[v, \epsilon] \mapsto \text{Apply}_A[\text{Apply}_A^*[v, \epsilon/(2\bar{\sigma}_A)], \epsilon/2], \quad (4.3.41)
\]

see e.g. [SS09, Cor. 4.6].
Chapter 5.

Adaptive Iterative Solvers

In Chapter 4, we applied adaptive wavelet methods to parametric operator equations in a fully discrete setting, using tensor product polynomials on the parameter domain and suitable Riesz bases or frames in the physical space. In principle, these methods satisfy all the requirements for adaptive solution procedures for parametric or stochastic operator equations. However, full adaptivity requires complex data structures tailored for manipulating indices of the product of an infinite tensor product polynomial basis and, for example, a wavelet basis.

We suggest applying an adaptive wavelet method only for the tensor product polynomial bases on the parameter domain. This can be combined with any discretization of the physical domain, leading to a modular adaptive solver which takes advantage of the structure of the discretized parametric operator equation.

We consider a basic adaptive wavelet method along the lines of [CDD02], which consists of perturbing a linear iteration for the full operator equation. We extend this approach to the semidiscrete setting of vector coefficients, and customize it to tensor product polynomial bases. In this setting, approximate application of the discretized parametric operator amounts to truncating the series expansion of parametric perturbations for each coefficient, possibly at different numbers of terms.

Iterations can easily be derived that converge uniformly in the parameter. If the error induced by approximate application of the operator is controlled with respect to the maximum norm, the resulting adaptive method also converges uniformly, with an explicit bound on the maximal error over the entire parameter domain. This is highly desirable for parametric equations in which no measure is given on the parameter domain, and for stochastic equations with probability measures that are not product measures on the parameter domain. In the latter case, uniform convergence dispenses with the need for elusive absolute continuity assumptions on the probability measure.

We develop this adaptive method in Section 5.1 with no discretization of the physical domain, for both mean square and uniform convergence. In Section 5.2, we discuss two extensions. For symmetric distributions, reordering some computations leads to faster convergence. Also, adding a coarsening step to the iteration prevents the approximate solution from amassing too many insignificant coefficients.

In Section 5.3, we add a spatial discretization to the adaptive method. We consider both a single discretization for all coefficients, and an independent adaptive discretization for each coefficient. Due to its modular design, our method can be coupled with any solver for the nonparametric operator equation, including adaptive wavelet methods.
Chapter 5. Adaptive Iterative Solvers

and adaptive finite element methods.

5.1. A General Iterative Method

5.1.1. An Abstract Perturbed Stationary Linear Iteration

Let $X$ and $Y$ be Banach spaces, and let $A \in \mathcal{L}(X, Y)$ be of the form

$$A = D + R \quad (5.1.1)$$

with $D \in \mathcal{L}(X, Y)$ boundedly invertible, and $R \in \mathcal{L}(X, Y)$ satisfying

$$\|D^{-1}R\|_{X \to Y} \leq \gamma < 1. \quad (5.1.2)$$

In the setting of Section 1.3.2, in which $A$ is a parametric operator with affine parameter dependence, $D$ is the constant part of $A$ and $R$ is the linear part, (5.1.2) holds with $\gamma$ from (1.3.16) for $X = L^p(\Gamma) \otimes \ov{\alpha} V$ and $Y = L^p(\Gamma) \otimes \ov{\alpha} W^*$ for $1 \leq p < \infty$, or $X = C(\Gamma) \otimes \ov{\alpha} V$ and $Y = C(\Gamma) \otimes \ov{\alpha} W^*$, where $\ov{\alpha}$ is any tensor norm.

The condition (5.1.2) implies that $D^{-1}A$ can be inverted by a Neumann series in $X$. Therefore, for any $f \in Y$, the operator equation

$$Au = f \quad (5.1.3)$$

can be solved by the iteration $u_0 := 0 \in X$, $u_k := D^{-1}(f - Ru_{k-1})$, $k \in \mathbb{N}$. \quad (5.1.4)

More specifically, for all $k \in \mathbb{N}_0$,

$$\|u - u_k\|_X \leq \gamma^k \|u\|_X \leq \gamma^k \frac{\|D^{-1}\|_{Y \to X}}{1 - \gamma} \|f\|_Y \quad (5.1.5)$$

since

$$u - u_k = -D^{-1}R(u - u_{k-1})$, \quad k \in \mathbb{N}. \quad (5.1.6)$$

We generalize (5.1.4) by allowing errors in the computation of $f$, the evaluation of $R$, and the inversion of $D$.

Let $u_0 := 0 \in X$ as above, and

$$\delta_0 := \frac{\|D^{-1}\|_{Y \to X}}{1 - \gamma} \|f\|_Y. \quad (5.1.7)$$

Then by (5.1.2),

$$\|u - u_0\|_X = \|u\|_X \leq \delta_0. \quad (5.1.8)$$

\footnote{We denote the general tensor norm by $\ov{\alpha}$ in this chapter to distinguish it from the parameter $\alpha$ from (5.1.10).}
5.1. A General Iterative Method

For all $k \in \mathbb{N}$, let $g_k \in \mathcal{Y}$ with

$$
\|g_k - (f - \mathcal{A}u_{k-1})\|_{\mathcal{Y}} \leq \beta \delta_{k-1} \|D^{-1}\|_{\mathcal{Y} \to \mathcal{X}}, \tag{5.1.9}
$$

and let $u_k \in \mathcal{X}$ satisfy

$$
\|u_k - D^{-1}g_k\|_{\mathcal{X}} \leq \alpha \delta_{k-1}, \tag{5.1.10}
$$

where $\delta_{k-1}$ is an upper bound for $\|u - u_{k-1}\|_{\mathcal{X}}$ and $\alpha, \beta \geq 0$ are independent of $k$.

**Theorem 5.1.1.** Let $u_k$ and $g_k$ satisfy (5.1.9) and (5.1.10) for any upper bound $\delta_{k-1}$ of $\|u - u_{k-1}\|_{\mathcal{X}}$. Then

$$
\|u - u_k\|_{\mathcal{X}} \leq \delta_k := (\alpha + \beta + \gamma) \delta_{k-1}. \tag{5.1.11}
$$

In particular, if $\alpha + \beta + \gamma < 1$, then $u_k \to u$ in $\mathcal{X}$, and

$$
\|u - u_k\|_{\mathcal{X}} \leq (\alpha + \beta + \gamma) \left( \frac{\|D^{-1}\|_{\mathcal{Y} \to \mathcal{X}}}{1 - \gamma} \right) \|f\|_{\mathcal{Y}} \quad \forall k \in \mathbb{N}_0. \tag{5.1.12}
$$

**Proof.** Since $\mathcal{A}u = f - \mathcal{A}u$,

$$
u - u_k = D^{-1}(f - \mathcal{A}u) - D^{-1}(f - \mathcal{A}u_{k-1}) + D^{-1}(f - \mathcal{A}u_{k-1} - g_k) + D^{-1}g_k - u_k.
$$

By triangle inequality,

$$
\|u - u_k\|_{\mathcal{X}} \leq \|D^{-1}(u - u_{k-1})\|_{\mathcal{X}} + \|D^{-1}\|_{\mathcal{Y} \to \mathcal{X}} \|g_k - (f - \mathcal{A}u_{k-1})\|_{\mathcal{Y}} + \|u_k - D^{-1}g_k\|_{\mathcal{X}}
\leq \gamma \|u - u_{k-1}\|_{\mathcal{X}} + \beta \delta_{k-1} + \alpha \delta_{k-1}.
$$

Equation (5.1.11) follows by the assumption that $\delta_{k-1}$ is greater than $\|u - u_{k-1}\|_{\mathcal{X}}$. If $\alpha + \beta + \gamma < 1$, repeated application with $\delta_k$ defined as in (5.1.11), using (5.1.7) to estimate the initial error, leads to (5.1.12). □

**Remark 5.1.2.** Theorem 5.1.1 uses a priori known quantities $\delta_k = (\alpha + \beta + \gamma)^k \delta_0$ as upper bounds for the error at iteration $k \in \mathbb{N}_0$. However, better estimates may be available or computable during an iteration. The residual at iteration $k \in \mathbb{N}_0$ is given by

$$
r_k := f - \mathcal{A}u_k = \mathcal{A}(u - u_k) \in \mathcal{Y}. \tag{5.1.13}
$$

Since $\mathcal{A}$ is invertible by a Neumann series,

$$
\|u - u_k\|_{\mathcal{X}} \leq \|D^{-1}\|_{\mathcal{Y} \to \mathcal{X}} \|r_k\|_{\mathcal{Y}} \leq \frac{1}{1 - \gamma} \|D^{-1}\|_{\mathcal{Y} \to \mathcal{X}} \|r_k\|_{\mathcal{Y}}. \tag{5.1.14}
$$

Therefore, if it is known that $\|r_k\|_{\mathcal{Y}} \leq \rho_k$, we may define $\delta_k$ as

$$
\delta_k := \min \left( (\alpha + \beta + \gamma) \delta_{k-1}, \frac{1}{1 - \gamma} \|D^{-1}\|_{\mathcal{Y} \to \mathcal{X}} \rho_k \right) \tag{5.1.15}
$$

for all $k \in \mathbb{N}$. This combines the a priori bound with a posteriori information on the residual $r_k$. □
Chapter 5. Adaptive Iterative Solvers

5.1.2. Realization by Adaptive Subroutines

We consider the setting of Sections 1.2.2 and 1.3.2. For a tensor norm \( \alpha \) and Banach spaces \( V \) and \( W \), \( \mathcal{A} = \mathcal{D} + \mathcal{B} \) is a continuous linear operator from \( L^2_\pi(\Gamma) \otimes_{\mathcal{A}} V \) to \( L^2_\pi(\Gamma) \otimes_{\mathcal{A}} W \), which satisfies (1.3.16) and thus is invertible. Let \( P = (P_\nu)_{\nu \in \Lambda} \) be a tensor product orthonormal polynomial basis of \( L^2_\pi(\Gamma) \) as in (3.2.4). By Proposition 3.2.7, the representation of \( \mathcal{A} \) in the basis \( P \) is

\[
\mathcal{A} = T^W_P \mathcal{A} T^V_P = I \otimes D + \sum_{m \in \mathcal{M}} K_m \otimes R_m
\]  

(5.1.16)

for \( K_m := T^*_m T_P \) and \( I := \text{id}_{\mathcal{D}(\Lambda)} \). We decompose \( \mathcal{A} \) as \( \mathcal{A} = \mathcal{D} + \mathcal{R} \) with \( \mathcal{D} := T^W_P \mathcal{A} T^V_P = I \otimes D \) and \( \mathcal{R} := T^W_P \mathcal{A} T^V_P = \sum_{m \in \mathcal{M}} K_m \otimes R_m \). Thus we are in the setting of Section 5.1.1 with \( \mathcal{R} = \ell^2(\Lambda) \otimes_{\mathcal{A}} V \) and \( \mathcal{Y} = \ell^2(\Lambda) \otimes_{\mathcal{A}} W \). Condition (5.1.2) is ensured by (1.3.16).

As in Section 4.3.1, we assume without loss of generality that \( \mathcal{M} = \mathbb{N} \) and the sequence \( (\|R_m\|_{V \rightarrow W^*})_{m \in \mathbb{N}} \) is nonincreasing. For all \( M \in \mathbb{N}_0 \), we define the operator

\[
\mathcal{R}_{[M]} := \sum_{m=1}^{M} K_m \otimes R_m
\]  

(5.1.18)

in \( \mathcal{L}(\ell^2(\Lambda) \otimes_{\mathcal{A}} V, \ell^2(\Lambda) \otimes_{\mathcal{A}} W^*) \). Since \( T^V_P \) is an isometric isomorphism from \( L^2_\pi(\Gamma) \otimes_{\mathcal{A}} V \) to \( \ell^2(\Lambda) \otimes_{\mathcal{A}} V \), and \( T^W_P \) is an isometric isomorphism from \( \ell^2(\Lambda) \otimes_{\mathcal{A}} W^* \) to \( L^2_\pi(\Gamma) \otimes_{\mathcal{A}} W^* \), we may equivalently consider

\[
\mathcal{R}_{[M]} := \sum_{m=1}^{M} K_m \otimes R_m
\]  

(5.1.19)

in \( \mathcal{L}(L^2_\pi(\Gamma) \otimes_{\mathcal{A}} V, L^2_\pi(\Gamma) \otimes_{\mathcal{A}} W^*) \). These operators are related by \( \mathcal{R}_{[M]} = T^W_P \mathcal{R}_{[M]} T^V_P \).

**Remark 5.1.3.** Under the conditions of Theorem 4.3.2 for an \( s > 0 \),

\[
\left\| \mathcal{R} - \mathcal{R}_{[M]} \right\|_{\ell^2(\Lambda) \otimes_{\mathcal{A}} V \rightarrow \ell^2(\Lambda) \otimes_{\mathcal{A}} W^*} \leq \delta_{\mathcal{R},s}(M + 1)^{-s} \quad \forall M \in \mathbb{N}_0
\]  

(5.1.20)

since \( \mathcal{R} = \mathcal{A} - \mathcal{D} \) and \( \mathcal{R}_{[M]} = \mathcal{A}_{[M]} - \mathcal{D} \). If the assumptions hold for all \( s \in (0, s') \), then it follows as in Corollary 4.3.3 that \( \mathcal{R} \) is \( s' \)-compressible with \( c_{\mathcal{R}} = 2 \) and \( d_{\mathcal{R},s} \leq 3\delta_{\mathcal{R},s} \) since \( \mathcal{R}_{[M]} \) is \( 3M \)-sparse. If all of the distributions \( \pi_m, m \in \mathbb{N} \), are symmetric, then \( d_{\mathcal{R},s} \leq 2\delta_{\mathcal{R},s} \) since \( \mathcal{R}_{[M]} \) is \( 2M \)-sparse.

We extend the adaptive application routine from Section 4.2.2 to the Banach space setting and specialize to the operator \( \mathcal{R} \). For all \( M \in \mathbb{N}_0 \), let explicit values \( N_M \in \mathbb{N}_0 \) and \( \varepsilon_{\mathcal{R},M} \geq 0 \) be given such that \( \mathcal{R}_{[M]} \) is \( N_M \)-sparse and

\[
\left\| \mathcal{R} - \mathcal{R}_{[M]} \right\|_{\ell^2(\Lambda) \otimes_{\mathcal{A}} V \rightarrow \ell^2(\Lambda) \otimes_{\mathcal{A}} W^*} \leq \varepsilon_{\mathcal{R},M}.
\]  

(5.1.21)
5.1. A General Iterative Method

Possible values are mentioned in Remark 5.1.3; in particular, \( N_M \leq 3M \) for all \( M \in \mathbb{N}_0 \). The bounds \( \bar{\varepsilon}_{R,M} \) can be chosen, for example, as

\[
\bar{\varepsilon}_{R,M} := \sum_{m=M+1}^{\infty} ||R_m||_{V \rightarrow \mathcal{W}^*} . \tag{5.1.22}
\]

Let \( v = (v_\mu)_{\mu \in \Lambda} \in \ell^2(\Lambda) \otimes_3 V \) and \( v|_{\Lambda_p} := (v_\mu)_{\mu \in \Lambda_p} \) for some sets \( \Lambda_p \subset \Lambda, p = 1, \ldots, \ell. \) For any \( M = (M_p)_{p=1}^\ell \in \mathbb{N}_0^\ell \), define

\[
\zeta_M := \sum_{p=1}^{\ell} \bar{\varepsilon}_{R,M_p} \|v_p\|_{\ell^2(\Lambda) \otimes_3 V} \quad \text{and} \quad \sigma_M := \sum_{p=1}^{\ell} N_{M_p}(\# \text{supp } v_p) . \tag{5.1.23}
\]

By (5.1.21), \( \bar{\varepsilon}_{R,M_p} \|v_p\|_{\ell^2(\Lambda) \otimes_3 V} \) is an upper bound for the error resulting from approximating \( \mathcal{R}v|_{\Lambda_p} \) by \( \mathcal{R}_M|_{\Lambda_p} v_p \) in either \( \ell^2(\Lambda) \otimes_3 W^* \) or \( L^2_n(I) \otimes_3 W^* \). The term \( N_{M_p}(\# \text{supp } v_p) \) is a measure for the cost of the computation \( \mathcal{R}_M|_{\Lambda_p} v_p \).

---

Apply\(_\mathcal{R}\)[\(v, \epsilon\)] \(\mapsto \hat{z}\)

\[
\lfloor \ell, (\Lambda_p)_{p=1}^\ell \rfloor \leftarrow \text{BucketSort} \left( \left\{ \|v_p\|_{\mathcal{W}} \right\}_{\mu \in \Lambda_p}, \frac{\epsilon}{2\bar{\varepsilon}_{R,0}} \right)
\]

for \(p = 1, \ldots, P\) do

\(v_p := (v_\mu)_{\mu \in \Lambda_p}\),

Compute the minimal \(\ell \in \{0, 1, \ldots, P\}\) s.t. \(\delta := \bar{\varepsilon}_{R,0} \left\| v - \sum_{p=1}^{\ell} v_p \right\|_{\ell^2(\Lambda) \otimes_3 V} \leq \frac{\epsilon}{2}\).

\(M := (M_p)_{p=1}^\ell := (0)_{p=1}^\ell\)

while \(\zeta_M > \epsilon - \delta\) do

\(M \leftarrow \text{NextOpt}[M]\) with objective \(-\zeta_M\) and cost \(\sigma_M\)

end

\(\hat{z} := (z_\mu)_{\mu \in \Lambda} \leftarrow 0\)

for \(p = 1, \ldots, \ell\) do

forall \(\mu \in \Lambda_p\) do

for \(m = 1, \ldots, M_p\) do

\(w \leftarrow R_m v_\mu\)

\(z_{\mu+\epsilon_m} := z_{\mu} + \beta_m^{m+1} w\)

if \(\mu_m \geq 1\) then \(z_{\mu-\epsilon_m} := z_{\mu} - \epsilon_m + \beta_m^{m+1} w\)

if \(\alpha_m^{m+1} \neq 0\) then \(z_{\mu} := z_{\mu} + \alpha_m^{m+1} w\)

end

end

end

Under Assumption 4.2.A, Apply\(_\mathcal{R}\)[\(v, \epsilon\)] constructs \(\hat{z} \in \ell^2(\Lambda) \otimes_3 W^*\) with

\[
\|\mathcal{R}v - \hat{z}\|_{\ell^2(\Lambda) \otimes_3 W^*} \leq \epsilon \tag{5.1.24}
\]
Chapter 5. Adaptive Iterative Solvers

and

\[ \# \text{supp} \tilde{z} \leq \sum_{j=1}^{\ell} N_{j}(\# \text{supp} v_{[j]}) = \sigma_{M} \]  

(5.1.25)

for any finitely supported \( v \in \ell^{2}(\Lambda) \otimes_{\alpha} V \) and any \( \varepsilon > 0 \), see Proposition 4.2.5 and Remark 4.2.7. The method is optimal in the sense that if \( j \in \mathbb{N}^{\ell}_{0} \) with \( \zeta_{j} \leq \sigma_{M} \), then \( \zeta_{j} \geq \zeta_{M} \), and if \( \zeta_{j} \leq \zeta_{M} \), then \( \sigma_{j} \geq \sigma_{M} \), where \( M \) is the final value of \( M \) in \( \text{Apply}_{R}[v, \varepsilon] \), see Proposition 4.2.6.

Let \( f = (f_{\nu})_{\nu \in \Lambda} \in \ell^{2}(\Lambda) \otimes_{\alpha} W^{*} \). We assume that a routine

\[ \text{RHS}_{f}[\varepsilon] \mapsto \tilde{f} \]  

(5.1.26)

is available to compute approximations \( \tilde{f} = (\tilde{f}_{\nu})_{\nu \in \Lambda} \) of \( f \) with \( \# \text{supp} \tilde{f} < \infty \) and

\[ \| f - \tilde{f} \|_{\ell^{2}(\Lambda) \otimes_{\alpha} W} \leq \varepsilon \]  

(5.1.27)

for any \( \varepsilon > 0 \).

Neglecting errors arising from the inversion of \( D \), the routines \( \text{Apply}_{R} \) and \( \text{RHS}_{f} \) combine to a realization of the abstract iterative solver for

\[ \mathfrak{A} u = f \]  

(5.1.28)

from Section 5.1.1. Details are given in \( \text{SolveDirect}_{R, f} \).

\begin{center}
\textbf{SolveDirect}_{R, f}[\varepsilon, \beta_{0}, \beta_{1}, \gamma] \mapsto u_{\varepsilon}
\end{center}

\begin{verbatim}
\begin{verbatim}
\begin{aligned}
    u_{\varepsilon} &\leftarrow 0 \\
    \delta &\leftarrow (1 - \gamma)^{-1} \| D^{-1} \|_{W^{*} \rightarrow V} \| f \|_{\ell^{2}(\Lambda) \otimes_{\alpha} W} \\
    \text{while} \ \delta > \varepsilon \ \text{do} \\
        \eta &\leftarrow \delta \| D^{-1} \|_{W^{*} \rightarrow V}^{-1} \\
        g_{\nu} &\leftarrow \text{RHS}_{f}[\beta_{0}\eta] - \text{Apply}_{R}[u_{\varepsilon}, \beta_{1}\eta] \\
        u_{\varepsilon} &\leftarrow (u_{\nu})_{\nu \in \Lambda} - D^{-1} g_{\nu} \\
        \delta &\leftarrow (\beta_{0} + \beta_{1} + \gamma)\delta \\
    \text{end}
\end{aligned}
\end{verbatim}
\end{verbatim}

\textbf{Proposition 5.1.4.} For any \( \varepsilon > 0 \), if \( \beta_{0}, \beta_{1} > 0 \) and \( \beta_{0} + \beta_{1} + \gamma < 1 \), then the iteration in \( \text{SolveDirect}_{R, f}[\varepsilon, \beta_{0}, \beta_{1}, \gamma] \) terminates and

\[ \| u - u_{\varepsilon} \|_{\ell^{2}(\Lambda) \otimes_{\alpha} V} \leq \varepsilon . \]  

(5.1.29)

Furthermore, at every iteration, the error in \( \ell^{2}(\Lambda) \otimes_{\alpha} V \) is at most \( \delta \).

\textbf{Proof.} It suffices to show the second assertion. This follows from Theorem 5.1.1 with \( \alpha = 0 \) and \( \beta = \beta_{0} + \beta_{1} \). \hfill \Box

86
5.1. A General Iterative Method

Remark 5.1.5. As in Remark 5.1.2, a posteriori information can be used to improve the estimates $\delta$ of the error in SolveDirect$_{\mathbb{R},I}$. Let $u_0$ and $u_1$ be two successive values of $u_c$ in SolveDirect$_{\mathbb{R},I}$, $g_i = \mathcal{P}u_i$, and let $\delta_0$ be an upper bound for the error of $u_0$. We observe that for the residual $r_0 := \mathcal{P}(u - u_0)$,

$$\|u - u_0\|_{\ell^2(\Lambda) \otimes_3 V} \leq \|D^{-1}q\|^{-1} \|D^{-1}r_0\|_{\ell^2(\Lambda) \otimes_3 V} \leq \frac{1}{1 - \gamma} \|D^{-1}r_0\|_{\ell^2(\Lambda) \otimes_3 V} .$$

(5.1.30)

Furthermore, $D^{-1}r_0$ can be approximated by known quantities since

$$\|u_1 - D^{-1}(f - \mathcal{P}u_0)\|_{\ell^2(\Lambda) \otimes_3 V} \leq \beta \delta_0$$

(5.1.31)

with $\beta := \beta_0 + \beta_1$, and therefore

$$\|D^{-1}r_0\|_{\ell^2(\Lambda) \otimes_3 V} \leq \|u_1 - u_0\|_{\ell^2(\Lambda) \otimes_3 V} + \beta \delta_0 .$$

(5.1.32)

Consequently, we have a refined upper bound for $\|u - u_0\|_{\ell^2(\Lambda) \otimes_3 V}$ given by $\min(\delta_0, \delta_0)$ for

$$\delta_0 := \frac{1}{1 - \gamma} \left(\|u_1 - u_0\|_{\ell^2(\Lambda) \otimes_3 V} + \beta \delta_0\right) .$$

(5.1.33)

If $V$ and $W$ are Hilbert spaces, $\alpha$ is the Hilbert tensor norm, and $D$ is the inner product on $V$, then

$$\|u_1 - u_0\|^2_{\ell^2(\Lambda; V)} = \sum_{\mu \in \Lambda} \langle g_{1,\mu} - g_{0,\mu}, u_{1,\mu} - u_{0,\mu}\rangle_V$$

(5.1.34)

is particularly simple to evaluate.

5.1.3. Uniform Convergence

The iterative method from Section 5.1.2 controls and ensures convergence in $\ell^2(\Lambda) \otimes_3 V$. Since $P = (P_i)_{i \in \Lambda}$ is a tensor product orthonormal polynomial basis of $L^2_n(\Gamma)$, this is equivalent to convergence in $L^2_n(\Gamma) \otimes_3 V$. With minor modifications, convergence is also ensured in $C(\Gamma; V)$, i.e. uniformly in the parameter $y \in \Gamma$.

Proposition 5.1.6. For all $M \in \mathbb{N}_0$,

$$\|\mathcal{R} - \mathcal{R}_M\|_{C(\Gamma; V) \to C(\Gamma; W')} \leq \sum_{m=M+1}^{\infty} \|R_m\|_{V \to W'} .$$

(5.1.35)

If the assumptions of Theorem 4.3.2 are satisfied for an $s > 0$, then this bound is at most $\bar{\delta}_{\mathbb{R},s}(M + 1)^{-s}$.

Proof. By the triangle inequality,

$$\|\mathcal{R} - \mathcal{R}_M\|_{C(\Gamma; V) \to C(\Gamma; W')} \leq \sum_{m=M+1}^{\infty} \|K_m \otimes R_m\|_{C(\Gamma; V) \to C(\Gamma; W')} .$$
Chapter 5. Adaptive Iterative Solvers

Due to Lemma 1.3.3 and (A.2.9), using the tensor product structure Theorem A.3.3,

\[ \|K_m \otimes R_m\|_{C(I;V) \rightarrow C(I;W^*)} = \|K_m\|_{C(I) \rightarrow C(I)} \|R_m\|_{V \rightarrow W^*} = \|R_m\|_{V \rightarrow W^*}, \]

which implies (5.1.35). The second part of the assertion is shown in the proof of Theorem 4.3.2. \(\square\)

The bound in Proposition 5.1.6 on the spaces \(C(I;V)\) and \(C(I;W^*)\) is exactly the same as on \(L^2(I) \otimes \tilde{\mathfrak{a}} V\) and \(L^2(I) \otimes \tilde{\mathfrak{a}} W^*\). For all \(M \in \mathbb{N}_0\), let \(\varepsilon_{\mathfrak{R},M} \geq 0\) be given such that

\[ \|\mathfrak{R} - \mathfrak{R}[M]\|_{C(I;V) \rightarrow C(I;W^*)} \leq \varepsilon_{\mathfrak{R},M}. \quad (5.1.36) \]

Due to (5.1.35), (5.1.22) is a valid choice.

The estimate

\[ \|\mathfrak{R}v - \mathfrak{R}[M]v\|_{C(I;W^*)} \leq \varepsilon_{\mathfrak{R},M} \|v\|_{C(I;V)} \leq \varepsilon_{\mathfrak{R},M} \sum_{\mu \in \Lambda} \|P_\mu\|_{C(I)} \|v_\mu\|_V \quad (5.1.37) \]

introduces an explicit dependence on the polynomial basis \(P = (P_\nu)_{\nu \in \Lambda}\). Due to the product structure of \(I\) and \(P_\mu\),

\[ \|P_\mu\|_{C(I)} = \max_{y \in I} \prod_{m \in \text{supp } \mu} |P^n_{\nu_m}(y_m)| = \prod_{m \in \text{supp } \mu} \max_{\varepsilon \in [-1,1]} |P^n_{\nu_m}(\varepsilon)|, \quad \mu \in \Lambda. \quad (5.1.38) \]

We note that (5.1.37) is independent of the normalization of the polynomials \(P\), i.e. the estimate does not rely on normalization in \(L^2(I)\), nor would a different normalization improve the estimate.

**Example 5.1.7 (Legendre Polynomials).** Legendre polynomials with the normalization from Example 3.2.4 take their maximum on \([-1,1]\) at the right endpoint. It follows from the three term recursion (3.2.6) that \(L_n(1) = \sqrt{2n+1}\) for all \(n \in \mathbb{N}_0\) since, by induction,

\[ \frac{n}{\sqrt{2n+1} \sqrt{2n-1}} L_n(1) = \sqrt{2n-1} - \frac{n-1}{\sqrt{2n-1} \sqrt{2n-3}} \sqrt{2n-3} = \frac{n}{\sqrt{2n-1}}. \]

Therefore, (5.1.38) becomes

\[ \|P_\mu\|_{C(I)} = \prod_{m \in \text{supp } \mu} \sqrt{2^{\mu_m} + 1}, \quad \mu \in \Lambda, \quad (5.1.39) \]

which grows when the support \(\text{supp } \mu\) is increased and when an individual polynomial degree \(\mu_m\) is increased.

**Example 5.1.8 (Chebyshev Polynomials).** The recursion coefficients for Chebyshev polynomials of the first kind are given in Table 3.1. We denote these polynomials by \((T_n)_{n \in \mathbb{N}_0}\). Just like Legendre polynomials, Chebyshev polynomials of the first kind
5.1. A General Iterative Method

take on their maximum at one. Using the three term recursion (3.2.2), one can show inductively that $T_0(1) = 1$ and $T_n(1) = \sqrt{2}$ for all $n \in \mathbb{N}$, since the recursion reduces to

$$T_n(1) = 2 \left( \sqrt{2} - \frac{1}{\sqrt{2}} \right) = 2 \sqrt{2} - \sqrt{2} = \sqrt{2}, \quad n \in \mathbb{N}.$$ 

By (5.1.38), the maximum of the tensor product polynomial $T_\mu$ is

$$\|T_\mu\|_{C(I)} = 2^{(\# \text{ supp } \mu)/2}, \quad \mu \in \Lambda,$$  

which depends only on the size of the support of $\mu$. 

Let $v = (v_\mu)_{\mu \in \Lambda} \in c_0(\Lambda; V)$ and $v_{[p]} := v|_{\Lambda_p} = (v_\mu)_{\mu \in \Lambda_p}$ for some sets $\Lambda_p \subset \Lambda$, $p = 1, \ldots, \ell$. Motivated by (5.1.37), we define

$$\zeta_\mathbf{M} := \sum_{p=1}^\ell \sum_{\mu \in \Lambda_p} \left\|P_\mu\right\|_{C(I)} \|v_\mu\|_V$$ 

for any $\mathbf{M} = (M_p)_{p=1}^\ell \in \mathbb{N}_0^\ell$. With these values, Apply$_R$ from Section 5.1.2 controls the error in $C(I; W^*)$ instead of $L^2(I) \otimes_\pi W^*$.

We consider the operator equation

$$\mathcal{A}u = f,$$  

where $u \in C(I; V)$ and

$$f = \sum_{\nu \in \Lambda} f_\nu P_\nu \in C(I; W^*).$$ 

Let $\hat{f} := (f_\nu)_{\nu \in \Lambda}$. We assume that a routine

$$\text{RHS}_f[e] \mapsto \hat{f}$$ 

is available to compute approximations $\hat{f} = (\hat{f}_\nu)_{\nu \in \Lambda}$ of $\hat{f}$ with $\# \text{ supp } \hat{f} < \infty$ and

$$\|T^W P f - T^W P \hat{f}\|_{C(I; W^*)} \leq \epsilon$$  

for any $\epsilon > 0$.

Equation (5.1.42) can be solved in $C(I; V)$ using the method SolveDirect$_R, \hat{f}$ from Section 5.1.2 with the following modifications. The estimate for the error is initialized as

$$\delta := \frac{1}{1 - \gamma} \left\|D^{-1}\|_{W^* \rightarrow V} \left\|f\right\|_{C(I; W^*)} \right.$$ 

or some upper bound of this quantity. Furthermore, in the call of Apply$_R$, the values $\zeta_\mathbf{M}$ from (5.1.41) are used, and the routine RHS is assumed to satisfy (5.1.45) for $\hat{f}$ from (5.1.44).
Chapter 5. Adaptive Iterative Solvers

**Proposition 5.1.9.** For any $\varepsilon > 0$, if $\beta_0, \beta_1 \geq 0$ and $\beta_0 + \beta_1 + \gamma < 1$, then the routine SolveDirect$_{\mathbb{R}^j} \{\varepsilon, \beta_0, \beta_1, \gamma\}$ with the above modifications computes $u_\varepsilon$ such that

$$\|u - T^V_\mathcal{P} u_\varepsilon\|_{C(\Gamma; V)} \leq \varepsilon. \quad (5.1.47)$$

Furthermore, at every iteration, the error in $C(\Gamma; V)$ is at most $\delta$.

**Proof.** It suffices to show the second assertion. This follows from Theorem 5.1.1 with $\alpha = 0$ and $\beta = \beta_0 + \beta_1$ for the spaces $\mathcal{X} = C(\Gamma; V)$ and $\mathcal{Y} = C(\Gamma; W)$.

**Remark 5.1.10.** As in Remarks 5.1.2 and 5.1.5, a posteriori information can be used to improve the estimates $\delta$ of the error in SolveDirect$_{\mathbb{R}^j}$. Let $u_i = T^V_{\mathcal{P}} u_i$ for two successive values of $u_\varepsilon$ in the modified routine SolveDirect$_{\mathbb{R}^j}$, and let $\delta_0$ be an upper bound for the error of $u_i$ in $C(\Gamma; V)$. By construction,

$$\|u_1 - \mathcal{D}^{-1}(f - \mathcal{A} u_0)\|_{C(\Gamma; V)} \leq \beta \delta_0 \quad (5.1.48)$$

with $\beta := \beta_0 + \beta_1$. Therefore, for the residual $r_0 := \mathcal{A}(u - u_0)$,

$$\|u - u_0\|_{C(\Gamma; V)} \leq \frac{1}{1 - \gamma} \|\mathcal{D}^{-1} r_0\|_{C(\Gamma; V)} \leq \frac{1}{1 - \gamma} \left(\|u_1 - u_0\|_{C(\Gamma; V)} + \beta \delta_0\right). \quad (5.1.49)$$

Consequently, an alternative upper bound for $\|u - u_0\|_{C(\Gamma; V)}$ is given by

$$\delta_0 := \frac{1}{1 - \gamma} \left(\beta \delta_0 + \sum_{\mu \in \Lambda} \|P_\mu\|_{C(\Gamma)} \|u_{1,\mu} - u_{0,\mu}\|_V\right). \quad (5.1.50)$$

Of course, the smaller of the two values $\delta_0$ and $\delta_0$ should be used to estimate the error

$$\|u - u_1\|_{C(\Gamma; V)} \leq (\beta + \gamma) \min(\delta_0, \delta_0) =: \delta_1 \quad (5.1.51)$$

of the following iterate.

### 5.2. Extensions

#### 5.2.1. Alternating Subspace Correction

Let $P = \{P_\mu\}_{\mu \in \Lambda}$ be a tensor product orthonormal polynomial basis of $L^2_\mathcal{A}(\Gamma)$. We assume that $\pi_m$ is symmetric for all $m \in \mathcal{M}$. Then by Remark 3.2.10, $\alpha_m = 0$ for all $n \in \Lambda_m$, and due to Lemma 3.2.9, the operator $\mathcal{R}$ has the form

$$(\mathcal{R} v)_\mu = \sum_{m \in \Lambda} R_m (P_{v_{\mu + \varepsilon_m}}^m + P_{v_{\mu - \varepsilon_m}}^m) \quad (5.2.1)$$

for any $v = \{v_\mu\}_{\mu \in \Lambda} \in L^2(\Lambda) \otimes_\mathcal{A} V$, where $v_\mu := 0$ if $\mu_m < 0$ for any $m \in \mathcal{M}$, see Proposition 3.2.11. If $|v| := \|v\|_{L^2(\Lambda)}$ is even, then all the indices $\mu \in \Lambda$ that appear on the right hand side of (5.2.1) have odd $|\mu|$. Similarly, $|\mu|$ is always even if $|v|$ is odd.
5.2. Extensions

Let \([n] := n + 2\mathbb{Z}\) denote the equivalence class modulo two of \(n \in \mathbb{Z}\), i.e. \([n] = [m]\) if \(n - m\) is even. We define the index sets

\[
\Lambda^{[n]} := \{ \mu \in \Lambda ; \| \mu \| = [n] \} , \quad n \in \mathbb{Z} ,
\]

where \(\| \mu \| = \| \mu \|_{\ell_p(\mathbb{M})}\). Then

\[
\Lambda = \Lambda^{[0]} \sqcup \Lambda^{[1]} .
\]

We call \(\mu \in \Lambda\) even if \(\mu \in \Lambda^{[0]}\) and odd if \(\mu \in \Lambda^{[1]}\).

For \(v \in C(\Gamma)\), let \(\hat{\vartheta}\) be given by \(\hat{\vartheta}(y) := v(-y)\) for \(y \in \Gamma\). We define the projections

\[
\Pi^{[0]} v := \frac{1}{2} (v + \hat{\vartheta}) \quad \text{and} \quad \Pi^{[1]} v := \frac{1}{2} (v - \hat{\vartheta}) .
\]

Clearly, \(\Pi^{[n]} \in \mathcal{L}(C(\Gamma))\) for \(n \in \mathbb{Z}\), and

\[
\| \Pi^{[n]} \|_{\mathcal{L}(C(\Gamma))} \leq 1
\]

since \(\| \hat{\vartheta} \|_{C(\Gamma)} = \| v \|_{C(\Gamma)}\). Moreover, for any \(1 \leq p < \infty\), \(\Pi^{[n]}\) extends uniquely by continuity and density to a continuous linear operator on \(L^p_\mathbb{F}(\Gamma)\) with

\[
\| \Pi^{[0]} \|_{L^p_\mathbb{F}(\Gamma) \to L^p_\mathbb{F}(\Gamma)} \leq 1
\]

since \(\pi\) is invariant under the transformation \(y \mapsto -y\).

For any \(v \in C(\Gamma)\), \(\Pi^{[0]} v\) is an even function and \(\Pi^{[1]} v\) is an odd function. By definition,

\[
\Pi^{[0]} + \Pi^{[1]} = \text{id} .
\]

Consequently,

\[
\Pi^{[0]} \Pi^{[1]} = \Pi^{[1]} \Pi^{[0]} = 0 , \quad \Pi^{[n]} \Pi^{[m]} = \Pi^{[n]} ,
\]

since

\[
\Pi^{[0]} \Pi^{[0]} v = \frac{1}{4} (v + \hat{\vartheta} + \hat{\vartheta} + \hat{\vartheta}) = \frac{1}{4} (v + \hat{\vartheta} + \hat{\vartheta} + v) = \Pi^{[0]} v .
\]

**Lemma 5.2.1.** For all \(v \in L^2_\mathbb{F}(\Gamma)\),

\[
\Pi^{[n]} v = \sum_{\mu \in \Lambda^{[n]}} v_\mu \mu, \quad n \in \mathbb{Z} ,
\]

where \((v_\mu)_{\mu \in \Lambda} = T_F v\).

**Proof.** Due to the symmetry of \(\pi_m\), \(\alpha_{m}^n = 0\) for all \(n \in \mathbb{N}_0\), and it follows by induction using the three term recursion (3.2.2) that \(P^m_n\) is an even function if \(n\) is even and an odd function if \(n\) is odd. Consequently, for all \(\mu \in \Lambda\) and all \(y \in \Gamma\),

\[
P_\mu(-y) = \prod_{\text{mesupp } \mu} P^m_{\mu_m}(-y_m) = \prod_{\text{mesupp } \mu} (-1)^{\mu_m} P^m_{\mu_m}(y_m) = (-1)^{|\mu|} P_\mu(y) .
\]
Chapter 5. Adaptive Iterative Solvers

The function $\hat{v} \in L^2_\pi(\Gamma)$ given by $\hat{v}(y) := v(-y)$ satisfies

$$\hat{v} = \sum_{\mu \in \Lambda} \hat{v}_\mu P_\mu$$

for the coefficients

$$\hat{v}_\mu = \int_\Gamma v(-y) P_\mu(y) \, d\pi(y) = \int_\Gamma v(y) P_\mu(-y) \, d\pi(y) = (-1)^{l_\mu} |\mu| v_\mu$$

since $\pi$ is invariant under the transformation $y \mapsto -y$. The assertion follows using (5.2.4). □

Let $S$ be a Banach space of functions on $\Gamma$ on which $\Pi^{[n]}$ is a continuous linear operator with norm at most one. For example, due to (5.2.5) and (5.2.6), $S$ could be $C(\Gamma)$ or $L^p_\pi(\Gamma)$ for any $1 \leq p < \infty$. Furthermore, let $\alpha$ be a tensor norm.

