Adaptive Methods and Collocation by Splines for Solving Differential Equations

By

Shiying Li

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Mathematics

August 9, 2019

Nashville, Tennessee

Approved:

Larry Schumaker, Ph.D.

Marian Neamtu, Ph.D.

Akram Aldroubi, Ph.D.

Caglar Oskay, Ph.D.

Alexander Powell, Ph.D.

To my dear parents, Qiang Li and Li Dai

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor Prof. Larry Schumaker and co-advisor Prof. Mike Neamtu. I am grateful for the freedom that Prof. Schumaker gave me to explore areas of interest on my own as well as his generous help in supporting me to attend many wonderful workshops and conferences both in the US and Europe. Through his wise guidance in our first project, I learned to write relatively complicated codes and became motivated and comfortable in programming. I also appreciate his patience and tremendous help when we were writing our first paper; we had to communicate via emails since I was in China during that time. I am also very grateful for the financial support Prof. Schumaker kindly provided in several summers and my last year at Vanderbilt. The observation of his diligence and organization in daily research and always giving honest opinions will have a lasting influence in my future career and life. I am also extremely grateful to my co-advisor Prof. Neamtu for his genuine care for the development of students and time in many insightful discussions about IGA and collocation methods, despite an overwhelming schedule. He made an effort in identifying good summer schools and workshops for me and generously supported my travel. His optimistic attitude and fearlessness when facing a challenging research problem has been a constant encouragement to me. Lastly, I would like to thank Prof. Neamtu for his sincere effort in helping and care for my fellow student Ziliang and his family last winter, which comforted his family and friends including me greatly during that difficult time.

Many thanks to my committee members, Professors Akram Aldroubi, Caglar Oskay and Alexander Powell, for spending precious time reading and revising this manuscript. I especially thank Prof. Aldroubi for his encouragement through the years and introducing me to a very nice lab and research area for my postdoc career. I would also like to thank Professors Bruce Hughes and Jesse Peterson for writing and sending recommendation letters for me. Prof. Hughes and Dr. Jakayla Robbins's instruction style has had an impact on the way I teach and I thank them for their genuine care for me and other students.

I would also like to thank my fellow graduate students in our department for their friendship and support. Special thanks to Longxiu Huang and Bin Sun who offered me some insightful ideas related to a lemma in the thesis and generously shared their experiences related to the graduation and OPT application process with me.

My very special thanks to my family in China and Nashville Chinese Baptist Church (NCBC) family

here. My parents have always supported me with their love and understanding. I also thank my younger brother for taking care of them when I am in the US and his encouragement. I am deeply indebted to Andrew Siao and Iwen Chia, who treated me like their daughter in faith and supported me with constant prayer, encouragement and care, along with many other elders and friends in NCBC. In particular, I am grateful to Mengying Xi for her sincere company, Joy Wang for her humor and positive words, and Chunxue Wang for her wisdom in clarifying various issues for me. Most importantly, I thank God for being my guide and strength.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Chapter 1

Introduction

1.1 Adaptive Methods by Splines on H-Triangulations

The discussion about adaptive methods by splines on H-triangulations is in Chapter 3. Polynomial splines defined on ordinary triangulations are a well known and highly effective tool in numerical mathematics, and are used in a variety of settings for approximating data and solving partial differential equations numerically, among other things. The recent book [38] explains in detail how to compute with such splines efficiently using Bernstein–Bézier techniques, and even includes an extensive Matlab library.

Polynomial splines defined on triangulations with hanging vertices (H-triangulations) have been used for some time by engineers to solve boundary value problems (BVP), see [1, 3, 41] to cite just a few examples. The main advantage of allowing hanging vertices is that it allows much simpler local adaptive refinement techniques than is possible in the framework of ordinary triangulations.

On the other hand, splines on triangulations with hanging vertices have been treated in the mathematical literature only recently — see [40], where questions such as dimension, construction of stable local bases, and approximation power are dealt with using Bernstein–Bézier techniques. The purpose of our study is to show how these same techniques can effectively be used for computing with such splines, thus avoiding the use of parametric maps as is common in the engineering literature.

All of the numerical algorithms described in [38] for working with splines on ordinary triangulations can be carried over to the case of H-triangulations. These include a variety of scattered data fitting methods such as minimal energy methods, local macro-element based methods, and local methods based on derivative estimation. They also include least-squares fitting and penalized least-squares for fitting noisy data. Chapter 3 of this manuscript is focused on adaptive computation using splines on H-triangulations, with two specific applications :1) function approximation, and 2) solution of boundary value problems for partial differential equations. This part is largely from the paper [26].

1.2 Collocation Methods

The discussion about collocation methods is in Chapter 4. The application of the collocation methods to differential equations appeared as early as 1930s [16]. Two basic ingredients of collocation are a finite

dimensional space of trial functions and a set of points to be collocated by the differential equations. Collocating with cubic splines were first seen in [8, 17] to solve two-point boundary value problems. Russell and Shampine [32] developed a general theory for solving two-point boundary problems involving m-th order ordinary differential equations using splines in C^m. Existence, uniqueness and convergence properties of a solution was treated. In general, a convergence order of $(d+1-m)$ is obtained for sufficiently smooth solutions , where d is the degree of spline space involved. deBoor and Swartz [11] generalized their argument to collocation using C^{m-1} splines and proposed the use of Gauss-Legendre points in each subinterval to obtain optimal order accuracy, which is also known as the *orthogonal spline collocation* (OSC) method nowadays, see the survey paper [14] by Fairweather and Meade and references therein. Applications and analysis of the OSC method to partial differential equations can be found in [7, 31] among many others. In particular, for an overview of convergence theory of the OSC methods, see Sect. 3.2 in [7], where references of optimal order H^1, H^2 or L^2 error estimates for different types of elliptic boundary value problems are directed. A recent application of the OSC method to the Navier-Stokes equations can be found in [15]. Collocating at other special collocation sets such as Cauchy-Galerkin points in [19] leads to production of the Galerkin solution. One drawback of this idea is that these points are not easy to obtain in general and can only be estimated in practice. A comparison of the numerical performance of various collocation sets can be found in [29].

In this thesis, we presented an alternative proof of existence and error bounds to that in [32, 11] in the case of of cubic collocation on the knots for linear second-order two-point boundary value problems and extended the proof to the case when the collocation points are the associated Greville points. While their proofs deal with more general cases, i.e., higher-order ordinary differential equations (both linear and nonlinear), our proof does not rely heavily on the properties of the Green's function associated with a specific differential operator and the constants in our error bounds are more tractable. To achieve higher order convergence, instead of collocating at special points such as the OSC method, we introduce a generalized collocation model by modifying the collocation equations based on certain quasi-interpolation operators. The collocation points being used can be simply uniformly distributed points or Greville points. Algorithms using spline spaces of maximum smoothness for both 1D and 2D problems (second-order elliptic partial differential equations in the unit square) have been proposed and numerical examples showing optimal convergence results for spline spaces of various degrees are given. A detailed analysis for the generalized collocation using C^2 cubic splines in the case when the differential operator $L = D^2$ is also given. After

completing the research, we found out that similar results had been proved for the C^2 cubic case in Archer's Ph.D. thesis [4], in which the term *modified collocation* is used. Archer later generalized this idea to obtain optimal convergence rate using C^1 cubic splines in [5]. Irodotou-Ellina and Houstis developed an $O(h^6)$ modified collocation method with quintic splines for linear fourth-order two-point boundary value problems in [22] and extended the method to more general problems in [21]. Houstis et al. [20] presented a modified collocation method using C^2 bicubic splines for solving elliptic BVP in the unit square, the numerical results of which can be compared with ours in Sec. 4.5.2. A connection between the modified collocation and the so-called deferred correction process [17] has been noted in several papers, e.g. [22, 20], which is considered an alternative way to achieve optimal convergence rates.

In Sec. 4.6, a least-squares collocation problem using C^2 spline spaces defined on triangulations is formulated for the Poisson problem. A sufficient condition on the choice of collocation points which guarantees the existence of a unique collocation solution is given and an example of such choice is also provided. In Theorem 4.6.1, an error bound for the collocation solution comparing the error with the error of best approximation in $W^{2,\infty}$-norm to the solution of the Poisson problem is given. Results on convergence rates are given in Theorem 4.6.2. Numerical examples are tested for $S_9^{2,4}(\triangle)$, a macro-element spline space of degree 9.

Chapter 2

Spline Tools

2.1 B-splines

Univariate splines are piecewise polynomials defined on a partition $\triangle$ of an interval $I = [a,b]$, where the partition is defined by a set of points $\triangle := \{x_i\}_{i=0}^{k+1}$ with $a = x_0 < x_1 < \cdots < x_k < x_{k+1} = b$. The mesh size is defined to be $h := \max\limits_{i=1,\ldots k+1} |x_i - x_{i-1}|$. We recall the following definitions and properties guided by classical spline literature, see e.g., [10, 35].

Given integers $0 \leq r < d$ and a partition $\triangle$, the space of univariate splines of smoothness r and degree d is defined as

$$S_d^r(\triangle) := \{s \in C^r(I) : s|_{(x_i,x_{i+1})} \in \mathscr{P}_d, i = 0,...,k\}, \tag{2.1}$$

where $\mathscr{P}_d$ is the space of univariate polynomials of degree at most d. It is well known that the dimension of spline space $S_d^r(\triangle)$ is

$$n := \dim S_d^r(\triangle) = k(d-r) + d + 1. \tag{2.2}$$

The associated extended partition $\triangle_e$ is defined to be $\{y_i\}_{i=1}^{n+d+1}$, where n is the dimension of $S_d^r(\triangle)$,

$$a = y_1 = \cdots = y_{d+1}, \quad y_{n+1} = \cdots = y_{n+d+1} = b,$$

and

$$y_{d+2} \leq \cdots \leq y_n = \overbrace{x_1,...,x_1}^{d-r}, \cdots, \overbrace{x_k,...,x_k}^{d-r}.$$

Given an extended partition $\triangle_e$, starting from $m = 1$, the B-splines of order m can be generated recursively by the Cox-de Boor formula, where

$$N_i^1(t) := \begin{cases} 1, & y_i \leq t < y_{i+1}, \\ 0, & \text{otherwise}, \end{cases} \tag{2.3}$$

and

$$N_i^m := \begin{cases} \frac{t-y_i}{y_{i+m-1}-y_i} N_i^{m-1} + \frac{y_{i+m}-t}{y_{i+m}-y_{i+1}} N_{i+1}^{m-1}, & y_i \le t < y_{i+m}, \\ 0, & \text{otherwise}, \end{cases} \tag{2.4}$$

for $2 \le m \le d+1$ and $i = 1, ..., n+d-m+1$.

2.2 Splines on Triangulations

Bernstein–Bézier techniques are the key tools for dealing with splines on ordinary triangulations, see [24]. Given a non-degenerate triangle $T := < v_1, v_2, v_3 >$ in $\mathbb{R}^2$ and an integer $d > 0$, the associated Bernstein basis polynomial of degree d relative to T are defined as

$$B_{ijk}^d := \frac{d!}{i!\,j!\,k!} b_1^i b_2^j b_3^k, \quad i+j+k = d, \tag{2.5}$$

where $b_1(x,y), b_2(x,y), b_3(x,y)$ are the linear functions giving the barycentric coordinates of the point (x,y) relative to T and i, j, k are non-negative integers. The associated set of *domain points of degree d* is

$$\mathscr{D}_{d,T} := \{\xi_{ijk}^T := \frac{iv_1 + jv_2 + kv_3}{d}, \quad i+j+k = d\}.$$

It is easy to associate each domain point with a Bernstein basis polynomial. We may also index the basis polynomials as $\{B_\xi^T\}_{\xi \in \mathscr{D}_{d,T}}$. Note that the number of these basis polynomials associated to a triangle T is $n_d := (d+1)(d+2)/2$.

A collection $\triangle := \{T_i\}_{i=1}^{n_t}$ of triangles in the plane is called a triangulation of $\Omega = \bigcup_{i=1}^{n_t} T_i$, provided the non-empty intersection of any pair of triangles is either a common vertex or a common edge. We denote the mesh size of $\triangle$ as $|\triangle| := \max_{T_i \in \triangle} |T_i|$. Given a triangulation $\triangle$, we write $\mathscr{PP}(\triangle)$ for the linear space of all piecewise polynomials defined on $\triangle$. Then for any $s \in \mathscr{PP}(\triangle)$ and any $T \in \triangle$, we can write

$$s|_T = \sum_{\xi \in \mathscr{D}_{d,T}} c_\xi B_\xi^T.$$

The coefficients c_ξ are called the *B-coefficients* of s. It is easy to see that $\mathscr{PP}(\triangle)$ is in 1-1 correspondence with the set of domain points

$$\mathscr{D}_{d,\triangle} := \bigcup_{T \in \triangle} \mathscr{D}_{d,T}$$

associated with $\triangle$, where here the union is to be understood in the sense that multiple appearances of the same point are allowed. This implies immediately that $\dim \mathscr{P}\mathscr{P}(\triangle) = n_t n_d$.

2.3 Tensor-Product Splines

In spaces of dimension two or larger, tensor-product splines are convenient approximation tools. These splines are simply tensor products of univariate B-splines. Consider a (hyper-) cube $\Omega := \bigotimes_{i=1}^{N}[a_j, b_j]$ in $\mathbb{R}^N$, and partitions $\triangle_j := \{x_i^j\}_{i=0}^{k_j+1}$ with $a_j = x_0^j < x_1^j < \cdots < x_{k_j}^j < x_{k_j+1}^j = b_j$ with $j = 1,...,N$, and let $\alpha := (\alpha_1,...,\alpha_N)$ be a multi-index and

$$\Omega_\alpha := \bigotimes_{j=1}^{N}[x_{\alpha_j}^j, x_{\alpha_j+1}^j),$$

where $0 \le \alpha_j \le k_j$ and we replace the corresponding half-open intervals by closed intervals when $\alpha_j = k_j$. This partitions Ω into $\prod_{i=1}^{N}(k_i+1)$ subcubes. We denote the space of tensor-product polynomials of degree $\mathbf{d} := (d_1,...,d_N)$ as

$$\mathscr{P}^{\mathbf{d}} := \{p(\mathbf{x}) : p(\mathbf{x}) = \sum_{\beta \le \mathbf{d}} c_\beta \mathbf{x}^\beta\},$$

where $\mathbf{x}^\beta$ follows the usual multi-index power rule with $x = (x_1,...,x_N)$ and $\beta = (\beta_1,...,\beta_N)$, and "$\le$" is a partial order induced by the component-wise comparison.

Given multi-index N-tuples $\mathbf{d}$ and $\mathbf{r}$, let

$$S_{\mathbf{d}}^{\mathbf{r}} := \{s(\mathbf{x}) \in C^{\mathbf{r}}(\Omega) : s|_{\Omega_\alpha} \in \mathscr{P}^{\mathbf{d}}, \alpha \le \mathbf{k}\}$$

where $\mathbf{k} = (k_1,...,k_N)$ and $C^{\mathbf{r}}(\Omega)$ is the set of all functions on Ω such that the partial derivatives $D^\gamma u$ are continuous for all N-tuple $\gamma \le \mathbf{r}$. We call this space of the tensor-product splines of degree $\mathbf{d}$ and smoothness $\mathbf{r}$. It is common to denote the above space as $\bigotimes_{j=1}^{N} S_{d_j}^{r_j}(\triangle_j)$, where each $S_{d_j}^{r_j}(\triangle_j)$ is a univariate spline space defined in Section 2.1. A basis for this space is of the form

$$\mathscr{N}_{\mathbf{i}}^{\mathbf{d+1}}(\mathbf{x}) := \prod_{j=1}^{N} N_{i_j}^{d_j+1}(x_j), \quad 1 \le i_j \le n_j, \tag{2.6}$$

where $\mathbf{i} = (i_1,...,i_N)$, $N_{i_j}^{d_j+1}$ is a univariate B-spline basis function in the corresponding space dimension and n_j is the dimension of the space $S_{d_j}^{r_j}(\triangle_j)$. For more properties of these *tensor-product B-splines*, see [38].

2.4 Spline Quasi-interpolation

Spline quasi-interpolation (QI) is an approximation scheme of the following form:

$$Qu := \sum \lambda_i(u) N_i, \tag{2.7}$$

where u is a continuous function defined on some domain, λ_i is a linear functional and $\{N_i\}$ is a basis of a certain spline space $S_d^r(\triangle)$. A construction theory and properties of such schemes can be found in [27]. Unlike Lagrange interpolation or least-squares approximation, they can be constructed directly with local information on the function u. Here we introduce the case when the linear functional λ_i takes local values of u in a neighborhood of the support of N_i, in contrast to other possibilities like derivatives or local integrals of u. In general, we impose that Q is a projection on the space of polynomials of total degree at most d. Various constructions of such Q have an order of accuracy comparable to the best approximation in the spline space used.

The following notion of local boundedness is useful in proving error bounds for a quasi-interpolation operator. We repeat Definition 1.52 in [38]:

Definition 2.4.1. *Suppose Q is a linear operator mapping $C[a,b]$ to $S_d^r(\triangle)$, and suppose there exist constants K_1, K_2 independent of u such that for every subinterval I_i of the partition $\triangle$, there exists an interval J_i containing I_i with*

1) $|J_i| \leq K_1 |I_i|$

2) $||Qu||_{L^\infty(I_i)} \leq K_2 ||u||_{L^\infty(J_i)}$ *for all* $u \in C[a,b]$.

Then we say that Q is locally bounded.

According to Theorem 1.53 in [38], a locally bounded quasi-interpolation operator Q which reproduces polynomials up to degree d has the following property: for every $u \in C^m[a,b]$ with $1 \leq m \leq d+1$,

$$||D^j(f - Qu)||_{L^\infty[a,b]} \leq K h^{m-j} ||D^m u||_{L^\infty[a,b]}, \tag{2.8}$$

for all $0 \leq j \leq \min(m-1, r)$. For $j = 0$, the constant K depends only on d and the constants K_1, K_2 in the above definition. For $j > 0$, it also depends on the global mesh ratio $\sigma := \frac{\max\limits_{0 \leq i \leq k}(x_{i+1} - x_i)}{\min\limits_{0 \leq i \leq k}(x_{i+1} - x_i)}$ associated the partition $\triangle$ being used.

We give a few explicit schemes in the one-dimensional setting, see more details in [33]. For multidimensional quasi-interpolation schemes, see [27]. In the following examples, let $I = [a,b]$ and $\triangle = a = x_0 < x_1 < \cdots < x_k < x_{k+1} = b$ be a uniform partition of I, i.e., $x_i = a + ih$ and $h = \frac{b-a}{k+1}$. Given a spline space $S_d^{d-1}(\triangle)$, define a quasi-interpolation operator Q of the form

$$Qu := \sum_{j=1}^{n} \lambda_j(u) N_j, \tag{2.9}$$

where n is the dimension of the spline space and $\{N_i\}_{i=1}^n$ is the associated B-spline basis.

Example 2.4.1. *Consider the cubic spline space $S_h := S_3^2(\triangle)$ where $n = k+4$. Given a continuous function u, a cubic quasi-interpolation operator $Q_3 : C[a,b] \to S_h$ is defined such that the coefficient functionals are: $\lambda_1(u) = u(x_0)$, $\lambda_2(u) = \frac{1}{18}(7u(x_0) + 18u(x_1) - 9u(x_2) + 2u(x_3))$, $\lambda_{n-1}(u) = \frac{1}{18}(2u(x_{n-6}) - 9u(x_{n-5}) - 18u(x_{n-4}) + 7u(x_{n-3}))$, $\lambda_n(u) = u(x_{n-3})$, and for $3 \le j \le n-2$,*

$$\lambda_j(u) = \frac{1}{6}(-u(x_{j-3}) + 8u(x_{j-2}) - u(x_{j-1})). \tag{2.10}$$

Discussion: In this construction, the coefficient linear functionals only take values of u on the knots $\{x_j\}_{j=0}^{k+1}$ of the partition. According to [33], $||Q_3||_\infty \le 2$ and $Q_3 p = p$ for all $p \in \mathscr{P}_3$.

Claim 2.4.1. *The quasi-interpolation operator Q_3 is locally bounded.*

Proof. Let $I_i = [x_{i-1}, x_i], i = 1, ..., k+1$ and $\{N_j\}_{j=1}^n$ be the B-spline basis for the cubic spline space in this example. For convenience we employ the convention that $x_i = x_0$ when $i < 0$, $x_i = x_{k+1}$ when $i > k+1$, and $N_j(x) = 0$ if $j < 0$ or $j > n$. Using the local properties of the B-splines, it is easy to verify that for $x \in I_i$, $Qu(x) = \sum_{j=i}^{i+3} \lambda_j(u) N_j$. Recall the definition of λ_j and the partition of unity property of the basis $\{N_j\}_{j=1}^n$, we have

$$||Qu||_{L^\infty(I_i)} \le \max_{i \le j \le i+3} |\lambda_j| \le 2||u||_{L^\infty[x_{j-3}, x_{j+2}]}. \tag{2.11}$$

Hence by letting $J_i = [x_{j-3}, x_{j+2}], K_1 = 5$ and $K_2 = 2$, the claim is proved. □

Hence by (2.8) we obtain that for $u \in C^m[a,b]$ with $1 \le m \le 4$

$$||D^j(u - Q_3 u)||_{L^\infty[a,b]} \le K h^{m-j} ||D^m u||_{L^\infty[a,b]}, \tag{2.12}$$

for all $0 \leq j \leq \min(m-1,2)$ and the constant K depends only on $d=3$ and the constants K_1, K_2 as in Definition 2.4.1. Note that $||u - Q_3 u)||_{L^\infty[a,b]} \leq Kh^4 ||D^4 u||_{L^\infty[a,b]}$ for all $u \in C^4[a,b]$, which turns out to have the same convergence rate in terms of the mesh size h as the best spline approximation in $S_3^2(\triangle)$.