We consider an operator $\mathcal{A} = D + R$ in $L(S \otimes \bar{\alpha}_V, S \otimes \bar{\alpha}_W^*)$ with $D = \text{id} \otimes D$ boundedly invertible and

$$\|\mathcal{D}^{-1}\|_{S \otimes \bar{\alpha}_V \rightarrow S \otimes \bar{\alpha}_V} \leq \gamma < 1,$$

(5.2.10)

which implies

$$\|\mathcal{A}^{-1}\|_{S \otimes \bar{\alpha}_W \rightarrow S \otimes \bar{\alpha}_V} \leq \frac{1}{1 - \gamma} \|D^{-1}\|_{W^* \rightarrow V}.$$

(5.2.11)

Furthermore, we assume that

$$\Pi^{[n]} R = R \Pi^{[n+1]}, \quad n \in \mathbb{Z}.$$

(5.2.12)

Due to the symmetry of $\pi_m$ for all $m \in \mathcal{M}$, this is satisfied in the setting of Sections 1.2.2 and 1.3.2. Note that, by definition, $\Pi^{[n]} \mathcal{D} = \mathcal{D} \Pi^{[n]}$.

Let $f \in S \otimes \bar{\alpha}_W$. The solution $u \in S \otimes \bar{\alpha}_V$ of

$$\mathcal{A} u = f$$

(5.2.13)

can be decomposed as

$$u = u^{[0]} + u^{[1]}, \quad u^{[0]} = \Pi^{[0]} u.$$

(5.2.14)

We consider a linear iteration similar to that from Section 5.1.1, with corrections computed alternately in range($\Pi^{[0]} \otimes \text{id}_V$) and range($\Pi^{[1]} \otimes \text{id}_V$).

Let $u_0 = u_0^{[0]} = u_0^{[1]} = 0 \in S \otimes \bar{\alpha}_V$, and

$$\delta_0 := \frac{\|D^{-1}\|_{W^* \rightarrow V}}{1 - \gamma} \|f\|_{S \otimes \bar{\alpha}_W}.$$

(5.2.15)

Due to (5.2.11),

$$\|u - u_0\|_{S \otimes \bar{\alpha}_V} = \|u\|_{S \otimes \bar{\alpha}_V} \leq \delta_0.$$

(5.2.16)

Consequently, also

$$\|u^{[n]} - u^{[n]}_0\|_{S \otimes \bar{\alpha}_V} \leq \delta_0 =: \delta^{[n]}_0, \quad n \in \mathbb{Z}.$$

(5.2.17)
5.2. Extensions

For all \( k \in \mathbb{N} \), let \( g_k = g_k^{[0]} + g_k^{[1]} \in S \otimes_\alpha W \) and \( u_k = u_k^{[0]} + u_k^{[1]} \in S \otimes_\alpha V \) with
\[
g_k^{[n]} := (I^{[n]} \otimes \text{id}_{W^*})g_k \quad \text{and} \quad u_k^{[n]} := (I^{[n]} \otimes \text{id}_V)u_k . \tag{5.2.18}
\]
For a fixed \( k \in \mathbb{N} \), let \( \delta_{k-1}^{[1]} \geq 0 \) with
\[
\left\| u^{[1]}_k - u^{[1]}_{k-1} \right\|_{S \otimes_\alpha V} \leq \delta_{k-1}^{[1]} , \tag{5.2.19}
\]
and let \( g_k^{[0]} \) satisfy
\[
\left\| g_k^{[0]} - (f^{[0]} - \mathcal{R} u^{[0]}_{k-1}) \right\|_{S \otimes_\alpha W^*} \leq \beta \delta_{k-1}^{[1]} \left\| D^{-1} \right\|_{W^* \to V} \tag{5.2.20}
\]
for a \( \beta \geq 0 \), where
\[
f^{[n]} := (I^{[n]} \otimes \text{id}_{W^*})f , \quad n \in \mathbb{Z} . \tag{5.2.21}
\]
Then, for a parameter \( \alpha \geq 0 \), let \( u_k^{[0]} \) satisfy
\[
\left\| u^{[0]}_k - \mathcal{R}^{-1} g_k^{[0]} \right\|_{S \otimes_\alpha V} \leq \alpha \delta_{k-1}^{[1]} , \tag{5.2.22}
\]
Similarly, for a \( \delta_{k}^{[0]} \geq 0 \) such that
\[
\left\| u^{[0]}_k - u^{[0]}_{k} \right\|_{S \otimes_\alpha V} \leq \delta_{k}^{[0]} , \tag{5.2.23}
\]
let \( g_k^{[1]} \) and \( u_k^{[1]} \) satisfy
\[
\left\| g_k^{[1]} - (f^{[1]} - \mathcal{R} u_k^{[0]} \right\|_{S \otimes_\alpha W^*} \leq \beta \delta_{k}^{[0]} \left\| D^{-1} \right\|_{W^* \to V} \tag{5.2.24}
\]
and
\[
\left\| u^{[1]}_k - \mathcal{R}^{-1} g_k^{[1]} \right\|_{S \otimes_\alpha V} \leq \alpha \delta_{k}^{[0]} , \tag{5.2.25}
\]

**Theorem 5.2.2.** Let \( u_k = u_k^{[0]} + u_k^{[1]} \in S \otimes_\alpha V \) be as above. Then
\[
\left\| u^{[0]}_k - u^{[0]}_{k} \right\|_{S \otimes_\alpha V} \leq (\alpha + \beta + \gamma) \delta_{k-1}^{[1]} , \tag{5.2.26}
\]
\[
\left\| u^{[1]}_k - u^{[1]}_{k} \right\|_{S \otimes_\alpha V} \leq (\alpha + \beta + \gamma) \delta_{k}^{[0]} . \tag{5.2.27}
\]
If \( \delta_{k}^{[0]} := (\alpha + \beta + \gamma) \delta_{k-1}^{[1]} \) and \( \delta_{k}^{[1]} := (\alpha + \beta + \gamma) \delta_{k}^{[0]} \) for all \( k \in \mathbb{N} \), then
\[
\left\| u^{[n]}_k - u^{[n]}_{k} \right\|_{S \otimes_\alpha V} \leq (\alpha + \beta + \gamma)^{2k} \delta_{0}^{[n]} , \quad n \in \mathbb{Z} , \tag{5.2.28}
\]
and, if \( \alpha + \beta + \gamma < 1 \), then \( u_k \to u \) in \( S \otimes_\alpha V \).
Chapter 5. Adaptive Iterative Solvers

Proof. Since \( \mathcal{D}u^{[0]} = f^{[0]} - \mathcal{R}u^{[1]} \),
\[
\begin{align*}
u^{[0]}_k - u^{[0]}_k &= \mathcal{D}^{-1}(f^{[0]} - \mathcal{R}u^{[1]}) - \mathcal{D}^{-1}(f^{[0]} - \mathcal{R}u^{[0]})_k \\
&+ \mathcal{D}^{-1}(f^{[0]} - \mathcal{R}u^{[0]})_k - \delta_k^{[0]} + \mathcal{D}^{-1} \delta_k^{[0]} - u^{[0]}_k.
\end{align*}
\]
By triangle inequality, using (5.2.20) and (5.2.22),
\[
\begin{align*}
\|u^{[0]}_k - u^{[0]}_k\|_{S^{\beta}_0V} &\leq \|\mathcal{D}^{-1}(\mathcal{R}u^{[1]} - u^{[1]}_k)\|_{S^{\beta}_0V} + \|\mathcal{D}^{-1}\| \|\mathcal{D}^{-1}(f^{[0]} - \mathcal{R}u^{[0]})_k\|_{S^{\beta}_0V} \\
&+ \|u^{[0]}_k - \mathcal{D}^{-1}\delta_k^{[0]}\|_{S^{\beta}_0V} \\
&\leq \gamma \|u^{[1]}_k - u^{[1]}_k\|_{S^{\beta}_0V} + \beta\delta_k^{[1]} + \alpha\delta_k^{[1]}.
\end{align*}
\]
Equation (5.2.26) follows since \( \delta_k^{[1]} \) is an upper bound for \( \|u^{[1]}_k - u^{[1]}_k\|_{S^{\beta}_0V} \). The derivation of (5.2.27) is analogous, using (5.2.24) and (5.2.25).

Comparing Theorems 5.1.1 and 5.2.2, we note that the iteration in the latter converges faster. By triangle inequality and (5.2.15), (5.2.28) implies
\[
\|u - u_k\|_{S^{\beta}_0V} \leq 2(\alpha + \beta + \gamma)^2 \frac{\|D^{-1}\|_{W^{-\infty}V} \|f\|_{S^{\beta}_0W^*}}{1 - \gamma}.
\]

As in Sections 5.1.2 and 5.1.3, we neglect errors arising from the inversion of \( \mathcal{D} \), i.e. we initially consider the idealized case \( \alpha = 0 \). The general case is studied in Section 5.3.

\[
\text{SolveAlternate}(\epsilon, \beta_0, \beta_1, \gamma) \mapsto u_c
\]

\[
\text{for } n = 0, 1 \text{ do}
\]
\[
u_c^{[n]} \leftarrow 0
\]
\[
\delta^{[n]} \leftarrow (1 - \gamma)^{-1} \|D^{-1}\|_{W^{-\infty}V} \|f\|_{S^{\beta}_0W^*}
\]
\end{align*}
\]
\[
\text{end}
\]
\[
n \leftarrow 0
\]
\[
\text{while } \delta^{[0]} + \delta^{[1]} > \epsilon \text{ do}
\]
\[
\eta \leftarrow \delta^{[n-1]} \|D^{-1}\|_{W^{-\infty}V}
\]
\[
\delta^{[n]} = (s_{\mu})_{\mu \in A[n]} \leftarrow \text{RHS}_{\mu \in A[n]}[\beta_0 \eta] - \text{Apply}_{\mu}[u_c^{[n-1]}, \beta_1 \eta]
\]
\[
u_c^{[n]} = (u_{\mu})_{\mu \in A[n]} \leftarrow \text{D}^{-1} \delta^{[n]} = (D^{-1} s_{\mu})_{\mu \in A[n]}
\]
\[
\delta^{[n]} \leftarrow (\beta_0 + \beta_1 + \gamma)\delta^{[n-1]}
\]
\[
n \leftarrow n + 1
\]
\end{align*}
\]
\[
\text{end}
\]
\[
u_c \leftarrow u_c^{[0]} + u_c^{[1]} = (u_{\mu})_{\mu \in A}
\]

If \( S = L^2_0(I) \), progressing to the coefficients with respect to the polynomial basis \( P = (P_\alpha)_{\alpha \in A} \), we arrive at the setting of Section 5.1.2. We assume in particular that \( \mathcal{R} \) has the form (1.3.14), such that its operator matrix representation \( \mathcal{R} \) is as in (5.1.17).
5.2. Extensions

Details of the alternating subspace correction method described above are given in SolveAlternateR alf. This method is very similar to SolveDirectR alf from Section 5.1.2, and uses the same building blocks App1yR alf and RHS alf. The latter is applied separately to the even and odd indices t[u] = (f[µ])µ∈ alf of t, and we assume that the approximations constructed by this routine are supported on the appropriate set alf.

Remark 5.2.3. If V and W are separable Hilbert spaces, and the tensor norm α is the Hilbert tensor norm, then ℓ²(alf; V) and ℓ²(alf; V) are orthogonal subspaces of ℓ²(alf; V). Therefore, the error bound (5.2.29) improves to

\[ \|u - u_k\|_{l^2(alf; V)} = \left( \left\| u^{[0]}_k - u^{[0]}_k \right\|^2_{l^2(alf; V)} + \left\| u^{[1]}_k - u^{[1]}_k \right\|^2_{l^2(alf; V)} \right)^{1/2} \leq \sqrt{2}(\alpha + \beta + \gamma)2^k \frac{\|D^{-1}\|_{W^\rightarrow V} \|f\|_{l^2(alf; W)}}{1 - \gamma}. \]  
(5.2.30)

Consequently, the termination criterion of the loop in SolveAlternateR alf can be replaced by

\[ (\delta^{[0]}_k)^2 + (\delta^{[1]}_k)^2 > \epsilon^2. \]  
(5.2.31)

Remark 5.2.4. As in Section 5.1.3, slight modifications to SolveAlternateR alf ensure convergence in C(alf; V) for arbitrary Banach spaces V and W. Theorem 5.2.2 applies to this space if S = C(alf) and α is the injective tensor norm. The error bounds should be initialized as

\[ \delta^{[n]} := \frac{1}{1 - \gamma} \|D^{-1}\|_{W^\rightarrow V} \|f\|_{C(alf; W)} \]  
(5.2.32)

as in (5.1.46). Also, in the call of App1yR alf, the values ζM from (5.1.41) need to be used, and the routine RHS alf must approximate t[u] in C(alf; W^*).

Remark 5.2.5. The routine SolveAlternateR alf uses separate error bounds δ^{[0]} and δ^{[1]} for the even and odd components of the solution. The improved error bounds from Remark 5.1.5 and Remark 5.1.10 based on estimates of the residual can be used for each of these components separately. However, since these bounds apply to the total error, and not just directly to even or odd part, we expect them to be less useful for SolveAlternateR alf than for SolveDirectR alf.

5.2.2. Coarsening of the Approximate Solution

We consider the setting of Section 3.2.2 for separable Hilbert spaces V and W and the Hilbert tensor norm. If the solution u = (u[µ])µ∈ alf of alf = f is known, approximations can be constructed by restricting u to those indices µ ∈ alf for which \( \|u[µ]\|_V \) is largest. Such approximations are optimal in ℓ²(alf; V) in that they minimize the approximation error for any given support size. It is desirable for numerical approximations to mimic this behavior. For any tolerance ε, the number of nonzero coefficients in the numerical approximation uε should exceed the optimal value only by a constant factor.
Chapter 5. Adaptive Iterative Solvers

We generalize the approximation spaces from Section 4.2.3 to sequences in \( V \). For \( v \in \ell^2(\Lambda; V) \) and \( N \in \mathbb{N}_0 \), let \( P_N(v) \) be a best \( N \)-term approximation of \( v \), that is, \( P_N(v) \) is an element of \( \ell^2(\Lambda; V) \) that minimizes \( \| v - v_N \|_{\ell^2(\Lambda; V)} \) over \( v_N \in \ell^2(\Lambda; V) \) with \# supp \( v_N \leq N \). For \( s \in (0, \infty) \), we define

\[
\| v \|_{A^s(\Lambda; V)} := \sup_{N \in \mathbb{N}_0} (N + 1)^s \| v - P_N(v) \|_{\ell^2(\Lambda; V)}
\]

and

\[
A^s(\Lambda; V) := \left\{ v \in \ell^2(\Lambda; V) \mid \| v \|_{A^s(\Lambda; V)} < \infty \right\}.
\]

By definition, an optimal approximation in \( \ell^2(\Lambda; V) \) of \( v \in A^s(\Lambda; V) \) with error tolerance \( \varepsilon > 0 \) consists of \( O(\varepsilon^{-1/s}) \) nonzero coefficients in \( V \).

**Proposition 5.2.6.** Let \( v \in A^s(\Lambda; V) \) for an \( s \in (0, \infty) \), and let \( w \in \ell^2(\Lambda; V) \) with

\[
\| v - w \|_{\ell^2(\Lambda; V)} \leq \varepsilon
\]

for an \( \varepsilon > 0 \). Furthermore, let \( N \in \mathbb{N}_0 \) be minimal with

\[
\| w - w_N \|_{\ell^2(\Lambda; V)} \leq 4\varepsilon
\]

for \( w_N := P_N(w) \). Then

\[
\| v - w_N \|_{\ell^2(\Lambda; V)} \leq 5\varepsilon, \quad (5.2.37)
\]

\[
\| v - w_N \|_{\ell^2(\Lambda; V)} \leq C \| v \|_{A^s(\Lambda; V)} N^{-s}, \quad (5.2.38)
\]

and

\[
\| w_N \|_{A^s(\Lambda; V)} \leq C \| v \|_{A^s(\Lambda; V)} \quad (5.2.39)
\]

with a constant \( C \) depending only on \( s \).

**Proof.** Equations \((5.2.37)\) and \((5.2.38)\) are a straightforward generalization of [CDD01, Corollary 5.2] to \( \ell^2(\Lambda; V) \). To show \((5.2.39)\), we note that for all \( n \leq N - 1 \),

\[
\| w_N - P_n(w_N) \|_{\ell^2(\Lambda; V)} \leq \| w_N - P_n(v) \|_{\ell^2(\Lambda; V)} \leq \| v - w_N \|_{\ell^2(\Lambda; V)} + \| v - P_n(v) \|_{\ell^2(\Lambda; V)}
\]

\[
\leq \| v - w_N \|_{\ell^2(\Lambda; V)} + (n + 1)^{-s} \| v \|_{A^s(\Lambda; V)},
\]

and the assertion follows using \((5.2.38)\).

\[ \square \]

Let the routine

\[
\text{Coarsen}[w, \eta] \mapsto w_N
\]

truncate \( w \in \ell^2(\Lambda; V) \) to \( P_N(w) \) for the minimal \( N \in \mathbb{N}_0 \) with

\[
\| w - w_N \|_{\ell^2(\Lambda; V)} \leq \eta.
\]
5.2. Extensions

\[
\text{SolveCoarse}_{\mathbb{R},f}[\epsilon, \beta_0, \beta_1, \gamma, \delta, \chi] \mapsto u_e
\]

\[
u_e \leftarrow 0
\]

\[
\delta \leftarrow (1 - \gamma) - D^{-1} \|g\|_{W^* \rightarrow V}\|f\|_{C(V,W^*)}
\]

while \(\delta > \epsilon\) do

\[
\delta \leftarrow \delta
\]

while \(\delta > \chi \delta\) do

\[
\zeta \leftarrow \delta - D^{-1} \|g\|_{W^* \rightarrow V}
\]

\[
g = (g_\mu)_{\mu \in \Lambda} \leftarrow \text{RHS}_{\mathbb{R}}[u_e, \beta_0 \zeta] - \text{Apply}_{\mathbb{R}}[u_e, \beta_1 \zeta]
\]

\[
u_e = (u_\mu)_{\mu \in \Lambda} \leftarrow D^{-1} g = (D^{-1} g_\mu)_{\mu \in \Lambda}
\]

\[
\delta \leftarrow (\beta_0 + \beta_1 + \gamma) \delta
\]

end

\[
u_e \leftarrow \text{Coarsen}[u_e, (1 - \chi) \delta \delta]
\]

\[
\delta \leftarrow \delta
\]

end

This can be realized by sorting the coefficients of \(w\). More generally, it would suffice for \(N\) to be minimal only up to a constant factor, in which case the approximate sorting routine \(\text{BucketSort}\) can be employed to achieve optimal complexity.

Due to Proposition 5.2.6, using \(\text{Coarsen}[u_e, \eta]\) in iterations such as \(\text{SolveDirect}_{\mathbb{R},f}\) and \(\text{SolveAlternate}_{\mathbb{R},f}\) can ensure optimal \#supp\(u_e\) compared to the error tolerance. Details of \(\text{SolveDirect}_{\mathbb{R},f}\) with a coarsening step are given in \(\text{SolveCoarse}_{\mathbb{R},f}\).

**Proposition 5.2.7.** For any \(\epsilon > 0\), if \(\beta_0, \beta_1 \geq 0, \beta_0 + \beta_1 + \gamma < 1, 0 < \delta < 1\), and \(0 < \chi \leq 1\), then the iteration in \(\text{SolveCoarse}_{\mathbb{R},f}[\epsilon, \beta_0, \beta_1, \gamma, \delta, \chi]\) terminates and

\[
\|u - u_e\|_{C(\Lambda; V)} \leq \epsilon.
\] (5.2.42)

Furthermore, at the end of every outer iteration, the error in \(\ell^2(\Lambda; V)\) is at most \(\tilde{\delta}\), and for \(\chi = 1/5\),

\[
\#\text{supp } u_e \leq C \|u\|_{\mathcal{A}^s(\Lambda; V)}^{1/s} \tilde{\delta}^{-1/s}
\] (5.2.43)

\[
\|u - u_e\|_{\ell^2(\Lambda; V)} \leq C \|u\|_{\mathcal{A}^s(\Lambda; V)}
\] (5.2.44)

if \(u \in \mathcal{A}^{s}(\Lambda; V)\), for a constant \(C\) depending only on \(s\).

**Proof.** The inner iteration of \(\text{SolveCoarse}_{\mathbb{R},f}\) is a copy of \(\text{SolveDirect}_{\mathbb{R},f}\) with different initialization. Proposition 5.1.4 implies that \(\delta\) is an upper bound for \(\|u - u_e\|_{\ell^2(\Lambda; V)}\) in the inner iteration, and thus the last iterate has an error of at most \(\chi \delta\). By construction, the coarsened approximation \(u_e\) satisfies

\[
\|u - u_e\|_{\ell^2(\Lambda; V)} \leq \delta\delta.
\]

If \(\chi = 1/5\), Proposition 5.2.6 with \(v = u, w = u_e\) and \(\epsilon = \delta\delta/5\) implies (5.2.43) and (5.2.44). \(\square\)
Chapter 5. Adaptive Iterative Solvers

In SolveCoarse$_{\delta,\ell}$, the a priori definitions of the error bounds $\delta$ and $\delta'$ can be improved by using the subsequently computed approximation of the residual as in Remark 5.1.5.

**Remark 5.2.8.** The operator $\mathcal{R}$ is $s'$-compressible with sparse approximations $\mathcal{R}[M]$, e.g. if $(\|R_m\|_{V \to W})_{m \in \mathbb{N}}$ is in $l^{1/(s'+1)}$ for all $s \in (0, s')$, see Remark 5.1.3. In this case, $\mathcal{R}$ is a bounded linear map from $\mathcal{A}^s(\Lambda; V)$ to $\mathcal{A}^s(\Lambda; W)$ for all $s \in (0, s')$. This carries over to the routine App1$_{\mathcal{R}}$ in that if $v \in \mathcal{A}^s(\Lambda; V)$ and $\tilde{\mathbf{z}}$ is the output of App1$_{\mathcal{R}}[v, \epsilon]$ for an $\epsilon > 0$, then

\[
\# \text{supp } \tilde{\mathbf{z}} \leq \|v\|_{\mathcal{A}^s(\Lambda; V)}^{1/s} \epsilon^{-1/s},
\]

\[
\|\mathbf{z}\|_{\mathcal{A}^s(\Lambda; W)} \leq \|v\|_{\mathcal{A}^s(\Lambda; V)},
\]

with constants depending only on $s$ and $\mathcal{R}$. Moreover, (5.2.45) is an upper bound for the total number of applications of operators $R_m$ in App1$_{\mathcal{R}}[v, \epsilon]$. This follows as in the scalar case, see Theorem 4.2.9 and Corollary 4.2.10.

We assume similar properties for the routine RHS$_{\delta}$. If $f \in \mathcal{A}^s(\Lambda; W^*)$ and $\bar{f}$ is the output of RHS$_{\delta}[\epsilon]$ for an $\epsilon > 0$, then $\bar{f}$ should satisfy

\[
\# \text{supp } \bar{f} \leq \|f\|_{\mathcal{A}^s(\Lambda; W^*)}^{1/s} \epsilon^{-1/s}.
\]

It follows as in Corollary 4.2.10 that

\[
\|\bar{f}\|_{\mathcal{A}^s(\Lambda; W^*)} \leq \|f\|_{\mathcal{A}^s(\Lambda; W^*)}.
\]

Note that if $u \in \mathcal{A}^s(\Lambda; V)$ and $\mathcal{R}$ is $s'$-compressible with $s < s'$, then also $\mathcal{R}$ is $s'$-compressible, and therefore $\|\bar{u}\|_{\mathcal{A}^s(\Lambda; V)} \leq \|u\|_{\mathcal{A}^s(\Lambda; V)}$.

**Proposition 5.2.9.** Let $\beta_0, \beta_1 > 0$, $\beta_0 + \beta_1 + \gamma < 1$, $\beta < 1$, $\delta < 1$, and $0 < \chi \leq 1$. If $\mathcal{R}$ is $s'$-compressible and $s \in (0, s')$ such that $u \in \mathcal{A}^s(\Lambda; V)$ and (5.2.47) holds, then at the end of each iteration of the inner loop of SolveCoarse$_{\delta,\ell}$, $\mathcal{R}$, $\|u - u_\ell\|_{\mathcal{A}^s(\Lambda; V)} \leq \delta$, $\# \text{supp } u_\ell \leq C \|u\|_{\mathcal{A}^s(\Lambda; V)}^{1/s} \delta^{-1/s}$, $\|u_\ell\|_{\mathcal{A}^s(\Lambda; V)} \leq C \|u\|_{\mathcal{A}^s(\Lambda; V)}$, and $\|u_\ell\|_{\mathcal{A}^s(\Lambda; V)} \leq C \|u\|_{\mathcal{A}^s(\Lambda; V)}$.

for a constant $C$ depending only on $s$, $\mathcal{R}$, $D$, $\beta_0$, $\beta_1$, $\gamma$, $\delta$ and $\chi$.

**Proof.** Equation (5.2.49) follows from Proposition 5.1.4 since, apart from the initialization, the inner loop of SolveCoarse$_{\delta,\ell}$ is identical to SolveDirect$_{\delta,\ell}$. We show (5.2.50) and (5.2.51) by induction.

The value of $u_\ell$ before the inner loop is either zero, or the output of Coarsen. Proposition 5.2.7 implies that (5.2.51) holds for these initial values.

Let $u_\ell$ be the approximation at the end of some iteration of the inner loop, and let $u_{\ell}'$ denote the previous approximation. By induction,

\[
\|u_{\ell}'\|_{\mathcal{A}^s(\Lambda; V)} \leq \|u\|_{\mathcal{A}^s(\Lambda; V)}.
\]
5.2. Extensions

Let \( g_0 := \text{RHS}[\beta_0 \zeta] \) and \( g_1 := \text{Apply}_{g}[u_{c}, \beta_1 \zeta] \). By (5.2.47), using \( \|u\|_{\mathcal{A}^\prime(A ; V)} \leq \|u\|_{\mathcal{A}^\prime(A ; V)} \),

\[
\# \text{supp} \; g_0 \leq \|u\|_{\mathcal{A}^\prime(A ; V)}^{1/s} (\beta_0 \zeta)^{-1/s} .
\]

Furthermore, using (5.2.45) and (5.2.51) for \( u_{c} \),

\[
\# \text{supp} \; g_1 \leq \|u_{c}\|_{\mathcal{A}^\prime(A ; V)}^{1/s} (\beta_1 \zeta)^{1/s} \leq \|u\|_{\mathcal{A}^\prime(A ; V)}^{1/s} (\beta_1 \zeta)^{1/s} .
\]

Since \( g = g_0 - g_1 \) and \( u_{c} = D^{-1} g_0 \),

\[
\# \text{supp} \; u_{c} = \# \text{supp} \; g \leq \|u\|_{\mathcal{A}^\prime(A ; V)}^{1/s} \delta^{-1/s}
\]

for the final value of \( \delta \).

By (5.2.48), \( \|g_0\|_{\mathcal{A}^\prime(A ; W)} \leq \|u\|_{\mathcal{A}^\prime(A ; V)} \). Similarly, (5.2.46) implies

\[
\|g_1\|_{\mathcal{A}^\prime(A ; W)} \leq \|u_{c}\|_{\mathcal{A}^\prime(A ; V)} \leq \|u\|_{\mathcal{A}^\prime(A ; V)} .
\]

Consequently, by the weakened triangle inequality,

\[
\|u_{c}\|_{\mathcal{A}^\prime(A ; V)} \leq \|g_0\|_{\mathcal{A}^\prime(A ; W)} \leq \|g_0\|_{\mathcal{A}^\prime(A ; W)} + \|g_1\|_{\mathcal{A}^\prime(A ; W)} \leq \|u\|_{\mathcal{A}^\prime(A ; V)} .
\]

This shows that (5.2.50) and (5.2.51) hold separately at every iteration of the inner loop. However, at each step of the iteration, the constants in these estimates may get larger. Since the loop terminates once an error reduction of \( \chi \delta \) is reached, and the error is reduced by a factor of \( \beta_0 + \beta_1 + \gamma \) at every iteration, the loop terminated after at most

\[
\left\lceil \frac{\log(\chi \delta)}{\log(\beta_0 + \beta_1 + \gamma)} \right\rceil
\]

iterations. The constants are reset for the next inner loop by the call of Coarsen. \( \Box \)

**Theorem 5.2.10.** Let \( \beta_0, \beta_1 > 0, \beta_0 + \beta_1 + \gamma < 1, 0 < \delta < 1, \) and \( \chi = 1/5 \). If \( R \) is \( s^* \)-compressible and \( s \in (0, s^*) \) such that \( u \in \mathcal{A}^s(A ; V) \) and (5.2.47) holds, then for any \( \varepsilon > 0 \), the total number of evaluations of \( D^{-1} \) and \( R_m, m \in \mathbb{N} \), in SolveCoarse\(_{a,t}\)(\( \varepsilon, \beta_0, \beta_1, \gamma, \delta, \chi \)) is bounded by \( \|u\|_{\mathcal{A}^\prime(A ; V)}^{1/s} e^{-1/s} \) up to a constant factor depending only on \( s, R, D, \beta_0, \beta_1, \delta \) and \( \gamma \).

**Proof.** The number of calls of \( D^{-1} \) in SolveCoarse\(_{a,t}\) is equal to the sum of \( \# \text{supp} \; u_{c} \) over all values of \( u_{c} \) in the inner loop of SolveCoarse\(_{a,t}\). We consider an inner loop with initial error bound \( \delta \) and target accuracy \( \delta \delta / 5 \). Due to (5.2.50), the number of calls \( N_\delta \) of \( D^{-1} \) in this loop is no more than

\[
N_\delta \leq C \|u\|_{\mathcal{A}^\prime(A ; V)}^{1/s} \delta^{-1/s} \sum_{k=1}^{K} (\beta_0 + \beta_1 + \gamma)^{-k/s} \leq \|u\|_{\mathcal{A}^\prime(A ; V)}^{1/s} \delta^{-1/s} (\beta_0 + \beta_1 + \gamma)^{-(K+1)/s}
\]

for \( K := \lceil \log(\delta / 5) / \log(\beta_0 + \beta_1 + \gamma) \rceil \). Since \( K \) is minimal with \( (\beta_0 + \beta_1 + \gamma)^K \leq \delta / 5 \), this term is bounded by

\[
N_\delta \leq \|u\|_{\mathcal{A}^\prime(A ; V)}^{1/s} (\delta \delta / 5)^{-1/s} .
\]
Chapter 5. Adaptive Iterative Solvers

This loop is called with $\delta = \delta^j \delta_0$ for all $j \in \mathbb{N}_0$ with $\delta^j \delta_0 > \epsilon$. Consequently, the total number $N$ of calls of $D^{-1}$ is no more than

$$N \leq \sum_{j=0}^{J} N \delta^j \delta_0 \leq \|u\|_{L^s(V;V)}^{1/s} \sum_{j=0}^{J} (\delta^{j+1} \delta_0/5)^{-1/s} \leq \|u\|_{L^s(V;V)}^{1/s} (\delta^{J+1} \delta_0/5)^{-1/s}$$

with $J := \lceil \log(\epsilon/\delta_0)/ \log(\delta) \rceil - 1$. Since $J$ is minimal with $\delta^{J+1} \leq \epsilon/\delta_0$,

$$N \leq \|u\|_{L^s(V;V)}^{1/s} (\epsilon/5)^{-1/s}.$$

By Remark 5.2.8, (5.2.50) is an upper bound for the total number of evaluations of operators $R_m$, $m \in \mathbb{N}$, in the call of $\text{Apply}_{R}$. Therefore, the above bound holds also for the total number of such operations in $\text{SolveCoarse}_{R}$. \hfill \Box

5.3. Discretization of Deterministic Problems

5.3.1. Single Level Discretization

Sections 5.1 and 5.2 focus on approximations only in the parameter domain $\Gamma$, with exact computations in $V$ and $W$. However, they also apply to fully discrete equations if the discretization of $V$ and $W$ is performed first, at which point these spaces are replaced by finite dimensional approximations, and all subsequent manipulations can indeed be assumed to be exact. Of course, the approximate solutions $u_e$ computed by the resulting method $\text{SolveDirect}_{R,F}$ and its extensions do not converge to the solution of the original problem. Rather, the error tolerance $\epsilon$ only applies to the difference between $u_e$ and the exact solution after discretization of $V$ and $W$.

A simple abstract formalization of the discretization of $V$ and $W$ is to replace $D^{-1}$ by some approximation $S \in \mathcal{L}(W^*,V)$, which can be thought of as having finite dimensional range. Then replacing $f$ by $Sf$, $D$ by the identity and $R$ by $SR$ leads to a parametric operator equation on the range of $S$.

The adaptive methods from Sections 5.1 and 5.2 iteratively approximate the solution $u_S$ of

$$u_S = (\text{id} \otimes S)(f - R u_S). \quad (5.3.1)$$

Under the assumption

$$\sum_{m=1}^{\infty} \|SR_m\|_{V \rightarrow V} \leq \gamma < 1, \quad (5.3.2)$$

a Neumann series argument shows that $u_S$ is well defined for example in $L^2_{\alpha}(\Gamma) \otimes_{\alpha} V$ or $C(\Gamma; V)$ if $f$ is in $L^2_{\alpha}(\Gamma) \otimes_{\alpha} W^*$ or $C(\Gamma; W^*)$, respectively, where $\alpha$ is any tensor norm.

The remaining error $u - u_S$ can be estimated independently of $\mathcal{R}$, given sufficient regularity of $u$. By definition,

$$(\text{id} +(\text{id} \otimes S)\mathcal{R})(u - u_S) = u - (\text{id} \otimes S)\mathcal{R} u. \quad (5.3.3)$$
5.3. Discretization of Deterministic Problems

Due to (5.3.2), the operator on the left of (5.3.3) can be inverted by a Neumann series in $C(\Gamma; V)$, implying
\[\|u - u_S\|_{C(\Gamma; V)} \leq \frac{1}{1 - \gamma} \max_{y \in \Gamma} \|u(y) - SDu(y)\|_V,\] (5.3.4)
and in $L^2_\pi(\Gamma) \otimes \bar{\alpha} V$, leading to
\[\|u - u_S\|_{L^2_\pi(\Gamma) \otimes \bar{\alpha} V} \leq \left(\frac{1}{1 - \gamma} \sum_{\mu \in \Lambda} \|u_\mu - SDu_\mu\|_V^2\right)^{1/2}.\] (5.3.5)

In particular, if $V$ is a separable Hilbert space, (5.3.5) applies to $\|u - u_S\|_{L^2_\pi(\Gamma; V)}$.

**Remark 5.3.1.** The above approach generalizes to multiple level discretizations in which a different approximation $S_\mu$ of $D^{-1}$ is chosen for each $\mu \in \Lambda$. We consider a different method based on adaptive solvers in Section 5.3.2.

5.3.2. Adaptive Solvers

Using a single approximation of $D^{-1}$ for all $\mu \in \Lambda$, though simple, is inefficient. The coefficients $u_\mu \in V$ vary in size, and thus finer discretizations should be chosen for $\mu \in \Lambda$ with larger $u_\mu$. This is achieved by specifying not a single discretization, as in Section 5.3.1, but a single absolute tolerance for all indices $\mu$.

We assume a solver for $D$ is available such that for any $g \in W^*$ and any $\epsilon > 0$,
\[\text{Solve}_D[g, \epsilon] \mapsto v, \quad \|v - D^{-1}g\|_V \leq \epsilon.\] (5.3.6)

For example, $\text{Solve}_D$ could be an adaptive wavelet method, see e.g. [CDD01, CDD02, GHS07], an adaptive frame method, see e.g. [Ste03, DFR07, DRW+07], or a finite element method with a posteriori error estimation, see e.g. [Dör96, MNS00, BDD04].

**Remark 5.3.2.** Suppose we wish to compute $v := D^{-1}g$ up to a tolerance $\eta$ in $\ell^2(\Lambda) \otimes_{\alpha} V$ for a finitely supported $g \in \ell^2(\Lambda) \otimes_{\alpha} W^*$. We approximate $v$ by $\bar{v} = (v_\mu)_{\mu \in \Lambda}$ with $\bar{v}_\mu = 0$ if $\mu \notin \text{supp } g$ and
\[\bar{v}_\mu := \text{Solve}_D[g_\mu, \zeta], \quad \mu \in \text{supp } g,\] (5.3.7)
for a $\zeta > 0$ to be determined. If $V$ is a separable Hilbert space and $\alpha$ is the Hilbert tensor product, then
\[\|v - \bar{v}\|_{\ell^2(\Lambda; V)} = \left(\sum_{\mu \in \text{supp } g} \|v_\mu - \bar{v}_\mu\|_V^2\right)^{1/2} \leq \zeta \sqrt{\# \text{supp } g},\] (5.3.8)
and therefore we choose
\[\zeta := \eta(\# \text{supp } g)^{-1/2}\] (5.3.9)
Chapter 5. Adaptive Iterative Solvers

to ensure a total absolute error of \( \eta \). More generally, if \( V \) is a Banach space and \( \alpha \) is any tensor norm, then by triangle inequality,

\[
\| v - \bar{v} \|_{\mathcal{E}(\Lambda; V)} \leq \sum_{\mu \in \text{supp } g} \| v_\mu - \bar{v}_\mu \|_V \leq \zeta \# \text{supp } g ,
\]

(5.3.10)

and accordingly we choose

\[
\zeta := \eta (\# \text{supp } g)^{-1}
\]

(5.3.11)

to reach the target accuracy \( \eta \).

If the application of \( D^{-1} \) is replaced by calls of \( \text{Solve}_D \) with equidistributed error tolerances as in Remark 5.3.2, the adaptive methods from Sections 5.1 and 5.2 become computationally accessible. We illustrate this for \( \text{SolveDirect}_{\|,f} \) in the case that \( V \) and \( W \) are separable Hilbert spaces.

\[
\text{SolveDirect}_{\|,f}[\varepsilon,\alpha,\beta_0,\beta_1,\gamma] \mapsto u_\varepsilon
\]

\[
\begin{align*}
\delta & \leftarrow (1 - \gamma)^{-1} \| D^{-1} \|_{W^* \rightarrow V} \| f \|_{\mathcal{E}(\Lambda; W^*)} \\
\text{while } \delta > \varepsilon \text{ do} & \\
\eta & \leftarrow \delta \| D^{-1} \|_{W^* \rightarrow V}^{-1} \\
g & \leftarrow \text{RHS}[\beta_0 \eta] - \text{Apply}_{\|}[u_\varepsilon] \\
\zeta & \leftarrow \alpha \delta (\# \text{supp } g)^{-1/2} \\
u_\varepsilon & \leftarrow (\bar{u}_\mu)_{\mu \in \Lambda} - (\text{Solve}_D[g_\mu, \zeta])_{\mu \in \Lambda} \\
\delta & \leftarrow (\alpha + \beta_0 + \beta_1 + \gamma) \delta
\end{align*}
\]

end

\textbf{Proposition 5.3.3.} Let \( V \) and \( W \) be separable Hilbert spaces. For any \( \varepsilon > 0 \), if \( \alpha, \beta_0, \beta_1 > 0 \) and \( \alpha + \beta_0 + \beta_1 + \gamma < 1 \), then the iteration in \( \text{SolveDirect}_{\|,f}[\varepsilon,\alpha,\beta_0,\beta_1,\gamma] \) terminates, and

\[
\| u - u_\varepsilon \|_{\mathcal{E}(\Lambda; V)} \leq \varepsilon .
\]

(5.3.12)

Furthermore, at every iteration, the error in \( \ell^2(\Lambda; V) \) is at most \( \delta \).

\textbf{Proof.} It suffices to show the second assertion. This follows from Remark 5.3.2 and Theorem 5.1.1 with \( \beta = \beta_0 + \beta_1 \).

If \( V \) and \( W \) are not separable Hilbert spaces, then by Remark 5.3.2, the definition of \( \zeta \) in \( \text{SolveDirect}_{\|,f} \) should be changed to \( \zeta \leftarrow \alpha \delta (\# \text{supp } g)^{-1} \).

\textbf{Remark 5.3.4.} The fully discrete routine \( \text{SolveDirect}_{\|,f} \) can be altered to ensure uniform convergence. In addition to the modifications from Section 5.1.3, this requires a different distribution of the tolerances for \( \text{Solve}_D \). If \( g \in \ell^2(\Lambda; W^*) \) is finitely supported.
5.3. Discretization of Deterministic Problems

and \( \tilde{v} = (\tilde{v}_\mu)_{\mu \in \Lambda} \) is an approximation to \( v := D^{-1}g \) with \( \tilde{v}_\mu = 0 \) for all \( \mu \in \Lambda \setminus \text{supp} \, g \), then by triangle inequality,

\[
\|v - \tilde{v}\|_{C(\Gamma; V)} \leq \sum_{\mu \in \text{supp} \, g} \|v_\mu - \tilde{v}_\mu\|_V \|P_\mu\|_{C(\Gamma)},
\]

(5.3.13)

where \( P = (P_\mu)_{\mu \in \Lambda} \) is a tensor product orthonormal polynomial basis of \( L^2_\Gamma \). Therefore, in order to ensure an error tolerance of \( \eta \) in \( v \), we define

\[
\zeta := \eta(\# \text{supp} \, g)^{-1}
\]

(5.3.14)

and compute \( \tilde{v} \) by

\[
\tilde{v}_\mu := \text{Solve}_D[g_\mu, \zeta, \tilde{u}_\mu], \quad \mu \in \text{supp} \, g.
\]

(5.3.15)

Note that this distribution of tolerances is independent of the scaling of the polynomials \( P_\mu \). It can be used in \( \text{SolveDirect}_{\tilde{u}, \zeta} \), in which case \( \zeta = \alpha \delta \).

\[ \text{Remark 5.3.5.} \] If \( \text{Solve}_D \) is an iterative solver, as all of the above examples are, then the previous approximation to \( u_\mu \) can be used as an initial guess. Adding the initial guess as an explicit argument, the call of \( \text{Solve}_D \) in \( \text{SolveDirect}_{\tilde{u}, \zeta} \) becomes

\[
u_\epsilon = (\tilde{u}_\mu)_{\mu \in \Lambda} \leftarrow (\text{Solve}_D[g_\mu, \zeta, \tilde{u}_\mu])_{\mu \in \Lambda}.
\]

Alternatively, a coarsened version of the previous approximation \( \tilde{u}_\mu \) can be used to prevent superfluous refinements from propagating to the next iteration.

\[ \text{Remark 5.3.6.} \] As in Remarks 5.1.2 and 5.1.5, a posteriori information can be used to improve the estimates \( \delta \) of the error in \( \text{SolveDirect}_{\tilde{u}, \zeta} \). Due to the additional source of error, the bound in (5.1.31) becomes \((\alpha + \beta)\delta_0 \), with \( \beta = \beta_0 + \beta_1 \), and consequently \( \min(\delta_0, \bar{\delta}_0) \) is an upper bound for \( \|u - u_0\|_{C(\Lambda; V)} \), where

\[
\bar{\delta}_0 := \frac{1}{1 - \gamma} \left(\|u_1 - u_0\|_{C(\Lambda; V)} + (\alpha + \beta)\delta_0 \right),
\]

(5.3.16)

and \( u_0 \) and \( u_1 \) are two consecutive values of \( u_\epsilon \) in \( \text{SolveDirect}_{\tilde{u}, \zeta} \). Equation (5.1.34) no longer holds exactly, but may serve as a useful approximation.

\[ \text{Remark 5.3.7.} \] The methods \( \text{SolveAlternate}_{\tilde{u}, \zeta} \) and \( \text{SolveCoarse}_{\tilde{u}, \zeta} \) can be extended analogously to use an adaptive solver \( \text{Solve}_D \) rather than the exact inverse of \( D \).

5.3.3. Multiple Level Coarsening

In principle, the routine \( \text{Coarsen} \) from Section 5.2.2 can be used in combination with an adaptive solver \( \text{Solve}_D \) to construct a fully adaptive version of \( \text{SolveCoarse}_{\tilde{u}, \zeta} \). However, this only truncates the set of active indices, and thereby forgoes the possibility to coarsen individual indices. In the following, let \( V \) and \( W \) be separable Hilbert spaces.
Chapter 5. Adaptive Iterative Solvers

We assume that, in addition to \( v \), the method \( \text{Solve}_D[g, \epsilon] \) from (5.3.6) also constructs a sequence \( (v^j)_{j=0}^n \) of coarser approximations to \( v \). These can be computed after \( v \), as projections of \( v \) onto a nested sequence of subspaces of \( V \), or during the solution process, as intermediate approximations of \( D^{-1}g \) in an adaptive iterative method. To each \( v^j \) is associated an error \( \|v - v^j\|_V \) and a cost \( c_j \), which represents the size of \( v^j \), e.g. \( c_j \) may be the number of basis functions used to represent \( v^j \). We assume that \( v^0 \) is always \( 0 \in V \), with \( c_0 = 0 \).

With this additional information, the coarsening step of \( \text{SolveCoarse}_D[g_\mu, \epsilon] \) should not simply restrict \( \Lambda_\epsilon = (\tilde{u}_\mu)_{\mu \in \Lambda} \) to a subset of \( \Lambda \); rather, it should construct a vector \( j = (j_\mu)_{\mu \in \text{supp} u_\mu} \) describing a coarsened approximation

\[
u_j = (\tilde{u}_\mu^{j_\mu})_{\mu \in \Lambda}
\]

which minimizes the total cost

\[c_j = \sum_{\mu \in \text{supp} u_\mu} c_\mu^{j_\mu}
\]

under the conditions that the total error

\[
\|u_\epsilon - u_j\|_{C(\Lambda; V)}^2 = \sum_{\mu \in \text{supp} u_\epsilon} \|\tilde{u}_\mu - \tilde{u}_\mu^{j_\mu}\|_V^2
\]

satisfies a predefined bound. Here, \( \tilde{u}_\mu^{j_\mu} \) are preliminary approximations of \( D^{-1}g_\mu \) from the previous call of \( \text{Solve}_D[g_\mu, \epsilon] \), and \( c_\mu^{j_\mu} \) is the cost of \( \tilde{u}_\mu^{j_\mu} \).

This optimization problem has the structure considered in Section 4.1. A solution can be computed efficiently by the recursive calls of the simple method \( \text{NextOpt} \) under the assumption that for each \( \mu \), the sequence of costs \( (c_\mu^{j_\mu})_{j=0}^n \) is strictly increasing, the errors \( \|\tilde{u}_\mu - \tilde{u}_\mu^{j_\mu}\|_V \) are decreasing in \( j \), and

\[
\frac{\|\tilde{u}_\mu - \tilde{u}_\mu^{j_\mu}\|_V^2 - \|\tilde{u}_\mu - \tilde{u}_\mu^{j_{i+1}}\|_V^2}{c_\mu^{j_{i+1}} - c_\mu^{j_i}} \geq \frac{\|\tilde{u}_\mu - \tilde{u}_\mu^{j_\mu}\|_V^2 - \|\tilde{u}_\mu - \tilde{u}_\mu^{j_{i+1}}\|_V^2}{c_\mu^{j_{i+1}} - c_\mu^{j_i}} \quad \text{if} \quad i \geq j,
\]

see Proposition 4.1.6. The technical condition (5.3.20) can easily be enforced by passing to a subsequence of \( (\tilde{u}_\mu^{j_\mu})_{j=0}^n \) if necessary since the numerators are bounded.

Remark 5.3.8. This approach can also be used to coarsen \( u_\epsilon \) in \( C(\Gamma; V) \). In this case, by triangle inequality, the error is bounded by

\[
\|T^V_P u_\epsilon - T^V_P u_j\|_{C(\Gamma; V)} \leq \sum_{\mu \in \text{supp} u_\epsilon} \|P_\mu\|_{C(\Gamma)} \|\tilde{u}_\mu - \tilde{u}_\mu^{j_\mu}\|_V,
\]

and thus these values should be used in \( \text{NextOpt} \). The condition (5.3.20) becomes

\[
\frac{\|\tilde{u}_\mu - \tilde{u}_\mu^{j_\mu}\|_V - \|\tilde{u}_\mu - \tilde{u}_\mu^{j_{i+1}}\|_V}{c_\mu^{j_{i+1}} - c_\mu^{j_i}} \geq \frac{\|\tilde{u}_\mu - \tilde{u}_\mu^{j_\mu}\|_V - \|\tilde{u}_\mu - \tilde{u}_\mu^{j_{i+1}}\|_V}{c_\mu^{j_{i+1}} - c_\mu^{j_i}} \quad \text{if} \quad i \geq j.
\]
5.3. Discretization of Deterministic Problems

Here, as in Section 5.1.3, \( P = \left( P_\mu \right)_{\mu \in \Lambda} \) denotes the tensor product orthonormal polynomial basis of \( L_\pi^2(\Gamma) \) with respect to any product probability measure \( \pi \).
Chapter 6.

Galerkin Methods for Elliptic Problems

The adaptive wavelet methods [CDD01, GHS07, DSS09] for symmetric elliptic problems use approximations of the residual to control refinements and to estimate errors of Galerkin projections. We apply a variant of these to symmetric parametric elliptic problems discretized by tensor product orthonormal polynomial bases on the parameter domain. This leads to a modular solver, which combines with any Galerkin discretization on the physical domain. For example, using adaptive finite elements or adaptive wavelet methods gives rise to a fully adaptive solver.

We begin Section 6.1 with some theoretical foundations, adapting our general abstract problem to the setting of symmetric operators on Hilbert spaces. We then consider algorithmic aspects of computing Galerkin approximations to the solution of such symmetric parametric equations.

Galerkin projections combine with elements of the adaptive methods from Chapter 5 to an adaptive solver, which is formulated first for a semidiscrete problem in Section 6.2, then with a Galerkin approximation also on the physical domain in Section 6.3.

6.1. Galerkin Projection

6.1.1. Theoretical Foundations

Let \( \mathcal{V} = \mathcal{W} \) be a separable Hilbert space. Furthermore, let \( D \) be positive and \( R_m \) be symmetric for all \( m \in \mathbb{N} \), such that (1.2.20) holds. By Remark 1.2.2, the parametric operator \( A(y) \) is positive for all \( y \in \Gamma \), and Remark 1.1.9 implies that \( A \) is a positive operator on \( L^2(\Gamma; V) \). It induces a norm that is equivalent to the standard norm on this space.

For any closed subspace \( \mathcal{V} \) of \( L^2(\Gamma; V) \), the Galerkin projection \( u_{\mathcal{V}} \) of \( u \) onto \( \mathcal{V} \) is the orthogonal projection of \( u \) onto \( \mathcal{V} \) with respect to the inner product \( \langle \cdot, \cdot \rangle_{\mathcal{V}} \) on \( L^2(\Gamma; V) \) from (1.1.16), i.e. due to Corollary 1.1.8, \( u_{\mathcal{V}} \) is the unique element of \( \mathcal{V} \) such that

\[
\int_{\Gamma} \langle A(y)u_{\mathcal{V}}(y), w(y) \rangle_{\mathcal{V}} \, d\mu(y) = \int_{\Gamma} \langle f(y), w(y) \rangle_{\mathcal{V}} \, d\mu(y) \quad \forall w \in \mathcal{V}.
\]  

(6.1.1)

Let \( P = (P_v)_{v \in \Lambda} \) be a tensor product orthonormal polynomial basis of \( L^2(\Gamma) \) as in (3.2.4). The synthesis operator \( T^V_P \) is an isometric isomorphism from \( \ell^2(\Lambda; V) \) to \( L^2(\Gamma; V) \). It transforms the inner product \( \langle \cdot, \cdot \rangle_{\mathcal{V}} \) to the inner product

\[
(v, w)_{\mathbb{R}} := \langle \Pi v, w \rangle_{\ell^2(\Lambda; V)} , \quad v, w \in \ell^2(\Lambda; V) ,
\]

(6.1.2)
Chapter 6. Galerkin Methods for Elliptic Problems

which induces the norm \( \| \cdot \|_{\mathcal{R}} \) on \( \ell^2(\Lambda; V) \).

Let \( \Lambda^\circ \subseteq \Lambda \). Then \( \ell^2(\Lambda^\circ; V) \) is a closed subspace of \( \ell^2(\Lambda; V) \). We denote by \( \Pi_{\Lambda^\circ} \) the orthogonal projection in \( \ell^2(\Lambda; V) \) onto \( \ell^2(\Lambda^\circ; V) \). For any \( v = (v_\mu)_{\mu \in \Lambda^\circ} \), \( \Pi_{\Lambda^\circ} v = (v_\mu)_{\mu \in \Lambda^\circ} \). We use the same notation if \( V \) is replaced by \( V^\circ \). The adjoint of \( \Pi_{\Lambda^\circ} \) is the natural embedding of \( \ell^2(\Lambda^\circ; V) \) into \( \ell^2(\Lambda; V) \).

We introduce the notation \( \mathfrak{A}_{\Lambda^\circ} := \Pi_{\Lambda^\circ} \mathfrak{R}_{\Lambda^\circ} \), \( \mathfrak{R}_{\Lambda^\circ} := \Pi_{\Lambda^\circ} \mathfrak{R}_{\Lambda^\circ} \), and \( \mathfrak{T}_{\Lambda^\circ} := \Pi_{\Lambda^\circ} \mathfrak{T} \). Since \( \mathfrak{T} \) maps \( \ell^2(\Lambda^\circ; V) \) onto \( \ell^2(\Lambda^\circ; V^\circ) \), we refrain from defining \( \mathfrak{T}_{\Lambda^\circ} \). By definition, \( \mathfrak{R}_{\Lambda^\circ} = \mathfrak{T} + \mathfrak{R}_{\Lambda^\circ} \). Furthermore, for \( \mathfrak{T}' := T'_{\mathfrak{B}} \ell^2(\Lambda^\circ; V) \), the Galerkin projection \( u_\mathfrak{T}' \) satisfies

\[ u_\mathfrak{T}' = T'_{\mathfrak{B}} u_{\Lambda^\circ} \]  

for the solution \( u_{\Lambda^\circ} = (u_{\Lambda^\circ, \mu})_{\mu \in \Lambda^\circ} \in \ell^2(\Lambda^\circ; V) \) of

\[ \mathfrak{A}_{\Lambda^\circ} u_{\Lambda^\circ} = \mathfrak{T}_{\Lambda^\circ} . \]  

Due to Proposition 3.2.11, (6.1.4) is equivalent to

\[ Du_{\Lambda^\circ, \nu} + \sum_{m=1}^{\infty} R_m (p_m u_{\Lambda^\circ, \nu + \varepsilon_m} + \alpha_m u_{\Lambda^\circ, \nu} + \beta_m u_{\Lambda^\circ, \nu - \varepsilon_m}) = f_\nu , \quad \nu \in \Lambda^\circ , \]  

where \( u_\mu := 0 \) for \( \mu \notin \Lambda^\circ \), see Theorem 3.2.12.

It is of some use to identify \( V \) and \( V^\circ \) via the Riesz isomorphism. It is with respect to this identification that \( D \) is a positive operator and \( R_m \), \( m \in \mathbb{N} \), are symmetric. This identification induces an identification of \( \ell^2(\Lambda; V) \) with \( \ell^2(\Lambda; V^\circ) \), with respect to which \( \mathfrak{A}, \mathfrak{T} \) and \( \mathfrak{R} \) are symmetric. The operator \( \mathfrak{T} \) is also positive, and thus has a positive square root \( \mathfrak{T}^{1/2} := \operatorname{id}_{\ell^2(\Lambda; V)} \otimes D^{1/2} \). We adapt condition (1.2.19) to this Hilbert space setting.

**Assumption 6.1.A.** There is a \( \gamma < 1 \) such that \( -\gamma \mathfrak{T} \leq \mathfrak{R} \leq \gamma \mathfrak{T} \).

The inequalities in Assumption 6.1.A are in the sense of symmetric operators, i.e.

\[ \left| \langle \mathfrak{R} v, v \rangle_{\mathcal{E}(\Lambda; V)} \right| \leq \gamma \left| \langle \mathfrak{T} v, v \rangle_{\mathcal{E}(\Lambda; V)} \right| \quad \forall v \in \ell^2(\Lambda; V) . \]  

Equivalently, we have

\[ \left\| \mathfrak{T}^{-1/2} \mathfrak{R} \mathfrak{T}^{-1/2} \right\|_{\mathcal{E}(\Lambda; V) \to \mathcal{E}(\Lambda; V)} \leq \gamma < 1 . \]  

**Remark 6.1.1.** Assumption 6.1.A is similar to the condition (1.2.19), (1.3.16) and (1.2.20). All are implied by

\[ \sum_{m=1}^{\infty} \|D^{-1}\|_{V^* \to V^*} \|R_m\|_{V \to V^*} \leq \gamma < 1 . \]  

Due to

\[ \mathfrak{T}^{-1/2} \mathfrak{R} \mathfrak{T}^{-1/2} = \sum_{m=1}^{\infty} K_m \otimes (D^{-1/2} R_mD^{-1/2}) \]
6.1. Galerkin Projection

and \( \|K_m\|_{\ell^2(\Lambda)} \leq 1 \), (6.1.8) implies

\[
\|D^{-1/2}K D^{-1/2}\| \leq \sum_{m=1}^{\infty} \|D^{-1/2}\|^2 \|R_m\| = \sum_{m=1}^{\infty} \|D^{-1}\| \|R_m\| \leq \gamma < 1.
\]

If \( D \) induces the inner product of \( V \), then (6.1.8) is equivalent to (1.2.19), and thus Assumption 6.1.A is a consequence of this condition.

Lemma 6.1.2. For any \( \Lambda^o \subset \Lambda \),

\[
(1 - \gamma)D \leq \mathbb{A}_{\Lambda^o} \leq (1 + \gamma)D,
\]

(6.1.9)

in the sense of (6.1.6).

Proof. Assumption 6.1.A implies \( R_{\Lambda^o} = \Pi_{\Lambda^o} \Pi_{\Lambda^o}^* \leq \gamma D \). For any \( v \in \ell^2(\Lambda^o; V) \),

\[
\langle \mathbb{A}_{\Lambda^o} v, v \rangle = \langle Dv, v \rangle + \langle \mathbb{R}_{\Lambda^o} v, v \rangle \leq (1 + \gamma) \langle Dv, v \rangle.
\]

This shows the second inequality of (6.1.9), and the first follows with \( \mathbb{R}_{\Lambda^o} \geq -\gamma D \).

Property (6.1.9) implies that the spectrum of \( D^{-1/2} \mathbb{A}_{\Lambda^o} D^{-1/2} \) is in \([1 - \gamma, 1 + \gamma] \). Due to the spectral mapping theorem, the spectrum of \( D^{1/2} \mathbb{R}_{\Lambda^o}^{-1} D^{1/2} \) is in \([(1 + \gamma)^{-1}, (1 - \gamma)^{-1}] \). This is equivalent to the statement (6.1.10). \( \square \)

In particular,

\[
\frac{1}{1 + \gamma} \mathbb{A}_{\Lambda^o} D^{-1} \mathbb{A}_{\Lambda^o} \leq \mathbb{A}_{\Lambda^o} \leq \frac{1}{1 - \gamma} \mathbb{A}_{\Lambda^o} D^{-1} \mathbb{A}_{\Lambda^o}
\]

(6.1.10)

follows from Lemma 6.1.2 using \( \mathbb{A}_{\Lambda^o} = \mathbb{A}_{\Lambda^o} \mathbb{R}_{\Lambda^o}^{-1} \mathbb{A}_{\Lambda^o} \) and applying (6.1.10).

6.1.2. Algorithmic Aspects

Let \( \Lambda^o \) be a finite subset of \( \Lambda \). Even under the simplifying assumption that we are able to perform operations exactly in the Hilbert spaces \( V \) and \( V^* \), it is generally infeasible to compute the solution \( u_{\Lambda^o} \in \ell^2(\Lambda^o; V) \) of (6.1.4).

The source term \( f_{\Lambda^o} = \Pi_{\Lambda^o} f \) may not be accessible. We assume the availability of a routine

\[
\text{InRHS}[\Lambda^o, \epsilon] \mapsto \tilde{f}_{\Lambda^o}
\]

(6.1.12)

which, for any \( \epsilon > 0 \), computes an approximation \( \tilde{f}_{\Lambda^o} \in \ell^2(\Lambda^o; V^*) \) of \( f_{\Lambda^o} \) for any \( v \in \ell^2(\Lambda^o; V) \) satisfying

\[
\|f_{\Lambda^o} - \tilde{f}_{\Lambda^o}\|_{\ell^2(\Lambda^o; V^*)} \leq \epsilon.
\]

Iterative solvers for (6.1.4) require a routine for evaluating \( \mathbb{A}_{\Lambda^o} v \) for any \( v \in \ell^2(\Lambda^o; V) \). Such a method is provided by \( \text{InApply}_{\eta} \). Due to the sparsity of \( \mathbb{A}_{\Lambda^o} \), we are able to compute these products exactly.
Chapter 6. Galerkin Methods for Elliptic Problems

\[
\text{InApply}_q[\Lambda^o, v] \mapsto \delta
\]

\[
\delta = (z_\nu)_{\nu \in \Lambda^o} \leftarrow \left( D + \sum_{m=1}^{\infty} \alpha_{\nu m}^m R_m \right)_{\nu \in \Lambda^o}
\]

forall \( \mu \in \Lambda^o \) do

forall \( \nu \in \Lambda^o, \nu = \mu - \epsilon_m, m \in \text{supp } \mu \) do

\[ z_\nu \leftarrow z_\nu + \beta_{\nu m}^m R_m v_\mu \]
end

forall \( \nu \in \Lambda^o, \nu = \mu + \epsilon_m, m \in \mathbb{N} \) do

\[ z_\nu \leftarrow z_\nu + \beta_{\nu m}^m R_m v_\mu \]
end
end

\[\text{Remark 6.1.3.} \quad \text{In the first line of InApply}_q[\Lambda^o, v], \text{the diagonal components} \]

\[ A_{\nu \nu} = D + \sum_{m=1}^{\infty} \alpha_{\nu m}^m R_m, \quad \nu \in \Lambda^o, \quad (6.1.14) \]

\[ \text{of } q \text{ are applied to the coefficients of } v. \text{ If all of the distributions } \pi_m, m \in \mathbb{N}, \text{ are symmetric, then } A_{\nu \nu} = D \text{ for all } \nu \in \Lambda^o \text{ by Remark 3.2.10. More generally, we assume that the operators } A_{\nu \nu} \text{ are available, and can be accessed without computing the infinite sum in (6.1.14).} \]

\[ \text{Alternatively, the operator } q \text{ could be replaced by } q_{[M]} \text{ for some } M \in \mathbb{N}. \text{ These operators approximate } q \text{ by Theorem 4.3.2. However, this introduces an additional error, which we prefer to avoid.} \]

\[ \text{In InApply}_q[\Lambda^o, v], \text{the product } R_m v_\mu \text{ is computed for all } \mu \in \Lambda^o \text{ for which } \mu + \epsilon_m \in \Lambda^o \text{ or } \mu - \epsilon_m \in \Lambda^o. \text{ Let} \]

\[ \mathcal{N}(\Lambda^o) := \left\{ (\mu, \nu) \mid \mu, \nu \in \Lambda^o, |\mu - \nu| = 1 \right\}. \quad (6.1.15) \]

Thus \( (\mu, \nu) \in \mathcal{N}(\Lambda^o) \) if and only if \( \nu = \mu \pm \epsilon_m \). We call such indices neighbors. Furthermore, we call a set \( \Lambda^o \subset \Lambda \) monotonic if for any \( \mu \in \Lambda^o \) and any \( \nu \in \Lambda, \nu_m \leq \mu_m \) for all \( m \in \mathbb{N} \) implies \( \nu \in \Lambda^o \). The average length of indices in \( \Lambda^o \),

\[ \bar{\lambda}(\Lambda^o) := \frac{1}{\# \Lambda^o} \sum_{\mu \in \Lambda^o} \# \text{supp } \mu, \quad (6.1.16) \]

provides a bound for the size of \( \mathcal{N}(\Lambda^o) \) compared to the size of \( \Lambda^o \).

\[ \text{Lemma 6.1.4.} \quad \text{For any finite } \Lambda^o \subset \Lambda, \]

\[ \# \mathcal{N}(\Lambda^o) \leq \bar{\lambda}(\Lambda^o) \# \Lambda^o. \quad (6.1.17) \]

Equality holds if and only if \( \Lambda^o \) is monotonic.

110
6.1. Galerkin Projection

Proof. We assume initially that $\Lambda^0$ is monotonic. Then $\mu \in \Lambda^0$ has the neighbors $v = \mu - \epsilon_m$ for all $m \in \text{supp } \mu$. All neighbor pairs in $\Lambda^0$ are of this form since if $v = \mu + \epsilon_m \in \Lambda^0$ for some $m \in \mathbb{N}$, then $\mu = v - \epsilon_m$ and $m \in \text{supp } v$. Consequently,

$$\#_{\mathcal{V}}(\Lambda^0) = \sum_{\mu \in \Lambda^0} \# \text{supp } \mu = \bar{\lambda}(\Lambda^0) \# \Lambda^0.$$ 

If $\Lambda^0$ is not monotonic, then there is a $\mu \in \Lambda^0$ and an $m \in \text{supp } \mu$ such that $v = \mu - \epsilon_m$ is not in $\Lambda^0$. Therefore,

$$\#_{\mathcal{V}}(\Lambda^0) < \sum_{\mu \in \Lambda^0} \# \text{supp } \mu = \bar{\lambda}(\Lambda^0) \# \Lambda^0. \quad \Box$$

Proposition 6.1.5. The routine $\text{InApply}_{\mathbb{R}}[\Lambda^0, v]$ computes $\mathbb{U}_{\Lambda^0} v$ using one application of $A_{\Lambda^0 v}$ for each $v \in \Lambda^0$ and a total of no more than $2\bar{\lambda}(\Lambda^0) \# \Lambda^0$ applications of $R_m$, $m \in \mathbb{N}$, for any $v \in \ell^2(\Lambda^0; V)$.

Proof. It follows from Proposition 3.2.11 that $\text{InApply}_{\mathbb{R}}[\Lambda^0, v]$ does indeed compute $\mathbb{U}_{\Lambda^0} v$. The multiplications $A_{\Lambda^0 v} v$ appear in the first line of the algorithm. The total number of subsequent products $R_m v_{\mu}$ is bounded by $2\#_{\mathcal{V}}(\Lambda^0)$. Thus the assertion follows using Lemma 6.1.4.

Proposition 6.1.5 bounds the total number of operator-vector multiplications in a call of $\text{InApply}_{\mathbb{R}}$ by $(1 + 2\bar{\lambda}(\Lambda^0)) \# \Lambda^0$. The average index length $\bar{\lambda}(\Lambda^0)$ is generally small compared to $\# \Lambda^0$. For certain monotonic sets $\Lambda^0$, [BAS10, Corollary 4.9] estimates the maximal index length by $\log \# \Lambda^0$.

We assume that an iterative method

$$\text{PCG}_{\mathbb{R}}[\Lambda^0, \tilde{f}, \lambda^0_{\Lambda^0}, \epsilon] \mapsto \tilde{u}_{\Lambda^0} \quad (6.1.18)$$

is available which, starting from the initial approximation $\tilde{u}^0_{\Lambda^0} \in \ell^2(\Lambda^0; V)$, computes $\tilde{u}_{\Lambda^0} \in \ell^2(\Lambda^0; V)$ satisfying

$$\|\tilde{u}_{\Lambda^0} - \tilde{u}^*_{\Lambda^0}\|_{\mathbb{R}} \leq \frac{\epsilon}{\sqrt{1 - \gamma}}, \quad (6.1.19)$$

where $\tilde{u}^*_{\Lambda^0} = \mathbb{U}_{\Lambda^0}^{-1} f_{\Lambda^0}$. Such a method would call the function $\text{InApply}_{\mathbb{R}}$ to evaluate the application of the operator $\mathbb{U}_{\Lambda^0}$ to a $v \in \ell^2(\Lambda^0; V)$. A realization of $\text{PCG}_{\mathbb{R}}$ by a preconditioned conjugate gradient iteration is provided in Section 6.1.3, see Proposition 6.1.9.

The method $\text{Galerkin}_{\mathbb{R}}$ combines $\text{PCG}_{\mathbb{R}}$ with $\text{InRHS}_{\mathbb{R}}$ to approximate $u_{\Lambda^0}$ with an ensured error bound in the norm $\|\cdot\|_{\mathbb{R}}$.

Proposition 6.1.6. For any finite $\Lambda^0 \subset \Lambda$, $\tilde{u}^0_{\Lambda^0} \in \ell^2(\Lambda^0; V)$, $\epsilon > 0$ and $\delta \in (0, 1)$, a call of $\text{Galerkin}_{\mathbb{R}}[\Lambda^0, \tilde{u}^0_{\Lambda^0}, \epsilon, \delta, \gamma]$ computes $\tilde{u}_{\Lambda^0} \in \ell^2(\Lambda^0; V)$ with

$$\|\tilde{u}_{\Lambda^0} - u_{\Lambda^0}\|_{\mathbb{R}} \leq \epsilon. \quad (6.1.20)$$

If $f_{\Lambda^0}$ is available, $\delta = 0$ is admissible.
Chapter 6. Galerkin Methods for Elliptic Problems

\[ \text{Galerkin}_{\mathbb{R}}[\Lambda^0, \tilde{u}^0_{\Lambda^0}, \varepsilon, \delta, \gamma] \mapsto \tilde{u}_{\Lambda^0} \]

\[ \epsilon_f \leftarrow \delta \sqrt{1 - \gamma} \| D^{-1} \|_{V' \to V} \epsilon \]
\[ \epsilon_1 \leftarrow (1 - \delta) \sqrt{1 - \gamma} \epsilon \]
\[ \tilde{f}_{\Lambda^0} \leftarrow \text{InRHS}[\Lambda^0, \epsilon_f] \]
\[ \tilde{u}_{\Lambda^0} \leftarrow \text{PCG}_{\mathbb{R}}[\Lambda^0, \tilde{f}_{\Lambda^0}, \tilde{u}^0_{\Lambda^0}, \epsilon_1] \]

**Proof.** Due to the assumption (6.1.13), \( \| \tilde{f}_{\Lambda^0} - \tilde{f}_{\Lambda^0} \|_{C(\Lambda^0; V')} \leq \epsilon_f \). Since \( u_{\Lambda^0} = \mathbb{R}_{\Lambda^0}^{-1} f_{\Lambda^0} \) and \( \tilde{u}^*_{\Lambda^0} = \mathbb{R}_{\Lambda^0}^{-1} \tilde{f}_{\Lambda^0} \), Lemma 6.1.2 implies

\[
\| u_{\Lambda^0} - \tilde{u}^*_{\Lambda^0} \|_{\mathbb{R}}^2 = \| \mathbb{R}_{\Lambda^0}^{-1} (f_{\Lambda^0} - \tilde{f}_{\Lambda^0}) \|_{\mathbb{R}}^2 = \langle (f_{\Lambda^0} - \tilde{f}_{\Lambda^0}), \mathbb{R}_{\Lambda^0}^{-1} (f_{\Lambda^0} - \tilde{f}_{\Lambda^0}) \rangle \\
\leq \frac{1}{1 - \gamma} \langle f_{\Lambda^0} - \tilde{f}_{\Lambda^0}, D^{-1}(f_{\Lambda^0} - \tilde{f}_{\Lambda^0}) \rangle \\
\leq \frac{1}{1 - \gamma} \| D^{-1} \| e_f^2 = (\delta \varepsilon)^2.
\]

Furthermore, (6.1.19) implies

\[
\| \tilde{u}_{\Lambda^0} - \tilde{u}^*_{\Lambda^0} \|_{\mathbb{R}} \leq \frac{\epsilon_1}{\sqrt{1 - \gamma}} = (1 - \delta) \varepsilon .
\]

The assertion follows by triangle inequality. \( \square \)

### 6.1.3. Conjugate Gradient Iteration

We use the preconditioned conjugate gradient method with preconditioner \( D^{-1} \) to approximate the Galerkin projection \( u_{\Lambda^0} \) for any finite \( \Lambda^0 \subset \Lambda \).

**Theorem 6.1.7.** The conjugate gradient iteration for \( \mathbb{R}_{\Lambda^0} u_{\Lambda^0} = \tilde{f}_{\Lambda^0}, \tilde{f}_{\Lambda^0} \in \ell^2(\Lambda^0; V') \) with initial approximation \( \tilde{u}^0_{\Lambda^0} \in \ell^2(\Lambda^0; V) \) and preconditioner \( D^{-1} \) constructs \( \tilde{u}^i_{\Lambda^0} \in \ell^2(\Lambda^0; V) \) satisfying

\[
\| \tilde{u}^i_{\Lambda^0} - \tilde{u}^*_{\Lambda^0} \|_{\mathbb{R}} \leq 2 \frac{q^i}{1 + q^i} \| \tilde{u}^0_{\Lambda^0} - \tilde{u}^*_{\Lambda^0} \|_{\mathbb{R}}, \quad q = \frac{\gamma}{1 + \sqrt{1 - \gamma^2}}, \quad (6.1.21)
\]

for all \( i \in \mathbb{N}_0 \).

**Proof.** The assertion follows from [Hac91, Satz 9.4.14], which also holds in separable Hilbert spaces, with

\[
q = \frac{\sqrt{1 + \gamma} - \sqrt{1 - \gamma}}{\sqrt{1 + \gamma} + \sqrt{1 - \gamma}} = \frac{\gamma}{1 + \sqrt{1 - \gamma^2}},
\]

and using (6.1.9) from Lemma 6.1.2. \( \square \)

A version of the preconditioned conjugate gradient method is given in PCG\( _{\mathbb{R}} \). It uses a termination criterion based on the following norm equivalence.
Lemma 6.1.8. For all $i \in \mathbb{N}_0$,\begin{equation}
abla \sum \frac{1 + \gamma}{1 + \gamma} \eta_i \leq \left\| \tilde{u}_{A^0}^i - \tilde{u}_{A^0}^i \right\|^2_R \leq \frac{1}{1 - \gamma} \eta_i, \tag{6.1.22}
abla \end{equation}
where $\tilde{u}_{A^0}^i \in \ell^2(A^0; V)$ is the solution of $\mathcal{A}_{V} \tilde{u}_{A^0}^i = \tilde{f}_{A^0}$.

Proof. By definition, $\eta_i = \left\langle \tilde{r}^i, s^i \right\rangle_{\mathcal{E}(A^0; V)}$, $\tilde{r}^i = \tilde{f}_{A^0} - \mathcal{A}_{V} \tilde{u}_{A^0}^i$, and $s^i = D^{-1} \tilde{r}^i$. We abbreviate $e^i := \tilde{u}_{A^0}^i - \tilde{u}_{A^0}^i$. The assertion follows from Lemma 6.1.2 since\begin{equation}
abla \left\| e^i \right\|^2_R = \left\langle \mathcal{A}_{V} e^i, \tilde{u}_{A^0}^i - \tilde{u}_{A^0}^i \right\rangle_{\mathcal{E}(A^0; V)} \tag{6.1.23}
abla \end{equation}
and\begin{equation}
abla \eta_i = \left\langle \mathcal{A}_{V} e^i, D^{-1} \mathcal{A}_{V} e^i \right\rangle_{\mathcal{E}(A^0; V)}. \tag{6.1.24}
abla \end{equation}

Proposition 6.1.9. The method $\text{PCG}_R[A^0, \tilde{f}_{A^0}, \tilde{u}_{A^0}^0, \epsilon]$ terminates and returns $\tilde{u}_{A^0}$ satisfying\begin{equation}
abla \left\| \tilde{u}_{A^0} - \tilde{u}_{A^0}^* \right\|^2_R \leq \frac{e}{\sqrt{1 - \gamma}}. \tag{6.1.25}
abla \end{equation}

At most\begin{equation}
abla 1 + \left[ \frac{\log(2 \sqrt{1 + \gamma} \left\| \tilde{u}_{A^0}^0 - \tilde{u}_{A^0}^* \right\|_R)}{\log q} \right]. \tag{6.1.26}
abla \end{equation}
Chapter 6. Galerkin Methods for Elliptic Problems

iterations are performed, with \( q \) from (6.1.21). Each iteration contains \( \#\Lambda^o \) evaluations of \( D^{-1} \), one application of \( A_v \) for each \( v \in \Lambda^o \), and a total of no more than \( 2\bar{\lambda}(\Lambda^o)\#\Lambda^o \) applications of \( R_m, m \in \mathbb{N} \).

**Proof.** Equation (6.1.23) follows from Lemma 6.1.8. Let the final iterate be \( \hat{u}_{\Lambda^o} = \hat{u}^N_{\Lambda^o} \). Then provided \( N \geq 1 \), using the other inequality in Lemma 6.1.8,

\[
\|\hat{u}^{N-1}_{\Lambda^o} - \hat{u}^*_{\Lambda^o}\|_\# \geq \frac{1}{\sqrt{1+\gamma}} \eta_{N-1} \geq \frac{\epsilon}{\sqrt{1+\gamma}} .
\]

By Theorem 6.1.7,

\[
\epsilon \leq \sqrt{1+\gamma} \|\hat{u}^{N-1}_{\Lambda^o} - \hat{u}^*_{\Lambda^o}\|_\# \leq 2\sqrt{1+\gamma} q^{N-1} \|\hat{u}^0_{\Lambda^o} - \hat{u}^*_{\Lambda^o}\|_\# .
\]

Solving for \( N \) leads to

\[
N - 1 \leq \frac{\log(2\epsilon^{-1}\sqrt{1+\gamma} \|\hat{u}^0_{\Lambda^o} - \hat{u}^*_{\Lambda^o}\|_\#)}{\log q} .
\]

The final part of the assertion is a consequence of Proposition 6.1.5. \( \square \)

### 6.2. An Adaptive Iterative Solver

#### 6.2.1. Successive Refinement of the Active Index Set

The error \( \|u - u_{\Lambda^v}\|_\# \) of the Galerkin projection \( u_{\Lambda^v} \) onto \( \ell^2(\Lambda^v; V) \) can be estimated, given any other approximation \( v \) of \( u \) in \( \ell^2(\Lambda^v; V) \), as

\[
\|u - u_{\Lambda^v}\|_\# \leq \|u - v\|_\# . \tag{6.2.1}
\]

The following proposition strengthens (6.2.1) under the assumption that the residual \( \tilde{f} - \mathfrak{f}v \) of \( v \) is sufficiently resolved in \( \ell^2(\Lambda^v; V^*) \). It is adapted from [CDD01, Lemma 4.1], [GHS07, Lemma 1.2] or [DSS09, Lemma 4.1].

**Proposition 6.2.1.** Let \( \Lambda^v \subset \Lambda \) and \( \delta \in [0, 1] \). Let \( v \in \ell^2(\Lambda^v; V) \) with

\[
\|\Pi_{\Lambda^v}(\tilde{f} - \mathfrak{f}v)\|_{\ell^2(\Lambda^v; V^*)} \geq \delta \|\tilde{f} - \mathfrak{f}v\|_{\ell^2(\Lambda^v; V^*)} . \tag{6.2.2}
\]

Then the Galerkin projection \( u_{\Lambda^v} \in \ell^2(\Lambda^v; V) \) satisfies

\[
\|u - u_{\Lambda^v}\|_\# \leq \sqrt{1 - \frac{1-\gamma}{1+\gamma} \kappa(D)^{-1} \delta^2 \|u - v\|_\#} , \tag{6.2.3}
\]

where \( \kappa(D) := \|D\|_{V \rightarrow V^*} \|D^{-1}\|_{V^* \rightarrow V} \).
6.2. An Adaptive Iterative Solver

Proof. Due to (6.2.2),
\[\|u - u_A\|_V \geq \|u\|_V^{1/2} \|\varphi(u - u_A)\|_{\mathcal{L}(A^*, V')} \geq \|u\|_V^{1/2} \|\varphi f - \varphi u\|_{\mathcal{L}(A^*, V')} \geq \|u\|_V^{1/2} \|f - u\|_\mathbb{R}.\]

By Galerkin orthogonality,
\[\|u - u_A\|_\mathbb{R}^2 = \|u - v\|_\mathbb{R}^2 - \|u - u_A\|_\mathbb{R}^2 \leq (1 - \|\varphi\|_V^{-1}) \|f - u\|_\mathbb{R}^2.\]

The assertion follows using the estimates
\[\|\varphi\| \leq (1 + \gamma) \|D\|_{V \to V'}, \quad \|\varphi^{-1}\| \leq \frac{1}{1 - \gamma} \|D^{-1}\|_{V' \to V'},\]
see Theorem 1.3.5.\hfill \Box

We use Proposition 6.2.1 as follows. Let \(A(0) \subset A\) and let \(u_A(0) \in \ell^2(A(0); V)\) such that an upper bound for \(\|u - u_A(0)\|_\mathbb{R}\) is available. We construct a set \(A(1) \subset A\) such that \(A(0) \subset A(1)\), and such that (6.2.2) holds with \(A^0 = A(1)\) and \(v = u_A(0)\). Then (6.2.3) provides an explicit upper bound for \(\|u - u_A(1)\|_\mathbb{R}\), where \(u_A(1) = \varphi^{-1} u_A(0)\). We approximate the Galerkin projection \(u_A(1)\) by an iterative method, and repeat the process.