Example 2.4.2. *Consider the quartic spline space $S_h := S_4^3(\triangle)$ where $n = k+5$. Let $T := \{t_j\}_{j=1}^{n-2}$ be a set of discrete points such that $t_1 = x_0$, $t_{n-2} = x_{k+1}$ and $t_j = \frac{1}{2}(x_{j-2} + x_{j-1})$ otherwise. Given a continuous function u, a quartic quasi-interpolation operator $Q_4 : C[a,b] \to S_h$ is defined such that the coefficient functionals are:* $\lambda_1(u) = u(t_1)$, $\lambda_2(u) = \frac{17}{105}u(t_1) + \frac{35}{32}u(t_2) - \frac{35}{96}u(t_3) + \frac{21}{160}u(t_4) - \frac{5}{224}u(t_5)$, $\lambda_3(u) = -\frac{19}{45}u(t_1) + \frac{377}{288}u(t_2) + \frac{61}{288}u(t_3) - \frac{59}{480}u(t_4) + \frac{7}{288}u(t_5)$, $\lambda_4(u) = \frac{47}{315}u(t_1) - \frac{77}{144}u(t_2) + \frac{251}{133}u(t_3) - \frac{97}{240}u(t_4) + \frac{47}{1008}u(t_5)$, *and for* $5 \leq j \leq n-4$,

$$\lambda_j(u) = \frac{47}{1152}[u(t_{j-4}) + u(t_j)] - \frac{107}{288}[u(t_{j-3}) + u(t_{j-1})] + \frac{319}{192}u(t_{j-2}),$$

while $\lambda_{n-3}, \lambda_{n-2}, \lambda_{n-1}, \lambda_n$ are defined accordingly by symmetry (of which we omit the explicit forms).

Discussion: In this construction, the coefficient linear functionals take values of u on the the set T, i.e., midpoints of subintervals of the partition and endpoints of I. According to [33], $||Q_4||_\infty \leq 3$ and Q_4 reproduces polynomials up to degree 4. It can be shown that Q_4 is locally bounded using similar arguments as in Claim 2.4.1. Hence for $u \in C^m[a,b]$ with $1 \leq m \leq 5$,

$$||D^j(u - Q_4 u)||_{L^\infty[a,b]} \leq Kh^{m-j} ||D^m u||_{L^\infty[a,b]}, \tag{2.13}$$

for all $0 \leq j \leq \min(m-1,3)$ and the constant K depends only on $d=4$ and the constants K_1, K_2 as in Definition 2.4.1. Note that $||u - Q_4 u||_{L^\infty} \leq Ch^5 ||D^5 u||_{L^\infty[0,1]}$ for all $u \in C^5[a,b]$, which shows the same convergence rate in terms of the mesh size h as the best spline approximation in $S_4^3(\triangle)$.

2.5 Spline Approximation Theory

In this section we present some results about the approximation power of spline spaces defined on partitions of an interval and on partitioned rectangles. For proofs of these results, see [35, 38]. In the two theorems below, we follow the notations in Section 2.1 and 2.3 respectively.

Theorem 2.5.1. *(cf. Sect. 6.5, 6.6 in [35] or Theorem 1.51 in [38]) Let $S_d^r(\triangle)$ be a space of univariate splines defined on $[a,b]$. Suppose $u \in C^m[a,b]$ with $1 \leq m \leq d+1$. Then there exists a spline $s \in S_d^r(\triangle)$ such*

that

$$||D^j(s-u)||_{L^\infty[a,b]} \leq Ch^{m-j}||D^m u||_{L^\infty[a,b]} \tag{2.14}$$

for all $0 \leq j \leq \min(m-1, r)$*, where C depends only on d.*

We remark that for $u \in C^{d+1}[a,b]$, the above error bound for $||s-u||_{L^\infty[a,b]}$ is of order $\mathcal{O}(h^{d+1})$, which is known to be the optimal order of approximation, see Sect. 6 in [35].

Theorem 2.5.2. *(cf. Theorem 12.7 in [35] or Theorem 2.22 in [38]) Let* $S_{d_1}^{r_1}(\triangle_1) \bigotimes S_{d_2}^{r_2}(\triangle_2)$ *be a space of tensor-product splines defined on a rectangle* Ω*. Suppose* $u \in C^{m_1,0}(\Omega) \cap C^{0,m_2}(\Omega)$ *with* $1 \leq m_1 \leq d_1+1$ *and* $1 \leq m_2 \leq d_2+1$*. Then there exists a spline* $S_{d_1}^{r_1}(\triangle_1) \bigotimes S_{d_2}^{r_2}(\triangle_2)$ *such that*

$$||s-u||_{L^\infty(\Omega)} \leq C(h_1^{m_1}||D_x^{m_1}u||_{L^\infty(\Omega)} + {h_2}^{m_2}||D_y^{m_2}u||_{L^\infty(\Omega)}), \tag{2.15}$$

where the constant C depends only on d_1, d_2 *and* h_i *is the mesh size corresponding to the univariate partition* $\triangle_i$ *with* $i = 1, 2$*.*

Chapter 3

Spline-Based Adaptive Methods

3.1 Introduction

3.1.1 Splines on H-Triangulations

Let $\triangle := \{T_i\}_{i=1}^{n_t}$ be a collection of triangles such that the interior of the domain $\Omega := \cup T_i$ is connected. In addition, suppose that any pair of distinct triangles can intersect each other only at points on their edges. Then we call $\triangle$ an *H-triangulation* of Ω.

This definition allows triangulations to have *hanging vertices*, i.e., a vertex v of a triangle may lie in the interior of an edge of another triangle. Whenever we refer to an H-triangulation, we assume it is regular and without cycles in the sense of the paper [39].

It turns out that for splines defined on H-triangulations we can use the same Bernstein–Bézier techniques that are the key tools for dealing with splines on ordinary triangulations.

For most applications we prefer to work with spaces of splines that are at least continuous, i.e., that are subspaces of $S_d^0(\triangle)$. Indeed, for many applications it suffices to work with $S_d^0(\triangle)$ itself. This space is well understood in the case of ordinary triangulations, but has to be treated with care when working on H-triangulations.

First, we describe an efficient scheme for storing splines $s \in S_d^0(\triangle)$. Rather than storing s as a member of $\mathscr{P}\mathscr{P}(\triangle)$, which would involve $n_t n_d$ coefficients, we can take advantage of the continuity of s and store a smaller set of coefficients. Let $\hat{D}_{d,\triangle}$ be the subset of $\mathscr{D}_{d,\triangle}$ obtained by choosing just one point at each vertex, $d-1$ points on each edge, and the $\binom{d-1}{2}$ points of $\mathscr{D}_{d,T}$ that lie inside of T for each triangle $T \in \triangle$. It is clear the cardinality of $\hat{D}_{d,\triangle}$ is

$$n_c = n_v + (d-1)n_e + \binom{d-1}{2} n_t,$$

where n_v, n_e, and n_t are the numbers of vertices, edges, and triangles of $\triangle$.

The number n_c is not generally equal to the dimension of $S_d^0(\triangle)$. Indeed, if the coefficients of s are known for domain points on a composite edge e, then the C^0 continuity of s determines all of its coefficients associated with domain points on subedges of e. For example, consider a spline $s \in S_1^0(\triangle)$ on the triangu-

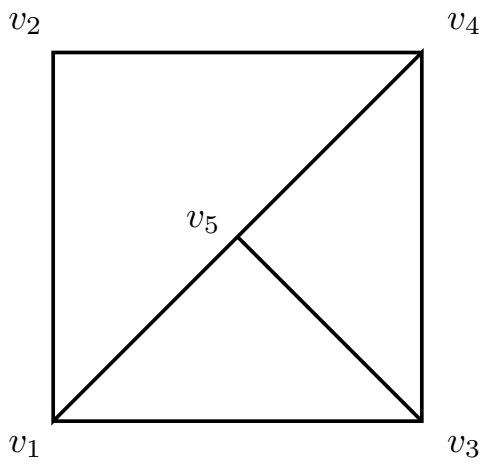


Figure 3.1.1: An H-triangulation

lation in Figure 3.1.1. Then since s reduces to a linear polynomial on the edge $\langle v_1, v_4 \rangle$, it follows that the value of s at the vertex v_5 is determined by its values at the vertices v_1 and v_4.

To get the dimension of $S_d^0(\triangle)$, we now look for a smallest set $\mathscr{M}$ of points in $\hat{D}_{d,\triangle}$ such that setting the corresponding coefficients of a spline $s \in S_d^0(\triangle)$ uniquely determines all other coefficients. Such a set is called a *minimal determining set* (MDS), see e.g. [24, 38, 40]. Let $\mathscr{M}$ be the subset of $\hat{D}_{d,\triangle}$ which is obtained from $\hat{D}_{d,\triangle}$ by dropping the domain points at hanging vertices along with those lying in the interior of edges which are proper subedges of composite edges. We quote the following theorem from [40].

Theorem 3.1.1. *For any H-triangulation $\triangle$, the set $\mathscr{M}$ is a MDS for $S_d^0(\triangle)$, and*

$$\dim S_d^0(\triangle) = \hat{n}_v + (d-1)\hat{n}_e + \binom{d-1}{2} n_t, \tag{3.1}$$

where $\hat{n}_v$ is the number of nonhanging vertices of $\triangle$, and $\hat{n}_e$ is the number of composite edges of $\triangle$.

Proof. It is easy to verify that the set $\mathscr{M}$ described above is a minimal determining set for $S_d^0(\triangle)$. It follows that the dimension of $S_d^0(\triangle)$ is just the cardinality of $\mathscr{M}$, which is given by the formula in (3.1). □

Suppose $\triangle = \{T_i\}_{i=1}^{n_t}$ is an H-triangulation of a polygonal domain Ω, and let $0 \le r < d$ be given integers. Then we define the associated *space of splines of degree d and smoothness r* to be

$$S_d^r(\triangle) := \{s \in C^r(\Omega) : s|_{T_i} \in \mathscr{P}_d \text{ for all } i = 1, \ldots, n_t\},$$

where $\mathscr{P}_d := \text{span}\{x^i y^j\}_{0 \le i+j \le d}$ is the usual space of polynomials of degree d. Clearly $S_d^r(\triangle)$ is a finite dimensional linear space. For $d \ge 4r+1$ an explicit formula for its dimension in terms of d, r, and the number of both hanging and nonhanging vertices of $\triangle$ can be found in [40].

Certain superspline subspaces of $S_d^r(\triangle)$ defined on H-triangulations are also useful for applications. Here we only recall their definition, leaving a disussion of computational methods for another time. Fix $0 \le 2r \le \rho < d$. Then

$$S_d^{r,\rho}(\triangle) := \{s \in S_d^r(\triangle) : s \in C^\rho(v) \text{ for all vertices of } \triangle\}, \tag{3.2}$$

is called a *superspline* space, where $s \in C^\rho(v)$ means that the polynomial pieces $s|_T$ associated with triangles T with a vertex at v have common derivatives up to order ρ at the point v. They were introduced in [40] for H-triangulations.

For a detailed theoretical treatment of supersplines on ordinary triangulations, see [24], and for computational methods based on them, see [36, 38]. Superspline spaces like $S_d^{r,\rho}(\triangle)$ have several advantages. In particular, they have the same approximation properties as $S_d^0(\triangle)$, but with fewer degrees of freedom, and often they are easier to work with than the spaces $S_d^r(\triangle)$.

3.1.2 Approximation Power of Splines on H-triangulations

The approximation power of spline spaces defined on H-triangulations was investigated in [40]. Suppose $S_d^{r,\rho}(\triangle)$ is the super-spline space defined in (3.2) with $0 \le 2r \le \rho < d$. Let $|\triangle| = \max_{T \in \triangle} |T|$ be the mesh size of $\triangle$, and for any $m \ge 0$ and $1 \le q \le \infty$, let $W_q^{m+1}(\Omega)$ be the usual Sobolev space defined on the domain Ω. Let $\|\cdot\|_q$ be the usual q-norm on Ω, and let $|\cdot|_{m+1,q,\Omega}$ be the seminorm measuring $m+1$-st order derivatives in the q-norm.

Theorem 3.1.2. *(cf. [40]) Suppose $f \in W_q^{m+1}(\Omega)$ with $0 \le m \le d$ and $1 \le q \le \infty$. Then there exists a spline*

$s \in S_d^{r,\rho}(\triangle)$ *such that*

$$\|D_x^\nu D_y^\mu (f-s)\|_q \le K|\triangle|^{m+1-\nu-\mu}\ |f|_{m+1,q,\Omega},$$

for all $0 \le \nu+\mu \le m$.

The constant K depends on d, the size of the smallest angle in $\triangle$, the length of the longest chain of hanging vertices (see Remark 6), and the constant

$$\alpha_\triangle := \max_{e\in\mathscr{E}_c} \max_{\tilde{e}\subset e} \frac{|e|}{|\tilde{e}|}, \tag{3.3}$$

where $\mathscr{E}_c$ is the set of all composite edges containing two or more edge segments. If Ω is not convex, then K also depends on the Lipschitz constant of the boundary of Ω.

3.2 Local Refinements and A Posteriori Error Estimators

Given an H-triangulation $\triangle$, we can locally refine it by choosing a triangle T and splitting it into two or more subtriangles. This can be done in several different ways. Here are two commonly used approaches:

1. *S1: Edge split.* Insert an edge segment connecting one vertex v of T to the midpoint w of the edge e opposite to v. This splits T into two subtriangles.

2. *S2: Midpoint refinement.* Connect the midpoints of the edges of T to each other to split T into four subtriangles.

For other refinement possibilities, see Remark 9. Each of these methods has advantages and disadvantages. Method S1 may reduce the size of the smallest angle in the triangulation, but this can be somewhat mitigated by always splitting the longest edge of the triangle. Method S2 leaves the smallest angle unchanged.

In both of these methods, if a new vertex is introduced, it will be either a boundary vertex, or a new hanging interior vertex of the refined triangulation. If a split point falls on an existing vertex, then that vertex becomes a non-hanging vertex.

In [36, 38] it was shown how various spline spaces can be used in conjunction with the Ritz-Galerkin method to compute approximations to the solutions of elliptic boundary value problems. The discussion there was for spline spaces on ordinary triangulations, but there is no essential difficulty in carrying out the

same program for splines defined on H-triangulations. We just need to work with spline spaces for which we know a minimal determining set, and for which we can compute the transformation matrix A. In this section we illustrate the method using the spaces $S_d^0(\triangle)$ along with local refinement to find good triangulations.

We first illustrate how to couple adaptive methods with H-triangulations in function approximation, assuming that we can sample the function at arbitrary points. Suppose f is a given continuous function defined on a domain Ω. We propose to use an adaptive algorithm to compute a spline approximant of f. In this section we choose an intitial triangulation of Ω, and work with splines of degree d on $\triangle$ and refinements of it. Here is a general algorithm for carrying this out. Suppose `Appx` is some approximation process producing an approximation in a spline space $S(\triangle)$. Suppose `Err` is a process to compute a vector $err = (err_1, \ldots, err_{n_t})$, where n_t is the number of triangles in $\triangle$. Choose a maximal number of iterations n_r.

Algorithm 1.

For $i = 1$ *until* n_r

1. *Pick a type of spline space* $S(\triangle)$
2. *Pick an initial triangulation* $\triangle$
3. *Use* `Appx` *to compute a spline* s *in* $S(\triangle)$
4. *Use* `Err` *to compute the error vector* err
5. *Sort* err
6. *Split the triangle with the largest error*
7. *Replace* $\triangle$ *by the new triangulation and repeat this process*

Refining one triangle at a time can be slow since at each step we have to reconstruct the approximating spline and compute the vector *err* of triangle errors. Thus, in practice we suggest using the following variant of this algorithm where we refine groups of triangles. Pick a number $0 < p < 100$.

Algorithm 2. *Replace step 4 in Algorithm 1 by*

4'. Choose the triangles corresponding to the largest p percent of the errors, and split them.

For the experiments below, we will primarily use Algorithm 2 with $p = 5$. The choice of the approximation process and error computation will depend on the application. Here we define the approximation scheme `Appx` as follows: Find a spline $s \in S(\triangle)$ that interpolates f at all of the domain points $\mathscr{D}_{d,\triangle}$. Next

we have to choose a scheme `Err` for computing the error on each triangle. Here are some possible choices. Let m be an integer greater than d, and define

$$e1(T) := A(T) \sum_{\xi \in \mathscr{D}_{m,T}} |s(\xi) - f(\xi)|, \tag{3.4}$$

$$e2(T) := A(T) \sum_{\xi \in \mathscr{D}_{m,T}} |s(\xi) - f(\xi)|^2, \tag{3.5}$$

$$em(T) := A(T) \max_{\xi \in \mathscr{D}_{m,T}} |s(\xi) - f(\xi)|, \tag{3.6}$$

$$E1(T) := \int_T |s - f|, \tag{3.7}$$

$$E2(T) := \int_T |s - f|^2, \tag{3.8}$$

where $A(T)$ is the area of the triangle T. In practice we typically take $m = d + 2$.

In order to couple adaptive methods with the Ritz-Galerkin method (to be defined in the next section) for solving our boundary value Problem 1, we make use of either Algorithm 1 or Algorithm 2. To use them, we need to have a way of computing an error value associated with each triangle in $\triangle$. Here is the problem to be attacked. We follow the notation of [38]. Suppose we are given functions f and κ defined on a domain Ω, and suppose g is a function defined on the boundary $\partial\Omega$ of Ω.

Problem 1. *Find a function u defined on Ω such that*

$$Lu := -\nabla \cdot (\kappa \nabla u) = f, \qquad on\ \Omega, \tag{3.9}$$

$$u = g, \qquad on\ \partial\Omega. \tag{3.10}$$

Here ∇ denotes the vector-valued differential operator $[D_x, D_y]^T$, and the dot denotes the vector inner-product. Now given a spline space $S(\triangle)$, suppose $\{\phi_1, ..., \phi_{n_0}\}$ is a basis for $U_0 := \{s \in S(\triangle) : s \equiv 0 \text{ on } \partial\Omega\}$. Then we look for an approximate solution of Problem 1 in the form

$$s = \sum_{i=1}^{n_0} c_i \phi_i + s_b, \tag{3.11}$$

where s_b is a spline in $S(\triangle)$ such that s_b is approximately equal to g on the boundary $\partial\Omega$.

Since we do not know the true solution of the boundary value problem, this error will have to be con-

structed numerically without using u. Computed error estimates of this type are usually called *a posteriori error estimates*, and have been studied extensively in the PDE literature, see e.g. [2, 6, 9]. Here are some possibilities in our setting, based on estimating the size of the residual $Ls - f$. Given a triangle T and a spline s along with an integer $m > d$, let

$$r_1(T) := A(T) \sum_{\xi \in \mathscr{D}_{m,T}} |Ls(\xi) - f(\xi)|, \tag{3.12}$$

$$r_2(T) := A(T) \sum_{\xi \in \mathscr{D}_{m,T}} |Ls(\xi) - f(\xi)|^2, \tag{3.13}$$

$$R_1(T) := \int_T |Ls - f|, \tag{3.14}$$

$$R_2(T) := \int_T |Ls - f|^2, \tag{3.15}$$

where $A(T)$ is the area of T. The last two of these are integrals, and need to be computed by numerical quadrature, see e.g. Sect. 4.6 of [38]. In practice we typically take $m = d + 2$.

The adaptive Ritz-Galerkin method can be summarized in the following:

Algorithm 3.

1. *Pick a type of spline space $S(\triangle)$*
2. *Pick an initial triangulation $\triangle$*
3. *Use the Ritz-Galerkin method to find a spline s in $S(\triangle)$ giving an approximate solution of the boundary value problem*
4. *Apply Algorithm 2 to successively refine $\triangle$ either a fixed number of times, or until some error measure is sufficiently small, where the* `Appx` *step in Algorithm 2 employs the Ritz-Galerkin approximation process in Step 3*

3.3 The Ritz-Galerkin Method

We recall briefly here the Ritz-Galerkin Method coupled with splines on triangulations in solving our model boundary value Problem 1.

Definition 3.3.1. *Suppose the spline s in (3.11) is such that*

$$\int_\Omega [Ls(x,y) - f(x,y)]\phi_i(x,y)\,dx\,dy = 0, \qquad i = 1,\ldots,n_0. \tag{3.16}$$

Then s is called the Ritz-Galerkin approximation to the solution u of Problem 1.

We recall some notation from [38]. For each triangle $T \in \triangle$, let

$$\langle \phi, \psi \rangle_{2,T} := \int_T \phi\,\psi\,dx\,dy,$$

and

$$\begin{aligned}\langle \phi, \psi \rangle_{G,T} &:= \int_T \kappa(x,y)\nabla\phi(x,y)\cdot\nabla\psi(x,y)\,dx\,dy \\ &= \int_T \kappa(x,y)[\phi_x(x,y)\psi_x(x,y) + \phi_y(x,y)\psi_y(x,y)]\,dx\,dy.\end{aligned}$$

Let

$$\langle \phi, \psi \rangle_2 := \sum_{T\in\triangle} \langle \phi, \psi \rangle_{2,T}, \qquad \langle \phi, \psi \rangle_G := \sum_{T\in\triangle} \langle \phi, \psi \rangle_{G,T}.$$

Then as shown in Theorem 9.5 of [38], the coefficients of the spline s in (3.11) approximating u via the Ritz-Galerkin method can be computed by solving the linear system of equations

$$Mc = r,$$

with $M = [\langle \phi_i, \phi_j \rangle_G]_{i,j=1}^{n_0}$ and

$$r_i = \langle f, \phi_i \rangle_2 - \langle s_b, \phi_i \rangle_G, \qquad i = 1,\ldots,n_0.$$

To use this approach with the space $S_d^0(\triangle)$ on an H-triangulation, we can follow the algorithm outlined in Sect. 9.3.1 of [38] to set up the stiffness matrix M and right-hand side vector r. This algorithm makes use of the transformation matrix A for the spline space.

3.4 Numerical Examples

Here are a few examples showing the performance of the above algorithms. We first illustrate the capabilities of adaptive methods coupled with H-triangulations in function approximation.

Example 3.4.1. *Fit the function* $f_1 = e^{-500[(x-.375)^2+(y-.375)^2]}$ *on the unit square by a spline in* $S_d^0(\triangle)$.

Discussion: A plot of the function f_1 is shown in Fig.3.4.1b. We begin with the type-2 triangulation of $[0,1]^2$ shown in Fig.3.4.1a. It has 64 triangles. Then for fixed d, we apply Algorithm 2 using the local error indicator defined in (3.6) with $m = d+2$. In each pass through the algorithm we identify the triangles corresponding to the top $p = 5$ percent of the errors, and split each such triangle uniformly into four subtriangles, see method S2 in Sect. 3.2. Table 3.4.1 illustrates the behaviour of the algorithm for $d = 4$, where

$nr :=$ number of passes through the algorithm,

$nt :=$ number of triangles of the final triangulation,

$ndof :=$ number of degrees of freedom of the associated spline space,

$emax :=$ maximum error on a 1001×1001 grid,

$RMS :=$ the root mean square error over that grid,

$time :=$ number of seconds to perform the computation.