An important component of this method is provided by the routine \(\text{Residual}_{\mathbb{R}, f}\), which approximates the residual up to a prescribed relative tolerance. The residual provides the basis for refining the index set \(A(0)\); furthermore, the routine checks for convergence of the iteration.

\[
\text{Residual}_{\mathbb{R}, f}\left[\epsilon, v, \eta_0, \chi, \alpha, \beta\right] \mapsto [\rho, \omega, \eta, \zeta]
\]

\[
A(0) \leftarrow \text{supp } v
\]
\[
\varrho = (g_\mu)_{\mu \in A(0)} \leftarrow \mathcal{D} v = (Dv_\mu)_{\mu \in A(0)}
\]
\[
\zeta \leftarrow 2\chi\eta_0
\]

repeat
\[
\zeta \leftarrow \frac{\zeta}{2}
\]
\[
b = (h_\nu)_{\nu \in A} \leftarrow \text{RHS} [\beta \zeta \|D^{-1}\|_{V' \to V'}^{1/2}] - \text{Apply}_{\mathbb{R}} [v, (1 - \beta)\zeta \|D^{-1}\|_{V' \to V'}^{1/2}]
\]
\[
\omega = (w_\nu)_{\nu \in A} \leftarrow D^{-1} \hat{b} = (D^{-1} h_\nu)_{\nu \in A}
\]
\[
\rho = (\varrho_\nu)_{\nu \in A} \leftarrow (\sqrt{\langle h_\nu - g_\nu, w_\nu - v_\nu \rangle_{V'}})_{\nu \in A}
\]
\[
\eta \leftarrow \|\rho\|_{\mathcal{L}(A)}
\]

until \(\zeta \leq \alpha\eta\) or \(\eta + \zeta \leq \epsilon\)

Let \(V_D\) denote the Hilbert space \(V\) with inner product \(\langle \cdot, \cdot \rangle_V\). Its dual \(V_D^*\) is equal to \(V^*\) with inner product \(\langle \cdot, D^{-1} \cdot \rangle_{V'}\). We note that the inner product on \(\ell^2(A; V_D)\) is given by \(\langle \cdot, \cdot \rangle_{\mathcal{L}(A; V')}\).
Proposition 6.2.2. For any finitely supported $v = (v_r)_{r \in \Lambda} \in \ell^2(\Lambda; V), \varepsilon > 0, \eta_0 > 0, \chi > 0, \omega > 0$ and $0 < \beta < 1$, a call of Residual$_{\mathfrak{R}, \mathfrak{I}}([e, v, \eta_0, \chi, \omega, \beta])$ computes $\rho \in \ell^2(\Lambda)$ and $\eta = \|\rho\|_{\ell^2(\Lambda)}$ with
\[
\left| \eta - \|v\|_{\ell^2(\Lambda; V_D)} \right| \leq \left( \sum_{r \in \Lambda} \|v_r - \|v_r\|_{V_D} \|^2 \right)^{1/2} \leq \zeta,
\]
where $r = (r_v)_{r \in \Lambda} \in \ell^2(\Lambda; V^*)$ is the residual $r = \rho - \mathfrak{R}v$. Furthermore, $w \in \ell^2(\Lambda; V)$ is finitely supported in $\Lambda$ and
\[
\|w - \mathfrak{D}^{-1}(f - \mathfrak{R}v)\|_{\ell^2(\Lambda; V_D)} \leq \zeta.
\]
In particular,
\[
\|u - w\|_{\ell^2(\Lambda; V_D)} \leq \gamma \|u - v\|_{\ell^2(\Lambda; V_D)} + \zeta
\]
with $\gamma$ as in Assumption 6.1.A. Finally, $\zeta$ satisfies either $\zeta \leq \omega \eta$ or $\eta + \zeta \leq \varepsilon$.

Proof. By construction,
\[
\|\mathfrak{B} - (f - \mathfrak{R}v)\|_{\ell^2(\Lambda; V_D)} \leq \|\mathfrak{D}^{-1}\|_{V^* \rightarrow V}^{1/2} \|\mathfrak{B} - (f - \mathfrak{R}v)\|_{\ell^2(\Lambda; V^*)} \leq \zeta.
\]
Since $g = \mathfrak{D}v$ and $r = \rho - (\mathfrak{D}v + \mathfrak{R}v)$,
\[
\|\mathfrak{B} - g - v\|_{\ell^2(\Lambda; V_D)} = \|\mathfrak{B} - (f - \mathfrak{R}v)\|_{\ell^2(\Lambda; V_D)} \leq \zeta.
\]
Equation (6.2.4) follows by triangle inequality with $\rho_v = \|h_v - g_v\|_{V_D}$. Also, using $w = \mathfrak{D}^{-1}\mathfrak{B}$,
\[
\|w - \mathfrak{D}^{-1}(f - \mathfrak{R}v)\|_{\ell^2(\Lambda; V_D)} = \|\mathfrak{B} - (f - \mathfrak{R}v)\|_{\ell^2(\Lambda; V_D)} \leq \zeta.
\]
Due to Assumption 6.1.A, using that $\mathfrak{D}^{-1}\mathfrak{R}$ is self-adjoint on $\ell^2(\Lambda; V_D)$,
\[
\|\mathfrak{D}^{-1}\mathfrak{R}\|_{\ell^2(\Lambda; V_D) \rightarrow \ell^2(\Lambda; V_D)} = \sup_{\|u\|_{\ell^2(\Lambda; V_D)} = 1} \left\{ \mathfrak{D} \mathfrak{D}^{-1}\mathfrak{R}u, u \right\} \leq \gamma.
\]
Therefore, (6.2.6) is a consequence of Theorem 5.1.1. \[\square\]

Remark 6.2.3. The geometric decrease of the tolerance $\zeta$ in Residual$_{\mathfrak{R}, \mathfrak{I}}$ ensures that the total cost is proportional to that of the final iteration of the loop. Nevertheless, to improve efficiency, it is still useful to minimize the number of iterations. We suggest the update
\[
\zeta \leftarrow \omega \left\{ \frac{1 - \omega}{1 + \omega} (\eta + \zeta) \right\} = \zeta_1
\]
in place of $\zeta \leftarrow \zeta/2$. This still ensures a geometric decrease of $\zeta$ since if $\zeta > \omega \eta$, then
\[
\zeta_1 = \omega \left\{ \frac{1 - \omega}{1 + \omega} (\eta + \zeta) \right\} < \frac{1 - \omega}{1 + \omega} (\zeta + \omega \zeta) = (1 - \omega) \zeta.
\]
Furthermore, if $\zeta > \omega \eta$, then also
\[
\zeta_1 = \omega \left\{ \frac{1 - \omega}{1 + \omega} (\eta + \zeta) \right\} > \omega (1 - \omega) \eta > \omega (\eta - \zeta).
\]
6.2. An Adaptive Iterative Solver

The term $\eta - \zeta$ in the last expression of (6.2.9) is a lower bound for the true residual $\|r\|_{\ell^2(A;V')}$. In this sense, the prescription (6.2.7) does not select an unnecessarily small tolerance.

Finally, if $\zeta \leq 2 \omega (1 - \omega)^{-1} \eta$, then $\zeta_1 \leq \omega \eta$. If the next value of $\eta$ is greater than or equal to the current value, this ensures that the termination criterion is met in the next iteration. For example, under the mild condition $\zeta \leq (1 + 4 \omega - \omega^2) (1 - \omega)^{-2} \eta$, we have $\zeta_1 \leq 2 \omega (1 - \omega)^{-1} \eta$. The loop can therefore be expected to terminate within three iterations.

We recall the routine Coarsen from Section 5.2.2, modifying its syntax to emphasize index sets rather than vectors. For any finitely supported $c = (c_v)_{v \in A} \in \ell^2(A)$,

$$\text{Coarsen}[c, \eta] \mapsto A_\eta$$

(6.2.10)

constructs a set $A_\eta$ of minimal size such that

$$\| c - \Pi_{A_\eta} c \|_{\ell^2(A)} \leq \eta.$$  

(6.2.11)

This can be realized by sorting $c$, and then truncating suitably. More generally, it suffices for $A_\eta$ to be of minimal size only up to a constant factor, in which case the more efficient approximate sorting routine BucketSort can be employed.

The routine SolveGalerkin[f] combines Residual[f] and Coarsen with the method Galerkin[f] from Section 6.1.2 to an adaptive iterative solver for $\forall u = f$.

---

**SolveGalerkin[f]($c, \gamma, \chi, \delta, \omega, \sigma, \beta$) $\mapsto u_e$**

\[
A^{(0)} \leftarrow \emptyset \\
\tilde{u}_{A^{(0)}} \leftarrow 0 \\
\delta_0 \leftarrow \sqrt{(1 - \gamma)^{-1} \|D^{-1}\|_{V' \rightarrow V} \|\|_{\ell^2(A;V')}} \\
\text{for all } k \in \mathbb{N} \text{ do} \\
[\rho, w, \eta, \zeta] \leftarrow \text{Residual[f]}(\sqrt{1 - \gamma}, \tilde{u}_{A^{(k)}}, \delta_k, \chi, \omega, \beta) \\
\delta_k \leftarrow (\eta + \zeta)/\sqrt{1 - \gamma} \\
\text{if } \min(\delta_k, \bar{\delta}_k) \leq \epsilon \text{ then break} \\
\bar{\Delta} \leftarrow \text{Coarsen}[\rho, \eta, \sqrt{1 - (\omega + \delta + \omega \delta)^2}] \\
A^{(k+1)} \leftarrow A^{(k)} \cup \bar{\Delta} \\
\tilde{u}_{A^{(k+1)}} \leftarrow \text{Galerkin[f]}(A^{(k+1)}, w|_{A^{(k+1)}}, \sigma \min(\delta_k, \bar{\delta}_k), \beta, \gamma) \\
\bar{\delta}_{k+1} \leftarrow (\sigma + \sqrt{1 - \delta^2 (1 - \gamma)(1 + \gamma)^{-1}}) \min(\delta_k, \bar{\delta}_k) \\
\text{end} \\
u_e \leftarrow \tilde{u}_{A^{(0)}}
\]

---

**Lemma 6.2.4.** If $\delta > 0$, $\omega > 0$, and $\omega + \delta + \omega \delta \leq 1$, then the set $\bar{\Delta}$ constructed by Coarsen[$\rho, \eta, \sqrt{1 - (\omega + \delta + \omega \delta)^2}]$ in SolveGalerkin[f] is such that

$$\| \Pi_{\bar{\Delta}} v_k \|_{\ell^2(A;V')} \geq \delta \| v_k \|_{\ell^2(A;V')} ,$$

(6.2.12)

117
Chapter 6. Galerkin Methods for Elliptic Problems

where \( r_k := \| f - \mathcal{R} u_{\mathcal{A}^{(k)}} \| \) is the residual at iteration \( k \in \mathbb{N}_0 \).

Proof. By Proposition 6.2.2,

\[
\| r_k \|_{\mathcal{E}(A; V)} \leq \eta + \zeta \leq (1 + \omega) \| \varrho \|_{\mathcal{E}(A)} .
\]

Furthermore, abbreviating \( \alpha := \omega + \delta + \omega \delta \), (6.2.11) implies

\[
\| \Pi_A \varrho \|_{\mathcal{E}(A)}^2 = \| \varrho \|_{\mathcal{E}(A)}^2 - \| \Pi_A \varrho \|_{\mathcal{E}(A)}^2 \geq \alpha^2 \| \varrho \|_{\mathcal{E}(A)}^2 .
\]

Consequently, using (6.2.4) and \( \zeta \leq \omega \| \varrho \|_{\mathcal{E}(A)} \),

\[
\| \Pi_A r_k \|_{\mathcal{E}(A; V)} \geq \| \Pi_A \varrho \|_{\mathcal{E}(A)} - \zeta \geq (\alpha - \omega) \| \varrho \|_{\mathcal{E}(A)} \geq \frac{\alpha - \omega}{1 + \omega} \| r_k \|_{\mathcal{E}(A; V)} ,
\]

and the last term is equal to \( \delta \| r_k \|_{\mathcal{E}(A; V)} \) by definition of \( \alpha \).

**Theorem 6.2.5.** If \( \epsilon > 0, \chi > 0, \delta > 0, \omega > 0, \omega + \delta + \omega \delta \leq 1, 0 < \beta < 1 \) and \( 0 < \sigma < 1 - \sqrt{1 - \delta^2(1 - \gamma)(1 + \gamma)^{-1}} \), then \( \text{SolveGal}^{\mathcal{A}_{(\epsilon, \gamma, \chi, \delta, \omega, \alpha, \beta)}} \) constructs a finitely supported \( u_{\epsilon} \in \ell^2(\mathcal{A}; V) \) with

\[
\| u - u_{\epsilon} \|_{\mathcal{R}} \leq \epsilon .
\]

Moreover,

\[
\sqrt{1 - \gamma} \frac{1 - \omega}{1 + \gamma + \omega} \delta_k \leq \| u - \tilde{u}_{\mathcal{A}^{(k)}} \|_{\mathcal{R}} \leq \min(\delta_k, \delta_k) \quad (6.2.14)
\]

for all \( k \in \mathbb{N}_0 \) reached by \( \text{SolveGal}^{\mathcal{A}_{(\epsilon, \gamma, \chi, \delta, \omega, \alpha, \beta)}} \).

Proof. Due to the termination criterion of \( \text{SolveGal}^{\mathcal{A}_{(\epsilon, \gamma, \chi, \delta, \omega, \alpha, \beta)}} \), it suffices to show (6.2.14).

For \( k = 0 \), since \( \| u \|_{\mathcal{E}(A; V)} \leq \| \varrho \|_{\mathcal{R}}^{1/2} \| u \|_{\mathcal{R}} \),

\[
\| u - \tilde{u}_{\mathcal{A}^{(0)}} \|_{\mathcal{R}}^2 = \| u \|_{\mathcal{R}}^2 = \langle f, u \rangle_{\mathcal{E}(A; V)} \leq \| f \|_{\mathcal{E}(A; V)} \| u \|_{\mathcal{E}(A; V)} \leq \delta_0 \| u \|_{\mathcal{R}} .
\]

Let \( \| u - \tilde{u}_{\mathcal{A}^{(0)}} \|_{\mathcal{R}} \leq \delta_k \) for some \( k \in \mathbb{N}_0 \). Abbreviating \( r_k := f - \mathcal{A} \tilde{u}_{\mathcal{A}^{(k)}} \), using (6.1.11) then (6.2.4), we have

\[
\| u - \tilde{u}_{\mathcal{A}^{(k)}} \|_{\mathcal{R}} \leq \frac{1}{\sqrt{1 - \gamma}} \| r_k \|_{\mathcal{E}(A; V)} \leq \frac{\zeta + \eta}{\sqrt{1 - \gamma}} = \delta_k .
\]

If \( \min(\delta_k, \delta_k) > \epsilon \), then \( \zeta \leq \omega \eta \) by Proposition 6.2.2. Due to Lemma 6.2.4, Proposition 6.2.1 with \( V_D \) in place of \( V \) implies

\[
\| u - u_{\mathcal{A}^{(\epsilon, \gamma, \chi, \delta, \omega, \alpha, \beta)}} \|_{\mathcal{R}} \leq \sqrt{1 - \frac{1 - \gamma}{1 + \gamma}} \delta^2 \min(\delta_k, \delta_k) .
\]
6.2. An Adaptive Iterative Solver

where $u_{A^{(k+1)}}$ is the exact Galerkin projection of $u$ onto $L^2(A^{(k+1)}; V)$. By Proposition 6.1.6, $\bar{u}_{A^{(k+1)}}$ approximates $u_{A^{(k+1)}}$ up to an error of at most $\sigma \min(\delta_k, \bar{\delta}_k)$ in the norm $\|\|_\mathbb{R}$. It follows by triangle inequality that $\left\| u - \bar{u}_{A^{(k+1)}} \right\|_\mathbb{R} \leq \delta_{k+1}$.

To show the other inequality in (6.2.14), we note that for any $k \in \mathbb{N}_0$,

$$\left\| u - \bar{u}_{A^{(0)}} \right\|_\mathbb{R} \geq \frac{1}{\sqrt{1 + \gamma}} \|u_k\|_{L^2(A^{(0)}; V)} \geq \frac{\eta - \zeta}{\sqrt{1 + \gamma}} = \sqrt{1 - \gamma \eta - \zeta} \delta_k,$$

and $(\eta - \zeta)(\eta + \zeta)^{-1} \geq (1 - \omega)(1 + \omega)^{-1}$. Finally, since

$$\delta_k \leq \left( \sigma + \sqrt{1 - \delta^2(1 - \gamma)(1 + \gamma)^{-1}} \right)^k \delta_0$$

and $\sigma + \sqrt{1 - \delta^2(1 - \gamma)(1 + \gamma)^{-1}} < 1$ by assumption, the iteration does terminate. \hfill \Box

6.2.2. Optimality Properties

The following statement is adapted from [GHS07, Lemma 2.1] and [DSS09, Lemma 4.1].

**Lemma 6.2.6.** Let $A^{(0)}$ be a finite subset of $A$ and $v \in L^2(A^{(0)}; V)$. If

$$0 \leq \delta < \frac{1}{\sqrt{1 + \gamma}} \kappa(D)^{-1} \quad (6.2.15)$$

for $\kappa(D) = \|D\|_{V \to V'} \|D^{-1}\|_{V' \to V'}$ and $A^{(0)} \subseteq A^{(1)} \subseteq A$ with

$$\#A^{(1)} \leq \bar{c} \min \left\{ \#\hat{A} \ ; \ A^{(0)} \subseteq \hat{A} , \ \|\Pi_{\hat{A}}(f-\mathbb{A}v)\|_{L^2(A^{(0)}; V)} \geq \delta \|f-\mathbb{A}v\|_{L^2(A^{(0)}; V)} \right\} \quad (6.2.16)$$

for $\bar{c} \geq 1$, then

$$\#(A^{(1)} \setminus A^{(0)}) \leq \bar{c} \min \left\{ \#\hat{A} \ ; \ \hat{A} \subseteq A , \ \|u - u_{\hat{A}}\|_{\mathbb{R}} \leq \tau \|u - v\|_{\mathbb{R}} \right\} \quad (6.2.17)$$

for $\tau = \sqrt{1 - \delta^2(1 + \gamma)(1 - \gamma)^{-1}} \kappa(D)$.

**Proof.** Let $\hat{A}$ be as in (6.2.17) and $\hat{A} := A^{(0)} \cup \hat{A}$. Since $\hat{A} \subseteq \hat{A}, \|u - u_{\hat{A}}\|_{\mathbb{R}} \leq \|u - u_{\hat{A}}\|_{\mathbb{R}}$, and by Galerkin orthogonality,

$$\|u_{\hat{A}} - v\|^2_{\mathbb{R}} = \|u - v\|^2_{\mathbb{R}} - \|u - u_{\hat{A}}\|^2_{\mathbb{R}} \geq (1 - \tau^2) \|u - v\|^2_{\mathbb{R}} = \delta^2(1 + \gamma)(1 - \gamma)^{-1} \kappa(D) \|u - v\|^2_{\mathbb{R}} .$$

Therefore, using $\kappa(\mathbb{A}) = \|\mathbb{A}\| \|\mathbb{A}^{-1}\| \leq (1 + \gamma)(1 - \gamma)^{-1} \kappa(D)$,

$$\|\Pi_{\hat{A}}(f - \mathbb{A}v)\|_{L^2(A^{(0)}; V)} = \|\mathbb{A}(u_{\hat{A}} - v)\|_{L^2(A^{(0)}; V)} \geq \|\mathbb{A}^{-1}\|^{-1/2} \|u_{\hat{A}} - v\|_{\mathbb{R}} \geq \delta \|\mathbb{A}\|^{1/2} \|u - v\|_{\mathbb{R}} \geq \delta \|f - \mathbb{A}v\|_{L^2(A^{(0)}; V)} .$$

By (6.2.16), $\#(A^{(1)} \setminus A^{(0)}) \leq \bar{c} \#(\hat{A} \setminus A^{(0)})$, and consequently

$$\#(A^{(1)} \setminus A^{(0)}) \leq \bar{c} \#(\hat{A} \setminus A^{(0)}) \leq \bar{c} \#(\hat{A} \setminus A^{(0)}) \leq \bar{c} \#(\hat{A} \setminus A^{(0)}) .$$

\hfill \Box

119
Chapter 6. Galerkin Methods for Elliptic Problems

We use Lemma 6.2.6 to show that, under additional assumptions on the parameters, the index sets $A^{(k)}$ generated by SolveGalerkin$_{q,f}$ are of optimal size, up to a constant factor.

**Theorem 6.2.7.** If the conditions of Theorem 6.2.5 are satisfied,

$$\hat{\delta} := \frac{\delta(1 + \omega) + 2\omega}{1 - \omega} < \sqrt{\frac{1 - \gamma'}{1 + \gamma'}},$$

(6.2.18)

and $u \in \mathcal{A}(\Lambda; V_D)$ for an $s > 0$, then for all $k \in \mathbb{N}_0$ reached by SolveGalerkin$_{q,f}$,

$$\#A^{(k)} \leq \frac{2(\rho/\tau)^{1/s}}{1 - \rho^{1/s} \left(\frac{1 + \gamma}{1 - \gamma}(1 + \omega)\right)} \left\| u - \tilde{u}_{A^{(k)}} \right\|_{L^1} \left\| u \right\|_{L^1}^{1/s}$$

(6.2.19)

with $\rho = \sigma + \sqrt{1 - \hat{\delta}^2(1 + \gamma)(1 + \gamma')^{-1}}$ and $\tau = \sqrt{1 - \hat{\delta}^2(1 + \gamma)(1 - \gamma)^{-1}}$.

**Proof.** Let $k \in \mathbb{N}_0$, $v_k = u - \Pi_{A^{(k)}} u$, with corresponding $\rho$ and $\Delta$ as in Lemma 6.2.4. We note that for $\alpha := \omega + \delta + \omega \delta$, we have $\delta = \frac{\sigma - \omega}{1 + \omega}$ and $\hat{\delta} = \frac{\sigma + \omega}{1 - \omega}$. Let $A^{(k)} \subset \Lambda \subset \Lambda$ satisfy $\left\| \Pi_{A^{(k)}} u \right\|_{C(A; V_D)} \geq \hat{\delta} \left\| u \right\|_{C(A; V_D)}$. Then

$$\hat{\delta} \left\| \rho \right\|_{C(A; V_D)} \leq \hat{\delta} \left\| v_k \right\|_{C(A; V_D)} + \hat{\delta} \omega \left\| \rho \right\|_{C(A; V_D)} \leq \left\| \Pi_{A^{(k)}} v_k \right\|_{C(A; V_D)} + \hat{\delta} \omega \left\| \rho \right\|_{C(A; V_D)} \leq \left\| \Pi_{A^{(k)}} u \right\|_{C(A; V_D)} + (1 + \hat{\delta}) \omega \left\| \rho \right\|_{C(A; V_D)}$$

and since $\hat{\delta} - (1 + \hat{\delta}) \omega = \alpha$, it follows that $\left\| \Pi_{A^{(k)}} u \right\|_{C(A; V_D)} \geq \alpha \left\| u \right\|_{C(A; V_D)}$. By construction, $\Delta$ is a set of minimal cardinality with this property. Consequently, $\#(A^{(k+1)} \setminus A^{(k)}) \leq \#\Lambda \leq \#\Lambda$. Since this holds for any $\Lambda$, using $\#A^{(k)} \leq \#\Lambda$, it follows that

$$\#A^{(k+1)} \leq 2 \min \left\{ \#\Lambda; A^{(k)} \subset \Lambda, \left\| \Pi_{A^{(k)}} v_k \right\|_{C(A; V_D)} \geq \hat{\delta} \left\| v_k \right\|_{C(A; V_D)} \right\}.$$

Lemma 6.2.6 implies

$$\#(A^{(k+1)} \setminus A^{(k)}) \leq \min \left\{ \#\Lambda; A^{(k)} \subset \Lambda, \left\| u - u_{A^{(k)}} \right\|_{\mathbb{R}} \leq \tau \left\| u - \tilde{u}_{A^{(k)}} \right\|_{\mathbb{R}} \right\}$$

with $\tau = \sqrt{1 - \hat{\delta}^2(1 + \gamma)(1 - \gamma)^{-1}}$.

Let $N \in \mathbb{N}_0$ be maximal with $\left\| u - P_N(u) \right\|_{C(A; V_D)} > \tau(1 + \gamma)^{-1/2} \left\| u - \tilde{u}_{A^{(k)}} \right\|_{\mathbb{R}}$, where $P_N(u)$ is a best $N$-term approximation of $u$. By (5.2.33),

$$N + 1 \leq \left\| u - P_N(u) \right\|_{C(A; V_D)}^{1/s} \left\| u \right\|_{L^1}^{1/s} \leq \tau^{-1/s}(1 + \gamma)^{1/2} \left\| u - \tilde{u}_{A^{(k)}} \right\|_{\mathbb{R}}^{1/s} \left\| u \right\|_{L^1}^{1/s}.$$

For $A_{N+1} := \text{supp} P_{N+1}(u)$, by maximality of $N$,

$$\left\| u - u_{A_{N+1}} \right\|_{\mathbb{R}} \leq \left\| u - P_{N+1}(u) \right\|_{\mathbb{R}} \leq (1 + \gamma)^{1/2} \left\| u - P_{N+1}(u) \right\|_{C(A; V_D)} \leq \tau \left\| u - \tilde{u}_{A^{(k)}} \right\|_{\mathbb{R}}.$$
6.2. An Adaptive Iterative Solver

and thus

$$\#(\Lambda^{(k+1)} \setminus \Lambda^{(k)}) \leq 2(N + 1) \leq 2 \tau^{-1/s}(1 + \gamma)^{1/2s} \big\| u - \tilde{u}_{\Lambda^{(n)}} \big\|_\mathbb{Q}^{-1/s} \big\| u \big\|_{s^2(\Lambda;V_D)}^{1/s}$$

Furthermore, by Theorem 6.2.5,

$$\big\| u - \tilde{u}_{\Lambda^{(n)}} \big\|_\mathbb{Q}^{-1/s} \leq \left( \frac{1 - \gamma}{1 + \gamma} \right)^{-1/s}$$

We estimate the cardinality of $\Lambda^{(k)}$ by slicing it into increments and applying the above estimates,

$$\#\Lambda^{(k)} = \sum_{j=0}^{k-1} \#(\Lambda^{(j+1)} \setminus \Lambda^{(j)}) \leq 2 \tau^{-1/s}(1 + \gamma)^{1/2s} \big\| u \big\|_{s^2(\Lambda;V_D)}^{1/s} \sum_{j=0}^{k-1} \big\| u - \tilde{u}_{\Lambda^{(n)}} \big\|_\mathbb{Q}^{-1/s} \sum_{j=0}^{k-1} \delta_j^{-1/s}$$

By definition, $\delta_k \leq \rho^{k-1}\delta_j$. Therefore,

$$\sum_{j=0}^{k-1} \delta_j^{-1/s} \leq \delta_k^{-1/s} \sum_{j=0}^{k-1} \rho^{j-1/s} = \delta_k^{-1/s} \sum_{i=1}^{k} \rho^{i/s} = \frac{\rho^{1/s}\delta_k^{-1/s}}{1 - \rho^{1/s}}$$

The assertion follows using

$$(1 - \gamma)^{1/2} \big\| u - \tilde{u}_{\Lambda^{(n)}} \big\|_{s^2(\Lambda;V_D)} \leq \big\| u - \tilde{u}_{\Lambda^{(n)}} \big\|_\mathbb{Q} \leq \delta_k \ .$$

**Corollary 6.2.8.** Under the conditions of Theorem 6.2.7,

$$\#\Lambda^{(k)} \leq 2 \left( \frac{\rho/\tau}{1 - \rho^{1/s}} \right)^{1/s} \left( \frac{(1 + \gamma)(1 + \omega)}{(1 - \gamma)(1 - \omega)} \right)^{1/s} \kappa(D)^{1/2} \big\| u - \tilde{u}_{\Lambda^{(n)}} \big\|_{s^2(\Lambda;V)}^{-1/s} \big\| u \big\|_{s^2(\Lambda;V)}^{1/s} \tag{6.2.20}$$

with $\kappa(D) = \|D\|_{V \rightarrow V^*} \|D^{-1}\|_{V^* \rightarrow V}$.

**Proof.** The assertion follows from Theorem 6.2.7 using the norm equivalence

$$\|D^{-1}\|^{-1/2} \|v\|_V \leq \|v\|_{V_D} \leq \|D\|^{1/2} \|v\|_V \quad \forall v \in V \ .$$

**Remark 6.2.9.** Comparing Theorem 6.2.7 and Corollary 6.2.8, we note that the choice of norm on $V$ does not influence the condition (6.2.18) under which the cardinality of the sets $\Lambda^{(k)}$ is shown to be optimal. Rather, the norm on $V$ only affects the constant in the bounds (6.2.19) and (6.2.20).
Chapter 6. Galerkin Methods for Elliptic Problems

Lemma 6.2.10. Under the conditions of Theorem 6.2.7,
\[ \| u_k\|_{s^\prime(\Lambda; V)} \leq C \| u\|_{s^\prime(\Lambda; V)} \quad \forall k \in \mathbb{N}_0, \tag{6.2.21} \]
with
\[ C = 1 + \frac{2^{1+s}\rho(1 + \gamma)(1 + \omega)\kappa(D)^{1/2}}{\tau(1 - \rho^{1/2})(1 - \gamma)(1 - \omega)}, \tag{6.2.22} \]
where \( \rho = \sigma + \sqrt{1 - 8^2(1 - \gamma)(1 + \gamma)^{-1}} \) and \( \tau = \sqrt{1 - 8^2(1 + \gamma)(1 - \gamma)^{-1}}. \)

Proof. Let \( k \in \mathbb{N}_0 \). For any \( N \geq \#A^{(k)} \), \( \| \tilde{u}_{A^{(k)}} - P_N(\tilde{u}_{A^{(k)}})\|_{\ell^2(\Lambda; V)} = 0. \) For \( N \leq \#A^{(k)} - 1 \),
\[ \| \tilde{u}_{A^{(k)}} - P_N(\tilde{u}_{A^{(k)}})\|_{\ell^2(\Lambda; V)} \leq \| u - \Pi_N u\|_{\ell^2(\Lambda; V)} + 2 \| u - \tilde{u}_{A^{(k)}}\|_{\ell^2(\Lambda; V)}, \]
where \( \Pi_N := \text{supp } P_N(u) \), such that \( \Pi_N u = P_N(u) \) and
\[ \| u - \Pi_N u\|_{\ell^2(\Lambda; V)} \leq (N + 1)^{-s} \| u\|_{s^\prime(\Lambda; V)}. \]
Furthermore, Corollary 6.2.8 implies
\[ \| u - \tilde{u}_{A^{(k)}}\|_{\ell^2(\Lambda; V)} \leq \frac{2^s\rho(1 + \gamma)(1 + \omega)\kappa(D)^{1/2}}{\tau(1 - \rho^{1/2})(1 - \gamma)(1 - \omega)} \cdot \#A^{(k)} \cdot (N + 1)^{-s} \| u\|_{s^\prime(\Lambda; V)}, \]
and \( \#A^{(k)} \cdot (N + 1)^{-s} \leq (N + 1)^{-s} \) by definition of \( N \). Consequently,
\[ \| \tilde{u}_{A^{(k)}}\|_{s^\prime(\Lambda; V)} = \sup_{N \in \mathbb{N}_0} (N + 1)^{-s} \| \tilde{u}_{A^{(k)}} - P_N(\tilde{u}_{A^{(k)}})\|_{\ell^2(\Lambda; V)} \leq C \| u\|_{s^\prime(\Lambda; V)}, \]
with \( C \) from (6.2.22).

As in Section 5.2.2, we make additional assumptions on the routine \( \text{RHS}_t \). If \( t \in \mathcal{A}^s(\Lambda; V^*) \) and \( \tilde{t} \) is the output of \( \text{RHS}_t(e) \) for an \( e > 0 \), then \( \tilde{t} \) should satisfy
\[ \# \text{supp } \tilde{t} \leq \|t\|_{s^\prime(\Lambda; V^*)} e^{-1/s}. \tag{6.2.23} \]
Note that if \( u \in \mathcal{A}^s(\Lambda; V) \) and \( \mathcal{R} \) is \( s^* \)-compressible with \( s < s^* \), then also \( \mathcal{R} \) is \( s^* \)-compressible, and therefore \( \|u\|_{s^\prime(\Lambda; V^*)} \leq \|u\|_{\mathcal{A}^s(\Lambda; V^*)}. \)

Theorem 6.2.11. Let the conditions of Theorem 6.2.7 be satisfied. If \( \mathcal{R} \) is \( s^* \)-compressible with \( s^* > s \) and (6.2.23) holds, then for any \( e > 0 \) and any \( s \in (0, s^*) \), the total number of applications of \( D, A_V, \) and \( D^{-1} \) in \( \text{SolveGalerkin}_{\mathcal{R}, \mathcal{A}}(e, \chi, \delta, \omega, \sigma, \beta) \) is bounded by \( \|u\|_{s^\prime(\Lambda; V^*)} e^{-1/s} \) up to a constant factor depending only on the input arguments other than \( e \). The same bound holds for the total number of applications of \( R_m, m \in \mathbb{N}, \) up to an additional factor of \( \max_{t \in \text{supp } u} \# \text{supp } u \).
6.3. Fully Discrete Variant

Proof. Let $k \in \mathbb{N}_0$. The routine $\text{Residual}_{\text{Gal}}[\varepsilon \sqrt{1 - \gamma}, \tilde{u}_{A(0)}, \delta_k, \chi, \omega, \beta]$ begins with $\#A^{(k)}$ applications of $D$. Due to the geometric decrease in tolerances, the complexity of the following loop is dominated by that of the last iteration. By Remark 5.2.8 and Lemma 6.2.10, the number of applications of $D^{-1}$ and $R_m$ is bounded by $|u||_{1/\gamma; A(V)} \zeta^{-1/s}$, and $\zeta \geq \delta_k$.

Next, assuming the termination criterion of $\text{SolveGalerkin}_{\text{Gal}, f}$ is not satisfied, the routine $\text{Galerkin}_{\text{Gal}, f}[A^{(k+1)}, w, \sigma \min(\delta_k, \delta_f), \beta, \gamma]$ is called to iteratively approximate the Galerkin projection onto $A^{(k+1)}$. Since only a fixed relative error reduction is required, the number of iterations remains bounded. Therefore, by Proposition 6.1.5, the number of applications of $D^{-1}$ and $A_{\nu}$ is bounded by $\#A^{(k+1)}$ and the total number of applications of $R_m$, $m \in \mathbb{N}$, is bounded by $2\lambda(A^{(k+1)}) \#A^{(k+1)}$. Since the sets $A^{(k)}$ are nested, $\lambda(A^{(k+1)}) \leq \max_{\mu \in \text{supp } u_c} \# \text{supp } \mu$. Furthermore, by Theorem 6.2.5 and Corollary 6.2.8, $\#A^{(k+1)} \leq \|u\|_{1/\gamma; A(V)} \delta^{-1/s}_{k+1}$. Let $k$ be such that $u_c = \tilde{u}_{A(0)}$. Due to the different termination criterion, the complexity of the last call of $\text{Residual}_{\text{Gal}, f}$ can be estimated by $|u||_{1/\gamma; A(V)} \zeta^{-1/s}$ with $\zeta \geq \varepsilon$. This bound obviously also holds for $\#A^{(k)}$, and thus for the complexity of the final call of $\text{Galerkin}_{\text{Gal}, f}$.

Combining all of the above estimates, the number of applications of $D^{-1}$, $D$, $A_{\nu}$ and $R_m$, $m \in \mathbb{N}$, in $\text{SolveGalerkin}_{\text{Gal}, f}$ is bounded by

$$|u||_{1/\gamma; A(V)} \left( \varepsilon^{-1/s} + \sum_{j=0}^{k-1} \delta^{-1/s}_j \right).$$

Furthermore, $\delta_{k-1} \geq \varepsilon$, and using $\delta_{k-1} \leq \rho^{k-1-j}\delta_j$ for $\rho = \sigma + \sqrt{1 - \delta^2(1 - \gamma)(1 + \gamma)^{-1}} < 1$, \[
\sum_{j=0}^{k-2} \delta^{-1/s}_j \leq \delta^{-1/s}_{k-1} \sum_{j=0}^{k-2} \rho^{(k-1-j)/s} \leq \delta^{-1/s}_{k-1} \sum_{i=1}^{k-1} \rho^{i/s} \leq \delta^{-1/s}_{k-1} \frac{\rho^{1/s}}{1 - \rho^{1/s}}.
\]

The assertion follows since $\delta_{k-1} \geq \varepsilon$. \hfill  \qed

6.3. Fully Discrete Variant

6.3.1. Computation of the Residual

As in Section 5.3.2, we introduce a discretization of $V$ into $\text{SolveGalerkin}_{\text{Gal}, f}$ through a solver $\text{Solve}_D$ for $D$ such that for any $g \in V^*$ and any $\varepsilon > 0$,

$$\text{Solve}_D[g, \varepsilon] \mapsto v, \quad \|v - D^{-1}g\|_{V_D} \leq \varepsilon . \quad (6.3.1)$$

Such a method can be realized by various adaptive techniques.

We first consider a discrete version of the routine $\text{Residual}_{\text{Gal}, f}$, which approximates the residual up to a prescribed relative tolerance.
Chapter 6. Galerkin Methods for Elliptic Problems

\[ \text{Residual}_{\mathbb{R}, f}[\varepsilon, v, \eta_0, \chi, \omega, \alpha, \beta] \mapsto [w, \eta, \zeta] \]

\[ \zeta \leftarrow 2 \chi \eta_0 \]

repeat

\[ \zeta \leftarrow \zeta/2 \]

\[ \xi \leftarrow (1 - \alpha)\zeta \left\| D^{-1/2} \right\|_V \]

\[ b = (h_v)_{v \in \Lambda} \leftarrow \text{RHS}[(\beta \xi) - \text{Apply}_{\mathbb{R}}[v, (1 - \beta)\xi]] \]

\[ w = (w_v)_{v \in \Lambda} \leftarrow (\text{Solve}_{\mathbb{D}}[h_v, \alpha\zeta(\text{supp } b)^{-1/2}])_{v \in \Lambda} \]

\[ \eta \leftarrow \|w - v\|_{\mathcal{E}(\Lambda; V_D)} \]

until \( \zeta \leq \omega \eta \) or \( \eta + \zeta \leq \epsilon \)

**Proposition 6.3.1.** For any finitely supported \( v = (v_v)_{v \in \Lambda} \in \ell^2(\Lambda; V), \epsilon > 0, \eta_0 > 0, \chi > 0, \omega > 0, 0 < \alpha < 1 \) and \( 0 < \beta < 1 \), a call of \( \text{Residual}_{\mathbb{R}, f}[\varepsilon, v, \eta_0, \chi, \omega, \alpha, \beta] \) computes \( w \in \ell^2(\Lambda; V), \eta \geq 0 \) and \( \zeta \geq 0 \) with

\[ \left\| \eta - \|w\|_{\mathcal{E}(\Lambda; V_D)} \right\|_{\mathcal{E}(\Lambda; V_D)} = \left\| w - D^{-1}(f - \mathcal{R}v) \right\|_{\mathcal{E}(\Lambda; V_D)} \leq \zeta, \quad (6.3.2) \]

where \( \tau = (r_v)_{v \in \Lambda} \in \ell^2(\Lambda; V^*) \) is the residual \( \tau = f - \mathcal{R}v \), and \( \zeta \) satisfies either \( \zeta \leq \omega \eta \) or \( \eta + \zeta \leq \epsilon \).

**Proof.** By construction,

\[ \|b - (f - \mathcal{R}v)\|_{\mathcal{E}(\Lambda; V_D)} \leq \left\| D^{-1/2} \right\|_V \|b - (f - \mathcal{R}v)\|_{\mathcal{E}(\Lambda; V)} \leq (1 - \alpha)\zeta. \]

Furthermore, using \( \|w - D^{-1}b\|_{\mathcal{E}(\Lambda; V_D)} \leq \alpha\zeta,

\[ \left\| w - D^{-1}(f - \mathcal{R}v) \right\|_{\mathcal{E}(\Lambda; V_D)} \leq \left\| w - D^{-1}b \right\|_{\mathcal{E}(\Lambda; V_D)} + \|b - (f - \mathcal{R}v)\|_{\mathcal{E}(\Lambda; V_D)} \leq \zeta. \]

The rest of (6.3.2) follows by triangle inequality with \( ||\tau||_{\mathcal{E}(\Lambda; V_D)} = ||D^{-1}\tau||_{\mathcal{E}(\Lambda; V_D)} \).

**Remark 6.3.2.** In \( \text{Residual}_{\mathbb{R}, f} \), the tolerances of \( \text{Solve}_{\mathbb{D}} \) are chosen such that the error tolerance \( \alpha \zeta \) is equidistributed among all the nonzero indices of \( w \), see Remark 5.3.2. This property is not required anywhere; Proposition 6.3.1 only uses that the total error in the computation of \( D^{-1}b \) is no more than \( \alpha \zeta \). Indeed, other strategies for selecting tolerances, e.g. based on additional a priori information, may be more efficient. Equidistributing the error among all the indices is a simple, practical starting point.

**Remark 6.3.3.** For all \( v \in \Lambda \), let \( w_v \) be the Galerkin projection of \( h_v \) onto a space \( V_v \subset V \), i.e. the orthogonal projection in \( V_D \), and let \( v_v \in V_v \). Then for \( g_v := Dv_v \), by Galerkin orthogonality,

\[ \eta^2 = \sum_{v \in \Lambda} \|w_v - v_v\|_{V_v}^2 = \sum_{v \in \Lambda} \left\langle h_v - g_v, w_v - v_v \right\rangle_V, \quad (6.3.3) \]

which has the same form as the definition of \( \eta \) in Section 6.2.1.
6.3. Fully Discrete Variant

6.3.2. A Discrete Iterative Solver

For every \( \mu \in \Lambda \), let \( V_\mu \) be a closed subspace of \( V \). These constitute a closed subspace

\[
V := \left\{ v = (v_\mu)_{\mu \in \Lambda} \in \ell^2(\Lambda; V) \mid v_\mu \in V_\mu \right\}
\]  

(6.3.4)
of \( \ell^2(\Lambda; V) \). Let \( \Pi_V \) denote the orthogonal projection in \( \ell^2(\Lambda; V_D) \) onto \( V \). Since \( \mathfrak{D} \) is the Riesz isomorphism from \( \ell^2(\Lambda; V_D) \) to \( \ell^2(\Lambda; V_D^\gamma) \), \( \tilde{\Pi}_V := \mathfrak{D} \Pi_V \mathfrak{D}^{-1} \) is the orthogonal projection in \( \ell^2(\Lambda; V_D^\gamma) \) onto \( \mathfrak{D} V \).

Proposition 6.2.1, which is based on [CDD01, Lemma 4.1], generalizes to Galerkin projections on spaces \( V \) of the form (6.3.4).

**Proposition 6.3.4.** Let \( V \) be as in (6.3.4), and \( \delta \in [0, 1] \). Let \( v \in V \) with

\[
\| \tilde{\Pi}_V (f - \Lambda v) \|_{\ell^2(\Lambda; V_D^\gamma)} \geq \delta \| f - \Lambda v \|_{\ell^2(\Lambda; V_D^\gamma)} .
\]  

(6.3.5)

Then the Galerkin projection \( v_V \) of \( u \) onto \( V \) satisfies

\[
\| u - u_V \|_{\Omega} \leq \sqrt{1 - \frac{\gamma^2}{1 + \gamma^2}} \| u - v \|_{\Omega} .
\]  

(6.3.6)

**Proof.** The proof is analogous to that of Proposition 6.2.1. Due to (6.3.5),

\[
\| u_V - v \|_{\Omega} \geq \| \Omega \|^{-1/2} \| \Omega (u_V - v) \|_{\ell^2(\Lambda; V_D^\gamma)} \geq \| \Omega \|^{-1/2} \| \tilde{\Pi}_V (f - \Lambda v) \|_{\ell^2(\Lambda; V_D^\gamma)} \geq \| \Omega \|^{-1/2} \| f - \Lambda v \|_{\ell^2(\Lambda; V_D^\gamma)} \geq \| \Omega \|^{-1/2} \| \Omega \|^{-1/2} \delta \| u - v \|_{\Omega} .
\]  

By Galerkin orthogonality,

\[
\| u - u_V \|_{\Omega}^2 = \| u - v \|_{\Omega}^2 - \| u_V - v \|_{\Omega}^2 \leq (1 - \| \Omega \|^{-1} \| \Omega \|^{-1} \delta^2) \| u - v \|_{\Omega}^2 .
\]  

The assertion follows using the estimates \( \| \Omega \| \leq (1 + \gamma) \) and \( \| \Omega \|^{-1} \leq (1 - \gamma)^{-1} \) from Theorem 1.3.5. \( \square \)

Our discrete variant of \texttt{SolveGalerkin}_{\mu, t} uses a multiple level coarsening step similar to that in Section 5.3.3. Let \( V \) be of the form (6.3.4) with component spaces \( V_\mu \), and let \( w \in \ell^2(\Lambda; V) \) be finitely supported. For any target accuracy \( \eta > 0 \), let

\[
\text{Refine}_D(V, w, \eta) \mapsto V
\]  

(6.3.7)

construct a space \( V \) of the form (6.3.4) with component spaces \( \bar{V}_\mu \) such that \( V_\mu \subset \bar{V}_\mu \) for all \( \mu \in \Lambda \) and

\[
\| w - \bar{\Pi} \|_{\ell^2(\Lambda; V_D)} \leq \eta .
\]  

(6.3.8)

In principle, one could choose simply \( \bar{V}_\mu := V_\mu + \text{span} \, w_\mu \).

More practically, for each \( \mu \in \text{supp} \, w \), let \( V_\mu := V_\mu^0 \subset V_\mu^1 \subset \cdots \) be finite dimensional subspaces of \( V \). To each space, we associate the cost \( \text{dim} \, V_\mu \) and the error \( \| w_\mu - \bar{\Pi} w_\mu \|_{V_D^\gamma} \).
Chapter 6. Galerkin Methods for Elliptic Problems

where \( \Pi_{V_{\mu}} \) is the orthogonal projection in \( V_D \) onto \( V_{\mu} \). For any \( j = (j_\mu)_{\mu \in \text{supp } w} \), setting \( V_{\mu} := V_{j_\mu} \) leads to

\[
\dim V = \sum_{\mu \in \text{supp } w} \dim V_{j_\mu},
\]

and a total approximation error

\[
\|w - \Pi_V w\|^2_{\mathcal{E}(\Lambda; V_D)} = \sum_{\mu \in \text{supp } w} \left\| w_\mu - \Pi_{V_{j_\mu}} w_\mu \right\|^2_{V_D}.
\]

Minimizing the dimension (6.3.9) under the condition that the square (6.3.10) of the approximation error is less than \( \eta^2 \) is an optimization problem of the form studied in Section 4.1. As in Section 5.3.2, a solution can be computed efficiently by the recursive calls of the simple method NextOpt under the mild technical assumption

\[
\frac{\|w_\mu - \Pi_{V_{j_\mu}} w_\mu\|^2_{V_D} - \|w_{j_\mu} - \Pi_{V_{j_\mu}} w_{j_\mu}\|^2_{V_D}}{\dim V_{j_\mu + 1} - \dim V_{j_\mu}} \geq \frac{\|w_\mu - \Pi_{V_{j_\mu + 1}} w_\mu\|^2_{V_D} - \|w_{j_\mu + 1} - \Pi_{V_{j_\mu + 1}} w_{j_\mu + 1}\|^2_{V_D}}{\dim V_{j_\mu + 1} - \dim V_{j_\mu + 1}}
\]

if \( i \leq j \) for all \( \mu \in \text{supp } w \), see Proposition 4.1.6.

The routine Galerkin\_f from Section 6.1.2 easily generalizes from spaces \( \ell^2(\Lambda^0; V) \) to arbitrary \( V \) of the form (6.3.4). Applications of \( D^{-1} \) are simply replaced by the Galerkin projection onto the appropriate component space \( V_{\mu} \) of \( V \). We use the notation

\[
\text{Galerkin}_{\mathfrak{f}}[V, w, \epsilon, \delta, \gamma] \mapsto \tilde{u}_V
\]

for this generalization. For any \( w \in V \), \( \epsilon > 0 \) and \( 0 < \delta < 1 \), it computes \( \tilde{u}_V \in V \) satisfying

\[
\|\tilde{u}_V - u_V\|_{\mathfrak{f}} \leq \epsilon,
\]

where \( u_V \) is the Galerkin projection of \( u \) onto \( V \).

The discrete routines Residual\_f, Refine\_f and Galerkin\_f combine to form a discrete variant of the adaptive solver SolveGalerkin\_f.

**Lemma 6.3.5.** If \( \delta > 0 \), \( \omega > 0 \), and \( \omega + \delta + \omega \delta \leq 1 \), then the space \( V^{(k+1)} \) in SolveGalerkin\_f is such that

\[
\|\hat{r}_V^{(k+1)} r_k\|_{\mathcal{E}(\Lambda; V_{k+1}^+)} \geq \delta \|r_k\|_{\mathcal{E}(\Lambda; V_{k+1}^+)}
\]

where \( r_k := \hat{r}_V^{(k)} r_k \) is the residual at iteration \( k \in \mathbb{N}_0 \).

**Proof.** The proof follows that of Lemma 6.2.4. We abbreviate \( \hat{z} := w - \hat{u}_V^{(0)} \). By Proposition 6.3.1, using \( \eta = \|\hat{z}\|_{\mathcal{E}(\Lambda; V_D)^r} \)

\[
\|r_k\|_{\mathcal{E}(\Lambda; V_{k+1}^+)} \leq \eta + \zeta \leq (1 + \omega) \|\hat{z}\|_{\mathcal{E}(\Lambda; V_D)}.
\]
6.3. Fully Discrete Variant

\[ \text{SolveGalerkin}_{\mathbb{R},1}[\varepsilon, \gamma, \chi, \delta, \omega, \sigma, \alpha, \beta] \mapsto u_e \]

\[ \mathbf{V}^{(0)} \leftarrow [0] \]
\[ \bar{\mathbf{u}}_{\mathbf{V}^{(0)}} \leftarrow 0 \]
\[ \delta_0 \leftarrow \sqrt{(1 - \gamma)^{-1} \| D^{-1} \|_{V^\prime \to V} \| H \|^2_{C(\Lambda; V')}} \]

\text{forall} \ k \in \mathbb{N}_0 \ \text{do}

\[ \left[ w, \eta, \zeta \right] \leftarrow \text{Residual}_{\mathbb{R},1}[\varepsilon, \sqrt{1 - \gamma}, \bar{\mathbf{u}}_{\mathbf{V}^{(0)}}, \delta_k, \chi, \omega, \alpha, \beta] \]
\[ \delta_k \leftarrow (\eta + \zeta) / \sqrt{1 - \gamma} \]
\[ \text{if} \ \min(\delta_k, \bar{\delta}_k) \leq \varepsilon \ \text{then break} \]
\[ \mathbf{V}^{(k+1)} \leftarrow \text{Refine}_{\mathbb{D}}[\mathbf{V}^{(k)}, w, \eta \sqrt{1 - (\omega + \delta + \omega \delta)^2}] \]
\[ \bar{\mathbf{u}}_{\mathbf{V}^{(k+1)}} \leftarrow \text{Galerkin}_{\mathbb{R},1}[\mathbf{V}^{(k+1)}], \mathbf{c}^{(k+1)} w, \sigma \min(\delta_k, \bar{\delta}_k), \beta, \gamma \]
\[ \delta_{k+1} \leftarrow (\sigma + \sqrt{1 - \delta^2 (1 - \gamma)(1 + \gamma)^{-3}}) \min(\delta_k, \bar{\delta}_k) \]

\[ u_e \leftarrow \bar{\mathbf{u}}_{\mathbf{V}^{(0)}} \]

Since \( \bar{\mathbf{u}}_{\mathbf{V}^{(0)}} \in \mathbf{V}^{(k)} \), (6.3.8) implies

\[ \| \Pi_{\mathbf{V}^{(k+1)}} \|^2_{C(\Lambda; V)} = \| \Pi_{\mathbf{V}^{(k+1)}} \|^2_{C(\Lambda; V)} \geq (\omega + \delta + \omega \delta)^2 \| \mathbf{h} \|^2_{C(\Lambda; V)} . \]

Furthermore, since \( \Pi_{\mathbf{V}^{(k+1)}} \) has norm one, Proposition 6.3.1 implies

\[ \| \Pi_{\mathbf{V}^{(k+1)}} \|^2_{C(\Lambda; V)} = \| \Pi_{\mathbf{V}^{(k+1)}} \|^2_{C(\Lambda; V)} \leq \| \Pi_{\mathbf{V}^{(k+1)}}(\mathbf{h} - \mathbf{r}_k) \|^2_{C(\Lambda; V)} \]
\[ \leq \| \mathbf{h} - \mathbf{r}_k \|^2_{C(\Lambda; V)} \leq 2 \cdot \| \mathbf{h} - \mathbf{r}_k \|^2_{C(\Lambda; V)} . \]

Consequently, using \( \zeta \leq \omega \| \mathbf{h} \|^2_{C(\Lambda; V)} \),

\[ \| \mathbf{r}_k \|^2_{C(\Lambda; V)} \geq \| \Pi_{\mathbf{V}^{(k+1)}}(\mathbf{h} - \mathbf{r}_k) \|^2_{C(\Lambda; V)} \geq \delta \| \mathbf{r}_k \|^2_{C(\Lambda; V)} . \]

\[ \square \]

Theorem 6.2.5 extends to the discrete setting with no significant modifications. We provide the proof for completeness.

\textbf{Theorem 6.3.6.} If \( \varepsilon > 0, \chi > 0, \delta > 0, \omega > 0, \omega + \delta + \omega \delta \leq 1, 0 < \alpha < 1, 0 < \beta < 1 \) and \( 0 < \sigma < 1 - \sqrt{1 - \delta^2 (1 - \gamma)(1 + \gamma)^{-3}} \), then \( \text{SolveGalerkin}_{\mathbb{R},1}[\varepsilon, \gamma, \chi, \delta, \omega, \sigma, \alpha, \beta] \) constructs a finitely supported \( u_e \in \ell^2(\Lambda; V) \) with

\[ \| u - u_e \|_{\mathbb{R}} \leq \varepsilon . \]

Moreover,

\[ \sqrt{\frac{1 - \gamma}{1 + \gamma} 1 - \omega} \delta_k \leq \| u - \bar{\mathbf{u}}_{\mathbf{V}^{(0)}} \|_{\mathbb{R}} \leq \min(\delta_k, \bar{\delta}_k) \]

for all \( k \in \mathbb{N}_0 \) reached by \( \text{SolveGalerkin}_{\mathbb{R},1} \).
Chapter 6. Galerkin Methods for Elliptic Problems

Proof. Due to the termination criterion of \(\text{SolveGalerkin}_{\| \cdot \|}\), it suffices to show (6.3.16).

For \(k = 0\), since \(\| u \|_{C(\Omega; V)} \leq \| \mathcal{M}^{-1} \|_{1/2} \| u \|_{\mathcal{M}}\),

\[
\| u - \tilde{u}_{V(0)} \|_{\mathcal{M}}^2 = \| u \|_{\mathcal{M}}^2 = \langle f, u \rangle_{C(\Omega; V')} \leq \| \mathcal{M} \|_{C(\Omega; V')} \| u \|_{C(\Omega; V)} \leq \delta_0 \| u \|_{\mathcal{M}}.
\]

Let \(\| u - \tilde{u}_{V(0)} \|_{\mathcal{M}} \leq \delta_k\) for some \(k \in \mathbb{N}_0\). Abbreviating \(r_k := f - \mathcal{M}\tilde{u}_{V(0)}\), using (6.1.11) then (6.3.2), we have

\[
\| u - \tilde{u}_{V(0)} \|_{\mathcal{M}} \leq \frac{1}{\sqrt{1-\gamma}} \| r_k \|_{C(\Omega; V')} \leq \frac{\zeta + \eta}{\sqrt{1-\gamma}} = \delta_k.
\]

If \(\min(\delta_k, \delta_k) > \epsilon\), then \(\zeta \leq \omega \eta\) by Proposition 6.3.1. Due to Lemma 6.3.5, Proposition 6.3.4 implies

\[
\| u - u_{V(k+1)} \|_{\mathcal{M}} \leq \sqrt{1 - \frac{1-\gamma}{1+\gamma} \delta^2 \min(\delta_k, \delta_k)},
\]

where \(u_{V(k+1)}\) is the exact Galerkin projection of \(u\) onto \(V^{(k+1)}\). By (6.1.13), \(\tilde{u}_{V(k+1)}\) approximates \(u_{V(k+1)}\) up to an error of at most \(\sigma \min(\delta_k, \delta_k)\) in the norm \(\| \cdot \|_{\mathcal{M}}\). It follows by triangle inequality that \(\| u - \tilde{u}_{V(k+1)} \|_{\mathcal{M}} \leq \delta_{k+1}\).

To show the other inequality in (6.3.16), we note that for any \(k \in \mathbb{N}_0\),

\[
\| u - \tilde{u}_{V(0)} \|_{\mathcal{M}} \geq \frac{1}{\sqrt{1+\gamma}} \| r_k \|_{C(\Omega; V')} \geq \frac{\eta - \zeta}{\sqrt{1+\gamma}} = \sqrt{\frac{1-\gamma}{1+\gamma} \frac{\eta - \zeta}{\eta + \zeta}} \delta_k,
\]

and \((\eta - \zeta)(\eta + \zeta)^{-1} \geq (1 - \omega)(1 + \omega)^{-1}\).

Finally, since

\[
\delta_k \leq \left(\sigma + \sqrt{1 - \delta^2(1-\gamma)(1+\gamma)^{-1}}\right)^k \delta_0
\]

and \(\sigma + \sqrt{1 - \delta^2(1-\gamma)(1+\gamma)^{-1}} < 1\) by assumption, the iteration does terminate. \(\square\)
Chapter 7.
Computational Examples and Applications

As an application of the abstract theory in the previous chapters, we consider the isotropic diffusion equation with homogeneous Dirichlet boundary conditions and a parametric or stochastic diffusion coefficient. We illustrate in Section 7.1 how the abstract assumptions from Chapter 1 apply to boundary value problems depending on unknown coefficients. Furthermore, we present suitable spatial discretizations for the diffusion equation, including a residual-based a posteriori error estimator adapted to our needs. In a one dimensional setting, we consider a well-posed weak formulation of the diffusion equation with weakened integrability conditions on the forcing term. This requires Banach spaces, and thereby makes use of the generality of our abstract approach.

We present computational results on the adaptive methods from Chapters 5 and 6 applied to the diffusion equation in Section 7.3. Numerical experiments address mean square and uniform convergence, compare various versions of our adaptive solvers, describe index sets generated by these methods, and provide empirical estimates of achieved convergence rates. Details of the implementation are given in Section 7.2.

7.1. The Isotropic Diffusion Equation

7.1.1. A Differential Equation with Stochastic Coefficients

Let $\mathbb{K} = \mathbb{R}$ and let $G \subset \mathbb{R}^d$ be a bounded Lipschitz domain. For $a \in L^\infty(G)$ uniformly positive and $f \in L^2(G)$, we consider the isotropic diffusion equation on $G$ with homogeneous Dirichlet boundary conditions,

\begin{align}
-\nabla \cdot (a(x) \nabla u(x)) &= f(x), \quad x \in G, \\
\quad u(x) &= 0, \quad x \in \partial G.
\end{align}

We consider (7.1.1) to be a parametric operator equation in the sense of (1.2.1), with parameter domain

\begin{align}
Z := L^\infty(G) \times L^2(G).
\end{align}

For all $(a, f) \in Z$, we define the parametric operator

\begin{align}
A_0(a, f) : H^1_0(G) \to H^{-1}(G), \quad v \mapsto -\nabla \cdot (a \nabla v),
\end{align}

(7.1.3)
Chapter 7. Computational Examples and Applications

and the parametric functional \( f_0(a, f) := f \), which we interpret as an element of \( H^{-1}(G) \). Since \( u_0(a, f) \) only depends on \( a \) and \( f_0(a, f) \) only depends on \( f \), we abbreviate \( u_0(a) := u_0(a, f) \) and \( f_0(f) := f_0(a, f) \).

**Remark 7.1.1.** In the case of inhomogeneous boundary conditions \( u(x) = g(x), x \in \partial G \), the parametric functional \( f_0(a, f) \) does depend on \( a \). If \( g \in H^{1/2}(\partial G) \) and \( \tilde{g} \in H^1(G) \) is an extension of \( g \), then \( u_0 := u - \tilde{g} \) solves (7.1.1) with \( f \) replaced by \( f_0(a, f) := f - \nabla \cdot (aV \tilde{g}) \). The Dirichlet boundary data \( g \) may be added as a third parameter, in which case \( Z := L^\infty(G) \times L^2(G) \times H^{1/2}(\partial G) \).

**Lemma 7.1.2.** \( A_0 \) is a continuous linear map with

\[
\|A_0(a)\|_{H^1_0(G) \to H^{-1}(G)} \leq \|a\|_{L^\infty(G)}, \quad a \in L^\infty(G). \tag{7.1.4}
\]

**Proof.** It is clear from (7.1.3) that \( A_0 \) is linear. By Hölder’s inequality,

\[
\langle A_0(a)v, w \rangle = \left| \int_G a(x)Vv(x) \cdot Vw(x) \, dx \right| \leq \|a\|_{L^\infty(G)} \|v\|_{H^1_0(G)} \|w\|_{H^1_0(G)}
\]

for all \( v, w \in H^1_0(G) \) and \( a \in L^\infty(G) \). \( \square \)

Identifying \( H^{-1}(G) \) with the dual space of \( H^1_0(G) \), \( A_0 \) and \( f_0 \) are continuous maps from \( Z \) into \( \mathcal{L}(V, W^*) \) and \( W^* \), respectively, for \( V = W = H^1_0(G) \). Moreover, \( u_0(a) \) depends linearly on the parameter \((a, f) \in Z \), i.e. it is of the form (1.2.16) with \( A_0^0 = 0 \).

We model the parameter \((a, f) \in Z \) as a random variable on a probability space \((\Omega, \mathcal{F}, P)\),

\[
\hat{a}: (\Omega, \mathcal{F}) \to L^\infty(G), \quad \hat{f}: (\Omega, \mathcal{F}) \to L^2(G). \tag{7.1.5}
\]

The isotropic diffusion equation on \( G \) with a stochastic diffusion coefficient and a stochastic right hand side is

\[
A_0(\hat{a}(\omega))U(\omega) = f_0(\hat{f}(\omega)) \quad \forall \omega \in \Omega. \tag{7.1.6}
\]

The solution is a random variable \( U \) on \((\Omega, \mathcal{F})\) with values in \( H^1_0(G) \). For all \( \omega \in \Omega \), \( U(\omega) \) is the weak solution of (7.1.1) with \( a = \hat{a}(\omega) \) and \( f = \hat{f}(\omega) \).

Let \( \tilde{a}(\omega) \) be given by a series

\[
\tilde{a}(\omega) = \bar{a} + \sum_{m \in \mathcal{M}_d} Y_m(\omega)a_m, \quad \omega \in \Omega, \tag{7.1.7}
\]

with convergence in \( L^\infty(G) \), for a sequence \( Y_m := (Y_m)_{m \in \mathcal{M}_d} \) of bounded random variables on \((\Omega, \mathcal{F})\), as in (1.2.13). We assume without loss of generality that \( Y_m \) and \( a_m \) are normalized such that \( Y_m(\Omega) \subset [-1, 1] =: \Gamma_m \) for all \( m \in \mathcal{M}_d \). Let

\[
a(y) := \bar{a} + \sum_{m \in \mathcal{M}_d} y_m a_m, \quad y = (y_m)_{m \in \mathcal{M}_d} \in [-1, 1]^{\mathcal{M}_d}, \tag{7.1.8}
\]
7.1. The Isotropic Diffusion Equation

as in (1.2.14). Clearly, \( \bar{a}(\omega) = a(Y_0(\omega)) \) for all \( \omega \in \Omega \). We assume for the moment that the series in (7.1.8) converges in \( L^\infty(G) \) uniformly in \( y \in [-1,1]^d \), and \( \text{ess inf} \ a(y) > 0 \) for all \( y \in [-1,1]^d \). This is discussed in Section 7.1.2.

Define the parametric operator \( A(y) := A_0(a(y)) \) for \( y \in [-1,1]^d \). By the above assumptions, \( A(y) \in \mathcal{L}(H^1_0(G), H^{-1}(G)) \) is boundedly invertible for all \( y \in [-1,1]^d \). Due to the linearity of \( A_0 \), by (1.2.17),

\[
A(y) = D + \sum_{m \in \mathcal{M}} y_m R_m \quad \forall y \in [-1,1]^d \tag{7.1.9}
\]

with convergence in \( \mathcal{L}(H^1_0(G), H^{-1}(G)) \), for

\[
D = A_0(\bar{a}): H^1_0(G) \to H^{-1}(G) \ , \quad v \mapsto -\nabla \cdot (\bar{a} \nabla v) \ ,
\]

\[
R_m = A_0(a_m): H^1_0(G) \to H^{-1}(G) \ , \quad v \mapsto -\nabla \cdot (a_m \nabla v) \ , \quad m \in \mathcal{M}.
\]

Similarly, let \( \hat{f}(\omega) \) depend continuously on a separate sequence \( Y_f := (Y_m)_{m \in \mathcal{M}} \) of bounded random variables. We assume that these are normalized such that \( Y_m(\Omega) \subset [-1,1] =: \Gamma_m \) for all \( m \in \mathcal{M} \), and there is a continuous map \( f: [-1,1]^d \to L^2(G) \) such that \( \hat{f}(\omega) = f(Y_f(\omega)) \) for all \( \omega \in \Omega \).

We combine these two sequences to \( Y := (Y_m)_{m \in \mathcal{M}} \) for \( \mathcal{M} := \mathcal{M} \cup \mathcal{M} \), which we interpret as a map \( Y: \Omega \to \Gamma := [-1,1]^d \). Thus we have a parametric operator equation

\[
A(y)u(y) = f(y) \quad \forall y \in \Gamma \tag{7.1.10}
\]

with \( A(y) \) from (7.1.9). The solution is related to the solution \( U \) of (7.1.6) by \( U(\omega) = u(Y(\omega)) \) for all \( \omega \in \Omega \).

**Remark 7.1.3.** All of the above extends to the nonisotropic diffusion equation. Let \( S^d \) denote the space of symmetric \( d \times d \) matrices. For any \( S \in S^d \), let \( \lambda_{\min}(S) \) and \( \lambda_{\max}(S) \) be the smallest and largest eigenvalue of \( S \), respectively, and let the norm on \( S^d \) be defined as \( |S| := \max(-\lambda_{\min}(S), \lambda_{\max}(S)) \). Then the real line \( \mathbb{R} \) embeds isometrically into \( S^d \) as the span of the identity matrix. Replacing \( L^\infty(G) \) by \( L^\infty(G; S^d) \), the above discussion extends verbatim to arbitrary bounded symmetric diffusion coefficients.

### 7.1.2. Series Expansion of the Stochastic Coefficient

We assume that the stochastic diffusion coefficient \( \bar{a}(\omega) \) is uniformly bounded from above and away from 0,

\[
0 < \hat{a} \leq \bar{a}(\omega, x) \leq \hat{a} < \infty \quad \forall x \in G , \quad \forall \omega \in \Omega . \tag{7.1.11}
\]

Let \( \bar{a} \in L^\infty(G) \) be some deterministic approximation of \( \bar{a} \). For example, \( \bar{a} \) can be the mean field

\[
\bar{a}: G \to \mathbb{R} , \quad \bar{a}(x) := \int_\Omega \bar{a}(\omega, x) \, dP(\omega) , \tag{7.1.12}
\]
Chapter 7. Computational Examples and Applications

or simply a constant $\bar{a} := (\hat{a} + \tilde{a})/2$, $\bar{a} := \sqrt{\bar{a}\bar{b}}$, or $\bar{a} := 1$.

Let $(q_m)_{m \in \Omega}$ be a frame of $L^2(G)$ with dual frame $(q^*_m)_{m \in \Omega}$, which we interpret also as a sequence in $L^2(G)$ using the Riesz isomorphism. Define the random variables

$$Y_m(\omega) := \frac{1}{\alpha_m} \int_G (\tilde{a}(\omega, x) - \bar{a}(x)) q^*_m(x) \, dx, \quad m \in \mathcal{M}. \quad (7.1.13)$$

Note that $Y_m$ is bounded due to Hölder’s inequality and (7.1.11). As above, we assume that $\alpha_m$ is chosen such that $Y_m(\Omega) \subset [-1, 1]$ for all $m \in \mathcal{M}_a$. For example, this holds for

$$\alpha_m := \sup_{\omega \in \Omega} \|\tilde{a}(\omega) - \bar{a} \|_{L^\infty(G)} \|q^*_m\|_{L^1(G)} , \quad m \in \mathcal{M}. \quad (7.1.14)$$

We abbreviate

$$a_m := \alpha_m q_m , \quad m \in \mathcal{M}. \quad (7.1.15)$$

By (2.1.13),

$$\tilde{a}(\omega, x) = \bar{a}(x) + \sum_{m \in \mathcal{M}} Y_m(\omega) a_m(x) \quad (7.1.16)$$

for all $\omega \in \Omega$ with convergence in $L^2(G)$. Let

$$a(y, x) := \bar{a}(x) + \sum_{m \in \mathcal{M}} y_m a_m(x) , \quad y = (y_m)_{m \in \mathcal{M}} \in [-1, 1]^{\mathcal{M}}, \quad (7.1.17)$$

as in (7.1.8). Then $\tilde{a}(\omega, x) = a(Y_\omega(\omega), x)$ for all $\omega \in \Omega$, where $Y_\omega := (Y_m)_{m \in \mathcal{M}}$.

**Example 7.1.4.** Let $\tilde{a}$ be given by (7.1.12). The covariance operator of $\tilde{a}$ is the map $Q: L^2(G) \to L^2(G)$ given by

$$(Q\phi)(x) := \int_G k(x, x') \phi(x') \, dx' , \quad x \in G , \quad \phi \in L^2(G) , \quad (7.1.18)$$

with kernel $k: G \times G \to \mathbb{R}$,

$$k(x, x') := \int_\Omega (\tilde{a}(\omega, x) - \bar{a}(x))(\tilde{a}(\omega, x') - \bar{a}(x')) \, dP(\omega) , \quad x, x' \in G . \quad (7.1.19)$$

The operator $Q$ is symmetric, positive and nuclear. It therefore possesses a finite or countably infinite sequence $(\varphi_m)_{m \in \mathcal{M}}$ of eigenfunctions, which form an orthonormal system in $L^2(G)$, and a positive sequence $(\lambda_m)_{m \in \mathcal{M}}$ accumulating only at 0 such that

$$(Q\varphi_m)(x) = \int_G k(x, x') \varphi_m(x') \, dx' = \lambda_m \varphi_m(x) , \quad x \in G , \quad m \in \mathcal{M}. \quad (7.1.20)$$

For $(Y_m)_{m \in \mathcal{M}}$ defined by (7.1.13) with $q^*_m = \varphi_m$, (7.1.16) is the Karhunen–Loève series

$$\tilde{a}(\omega, x) = \bar{a}(x) + \sum_{m \in \mathcal{M}} Y_m(\omega) \sqrt{\lambda_m} \alpha_m \varphi_m(x) . \quad (7.1.21)$$

It converges in $L^2(G)$ for all $\omega \in \Omega$ and in $L^2_\mathcal{P}(\Omega; L^2(G))$. \hfill \qed
7.1. The Isotropic Diffusion Equation

Remark 7.1.5. More generally, one might apply a transformation to $a$ before expanding it in a series. For example, since the diffusion coefficient is positive, it seems natural to take the mean $\bar{a}$ and a series expansion of its logarithm. This leads to

$$a(\omega, x) = \exp \left( \bar{a}(x) + \sum_{m=1}^{\infty} Y_m(\omega) a_m(x) \right).$$  (7.1.22)

Finite expansions of this form are considered in [BNT07, Bie09a]. Foundations for the stochastic Galerkin method for a general expansion (7.1.22) with Gaussian $Y_m$ are developed in [Git10, SG11].

For all $m \in \mathcal{M}_a$, we denote the absolute value of $a_m$ by $|a_m|(x) := |a_m(x)|$.

Assumption 7.1.A. The series

$$\sum_{m \in \mathcal{M}_a} |a_m|$$  (7.1.23)

converges in $L^\infty(G)$.

Lemma 7.1.6. The map $a$: $[-1, 1]^{\mathcal{M}_a} \rightarrow L^\infty(G)$ given by (7.1.17) is well-defined and continuous with respect to the product topology on $[-1, 1]^{\mathcal{M}_a}$.

Proof. For all $y \in [-1, 1]^{\mathcal{M}_a}$, since $|y_m| \leq 1$

$$\|a(y)\|_{L^\infty(G)} \leq \|\bar{a}\|_{L^\infty(G)} + \left\| \sum_{m \in \mathcal{M}_a} y_m a_m \right\|_{L^\infty(G)} \leq \|\bar{a}\|_{L^\infty(G)} + \left\| \sum_{m \in \mathcal{M}_a} |a_m| \right\|_{L^\infty(G)},$$

which is finite by Assumption 7.1.A, so $a$ is well-defined.

A sequence $(y^n)_{n \in \mathbb{N}}$ converges to $y$ in $[-1, 1]^{\mathcal{M}_a}$ if $y^n_m \rightarrow y_m$ for all $m \in \mathcal{M}_a$. For any $\epsilon > 0$, let $\mathcal{M}_\epsilon \subset \mathcal{M}_a$ with $\mathcal{M}_a \setminus \mathcal{M}_\epsilon$ finite and

$$\left\| \sum_{m \in \mathcal{M}_\epsilon} |a_m| \right\|_{L^\infty(G)} \leq \epsilon.$$

Then let $n_\epsilon \in \mathbb{N}$ such that for all $n \geq n_\epsilon$,

$$\sum_{m \in \mathcal{M}_a \setminus \mathcal{M}_\epsilon} |y^n_m - y_m| \|a_m\|_{L^\infty(G)} \leq \epsilon.$$

Similarly to above, for all $n \geq n_\epsilon$, since $|y^n_m - y_m| \leq 2$ for all $m \in \mathcal{M}_\epsilon$,

$$\|a(y^n) - a(y)\|_{L^\infty(G)} \leq \sum_{m \in \mathcal{M}_\epsilon} (y^n_m - y_m)a_m \|a_m\|_{L^\infty(G)} + \left\| \sum_{m \in \mathcal{M}_a \setminus \mathcal{M}_\epsilon} (y^n_m - y_m)a_m \right\|_{L^\infty(G)} \leq 2 \left\| \sum_{m \in \mathcal{M}_\epsilon} |a_m| \right\|_{L^\infty(G)} + \sum_{m \in \mathcal{M}_a \setminus \mathcal{M}_\epsilon} |y^n_m - y_m| \|a_m\|_{L^\infty(G)} \leq 3\epsilon.$$

This shows continuity of $a$. \qed
Chapter 7. Computational Examples and Applications

By Lemma 7.1.2 and Lemma 7.1.6, the parametric operator $A(y)$ has the decomposition (7.1.9). Note that Assumption 7.1.A is less restrictive than the general condition (1.2.15). We abbreviate

$$R(y) := A_0 \left( \sum_{m \in A_d} y_m a_m \right) = \sum_{m \in A_d} y_m R_m .$$

(7.1.24)

Then $A(y) = D + R(y)$ for all $y \in [-1, 1]$.

**Assumption 7.1.B.** There is a constant $\delta > 0$ such that

$$\text{ess inf}_{x \in G} \bar{a}(x) \geq \delta^{-1} .$$

(7.1.25)

By the Lax–Milgram lemma, Assumption 7.1.B implies that $D = A_0(\bar{a})$ is boundedly invertible and the norm of $D^{-1}$ is at most $\delta$.

**Lemma 7.1.7.** For all $y \in [-1, 1]$, $\delta$

$$\|D^{-1} R(y)\|_{H^1_0(G) \to H^1_0(G)} \leq \delta \left\| \sum_{m \in A_d} |a_m| \right\|_{L^\infty(G)} =: \gamma .$$

(7.1.26)

**Proof.** By Lemma 7.1.2 and since $y_m \leq 1$ for all $m$,

$$\|R(y)\| = \left\| A_0 \left( \sum_{m \in A_d} y_m a_m \right) \right\| \leq \left\| \sum_{m \in A_d} y_m a_m \right\|_{L^\infty(G)} \leq \left\| \sum_{m \in A_d} |a_m| \right\|_{L^\infty(G)} .$$

The assertion follows since $\|D^{-1}\| \leq \delta$ by the Lax–Milgram lemma, and

$$\|D^{-1} R(y)\| \leq \|D^{-1}\| \|R(y)\| .$$

□

In particular, if $\gamma < 1$, then $A(y)$ is boundedly invertible for all $y \in [-1, 1]$ by Example 1.1.2. Note that the constant $\gamma$ given in (7.1.26) is smaller than in the general estimate (1.2.19).

7.1.3. Frame Representation

Let $G$ be a polytope, and let $\mathcal{X}_0$ be a regular simplicial mesh of $G$. For all $j \in \mathbb{N}$, let $\mathcal{X}_j$ be the mesh of $G$ constructed by $j$ regular refinements of $\mathcal{X}_0$. Denote by $\mathcal{M}_j$ the interior nodes of the mesh $\mathcal{X}_j$ and by $\mathcal{M}_j := \mathcal{M}_j \setminus \mathcal{M}_{j-1}$ the new nodes on discretization level $j \in \mathbb{N}$. On the coarsest level, we have $\mathcal{M}_0 := \emptyset$. Also, denote by $\mathcal{E}_j$ the set of edges of $\mathcal{X}_j$. Let $V_j := P_1(\mathcal{X}_j)$ be the space of continuous piecewise polynomials on the mesh $\mathcal{X}_j$. The standard basis of $V_j$ consists of the piecewise linear nodal basis functions

$$\lambda^j_m \in \mathcal{M}_j , \quad \delta_{nm} \forall m \in \mathcal{M}_j .$$

(7.1.27)

We will construct alternative bases of $V_j$ whose unions form frames $\Psi = \Theta$ of $H^1_0(G)$. These can be used to derive discrete operator equations for (7.1.10) as in Section 3.1.2.
7.1. The Isotropic Diffusion Equation

Example 7.1.8 (Finite element wavelets). We follow the construction in [DS99, NS09]. See also [BAS10] for an application of this basis to (7.1.1) and [Coh03] for a general overview of wavelets.

Define auxiliary functions \((\eta_{jn}^i)_{n \in \mathbb{N}}\) in \(V_j\) satisfying

\[
(\eta_{jn}^i, \lambda_{m}^{j-1})_{L^2(G)} = \delta_{nm} \left\| \eta_{jn}^i \right\|_{L^2(G)} \left\| \lambda_{m}^{j-1} \right\|_{L^2(G)},
\]

for all \(n, m \in \mathcal{N}_{j-1}\) and \(j \in \mathbb{N}\). For \(d = 1\),

\[
\eta_{jn}^i(m) := \begin{cases} 
3 & , m = n \\
-1/2 & , [m, n] \in \mathcal{E}_j \quad , n \in \mathcal{R}_{j-1}, \\
0 & , \text{all other } m \in \mathcal{R}_j
\end{cases}
\]

and for \(d = 2\),

\[
\eta_{jn}^i(m) := \begin{cases} 
14 & , m = n \\
-1 & , [m, n] \in \mathcal{E}_j \quad , n \in \mathcal{R}_{j-1}.
\end{cases}
\]

Define \(\tilde{\psi}_{m}^0 := \lambda_{m}^0\) for \(m \in \mathcal{R}_0\) and

\[
\tilde{\psi}_{m}^j := \lambda_{m}^j - \sum_{n \in \mathcal{R}_{j-1}} \left(\frac{(\eta_{jn}^i, \lambda_{m}^{j-1})_{L^2(G)}}{(\eta_{jn}^i, \lambda_{n}^{j-1})_{L^2(G)}}\right) \eta_{jn}^i, \quad m \in \mathcal{R}_j.
\]

Finally, we define the wavelets

\[
\psi_{m}^j := 2^{j(d-2)/2} \tilde{\psi}_{m}^j \in V_j, \quad m \in \mathcal{R}_j, \quad j \in \mathbb{N}_0.
\]

Then \(\Psi := (\psi_{m}^j)_{m \in \mathcal{R}_j, j \in \mathbb{N}_0}\) is a Riesz basis of \(H^1_0(G)\).

Example 7.1.9 (Multilevel frame). By [HSS08], the collection of properly scaled nodal basis functions on the meshes \(T_j\) form a frame of \(H^1_0(G)\), though not a Riesz basis. Define

\[
\psi_{n}^j := 2^{j(d-2)/2} \tilde{\psi}_{n}^j \in V_j, \quad n \in \mathcal{R}_j, \quad j \in \mathbb{N}_0.
\]

By [HSS08, Thm. 5], \(\Psi := (\psi_{n}^j)_{n \in \mathcal{R}_j, j \in \mathbb{N}_0}\) is a frame of \(H^1_0(G)\).