With $nr = 5$ we get the triangulation shown in Fig.3.4.1c. It has 136 triangles, and gives maximum and RMS errors of 1.54(-3) and 4.96(-5). We do not show a plot of the associated spline surface since it is visually indistinguishable from a plot of f_1 itself. With $nr = 10$, we get the triangulation shown in Fig.3.4.1d which has 220 triangles. Now the errors are 8.64(-5) and 3.49(-6). As we can see from the table, as nr increases, we get more triangles, but the accuracy improves rapidly in both the maximum and RMS norms.

Table 3.4.2 shows the case $d = 5$. We get essentially the same accuracy with 340 triangles and 4161 degrees of freedom as we do with $d = 4$ and 808 triangles and 6369 degrees of freedom. For these examples we have used the local error indicator given in (3.6). Using the other local error indicators defined above give similar results. □

nr	nt	ndof	emax	RMS	time
5	136	1089	1.54e-03	4.96e-05	0.14
10	220	1737	8.64e-05	3.49e-06	0.40
15	388	3017	3.65e-05	1.16e-06	0.85
20	508	3993	2.04e-05	4.84e-07	1.56
30	808	6369	2.97e-06	1.04e-07	3.34

Table 3.4.1: Table of errors for Example 3.4.1 with $d = 4$

nr	nt	ndof	emax	RMS	time
5	136	1701	6.12e-04	1.22e-05	0.15
10	244	3001	1.53e-05	4.82e-07	0.47
15	340	4161	6.04e-06	1.52e-07	0.95
20	532	6561	1.71e-06	4.01e-08	1.71
30	952	11701	2.53e-07	5.47e-09	4.93

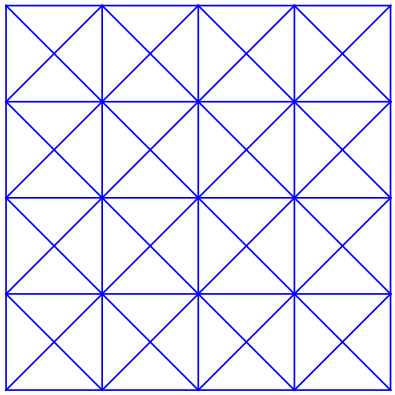
Table 3.4.2: Table of errors for Example 3.4.1 with $d = 5$

To illustrate the amount of compression achieved by the adaptive algorithm, one can compare with the interpolants from $S_4^0(\triangle)$ based on large type-2 triangulations. For example, using a type-2 triangulation defined on a 17×17 grid gives an RMS error of 7.52(-5), but with 1024 triangles and 8321 degrees of freedom. We achieve a better fit with only 136 triangles and 1089 degrees of freedom. Similarly, using a type-2 triangulation on a 33×33 grid gives an RMS error of 2.42(-6), but with 4096 triangles and 33025 degrees of freedom. We achieve a better fit with only 388 triangles and 3017 degrees of freedom.

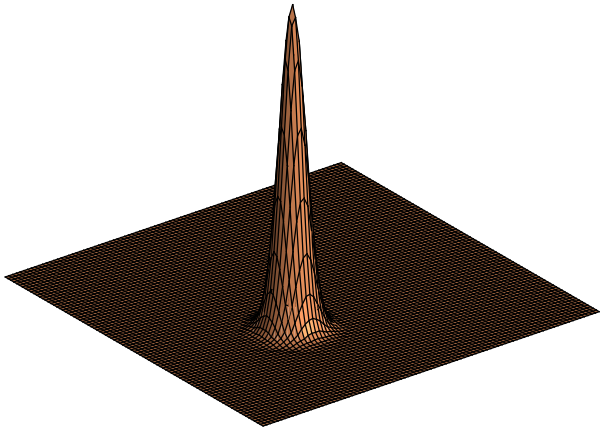
As a second example we take a test function which we shall use later in our study of boundary value problems.

Example 3.4.2. *Fit the function* $f_2 := \tanh(40y - 80x^2) - \tanh(40x - 80y^2)$ *on the unit square by a spline in* $S_d^0(\triangle)$.

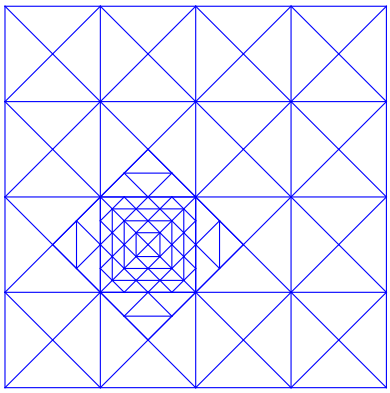
Discussion: A plot of the function f_2 is shown in Fig.3.4.2d. This time we begin with the type-1 triangulation of $[0,1]^2$ on a 5×5 grid shown in Fig.3.4.2a. It has 32 triangles. We then apply Algorithm 2 using the local error indicator defined in (3.6) with $m = d + 2$. In each run of the algorithm we identify the triangles corresponding to the top $p = 5$ percent of the errors, and split each such triangle uniformly into four subtriangles, see method S2 in Sect. 3.2. Tables 3.4.3 – 3.4.5 give numerical results for $d = 3, 5, 7$. They contain the same information as given in Tables 3.4.1 – 3.4.2. As to be expected, larger values of d give smaller errors. We have also experimented with the other local error indicators defined earlier, and get



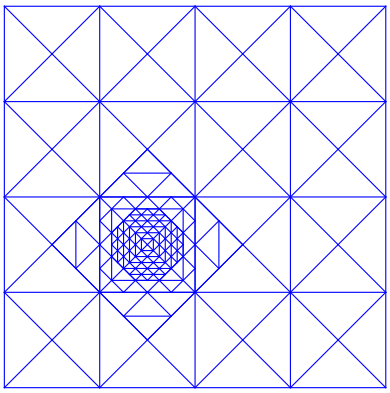
(a) $nt = 64$



(b) The peak function f_1



(c) $nt = 136$



(d) $nt = 220$

Figure 3.4.1: Results for Example 3.4.1 with $d = 4$

similar results.

For comparison purposes, we note that interpolation with $S_7^0(\triangle)$ on a type-1 triangulation of a 17×17 grid has 512 triangles and an RMS error of 3.85(-3). This can be compared with our adapted triangulation with 230 triangles which has an RMS error 8.71(-4). Interpolation with $S_7^0(\triangle)$ on a type-1 triangulation of a 33×33 grid has 2048 triangles and an RMS error of 1.2(-4). This can be compared with our adapted triangulation with 1079 triangles which has an RMS error of 1.32(-6). □

nr	nt	ndof	emax	RMS	time
10	134	598	3.94e-01	2.89e-02	0.43
20	248	1072	1.60e-01	1.03e-02	0.70
30	422	1819	1.03e-01	3.93e-03	1.59
40	632	2704	5.71e-02	1.58e-03	3.08
50	860	3688	1.80e-02	8.20e-04	5.23
60	1100	4756	1.37e-02	5.08e-04	8.02
70	1448	6262	1.16e-02	3.24e-04	11.68
80	1928	8332	5.52e-03	1.61e-04	17.07

Table 3.4.3: Table of errors for Example 3.4.2 with $d = 3$

nr	nt	ndof	emax	RMS	time
5	68	876	7.14e-01	4.67e-02	0.10
10	128	1586	2.90e-01	1.02e-02	0.28
20	254	3101	8.40e-02	2.54e-03	1.01
30	398	4851	2.56e-02	5.81e-04	2.32
40	578	6951	1.01e-02	2.18e-04	4.49
50	806	9771	3.66e-03	6.75e-05	7.97
60	980	11931	1.44e-03	2.73e-05	12.70

Table 3.4.4: Table of errors for Example 3.4.2 with $d = 5$

We also give two examples of the use of the Ritz-Galerkin approach coupled with our algorithm for locally refining triangulations.

For the examples in this section we will work with the *a posteriori* error estimator R_2 defined in (3.15). Our first example deals with the Poisson equation where the differential operator is the Laplace operator.

Example 3.4.3. *Use the spline spaces $S_d^0(\triangle)$ to solve the boundary value Problem 1 on $\Omega := [0,1]^2$ with $\kappa = 1$, $f := -\Delta u$, and $g = u|_{\partial\Omega}$, where $u := \tanh(40y - 80x^2) - \tanh(40x - 80y^2)$.*

Discussion: The true solution of this boundary value problem u is the same as the function f_2 used in Example 3.4.2, see Fig.3.4.2d. In running Algorithm 2 we sort the triangles according to the size of the error estimates, then refine the top 5% using the refinement method S2 described above. We repeat this *nr* times to get our solution.

Table 3.4.6 gives numerical results for the case $d = 3$. It has the same columns as in our previous tables, where again we use a 1001×1001 grid on Ω to compute the max and RMS errors. The cases $d = 5, 7$ are shown in Tables 3.4.7 – 3.4.8.

nr	nt	ndof	emax	RMS	time
10	98	2395	3.53e-01	1.34e-02	0.31
20	230	5496	4.34e-02	8.71e-04	1.31
30	362	8723	1.17e-02	1.42e-04	3.27
40	536	12818	2.86e-03	4.10e-05	6.39
50	686	16430	9.43e-04	1.35e-05	11.82
60	884	21218	2.67e-04	3.80e-06	19.64
70	1079	26013	1.71e-04	1.32e-06	29.69

Table 3.4.5: Table of errors for Example 3.4.2 with $d = 7$

For comparison purposes, we note that working with $S_7^0(\triangle)$ on a type-1 triangulation of a 33×33 grid with 2048 triangles and 50625 degrees of freedom gives an RMS error of 8.26(-5). This can be compared with using our adapted triangulation with 488 triangles and 11614 degrees of freedom which leads to an RMS error 4.54(-5).

We do not give any plots of the splines produced in this example or their associated triangulations since they are very similar to those obtained in Example 3.4.2 where we computed an interpolating spline for u. For example, for $d = 7$ using $nr = 40$ here we get a triangulation with 374 triangles and an associated RMS error of 2.28(-4), while in Table 3.4.5 with $nr = 30$ we had 362 triangles, and an RMS error of 1.42(-4). Thus, suprisingly the Ritz-Galerkin solution has virtually the same accuracy as the interpolating spline fit to the true solution. □

nr	nt	ndof	emax	RMS	time
10	104	445	8.85e-01	1.16e-01	2.96
20	218	907	2.00e-01	1.99e-02	9.72
30	392	1612	8.93e-02	1.03e-02	22.32
40	518	2137	8.15e-02	7.74e-03	41.68
50	626	2587	4.55e-02	3.18e-03	66.32
60	872	3598	4.28e-02	2.43e-03	102.46

Table 3.4.6: Table of errors for Example 3.4.3 with $d = 3$

In this example we have worked with the space $S_7^0(\triangle)$ along with the *a posteriori* estimator R_1 given in (3.15). The performance is not as good with other error estimators listed above, see Remark 14. Our second example is a boundary value problem defined on an L-shaped domain with a reentrant corner.

Example 3.4.4. *Use the spline space $S_3^0(\triangle)$ to solve the boundary value Problem 1 on the L-shaped domain*

nr	nt	ndof	emax	RMS	time
10	92	1136	7.89e-01	7.55e-02	3.70
20	182	2181	6.40e-02	5.57e-03	12.31
30	284	3381	4.13e-02	2.19e-03	27.53
40	374	4446	2.34e-02	1.11e-03	51.45
50	458	5466	1.78e-02	8.01e-04	82.55
60	596	7136	6.59e-03	3.20e-04	123.81

Table 3.4.7: Table of errors for Example 3.4.3 with $d = 5$

nr	nt	ndof	emax	RMS	time
10	104	2500	1.98e-01	1.65e-02	5.87
20	188	4467	6.61e-02	2.88e-03	20.09
30	248	5909	7.96e-03	6.06e-04	44.11
40	374	8856	4.84e-03	2.28e-04	81.46
50	488	11614	1.02e-03	4.54e-05	139.34

Table 3.4.8: Table of errors for Example 3.4.3 with $d = 7$

shown in Fig.3.4.3a with $\kappa \equiv 1$*,* $f \equiv 0$*, and Dirichlet boundary values taken from the function*

$$u(x,y) = r(x,y)^{2/3} \sin(\frac{2}{3}\theta(x,y)),$$

where $(r(x,y),\theta(x,y))$ *are the polar coordinates of* (x,y) *for all* $\theta \in [0,\frac{3}{2}\pi]$.

Discussion: This example corresponds to Example 9.13 in [38]. Here we begin with the triangulation shown in Fig.3.4.3a. It has 12 triangles. As in the previous example we use the *a posteriori* error estimator R_2 given in (3.15), and run Algorithm 2 with $p = 5$, again using the local refinement method S2. We repeat this *nr* times to get our solution.

Table 3.4.9 gives numerical results for this example. where here we report errors on a 101×101 grid to save evaluation time. Otherwise, it has the same information as in our previous tables. We have stopped the table at $nr = 15$ since at this level the smallest triangle already has area 1.5(-5) and doing another round of refinement gives only a marginal improvement in errors. □

3.5 Remarks

Remark 1. *In the introduction we have referenced only a few of the many papers in the FEM literature which involve solving boundary value problems adaptively with some kind of piecewise polynomials defined*

nr	nt	ndof	emax	RMS	time
0	12	67	2.79e-02	2.85e-03	0.00
5	42	178	1.16e-02	1.07e-03	0.69
10	72	307	2.45e-03	1.40e-04	1.73
15	102	433	2.68e-04	6.37e-05	3.44

Table 3.4.9: Table of errors for Example 3.4.4 with $d = 3$

on triangulations with hanging vertices. However, as far as we know, the approach in all of these papers follows the standard engineering approach of working with a single reference triangle which is then mapped to the triangles of an H-triangulation of the physical domain. Here we are able to avoid working with such parametric maps by employing Bernstein–Bézier techniques.

Remark 2. *The reason for restricting our attention to H-triangulations that are regular is to prevent the situation where two triangles touch at a common vertex, while no other triangles share that vertex, see for example Fig. 4.1 in [24]. We need connectivity to connect the polynomial pieces on adjoining triangles.*

Remark 3. *It can be shown that if $\triangle$ is an H-triangulation with no cycles, then the H-triangulation $\widetilde{\triangle}$ obtained applying either of the refinement methods discussed in Sect. 3.2 does not have any cycles, see [40].*

Remark 4. *The construction of minimal determining sets for $S_d^r(\triangle)$ on H-triangulations was carried out in [40] only for $d \geq 4r+1$. It is possible to extend this to all $d \geq 3r+2$ by arguments similar to those used for splines on ordinary triangulations, see [24].*

Remark 5. *For splines on ordinary triangulations, the columns of the transformation matrix A give the coefficients of the individual dual basis functions $\{\phi_\xi\}_{\xi \in \mathscr{M}}$, see [36]. The same holds here.*

Remark 6. *Suppose e is a composite edge of $\triangle$, and suppose $e_1, \ldots, e_m$ is a maximal sequence of composite edges such that for each $i = 1, \ldots, m$, one end of e_i is in the interior of e_{i+1}, where $e_{m+1} = e$. Following [40], we call $e_1, \ldots, e_m$ a chain ending at e. We refer to m as the length of the chain.*

Remark 7. *Theorem 3.1.2 gives an error bound on the entire domain Ω. There is an analogous local result which gives a bound on a single triangle. For a constructive proof, see [40]. We have stated the global form here for simplicity.*

Remark 8. *If we start with an ordinary triangulation and repeatedly locally refine it using the refinement method S2, the constant $\alpha_\triangle$ in (3.3) can be bounded by $1/m$, where m is the maximum number of times a*

given edge is refined. We also note that if we use this refinement process, the smallest angle in the resulting triangulation cannot be smaller than the smallest angle in the original triangulation.

Remark 9. *Here we have focused only on two possible refinement strategies. As is well known in the literature, see e.g. [24], and references therein, in some applications it may be desirable to split triangles in different ways – e.g., with the so-called Powell-Sabin split into either 6 or 12 subtriangles, or the Clough-Tocher split into 3 subtriangles.*

Remark 10. *In using local refinement with H-triangulations, it will happen that eventually very small triangles will be introduced. Fortunately, if we use the refinement strategy S2, we don't have to worry about the angles, but if the triangles are too small it is hard to keep track of the triangle lists or even to decide whether a point is in a triangle or not. Thus, in our test codes we have avoided splitting any triangle whose area is below a threshold (which we took to be* 10^{-7}*).*

Remark 11. *All experiments here were run in Matlab on a typical desktop. The times reported are just for comparison purposes – they of course would be different for other machines.*

Remark 12. *In the interest of saving time, our experiments have all been run with Algorithm 2 where we process the top 5% of triangles in turn, before recomputing the approximation and associated error. If we work with a larger percentage, the algorithms will run faster, but with slightly less accurate results. Conversely, if we work with a smaller percentage, computational times will go up, but we will get somewhat better errors. The best would be to do only one triangle at a time.*

Remark 13. *For both function approximation and solution of boundary value problems, the approximation process involves computing a minimal determining set. When an H-triangulation is refined locally, it is not necessary to recompute the entire MDS – one can simply update it in the affected triangulations. This would save time, but was not done in the reported results. The same is true of the second stage of the process, the computation of the approximation itself. For example, when solving boundary value problems, instead of computing the entire stiffness matrix anew, we can locally update it each time we have completed a refinement cycle. This would also save time, but again was not implemented for the results reported here.*

Remark 14. *The performance of the adaptive method of Sect. 3.3 for solving boundary value problems is somewhat sensitive to which* a posteriori *error estimator is used. Using the one given in (3.13), which while*

quite similar to the one in (3.15) used here, gives significantly different results. The others do not seem to function as well either.

Remark 15. *The experiments in this paper were all conducted using the midpoint refinement method S2. We get similar results for all of the examples if we use S1 instead, although since it only splits a triangle into two subtriangles instead of four, we have to work with larger numbers nr of iterations to get comparable numbers of triangles in the final triangulation.*

Remark 16. *The experiments of function approximation in Sect. 3.4 were carried out with C^0 spline spaces. We get very similar results when working with piecewise polynomials, i.e., splines with no global continuity. In fact, the algorithms run faster, and as soon as we get a reasonably accurate result, the lack of continuity cannot be detected visually.*

Remark 17. *In refining a single triangle from a triangulation $\triangle$ with either of subdivision methods discussed here, we have chosen the new vertices to be midpoints of existing edges of $\triangle$. This helps maintain the shape of the triangles, but also has the effect that as we continue to refine, these new hanging vertices have a tendency to disappear as we refine a neighboring triangle. The practical effect is that our final H-triangulations don't usually include a large percentage of hanging vertices. This feature also helps avoid long chains of hanging vertices (see [40]).*

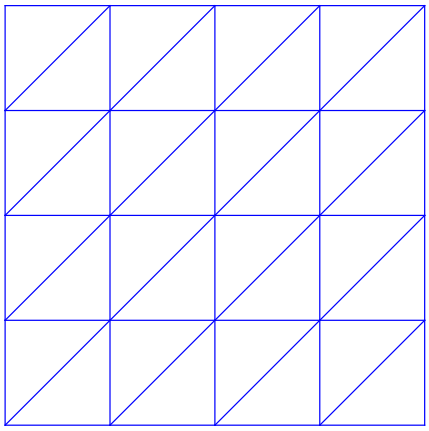
Remark 18. *The Ritz-Galerkin method can also be used to solve higher order boundary value problems. A typical example is the biharmonic equation which is of order four, see Sect. 9.6 of [38]. For conformality, this problem requires working with splines with global smoothness C^1. This is a good example where the space $S_5^{1,2}(\triangle)$ can be of use, and we can apply the same adaptive approach in this setting.*

Remark 19. *The key to computing with splines on H-triangulations using Bernstein–Bézier techniques is to work only with spaces where we know a minimal determining set and can construct the transformation matrix A. This is the case for the spaces $S_d^0(\triangle)$ for all d, as well as for the superspline spaces $S_d^{r,\rho}(\triangle)$ with $0 \le 2r \le \rho < d$ with $d > 2\rho$. It also holds for the analogs of all of the macro-element spaces discussed in [24] defined over Powell-Sabin refinements, Clough-Tocher refinements, etc.*

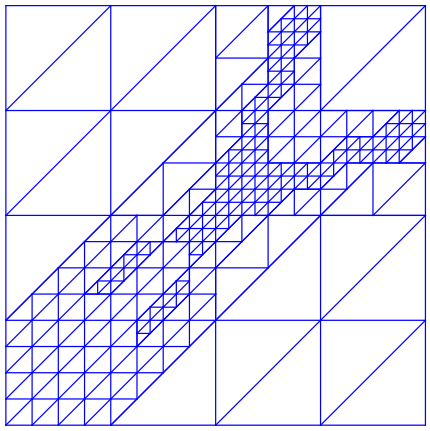
Remark 20. *In this paper we have focused on splines on H-triangulations in the plane. It is also possible to define splines on H-triangulations on the sphere, see [37]. Indeed, everything done there can just as easily be done on the sphere.*

Remark 21. *Splines on triangulations can also be used to approximate images, for example using piecewise linear polynomials with C^0 or C^{-1} global continuity, see e.g. [13, 12, 23]. Typically such algorithms involve starting with a coarse triangulation and adding new vertices adaptively, after which a Delaunay triangulation is constructed using the resulting vertices. Then the image is stored (and transmitted) by giving only the locations of the vertices and the associated coefficients. It is also possible to approximate images with splines on H-triangulations using the ideas of Sect. 3.2. With our approach, we need only store an ordered list of the triangles which have been split, along with the final set of coefficients. We have done some preliminary experimentation with this approach, but to compare with previous methods one has to properly encode this information. We plan to look into this further in later work.*

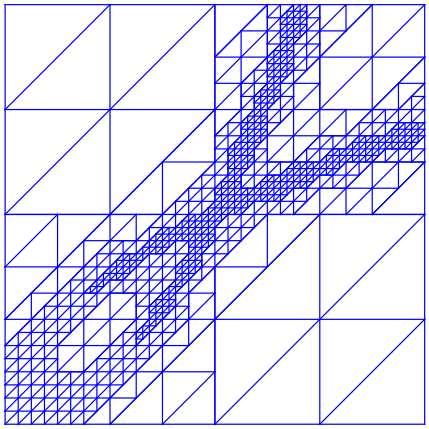
Remark 22. *A completely different approach to solving boundary value problems adaptively with splines has been proprosed recently in [25]. Their method works with ordinary triangulations, and the refinement algorithm is global rather than local. To make it easier to compare the results in [25] (which focus on the Poisson problem) with ours, in Example 3.4.3 we solve the same BVP as they do in Example 1 of their paper. For this BVP, their Table 2 shows an RMS error on a* 1001×1001 *grid of 3.31(-3) for a spline $s \in S_7^0(\triangle)$, based on an adapted triangulation with 1568 triangles. For this problem we got a spline in $S_7^0(\triangle)$ on an H-triangulation with only 200 triangles with an RMS accuracy of 2.64(-3), see our Table 3.4.6.*



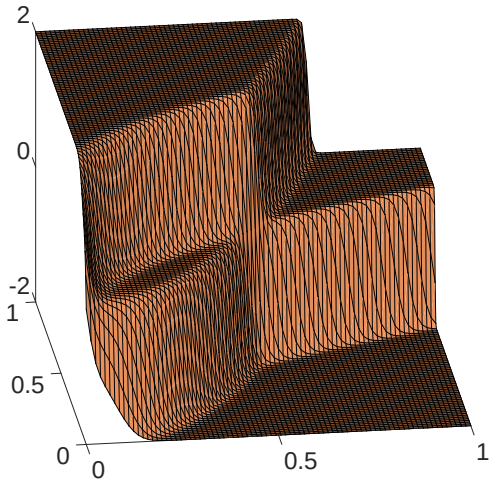
(a) $nt = 32$



(b) $nt = 362$

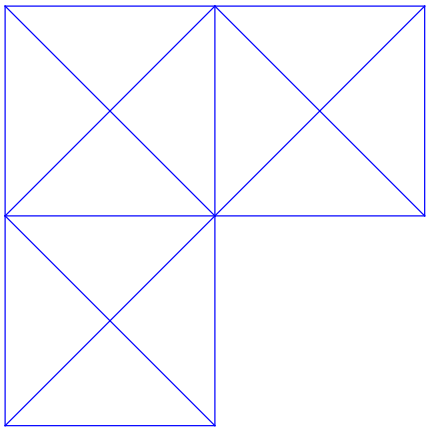


(c) $nt = 1079$

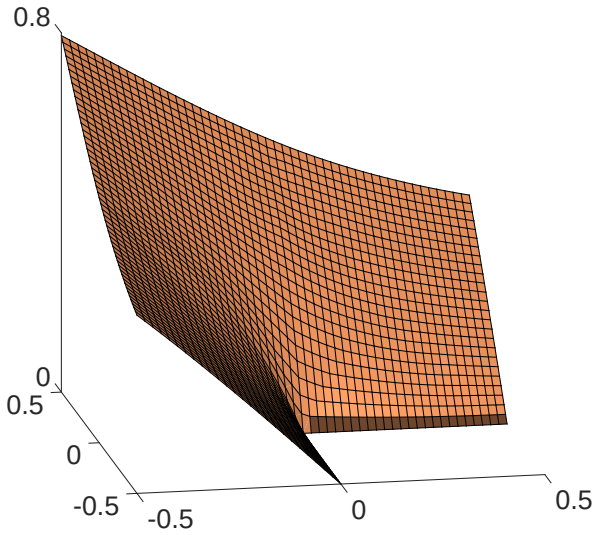


(d) The function f_2

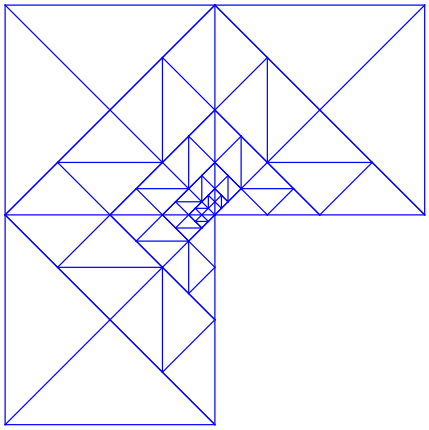
Figure 3.4.2: Results for Example 3.4.2 with $d = 7$



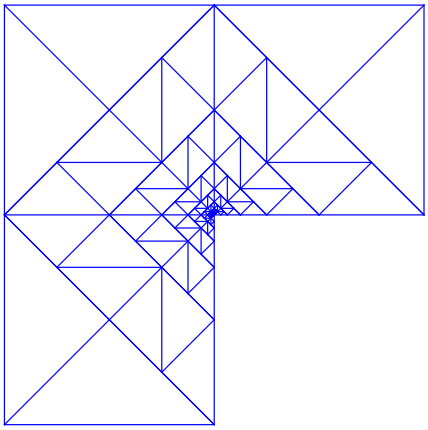
(a) $nt = 12$



(b) The solution u



(c) $nt = 72$



(d) $nt = 102$

Figure 3.4.3: Results for Example 3.4.4

Chapter 4

Collocation Methods

4.1 Notations

Let $I = [a,b]$ be an interval and $\{\xi_i\}_{i=1}^n \subseteq I$ be a set of discrete points. Define the following linear operators associated with a vector $\xi := (\xi_1, ..., \xi_n)^T$ and $\xi_\otimes := \big((\xi_1,\xi_1), ..., (\xi_1,\xi_n), ..., (\xi_n,\xi_1), ..., (\xi_n,\xi_n)\big)^T$:

- $\mathscr{R}_\xi : C(I) \to \mathbb{R}^n$, the restriction operator defined by $\mathscr{R}_\xi u := (u(\xi_1), ..., u(\xi_n))^T$.
- $\mathscr{D}_\xi^i : C^2(I) \to \mathbb{R}^n$, the "$i$-th derivative" operator defined by

  $$\mathscr{D}_\xi^i := (D^i u(\xi_1), \cdots, D^i u(\xi_n))^T,$$

  $i = 0, 1, 2$. Note that D^0 is the usual point evaluation operator and D^i denotes the i-th derivative operator.
- $\mathscr{L}_\xi : C^2(I) \to \mathbb{R}^n$, an evaluation operator associated with a second-order differential operator L acting on functions of one variable defined by

  $$\mathscr{L}_\xi u := (Lu(\xi_1), ..., Lu(\xi_n))^T.$$

- $\mathscr{L}_{\xi_\otimes} : C^2(I \times I) \to \mathbb{R}^{n^2}$, an evaluation operator associated with a second-order differential operator L acting on functions of two variables defined by

  $$\mathscr{L}_{\xi_\otimes} u := \big(Lu(\xi_1,\xi_1), ..., Lu(\xi_1,\xi_m), ..., Lu(\xi_m,\xi_1), ..., Lu(\xi_m,\xi_m)\big)^T.$$

  Definitions of linear operators $\mathscr{D}^x_{\xi_\otimes}, \mathscr{D}^y_{\xi_\otimes}, ..., \mathscr{D}^{yy}_{\xi_\otimes}$ etc. are analogous with L being the differential operators $D_x, D_y, ..., D_{yy}$ etc., respectively. Similarly, we define $\mathscr{R}_{\xi_\otimes} : C^2(I \times I) \to \mathbb{R}^{n^2}$ by

  $$\mathscr{R}_{\xi_\otimes} u := \big(u(\xi_1,\xi_1), ..., u(\xi_1,\xi_m), ..., u(\xi_m,\xi_1), ..., u(\xi_m,\xi_m)\big)^T.$$

Note that we will abuse notation and sometimes denote $\xi = \{\xi_i\}_{i=1}^n$ and $\xi_\otimes := \{(\xi_i,\xi_j)\}_{i=1,j=1}^{n,n}$ for conve-

nience in describing the set of points being used, the meaning of which will be clear in the context.

Other common notations and conventions:

- If A is a $n \times n$ matrix, the ∞-norm of A is defined as $||A||_\infty := Sup\{||Ax||_\infty : x \in \mathbb{R}^n, ||x||_\infty = 1\}$.
- Given a non-negative interger l, the $W^{l,\infty}$-norm of a function f on an interval I is the usual Sobolev norm, i.e., $||f||_{W^{l,\infty}(I)} = \max_{i=0,..l} ||D^i f||_{L^\infty(I)}$. The derivatives should be interpreted in the weak sense when needed.
- Let S_h be a spline space on a partition of mesh size h. If a method produces successive approximations $s_h \in S_h$ to a function u and the error bound, i.e., $s_h - u$ measured in a certain norm, is $\mathcal{O}(h^k)$, we say that the convergence rate (or order) of the method or of these successive approximations to u is k.

4.2 Collocation with Cubic Splines

We are interested in the following model problem:

Problem 2. *Find a function u defined on $I = [a,b]$ such that*

$$Lu := -u''(x) + \alpha(x)u'(x) + \beta(x)u(x) = f(x), \qquad on\ (a,b) \tag{4.1}$$

$$u(a) = u_a, \quad u(b) = u_b, \tag{4.2}$$

where $\alpha, \beta \in C^2(I)$ and $f \in C(I)$. Throughout this section, we assume that Problem 2 has a unique solution u in $C^2(I)$ for any real pairs $\{u_a, u_b\}$.

Given the knots of a partition $\triangle := \{x_i\}_{i=0}^{k+1}$, where $a = x_0 < x_1 < \cdots < x_k < x_{k+1} = b$, we define the mesh size $h := \max_{i=1,...k+1} |x_i - x_{i-1}|$ and chose the corresponding extended partition $\triangle_e := \{y_i\}_{i=1}^{n+d+1}$ for $S_d^r(\Delta)$, see section 2.1 for more details about these notations and definitions. We recall that d is the degree and n is the dimension of the spline space. One good choice of collocation points is the so-called *Greville points* $\{g_i\}_{i=1}^n$ associated with $\triangle_e$, defined as

$$g_i = \frac{y_{i+1} + \cdots + y_{i+d}}{d}, \quad i = 1, ..., n. \tag{4.3}$$

The *collocation method* associated with Problem 2 is as follows: given a partition $\triangle$ for $I = [a,b]$ and a

spline space $S_d^r(\triangle)$, choose a set of collocation points $\xi := \{\xi_i\}_{i=1}^m$ and find a spline s in $S_d^r(\triangle)$ such that

$$\mathscr{L}_\xi s = \mathscr{R}_\xi f \tag{4.4}$$

$$s(a) = u_a, \quad s(b) = u_b. \tag{4.5}$$

When the number of equations in this system is equal to the dimension of the spline space, i.e., $m = n-2$, the method is referred to as the exact collocation; while in the case where the number of equations is larger than the dimension of the spline space, i.e., $m > n-2$, we call the method the least-squares collocation if the above system is solved in the least-squares sense. We refer to the collocation method defined in (4.4) as ordinary collocation (OC) for convenience, since we will introduce a generalized collocation method later.

In this section we give three theorems related to the OC method using C^2 cubic splines in the case of exact collocation. The first two theorems are about the non-singularity of the collocation system with two different choices of the collocation points: i) the knots of the partition of the spline space being used ii) the interior Greville points associated with the partition of the spline space being used. A result about the convergence rates of the OC method using cubic splines defined on uniform partitions is presented in the third theorem below.

Lemma 4.2.1. *Let p be a cubic polynomial. Given a differential operator L of the form as in Problem 2, where α, β are C^2 functions defined on some interval $[a_1, a_2]$, there exists a constant C depending on the W_∞^2–norm of α and β on $[a_1, a_2]$ such that*

$$||Lp||_{L^\infty[a_1,a_2]} \leq C(h_1 + h_1^2)||p||_{L^\infty[a_1,a_2]} + (Ch_1 + Ch_1^2 + 2)\big(|Lp(a_1)| + |Lp(a_2)|\big), \tag{4.6}$$

where $h_1 = a_2 - a_1$.

Proof. By Markov brothers' inequality [28] (cf. Fact 4 in Appendix) we have $||p'||_{L^\infty[a_1,a_2]} \leq \frac{18}{h_1}||p||_{L^\infty[a_1,a_2]}$ for any $p \in \mathscr{P}_3$. Combining this estimate with the equality $p''(a_1) = \alpha(a_1)p'(a_1) + \beta(a_1)p(a_1) - Lp(a_1)$, we can see that there exists some constant C depending on the L^∞–norm of α and β such that

$$|p''(a_1)| \leq \frac{C(1+h_1)}{h_1}||p||_{L^\infty[a_1,a_2]} + |Lp(a_1)|.$$

A similar estimate works for $|p''(a_2)|$. We will abuse the notation of the constant C in this proof and C will

stand for a generic constant independent of h_1 and the polynomial p. Hence since p'' is a linear polynomial we have

$$||p''||_{L^\infty[a_1,a_2]} \leq \frac{C(1+h_1)}{h_1}||p||_{L^\infty[a_1,a_2]} + \max\{|Lp(a_1)|, |Lp(a_2)|\}.$$

We then estimate p''' which is a constant function. By the mean value theorem, there exists a number $t \in (a_1,a_2)$ such that $(Lp)'(t) = \frac{Lp(a_2)-Lp(a_1)}{a_2-a_1}$. Since $(Lp)' = -p''' + (\alpha p')' + (\beta p)'$, it follows that there exists a constant C depending on the $W^1_\infty(I)$–norm of α and β such that

$$||p'''||_{L^\infty[a_1,a_2]} \leq \frac{|Lp(a_2)-Lp(a_1)|}{h_1} + C\big(\frac{1+h_1}{h_1}||p||_{L^\infty[a_1,a_2]} + \max\{|Lp(a_1)|, |Lp(a_2)|\}\big).$$

Combining the equality $(Lp)'' = (\alpha p')'' + (\beta p)''$ with the estimates for p, p', p'', p''', we obtain that

$$||(Lp)''||_{L^\infty[a_1,a_2]} \leq C\frac{1+h_1}{h_1}\big(||p||_{L^\infty[a_1,a_2]} + |Lp(a_1)| + |Lp(a_2)|\big). \tag{4.7}$$

where C depends only on the W^2_∞-norm of α and β on $[a,b]$. Lastly, by the observation $||Lp||_{L^\infty[a_1,a_2]} \leq \frac{h_1^2}{2}||(Lp)''||_{L^\infty[a_1,a_2]} + |Lp(a_2)-Lp(a_1)| + |Lp(a_1)|$(cf. Fact 2 in Appendix), the proof is done.

□

In the remaining of this section, we denote the partition with mesh size h as $\triangle_h$ and the associated C^2 cubic spline space $S^2_3(\triangle_h)$ as S_h. In the following theorem, we consider the case when the collocation points coincide with the knots of the partition being used, i.e., $\xi_i = x_i,\ , i = 0, ..., k+1$.

Theorem 4.2.1. *Let $\mathscr{L}_\xi$ be the linear operator associated with Problem 2. For sufficiently small h , there exists a constant C independent of h such that*

$$||s||_{L^\infty[a,b]} \leq C||\mathscr{L}_\xi s||_\infty, \tag{4.8}$$

for any s in S_h such that $s(a) = s(b) = 0$, where $\xi = \{x_i\}_{i=0}^{k+1}$, i.e., the knots of $\triangle_h$.

Proof. Given a partition $\triangle_h$, let s_h be a cubic spline in S_h such that $s_h(a) = s_h(b) = 0$. Applying Lemma 4.2.1 to each subinterval $[x_{i-1}, x_i]$, there exists a constant C_1 depending on the W^2_∞-norm of α and β on $[x_{i-1}, x_i]$ such that

$$||Ls_h||_{L^\infty[x_{i-1},x_i]} \leq C_1(h+h^2)||s_h||_{L^\infty[x_{i-1},x_i]} + (C_1h + C_1h^2 + 2)\big(|Ls_h(x_{i-1})| + |Ls_h(x_i)|\big). \tag{4.9}$$

Hence there exists a constant C_2 depending on the W_∞^2–norm of α and β on $[a,b]$ such that

$$||Ls_h||_{L^\infty[a,b]} \leq C_2(h+h^2)||s_h||_{L^\infty[a,b]} + C_2(1+h+h^2)||\mathscr{L}_\xi s_h||_\infty. \tag{4.10}$$

Using the properties of the Green's function related to Problem 2 with homogeneous boundary conditions (cf. Fact 3 in Appendix), there exists a constant C_3 such that

$$||s_h||_{L^\infty[a,b]} \leq C_3||Ls_h||_{L^\infty[a,b]}, \tag{4.11}$$

where C_3 can be the maximum of the Green's function taken over the rectangle $[a,b]\times[a,b]$, and is hence independent of the mesh size h. Combining the previous two inequalities, we have

$$||s_h||_{L^\infty[a,b]} \leq C_4(h+h^2)||s_h||_{L^\infty[a,b]} + C_4(1+h+h^2)||\mathscr{L}_\xi s_h||_\infty, \tag{4.12}$$

where $C_4 = C_2C_3$. Let h be such that $C_4(h+h^2) \leq 1/2$, then

$$||s_h||_{L^\infty[a,b]} \leq (2C_4+1)||\mathscr{L}_\xi s_h||_{l^\infty}. \tag{4.13}$$

Since C_4 is independent of the mesh size h, the proof is finished. □

Now let us consider the case when the collocation points are the Greville points associated with the extended partitions, i.e., $\xi = \{g_i\}_{i=2}^{n-1}$ as defined in (4.3). Recall that here the degree $d=3$ and dimension of the spline space $n = k+4$. It is easy to see that

$$g_j = \begin{cases} \frac{2}{3}x_0 + \frac{1}{3}x_1, & j=2, \\ x_{j-2}, & j=3,\cdots,n-2, \\ \frac{1}{3}x_k + \frac{2}{3}x_{k+1}, & j=n-1. \end{cases} \tag{4.14}$$

Hence only the first and last collocation points are not knots of the partition being used.

Lemma 4.2.2. *We assume the same notations and assumptions as in Lemma 4.2.1. Let* $0<t<1$ *be a real number and c be a real number in interval* (a_1,a_2) *such that* $c=(1-t)a_1+ta_2$. *Assume that* h_1 *is small*

enough such that $\gamma = \frac{Ch_1(1+h_1)}{1-t} < 1$, *where C is the constant in Lemma 4.2.1. Then*

$$|Lp(a_1)| \leq \frac{\gamma}{1-\gamma}||p||_{L^\infty[a_1,a_2]} + \frac{|Lp(c)|}{(1-\gamma)(1-t)} + \frac{|Lp(a_2)|}{(1-\gamma)}(\gamma + \frac{t}{1-t}). \tag{4.15}$$

Proof. Since α, β are in $C^2[a_1,a_2]$ and applying Taylor's theorem to Lp,

$$|(1-t)Lp(a_1) - Lp(c) + tLp(a_2)| \leq \frac{h_1^2}{4}||(Lp)''||_{L^\infty[a_1,a_2]}. \tag{4.16}$$

Divide both sides by $1-t$ and recall inequality (4.7) and the constant C in Lemma 4.2.1, we have

$$|Lp(a_1) - \frac{Lp(c)}{1-t} + \frac{tLp(a_2)}{1-t}| \leq \frac{Ch_1^2}{1-t}||(Lp)''||_{L^\infty[a_1,a_2]} \tag{4.17}$$

$$\leq \frac{Ch_1(1+h_1)}{1-t}\big(||p||_{L^\infty[a_1,a_2]} + |Lp(a_1)| + |Lp(a_2)|\big). \tag{4.18}$$

Let h_1 be such that $\gamma < 1$. After using the triangle inequality and moving the term $\gamma|Lp(a_1)|$ to the left and dividing by $1-\gamma$ on both sides , we obtain the desired inequality (4.15). □

Theorem 4.2.2. *Let the notations and assumptions be the same as in Theorem 4.2.1 except that the collocation points are chosen as the corresponding interior Greville points, i.e.,* $\xi = \{g_i\}_{i=2}^{n-1}$ *in (4.14). Then, for sufficiently small h, there exists a constant C independent of h such that*

$$||s||_{L^\infty[a,b]} \leq C||\mathscr{L}_\xi s||_\infty \tag{4.19}$$

for any s in S_h *such that* $s(a) = s(b) = 0$.

Proof. Given a partition $\triangle_h$, let s_h be a cubic spline in S_h such that $s_h(a) = s_h(b) = 0$. Note that $g_2 = (1-t)x_0 + tx_1$ where $t = \frac{1}{3}$. Applying Lemma 4.2.2 to the first subinterval with $[a_1,a_2] = [x_0,x_1]$ we get the estimate for $Ls_h(x_0)$:

$$|Ls_h(x_0)| \leq \frac{\gamma}{1-\gamma}||s_h||_{L^\infty[x_0,x_1]} + \frac{3|Ls_h(g_2)|}{2(1-\gamma)} + \frac{|Ls_h(x_1)|}{(1-\gamma)}(\gamma + \frac{1}{2}), \tag{4.20}$$

where $h_1 = x_1 - x_0$ is chosen such that $\gamma = \frac{C_1h_1(1+h_1)}{1-t} < 1$ and C_1 is chosen as the constant in the statement of Lemma 4.2.2, which depends on the W_∞^2-norm of α and β on $[x_0,x_1]$. Similarly we can bound $|Ls_h(x_{k+1})|$

by an expression involving $||s_h||_{L^\infty[x_k,x_{k+1}]}, |Ls_h(g_{n-1})|$ and $|Ls_h(x_k)|$. Plugging (4.20) into (4.6) we obtain

$$||Ls_h||_{L^\infty[x_0,x_1]} \leq \frac{C_1h_1 + C_1h_1^2 + 2\gamma}{1-\gamma}||s_h||_{L^\infty[x_0,x_1]} + \frac{3(C_1h_1 + C_1h_1^2 + 2)}{2(1-\gamma)}\big(|Ls_h(g_2)| + |Ls_h(x_1)|\big).$$

Since h is the mesh size and $\gamma \leq \frac{3C_1h(1+h)}{2} \to 0$ as $h \to 0$, we can choose h sufficiently small such that for some constant C_2 independent of h

$$||Ls_h||_{L^\infty[x_0,x_1]} \leq C_2(h+h^2)||s_h||_{L^\infty[x_0,x_1]} + C_2(1+h+h^2)(|Ls_h(g_2) + Ls_h(g_3)|). \tag{4.21}$$

We note that $g_3 = x_1$. Similar estimates can be obtained for $||Ls_h||_{L^\infty[x_k,x_{k+1}]}$ and the estimates for $||Ls_h||_{L^\infty[x_{i-1},x_i]}$ are the same as in (4.9) for $i = 2,...,k$. The rest of the proof proceeds exactly as in Theorem 4.2.1.