Remark 7.1.10. The constructions in Examples 7.1.8 and 7.1.9 can be generalized to piecewise polynomial finite elements of arbitrary degree and to more general subspaces of \(H^1(G)\). See the respective references for details.
Chapter 7. Computational Examples and Applications

7.1.4. Finite Element Approximation

By Example 1.1.2, the parametric operator $A(y)$ can be inverted by a Neumann series in which only the operator $D$ is inverted. Replacing $D^{-1}$ by some computable approximation leads to a numerical method for inverting $A(y)$. We consider general Galerkin approximations. Consider a scale of finite element spaces

\[ \{0\} = V_0 \subset V_1 \subset \ldots \subset V_j \subset \ldots \subset H^1_0(G) \]  
(7.1.34)

whose union is dense in $H^1_0(G)$. For each $j \in \mathbb{N}_0$, we define the approximation $S_j \in L(H^{-1}(G), H^1_0(G))$ of $D^{-1}$ as the Galerkin solution in $V_j$, i.e.

\[ S_j g := w_j \in V_j : \left< Dw_j, v_j \right>_{H^1_0(G)} = \left< g, v_j \right>_{H^1_0(G)} \quad \forall v_j \in V_j. \]  
(7.1.35)

By coercivity of the bilinear form $\left< D(\cdot), \cdot \right>_{H^1_0(G)}$ on $H^1_0(G)$,

\[ \|S_j\|_{H^{-1}(G) \rightarrow H^1_0(G)} \leq \|D^{-1}\|_{H^{-1}(G) \rightarrow H^1_0(G)} \leq \delta. \]  
(7.1.36)

7.1.5. A Posteriori Error Estimation

In the adaptive solvers with multilevel discretizations from Sections 5.3.2 and 6.3, a generic solver Solve$_D$ is used to approximate $D^{-1}g_\mu$ to any desired accuracy, where $g_\mu$ has the form

\[ g_\mu = f_\mu - \sum_{i=1}^{k} \kappa_i R_m w_i, \]  
(7.1.37)

with $w_i \in V = H^1_0(G)$ equal to some coefficients $\tilde{u}_i$ of the previous approximate solution, see Proposition 3.2.11. If $D^{-1}g_\mu$ is approximated by the finite element method, an a posteriori error estimator is required to determine whether or not a given approximation attains the desired accuracy. Due to the unusual structure of $g_\mu$, standard error estimators cannot be applied directly. We derive a reliable residual-based estimator, following the standard argument from [MNS00, AO00, Ver96].

Let $\mathcal{T}$ be a regular mesh of $G$, and let $V_\mu$ be a finite element space of continuous, piecewise smooth shape functions on $\mathcal{T}$ which contains at least the piecewise linear functions.

We will denote the set of elements of $\mathcal{T}$ by $\mathcal{T}$ and the set of faces of $\mathcal{T}$ by $\mathcal{F}$. The set $\mathcal{T}$ can be decomposed into interior faces $\mathcal{T} \cap G$ and boundary faces $\mathcal{T} \cap \partial G$. For any $T \in \mathcal{T}$, let $h_T$ be the diameter of $T$, and similarly, define $h_F$ as the diameter of $F$ for any $F \in \mathcal{F}$. Furthermore, for any $T \in \mathcal{T}$, let $\mathcal{V}_T \subset G$ consist of all elements of $\mathcal{T}$ sharing at least a vertex with $T$. Analogously, let $\mathcal{V}_F \subset G$ consist of all elements of $\mathcal{T}$ sharing at least a vertex with the face $F \in \mathcal{F}$. Note that each element $T \in \mathcal{T}$ belongs to only a bounded number of domains $\mathcal{V}_T$ or $\mathcal{V}_F$. 

136
7.1. The Isotropic Diffusion Equation

By the above assumptions, there is a Clément interpolant for $V_{\mu}$, i.e. a continuous projection $\mathcal{I}_T : H^1_0(G) \rightarrow V_{\mu}$ such that for all $u \in H^1_0(G)$,

$$\|u - \mathcal{I}_T u\|_{L^2(T)} \leq c_1 h_T |u|_{H^1(T)} \quad \forall T \in \mathcal{I} \quad (7.1.38)$$

and

$$\|u - \mathcal{I}_T u\|_{L^2(F)} \leq c_2 h_T^{1/2} |u|_{H^1(T)} \quad \forall F \in \mathcal{F} \quad (7.1.39)$$

with constants $c_1$ and $c_2$ depending only on the shape regularity of $\mathcal{I}$, see e.g. [BS02].

Let each of the functions $w_i$ from (7.1.37) itself be an element of a finite element space $V_i$ of piecewise smooth functions on a mesh $\mathcal{I}_i$, which may differ from $\mathcal{I}$. We assume that these meshes are compatible in the sense that for any $T \in \mathcal{I}$ and $T_i \in \mathcal{I}_i$, the intersection $T \cap T_i$ is either empty, equal to $T$, or equal to $T_i$.

Standard error estimators run into problems on faces of $\mathcal{I}_i$ that are not in the skeleton of $\mathcal{I}$, since $g_\mu$ is singular on these faces. For all $i$, let $\bar{w}_i$ be an approximation of $w_i$ that is piecewise smooth on $\mathcal{I}$. Replacing $g_\mu$ by

$$\bar{g}_\mu := f_\mu - \sum_{i=1}^k \kappa_i R_{\mu_i} \bar{w}_i \quad (7.1.40)$$

induces an error

$$\|D^{-1} g_\mu - D^{-1} \bar{g}_\mu\|_D \leq \sum_{i=1}^k |\kappa_i| \|\frac{\alpha_m}{\alpha_0}\|_{L^\infty(G)} \|w_i - \bar{w}_i\|_D =: \text{EST}_D^p \quad (7.1.41)$$

where $\|v\|_D := \sqrt{(Dv, v)}$ is the norm on $H^1_0(G)$ induced by $D$, since

$$\sup_{\|v\|_D = 1} \left| \int_G a_\mu \nabla w \cdot \nabla v \, dx \right| \leq \left\| \frac{\alpha_m}{\alpha_0} \right\|_{L^\infty(G)} \sup_{\|v\|_D = 1} \int_G |a\nabla w \cdot \nabla v| \, dx = \left\| \frac{\alpha_m}{\alpha_0} \right\|_{L^\infty(G)} \|w\|_D$$

for all $m \in \mathbb{N}$ and all $w \in H^1_0(G)$.

Let $\bar{u}_\mu \in V_{\mu}$ be the Galerkin projection of $D^{-1} \bar{g}_\mu$, i.e.

$$\int_G a \nabla \bar{u}_\mu \cdot \nabla v \, dx = \int_G f_\mu v \, dx - \sum_{i=1}^k \kappa_i \int_G a_\mu \nabla \bar{w}_i \cdot \nabla v \, dx \quad \forall v \in V_{\mu} \quad (7.1.42)$$

Abbreviating

$$\sigma_\mu := a \nabla \bar{u}_\mu + \sum_{i=1}^k \kappa_i a_\mu \nabla \bar{w}_i \quad ,$$

the residual of $\bar{u}_\mu$ is the functional

$$r_\mu(\bar{u}_\mu; v) = \int_G \bar{g}_\mu - a \nabla \bar{u}_\mu \cdot \nabla v \, dx = \int_G f_\mu - \sigma_\mu \cdot \nabla v \, dx , \quad v \in H^1_0(G) \quad (7.1.44)$$
Chapter 7. Computational Examples and Applications

By Galerkin orthogonality, \( r_\mu(\tilde{u}_\mu; v) = 0 \) for all \( v \in V_\mu \). Furthermore, due to the Riesz isomorphism,

\[
\left\| D^{-1} \mathbf{g}_\mu - \tilde{u}_\mu \right\|_D = \sup_{v \in H^1_0(G) \setminus \{0\}} \left| \frac{r_\mu(\tilde{u}_\mu; v)}{\|v\|_D} \right| \leq \sqrt{\delta} \sup_{v \in H^1_0(G) \setminus \{0\}} \left| \frac{r_\mu(\tilde{u}_\mu; v)}{\|v\|_{H^1(G)}} \right| ,
\]

with \( \delta \) from Assumption 7.1.B.

For all \( T \in \mathcal{I} \), let

\[
R_{\mu,T}(\tilde{u}_\mu) := h_T \left\| f_\mu + \nabla \cdot \sigma_\mu \right\|_{L^2(T)} ,
\]

where the dependence on \( \tilde{u}_\mu \) is implicit in \( \sigma_\mu \). Note that \( \nabla \cdot \sigma_\mu \) is given by

\[
\nabla \cdot \sigma_\mu = \nabla \tilde{u} \cdot \nabla \tilde{u}_\mu + \tilde{u} \Delta \tilde{u}_\mu + \sum_{i=1}^k \kappa_i (\nabla a_{m_i} \cdot \nabla \tilde{w}_i + a_{m_i} \Delta \tilde{w}_i) .
\]

Also, let

\[
R_{\mu,F}(\tilde{u}_\mu) := h_T^{1/2} \left\| \sigma_\mu \right\|_{L^2(F)} ,
\]

where \( \left\| \cdot \right\| \) is the normal jump over the face \( F \in \mathcal{F} \cap \partial G \), i.e. if \( F = T_1 \cap T_2 \), and \( n_1 \) and \( n_2 \) are the respective exterior normal vectors, then

\[
\left\| \sigma_\mu \right\| := \sigma_\mu |_{T_1} \cdot n_1 + \sigma_\mu |_{T_2} \cdot n_2 ,
\]

and \( \left\| \sigma_\mu \right\| := \sigma_\mu \cdot n_G \) if \( F \in \mathcal{F} \cap \partial G \) for the exterior unit normal \( n_G \) of \( G \). These terms combine to

\[
\text{EST}^R_\mu(\tilde{u}_\mu) := \left( \sum_{T \in \mathcal{I}} R_{\mu,T}(\tilde{u}_\mu)^2 + \sum_{F \in \partial G} R_{\mu,F}(\tilde{u}_\mu)^2 \right)^{1/2} .
\]

Note that if \( d = 1 \), then \( h_T = 0 \) for all \( F \in \mathcal{F} \), and \( R_{\mu,F}(\tilde{u}_\mu) = 0 \). In this case, the Clément interpolation operator \( \mathcal{I}_\mu \) is simply the nodal interpolant, and \( \tilde{\omega}_T \) can be replaced by \( T \) in (7.1.38).

**Theorem 7.1.11.** For all \( v \in H^1_0(G) \),

\[
\left| r_\mu(\tilde{u}_\mu; v) \right| \leq C \text{EST}^R_\mu(\tilde{u}_\mu) \|v\|_{H^1(G)}
\]

with a constant \( C \) depending only on the shape regularity of \( \mathcal{I} \).

**Proof.** Let \( v \in H^1_0(G) \). Since \( \mathcal{I}_\mu v \in V_\mu \), by Galerkin orthogonality

\[
r_\mu(\tilde{u}_\mu; v) = r_\mu(\tilde{u}_\mu; v - \mathcal{I}_\mu v) .
\]

We abbreviate \( v_\delta := v - \mathcal{I}_\mu v \), and denote by \( n_T \) the exterior unit normal of \( T \in \mathcal{I} \). Using (7.1.38), (7.1.39), integration by parts and the Cauchy–Schwarz inequality,

\[
r_\mu(\tilde{u}_\mu; v - \mathcal{I}_\mu v) = \sum_{T \in \mathcal{I}} \int_T f_\mu v_\delta - \sigma_\mu \cdot \nabla v_\delta \, dx
\]
Corollary 7.1.12. The Galerkin projection \( \bar{u}_\mu \) from (7.1.42) satisfies
\[
\|D^{-1}g_\mu - \bar{u}_\mu\|_D \leq \text{EST}^p_\mu + \sqrt{\delta \text{EST}^R_\mu(\bar{u}_\mu)}
\] (7.52)
for \( \delta \) from Assumption 7.1.B and C from Theorem 7.1.11.

Proof. The assertion follows by triangle inequality using (7.1.41), (7.1.45) and (7.1.51).

Transforming the estimate (7.1.52) to the standard norm on \( H^1_0(G) \), we have
\[
\|D^{-1}g_\mu - \bar{u}_\mu\|_{H^1(G)} \leq \sqrt{\delta \text{EST}^p_\mu + \delta \text{EST}^R_\mu(\bar{u}_\mu)}.
\] (7.53)
A slightly different estimate is obtained by estimating (7.1.41) directly in this norm.

7.1.6. Weak Formulation on Banach Spaces

We consider the diffusion equation on the interval \( I := (0,1) \) with mixed boundary conditions,
\[
-(a(y,x)u'(y,x))' = f(y,x) , \quad x \in I , \quad y \in \Gamma ,
\]
\[
u(y,0) = u'(y,1) = 0 , \quad y \in \Gamma ,
\] (7.54)
where all the derivatives are with respect to the variable \( x \in I \). As in (7.1.17), we assume that \( a(\cdot, \cdot) \) is of the form
\[
a(y,x) = a(x) + \sum_{m \in \mathcal{A}} a_m(x) y_m , \quad y = (y_m)_{m \in \mathcal{A}} \in \Gamma ,
\] (7.55)
Chapter 7. Computational Examples and Applications

with \( \tilde{a} \in L^\infty(G) \) and \( \Gamma = [-1, 1]^d \). Furthermore, we assume

\[
\delta := \left( \text{ess inf}_{x \in G} \tilde{a}(x) \right)^{-1} < \infty .
\] (7.1.56)

Define the Sobolev spaces

\[
W_{00}^{1,p}(I) := \left\{ v \in W^{1,p}(I) ; v(0) = 0 \right\} , \quad 1 < p < \infty ,
\] (7.1.57)

with norms

\[
\|v\|_{W_{00}^{1,p}(I)} := \|v\|_{W^{1,p}(I)} = \left( \int_0^1 |v'(\xi)|^p \, d\xi \right)^{1/p} , \quad v \in W_{00}^{1,p}(I) .
\] (7.1.58)

Let \( V := W_{00}^{1,p}(I) \) for a \( 1 < p < \infty \) and \( W := W_{00}^{1,q}(I) \) for the exponent \( q \) conjugate to \( p \).

Lemma 7.1.13. The operator

\[
D : W_{00}^{1,p}(I) \to (W_{00}^{1,q}(I))^*, \quad v \mapsto -(\tilde{a}v')' ,
\] (7.1.59)

is an isomorphism with

\[
\|D\|_{W_{00}^{1,p}(I) \to (W_{00}^{1,q}(I))^*} \leq \|\tilde{a}\|_{L^\infty(G)} ,
\] (7.1.60)

\[
\|D^{-1}\|_{(W_{00}^{1,q}(I))^* \to W_{00}^{1,p}(I)} \leq \delta .
\] (7.1.61)

Proof. By Hölder’s inequality, for all \( v \in W_{00}^{1,p}(I) \) and \( w \in W_{00}^{1,q}(I) \),

\[
|\langle Dv, w \rangle| = \left| \int_0^1 \tilde{a}v'w' \, dx \right| \leq \|\tilde{a}\|_{L^\infty(G)} \|v\|_{W_{00}^{1,p}(I)} \|w\|_{W_{00}^{1,q}(I)} .
\]

This shows (7.1.60). Let \( v \in W_{00}^{1,p}(I) \) and define

\[
w(x) := \int_0^x \text{sign}(v'(\xi)) |v'(\xi)|^{p/q} \, d\xi , \quad x \in I .
\] (7.1.62)

Then \( |w'(x)|^q = |v'(x)|^p \) and in particular \( w \in W_{00}^{1,q}(I) \). Furthermore,

\[
\delta \langle Dv, w \rangle = \delta \int_0^1 \tilde{a} |v'|^{1+\frac{p}{q}} \, dx \geq \int_0^1 |v'|^p \, dx
\]

\[
= \left( \int_0^1 |v'|^p \, dx \right)^{1/p} \left( \int_0^1 |v'|^p \, dx \right)^{1/q} = \|v\|_{W^{1,p}(I)} \|w\|_{W^{1,q}(I)} .
\]

This shows

\[
\inf_{0 \neq v \in W_{00}^{1,p}(I)} \sup_{0 \neq w \in W_{00}^{1,q}(I)} \frac{\langle Dv, w \rangle}{\|v\|_{W^{1,p}(I)} \|w\|_{W^{1,q}(I)}} \geq \delta^{-1} .
\]
7.1. The Isotropic Diffusion Equation

Exchanging \( v \) and \( w \) as well as \( p \) and \( q \) in the above argument leads to
\[
\inf_{0 \neq v \in W_0^1(I)} \sup_{0 \neq w \in W_0^1(I)} \frac{\langle Dv, w \rangle}{\|v\|_{W^1_q(I)} \|w\|_{W^1_p(I)}} \geq \delta^{-1}.
\]
This implies that \( D \) is boundedly invertible and \( D^{-1} \) satisfies (7.1.61). \( \square \)

Similarly, if \( a_m \in L^\infty(G) \), the operators
\[
R_m: W_0^1(I) \to (W_0^1(I))^*, \quad v \mapsto -(a_m v')',
\]
are continuous with norm less than \( \|a_m\|_{L^\infty(I)} \) for all \( m \in M \). Assuming these norms are summable, as in (1.3.11), the operator
\[
R(y) := A(y) - D = \sum_{m \in M} y_m R_m
\]
is well-defined in \( \mathcal{L}(W_0^1(I), (W_0^1(I))^*) \).

**Lemma 7.1.14.** If
\[
\gamma := \esssup_{x \in I} \delta \sum_{m \in M} |a_m(x)| < 1,
\]
then (1.3.16) holds with this \( \gamma \) and \( V = W_0^1(I) \).

**Proof.** For all \( y = (y_m)_{m \in M} \), \( v \in W_0^1(I) \) and \( w \in W_0^1(I) \), by (7.1.63) and (7.1.64),
\[
\langle R(y) v, w \rangle = \int_0^1 \sum_{m \in M} y_m a_m(x)v'(x)w'(x) \, dx.
\]
Since \( |y_m| \leq 1 \) for all \( m \in M \), it follows by Hölder's inequality that
\[
\|R(y)\|_{W_0^1(I) \to (W_0^1(I))^*} \leq \esssup_{x \in I} \sum_{m \in M} |a_m(x)|.
\]
Then the assertion follows from Lemma 7.1.13 using
\[
\|D^{-1} R(y)\|_{W_0^1(I) \to W_0^1(I)} \leq \|D^{-1}\|_{(W_0^1(I))^* \to W_0^1(I)} \|R(y)\|_{W_0^1(I) \to (W_0^1(I))^*}. \quad \square
\]

Let \( V_0 := \{0\} \) and let
\[
V_1 \subset \cdots \subset V_j \subset V_{j+1} \subset \cdots \subset W_0^1(I)
\]
be the spaces of continuous piecewise linear finite elements on nested meshes of \( I \). For each \( j \in \mathbb{N}_0 \), we define the approximation \( S_j \in \mathcal{L}((W_0^1(I))^*, W_0^1(I)) \) of \( D^{-1} \) as the Galerkin solution in \( V_j \), i.e. for all \( g \in (W_0^1(I))^* \),
\[
S_j g := w_j \in V_j : \quad \langle Dw_j, v_j \rangle = \langle g, v_j \rangle \quad \forall v_j \in V_j,
\]
as in (7.1.35).
Lemma 7.1.15. For all $j \in \mathbb{N}_0$,
\[
\|S\|_{(W_0^1)^0(\gamma) \rightarrow (W_0^1)^0(I)} \leq \delta.
\] (7.1.68)

Proof. The assertion follows as in the proof of Lemma 7.1.13 since if $v \in V_j$, then $v'$ is piecewise constant and $w$ defined by (7.1.62) is also in $V_j$. □

7.2. Elements of a Matlab Implementation

7.2.1. Data Structure for the Index Set

For polynomial basis functions as in Section 3.2, assuming $\Lambda^m = \mathbb{N}_0$ for all $m \in \mathbb{N}$, the index set $\Lambda$ consists of all finitely supported sequences in $\mathbb{N}_0$. Elements $\mu \in \Lambda$ can be encoded by two vectors, one of which lists the $m \in \mathbb{N}$ with $\mu_m \neq 0$, and the other stores the values $\mu_m$. Assuming that the first vector is stored in ascending order, a nd contains exactly supp $\mu$ with no superfluous $m \in \mathbb{N}$, this representation is unique. In a matrix-based programming language such as Matlab, these two vectors can be stored together in a single matrix. In the following, we do not distinguish between $\mu \in \Lambda$ and a finite encoding of $\mu$ as a matrix or a pair of vectors.

Generally, the objects $\mu \in \Lambda$ cannot be used directly as indices. Rather, they must be mapped onto distinct integers, which can in turn be used to index an array.

Example 7.2.1. The index set $\Lambda$ can be identified with $\mathbb{N}$ through prime factorization. Let $p = (p_m)_{m \in \mathbb{N}}$ be the prime numbers, enumerated e.g. in increasing order. Then the map
\[
\Lambda \rightarrow \mathbb{N}, \quad \mu \mapsto p^\mu = \prod_{m \in \text{supp } \mu} p_m^{\mu_m} = \prod_{m \in \text{supp } \mu} p_m^{\mu_m}
\] (7.2.1)
is a bijection. However, it is of little practical use as $p^\mu$ is very large already for $\mu \in \Lambda$ with moderate supp $\mu$. Even considering only the first dimension, $\mu_1$ may not exceed 31 if $2^{\mu_1}$ is to be a valid 32 bit unsigned integer.

A more useful approach than Example 7.2.1 is to enumerate the elements of a given finite $\Lambda^0 \subset \Lambda$. Allowing the integer pointer $i_\mu \in \mathbb{N}$ associated to a given $\mu \in \Lambda$ to depend on the set $\Lambda^0$, the values $\{i_\mu; \mu \in \Lambda^0\}$ can always be chosen as $\{1, \ldots, \#\Lambda^0\}$. The map from $i_\mu$ to $\mu \in \Lambda^0$ can be realized simply by listing the indices $\mu$. The opposite map, from $\mu$ to $i_\mu$, requires a more sophisticated data structure. We consider two examples below and refer to e.g. [CLRS09] for further details.

Example 7.2.2 (Hash Table). A practical choice of data structure is a hash table, which typically requires $\mathcal{O}(\#\text{supp } \mu)$ operations to look up $i_\mu$, given $\mu \in \Lambda^0$. An unspecified function $h$ maps $\Lambda^0$ into a block of memory. Assuming that the memory is sufficiently large compared to $\#\Lambda^0$, $h$ can be assumed to map most $\mu \in \Lambda^0$ to distinct memory locations. The hash table stores a pointer to $\mu$ and $i_\mu$ at $h(\mu)$. Thus a lookup operation typically consists of evaluating $h(\mu)$, checking that the index $\mu'$ at $h(\mu)$ is indeed equal to $\mu$, and returning $i_{\mu'}$. The comparison with $\mu'$ is necessary because $h$ may map some
7.2. Elements of a Matlab Implementation

\( \mu, \mu' \in \Lambda^\circ \) onto the same memory location. In this case, the values can be stored in a linked list, which must be searched for the desired \( \mu \) in a lookup operation. Assuming that the hash function manages to avoid collisions, or more generally that the linked lists are of bounded size, a lookup operation consists of one evaluation of \( h(\mu) \) and a fixed number of comparisons with \( \mu \). The latter have complexity \( O(\# \text{supp } \mu) \), and the optimal complexity of the evaluation \( h(\mu) \) is also \( O(\# \text{supp } \mu) \) if the hash function is to distinguish between distinct \( \mu \in \Lambda^\circ \).

As of version 2008b, Matlab provides a native hash table data structure in form of the `map` container. However, it does not directly support matrices as keys. We recast \( \mu \) as a character array, which is supported, and use this hash table implementation. The character arrays provide sixteen bit precision, which limits both \( m \) and \( \mu_m \) to \( 2^{16} - 1 = 65535 \).

**Example 7.2.3 (Trie).** A trie is a tree data structure for vectors of arbitrary length, which stores these vectors in an overlapping format. We illustrate this data structure by an example. The table below enumerates the five elements \( \mu \) of a set \( \Lambda^\circ \subset \Lambda \); they are assumed to have \( \mu_m = 0 \) for \( m \geq 3 \). To the right, these indices are encoded in a trie. Values \( \mu_m \) are shown in circles, and the dimensions \( m \) are depicted on arrows connecting these circles in a tree. If all of the values leading to a given circle in this tree form an element \( \mu \) of \( \Lambda^\circ \), the value \( i_\mu \) is appended, shown here in a square.

<table>
<thead>
<tr>
<th>( \mu_1 )</th>
<th>( \mu_2 )</th>
<th>( i_\mu )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1 0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1 8</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>0 5</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>2 3</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

As the above example illustrates, exactly \( 2^{\# \text{supp } \mu} \) integer comparisons are required for a lookup operation in a trie. In comparison, a balanced tree storing a complete \( \mu \) in each node requires \( O(\# \text{supp } \mu \log(\# \Lambda^\circ)) \) integer comparisons. Also, the trie data structure does not need any rebalancing, and it is more memory efficient than a tree. Note that there may be arbitrarily many arrows emanating from a given node. We suggest storing these in an array of arrays, the first indexed by the next dimension \( m \), and the second by the following value \( \mu_m \). This allows \( i_\mu \) to be determined by just \( 2^{\# \text{supp } \mu} \) array lookups. The arrays of pointers are full if the dimensions \( m \) and values \( \mu_m \) are filled in increasing order, which is generally the case for our purposes. For more general sets \( \Lambda^\circ \) or more dynamic operations, the arrows starting at the same node can be stored in a tree or a hash table, which slightly increases the cost of a lookup operation.

As discussed in Section 6.1.2, computing the Galerkin projection requires access to the neighbors in \( \Lambda^\circ \) of a \( \mu \in \Lambda^\circ \). These can be stored in a directed graph, with an edge
from \( \mu \) to \( \nu \) if there is an \( m \in \mathbb{N} \) such that \( \nu = \mu - \epsilon_m \). Such a graph can be encoded as a sparse matrix, with a nonzero value \( m \) at \((i_\nu, i_\mu)\) in the above situation. Its construction uses just
\[
\sum_{\mu \in \Lambda^0} \# \text{supp } \mu = \bar{\lambda}(\Lambda^0) \# \Lambda^0
\] lookup operations, however \( \Lambda^0 \) is encoded. By Lemma 6.1.4, this is optimal if \( \Lambda^0 \) is monotonic.

### 7.2.2. Multilevel Finite Elements

We consider a piecewise linear finite element discretization of the diffusion equation from Section 7.1 in one dimension. For simplicity, we stick to uniform dyadic meshes of an arbitrary interval, and use the standard hat function basis.

In principle, the assembly of the stiffness matrices and load vectors is straightforward. To avoid quadrature errors, and since diffusion coefficients may be oscillatory, we use exact formulas to evaluate all integrals. For example, up to a sign, a generic component of the stiffness matrix has the form
\[
\int_{x_i}^{x_{i+1}} \frac{a(x)}{(x_{i+1} - x_i)^2} \, dx = \frac{1}{(x_{i+1} - x_i)^2} \int_{x_i}^{x_{i+1}} a(x) \, dx . \tag{7.2.3}
\]
Similarly, using integration by parts, elements of the load vector can be computed by e.g.
\[
\int_{x_i}^{x_{i+1}} f(x) \frac{x - x_i}{x_{i+1} - x_i} \, dx = F(x_{i+1}) - \frac{1}{x_{i+1} - x_i} \int_{x_i}^{x_{i+1}} F(x) \, dx , \tag{7.2.4}
\]
where \( F' = f \). We assume that all antiderivatives appearing in (7.2.3) and (7.2.4) are given, and use these formulas in the assembly procedure.

Using dyadic grids ensures that the finite element spaces associated to these grids are nested. Coefficient vectors can easily be transferred between levels. To project a vector from a fine grid to a coarser grid, we simply interpolate the function, which amounts to taking only every \( 2^j \)-th element of the vector, if the coarse grid is \( j \) levels below the fine grid. Prolonging a vector from level \( j \) to level \( j + 1 \) consists of inserting the arithmetic mean of every two successive coefficients between these. Both operations are easily vectorized, and execute quickly in Matlab.

In our adaptive solvers from Chapters 5 and 6, it is not clear initially which matrices are needed on which discretization levels. Therefore, assembly must be done during the solution procedure. We use Matlab’s handle classes to store all matrices and load vectors for future use whenever they are computed. Using an object oriented approach allows for an elegant syntax and easily understandable code.

An operator is encoded as a struct, with fields for all function handles required in its assembly. The struct also contains an infinite list of the matrices representing the operator on all discretization levels. This list is realized as a handle class, and can be indexed by an arbitrary integer. Internally, it contains a function handle to a function that assembles the matrix on any level, and a cell array for storing matrices that have
7.2. Elements of a Matlab Implementation

already been computed. The first time a matrix is used, it is assembled and stored in
the cell array, from where it is recovered on every subsequent call. As a handle class,
this list is always passed by reference in Matlab. This hides the assembly from the user,
and allows completely disjoint codes using the same operator to automatically share
matrices.

Load vectors are implemented analogously, using the same list class. We distinguish
between load vectors that are assembled directly by a function handle, load vectors
that are given by an operator applied to a vector, and load vectors that are linear
combinations of other load vectors. Each has its own assembly routines, but all follow
the same basic structure, and store every vector for future use the first time it is accessed.

Finite element vectors are also stored as structs, with a field for the level of dis-
cretization and a field for the vector of coefficients. Of course, the level could also be
determined from the length of the coefficient vector.

To a certain extent, the above abstraction hides the discretization from the user. For
example, when an operator is applied to a vector, no computations are performed
initially. Rather, a load vector structure is created, as described above. When the
coefficients of this vector are required on a certain level, the matrix representation of the
operator on this level is accessed, the finite element vector is prolonged or interpolated
to this level, and the two are multiplied.

7.2.3. Implementation of the Adaptive Solvers

Our implementation of the adaptive solvers from Chapters 5 and 6 differ slightly from
the original formulations in that we aim to make the error bounds as sharp as pos-
sible. We always use the a posteriori data on the residual, as in e.g. Remark 5.1.2
and Remark 5.1.5. However, rather than using tolerances such as \( \beta \delta_0 \) in (5.1.33), we
use the actual error bounds computed in the routines Apply, RHS, Coarsen, Solve, Residual, Refine,
Galerkin and PCG. These are guaranteed to be smaller than
their respective tolerances.

In all of our numerical experiments, the operator \( D \) is the Dirichlet Laplacian. There-
fore, \( \| \cdot \|_D \) coincides with the standard norm on \( H^1_0(G) \), and \( D \) and \( D^{-1} \) have unit norm
as maps between \( V \) and \( W^* \). Our codes always use \( \| \cdot \|_D \) as the norm on \( V = W \).

We make several changes to the a posteriori error estimator from Section 7.1.5. As
already mentioned, \( R_{\mu,F}(\bar{u}_\mu) = 0 \) for all \( F \in \mathcal{F} \) since \( d = 1 \). Also, \( \Delta \bar{u}_\mu = 0 \) and \( \Delta \bar{w}_i = 0 \)
for all \( i \) in (7.1.47) since \( \bar{u}_\mu \) and \( \bar{w}_i \) are piecewise linear. Moreover, for \( D \) equal to the
Laplacian, \( \bar{a} = 1 \) is constant, and thus \( R_{\mu,T}(\bar{u}_\mu) \) is actually independent of \( \bar{u}_\mu \).

Furthermore, we approximate the integrand

\[
 f_\mu + \nabla \cdot \sigma_\mu = f_\mu + \sum_{i=1}^k \kappa_i a'_{m_i} \bar{w}_i
\]

in (7.1.46) by a constant, and estimate the error incurred by this approximation. On an
element \( T = [x_i, x_{i+1}] \), let \( f_\mu \) denote the mean of \( f_\mu \), and let \( a'_{m_i} \) be the mean of \( a'_{m_i} \) on \( T \).
Chapter 7. Computational Examples and Applications

By triangle inequality, since $\bar{a}_i$ is constant on $T$,

$$\|f_\mu + \nabla \cdot a_\mu\|_{L^2(T)} \leq \left| \left| \left| f_\mu + \sum_{i=1}^{k} \kappa_i a'_m \bar{a}'_i \right| \right|_{L^2(T)} + \left| \left| \left| f_\mu - f_\mu\right| \right|_{L^2(T)} + \sum_{i=1}^{k} |\kappa_i\bar{a}'_i| \|a'_m - \bar{a}'_m\|_{L^2(T)} \right| .$$

(7.2.6)

Since the integrand is constant on $T$, the value

$$\left| \left| \left| f_\mu + \sum_{i=1}^{k} \kappa_i a'_m \bar{a}'_i \right| \right|_{L^2(T)} = h^{1/2}_T \left| \left| \left| f_\mu + \sum_{i=1}^{k} \kappa_i a'_m \bar{a}'_i \right| \right|_{L^2(T)} \right|$$

(7.2.7)

is simple to compute. Since the mean value $\tilde{f}_\mu$ is the orthogonal projection of $f_\mu$ onto constants in $L^2(T)$, we have the explicit bound

$$\|f_\mu - f_\mu\|_{L^2(T)} \leq \frac{h_T}{2\sqrt{3}} \|f'_\mu\|_{L^2(T)} ,$$

(7.2.8)

and similarly for $\|a'_m - \bar{a}'_m\|_{L^2(T)}$. The constant $2\sqrt{3}$ can be determined by expanding $f_\mu$ with respect to the Legendre polynomial basis and comparing the two sides of (7.2.8). Therefore,

$$R_{\mu,T}(\bar{a}_\mu) \leq \tilde{R}_{\mu,T} := h^{3/2}_T \left| \left| \left| f_\mu + \sum_{i=1}^{k} \kappa_i a'_m \bar{a}'_i \right| \right|_{L^2(T)} + \frac{h^2_T}{2\sqrt{3}} \left( \left| \left| \left| f'_\mu \right| \right|_{L^2(T)} + \sum_{i=1}^{k} |\kappa_i\bar{a}'_i| \|a''_m\|_{L^2(T)} \right| \right) ,$$

(7.2.9)

and thus, with no further constants,

$$\text{EST}_{\mu,T}(\bar{a}_\mu) \leq \left( \sum_{t \in \mathbb{Z}} \tilde{R}_{\mu,T}^2 \right)^{1/2} .$$

(7.2.10)

Similarly, we can compute the constant in Theorem 7.1.11. We use that $3_{\mu}$ is just the nodal interpolant in the one dimensional setting, and expand the derivative $v'$ from Theorem 7.1.11 with respect to Legendre polynomials on $T$. Following the proof of Theorem 7.1.11, we arrive at $C = 1/\sqrt{3}$. Consequently, Corollary 7.1.12 with $\delta = 1$ due to $\bar{a} = 0$ implies

$$\|D^{-1}g_\mu - \bar{a}_\mu\|_{L^2(T)} \leq \text{EST}_{\mu,T}^p + \frac{1}{\sqrt{3}} \left( \sum_{t \in \mathbb{Z}} \tilde{R}_{\mu,T}^2 \right)^{1/2} .$$

(7.2.11)

All of the remaining terms can be computed exactly, assuming the appropriate antiderivatives of $f_\mu$ and $a_m$ are available.

The routine $\text{Solve}_D$ consists of computing the right hand side of (7.2.11) on successive refinement levels, until the desired accuracy is reached, and then computing the Galerkin projection on this level using Matlab’s backslash solver. The error bound is guaranteed with no unknown factors or higher order terms.
7.3. Numerical Experiments

7.3.1. Model Problem

We consider as a model problem the diffusion equation (7.1.1) on the one dimensional domain $G = (0, 1)$. For two parameters $k$ and $\gamma$, the diffusion coefficient has the form

$$a(y, x) = 1 + \frac{1}{c} \sum_{m=1}^{\infty} y_m \frac{1}{m^k} \sin(m\pi x), \quad x \in (0, 1), \quad y \in \Gamma = [-1, 1]^N,$$

(7.3.1)

where $c$ is chosen as

$$c = \gamma \sum_{m=1}^{\infty} \frac{1}{m^k},$$

(7.3.2)

such that $|a(y, x) - 1|$ is always less than $\gamma$. We consider the countable product of uniform distributions on $[-1, 1]$; the corresponding family of orthonormal polynomials is the Legendre polynomial basis, see Example 3.2.4.

A few realizations of $a(y)$ are plotted in Figure 7.1 for $k = 4$ and $\gamma = 5/6$, along with the resulting solutions $u(y)$ of (7.1.1). Unless stated otherwise, we use these parameters. For comparison, we also consider $k = 2$ and $\gamma = 1/2$.

![Figure 7.1: Realizations of $a(y, x)$ (left) and $u(y, x)$ (right).](image)

7.3.2. Comparison of Methods

Figure 7.2 illustrates the convergence behavior of SolveDirect for $k = 4$ and $\gamma = 5/6$. The solver parameters are set to $\alpha = 1/36$, $\beta_1 = 1/18$, and $\beta_0 = 0$, since the right hand side can be evaluated exactly. Solid lines in the plots refer to the error bounds $\delta$, which are available during the computation. Dashed lines represent estimates of the actual errors in $L_2(\Gamma; V)$, computed by comparison to a reference solution with tolerance $10^{-6}$.

We consider a single level spatial discretization, using linear finite elements on a uniform mesh of $(0, 1)$ with 1024 elements to approximate all coefficients, and a multilevel
discretization in which the a posteriori error estimator from Section 7.1.5 is used to determine an appropriate discretization level for each coefficient, see also Section 7.2.3. A discretization level \( j_\mu \), which represents linear finite elements on a uniform mesh with \( 2^j \) cells, is assigned to each index \( \mu \) with the goal of equidistributing the estimated error among all coefficients.

On the left, the errors are plotted against the number of degrees of freedom, which refers to total number of basis functions used in the discretization, \textit{i.e.} 1023 times the number of active indices for the single level method, and the sum of \( 2^j - 1 \) over all \( \mu \) for the multilevel method. On the right, we plot the errors against an estimate of the computational cost. This estimate takes scalar products, matrix-vector multiplications and linear solves into account. The total number of each of these operations on each discretization level is tabulated during the computation, weighted by the number of degrees of freedom on the discretization level, and summed over all levels. The estimate is equal to seven times the resulting sum for linear solves, plus three times the value for matrix-vector multiplications, plus the sum for scalar products. These weights were determined empirically by timing the operations for tridiagonal sparse matrices in Matlab.

We note that the discretizations generated by the adaptive multilevel method are an order of magnitude more efficient than the single level method, with respect to the total number of degrees of freedom. The difference is less substantial when considering the estimated computational cost. Since indices \( \mu \) with high discretization levels \( j_\mu \) are likely to also have many neighbors, it is to be expected that the multilevel method performs many multiplication operations on fine meshes. Also, it requires some operations to estimate errors. As can be seen in the figures, the error component due to the finite element discretization is ignored by the single level method. Consequently, as this term becomes dominant, the method cedes to converge.

Figure 7.3 shows the convergence of SolveCoarse_{8;5}, with the same multilevel discretization as above. The parameters are the same as for SolveDirect_{8;5}, with \( \delta = 5/12 \),...
and $\chi$ equal to 1/5 or 1/2. In the former case, the conditions of Theorem 5.2.10 are satisfied, and thus the complexity of the solver is optimal, up to a constant factor. However, the method performs substantially better for the larger value $\chi = 1/2$. We therefore assume this value when referring to $\text{SolveCoarse}_{\text{Gf}}$ below.

Figure 7.4 depicts the convergence behavior of $\text{SolveGalerkin}_{\text{Gf}}$ with the multilevel linear finite element discretization in space. Even though for this solver the error estimate $\delta$ is an upper bound for the error in the energy norm, it is apparently also a good estimate of the error in $L^2_\pi(I; V)$.

We use $\chi = 5/72$, $\delta = 0.57$, $\omega = 1/4$, $\sigma = 0.0029758$, $\alpha = 1/36$ and $\beta = 0$. With these parameters, the assumptions of Theorem 6.2.7 are not satisfied. The method failed to converge in reasonable time with parameters satisfying condition (6.2.18).

The convergence curves in Figure 7.4 refer to two variants of $\text{SolveGalerkin}_{\text{Gf}}$. The first, with coarsening, is essentially as described in Chapter 6, up to the comments in Section 7.2.3. The second variant, without coarsening, skips the refinement step.
Instead, the entire support of $w$ is used as the next index set. In both cases, the value of $\vartheta$ is decreased to the minimal value consistent with the error in the estimation of the residual. Evidently, the former strategy is superior, and we use it in the following.

We compare the convergence of several solvers in Figure 7.5. All of them use multilevel linear finite elements as a spatial discretization. The parameters for the method SolveAlternate$_{A,f}$ are the same as for SolveDirect$_{A,f}$, and the parameters of the other solvers are given above.

For large error tolerances, SolveDirect$_{A,f}$ and SolveAlternate$_{A,f}$ seem to be most efficient in terms of the computational cost, with a slight advantage for the latter method. As the error tolerance is decreased, SolveGalerkin$_{A,f}$ catches up to these methods. For all tolerances, SolveCoarse$_{A,f}$ has the highest computational cost.

In terms of the number of degrees of freedom in the approximate solutions, for this example SolveCoarse$_{A,f}$ and SolveGalerkin$_{A,f}$ are the most efficient methods, in particular when considering the error bound $\delta$ as a measure of the error. Interestingly, the difference is much less pronounced when considering the actual error in $L^2_\pi(\Gamma; V)$. This suggests that the discretizations generated by SolveDirect$_{A,f}$ and SolveAlternate$_{A,f}$ are almost as efficient as those generated by methods with coarsening steps, but that, without coarsening, the error bounds are less accurate.

Figure 7.5.: Comparison of convergence behavior for $k = 4$ and $\gamma = 5/6$. 
7.3. Numerical Experiments

Table 7.1.: Convergence Statistics

<table>
<thead>
<tr>
<th></th>
<th>dofs (#)</th>
<th>A_N (10^3)</th>
<th>time (sec.)</th>
<th>cost (10^6)</th>
<th>δ error (10^-6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SolveDirect_{R,f}</td>
<td>1006</td>
<td>1004</td>
<td>649</td>
<td>4.30</td>
<td>8.29</td>
</tr>
<tr>
<td></td>
<td>222</td>
<td>263</td>
<td>1302</td>
<td>16.36</td>
<td>9.41</td>
</tr>
<tr>
<td>SolveAlternate_{R,f}</td>
<td>1034</td>
<td>966</td>
<td>406</td>
<td>3.79</td>
<td>8.96</td>
</tr>
<tr>
<td>SolveCoarse_{R,f}</td>
<td>285</td>
<td>301</td>
<td>6103</td>
<td>91.34</td>
<td>5.55</td>
</tr>
<tr>
<td></td>
<td>448</td>
<td>358</td>
<td>1798</td>
<td>24.85</td>
<td>5.62</td>
</tr>
<tr>
<td>SolveGalerkin_{R,f}</td>
<td>346</td>
<td>219</td>
<td>586</td>
<td>1.73</td>
<td>9.53</td>
</tr>
<tr>
<td></td>
<td>1912</td>
<td>512</td>
<td>2340</td>
<td>10.64</td>
<td>8.95</td>
</tr>
</tbody>
</table>

Some convergence statistics for \( k = 4, \gamma = 5/6 \) and a target accuracy of \( 10^{-5} \) are listed in Table 7.1. We consider a modified version of \( \text{SolveDirect}_{R,f} \), which first computes a solution with an error tolerance \( 1/5 \) of the desired accuracy, and then coarsens. Although the resulting approximation uses the fewest degrees of freedom, this method is by no means the most efficient. For \( \text{SolveCoarse}_{R,f} \) and \( \text{SolveGalerkin}_{R,f} \), the two variants described above are considered.

In this example, \( \text{SolveAlternate}_{R,f} \) is fastest in terms of execution time, and the method \( \text{SolveGalerkin}_{R,f} \) with coarsening has the lowest estimated computational cost. With respect to both measures, only \( \text{SolveDirect}_{R,f} \) without coarsening is competitive. It should be noted that \( \text{SolveGalerkin}_{R,f} \) is unique in being efficient both in terms of computational cost and the number of degrees of freedom.

Figure 7.6 compares the adaptive solvers for the same model problem with \( k = 2 \) and \( \gamma = 1/2 \). The parameters of \( \text{SolveDirect}_{R,f} \), \( \text{SolveAlternate}_{R,f} \), and \( \text{SolveCoarse}_{R,f} \) are \( \alpha = 1/20, \beta_0 = 0, \beta_1 = 1/10 \) and, for the last method, \( \delta = 1/4 \) and \( \chi = 1/2 \). For \( \text{SolveGalerkin}_{R,f} \), we have \( \chi = 1/8, \delta = 0.57, \omega = 1/4, \sigma = 0.01114, \alpha = 1/20 \) and \( \beta = 0 \), which do not satisfy the assumptions of Theorem 6.2.7.

In this example, \( \text{SolveGalerkin}_{R,f} \) performs less favorably, with a computational cost that is slightly higher even than that of \( \text{SolveCoarse}_{R,f} \). As above, \( \text{SolveCoarse}_{R,f} \) and \( \text{SolveGalerkin}_{R,f} \) are most efficient in terms of the number of degrees of freedom in the approximate solutions, while \( \text{SolveDirect}_{R,f} \) and \( \text{SolveAlternate}_{R,f} \) are most efficient with respect to the computational cost. The errors of the latter two methods with respect to the number of degrees of freedom are not only larger, the convergence rates also appear to be slightly lower.

7.3.3. Uniform Convergence

As discussed in Section 5.1.3, the method \( \text{SolveDirect}_{R,f} \) and its extensions can be modified to control the error in \( C(\Gamma; V) \) rather than \( L^2(\Gamma; V) \). As this space is free of any measure on the parameter domain \( \Gamma \), we may choose any tensor product polynomial basis. In the following, we consider Legendre polynomials and Chebyshev polynomials.
Chapter 7. Computational Examples and Applications

Figure 7.6.: Comparison of convergence behavior for $k = 2$ and $\gamma = 1/2$. 
Numerical Experiments

7.3. Numerical Experiments

Figure 7.7.: Convergence of SolveDirect_{\mathcal{Q},f} with error controlled in $C(\Gamma; V)$. Solid lines are error estimates $\delta$, dashed lines are estimates of the actual errors.

of the first kind.

Numerically evaluating the maximum norm on the domain $\Gamma$ is nontrivial. To estimate errors in $C(\Gamma; V)$, we compute the lower bound given by the maximal error over the 64 points $\{-1, 1\}^6$ in $\Gamma$, extended by zero.

Figure 7.7 shows the convergence of SolveDirect_{\mathcal{Q},f} with the error controlled in $C(\Gamma; V)$. We consider both Legendre and Chebyshev polynomial bases. The solid lines represent the error bound $\delta$, which is available during computation, and the dashed lines refer to the lower bound for the maximal error on $\Gamma$. In all cases, the two error estimates converge at the same rate. Estimating the error in $C(\Gamma; V)$ by the triangle inequality, as is done in this method to compute $\delta$, does not appear to cause problems. The flattening of the convergence curve in the case of Legendre polynomials and error control in $C(\Gamma; V)$ suggests a slight numerical instability of the Legendre basis. However, we did not observe this in other computations.

Figure 7.8 compares the uniform convergence of SolveDirect_{\mathcal{Q},f} with the error controlled in $L^2_\pi(\Gamma; V)$ and in $C(\Gamma; V)$. Again, we consider both Legendre and Chebyshev polynomials. The choice of basis has almost no effect when the error is controlled in the maximum norm. In comparison, the method with an error target in a space $L^2_\pi(\Gamma; V)$ converges somewhat slower in the case of Legendre polynomials, and only marginally so for the Chebyshev basis.

In Figure 7.9, we also compare SolveDirect_{\mathcal{Q},f} with the error controlled in $L^2_\pi(\Gamma; V)$ and in $C(\Gamma; V)$, but here we consider the error in $L^2_\pi(\Gamma; V)$. Surprisingly, the version of the solver which estimates the maximal error is more efficient in this example. Of course, this is of limited use since this method provides an error bound only for the maximal error. These computations were performed with the Legendre polynomial basis.

The approximations computed with SolveGalerkin_{\mathcal{Q},f} converge uniformly, as is demonstrated in Figure 7.10. We compare the convergence of SolveGalerkin_{\mathcal{Q},f}, using both Legendre and Chebyshev bases, with that of SolveAlternate_{\mathcal{Q},f} with Chebyshev polynomials, and controlling the maximal error on $\Gamma$. The convergence of all of these
Chapter 7. Computational Examples and Applications

Figure 7.8.: Convergence in $C(\Gamma; V)$ of $\text{SolveDirect}_q$ with error controlled in $L^2_\pi(\Gamma; V)$ and in $C(\Gamma; V)$. Solid lines are with the Legendre basis, dashed lines with Chebyshev polynomials.

Figure 7.9.: Convergence in $L^2_\pi(\Gamma; V)$ of $\text{SolveDirect}_q$ with error controlled in $L^2_\pi(\Gamma; V)$ and in $C(\Gamma; V)$. 
7.3. Numerical Experiments

Figure 7.10.: Uniform convergence of SolveGalerkin$\text{f}_N$ with Legendre basis (solid) and Chebyshev basis (dashed), compared to SolveAlternate$\text{f}_N$ with Chebyshev polynomials, and error control in $C(I,V)$.

Figure 7.11.: Evolution of the average index length and the total number of neighbors.

methods is quite similar. We note in particular that the approximations generated by SolveAlternate$\text{f}_N$, with no coarsening, use as few degrees of freedom as those computed by SolveGalerkin$\text{f}_N$.

7.3.4. Index Sets

Our adaptive solvers automatically select active indices in $\Lambda$, and, in the case of multilevel discretizations, assign a finite element level to each. We study the discretizations generated by these methods.

The complexity of iteratively computing a Galerkin projection depends on the number of neighbors $\#(\Lambda_N)$ in an index set $\Lambda_N$. In particular, this contributes to the computational cost of SolveGalerkin$\text{f}_N$. Lemma 6.1.4 bounds $\#(\Lambda_N)$ by the average index length $\bar{\lambda}(\Lambda_N)$ times the number of indices $\#\Lambda_N$, with equality if and only if $\Lambda_N$ is monotonic.
Chapter 7. Computational Examples and Applications

Figure 7.12.: Slices of index sets generated by SolveDirect_{\text{AF}} (left) and SolveGalerkin_{\text{AF}} (right) with single level discretization (top) and multilevel discretization (bottom). Active indices with support in \{1,2\} are plotted; the level of the finite element discretization is proportional to the radius of the circle.

Figure 7.11 plots the average index length \( \bar{\lambda}(\Lambda_N) \) against the number of indices \(#\Lambda_N\) for sets generated by several adaptive solvers. We observe a clear logarithmic dependence with a slight difference between methods with and without coarsening. Consequently, the total number of neighbors \(#\mathcal{N}(\Lambda_N)\) depends log-linearly on \(#\Lambda_N\). The relation from Lemma 6.1.4 is illustrated by the similarity of the two plots in Figure 7.11.

As index sets \( \Lambda_N \subset \Lambda \) are infinite dimensional in the sense that they can contain indices of arbitrary length, they are difficult to visualize in only two dimensions. In Figure 7.12, we plot two dimensional slices of sets generated by SolveDirect_{\text{AF}} and SolveGalerkin_{\text{AF}}. We consider only those indices which are zero in all dimensions after the second, and plot their values in the first two dimensions. The upper plots depict index sets generated by adaptive solvers using single level finite element discretizations of the physical domain. Dots refer to active indices. The lower plots illustrate the discretizations generated by adaptive solvers with multilevel finite element discretizations. The radii of the circles are proportional to the discretization level. We note that, for both solvers, the index sets generated by the multilevel versions are not monotonic, although each has just a single hole within the subset considered here.

Figure 7.13 shows the number of indices per finite element discretization level in the discretizations generated by SolveDirect_{\text{AF}} and SolveGalerkin_{\text{AF}}. We observe an initial exponential increase as the discretization level is decreased, a subsequent flattening and even a decrease for the coarsest discretizations.

7.3.5. Empirical Convergence Rates

We empirically determine convergence rates of our adaptive solvers by fitting lines to the convergence data using least squares approximation. Since the convergence curves are
7.3. Numerical Experiments

Figure 7.13.: Number of indices per finite element discretization level for SolveDirect_{\text{gf}} (left) and SolveGalerkin_{\text{gf}} (right).

Table 7.2.: Estimated Convergence Rates for $k = 4$, $\gamma = 5/6$

<table>
<thead>
<tr>
<th>Method</th>
<th>$\delta$</th>
<th>$L^2_\gamma(\Gamma; V)$ w.r.t. num. dofs</th>
<th>$C(\Gamma; V)$ w.r.t. num. dofs</th>
<th>$\delta$ w.r.t. $# \Lambda_N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SolveDirect_{\text{gf}}</td>
<td>1.00</td>
<td>0.95</td>
<td>0.75</td>
<td>1.90</td>
</tr>
<tr>
<td>SolveAlternate_{\text{gf}}</td>
<td>1.00</td>
<td>0.95</td>
<td>0.80</td>
<td>1.85</td>
</tr>
<tr>
<td>SolveCoarse_{\text{gf}}</td>
<td>0.90</td>
<td>0.90</td>
<td>0.80</td>
<td>2.00</td>
</tr>
<tr>
<td>SolveGalerkin_{\text{gf}}</td>
<td>0.95</td>
<td>0.95</td>
<td>0.80</td>
<td>2.00</td>
</tr>
</tbody>
</table>

not always well approximable by affine functions, we determine such approximations independently for overlapping sets of consecutive data points, thereby obtaining a sequence of convergence rates, which locally approximate the slope of the convergence curve in bilogarithmic scale. We expect this sequence to approach the asymptotic convergence rate.

Figure 7.14 shows empirically estimated convergence rates in $L^2_\gamma(\Gamma; V)$ of the multi-level variants of our solvers for $k = 4$ and $\gamma = 5/6$. We consider convergence with respect to the total number of degrees of freedom, and with respect to the estimated computational cost. For all methods, the former seems to approach 1, while the latter is somewhat smaller. We note that the perceived faster convergence of SolveGalerkin_{\text{gf}} in Figure 7.5 is only a preasymptotic effect; this method does not have a higher asymptotic convergence rate than our other adaptive solvers.

Similar statistics for the maximal errors over the parameter domain are plotted in Figure 7.15. We consider adaptive solvers that control the error in $C(\Gamma; V)$ as well as solvers with target accuracies in $L^2_\gamma(\Gamma; V)$. Convergence rates of solvers using a single finite element discretization level are studied in Figure 7.16.

Empirical estimates of asymptotic convergence rates are listed in Table 7.2 for $k = 4$
Chapter 7. Computational Examples and Applications

Figure 7.14.: Estimated convergence rates of the error bound $\delta$ and the error in $L^2_{\pi}(\Gamma; V)$, with respect to the number of degrees of freedom (solid) and the estimated computational cost (dashed). Convergence rates are computed by fitting a linear polynomial to 16, 24, 5 and 8 data points, respectively, for the methods SolveDirect$_{\mathbb{Q}, f}$, SolveAlternate$_{\mathbb{Q}, f}$, SolveCoarse$_{\mathbb{Q}, f}$ and SolveGalerkin$_{\mathbb{Q}, f}$.

Figure 7.15.: Estimated convergence rates of the bound $\delta$ for the error in $C(\Gamma; V)$ and the error in $L^2_{\pi}(\Gamma; V)$, with respect to the number of degrees of freedom (solid) and the estimated computational cost (dashed). Convergence rates are computed by fitting a linear polynomial to 16, 24, 5 and 8 data points, respectively, for the methods SolveDirect$_{\mathbb{Q}, f}$, SolveAlternate$_{\mathbb{Q}, f}$, SolveCoarse$_{\mathbb{Q}, f}$ and SolveGalerkin$_{\mathbb{Q}, f}$. On the left, the Chebyshev polynomial basis is used.
7.3. Numerical Experiments

Figure 7.16.: Estimated convergence rates of the error bound $\delta$ for single level methods, with respect to the number of active indices in $\Lambda$ (solid) and the estimated computational cost (dashed). On the left, errors are controlled in $L^2_\Omega(\Gamma; V)$, and convergence rates are computed by fitting a linear polynomial to 10, 10, 5 and 8 data points, respectively, for the methods $\text{SolveDirect}_{\Omega,f}$, $\text{SolveAlternate}_{\Omega,f}$, $\text{SolveCoarse}_{\Omega,f}$ and $\text{SolveGalerkin}_{\Omega,f}$. On the right, errors are controlled in $C(\Gamma; V)$, the Chebyshev polynomial basis is used, and convergence rates are computed using 16 and 24 data points for $\text{SolveDirect}_{\Omega,f}$ and $\text{SolveAlternate}_{\Omega,f}$, respectively.

Table 7.3.: Estimated Convergence Rates for $k = 2$, $\gamma = 1/2$

<table>
<thead>
<tr>
<th>Method</th>
<th>$\delta$</th>
<th>$L^2_\Omega(\Gamma; V)$ w.r.t. num. dofs</th>
<th>$C(\Gamma; V)$ w.r.t. num. dofs</th>
<th>$\delta$ w.r.t. $#\Lambda_N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SolveDirect$_{\Omega,f}$</td>
<td>0.50</td>
<td>0.50</td>
<td>0.35</td>
<td>0.70</td>
</tr>
<tr>
<td>SolveAlternate$_{\Omega,f}$</td>
<td>0.50</td>
<td>0.55</td>
<td>0.40</td>
<td>0.65</td>
</tr>
<tr>
<td>SolveCoarse$_{\Omega,f}$</td>
<td>0.65</td>
<td>0.70</td>
<td>0.60</td>
<td>1.15</td>
</tr>
<tr>
<td>SolveGalerkin$_{\Omega,f}$</td>
<td>0.65</td>
<td>0.65</td>
<td>0.60</td>
<td>1.05</td>
</tr>
</tbody>
</table>
and $\gamma = \frac{5}{6}$ and in Table 7.3 for $k = 2$ and $\gamma = \frac{1}{2}$. These are determined as in the previous figures, by fitting a line to the last few data points. The convergence with respect to the number of degrees of freedom refers to multilevel solvers, and the convergence rates with respect to the number of active indices is for single level solvers. All values are rounded to the nearest $\frac{1}{20}$.

In the former case $k = 4$, the convergence of the multilevel adaptive methods in $L^2_h(I; V)$ with respect to the total number of degrees of freedom is 1. This is equal to the approximation rate shown in [CDS10b, CDS10a], see e.g. [CDS10a, Thm. 5.5]. This theorem predicts a convergence rate of 1 also in $C(I; V)$; however, for our solvers, we observe slightly slower convergence in the maximum norm.

Empirical convergence rates with respect to the number of active indices do not agree with theoretical best approximation rates. We observe a rate of 2 for single level methods, whereas [CDS10a, Thm. 4.1] predicts an asymptotic rate of 3.5. Although our measured values are very close to 2, in view of the positive slope of the estimated convergence rates plotted in Figure 7.16, it is in principle possible that the asymptotic rate is higher. For practical purposes, our observed values are certainly more relevant. Similarly, Figure 7.16 suggests a convergence rate of 1.5 in the maximum norm, while [CDS10a, Thm. 4.1] predicts an asymptotic rate of 3.

In the case $k = 2$, we observe higher convergence rates than [CDS10b, CDS10a] suggest for the multilevel methods. The versions without coarsening converge with rate $1/2$, which is what is predicted by [CDS10a, Thm. 5.5]. However, the solvers with coarsening steps have a convergence rate of $2/3$ with respect to the number of degrees of freedom. As the latter measurements are equal to $2/3$ up to several decimal points, we conjecture that this is the exact asymptotic convergence rate.

Theory predicts a convergence rate of $3/2$ with respect to the number of active indices for $k = 2$. Our observed rates for single level methods are slightly to substantially below this value. Again, due to large constants in approximation estimates, the asymptotic rates may not be perceivable for computationally accessible tolerances in this example.
Appendix A.

Tensor Products and Vector-Valued Integration

The weak formulation of parametric and stochastic operator equations naturally leads to spaces of functions on the parameter domain, mapping into Banach spaces. Various distinct constructions of integrals exist for such functions, together with generally different notions of measurability. We give an overview in Section A.1.

For the discretization of parametric operator equations, it is crucial that the function spaces appearing in the weak formulation have a tensor product structure. Such results are presented in Section A.3. Section A.2 is a brief introduction to tensor products of Banach spaces.

A.1. Notions of Measurability and Integration

A.1.1. The Bochner Integral

Let \((\Omega, \Sigma, \mu)\) be a \(\sigma\)-finite measure space and \(X\) a Banach space. An \(X\)-valued simple function on \((\Omega, \Sigma, \mu)\) is a map \(s: \Omega \to X\) taking finitely many values, each on a \(\Sigma\)-measurable set of finite measure, i.e.

\[
s = \sum_{i=1}^{n} 1_{E_i} x_i
\]

with \(x_i \in X\) and \(E_i \in \Sigma\) with \(\mu(E_i) < \infty\). For simple functions, we define the \(X\)-valued integral with respect to \(\mu\) as

\[
\int_E s \, d\mu := \sum_{i=1}^{n} \mu(E_i \cap E) x_i , \quad E \in \Sigma .
\]

(A.1.2)

This is independent of the representation (A.1.1).

A map \(f: \Omega \to X\) is strongly \(\mu\)-measurable if there exists a sequence \((s_n)_{n \in \mathbb{N}}\) of simple functions such that \(s_n(\xi) \to f(\xi)\) for \(\mu\)-a.e. \(\xi \in \Omega\); it is Bochner integrable if, in addition,

\[
\int_{\Omega} \|f - s_n\|_X \, d\mu \to 0 , \quad n \to \infty .
\]

(A.1.3)
Appendix A. Tensor Products and Vector-Valued Integration

If \( f \) is Bochner integrable,

\[
\int_E f \, d\mu := \lim_{n \to \infty} \int_E s_n \, d\mu , \quad E \in \Sigma ,
\]

is the Bochner integral of \( f \) over \( E \) with respect to \( \mu \). It follows by triangle inequality and (A.1.3) that (A.1.4) is independent of the choice of \((s_n)_{n \in \mathbb{N}}\).

**Proposition A.1.1 (Bochner’s Theorem).** Let \( f : \Omega \to X \) be strongly \( \mu \)-measurable. Then \( f \) is Bochner integrable if and only if the scalar function \( \|f\|_X \) is \( \mu \)-integrable. In this case,

\[
\left\| \int_E f \, d\mu \right\|_X \leq \int_E \|f\|_X \, d\mu , \quad E \in \Sigma .
\]

For a proof of Proposition A.1.1, we refer to [Yos80, Thm. V.5.1] or [Rya02, Prop. 2.16].

A.1.2. Weak Measurability and the Pettis Integral

As above, let \((\Omega, \Sigma, \mu)\) be a \( \sigma \)-finite measure space and \( X \) a Banach space. A function \( f : \Omega \to X \) is weakly \( \Sigma \)-measurable if \( \varphi \circ f \) is \( \Sigma \)-measurable for all \( \varphi \in X^* \). Similarly, \( f \) is weakly \( \mu \)-integrable if \( \varphi \circ f \) is \( \mu \)-integrable for all \( \varphi \in X^* \). We follow [Rya02, Sec. 3.3] in constructing an integral for weakly \( \mu \)-integrable functions.

For a fixed weakly \( \mu \)-integrable \( f : \Omega \to X \), let

\[
S_f : X^* \to L^1_\mu(\Omega) , \quad S_f \varphi := \varphi \circ f .
\]

**Lemma A.1.2.** The operator \( S_f \) is bounded.

Proof. Let \((\varphi_n)_{n \in \mathbb{N}} \subset X^* \) with \( \varphi_n \to \varphi \) in \( X^* \) and \( S_f \varphi_n = \varphi_n \circ f \to g \) in \( L^1_\mu(\Omega) \). Then there is a subsequence such that \( \varphi_n(f(\xi)) \to g(\xi) \) for \( \mu \)-a.e. \( \xi \in \Omega \). Since \( \varphi_n(f(\xi)) \to \varphi(f(\xi)) \) for all \( \xi \in \Omega \), it follows that \( g = \varphi \circ f = S_f \varphi \) in \( L^1_\mu(\Omega) \). Then the assertion is a consequence of the closed graph theorem. \( \square \)

The Dunford operator is

\[
T_f := S_f^* : (L^1_\mu(\Omega))^* \to X^{**} ;
\]

\( T_f \) can be applied to bounded functions on \( \Omega \) using isometric isomorphism between \( L^\infty_\mu(\Omega) \) and \((L^1_\mu(\Omega))^*\). For all \( g \in L^\infty_\mu(\Omega) \) and \( \varphi \in X^* \),

\[
\langle \varphi, T_f g \rangle = \langle S_f \varphi, g \rangle = \int_\Omega \varphi \circ f \, \bar{g} \, d\mu .
\]

The Dunford integral of \( f \) over \( E \) is

\[
\int_E f \, d\mu := T_f 1_E \in X^{**} , \quad E \in \Sigma .
\]
A.1. Notions of Measurability and Integration

By (A.1.8), it satisfies
\[
\left\langle \varphi, \int_E f \, d\mu \right\rangle = \int_E \varphi \circ f \, d\mu \quad \forall \varphi \in X^*.
\] (A.1.10)

We recall that \( \varphi \) is an antilinear map. However, taking complex conjugates on both sides, it follows that (A.1.10) also holds for bounded linear maps \( \varphi \).

A weakly \( \mu \)-integrable function \( f \) is Pettis integrable if the Dunford integral \( \int_E f \, d\mu \) is in \( X \) for every \( E \in \Sigma \). In this case, the Pettis integral of \( f \) over \( E \in \Sigma \) is defined by (A.1.9), interpreted as an element of \( X \) instead of \( X^* \), i.e. it is characterized by (A.1.10). If \( X \) is reflexive, then the Dunford and Pettis integrals are equivalent.

**Remark A.1.3.** Let \( f : \Omega \to X \) be Bochner integrable, and \((s_n)_{n \in \mathbb{N}}\) a sequence of simple functions approximating \( f \) as in (A.1.3). By (A.1.10), each \( s_n \) is Pettis integrable and the Pettis integral of \( s_n \) over \( E \in \Sigma \) is given by (A.1.2). Passing to the limit, it follows that \( f \) is Pettis integrable, and its Pettis integral coincides with its Bochner integral. Therefore, the Pettis integral is a consistent extension of the Bochner integral.

**A.1.3. Equivalent Notions of Measurability**

The concepts of strong measurability and weak measurability of vector-valued functions are defined in Sections A.1.1 and A.1.2, respectively. A third way to define measurability for vector-valued functions is to generalize the abstract definition for scalar functions, i.e. the preimages of open sets are measurable.

Let \((\Omega, \Sigma, \mu)\) be a \( \sigma \)-finite measure space and \( X \) a Banach space, as above. We call a function \( f : \Omega \to X \) **topologically \( \Sigma \)-measurable** if \( f^{-1}(U) \in \Sigma \) for all open sets \( U \subset X \).

In general, these three definitions of measurability differ, and strong measurability is indeed stronger than the other two notions, see e.g. [Rya02, Prop. 2.15] or [Yos80, Sec. V.4]. We show a special case of Pettis’ theorem in which they coincide.

**Proposition A.1.4.** If \((\Omega, \Sigma, \mu)\) is a complete \( \sigma \)-finite measure space and \( X \) is a separable Banach space, then strong \( \mu \)-measurability, weak \( \Sigma \)-measurability and topological \( \Sigma \)-measurability are equivalent.

**Proof.** Let \( f : \Omega \to X \) be strongly \( \mu \)-measurable, and let \((s_n)_{n \in \mathbb{N}}\) be a sequence of simple functions such that \( s_n(\xi) \to f(\xi) \) for all \( \xi \in \Omega \) \( \setminus N \) for a \( N \in \Sigma \) with \( \mu(N) = 0 \). For all \( \varphi \in X^* \), \( \varphi \circ s_n \) is \( \Sigma \)-measurable, and \( \varphi(s_n(\xi)) \to \varphi(f(\xi)) \) for all \( \xi \in \Omega \setminus N \). Since \((\Omega, \Sigma, \mu)\) is complete, all subsets of \( N \) are in \( \Sigma \), and therefore \( \varphi \circ f \) is \( \Sigma \)-measurable. Hence \( f \) is weakly \( \Sigma \)-measurable.

Suppose \( f : \Omega \to X \) is weakly \( \Sigma \)-measurable. Since \( \varphi \circ f \) is \( \Sigma \)-measurable for all \( \varphi \in X^* \), and these \( \varphi \) generate the weak topology on \( X \), \( f^{-1}(U) \in \Sigma \) for all weakly open \( U \subset X \). Hence also \( f^{-1}(C) \in \Sigma \) for all weakly closed sets \( C \subset X \), and in particular for all closed balls \( C \) in \( X \), which, being convex, are also weakly closed. By separability of \( X \), every open \( U \subset X \) is a countable union of closed balls, and consequently \( f^{-1}(U) \in \Sigma \) for all open \( U \subset X \).
Appendix A. Tensor Products and Vector-Valued Integration

Let $f : \Omega \to X$ be topologically $\Sigma$-measurable. Since $f$ is strongly $\mu$-measurable if and only if $f|_E$ is strongly $\mu$-measurable for all $E \in \Sigma$ with $\mu(E) < \infty$, it suffices to consider the case of finite $\mu$. Let $(x_k)_{k \in \mathbb{N}}$ be a dense sequence in $X$, and denote by $B^n_k$ the open ball in $X$ centered at $x_k$ with radius $1/n$. By assumption, $E^n_k := f^{-1}(B^n_k)$ are $\Sigma$-measurable and cover $\Omega$ for each $n \in \mathbb{N}$. Then $F^n_k := E^n_k \setminus \bigcup_{j < k} E^n_j$ form a disjoint measurable cover of $\Omega$ for every $n$. Setting

$$g_n := \sum_{k=1}^{\infty} 1_{F^n_k} x_k, \quad n \in \mathbb{N},$$

we have $\|f(\xi) - g_n(\xi)\|_X \leq 1/n$ for all $n \in \mathbb{N}$ and all $\xi \in \Omega$. For each $n \in \mathbb{N}$, let $m_n$ be such that

$$\mu \left( \bigcup_{k=m_n+1}^{\infty} F^n_k \right) \leq 2^{-n}$$

and define the simple function

$$s_n := \sum_{k=1}^{m_n} 1_{F^n_k} x_k.$$ Then $s_n = 0$ on $C_n := \bigcup_{k=m_n+1}^{\infty} F^n_k$, and $s_n = g_n$ on $\Omega \setminus C_n$. Note that $C := \bigcap_{n=1}^{\infty} \bigcup_{j=n}^{\infty} C_j$ has $\mu$-measure zero. For all $\xi \in \Omega \setminus C$, there is an $n$ such that $\xi \notin C_j$ for all $j \geq n$, and therefore

$$\|f(\xi) - s_j(\xi)\|_X = \|f(\xi) - g_j(\xi)\|_X \leq \frac{1}{j} \quad \forall j \geq n.$$ \hfill \Box

Remark A.1.5. Note that the assumption of completeness in Proposition A.1.4 is only used to derive weak $\Sigma$-measurability from strong $\mu$-measurability. If $(\Omega, \Sigma, \mu)$ is not complete, and $f : \Omega \to X$ is strongly $\mu$-measurable, then the above proof shows that there is a weakly $\Sigma$-measurable $\tilde{f}$ that coincides with $f$ up to a set of measure zero, and of course $\tilde{f}$ is also strongly $\mu$-measurable.

A.2. Tensor Products of Banach Spaces

A.2.1. The Algebraic Tensor Product

Let $X$ and $Y$ be vector spaces over a field $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. We define the algebraic tensor product $X \otimes Y$ abstractly via its universal property. Let $X \otimes Y$ be a vector space and $\Phi : X \times Y \to X \otimes Y$ a bilinear map such that for all vector spaces $Z$ and all bilinear maps $B : X \times Y \to Z$, there exists a unique linear map $T : X \otimes Y \to Z$ satisfying $B = T \circ \Phi$.

$$\begin{align*}
X \times Y \xrightarrow{\Phi} X \otimes Y \\
B \downarrow \quad | \quad T \\
Y \downarrow \\
Z
\end{align*}$$

164
A.2. Tensor Products of Banach Spaces

Such a space $X \otimes Y$ exists, and is unique up to linear bijections, see e.g. [Rya02, Sec. 1]. Elements of the form

$$x \otimes y := \Phi(x, y) \in X \otimes Y, \quad (x, y) \in X \times Y,$$

(A.2.2)

are simple tensors. Due to the uniqueness of the map $T$ in (A.2.1), $X \otimes Y$ is the linear span of simple tensors, i.e. every element $q \in X \otimes Y$ can be written as a finite sum

$$q = \sum_{i=1}^{n} x_i \otimes y_i,$$

(A.2.3)

with $x_i \in X$ and $y_i \in Y$. However, the representation (A.2.3) is not unique.

Let $\varphi: X \to K$ and $\psi: Y \to K$ be linear maps. Then $(x, y) \mapsto \varphi(x)\psi(y) \in K$ is a bilinear map on $X \times Y$. The universal property (A.2.1) provides a unique linear product map

$$\varphi \otimes \psi: X \otimes Y \to K, \quad x \otimes y \mapsto \varphi(x)\psi(y).$$

(A.2.4)

More generally, if $A: X \to V$ and $B: Y \to W$ are linear maps into vector spaces $V$ and $W$, then $(x, y) \mapsto (Ax) \otimes (By) \in V \otimes W$ is bilinear, and (A.2.1) leads to the linear map

$$A \otimes B: X \otimes Y \to V \otimes W, \quad x \otimes y \mapsto (Ax) \otimes (By).$$

(A.2.5)

The assignments (A.2.4) and (A.2.5) extend to arbitrary elements (A.2.3) of $X \otimes Y$ by linearity.

In the complex case $K = \mathbb{C}$, the above construction extends to antilinear maps. If $A: X \to V$ and $B: Y \to W$ are antilinear, then their complex conjugates $\bar{A}$ and $\bar{B}$, given by e.g. $\bar{A}x := \overline{Ax}$, are linear, and thus $\bar{A} \otimes \bar{B}$ is a linear map from $X \otimes Y$ to $V \otimes W$. We define the tensor product of the antilinear maps $A$ and $B$ as the antilinear map

$$A \otimes B := \bar{A} \otimes \bar{B}: X \otimes Y \to V \otimes W.$$

(A.2.6)

It satisfies $(A \otimes B)(x \otimes y) = (Ax) \otimes (By)$ for all $x \in X$ and $y \in Y$. We will use (A.2.6) in particular for elements of dual spaces, which by our convention are antilinear functionals.

A.2.2. Cross Norms

Let $X$ and $Y$ be Banach spaces. Their algebraic tensor product $X \otimes Y$ defined in Section A.2.1 is a vector space with no topological structure. In order to call a Banach space a tensor product of $X$ and $Y$, we introduce an appropriate norm on $X \otimes Y$, and take the completion with respect to this norm.

A norm $\alpha$ on $X \otimes Y$ is a cross norm if

1. $\alpha(x \otimes y) = \|x\|_{X} \|y\|_{Y}, \quad \forall x \in X, \quad \forall y \in Y,$

(A.2.7)

2. $\sup_{q \in X \otimes Y, \alpha(q) = 1} |(\varphi \otimes \psi)(q)| = \|\varphi\|_{X^*} \|\psi\|_{Y^*}, \quad \forall \varphi \in X^*, \quad \forall \psi \in Y^*.$

(A.2.8)

In some references, the term cross norm refers to only the first property (A.2.7), and $\alpha$ is called a reasonable cross norm if it also satisfies (A.2.8).
Appendix A. Tensor Products and Vector-Valued Integration

For a cross norm $\alpha$ on $X \otimes Y$, the $\alpha$-Banach tensor product of $X$ and $Y$, denoted $X \otimes_\alpha Y$, is the completion of $X \otimes Y$ with respect to $\alpha$.

A uniform cross norm $\alpha$ is an assignment to each pair of Banach spaces $X,Y$, of a cross norm $\alpha_{X,Y}$ on $X \otimes Y$ such that for all linear operators $A: X \to V$ and $B: Y \to W$,

$$
\|A \otimes B\|_{(X \otimes_\alpha Y) \to (V \otimes_\alpha W,A \otimes B)} = \|A\|_{X \to V} \|B\|_{Y \to W}.
$$

(A.2.9)

We will abbreviate $\alpha := \alpha_{X,Y}$ if the spaces $X$ and $Y$ are clear from the context. A uniform cross norm $\alpha$ is finitely generated if for all $q \in X \otimes Y$,

$$
\alpha_{X,Y}(q) = \inf \{ \alpha_{M,N}(q) ; q \in M \otimes N, M \subset X, N \subset Y, \dim M < \infty, \dim N < \infty \},
$$

(A.2.10)

i.e. $\alpha$ is determined by its behavior on finite dimensional product spaces. A tensor norm is a finitely generated uniform cross norm. We refer to [Rya02, Chap. 6] for a general discussion, and consider a few important examples of tensor norms.

A.2.3. The Projective Tensor Product

Let $X$ and $Y$ be Banach spaces. The projective norm on $X \otimes Y$ is

$$
\pi(q) := \inf \left\{ \sum_{i=1}^{n} \|x_i\|_X \|y_i\|_Y ; q = \sum_{i=1}^{n} x_i \otimes y_i \right\}.
$$

(A.2.11)

The projective tensor product $X \otimes_\pi Y$ is the completion of $X \otimes Y$ with respect to $\pi$. Elements of $X \otimes_\pi Y$ can be expressed as countable sums of simple tensors. If $q \in X \otimes_\pi Y$, then for any $\varepsilon > 0$, there are sequences $(x_i)_{i \in \mathbb{N}} \subset X$ and $(y_i)_{i \in \mathbb{N}} \subset Y$ such that

$$
q = \sum_{i=1}^{\infty} x_i \otimes y_i
$$

(A.2.12)

with convergence in $X \otimes_\pi Y$, and

$$
\sum_{i=1}^{\infty} \|x_i\|_X \|y_i\|_Y < \pi(q) + \varepsilon.
$$

(A.2.13)

In particular,

$$
\pi(q) = \inf \left\{ \sum_{i=1}^{\infty} \|x_i\|_X \|y_i\|_Y ; q = \sum_{i=1}^{\infty} x_i \otimes y_i \right\}
$$

(A.2.14)

for all $q \in X \otimes_\pi Y$.

The projective norm is a cross norm. In fact, it is the largest cross norm, i.e. if $\alpha$ is any cross norm on $X \otimes Y$, then $\alpha(q) \leq \pi(q)$ for all $q \in X \otimes Y$. Accordingly, the completion $X \otimes_\pi Y$ is the smallest Banach tensor product of $X$ and $Y$.

The projective norm satisfies

$$
\pi(x \otimes y) = \|x\|_X \|y\|_Y \quad \forall x \in X, \quad \forall y \in Y.
$$

(A.2.15)
A.2. Tensor Products of Banach Spaces

Furthermore, for any linear operators \( A: X \to V \) and \( B: Y \to W \),

\[
||A \otimes B||_{X\otimes Y \to V\otimes W} = ||A||_{X \to V} \cdot ||B||_{Y \to W} .
\]  
(A.2.16)

It is clear from the definition that \( \pi \) is finitely generated. Therefore, \( \pi \) is a tensor norm in the sense of Section A.2.2.

An operator \( Q: X \to V \) is a quotient operator if it maps the open unit ball of \( X \) onto the open unit ball of \( V \), i.e. if \( V \) is isometrically isomorphic to \( X/\ker Q \). The projective tensor product has the property that if \( Q: X \to V \) and \( R: Y \to W \) are quotient operators, then

\[
Q \otimes R: X \otimes \pi Y \to V \otimes \pi W
\]  
(A.2.17)

is also a quotient operator. Any uniform cross norm with this property is also called projective. We refer to [Rya02, Chap. 2] or [LC85, Chap. 1] for proofs of the above statements and further properties of the projective tensor product.

A.2.4. The Injective Tensor Product

For Banach spaces \( X \) and \( Y \), elements of \( X \otimes Y \) can be interpreted as bilinear forms on \( X^* \times Y^* \) via (A.2.4), i.e.

\[
X^* \times Y^* \ni (\varphi, \psi) \mapsto B_q(\varphi, \psi) \equiv (\varphi \otimes \psi)(q) = \sum_{i=1}^n \varphi(x_i)\psi(y_i) \quad \text{(A.2.18)}
\]

for \( q \in X \otimes Y \) as in (A.2.3). The injective norm \( \iota \) on \( X \otimes Y \) is the norm induced by this embedding,

\[
\iota(q) \equiv \|B_q\|_{X^* \times Y^* \to \mathbb{K}} = \sup \left\{ \left| \sum_{i=1}^n \varphi(x_i)\psi(y_i) \right| ; \varphi \in B_{X^*}, \psi \in B_{Y^*} \right\} , \quad \text{(A.2.19)}
\]

where \( B_{X^*} \) and \( B_{Y^*} \) are the unit balls of \( X^* \) and \( Y^* \), respectively, and \( x_i \in X, y_i \in Y \) are elements in a series representation (A.2.3) of \( q \in X \otimes Y \). The injective tensor product \( X \otimes \pi Y \) is the completion of \( X \otimes Y \) with respect to \( \iota \). Identifying \( q \in X \otimes Y \) with the bilinear form \( B_q \) from (A.2.18), the injective tensor product can be interpreted as a closed subspace of the space of bilinear forms on \( X^* \times Y^* \).