□

We have the following corollary related to the existence and uniqueness of collocation solutions to Problem 2.

Corollary 4.2.1. *Let the assumptions and notations be the same as in Theorem 4.2.1. Given a partition $\triangle_h$ with small enough mesh size h and $\xi = \{x_i\}_{i=0}^{k+1}$ or $\xi = \{g_i\}_{i=2}^{n-1}$ as defined in the beginning of this section, the OC system (4.4) (4.5) has a unique solution for any given f in $C[a,b]$ and real numbers $\{u_a, u_b\}$.*

Proof. It is easy to see that the OC system (4.4) (4.5) is a square system given our choice of collocation points here. Hence to prove the existence of a unique solution to the system, it suffices to show that the only spline s in S_h that satisfies the following equations

$$\mathscr{L}_\xi s = \vec{0}, \tag{4.22}$$

$$s(a) = s(b) = 0 \tag{4.23}$$

is the zero spline, which follows directly from the inequality (4.8) or (4.19). □

We have shown that the OC system (4.4) and (4.5) has a unique solution in the cubic spline space $S_3^2(\triangle)$ if the mesh size of the partition $\triangle$ is sufficiently small and the collocation points are the knots or the associated Greville points. Now we present a result related to the convergence rate of the OC solutions in terms of the mesh size.

Theorem 4.2.3. *Let $\triangle_h = \{x_i\}_{i=0}^{k+1}$ be a uniform partition on $[a,b]$ with mesh size $h = \frac{1}{k+1}$ and let $S_h := S_3^2(\triangle_h)$ be the associated C^2 cubic spline space. Assume that for the OC system defined in (4.4) and (4.5) the knots or the associated interior Greville points are chosen as the collocation points. Let the true solution u of Problem 2 be in $C^j[a,b]$, $2 \leq j \leq 4$. Then, for sufficiently small h, there exists a constant C independent of h and u such that the solution $s_h \in S_h$ of (4.4) and (4.5) satisfies*

$$||s_h - u||_{L^\infty[a,b]} \leq Ch^{j-2}||D^j u||_{L^\infty[a,b]}. \tag{4.24}$$

Proof. Let Q be the quasi-interpolation operator defined in Example 2.4.1. Recall that $\mathscr{L}_\xi Qu = -D_\xi^2 Qu + (\mathscr{R}_\xi \alpha)\cdot D_\xi Qu + (\mathscr{R}_\xi \beta)\cdot \mathscr{R}_\xi Qu$ where "$\cdot$" denotes the entry-wise product between vectors. In light of the error bound for Q in (2.13), there exists a constant C depending on the L^∞-norm of α and β on $[a,b]$ such that

$$||\mathscr{L}_\xi(Qu-u)||_\infty \leq Ch^{j-2}||D^j u||_{L^\infty[a,b]}, \tag{4.25}$$

for $u \in C^j[a,b]$, with $2 \leq j \leq 4$. Choose h sufficiently small as in Theorem 4.2.1 or Theorem 4.2.2 so that the OC system (4.4) and (4.5) has a unique solution s_h. By the triangle inequality we have $||s_h - u||_{L^\infty[a,b]} \leq ||s_h - Qu||_{L^\infty[a,b]} + ||Qu-u||_{L^\infty[a,b]}$, where the second term on the right-hand side satisfies (2.13). By Theorem 4.2.1, there exists a constant C independent of h and $s_h - Qu$ such that

$$||s_h - Qu||_{L^\infty[a,b]} \leq C||\mathscr{L}_\xi(s_h - Qu)||_\infty, \tag{4.26}$$

since both s_h and Qu satisfy the boundary conditions of Problem 2 and therefore $s_h - Qu$ is zero on the boundary of $[a,b]$. Since s_h is the OC solution, i.e., $\mathscr{L}_\xi s_h = \mathscr{R}_\xi f = \mathscr{L}_\xi u$, we have

$$||s_h - Qu||_{L^\infty[a,b]} \leq C||\mathscr{L}_\xi(u - Qu)||_\infty \leq Ch^{j-2}||D^j u||_{L^\infty[a,b]}, \tag{4.27}$$

where the last inequality follows from (4.25). Combining (4.27) and (2.13) finishes the proof.

□

Hence in particular, if the true solution u of Problem 2 is in $C^4[a,b]$, the accuracy is $\mathscr{O}(h^2)$, which is verified by the numerical examples shown in later sections. These examples also show that the quadratic convergence rate is best possible in general using the OC coupled with cubic splines.

4.3 A Generalized Collocation Method for $L = D^2$

Let us start with a motivating example, with $S_h := S_3^2(\Delta)$ defined on a uniform partition Δ of interval $I = [0,1]$. Consider the following two-point boundary value problem

Problem 3. *Find a function u in $C^2(I)$ such that*

$$Lu := D^2 u = f, \qquad on\ (0,1) \tag{4.28}$$

$$u(0) = u_0, \quad u(1) = u_1, \tag{4.29}$$

where f is a continuous function on I.

The OC solution associated with S_h and interior Greville points $\xi = \{g_i\}_{i=2}^{n-1}$ is a spline $s \in S_h$ such that

$$D_\xi^2 s = \mathscr{R}_\xi f \tag{4.30}$$

$$s(0) = u_0, \quad s(1) = u_1. \tag{4.31}$$

From [38] it is clear that $DS_h := \{s'' |\ s \in S_h\}$ is the space of continuous piecewise linear functions defined on the partition Δ and hence, there exists a unique linear spline $\hat{s}$ such that

$$\hat{s}(g_i) = f(g_i), \quad i = 2, ..., n-1. \tag{4.32}$$

By elementary calculus we know that there exists a unique spline $s \in S_h$ such that $s'' = \hat{s}$, satisfying boundary condition (4.31). Hence, for a given function f, there exists a unique cubic spline in S_h solving the collocation equation (4.30) and satisfying boundary condition (4.31).

Although with the above choice of collocation points the non-singularity of the collocation system has been proved in Corollary 4.2.1, numerically we observe that the convergence rate of this method in terms of the mesh size h is only quadratic (see e.g., Tables **??** and **??**), despite the fact the approximation order of cubic splines is known to be four (cf. Sect. 2.5). This apparent drawback of OC leads us to consider a generalized collocation (**GC**) system:

$$BD_\xi^2 s = R_\xi f \tag{4.33}$$

subject to boundary condition (4.31), where $\xi = \{g_i\}_{i=2}^{n-1}$ and

$$B = \begin{bmatrix} \frac{172}{9} & -\frac{505}{18} & \frac{211}{27} & \frac{115}{54} & & & & \\ -\frac{9}{2} & 8 & -2 & -\frac{1}{2} & & & & \\ & \frac{1}{16} & \frac{85}{96} & \frac{1}{24} & \frac{1}{96} & & & \\ & & \frac{1}{12} & \frac{5}{6} & \frac{1}{12} & & & \\ & & & \frac{1}{12} & \frac{5}{6} & \frac{1}{12} & & \\ & & & & \ddots & \ddots & \ddots & \\ & & & & & & & \end{bmatrix}. \tag{4.34}$$

Here the $(n-2)\times(n-2)$ matrix B has been judiciously chosen to meet a number of criteria explained below. In particular, B depends on the location of the collocation points ξ and is essentially a Toeplitz matrix except for the first and last few rows. We prove several crucial properties of matrix B in the next two lemmas.

Lemma 4.3.1. *Matrix B defined above is non-singular and such that*

$$||B^{-1}||_\infty \leq 12, \tag{4.35}$$

i.e., the norm is not dependent on the size of the matrix (or the mesh size h).

Proof. First, let us eliminate the first entry in the second row and last entry in the second to the last row by performing two elementary row operations. We obtain a new matrix $\overline{B}$ such that $\overline{B} = EB$ where $||E||_\infty = 1 + \frac{9}{2} \times \frac{9}{172} \leq \frac{4}{3}$. We can write

$$\overline{B} = \begin{bmatrix} D_1 & B_1 & O_1 \\ O_2 & D_2 & O_2 \\ O_1 & B_2 & D_1 \end{bmatrix}, \tag{4.36}$$

where $D_1 = \begin{bmatrix} \frac{172}{9} \end{bmatrix}, B_1 = \begin{bmatrix} -\frac{505}{18} & \frac{211}{27} & \frac{115}{54} & 0 & \cdots & 0 \end{bmatrix}, B_2 = \begin{bmatrix} 0 & \cdots & 0 & \frac{115}{54} & \frac{211}{27} & -\frac{505}{18} \end{bmatrix}$, and O_1, O_2 are 1×1 are $(n-4)\times 1$ zero matrices, respectively. One can check that D_2 is a strictly diagonally dominant (SDD) matrix, hence is non-singular with the property $||D_2^{-1}||_\infty \leq \frac{3}{2}$, using a simple fact about SDD matrices, see Fact 1 in Appendix. Next we show that there exists some positive constant α, independent of the size of $\overline{B}$, such that for any $x \in \mathbb{R}^{n-2}$,

$$||x||_\infty \leq \alpha ||\overline{B}x||_\infty. \tag{4.37}$$

Let x_i denote the i-th entry of a vector x. We have for example $(\overline{B}x)_1 = ax_1 + bx_2 + cx_3 + dx_4$, where $a = \frac{172}{9}, b = -\frac{505}{18}$, etc. Noting that $\big((\overline{B}x)_2, \cdots, (\overline{B}x)_{n-3}\big)^T = D_2(x_2, \cdots, x_{n-3})^T$, it is easy to see that

$$||(x_2, \cdots, x_{n-3})^T||_\infty \leq \frac{3}{2}||\big((Bx)_2, \cdots, (Bx)_{n-3}\big)^T||_\infty, \tag{4.38}$$

using the estimate for $||D_2^{-1}||_\infty$. For the first entry of the vector x, let us consider two cases:

case 1: if $|x_1| > \frac{1}{a}|(\overline{B}x)_1|$, then $|x_1| > |x_1 + \frac{b}{a}x_2 + \frac{c}{a}x_3 + \frac{d}{a}x_4|$. Hence there exists some $0 < \beta < 2$ such that $\frac{b}{a}x_2 + \frac{c}{a}x_3 + \frac{d}{a}x_4 = -\beta x_1$. When $0 < \beta < \frac{1}{2}$, we have that $|x_1 + \frac{b}{a}x_2 + \frac{c}{a}x_3 + \frac{d}{a}x_4| = (1-\beta)|x_1|$, which indicates that $|x_1| < 2|(\overline{B}x)_1|$. When $\frac{1}{2} \leq \beta < 2$, it is easy to see that $|x_1| = \frac{1}{\beta}|\frac{b}{a}x_2 + \frac{c}{a}x_3 + \frac{d}{a}x_4| \leq 2(|\frac{b}{a}| + |\frac{c}{a}| + |\frac{d}{a}|)(x_2, \cdots, x_{n-3})^T||_\infty \leq 9||\big((\overline{B}x)_2, \cdots, (\overline{B}x)_{n-3}\big)^T||_\infty$, where the last inequality follows from (4.38). Obviously $|x_1| \leq 9||\overline{B}x||_\infty$.

case 2: if $|x_1| \leq \frac{1}{a}|(\overline{B}x)_1|$, then $|x_1| \leq \frac{1}{a}||\overline{B}x||_\infty$.

We can do a similar analysis for x_{n-2} and conclude that $||x||_\infty \leq \alpha||\overline{B}x||_\infty$, where $\alpha = \max\{\frac{1}{a}, 9, \frac{3}{2}\} = 9$. Thus $||\overline{B}^{-1}||_\infty \leq 9$. Since $||B^{-1}||_\infty \leq ||\overline{B}^{-1}||_\infty||E||_\infty$, it follows that $||B^{-1}||_\infty \leq 12$. □

Lemma 4.3.2. *Let $Q := Q_3$ be the quasi-interpolation operator defined in Sect. 2.4 and let ξ be the above choice of collocation points. Then $||BD_\xi^2 Qu - D_\xi^2 u||_\infty \leq Ch^4$, for any $u \in C^6(I)$, where the constant C depends only on u.*

Proof. Using the definition of Q, we can expand Qu as a linear combination of spline basis functions in S_h and evaluate the second derivatives of these basis functions at the collocation points ξ. Consequently, each entry of BD_ξ^2 becomes a linear combination of values of u. For each i, expanding each term of u around ξ_i of $(BD_\xi^2 Qu)_i$ up to the 6-th derivative $u^{(6)}$ by Taylor's theorem, one obtains that $(BD_\xi^2 Qu)_i = u''(\xi_i) + Ch^4$, where C is a constant related to the norm of the 6-th derivative of u. □

The following convergence result for (4.33) and (4.31) assumes the use of the spline space $S_h = S_3^2(\Delta)$ defined on a uniform partition Δ of I and interior Greville points as collocation points.

Theorem 4.3.1. *Let $u \in C^6(I)$ be the solution of Problem 3. The generalized collocation equation (4.33) has a unique solution $s_h \in S_h$ satisfying the boundary condition (4.31), and*

$$||s_h - u||_{L^\infty(I)} \leq Ch^4, \tag{4.39}$$

where the constant C depends only on u.

Proof. The existence and uniqueness of a solution to the generalized collocation system follows from that of the ordinary collocation system and the fact that matrix B is non-singular. Let $Q := Q_3$ as in Sec.2.4. We know from (2.13) that $||u - Qu||_{L^\infty} \leq C_u h^4$, where C_u is a constant depending on u. Since both s_h and Qu match the true solution u at the endpoints of the interval, by elementary calculus (see Fact 2 in Appendix) we have

$$||s_h - Qu||_{L^\infty} \leq C_1 ||D^2 s_h - D^2 Qu||_{L^\infty(I)} \tag{4.40}$$

$$\leq C_2 ||D^2_\xi s_h - D^2_\xi Qu||_\infty, \tag{4.41}$$

where $C_1 = \frac{1}{2}$ and $C_2 = 1$. The second inequality holds since $D^2 s_h$ and $D^2 Qu$ are continuous piecewise linear functions and all the collocation points except the first (which lies at one third of the first interval) are knots. Since $||B^{-1}||_\infty \leq 12$, inserting BB^{-1} into the last inequality above we get

$$||s_h - Qu||_{L^\infty(I)} \leq ||B^{-1}||_\infty ||BD^2_\xi s_h - BD^2_\xi Qu||_\infty \tag{4.42}$$

$$= ||B^{-1}||_\infty ||D^2_\xi u - BD^2_\xi Qu||_\infty \tag{4.43}$$

$$\leq 12Ch^4, \tag{4.44}$$

where (4.43) follows from the fact that s_h is the solution to the generalized collocation equation (4.33) and the constant C in the last inequality is the same as in Lemma 4.3.2. By the triangle inequality, $||s_h - u||_{L^\infty(I)} \leq ||s_h - Qu||_{L^\infty(I)} + ||Qu - u||_{L^\infty(I)}$, which finishes the proof. □

We next give a few numerical examples verifying the convergence results of the above theorem. In particular, we solve the system (4.33) and (4.31) for several test functions u using $S^2_3(\Delta)$ defined above by the GC method.

In all the tables in this section, `k` denotes the number of subintervals in the partition, `maxerr` and `rmserr` denote the maximum error and root-mean-square error respectively, measured on 1001 uniformly distributed points in I, and $(x-y)^n_+ := \max\{0, (x-y)^n\}$. The last three columns in the tables give convergence rates; `maxrate` and `rmsrate` denote the convergence rates of the maximum error and root-mean-square error, while the last column shows the convergence rate for $||BD^2_\xi Qu - D^2_\xi u||_\infty$. We observed

k	maxerr	rmserr	maxrate	rmsrate	$BD_\xi^2 Qu - D_\xi^2 u$ maxrate
10	1.00e-05	5.11e-06			
20	6.64e-07	3.36e-07	3.92	3.93	3.84
40	4.27e-08	2.14e-08	3.96	3.97	3.92
80	2.71e-09	1.35e-09	3.98	3.99	3.94
160	1.68e-10	8.46e-11	4.01	3.99	2.35
320	1.05e-11	4.82e-12	3.99	4.13	0.96
640	1.00e-11	6.89e-12	0.07	-0.51	-4.12

Table 4.3.1: $u = e^x$ on $[0,1]$

k	maxerr	rmserr	maxrate	rmsrate	$BD_\xi^2 Qu - D_\xi^2 u$ maxrate
10	1.67e-05	6.20e-06			
20	1.24e-06	5.13e-07	3.76	3.59	3.24
40	8.57e-08	3.61e-08	3.85	3.83	3.58
80	5.69e-09	2.35e-09	3.91	3.94	3.78
160	3.61e-10	1.49e-10	3.98	3.98	3.89
320	2.30e-11	9.45e-12	3.98	3.98	3.90
640	3.07e-12	2.28e-12	2.90	2.05	1.90

Table 4.3.2: $u = \log(1+x)$ on $[0,1]$

k	maxerr	rmserr	maxrate	rmsrate	$BD_\xi^2 Qu - D_\xi^2 u$ maxrate
10	2.28e-04	6.07e-05			
20	1.86e-05	7.31e-06	3.62	3.05	3.53
40	1.39e-06	5.71e-07	3.74	3.68	3.81
80	9.57e-08	3.80e-08	3.86	3.91	3.91
160	6.20e-09	2.42e-09	3.95	3.97	3.96
320	3.98e-10	1.52e-10	3.96	3.99	3.98
640	1.90e-11	9.53e-12	4.39	4.00	4.01

Table 4.3.3: $u = (x - \frac{1}{2})_+^7$ on $[0,1]$

optimal convergence (order 4) for cubic splines in all examples. Also, order 4 convergence is observed for $||BD_\xi^2 Qu - D_\xi^2 u||_\infty$ in Tables 4.3.1-4.3.4 when the true solution u is at least C^6, as assumed in Lemma 4.3.2. However, for u that is only C^4, this convergence rate is not clear, see Table 4.3.5. We remark that when the error goes down to around 1(-11), the convergence rate is most likely polluted by roundoff error.

4.4 A Model for the Generalized Collocation Methods

In the following let us consider a more general model problem:

k	maxerr	rmserr	maxrate	rmsrate	$BD_\xi^2 Qu - D_\xi^2 u$ maxrate
10	6.59e-04	2.27e-04			
20	5.17e-05	2.14e-05	3.67	3.41	3.69
40	3.71e-06	1.52e-06	3.80	3.81	3.87
80	2.50e-07	9.89e-08	3.89	3.94	3.94
160	1.59e-08	6.26e-09	3.97	3.98	3.97
320	1.02e-09	3.93e-10	3.97	3.99	3.98
640	4.83e-11	2.45e-11	4.39	4.00	3.92

Table 4.3.4: $u = (x - \frac{1}{\pi})_+^7$ on $[0,1]$

k	maxerr	rmserr	maxrate	rmsrate	$BD_\xi^2 Qu - D_\xi^2 u$ maxrate
10	2.74e-04	1.21e-04			
20	1.92e-05	8.01e-06	3.83	3.92	4.92
40	1.26e-06	5.14e-07	3.93	3.96	2.81
80	8.10e-08	3.21e-08	3.96	4.00	3.42
160	5.03e-09	2.05e-09	4.01	3.97	2.34
320	3.17e-10	1.28e-10	3.99	4.00	3.06
640	1.50e-11	7.89e-12	4.40	4.02	3.20

Table 4.3.5: $u = (x - \frac{1}{\pi})_+^5$ on $[0,1]$

Let $S_h := S_d^{d-1}(\Delta)$ be a spline space of dimension n defined on a partition Δ of the interval I and let $\xi = \{\xi_i\}_{i=1}^m$ ($m = n-2$) be a set of collocation points. The corresponding generalized collocation (GC) system associated with this problem is to find an $s \in S_h$, such that

$$\mathscr{B}_\xi s = \mathscr{R}_\xi f, \tag{4.45}$$

subject to the boundary condition $s(a) = u_a,\ s(b) = u_b$, with

$$\mathscr{B}_\xi s := -B_2 \mathscr{D}_\xi^2 s + (\mathscr{R}_\xi \alpha) \cdot B_1 \mathscr{D}_\xi^1 s + (\mathscr{R}_\xi \beta) \cdot \mathscr{R}_\xi s. \tag{4.46}$$

Here B_1, B_2 are $m \times m$ matrices and the "$\cdot$" means the entry-wise product. The following theorem gives a sufficient condition on the matrices B_1, B_2 that guarantees optimal convergence rate of the generalized collocation solutions.

Theorem 4.4.1. *Let C be a generic constant independent of the mesh size h or a specific spline s in S_h and let $I = [a,b]$. Let $Q : C(I) \to S_h$ be a quasi-interpolation operator such that*

i) $u(a) = Qu(a)$, $u(b) = Qu(b)$, *for all* $u \in C(I)$

ii) for all smooth enough u, $||u - Qu||_{L^\infty(I)} \leq Ch^{d+1}$.

Assume further that the matrix pair $\{B_1, B_2\}$ *in (4.46) has the property that for any* $s \in S_h$, *with* $s(a) = s(b) = 0$

$$||\mathscr{B}_\xi s||_\infty \geq C||s||_{L^\infty(I)}, \tag{4.47}$$

and

$$||\mathscr{B}_\xi Qu - \mathscr{L}_\xi u||_{l^\infty} \leq Ch^{d+1}. \tag{4.48}$$

Then the GC system is non-singular and the GC solution s_h *associated with the matrices* B_1, B_2 *satisfies*

$$||s_h - u||_{L^\infty} \leq Ch^{d+1}. \tag{4.49}$$

Proof. Consider the homogeneous system $\mathscr{B}_\xi s = \vec{0}$ subject to zero boundary conditions. It follows from (4.47) that s is the zero function. Since this GC system is a square linear system, i.e., the number of equations equals the dimension of the spline space, it has a unique solution for any input $\{\mathscr{R}_\xi f, u_a, u_b\}$. Using the triangle inequality we have $||s_h - u||_{L^\infty(I)} \leq ||s_h - Qu||_{L^\infty(I)} + ||Qu - u||_{L^\infty(I)}$, the second term of which is bounded by assumption ii) of Q. Hence we only need to estimate

$$\begin{aligned} ||s_h - Qu||_{L^\infty(I)} &\leq C||\mathscr{B}_\xi(s_h - Qu)||_\infty && (4.50)\\ &\leq C||\mathscr{L}_\xi u - \mathscr{B}_\xi Qu||_\infty && (4.51)\\ &\leq Ch^{d+1}, && (4.52) \end{aligned}$$

where (4.50) follows from (4.47) and the fact that s_h and Qu are both splines in S_h and they coincide at the endpoints of I. Inequality (4.51) follows from the fact that s_h is the GC solution to the problem, i.e., $\mathscr{B}_\xi s_h = \mathscr{R}_\xi f = \mathscr{L}_\xi u$, while the last inequality follows from (4.48). Again C stands for a generic constant. □

We summarize the GC method for solving Problem 2 as follows:

Algorithm 4.

1. *Pick a spline space $S_d^{d-1}(\Delta)(d \geq 3)$, where Δ is a uniform partition on $[a,b]$, and choose the collocation points ξ to be the interior Greville points associated with the corresponding extended partition $\triangle_e$.*

2. *Choose a quasi-interpolation operator Q with optimal approximation power, as in assumption ii) of Theorem 4.4.1.*

3. *Calculate the rows of B_1 and B_2 by forcing $B_1 D_\xi^1 Qp = D_\xi^1 p$ for polynomials p of degree $\leq d+2$ and $B_2 D_\xi^2 Qp = D_\xi^2 p$ for polynomials p of degree $\leq d+3$.*

4. *Set up the linear system according to (4.46) and solve it.*

We remark that this algorithm can be applied for non-uniform partitions and other choices of collocation points too. However, there are several advantages of using uniform partitions and Greville points:

i) there exists a local quasi-interpolation operator Q such that the coefficient matrix is essentially Toeplitz in the sense that its largest block is Toeplitz (cf. (4.34)). The matrix is also sparse.

ii) the relative position of the collocation points in each interval is fixed and all but the boundary the B-spline basis functions are translations of the cardinal B-splines, cf. e.g. [34], which makes the calculation of the values or derivatives of the basis functions at the collocation points very easy. In fact, there are only a few different non-zero values that needed to be calculated, the number of which depends on the degree d of the spline space only.

iii) by i) and ii) matrices B_1 and B_2 can be made into sparse matrices with some Toeplitz structure; hence except for the first and last few rows, the vectors consisting of the non-zero entries in each row are the same. These vectors of non-zero entries in each row can be calculated once and for all once the degree d is fixed.

Next, we show how to make use of the GC scheme to solve elliptic boundary value problems on a rectangle in $\mathbb{R}^2$. Consider the following model problem

Problem 4. *Find a function u defined on the unit square Ω such that*

$$Lu := -\nabla \cdot (\kappa \nabla u) = f \quad \text{on } \Omega, \tag{4.53}$$

$$u = g \quad \text{on } \partial\Omega, \tag{4.54}$$

where κ and f are given functions defined on Ω and g is a function defined on the boundary $\partial\Omega$.

Given a partition $\triangle$ on the interval $I = [0,1]$ and an integer $d \geq 3$, we employ a tensor-product spline space $S_\otimes := S_d^{d-1}(\triangle) \bigotimes S_d^{d-1}(\triangle)$ defined on Ω (cf. Sect. 2.3) as the approximation space. Note that the partitions in the x- and y- directions need not to be the same; we assume that $\triangle$ is the partition used in both directions for simplicity. Let $\xi = \{\xi_i\}_{i=1}^m$ be a set of collocation points in I according to Algorithm 4 and let $\xi_\otimes := \{(\xi_i, \xi_j)\}_{i=1,j=1}^{m,m}$. Expanding (4.53), we obtain

$$Lu = -\kappa(u_{xx} + u_{yy}) - \kappa_x u_x - \kappa_y u_y. \tag{4.55}$$

Let us define the following GC operator $\mathscr{B}_{\xi_\otimes} : S_\otimes \to \mathbb{R}^{m^2}$

$$\mathscr{B}_{\xi_\otimes} s := -(\mathscr{R}_{\xi_\otimes} \kappa) \cdot (B_2^x \mathscr{D}_{\xi_\otimes}^{xx} s + B_2^y \mathscr{D}_{\xi_\otimes}^{yy} s) - (\mathscr{R}_{\xi_\otimes} \kappa_x) \cdot (B_1^x \mathscr{D}_{\xi_\otimes}^{x} s) - (\mathscr{R}_{\xi_\otimes} \kappa_y) \cdot (B_1^y \mathscr{D}_{\xi_\otimes}^{y} s), \tag{4.56}$$

where $B_2^x, B_2^y, B_1^x, B_1^y$ are $m^2 \times m^2$ matrices and the "$\cdot$" means the entry-wise product. We note that B_2^x, B_2^y correspond to the $m \times m$ matrix B_2 in Algorithm 4 in the following sense: if the non-zero entries of the i-th row and j-th row of B_2 are $\{b_{i,i_1}, ..., b_{i,i_2}\}$ and $\{b_{j,j_1}, ..., b_{j,j_2}\}$ respectively, then, for $l = (i-1)m + j$

$$(B_2^x \mathscr{D}_{\xi_\otimes}^{xx} s)_l = \sum_{k=i_1}^{i_2} b_{i,k} s_{xx}(\xi_k, \xi_j), \tag{4.57}$$

$$(B_2^y \mathscr{D}_{\xi_\otimes}^{yy} s)_l = \sum_{k=j_1}^{j_2} b_{j,k} s_{yy}(\xi_i, \xi_k), \tag{4.58}$$

while B_1^x, B_1^y correspond to the matrix B_1 in Algorithm 4 in similar sense.

We propose the following GC scheme to solve Problem 4:

Algorithm 5.

1. *Pick a spline space $S_d^{d-1}(\triangle) \bigotimes S_d^{d-1}(\triangle) (d \geq 3)$, where Δ is a uniform partition on $[a,b]$, and choose the collocation points ξ to be the interior Greville points associated with the corresponding extended partition $\triangle_e$.*

2. *Calculate matrices B_1 and B_2 according to Step 3 in Algorithm 4.*

3. *Construct the sparse matrices $B_2^x, B_2^y, B_1^x, B_1^y$ according to (4.57) and (4.58).*

4. *Pick $n_b(\geq n)$ uniformly distributed points on each boundary edge of Ω and obtain the boundary degrees of freedom c_b by solving four least-squares problems using the data of g on each boundary edge, where $n = \dim S_d^{d-1}(\triangle)$.*

5. *Solve the following linear system (GC system)*

$$\mathscr{B}_{\xi_\otimes} s = \mathscr{R}_{\xi_\otimes} f, \tag{4.59}$$

subject to the condition that the boundary degrees of freedom of s coincide with c_b obtained in the previous step, where $\mathscr{B}_{\xi_\otimes}$ is as in (4.56) and $s \in S_d^{d-1}(\triangle) \bigotimes S_d^{d-1}(\triangle)$.

We remark that in the case when the partitions or collocation points in the x- and y- directions are different, matrices $B_2^x, B_2^y, B_1^x, B_1^y$ can be modified accordingly. In fact, matrices B_1, B_2 depend on the specific partition and collocation points being used.

4.5 Numerical Examples Using the Generalized Collocation

4.5.1 1D Examples

We present some examples of the numerical solutions of the two-point BVP in Problem 2. The spline spaces used below are $S_d^{d-1}(\Delta)$ where Δ is a uniform partition of an interval I. The collocation points ξ are interior Greville points as defined in Sect. 4.3. In the tables `k` denotes the number of subintervals in the partition, `maxerr` and `rmserr` denote the maximum error and root-mean-square error respectively, measured on 501 uniformly distributed points in I. `maxrate` and `rmsrate` denote convergence orders of the respective errors in terms of the mesh size h.

Example 4.5.1. *Consider the boundary value problem when $\alpha(x) = \beta(x) \equiv 1$ on $I = [0,1]$ with $f = e^x$. Note that the true solution of this problem is $u = e^x$.*

Discussion: We give the tables of errors for the collocation solutions when $d = 3,4,5$, see Tables 4.5.1 – 4.5.6. All tables of GC solutions (marked as "Gencol") show a convergence order of $d+1$, which is optimal for a spline space of degree d. Note that the orders are unclear once the errors are below 1(-12). These are polluted by the round-off errors due to the machine accuracy and condition numbers of the linear systems. For comparison purposes, the tables of errors for the OC solutions (marked as "Ordcol") are also provided;

as observed here, they agree with the odd/even discrepancy results (i.e., has convergence order $(d-1)$ for odd degree splines and d for even degree splines) noted in the literature, see e.g. [29].

k	maxerr	rmserr	maxrate	rmsrate
12	4.94e-06	2.24e-06		
24	3.25e-07	1.41e-07	3.93	3.98
48	2.08e-08	8.91e-09	3.96	3.99
96	1.32e-09	5.60e-10	3.98	3.99
192	8.28e-11	3.52e-11	3.99	3.99
384	3.64e-12	1.82e-12	4.51	4.28
768	4.47e-12	2.73e-12	-0.30	-0.59

Table 4.5.1: $u = e^x$ with Gencol, $d = 3$

k	maxerr	rmserr	maxrate	rmsrate
12	4.94e-06	3.17e-06		
24	3.07e-09	6.25e-10	10.65	12.31
48	9.87e-11	1.42e-11	4.96	5.46
96	3.09e-12	3.34e-13	5.00	5.41
192	1.43e-12	1.00e-12	1.11	-1.58
384	3.44e-12	2.54e-12	-1.27	-1.34
768	2.23e-11	1.62e-11	-2.69	-2.67

Table 4.5.2: $u = e^x$ with Gencol, $d = 4$

k	maxerr	rmserr	maxrate	rmsrate
12	1.77e-08	6.00e-09		
24	3.04e-10	9.70e-11	5.86	5.95
48	4.97e-12	1.81e-12	5.93	5.74
96	3.83e-13	1.37e-13	3.70	3.72
192	3.36e-13	1.84e-13	0.19	-0.43
384	3.44e-12	2.31e-12	-3.35	-3.65
768	2.69e-12	1.63e-12	0.36	0.51

Table 4.5.3: $u = e^x$ with Gencol, $d = 5$

We also give the matrix pair $\{B_1, B_2\}$ for the case $d = 3,4$. We do not display the matrices for $d = 5$; they possess similar structures in the sense that the non-zero entries of the middle rows (i.e., except the first and last few rows) are identical and the matrices are strictly diagonally dominant for these middle rows.

k	maxerr	rmserr	maxrate	rmsrate
12	1.07e-04	7.73e-05		
24	2.72e-05	1.96e-05	1.98	1.98
48	6.82e-06	4.93e-06	2.00	1.99
96	1.71e-06	1.23e-06	2.00	2.00
192	4.26e-07	3.08e-07	2.00	2.00
384	1.07e-07	7.71e-08	2.00	2.00
768	2.67e-08	1.93e-08	2.00	2.00

Table 4.5.4: $u = e^x$ with Ordcol, $d = 3$

k	maxerr	rmserr	maxrate	rmsrate
12	3.99e-08	2.78e-08		
24	2.61e-09	1.85e-09	3.93	3.91
48	1.68e-10	1.20e-10	3.96	3.95
96	1.02e-11	7.32e-12	4.04	4.03
192	7.99e-13	5.58e-13	3.67	3.71
384	3.45e-12	2.55e-12	-2.11	-2.19
768	1.92e-11	1.35e-11	-2.47	-2.40

Table 4.5.5: $u = e^x$ with Ordcol, $d = 4$

k	maxerr	rmserr	maxrate	rmsrate
12	1.17e-08	8.36e-09		
24	7.76e-10	5.59e-10	3.92	3.90
48	4.91e-11	3.55e-11	3.98	3.98
96	3.09e-12	2.23e-12	3.99	3.99
192	5.92e-13	4.24e-13	2.38	2.40
384	1.37e-12	9.12e-13	-1.21	-1.11
768	9.37e-12	6.78e-12	-2.77	-2.89

Table 4.5.6: $u = e^x$ with Ordcol, $d = 5$

For $d = 3$,

$$B_2 = \begin{bmatrix} \frac{172}{9} & -\frac{505}{18} & \frac{211}{27} & \frac{115}{54} & & & \\ -\frac{9}{2} & 8 & -2 & -\frac{1}{2} & & & \\ & \frac{1}{16} & \frac{85}{96} & \frac{1}{24} & \frac{1}{96} & & \\ & & \frac{1}{12} & \frac{5}{6} & \frac{1}{12} & & \\ & & & \frac{1}{12} & \frac{5}{6} & \frac{1}{12} & \\ & & & & \ddots & \ddots & \ddots \end{bmatrix}, \; B_1 = \begin{bmatrix} \frac{100}{93} & -\frac{140}{837} & \frac{112}{837} & -\frac{35}{837} & & & \\ -\frac{9}{31} & \frac{51}{31} & -\frac{16}{31} & \frac{5}{31} & & & \\ & & 1 & & & & \\ & & & 1 & & & \\ & & & & 1 & & \\ & & & & & \ddots & \end{bmatrix}.$$

For $d = 4$,

$$B_2 = \begin{bmatrix} \frac{4867}{95} & -\frac{17403}{163} & \frac{15999}{218} & -\frac{2345}{147} & -\frac{18}{121} & -\frac{838}{5935} & & & & & \\ -\frac{16873}{771} & \frac{5752}{121} & -\frac{22450}{701} & \frac{1285}{177} & \frac{133}{2549} & \frac{229}{3755} & & & & & \\ \frac{1605}{358} & -\frac{1732}{181} & \frac{2549}{333} & -\frac{1589}{1002} & \frac{140}{3691} & -\frac{244}{11659} & & & & & \\ & \frac{100}{5109} & -\frac{399}{6691} & \frac{1091}{1007} & -\frac{439}{6574} & \frac{191}{6617} & -\frac{447}{82114} & & & & \\ & \frac{33}{14488} & -\frac{122}{11809} & \frac{47}{2020} & \frac{1418}{1459} & \frac{207}{12244} & -\frac{56}{13935} & & & & \\ & & -\frac{5}{75351} & -\frac{110}{32793} & \frac{153}{10873} & \frac{1827}{1867} & \frac{523}{36359} & -\frac{69}{19112} & & & \\ & & & -\frac{1}{58433} & -\frac{55}{15409} & \frac{186}{12875} & \frac{1574}{1609} & \frac{136}{9361} & -\frac{28}{7701} & & \\ & & & & & -\frac{7}{1920} & \frac{7}{480} & \frac{313}{320} & \frac{7}{480} & -\frac{7}{1920} & \cdots \\ & & & & & & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix},$$

$$B_1 = \begin{bmatrix} \frac{919}{1050} & \frac{1375}{4172} & -\frac{235}{681} & \frac{583}{2592} & -\frac{242}{3699} & \frac{33}{4027} & & & & & \\ -\frac{152}{623} & \frac{2619}{1601} & -\frac{2104}{3243} & \frac{758}{2149} & -\frac{2397}{22480} & \frac{81}{7489} & & & & & \\ \frac{979}{4023} & -\frac{2923}{4499} & \frac{188}{111} & -\frac{443}{1076} & \frac{499}{3454} & -\frac{141}{7013} & & & & & \\ & \frac{19}{1783} & -\frac{158}{4527} & \frac{2461}{2338} & -\frac{207}{4778} & \frac{128}{7187} & -\frac{82}{28773} & & & & \\ & \frac{11}{35473} & \frac{15}{83054} & -\frac{68}{28097} & \frac{170}{169} & -\frac{51}{11953} & \frac{59}{52995} & & & & \\ & & \frac{1}{120766} & \frac{81}{68858} & -\frac{73}{15246} & \frac{976}{969} & -\frac{17}{3520} & \frac{11}{9095} & & & \\ & & & & \frac{59}{48633} & -\frac{95}{19559} & \frac{691}{686} & -\frac{62}{12759} & \frac{23}{18931} & & \\ & & & & & \frac{7}{5760} & -\frac{7}{1440} & \frac{967}{960} & -\frac{7}{1440} & \frac{7}{5760} & \cdots \\ & & & & & & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix},$$

where the non-zero entries of the middle rows have the patterns $\begin{bmatrix} -\frac{7}{1920} & \frac{7}{480} & \frac{313}{320} & \frac{7}{480} & -\frac{7}{1920} \end{bmatrix}$ (with $\frac{313}{320}$ on the diagonal) and $\begin{bmatrix} \frac{7}{5760} & -\frac{7}{1440} & \frac{967}{960} & -\frac{7}{1440} & \frac{7}{5760} \end{bmatrix}$ (with $\frac{967}{960}$ on the diagonal) respectively.

Example 4.5.2. *Consider the boundary value problem when* $\alpha(x) = 100\sin(x), \beta(x) = e^x$ *on* $I = [0,1]$ *with* $f(x) = 50\sin(2x) + \sin(x) + e^x\sin(x)$. *Note that the true solution of this problem is* $u = \sin(x)$.

Discussion: We give the tables of errors for the collocation solutions when $d = 3,4,5$, see Tables 4.5.7 –

4.5.12. All tables of GC solutions show a convergence order of $d+1$, while the tables for the OC solutions indicate the usual odd/even discrepancy results. Note that the orders are unclear after the errors are below 10(-13), due to the round-off errors and the condition numbers of the linear systems.

k	maxerr	rmserr	maxrate	rmsrate
12	2.30e-06	5.42e-07		
24	1.50e-07	2.56e-08	3.94	4.40
48	8.83e-09	1.09e-09	4.09	4.55
96	4.97e-10	4.61e-11	4.15	4.57
192	2.88e-11	2.20e-12	4.11	4.39
384	9.86e-13	1.26e-13	4.87	4.12
768	1.37e-13	8.50e-14	2.85	0.57

Table 4.5.7: $u = \sin(x)$ with Gencol, $d = 3$

k	maxerr	rmserr	maxrate	rmsrate
12	1.31e-06	7.57e-07		
24	1.21e-09	2.62e-10	10.08	11.50
48	3.76e-11	5.77e-12	5.00	5.50
96	1.18e-12	1.26e-13	5.00	5.51
192	3.70e-14	7.73e-15	4.99	4.03
384	2.49e-14	1.30e-14	0.57	-0.75
768	2.46e-13	1.19e-13	-3.31	-3.20

Table 4.5.8: $u = \sin(x)$ with Gencol, $d = 4$

k	maxerr	rmserr	maxrate	rmsrate
12	4.35e-09	1.36e-09		
24	7.37e-11	1.67e-11	5.88	6.35
48	1.24e-12	1.97e-13	5.90	6.40
96	3.11e-14	5.00e-15	5.31	5.30
192	3.15e-14	1.49e-14	-0.02	-1.58
384	1.22e-13	7.02e-14	-1.96	-2.23
768	3.65e-13	2.14e-13	-1.58	-1.61

Table 4.5.9: $u = \sin(x)$ with Gencol, $d = 5$

Example 4.5.3. *Consider the boundary value problem when $\alpha(x) = \sin(x), \beta(x) = x$ on $I = [0,1]$, with $f(x)$ being chosen such that the true solution of this problem is $u = \frac{1}{1+25x^2}$.*

Discussion: We give the tables of errors for the collocation solutions when $d = 3,4,5$, see Tables 4.5.13 – 4.5.18. We still observe a $d+1$ convergence order for $d = 3,5$ in the tables for the GC solutions when the mesh is fine enough ($k \geq 48$). However the convergence rate of the GC solutions for $d = 4$ seems to be 6

k	maxerr	rmserr	maxrate	rmsrate
12	5.65e-06	3.02e-06		
24	1.38e-06	7.97e-07	2.04	1.92
48	3.36e-07	2.00e-07	2.03	1.99
96	8.32e-08	5.00e-08	2.01	2.00
192	2.08e-08	1.25e-08	2.00	2.00
384	5.18e-09	3.12e-09	2.00	2.00
768	1.30e-09	7.80e-10	2.00	2.00

Table 4.5.10: $u = \sin(x)$ with Ordcol, $d = 3$

k	maxerr	rmserr	maxrate	rmsrate
12	5.85e-08	3.58e-08		
24	3.10e-09	2.02e-09	4.24	4.15
48	1.81e-10	1.18e-10	4.10	4.10
96	1.10e-11	7.11e-12	4.04	4.05
192	6.67e-13	4.26e-13	4.05	4.06
384	5.37e-14	2.38e-14	3.63	4.16
768	1.02e-13	4.83e-14	-0.93	-1.02

Table 4.5.11: $u = \sin(x)$ with Ordcol, $d = 4$

k	maxerr	rmserr	maxrate	rmsrate
12	5.24e-10	3.00e-10		
24	3.72e-11	2.23e-11	3.82	3.75
48	2.39e-12	1.43e-12	3.96	3.96
96	1.53e-13	9.19e-14	3.96	3.96
192	2.28e-14	1.41e-14	2.75	2.70
384	1.65e-14	9.35e-15	0.47	0.59
768	4.76e-14	2.49e-14	-1.53	-1.41

Table 4.5.12: $u = \sin(x)$ with Ordcol, $d = 5$

numerically, although we know the rate is 5 for quartic splines in general. The convergence orders for the OC solution agree with the usual even/odd discrepancy results. The second (or higher) derivatives of u have large variations near $x = 0$, which could cause instability in the convergence orders. Recall that the design of matrices B_1, B_2 are based on Taylor expansion around the collocation points and the constant C in the inequality $||\mathscr{B}_\xi Qu - \mathscr{L}_\xi u||_{l^\infty} \leq Ch^{d+1}$ depends on the derivatives of u.