We may also interpret \( X \otimes \pi Y \) as a space of operators. Consider the bilinear map

\[
\ell: X \times Y \to \mathcal{L}^*(X^*, Y) , \quad \ell(x, y) \varphi := \overline{\varphi(x)y} ,
\]

where \( \mathcal{L}^*(X^*, Y) \) is the space of bounded antilinear maps from \( X^* \) to \( Y \). We recall that \( X^* \) is the space of bounded antilinear functionals on \( X \). By the universal property (A.2.1), there is a unique linear map \( L: X \otimes Y \to \mathcal{L}^*(X^*, Y) \) such that \( L(x \otimes y) = \ell(x, y) \). Abbreviating \( L_q := Lq \) for \( q \in X \otimes Y \), we have

\[
\iota(q) \equiv \|L_q\|_{X^* \to Y} = \sup \left\{ \left| \sum_{i=1}^n \overline{\varphi(x_i)y_i} \right| ; \varphi \in B_{X^*} \right\} , \quad \text{(A.2.20)}
\]
Appendix A. Tensor Products and Vector-Valued Integration

for $q$ as in (A.2.3). Similarly, exchanging the roles of $X$ and $Y$, a $q \in X \otimes Y$ can be identified with an antilinear operator $R_q \in \mathcal{L}^\alpha(Y^*, X)$ such that $R_{x \otimes y} \psi = \overline{\psi(y)}x$ for $\psi \in Y^*$, and

$$\iota(q) = \|R_q\|_{Y^* \to X} = \sup \left\{ \left| \sum_{i=1}^n \overline{\psi(y_i)}x_i \right|_{X} ; \psi \in B_{Y^*} \right\} . \quad (A.2.21)$$

Thus the injective tensor product $X \otimes Y$ can also be identified with a closed subspace of $\mathcal{L}^\alpha(X^*, Y)$ or $\mathcal{L}^\alpha(Y^*, X)$.

The injective norm $\iota$ is the smallest cross norm on $X \otimes Y$, i.e. if $\alpha$ is any cross norm on $X \otimes Y$, then $\iota(q) \leq \alpha(q)$ for all $q \in X \otimes Y$. In particular, $\iota(q) \leq \pi(q)$, and therefore $X \otimes Y$ maps continuously into the Banach space $X \otimes Y$. Furthermore,

$$\iota(x \otimes y) = \|x\|_X \|y\|_{Y^*} \quad \forall x \in X , \quad \forall y \in Y , \quad (A.2.22)$$

and for any linear operators $A : X \to V$ and $B : Y \to W$,

$$\|A \otimes B\|_{X \otimes Y \to V \otimes W} = \|A\|_{X \to V} \|B\|_{Y \to W} . \quad (A.2.23)$$

Consequently, $\iota$ is a uniform cross norm. It is independent of the space in which an element is considered in the sense that if $X_0 \subset X$ and $Y_0 \subset Y$ are closed subspaces, then the norm induced on $X_0 \otimes Y_0$ in $X \otimes Y$ coincides with the norm of $X_0 \otimes Y_0$. Any uniform cross norm with this property is called injective. Trivially, injective cross norms, and in particular the injective norm, are finitely generated, and thus $\iota$ is a tensor norm in the sense of Section A.2.2. We refer to [Rya02, Chap. 3] and [LC85, Chap. 1] for proofs of the above statements and further properties of the injective tensor product.

A.2.5. The Hilbert Tensor Product

Let $X$ and $Y$ be Hilbert spaces. We define a sesquilinear form on $X \otimes Y$ by

$$(x \otimes y, x' \otimes y')_{X \otimes Y} := (x, x')_X (y, y')_Y , \quad x, x' \in X , \quad y, y' \in Y , \quad (A.2.24)$$

and extension to $X \otimes Y$ by linearity.

Lemma A.2.1. The sesquilinear form (A.2.24) is an inner product on $X \otimes Y$.

Proof. Let $0 = q \in X \otimes Y$ be expanded as in (A.2.3). Then for all bilinear forms $B : X \times Y \to \mathbb{K}$, since $B$ is associated to a linear map $T : X \otimes Y \to \mathbb{K}$ by (A.2.1),

$$\sum_{i=1}^n B(x_i, y_i) = Tq = 0 .$$

For any $x' \in X$ and $y' \in Y$, setting $B(x, y) := (x, x')_X (y, y')_Y$, it follows that

$$(q, x' \otimes y')_{X \otimes Y} = \sum_{i=1}^n (x_i, x')_X (y_i, y')_Y = 0 ,$$

168
A.2. Tensor Products of Banach Spaces

and thus \((q, q')_{X \otimes_\eta Y} = 0\) for all \(q' \in X \otimes Y\) by antilinearity. Consequently, \((\cdot , \cdot )_{X \otimes_\eta Y}\) is well-defined.

Since symmetry is clear from the definition, it remains to be shown that \((\cdot , \cdot )_{X \otimes_\eta Y}\) is positive definite. Let \(q \in X \otimes Y\) be as in (A.2.3). Furthermore, let \((e_i)_{i=1}^n\) be an orthonormal basis of span \((x_i)_{i=1}^n\) and \((f_j)_{j=1}^n\) an orthonormal basis of span \((y_i)_{i=1}^n\). Expanding \(x_i\) and \(y_i\) in these bases, we have

\[
q = \sum_{i,j} c_{ij} e_i \otimes f_j ,
\]

and

\[
(q, q)_{X \otimes_\eta Y} = \sum_{i,j,k,l} c_{ij} \overline{c}_{kl} (e_i, e_k)_X (f_j, f_l)_Y = \sum_{i,j} |c_{ij}|^2 \geq 0 .
\]

Equality holds only if \(c_{ij} = 0\) for all \(i, j\), i.e. if \(q = 0\). \(\square\)

Let \(\|\cdot\|_{X \otimes_\eta Y}\) denote the norm induced by \((\cdot , \cdot )_{X \otimes_\eta Y}\). The Hilbert tensor product \(X \otimes_\eta Y\) of \(X\) and \(Y\) is the completion of \(X \otimes Y\) with respect to \(\|\cdot\|_{X \otimes_\eta Y}\). By Lemma A.2.1, \(X \otimes_\eta Y\) is a Hilbert space.

As a direct consequence of (A.2.24),

\[
\| x \otimes y \|_{X \otimes_\eta Y} = \| x \|_X \| y \|_Y \quad \forall x \in X , \quad \forall y \in Y .
\]  

(A.2.25)

Also, the dual space of \(X \otimes_\eta Y\) is \(X^* \otimes_\eta Y^*\).

**Lemma A.2.2.** Let \(X, Y, V\) and \(W\) be Hilbert spaces, \(A \in \mathcal{L}(X, V)\) and \(B \in \mathcal{L}(Y, W)\). Then \(A \otimes B \in \mathcal{L}(X \otimes \eta Y, V \otimes \eta W)\), and

\[
\| A \otimes B \|_{X \otimes_\eta Y \rightarrow V \otimes_\eta W} = \| A \|_{X \rightarrow V} \| B \|_{Y \rightarrow W} .
\]  

(A.2.26)

**Proof.** Let \(q \in X \otimes Y\) be as in (A.2.3). Then

\[
\left\| (A \otimes B) q \right\|_{V \otimes_\eta W}^2 = \sum_{i=1}^n \left( \sum_{j=1}^n (Ax_i) \otimes (By_j) \right)_{V \otimes_\eta W}^2 = \sum_{i,j=1}^n (Ax_i, Ax_j)_V (By_j, By_j)_W .
\]

Since \((A, A)_V\) is a hermitian sesquilinear form on \(\text{span}(x_i)\), there is an orthonormal basis \((e_i)\) of \(\text{span}(x_i)\) such that \((Ae_i, Ae_j)_V = 0\) if \(i \neq j\). Similarly, there is an orthonormal basis \((f_j)\) of \(\text{span}(y_i)\) such that \((Bf_i, Bf_j)_W = 0\) if \(i \neq j\). Expanding \(x_i\) and \(y_i\) in these bases, we have

\[
q = \sum_{i,j} c_{ij} e_i \otimes f_j ,
\]

Therefore

\[
\left\| (A \otimes B) q \right\|_{V \otimes_\eta W}^2 = \sum_{i,j} |c_{ij}|^2 \| Ae_i \|_V^2 \| Bf_j \|_W^2 \leq \| A \|_{X \rightarrow V}^2 \| B \|_{Y \rightarrow W}^2 \sum_{i,j} |c_{ij}|^2 ,
\]
Appendix A. Tensor Products and Vector-Valued Integration

and the last sum is equal to $\|q\|_{X \otimes_0 Y}^2$. This shows one direction of (A.2.26). The other follows from (A.2.25) since

$$\|A\|_{X \rightarrow V} \|B\|_{Y \rightarrow W} = \sup_{\|x\|=1} \|Ax\|_V \sup_{\|y\|=1} \|Bx\|_W = \sup_{\|x\otimes y\|_{X \otimes_0 Y}=1} \|(A \otimes B)(x \otimes y)\|_{V \otimes_0 W} \cdot$$

$\square$

In particular, $\|\|_{X \otimes_0 Y}$ is a uniform cross norm. It is both injective and projective, and thus a tensor norm on Hilbert spaces.

A.2.6. Tensor Product Operators

In working with tensor product spaces, it is important to be aware of what properties of operators $A$ and $B$ are inherited by the tensor product operator $A \otimes B$.

We note first that, given vector spaces $X_i$ and $Y_i$ for $i = 1, 2, 3$ and linear maps $A_i: X_i \rightarrow X_{i+1}$ and $B_i: Y_i \rightarrow Y_{i+1}$ for $i = 1, 2$, the composition of $A_1 \otimes B_1$ and $A_2 \otimes B_2$ is

$$(A_2 \otimes B_2)(A_1 \otimes B_1) = (A_2A_1) \otimes (B_2B_1): X_1 \otimes Y_1 \rightarrow X_3 \otimes Y_3,$$  \hspace{1cm} (A.2.27)

since this holds for simple tensors $x \otimes y \in X_1 \otimes Y_1$. This extends directly to bounded linear operators on Banach tensor products for arbitrary tensor norms.

**Lemma A.2.3.** Let $X$, $Y$, $V$ and $W$ be vector spaces and $A: X \rightarrow V$ and $B: Y \rightarrow W$ linear maps. We consider the tensor product operator $A \otimes B$ on the algebraic tensor product $X \otimes Y$, mapping into $V \otimes W$.

1. If $A$ and $B$ are injective, then $A \otimes B$ is injective.
2. If $A$ and $B$ are surjective, then $A \otimes B$ is surjective.
3. If $A$ and $B$ are bijective, then $A \otimes B$ is bijective, and

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}\). \hspace{1cm} (A.2.28)$$

Moreover, the kernel of $A \otimes B$ is

$$\ker(A \otimes B) = (\ker A) \otimes Y + X \otimes (\ker B) \subset X \otimes Y. \hspace{1cm} (A.2.29)$$

**Proof.** If $A$ and $B$ are surjective, then all simple tensors $v \otimes w \in V \otimes W$ are in the range of $A \otimes B$, thus $A \otimes B$ is surjective. If $A \otimes B$ is bijective, then (A.2.28) follows from (A.2.27).

It remains to be shown that (A.2.29) holds. We factor

$$A \otimes B = (id_V \otimes B)(A \otimes id_Y).$$
A.2. Tensor Products of Banach Spaces

The tensor product \( X \otimes Y \) can be interpreted as a space of antilinear maps from the algebraic dual \( Y' \) to \( X \) by

\[
x \otimes y \colon Y' \ni \psi \mapsto \psi(y)x \in X.
\]

Then \( A \otimes \text{id}_Y \) acts on a general element of \( X \otimes Y \) as

\[
A \otimes \text{id}_Y \sum_{i=1}^n x_i \otimes y_i \colon Y' \ni \psi \mapsto \sum_{i=1}^n \psi(y_i)Ax_i = A \sum_{i=1}^n \psi(y_i)x_i.
\]

Since we can assume without loss of generality that the \( x_i \) are linearly independent, it follows that

\[
\ker(A \otimes \text{id}_Y) = (\ker A) \otimes Y \subset X \otimes Y.
\]

Similarly,

\[
\ker(\text{id}_V \otimes B) = V \otimes (\ker B) \subset V \otimes Y.
\]

Equation (A.2.29) follows since

\[
\ker(A \otimes B) = (A \otimes \text{id}_Y)^{-1}(\ker(\text{id}_V \otimes B)).
\]

**Proposition A.2.4.** Let \( \alpha \) be any tensor norm. Let \( X, Y, V \) and \( W \) be Banach spaces, \( A \in \mathcal{L}(X, V) \) and \( B \in \mathcal{L}(Y, W) \).

1. If \( A \) and \( B \) have dense range, then \( A \otimes B \) has dense range.

2. If \( A \) and \( B \) are boundedly invertible, \( A \otimes B \) is boundedly invertible.

In the latter case, the inverse of \( A \otimes B \) is \( A^{-1} \otimes B^{-1} \).

**Proof.** Let \( A \) and \( B \) have dense range. For all \( v \in V \) and \( w \in W \), there exist sequences \( (x_k) \subset X \) and \( (y_k) \subset Y \) with \( Ax_k \to v \) and \( By_k \to w \) in norm. Therefore, \( (A \otimes B)(x_k \otimes y_k) \to v \otimes w \), and thus the closure of the range of \( A \otimes B \) contains the algebraic tensor product \( V \otimes W \), which is dense in \( V \otimes_{\alpha} W \).

Suppose that \( A \) and \( B \) are boundedly invertible. By Lemma A.2.3, the tensor product operators \( A \otimes B \) and \( A^{-1} \otimes B^{-1} \), considered on the algebraic tensor products \( X \otimes Y \) and \( V \otimes W \), are inverse to each other. Since \( \alpha \) is a uniform cross norm, both operators are bounded, and thus extend by continuity to \( X \otimes_{\alpha} Y \) and \( V \otimes_{\alpha} W \). Since they are inverse to each other on dense subspaces, this still holds for the extensions, and the extended operators are still bounded.

**Proposition A.2.5.** Let \( \alpha \) be an injective tensor norm. Let \( X, Y, V \) and \( W \) be Banach spaces, and let \( A \in \mathcal{L}(X, V) \) and \( B \in \mathcal{L}(Y, W) \) be injective and have closed range. Then \( A \otimes B \) is an injective operator from \( X \otimes_{\alpha} Y \) to \( V \otimes_{\alpha} W \) with closed range

\[
\text{range}(A \otimes B) = (\text{range } A) \otimes_{\alpha} (\text{range } B) \subset V \otimes_{\alpha} W.
\]

(A.2.30)
Appendix A. Tensor Products and Vector-Valued Integration

Proof. The ranges \( V_0 := \text{range } A \subset V \) and \( W_0 := \text{range } B \subset W \) are closed by assumption. Due to the open mapping theorem, the operators \( A: X \to V_0 \) and \( B: Y \to W_0 \) are boundedly invertible. Proposition A.2.4 implies that \( A \otimes B: X \otimes_\alpha Y \to V_0 \otimes_\alpha W_0 \) is boundedly invertible, and thus in particular injective. Since \( V_0 \otimes_\alpha W_0 \) is a closed subspace of \( V \otimes_\alpha W \) by injectivity of \( \alpha \), the assertion follows. \[ \square \]

Proposition A.2.6. Let \( \alpha \) be a projective tensor norm. Let \( X, Y, V \) and \( W \) be Banach spaces, and let \( A \in \mathcal{L}(X, V) \) and \( B \in \mathcal{L}(Y, W) \) be surjective. Then \( A \otimes B \) is a surjective operator from \( X \otimes_\alpha Y \) to \( V \otimes_\alpha W \).

Proof. We assume initially that \( W = Y \) and \( B = \text{id}_Y \). Let \( K := \ker A \subset X \), and let \( Q: X \to X/K \) be the canonical projection. Then \( A = A_1 Q \) with \( A_1 \in \mathcal{L}(X/K, V) \) boundedly invertible by the open mapping theorem. Tensorizing with \( \text{id}_Y \), we have \( A \otimes \text{id}_Y = (A_1 \otimes \text{id}_Y)(Q \otimes \text{id}_Y) \). Since \( \alpha \) is projective, \( A \otimes \text{id}_Y \) is a quotient map onto \( (X/K) \otimes_\alpha Y \), and in particular surjective. By Proposition A.2.4, \( A_1 \otimes \text{id}_Y \) is an isomorphism from \( (X/K) \otimes_\alpha Y \) to \( V \otimes_\alpha Y \), so it too is surjective, and consequently \( A \otimes \text{id}_Y \) is a surjective linear operator from \( X \otimes_\alpha Y \) onto \( V \otimes_\alpha Y \).

In the general case, the same argument shows that \( \text{id}_V \otimes B \) is a surjective linear map from \( V \otimes_\alpha Y \) to \( V \otimes_\alpha W \), and the assertion follows since \( A \otimes B \) is the composition of \( A \otimes \text{id}_Y \) and \( \text{id}_V \otimes B \). \[ \square \]

The above statements hold in particular for the Hilbert tensor product since it is both injective and projective. Furthermore, the Hilbert tensor product is compatible with adjoints. If \( X, Y, V \) and \( W \) are Hilbert spaces, \( A \in \mathcal{L}(X, V) \) and \( B \in \mathcal{L}(Y, W) \), then the adjoint of

\[
A \otimes B: X \otimes_\eta Y \to V \otimes_\eta W 
\]

is the tensor product operator

\[
(A \otimes B)^* = A^* \otimes B^*: V^* \otimes_\eta W^* \to X^* \otimes_\eta Y^* .
\]

This follows by linearity and continuity since \((A \otimes B)^*\) and \(A^* \otimes B^*\) coincide for simple tensors.

A.3. Tensor Product Structure of Function Spaces

A.3.1. Lebesgue–Bochner spaces

Lebesgue–Bochner spaces are defined analogously to standard Lebesgue spaces of scalar-valued functions for the Bochner integral (A.1.4). For \( 1 \leq p \leq \infty \), the Lebesgue–Bochner space is the Banach space

\[
L^p_\mu(\Omega; X) := \{ f: \Omega \to X \mid f \text{ strongly } \mu\text{-measurable and } \| f \|_X \in L^p_\mu(\Omega) \} \quad (A.3.1)
\]
A.3. Tensor Product Structure of Function Spaces

with norm \( \| f \|_{L^p(\Omega;X)} := \| f \|_{\overline{L}^p(\Omega;X)} \), i.e.

\[
\| f \|_{L^p(\Omega;X)} = \left( \int_\Omega \| f \|_X^p \, d\mu \right)^{1/p}
\]  (A.3.2)

for \( p < \infty \) and \( \| f \|_{L^\infty(\Omega;X)} = \text{ess sup}_{\xi \in \Omega} \| f(\xi) \|_X \). Of course, we group functions into equivalence classes of \( \mu \)-almost everywhere identical functions. For \( 0 < p < 1 \), this definition leads to a quasi-Banach space. Note that by Proposition A.1.1 \( L^1_\mu(\Omega;X) \) is the space of Bochner integrable functions.

The following result is shown e.g. in [LC85, Thm. 1.15] and [Rya02, Ex. 2.19].

**Theorem A.3.1.** Let \( (\Omega, \Sigma, \mu) \) be a measure space and \( X \) a Banach space. Then the map \( f \otimes x \mapsto fx \) extends uniquely to an isometric isomorphism from the projective tensor product \( L^1_\mu(\Omega) \otimes_\pi X \) to the Lebesgue–Bochner space \( L^1_\mu(\Omega;X) \).

A similar result holds for Hilbert spaces; for a proof, we refer to [RS72, Thm. II.10].

**Theorem A.3.2.** Let \( (\Omega, \Sigma, \mu) \) be a measure space with \( L^2_\mu(\Omega) \) separable, and let \( X \) be a separable Hilbert space. Then the map \( f \otimes x \mapsto fx \) extends uniquely to an isometric isomorphism from the Hilbert tensor product \( L^2_\mu(\Omega) \otimes_\eta X \) to the Lebesgue–Bochner space \( L^2_\mu(\Omega;X) \).

Note that \( L^2_\mu(\Omega) \) is separable if \( \mu \) is \( \sigma \)-finite and, up to \( \mu \)-null sets, \( \Sigma \) has a countable generator, see [Doo94, Sec. 15].

A.3.2. Lebesgue–Pettis Spaces and the Injective Tensor Product

As in the case of the Bochner integral, we define spaces of \( p \)-integrable \( X \)-valued functions for the Pettis integral. For \( 1 \leq p < \infty \), define the norm

\[
\| f \|_{\overline{L}^p(\Omega;X)} := \sup_{\varphi \in B_{X^*}} \| \varphi \circ f \|_{L^p(\Omega)}
\]  (A.3.3)

for \( f : \Omega \to X \) weakly integrable, where \( B_{X^*} \) denotes the unit ball of \( X^* \). Let \( \mathcal{S}_\mu(\Omega;X) \) denote the vector space of equivalence classes of \( \mu \)-a.e. identical \( X \)-valued simple functions (A.1.1) on \( (\Omega, \Sigma, \mu) \). The Lebesgue–Pettis space \( \overline{L}^p_\mu(\Omega;X) \) of \( p \)-integrable functions is the completion with respect to \( \| \cdot \|_{\overline{L}^p(\Omega;X)} \) of \( \mathcal{S}_\mu(\Omega;X) \). Note that, since simple functions are dense in \( \overline{L}^p_\mu(\Omega) \), \( \overline{L}^p_\mu(\Omega;X) \) coincides with \( \overline{L}^p_\mu(\Omega) \).

Since the Lebesgue–Pettis space \( \overline{L}^p_\mu(\Omega;X) \) is defined abstractly as the completion of \( \mathcal{S}_\mu(\Omega;X) \), it is not clear initially if its elements can be interpreted as functions. Of course, any Pettis integrable map \( f : \Omega \to X \) that can be approximated with respect to \( \| \cdot \|_{\overline{L}^p(\Omega;X)} \) by simple functions is in \( \overline{L}^p_\mu(\Omega;X) \). However, in general not all elements of \( \overline{L}^p_\mu(\Omega;X) \) are of this form. Rather, under some additional assumptions, \( \overline{L}^p_\mu(\Omega;X) \) can be interpreted as a space of vector measures on \( (\Omega, \Sigma) \), see [Rya02, Cor. 5.19].

173
Appendix A. Tensor Products and Vector-Valued Integration

Abbreviating $\mathcal{F}_\mu(\Omega) := \mathcal{F}_\mu(\Omega; K), \mathcal{F}_\mu(\Omega; X)$ can be identified with the algebraic tensor product $\mathcal{F}_\mu(\Omega) \otimes X$ via the linear map determined by

$$\mathcal{F}_\mu(\Omega) \otimes X \ni s \otimes x \mapsto sx \in \mathcal{F}_\mu(\Omega; X).$$  \hspace{1cm} (A.3.4)

This map is a bijection since both spaces are spanned by elements of the form (A.3.4) where $s$ is an indicator function. The norm (A.3.3) restricted to $\mathcal{F}_\mu(\Omega; X)$ coincides with the injective norm $\iota$ on $\mathcal{F}_\mu(\Omega) \otimes X$ due to (A.2.21), where $\mathcal{F}_\mu(\Omega)$ is considered as a subspace of $L^p_{\mu}(\Omega)$. Therefore, (A.3.4) extends to an isometric isomorphism from the injective tensor product $L^p_{\mu}(\Omega) \otimes \iota X$ to the Lebesgue–Pettis space $\overline{L}^p_{\mu}(\Omega; X)$.

In particular, $\overline{L}^p_{\mu}(\Omega; X)$ can be interpreted as a closed subspace of both $L^p_{\mu}(\Omega, L^p_{\mu}(\Omega))$ and $\mathcal{L}^\infty((L^p_{\mu}(\Omega))^*, X)$. By the first embedding, if $f \in L^p_{\mu}(\Omega; X)$, then for any $\varphi \in X^*$, $\varphi \circ f \in L^p_{\mu}(\Omega)$. The second interpretation allows the computation of $X$-valued integrals against weights in $L^p_{\mu}(\Omega)$, where $p'$ is the Hölder conjugate to $p$.

A.3.3. Spaces of Continuous Functions

Besides Lebesgue spaces, also spaces of continuous vector-valued functions have a tensor product structure. The following is shown e.g. in [LC85, Thm. 1.13] and [Rya02, Sec. 3.2].

Theorem A.3.3. If $\Omega$ is a compact Hausdorff space and $X$ a Banach space, then the injective tensor product $C(\Omega) \otimes X$ is isometrically isomorphic to $C(\Omega; X)$ via the linear map determined by $f \otimes x \mapsto fx$.

Consequently, continuous functions are dense in Lebesgue–Pettis spaces.

Corollary A.3.4. Let $\Omega$ be a compact Hausdorff space, $\mu$ a Radon measure on $\Omega$, and $X$ a Banach space. Then the range of the linear map

$$i_X : C(\Omega; X) \to \overline{L}^p_{\mu}(\Omega; X)$$  \hspace{1cm} (A.3.5)

determined by $fx \mapsto fx$ is dense for all $1 \leq p < \infty$.

Proof. If $C(\Omega; X)$ is identified with $C(\Omega) \otimes X$ and $\overline{L}^p_{\mu}(\Omega; X)$ is identified with $L^p_{\mu}(\Omega) \otimes \iota X$, then $i_X = i \otimes \text{id}_X$, where $i$ is the embedding of $C(\Omega)$ into $L^p_{\mu}(\Omega)$. Since $i$ has dense range by e.g. [Bau92, Satz 29.14], the assertion follows using Proposition A.2.4. \quad \Box

Note that, depending on the measure $\mu$, the map $i_X$ in Corollary A.3.4 may or may not be injective.

A.3.4. Sequence Spaces

For an arbitrary index set $\Xi$, let $\ell^p(\Xi)$ denote the standard space of $p$-summable sequences indexed by $\Xi$ for $1 \leq p \leq \infty$, and let $c_0(\Xi)$ denote the subspace of $\ell^\infty(\Xi)$ of sequences for which, given any $\epsilon > 0$, only finitely many elements have absolute value...
A.3. Tensor Product Structure of Function Spaces

greater than \( c \). Furthermore, for a Banach space \( X \), let \( \ell^p(\Xi; X) \) and \( c_0(\Xi; X) \) denote their \( X \)-valued analogs, i.e. \( \ell^p(\Xi; X) \) is a Lebesgue–Bochner space for the counting measure on \( \Xi \), and \( c_0(\Xi; X) \) consists of all sequences in \( X \) indexed by \( \Xi \) such that for any \( \epsilon > 0 \), all but finitely many elements have norm less than \( \epsilon \), endowed with the maximum norm.

**Theorem A.3.5.** Let \( 1 \leq p < \infty \). For any Banach space \( X \) and any cross norm \( \alpha \) on \( \ell^p(\Xi) \otimes X \), there is a unique linear map \( \gamma \) from \( \ell^p(\Xi) \otimes X \to c_0(\Xi; X) \) satisfying

\[
\gamma(s \otimes x) = (s_i, x_i)_{i \in \Xi} \quad \forall s = (s_i) \in \ell^p(\Xi), \quad \forall x \in X,
\]

and \( \gamma \) is injective and has norm one.

**Proof.** On the algebraic tensor product \( \ell^p(\Xi) \otimes X \), we consider the map

\[
\gamma: \ell^p(\Xi) \otimes X \to c_0(\Xi; X), \quad \sum_{i=1}^n s_i \otimes x_i \mapsto \left( \sum_{i=1}^n s_i x_i \right)_{i \in \Xi}.
\]

This is well-defined since \( s_i \in \ell^p(\Xi) \subset c_0(\Xi) \) for all \( i \). Let \( \sum_{i=1}^n s_i \otimes x_i \in \ell^p(\Xi) \otimes X \) be fixed. For each \( \nu \in \Xi \), by Hahn–Banach there is a continuous linear functional \( \varphi_\nu \) on \( X \) with \( \| \varphi_\nu \|_{X^*} = 1 \) and

\[
\| \varphi_\nu \|_{\ell^p(\Xi)} = 1.
\]

Furthermore, the map \( \delta_\nu: \ell^p(\Xi) \to \mathbb{K} \) given by \( \delta_\nu s := s_\nu \) is a continuous linear functional on \( \ell^p(\Xi) \), and \( \| \delta_\nu \|_{\ell^p(\Xi)} = 1 \). By definition,

\[
(\delta_\nu \otimes \varphi_\nu) \left( \sum_{i=1}^n s_i \otimes x_i \right) = \sum_{i=1}^n \delta_\nu(s_i) \varphi_\nu(x_i) = \varphi_\nu \left( \sum_{i=1}^n s_i x_i \right) = \left\| \sum_{i=1}^n s_i x_i \right\|_{X^*}.
\]

Since \( \alpha \) is a cross norm, (A.2.8) implies

\[
\left( \| \delta_\nu \|_{\ell^p(\Xi)} \right) \left( \| \varphi_\nu \|_{X^*} \right) = \| \sum_{i=1}^n s_i x_i \|_{X^*}.
\]

Consequently,

\[
\sup_{\nu \in \Xi} \left\| \sum_{i=1}^n s_i x_i \right\|_{X} \leq \alpha \left( \sum_{i=1}^n s_i \otimes x_i \right),
\]

i.e. \( \gamma \) is continuous with norm one. Therefore, \( \gamma \) extends to a linear map

\[
\hat{\gamma}: \ell^p(\Xi) \otimes X \to c_0(\Xi; X)
\]

with norm one.

To show injectivity of \( \hat{\gamma} \), we embed \( \ell^p(\Xi) \otimes X \) into \( \mathcal{L}^p(X^* \otimes \ell^q(\Xi)) \). Let \( S := \ell^q(\Xi) \) for the Hölder conjugate \( q \) of \( p \) if \( p > 1 \), and \( S := c_0(\Xi) \) if \( p = 1 \). Then \( \ell^q(\Xi) \) is isomorphic to \( S^* \),
Appendix A. Tensor Products and Vector-Valued Integration

see e.g. [Wer00, Satz II.2.3]. Furthermore, \( \ell^p(\Xi) \) has the approximation property, [Rya02, Example 4.5]. Therefore, [Rya02, Prop. 8.7] implies that \( \ell^p(\Xi) \otimes_\alpha X \) is isomorphic to the space of antilinear \( \alpha \)-nuclear operators from \( S \) to \( X \). A simple tensor \( s \otimes x \) is mapped to the operator \( T_{s \otimes x} \in \mathcal{L}^\ast(S, X) \) given by

\[
T_{s \otimes x} t := \sum_{i \in \Xi} i_s v_i x, \quad t \in S.
\]

Furthermore, \( \mathcal{L}^\ast(S, X) \) embeds into \( \mathcal{L}^\ast(X^\ast, \ell^p(\Xi)) \) through

\[
\mathcal{L}^\ast(S, X) \ni T \mapsto (X^\ast \ni \varphi \mapsto (S \ni t \mapsto \varphi(Tt))) \in \mathcal{L}^\ast(X^\ast, \ell^p(\Xi))
\]

which is linear in \( T \) by antilinearity of \( \varphi \in X^\ast \), and injective since for any \( T \in \mathcal{L}^\ast(S, X) \) not equal to the zero map, there exists a \( t \in S \) with \( Tt \neq 0 \) in \( X \), and thus there is also a \( \varphi \in X^\ast \) with \( \varphi(Tt) \neq 0 \). The concatenation of these embeddings maps a simple tensor \( s \otimes x \) to \( \hat{T}_{s \otimes x} \in \mathcal{L}^\ast(X^\ast, S) \) given by

\[
\hat{T}_{s \otimes x}(\varphi)(t) = \varphi \left( \sum_{i \in \Xi} i_s v_i x \right) = \sum_{i \in \Xi} i_s v_i \varphi(x), \quad t \in S, \quad \varphi \in X^\ast.
\]

Moving from \( S^\ast \) back to \( \ell^p(\Xi) \), \( \hat{T}_{s \otimes x} \) maps \( \varphi \in X^\ast \) to \( \varphi(x)s \). Since all its factors are continuous and injective, the map from \( \ell^p(\Xi) \otimes_\alpha X \) into \( \mathcal{L}^\ast(X^\ast, \ell^p(\Xi)) \) is an embedding. We write \( \hat{T}_{q} \) for the image of \( q \in \ell^p(\Xi) \otimes_\alpha X \). For finite rank tensors,

\[
\hat{T}_{\sum_{i=1}^n s_i \otimes x_i}(\varphi) = \sum_{i=1}^n \hat{T}_{s_i \otimes x_i}(\varphi) = s_i \varphi(x_i) = \left( \varphi \left( \delta_{i} \left( \sum_{i=1}^n s_i \otimes x_i \right) \right) \right)_{\varphi \in \Xi}.
\]

Let \( q \in \ell^p(\Xi) \otimes_\alpha X \) with \( \hat{\gamma}(q) = 0 \), and let \( (q_m)_{m \in \mathbb{N}} \subset \ell^p(\Xi) \otimes_\alpha X \) with \( q_m \to q \) in \( \ell^p(\Xi) \otimes_\alpha X \). Then for all \( v \in \Xi \) and all \( \varphi \in X^\ast \), by continuity of \( \delta_v \) on \( \ell^p(\Xi) \) and on \( c_0(\Xi; X) \),

\[
\delta_v(\hat{T}_q \varphi) = \lim_{m \to \infty} \delta_v(\hat{T}_{q_m} \varphi) = \lim_{m \to \infty} \varphi(\delta_v(\gamma(q_m))) = \varphi(\delta_v(\gamma(q))) = 0.
\]

Consequently, \( \hat{T}_q \) is the zero operator in \( \mathcal{L}^\ast(X^\ast, \ell^p(\Xi)) \), and \( q = 0 \) due to the injectivity of the map \( q \mapsto \hat{T}_q \). \( \square \)

Theorem A.3.5 states that spaces \( \ell^p(\Xi) \otimes_\alpha X \) for \( 1 \leq p < \infty \) can always be interpreted as spaces of sequences in \( X \) converging to 0. These spaces are always tractable in that finite sequences are dense.

**Theorem A.3.6.** Let \( X \) be a Banach space, and let \( \alpha \) be a uniform cross norm on \( \ell^p(\Xi) \otimes_\alpha X \) for \( 1 \leq p < \infty \). For any \( x = (x_v)_{v \in \Xi} \in \ell^p(\Xi) \otimes_\alpha X \) and any sequence \( (\Xi_n)_{n \in \mathbb{N}} \) in \( \mathcal{P}(\Xi) \), denote by \( x^n = (x^n_v)_{v \in \Xi} \in \ell^p(\Xi) \otimes_\alpha X \) the restriction of \( x \) to \( \Xi_n \), i.e. \( x^n_v := x_v \) if \( v \in \Xi_n \) and \( x^n_v := 0 \) if \( v \in \Xi \setminus \Xi_n \). Then for any \( x \in \ell^p(\Xi) \otimes_\alpha X \), there exists a sequence \( (\Xi_n)_{n \in \mathbb{N}} \) in \( \mathcal{P}(\Xi) \) such that \( x^n \to x \). Furthermore, if \( (\Xi_n)_{n \in \mathbb{N}} \) is any sequence in \( \mathcal{P}(\Xi) \) such that \( \Xi_n \uparrow \Xi \), then \( x^n \to x \) for all \( x \in \ell^p(\Xi) \otimes_\alpha X \).

\(^2\)We recall that \( \mathcal{P}(\Xi) \) is the power set of \( \Xi \), and \( \mathcal{F}(\Xi) \subset \mathcal{P}(\Xi) \) is the set of all finite subsets of \( \Xi \).
A.3. Tensor Product Structure of Function Spaces

Proof. We first show that finitely supported sequences are dense in $\ell^p(\mathcal{Z}) \circledast_a X$. Let

$$x = \sum_{i=1}^n c_i \otimes x_i \in \ell^p(\mathcal{Z}) \circledast_a X, \quad c_i \in \ell^p(\mathcal{Z}), \quad x_i \in X.$$ 

For each $i$, there is a sequence $(c_i^k)_{k \in \mathbb{N}}$ of finitely supported scalar sequences such that $c_i^k \to c_i$ as $k \to \infty$. Consequently, $x^k := \sum_{i=1}^n c_i^k \otimes x_i \to \sum_{i=1}^n c_i \otimes x_i = x$ in $\ell^p(\mathcal{Z}) \circledast_a X$, and $x^k$ is supported on the union of the supports of the sequences $c_i^k$, $i = 1, \ldots, n$, which is finite. Therefore, any $x \in \ell^p(\mathcal{Z}) \circledast_a X$ can be approximated in $\ell^p(\mathcal{Z}) \circledast_a X$ by finite sequences, and since the algebraic tensor product is dense, it follows that finitely supported sequences are dense in $\ell^p(\mathcal{Z}) \circledast_a X$.

Let $x = (x_i)_{i \in \mathcal{Z}} \in \ell^p(\mathcal{Z}) \circledast_a X$. By the above argument, there exists a sequence $(\hat{x}_k)_{k \in \mathbb{N}}$ in $\mathcal{F}(\mathcal{Z})$ and for each $k \in \mathbb{N}$ a $\hat{x}_k = (x_i^k) \in \ell^p(\mathcal{Z}) \circledast_a X$ with $x_i^k = 0$ for all $i \in \mathcal{Z} \setminus \hat{x}_k$ such that $\hat{x}_k \to x$ in $\ell^p(\mathcal{Z}) \circledast_a X$. For all $k \in \mathbb{N}$, let $p_k : \ell^p(\mathcal{Z}) \to \ell^p(\mathcal{Z})$ denote the map given by

$$(p_k c)_v := c_v \quad \text{if} \quad v \in \hat{x}_k \quad \text{and} \quad (p_k c)_v := 0 \quad \text{if} \quad v \in \mathcal{Z} \setminus \hat{x}_k.$$ 

Then $p_k$ and $\text{id}_{\ell^p(\mathcal{Z})} - p_k$ have norm one on $\ell^p(\mathcal{Z})$, and by (A.2.9), $p_k \otimes \text{id}_X$ and $(\text{id}_{\ell^p(\mathcal{Z})} - p_k) \otimes \text{id}_X$ have norm one on $\ell^p(\mathcal{Z}) \circledast_a X$. We note that $(p_k \otimes \text{id}_X)\hat{x}_k = \hat{x}_k$, and $(p_k \otimes \text{id}_X)x = x^k$ for $x^k$ defined as in the statement of the assertion. Consequently,

$$x - \hat{x}_k = x - \hat{x}_k + \hat{x}_k - x^k = x - \hat{x}_k - (p_k \otimes \text{id}_X)(x - \hat{x}_k) = ((\text{id}_{\ell^p(\mathcal{Z})} - p_k) \otimes \text{id}_X)(x - \hat{x}_k),$$

and thus

$$\alpha(x - x^k) = \alpha(((\text{id}_{\ell^p(\mathcal{Z})} - p_k) \otimes \text{id}_X)(x - \hat{x}_k)) \leq \alpha(x - \hat{x}_k) \to 0,$$

which shows the first statement of the assertion.

Finally, let $(\Xi_n)_{n \in \mathbb{N}}$ be any sequence in $\mathcal{F}(\mathcal{Z})$ with $\Xi_n \uparrow \Xi$. Defining $\hat{\Xi}_0 := \emptyset$, we ensure that

$$k_n := \max\left\{ k \leq n ; \hat{\Xi}_k \subset \Xi_n \right\} \in \mathbb{N}_0, \quad n \in \mathbb{N},$$

are well-defined. Since for each $k \in \mathbb{N}$, there is a $n \in \mathbb{N}$ such that $\hat{\Xi}_k \subset \Xi_n$, we have $k_n \to \infty$ as $n \to \infty$. Therefore, $\hat{x}^{k_n}$ converges to $x$, and $\hat{x}^{k_n}$ is supported on $\Xi_n$. By the same argument as above, it follows that

$$\alpha(x - x^{k_n}) \leq \alpha(x - \hat{x}^{k_n}) \to 0, \quad n \to \infty,$$

and thus $x^{k_n} \to x$ in $\ell^p(\mathcal{Z}) \circledast_a X$. \hfill \Box
References


References


References


References


References


References


Curriculum Vitae

Personal details

Name Claude Jeffrey Gittelson
Date of birth May 1, 1985
Place of birth Worcester, Massachusetts, USA
Citizenship Switzerland and the United States of America

Education

2/2008–2011 Seminar for Applied Mathematics
Eidgenössische Technische Hochschule Zürich
Zurich, Switzerland

Zurich, Switzerland

3/2008 Master of Science ETH in Mathematics
12/2006 Bachelor of Science ETH in Mathematics

Wetzikon ZH, Switzerland

8/2003 Matura Type C (Mathematics & Physics)

8/1996–6/1999 Phillips Middle School
Chapel Hill, North Carolina, USA