4.5.2 2D Examples

In this section we show the performance of the GC method coupled with tensor-product splines (cf., Sec. 2.3). We abbreviate this method as TPGencol and the OC method coupled with tensor-product as

k	maxerr	rmserr	maxrate	rmsrate
12	9.70e-03	5.35e-03		
24	7.66e-04	4.13e-04	3.66	3.69
48	9.32e-05	5.12e-05	3.04	3.01
96	6.60e-06	2.92e-06	3.82	4.13
192	4.48e-07	1.69e-07	3.88	4.11
384	2.06e-08	1.03e-08	4.44	4.04
768	1.29e-09	6.40e-10	3.99	4.01

Table 4.5.13: $u = \frac{1}{1+25x^2}$ with Gencol, $d = 3$

k	maxerr	rmserr	maxrate	rmsrate
12	2.47e-03	1.18e-03		
24	3.43e-04	2.00e-04	2.85	2.56
48	1.29e-05	3.45e-06	4.73	5.86
96	2.44e-07	8.18e-08	5.72	5.40
192	3.71e-09	1.10e-09	6.04	6.22
384	5.08e-11	1.57e-11	6.19	6.13
768	8.25e-13	3.58e-13	5.94	5.45

Table 4.5.14: $u = \frac{1}{1+25x^2}$ with Gencol, $d = 4$

k	maxerr	rmserr	maxrate	rmsrate
12	5.32e-03	3.32e-03		
24	5.20e-04	1.16e-04	3.35	4.83
48	1.17e-05	6.76e-06	5.47	4.11
96	2.33e-07	1.13e-07	5.65	5.91
192	4.86e-09	1.87e-09	5.58	5.91
384	4.35e-11	2.46e-11	6.81	6.25
768	7.74e-13	5.14e-13	5.81	5.58

Table 4.5.15: $u = \frac{1}{1+25x^2}$ with Gencol, $d = 5$

k	maxerr	rmserr	maxrate	rmsrate
12	3.60e-02	2.13e-02		
24	8.46e-03	4.91e-03	2.09	2.12
48	1.99e-03	1.14e-03	2.09	2.10
96	4.88e-04	2.80e-04	2.03	2.03
192	1.21e-04	6.95e-05	2.01	2.01
384	3.03e-05	1.73e-05	2.00	2.00
768	7.57e-06	4.33e-06	2.00	2.00

Table 4.5.16: $u = \frac{1}{1+25x^2}$ with Ordcol, $d = 3$

TPOrdcol. We employ Algorithm 5 and present some examples of the numerical solution of second-order linear elliptic PDE's defined on a rectangular domain Ω with Dirichlet boundary conditions. In all of the

k	maxerr	rmserr	maxrate	rmsrate
12	1.21e-03	6.99e-04		
24	1.80e-04	7.87e-05	2.74	3.15
48	1.35e-05	6.36e-06	3.74	3.63
96	8.50e-07	4.04e-07	3.99	3.98
192	5.31e-08	2.53e-08	4.00	4.00
384	3.32e-09	1.58e-09	4.00	4.00
768	2.07e-10	9.86e-11	4.00	4.00

Table 4.5.17: $u = \frac{1}{1+25x^2}$ with Ordcol, $d = 4$

k	maxerr	rmserr	maxrate	rmsrate
12	8.61e-04	4.44e-04		
24	9.22e-05	4.00e-05	3.22	3.47
48	6.39e-06	3.17e-06	3.85	3.66
96	3.48e-07	1.69e-07	4.20	4.23
192	2.06e-08	9.85e-09	4.08	4.10
384	1.27e-09	6.04e-10	4.02	4.03
768	7.88e-11	3.75e-11	4.01	4.01

Table 4.5.18: $u = \frac{1}{1+25x^2}$ with Ordcol, $d = 5$

following examples, $\Omega = [0,1] \times [0,1]$ is the unit square and $n_b = 2n$ in Step 4 of Algorithm 5. In the tables, `kx` and `ky` denote the number of subintervals in the partitions for the intervals on the x- and y- axis, respectively, `maxerr` and `rmserr` denote the maximum error and root-mean-square error, measured on a 51×51 grid in Ω, `n` denotes the dimension of the spline space being used, and `maxrate` and `rmsrate` denote convergence orders of the respective errors in terms of the mesh size.

Example 4.5.4. *Consider the Poisson problem*

$$\Delta u = f \quad \text{on } \Omega,$$

$$u = g \quad \text{on } \partial\Omega,$$

where f and g are chosen such that the true solution is $u = \sin(x^2+y^2) + .1\sin(25(x^2+y^2))$. Given a partition $\triangle$ on $[0,1]$, find a cubic spline in $S_3^2(\triangle) \bigotimes S_3^2(\triangle)$ (cf. Sect. 2.3) that solves the GC system associated with this problem.

Discussion: We start with a partition with 12×14 sub-rectangles on Ω and decrease the mesh size by half in each loop. From the last two columns in Table 4.5.19, we see the convergence orders for both errors

are roughly 4, which is optimal for tensor-product cubic splines. For comparison purposes, in Table 4.5.20 we also give the errors using ordinary collocation. The convergence orders are roughly 2. We observe that the approximating spline in $S_3^2(\triangle) \bigotimes S_3^2(\triangle)$ obtained by TPGencol achieves a maxerr of 3.06(-4) with 3009 degrees of freedom. To achieve similar accuracy, the approximating spline obtained by TPOrdcol has at least 11385 degrees of freedom (see Row 4 of Table 4.5.20). A plot for the resulting spline associated with the last row of Table 4.5.19 is shown in Figure 4.5.1, which looks no different than the plot for the true solution.

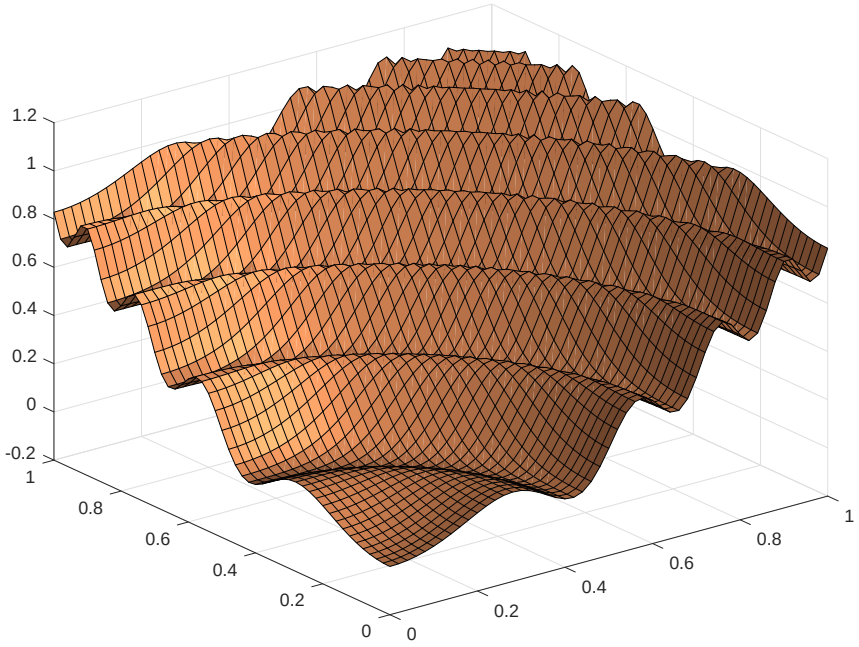


Figure 4.5.1: $u = \sin(x^2+y^2) + .1\sin(25(x^2+y^2))$ in Example 4.5.4

kx	ky	n	maxerr	rmserr	maxrate	rmsrate
12	14	255	3.21e-01	6.07e-02		
24	28	837	4.80e-02	8.09e-03	2.74	2.91
48	56	3009	3.06e-03	4.00e-04	3.97	4.34
96	112	11385	2.05e-04	2.52e-05	3.90	3.99
192	224	44265	1.10e-05	1.33e-06	4.22	4.24

Table 4.5.19: $u = \sin(x^2+y^2) + .1\sin(25(x^2+y^2))$, TPGencol with $S_3^2(\triangle) \bigotimes S_3^2(\triangle)$

Example 4.5.5. *Consider Problem 4 with* $\kappa(x,y) = e^{x+y}$ *and choose* f *and* g *such that the true solution is* $u = -\sin(4x) - \sin(4y)$. *Given a partition* $\triangle$ *on* $[0,1]$, *find a cubic spline in* $S_3^2(\triangle) \bigotimes S_3^2(\triangle)$ *that solves the*

kx	ky	n	maxerr	rmserr	maxrate	rmsrate
12	14	255	2.92e-01	5.13e-02		
24	28	837	5.37e-02	1.11e-02	2.44	2.21
48	56	3009	1.25e-02	2.92e-03	2.10	1.93
96	112	11385	3.04e-03	7.34e-04	2.04	1.99
192	224	44265	7.50e-04	1.83e-04	2.02	2.00

Table 4.5.20: $u = \sin(x^2+y^2) + .1\sin(25(x^2+y^2))$, TPOrdcol with $S_3^2(\triangle) \bigotimes S_3^2(\triangle)$

GC system associated with this problem.

Discussion: We start with a partition of 12×14 sub-rectangles and decrease the mesh size by half in each loop. From the last two columns in Tables 4.5.21 and 4.5.22, we see that the convergence orders of the spline solutions by TPGencol are (roughly) 4, compared with (roughly) 2 by TPOrdcol. Hence to achieve similar accuracy, the TPGencol scheme in general uses many fewer degrees of freedom than the TPOrdcol scheme. For example, TPGencol produces a spline with 837 degrees of freedom which has a maxerr of 2.44(-5), while TPordcol gives a spline of similar accuracy (maxerr = 4.99e(-5)) which has 44265 degrees of freedom. A plot for the resulting spline associated with the last row of Table 4.5.21 is shown in Figure 4.5.2, which looks no different than the plot for the true solution.

kx	ky	n	maxerr	rmserr	maxrate	rmsrate
12	14	255	3.11e-04	1.22e-04		
24	28	837	2.44e-05	7.54e-06	3.67	4.02
48	56	3009	1.01e-06	4.50e-07	4.59	4.07
96	112	11385	6.46e-08	2.95e-08	3.97	3.93
192	224	44265	4.15e-09	1.84e-09	3.96	4.00

Table 4.5.21: $u = -\sin(4x) - \sin(4y)$, TPGencol with $S_3^2(\triangle) \bigotimes S_3^2(\triangle)$

kx	ky	n	maxerr	rmserr	maxrate	rmsrate
12	14	255	1.28e-02	6.25e-03		
24	28	837	3.20e-03	1.55e-03	2.00	2.01
48	56	3009	7.99e-04	3.87e-04	2.00	2.00
96	112	11385	2.00e-04	9.67e-05	2.00	2.00
192	224	44265	4.99e-05	2.42e-05	2.00	2.00

Table 4.5.22: $u = -\sin(4x) - \sin(4y)$, TPOrdcol with $S_3^2(\triangle) \bigotimes S_3^2(\triangle)$

Example 4.5.6. *Consider Problem 4 with* $\kappa(x,y) = \sin(x+y^2)$ *and choose* f *and* g *such that the true solution is* $u = e^{x+2y^2} + 20\sin(5x^2+y)$. *Given a partition* $\triangle$ *on* $[0,1]$, *find a cubic spline in* $S_3^2(\triangle) \bigotimes S_3^2(\triangle)$ *that solves the GC system associated with this problem.*

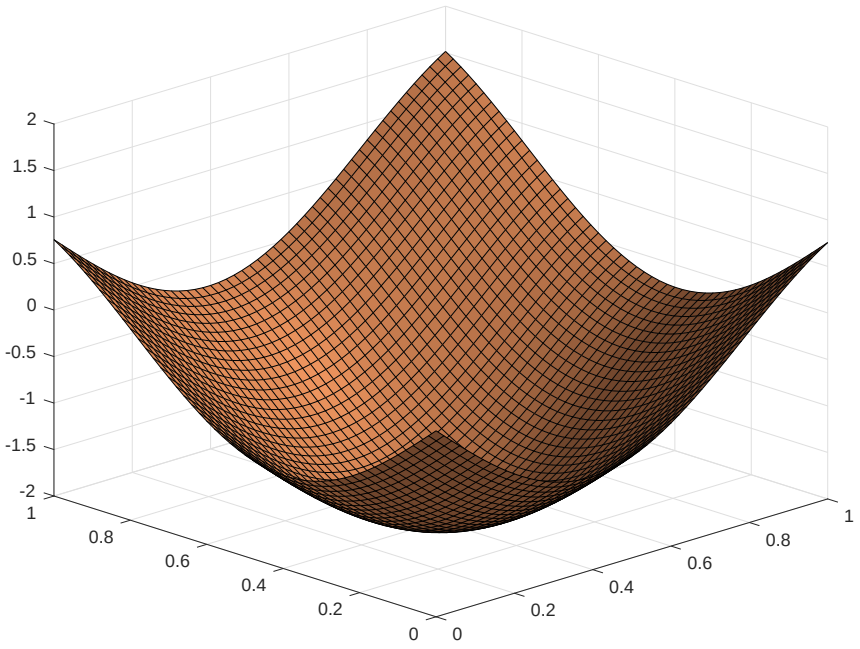


Figure 4.5.2: $u = -\sin(4x) - \sin(4y)$ in Example 4.5.5

Discussion: We use the same partitions and spline spaces as in the previous two examples. From the last two columns in Tables 4.5.23 and 4.5.24, we see that the convergence orders of the spline solutions by TPGencol and TPOrdcol are again (roughly) 4 and 2 respectively. Comparing row by row in these two tables, we observe that TPGencol gives much better accuracy than TPOrdcol. To give an idea of how many degrees of freedom a TPGencol scheme can save to achieve a certain accuracy, we compare a TPGencol spline with 3009 degrees of freedom which has a maxerr of 1.16(-3) with a TPOrdcol spline with 44265 degrees of freedom which has a maxerr of 2.89(-3). A plot for the resulting spline associated with the last row of Table 4.5.23 is shown in Figure 4.5.3, which looks no different than the plot for the true solution.

kx	ky	n	maxerr	rmserr	maxrate	rmsrate
12	14	255	2.48e-01	7.62e-02		
24	28	837	2.31e-02	6.75e-03	3.42	3.50
48	56	3009	1.16e-03	3.88e-04	4.32	4.12
96	112	11385	7.30e-05	2.51e-05	3.99	3.95
192	224	44265	4.60e-06	1.55e-06	3.99	4.02

Table 4.5.23: $u = e^{x+2y^2} + 20\sin(5x^2 + y)$, TPGencol with $S_3^2(\triangle) \bigotimes S_3^2(\triangle)$

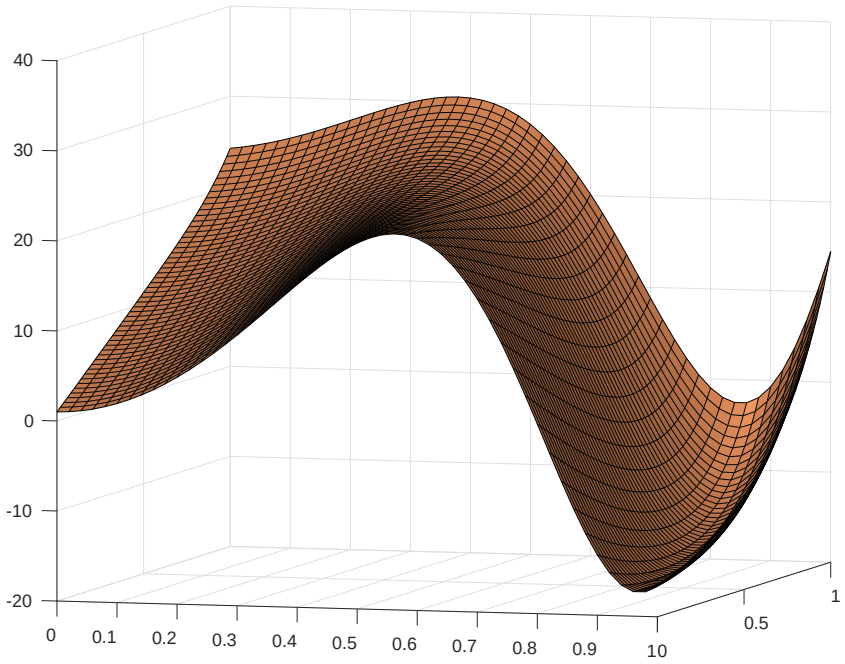


Figure 4.5.3: $u = e^{x+2y^2} + 20\sin(5x^2+y)$ in Example 4.5.6

kx	ky	n	maxerr	rmserr	maxrate	rmsrate
12	14	255	8.32e-01	3.46e-01		
24	28	837	1.92e-01	7.97e-02	2.12	2.12
48	56	3009	4.67e-02	1.95e-02	2.04	2.03
96	112	11385	1.16e-02	4.85e-03	2.01	2.01
192	224	44265	2.89e-03	1.21e-03	2.00	2.00

Table 4.5.24: $u = e^{x+2y^2} + 20\sin(5x^2+y)$, TPOrdcol with $S_3^2(\triangle) \bigotimes S_3^2(\triangle)$

4.6 Least-squares Collocation for Poisson's Problem Using Splines on Triangulations

Let Ω be a bounded domain in $\mathbb{R}^2$ and suppose f and g are functions defined on Ω and $\partial\Omega$, respectively. Let us consider the Poisson problem:

Problem 5. *Find a function u defined on Ω such that*

$$\Delta u = f \qquad \text{on } \Omega, \tag{4.60}$$

$$u = g \qquad \text{on } \partial\Omega. \tag{4.61}$$

Assume that Ω is a polygonal domain and let $\triangle$ be a triangulation of Ω with mesh size $|\triangle|$, i.e., the

maximum of the diameters of the triangles in $\triangle$. Let $X \subset \Omega$ and $Y \subset \partial\Omega$ be two finite sets of points and $S(\triangle)$ be a spline space defined on $\triangle$. The collocation system associated with Problem 5 is to find $s \in S(\triangle)$ such that

$$\Delta s|_X = f|_X, \tag{4.62}$$

$$s|_Y = g|_Y, \tag{4.63}$$

where X, Y are called the collocation points and boundary collocation points respectively. If the above system is over-determined, i.e., the number of equations is larger than the dimension of the spline space being used, it is solved in the least-squares sense. We first give an error bound for the collocation solution of (4.62), (4.63), assuming the existence and uniqueness of such a solution. Examples of conditions on the choice of collocation points X and Y which guarantee the existence and uniqueness of a solution will be given afterwards. For convenience, given positive integers p and m, let $||\phi||_{l^p(X)} := (\sum_{x\in X} |\phi(x)|^p)^{\frac{1}{p}}$ for any function ϕ defined on Ω and finite set X, and let $||v||_{l^p(\mathbb{R}^m)} := (\sum_{i=1}^{m} |v_i|^p)^{\frac{1}{p}}$ for any vector $v \in \mathbb{R}^m$. When $p = \infty$, the pertinent norms will be understood as the maximum norms. For a finite set X, we denote the number of elements in the set as $\#X$.

Theorem 4.6.1. *Let $S(\triangle) \subset C^2(\Omega)$ be a spline space and let $X \subset \Omega$ and $Y \subset \partial\Omega$ be finite sets such that the system (4.62) (4.63) has a unique solution in the least-squares sense. Let*

$$\hat{s} = \underset{s\in S(\triangle)}{\operatorname{argmin}} ||\Delta s - f||^2_{l^2(X)} + \lambda^2 ||s - g||^2_{l^2(Y)}, \tag{4.64}$$

where $1 \le \lambda \le K\sqrt{\frac{\#X}{\#Y}}$ for some fixed constant K. Assume further that there exist constants C_1, C_2, independent of the mesh size $|\triangle|$, such that $\sqrt{\#X} \le \frac{C_1}{|\triangle|}$ and

$$||s||_{L^\infty(\Omega)} \le C_2\big(||\Delta s||_{l^\infty(X)} + ||s||_{l^\infty(Y)}\big), \tag{4.65}$$

for all $s \in S(\triangle)$. If the solution u of Problem 5 is in $C^2(\Omega) \cap C^0(\overline{\Omega})$, then, for $|\triangle|$ sufficiently small, we have the following error bound

$$||u - \hat{s}||_{L^\infty(\Omega)} \le \frac{C}{|\triangle|} ||u - \tilde{s}||_{W^{2,\infty}(\Omega)}, \tag{4.66}$$

where $\tilde{s}$ is the best approximation to u in $S(\triangle)$ with respect to the $W^{2,\infty}$ norm, and C is a constant independent of the mesh size $|\triangle|$.

Proof. The assumption that the system (4.62) (4.63) has a unique solution guarantees that the observation matrix associated with the system has full rank; hence the minimization problem (4.64) has a unique solution. By (4.65) and since $\lambda \geq 1$,

$$||\tilde{s}-\hat{s}||_{L^\infty(\Omega)} \leq C_2(||\Delta(\hat{s}-\tilde{s})||_{l^\infty(X)} + ||\hat{s}-\tilde{s}||_{l^\infty(Y)}) \tag{4.67}$$

$$\leq C_2(||\Delta(\hat{s}-\tilde{s})||_{l^2(X)} + \lambda||\hat{s}-\tilde{s}||_{l^2(Y)}) \tag{4.68}$$

$$\leq C_2\sqrt{2}(||\Delta(\hat{s}-\tilde{s})||^2_{l^2(X)} + \lambda^2||\hat{s}-\tilde{s}||^2_{l^2(Y)})^{1/2}, \tag{4.69}$$

where the last inequality follows from the fact that $a+b \leq \sqrt{2(a^2+b^2)}$ for all non-negative real numbers a,b. Let $m = \#X + \#Y$ and let $S_F := \{(\Delta s|_X, \lambda s|_Y) \in \mathbb{R}^m |\ s \in S(\triangle)\}$. It is easy to see that $(\Delta\hat{s}|_X, \lambda\hat{s}|_Y)$ is the l^2-projection of $(f|_X, \lambda g|_Y)$ onto S_F. Since both $(\Delta\hat{s}|_X, \lambda\hat{s}|_Y)$ and $(\Delta\tilde{s}|_X, \lambda\tilde{s}|_Y)$ are in the linear subspace S_F, by the Pythagorean theorem, for any $s \in S(\triangle)$,

$$||(\Delta\hat{s}|_X - \Delta s|_X, \lambda\hat{s}|_Y - \lambda s|_Y)||_{l^2(\mathbb{R}^m)} \leq ||(f|_X - \Delta s|_X, \lambda g|_Y - \lambda s|_Y)||_{l^2(\mathbb{R}^m)}. \tag{4.70}$$

In particular, if we choose s to be $\tilde{s}$ and let $\tilde{f} = \Delta\tilde{s}|_\Omega$ and $\tilde{g} = \tilde{s}|_{\partial\Omega}$, we obtain

$$||(\Delta\hat{s}|_X - \Delta\tilde{s}|_X, \lambda\hat{s}|_Y - \lambda\tilde{s}|_Y)||_{l^2(\mathbb{R}^m)} \leq ||(f|_X - \tilde{f}|_X, \lambda g|_Y - \lambda\tilde{g}|_Y)||_{l^2(\mathbb{R}^m)}. \tag{4.71}$$

Combining (4.69) with (4.71) and since $\lambda\sqrt{\#Y} \leq K\sqrt{\#X}$, we get

$$||\tilde{s}-\hat{s}||_{L^\infty(\Omega)} \leq C_2\sqrt{2}(||f-\tilde{f})||^2_{l^2(X)} + \lambda^2||g-\tilde{g}||^2_{l^2(Y)})^{1/2} \tag{4.72}$$

$$\leq C_2\sqrt{2}(\sqrt{\#X}||f-\tilde{f}||_{l^\infty(X)} + \lambda\sqrt{\#Y}||g-\tilde{g}||_{l^\infty(Y)}) \tag{4.73}$$

$$\leq (K+1)C_2\sqrt{2\#X}(||f-\tilde{f}||_{l^\infty(X)} + ||g-\tilde{g}||_{l^\infty(Y)}) \tag{4.74}$$

$$\leq \frac{\sqrt{2}(K+1)C_2C_1}{|\triangle|}(||f-\tilde{f}||_{l^\infty(X)} + ||g-\tilde{g}||_{l^\infty(Y)}) \tag{4.75}$$

$$\leq \frac{\sqrt{2}(K+1)C_2C_1}{|\triangle|}(||f-\tilde{f}||_{L^\infty(\Omega)} + ||g-\tilde{g}||_{L^\infty(\partial\Omega)}) \tag{4.76}$$

$$\leq \frac{2\sqrt{2}(K+1)C_2C_1}{|\triangle|}||u-\tilde{s}||_{W^{2,\infty}(\Omega)}, \tag{4.77}$$

where the last inequality make use of the fact that

$$||\Delta v||_{L^\infty(\Omega)} \leq ||v||_{W^{2,\infty}(\Omega)}, \quad \forall v \in W^{2,\infty}(\Omega). \tag{4.78}$$

By the triangle inequality,

$$||u-\hat{s}||_{L^\infty(\Omega)} \leq ||u-\tilde{s}||_{L^\infty(\Omega)} + ||\tilde{s}-\hat{s}||_{L^\infty(\Omega)}. \tag{4.79}$$

Combining this with (4.77), we obtain

$$||u-\hat{s}||_{L^\infty(\Omega)} \leq \frac{2\sqrt{2}(K+1)C_2C_1+|\triangle|}{|\triangle|}||u-\tilde{s}||_{W^{2,\infty}(\Omega)}, \tag{4.80}$$

and the proof is finished by choosing a constant $C > 2\sqrt{2}(K+1)C_2C_1+|\triangle|$. □

Next, we present a lemma showing how certain Sobolev norms of a spline $s \in S(\triangle) \subseteq C^2(\Omega)$ depend on its boundary data and the L^∞-norm of Δs. Throughout this section, we assume that triangulation $\triangle$ is *β-quasi-uniform* in the following sense: there exists $\beta > 0$ such that

$$\frac{|\triangle|}{\rho_\triangle} \leq \beta, \tag{4.81}$$

where $\rho_\triangle$ is the minimum of the radii of the largest circles inscribed (incircles) in triangles of $\triangle$.

Lemma 4.6.1. *Let $S(\triangle) \subseteq C^2(\Omega)$ be a degree d spline space and $s \in S(\triangle)$. For any non-negative integer $\alpha \leq 2$, there exist constants K, C such that*

$$||s||_{W^{\alpha,\infty}(\Omega)} \leq \frac{K\beta^\alpha}{|\triangle|^\alpha}\left(||s||_{L^\infty(\partial\Omega)} + C||\Delta s||_{L^\infty(\Omega)}\right), \tag{4.82}$$

where K depends only on d, C depends only on diam Ω*, and β is the quasi-uniform parameter associated with $\triangle$.*

Proof. Applying Corollary 3.8 in [18] to the case where $Lu = \Delta u$, it is easy to see that i) the minimum eigenvalue of the coefficient matrix $(a^{ij}(x))$ is $\lambda = 1$ since for Laplace's equation $(a^{ij}(x))$ is simply the identity matrix ii) the coefficient functions $b, c = 0$ (for definitions of these variables, cf. Fact 5 in Appendix);

hence since $s \in S(\triangle)$ is in $C^2(\Omega)$, we obtain

$$||s||_{L^\infty(\Omega)} \leq ||s||_{L^\infty(\partial\Omega)} + C||\Delta s||_{L^\infty(\Omega)}, \tag{4.83}$$

where the constant C only depends on diam Ω. Now recall the Markov inequality for polynomials on triangles (cf. p.44 in [24]): Given a non-degenerate triangle T, there exists a constant K depending only on d such that $\forall\ p \in \mathscr{P}_d$, and any non-negative integers α_1 and α_2 with $0 \leq \alpha_1 + \alpha_2 \leq d$,

$$||D_x^{\alpha_1} D_y^{\alpha_2} p||_{L^\infty(T)} \ \leq \frac{K}{\rho_T^{\alpha+\alpha_2}} ||p||_{L^\infty(T)},$$

where ρ_T denotes the radius of the largest circle inscribed (incircle) in T. Hence for a spline $s \in S(\triangle)$ and $\alpha_1 + \alpha_2 \leq 2$,

$$\begin{aligned} ||D_x^{\alpha_1} D_y^{\alpha_2} s||_{L^\infty(\Omega)} = \sup_{T\in\triangle} ||D_x^{\alpha_1} D_y^{\alpha_2} s||_{L^\infty(T)} &\leq \sup_{T\in\triangle} \frac{K}{\rho_T^{\alpha_1+\alpha_2}} ||s||_{L^\infty(T)} \\ &\leq \frac{K}{\rho_\triangle^{\alpha_1+\alpha_2}} ||s||_{L^\infty(\Omega)}. \end{aligned}$$

Combining the above inequality with (4.83), we obtain

$$||s||_{W^{\alpha,\infty}(\Omega)} \ \leq \ \frac{K}{(\rho_\triangle)^\alpha} \ ||s||_{L^\infty(\Omega)} \tag{4.84}$$

$$\leq \frac{K\beta^\alpha}{|\triangle|^\alpha} \ \Big(||s||_{L^\infty(\partial\Omega)} + C||\Delta s||_{L^\infty(\Omega)} \Big), \tag{4.85}$$

where K depends only on d. □

We next give an example of choices of X and Y which guarantee a unique solution of the system (4.62) (4.63) in the least-squares sense. Given a degree d spline space $S(\triangle) \subseteq C^2(\Omega)$, a simple choice of X and Y is:

- X: domain points $\mathscr{D}_{m,T}$ of degree $m \geq d-2$ in each triangle $T \in \triangle$ (cf. Sect. 2.2)
- Y: $k \geq d+1$ uniformly spaced points on each boundary edge of $\triangle$

Claim 4.6.1. *Given a degree d spline space $S(\triangle) \subseteq C^2(\Omega)$ and X, Y defined above, there exists a constant C depending only on Ω and the degree d of the spline space $S(\triangle)$, m and k (cf. X, Y defined above) such*

that for all $s \in S(\triangle)$

$$||s||_{L^\infty(\Omega)} \leq C\Big(||\Delta s||_{l^\infty(X)} + ||s||_{l^\infty(Y)}\Big). \tag{4.86}$$

Proof. By letting $\alpha = 0$ and K, C be the constants as in Lemma 4.6.1, we obtain

$$||s||_{L^\infty(\Omega)} \leq K\Big(||s||_{L^\infty(\partial\Omega)} + C||\Delta s||_{L^\infty(\Omega)}\Big). \tag{4.87}$$

Hence it suffices to show that, for the given X and Y, there exists constants $C_1, C_2 > 0$ depending only on m and k such that for any $s \in S(\triangle)$

$$||s||_{L^\infty(\partial\Omega\cap T)} \leq C_1||s||_{l^\infty(Y\cap T)}, \tag{4.88}$$

$$||\Delta s||_{L^\infty(\Omega\cap T)} \leq C_2||\Delta s||_{l^\infty(X\cap T)}. \tag{4.89}$$

Note that $s|_T$ is a polynomial of total degree $\leq d$ and $\Delta s|_T$ is a polynomial of total degree $\leq d-2$. The first inequality holds since the interpolation matrix of one-dimensional Bernstein basis polynomials of degree $\leq (k-1)$ on k uniformly spaced points on a boundary edge is non-singular, and the constant C_1 depends only on the operator norm of the inverse of this matrix, which is determined by k. The second inequality holds for similar reasons, i.e., the interpolation matrix associated with the Bernstein–Bézier basis of degree m and degree m domain points in a triangle is also non-singular and depends only on m. □

The previous claim implies that the minimization problem in (4.64) is uniquely solvable and $\#X + \#Y \geq \dim S(\triangle)$.

We next show an upper bound for the size of X defined above. Let $n_t :=$ # of triangles in $\triangle$ and let $n_m := \frac{(m+2)(m+1)}{2}$ be the number of domain points of degree m. By the definition of X, we have

$$\#X \leq n_t \times n_m \tag{4.90}$$

$$\leq \frac{\text{Area}(\Omega)}{\min_{T\in\triangle} \text{Area}(T)} \times n_m \tag{4.91}$$

$$\leq \frac{\text{Area}(\Omega)}{\pi\rho_\triangle^2} \times n_m \tag{4.92}$$

$$\leq \frac{\text{Area}(\Omega)\beta^2}{\pi|\triangle|^2} \times n_m, \tag{4.93}$$

which shows that there exists a constant C independent of $|\triangle|$ such that $\sqrt{\#X} \leq \frac{C}{|\triangle|}$.

We have verified that given a C^2 spline space defined on a β-uniform triangulation and X, Y as defined above, the assumptions in Theorem 4.6.1 are satisfied. Furthermore, it is easy to see that an error bound similar to (4.66) in $W^{2,\infty}$-norm holds. Indeed, using the triangle inequality as in (4.79) while each term is measured in the $W^{2,\infty}$-norm and applying Lemma 4.6.1, we obtain that for any $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$

$$||u - \hat{s}||_{W^{2,\infty}(\Omega)} \leq \frac{C}{|\triangle|^3} ||u - \tilde{s}||_{W^{2,\infty}(\Omega)}, \tag{4.94}$$

where $\hat{s}$ is the least-squares collocation solution in (4.64) and C is constant depending on Ω, β, m, k and independent of $|\triangle|$. To show a convergence rate of the collocation approximations to u, we make use of the following approximation power property (cf. Theorem 10.10, Remark 10.11 in [24]): Let $d \geq 3r+2$, and suppose $\triangle$ is a regular triangulation (in the sense of Def. 4.7 in [24]) of Ω. Then for any $u \in W^{m,\infty}(\Omega)$, there exists a spline $s \in S_d^r(\triangle)$ such that

$$||D_x^\alpha D_y^\beta (u - s)||_{L^\infty(\Omega)} \leq K|\triangle|^{l-\alpha-\beta} \max_{\alpha+\beta=l} ||D_x^\alpha D_y^\beta u||_{L^\infty(\Omega)}, \tag{4.95}$$

for all $0 \leq \alpha + \beta \leq l-1$, where the constant K depends only on d, r, the smallest angle θ in $\triangle$, and the Lipschitz constant of $\partial\Omega$ if Ω is not convex. Assuming the smallest angle θ_Δ associated with a triangulation $\triangle$ is bounded below, i.e., there exists $\theta_0 > 0$ such that $\theta_\triangle \geq \theta_0$ for all $\triangle$ used in this section, we see that the triangulation is $\triangle$ is $\beta = \frac{2}{\sin(\theta_0/2)}$-quasi-uniform. Hence there exists a constant C depending on d, r, θ_0 and Ω such that for all $u \in W^{l,\infty}(\Omega)$ with $2 < l \leq d+1$,

$$||u - \tilde{s}||_{W^{2,\infty}(\Omega)} \leq C|\triangle|^{l-2} ||u||_{W^{l,\infty}(\Omega)}, \tag{4.96}$$

where $\tilde{s}$ is the best approximation to u in $S_d^r(\triangle)$ in $W^{2,\infty}$-norm as in Theorem 4.6.1. We summarize a convergence rate result of the least-squares collocation solutions in the following theorem:

Theorem 4.6.2. *Let $\triangle$ be a regular triangulation of Ω with the smallest angle of triangles $\geq \theta_0$ and let $S_d^r(\triangle)$ with $r \geq 2$ be a spline space with $d \geq 3r+2$. With the X, Y defined in this section, there exists a unique solution $\hat{s}$ to (4.64) such that for any $u \in C^l(\Omega) \cap C^0(\overline{\Omega})$*

$$||u - \hat{s}||_{L^\infty(\Omega)} \leq C|\triangle|^{l-3} ||u||_{W^{l,\infty}(\Omega)}, \quad 3 < l \leq d+1, \tag{4.97}$$

$$||u-\hat{s}||_{W^{2,\infty}(\Omega)} \leq C|\triangle|^{l-5}||u||_{W^{l,\infty}(\Omega)}, \quad 5 < l \leq d+1, \tag{4.98}$$

where l is an integer and C is a constant depending on d, r, θ_0, Ω and the Lipschitz constant of the boundary of Ω if Ω is not convex.

Proof. The result follows from Theorem 4.6.1, (4.94) and (4.96). □

Let us take the macro-element space $S_9^{2,4}(\triangle)$ (cf. p. 174 in [24]) as a specific example of C^2 splines spaces defined on a triangulation $\triangle$. It has been shown in [24] that this space has full approximation power, i.e., order 10 approximation to smooth enough functions. By Theorem 4.6.2, the least-squares collocation solutions in (4.64) approximate u in an order of at least $|\triangle|^7$ in L^∞-norm and in an order of at least $|\triangle|^5$ in $W^{2,\infty}$-norm. We give a numerical example below.

Example 4.6.1. *Let Ω be the unique square in $\mathbb{R}^2$ and $\{\triangle_i\}_{i=1}^4$ be a finite sequence of type-1 triangulations (cf. Definition 3.10 in [38]) of Ω. Consider Problem 5 where f and g are chosen such that the true solution $u = \sin(x^2+y^2) + .1\sin(25(x^2+y^2))$ (cf. Figure 4.5.1) and solve the problem (4.64) using $S_9^{2,4}(\triangle_i)$, $i = 1,...4$.*

Discussion: Let $\{\Delta_i\}_{i=1}^4$ be a sequence of type-1 triangulations with $|\Delta_i| = 2^{i-1}|\Delta_1|$. In particular, we make use of triangulation corresponding to the files `type1.25, type1.81, type1.289, type1.1089` and Matlab functions such as `mds29` in the Matlab software package associated with Schumaker's book [38]. For each triangulation $\triangle$, we tabulate $h_\triangle := \frac{|\triangle|}{\sqrt{2}}$, the number of degrees of freedom `ndof`, the maximum error `maxerr`, the RMS error `rmserr`, the condition number `cond` of the Gram matrix associated with the collocation systems and the convergence rate in the maximum norm `maxrate`. The errors are computed over a 501×501 grid on the unit square. Note that in this example, since $|\triangle_i| = 2|\triangle_{i+1}|$, the convergence rate between consecutive errors e_i, e_{i+1} is $\log_2(\frac{e_i}{e_{i+1}})$. We choose K in Theorem 4.6.1 to be 500 and choose $\lambda = K\mu$ in (4.64) where $\mu = \lfloor\sqrt{\frac{\#X}{\#Y}}\rfloor$. We have chosen such K so that the condition numbers of the Gram matrices involved are comparably small in our experiments. However from Table 4.6.1, we can see that the condition numbers are still relatively large. For the choice of X, degree d domain points are chosen in each triangle; for Y, $d+1$ uniformly distributed points are chosen on each boundary edge. For each boundary triangle T_b, the domain points chosen for X are associated with a similar triangle inside; the barycentric coordinates of the vertices of this triangle relative to T_b are $(0.8,0.1,0.1),(0.1,0.1,0.8),(0.1,0.8,0.1)$, respectively. We solved the minimization problem (4.64) by applying Matlab's built-in backslash function to

the Gram matrix, which may not be optimal considering the large condition numbers of the Gram matrices involved; we remark that iterative methods such as gradient descent or conjugate gradient descent etc. could be used and we leave the exploration of the performance using these methods for later work. From Table 4.6.1, it seems that the rate of convergence in the maximum norm is about 8, which is higher than the predicted rate 7 by Theorem 4.6.2. We expect to investigate this matter in more details in the near future.

$h_{\triangle}$	ndof	μ	maxerr	rmserr	cond	maxrate
.25	575	3	1.79e+00	2.90e-01	2.22e+08	
.125	1967	4	5.11e-02	2.69e-03	7.36e+08	5.13
.0625	7247	5	1.22e-04	4.96e-06	8.20e+09	8.71
.03125	27791	7	4.66e-07	1.82e-07	1.23e+11	8.03

Table 4.6.1: Least-squares collocation for Example 4.6.1 using $S_9^{2,4}(\triangle)$

Appendix

Fact 1: Let $A = \{a_{ij}\}_{i,j=1}^{n}$ be a strictly diagonally dominant(SDD) matrix, i.e., $|a_{ii}| > \sum_{j\neq i}|a_{ij}|$ for each $i = 1,\cdots,n$. Then A is non-singular. In addition,

$$||A^{-1}||_{\infty} \leq \frac{1}{\min\limits_{k=1,\cdots,m}(|a_{kk}| - \sum\limits_{j\neq k}|a_{kj}|)}, \tag{A.1}$$

see e.g., [42, 43].

Fact 2: For any $u \in C^2[a,b], ||u||_{L^{\infty}(I)} \leq |u(b)-u(a)| + \min\{|u(a)|,|u(b)|\} + \frac{||u''||_{L^{\infty}(I)}}{2}(b-a)^2$.

Proof. Let $x \in (a,b)$, by Taylor's Theorem, there exists real numbers $t_1 \in (a,x)$ and $t_2 \in (x,b)$ such that

$$u(a) = u(x) + u'(x)(a-x) + \frac{u''(t_0)}{2}(a-x)^2,$$
$$u(b) = u(x) + u'(x)(b-x) + \frac{u''(t_1)}{2}(b-x)^2.$$

Subtracting the first equality from the second we get

$$(b-a)u'(x) = u(b) - u(a) - \frac{u''(t_1)}{2}(b-x)^2 + \frac{u''(t_0)}{2}(a-x)^2,$$

and hence

$$||u'||_{L^{\infty}(I)} \leq \frac{1}{b-a}\big(|u(b)-u(a)| + \frac{b-a}{2}||u''||_{L^{\infty}(I)}\big).$$

Here we used a simple inequality that for $x \in [a,b]$, $(b-x)^2 + (a-x)^2 \leq (b-a)^2$. Since for any $y \in [a,b]$,

$$u(y) = \int_a^y u'(t)dt + u(a),$$

we have $|u(y| \leq (b-a)||u'||_{L^{\infty}(I)} + |u(a)|$. Combining the above estimate for $||u'||_{L^{\infty}(I)}$ we obtain $||u||_{L^{\infty}(I)} \leq |u(b)-u(a)| + |u(b)| + \frac{||u''||_{L^{\infty}(I)}}{2}(b-a)^2$. By symmetry, the desired inequality is observed.

□

Fact 3: Consider the two-point boundary problem where the coefficient functions $\alpha(x), \beta(x)$ and $f(x)$ are

continuous on $[a,b]$:

$$Lu := -u''(x) + \alpha(x)u'(x) + \beta(x)u(x) = f(x), \qquad \text{on } (a,b) \tag{A.2}$$

$$u(a) = u_a, \quad u(b) = u_b. \tag{A.3}$$

Assume that given any real numbers u_a, u_b, the above problem has a unique solution u in $C^2[a,b]$. Then there exists a constant C such that

$$||u||_{L^\infty[a,b]} \leq C||f||_{L^\infty[a,b]}, \tag{A.4}$$

for any $u \in C^2[a,b]$ such that $u(a) = u(b) = 0$.

Proof. From the theory of linear ordinary differential equations [30], the assumption on the existence of a unique solution of the problem implies the existence of the Green's function $G(x,t)$ associated with the homogeneous boundary condition, which is continuous on the square $[a,b] \times [a,b]$. Moreover, for any u in $C^2[a,b]$ such that $u(a) = u(b) = 0$,

$$u(x) = \int_a^b G(x,t)f(t)dt. \tag{A.5}$$

By the continuity of $G(x,t)$ on the square, we can take $C := \max_{(x,t)\in[a,b]\times[a,b]} |G(x,t)|$. □

Fact 4: We summarize Markov brothers' inequality [28]: Let p be a polynomial of degree less or equal to n and k be a positive integer. Then

$$\max_{-1\leq x\leq 1} |D^k p(x)| \leq \frac{n^2(n^2-1^2)(n^2-2^2)\cdots(n^2-(k-1)^2)}{1\cdot 3\cdot 5\cdots(2k-1)} \max_{-1\leq x\leq 1} |p(x)|. \tag{A.6}$$

Fact 5: We recall Corollary 3.8 in [18] : Let $Lu := \sum a^{ij}(x)D_{ij}u + \sum b^i(x)D_i u + c(x)u$ in a bounded domain Ω, where L is elliptic and $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$. Let C be the constant depending only on the diam Ω and $\beta := sup\frac{|b|}{\lambda}$ and suppose that $C_1 = 1 - C\sup_{\Omega} \frac{c^+}{\lambda} > 0$, where λ is the minimum of eigenvalues of the coefficient matrix $(a^{ij}(x))$ and $c^+ = \max\{c, 0\}$. In particular, if Ω lies between two parallel planes

with a distance d apart, let $C = e^{(\beta+1)d} - 1$. Then

$$\sup_{\Omega} |u| \leq \frac{1}{C_1} \Big(\sup_{\partial\Omega} |u| + C \sup_{\Omega} \frac{|f|}{\lambda} \Big). \tag{A.7}$$

BIBLIOGRAPHY

[1] AINSWORTH, M., DEMKOWICZ, L., AND KIM, C.-W. Analysis of the equilibrated residual method for a posteriori error estimation on meshes with hanging nodes. *Computer methods in applied mechanics and engineering 196*, 37-40 (2007), 3493–3507.

[2] AINSWORTH, M., AND ODEN, J. T. *A Posteriori Error Estimation in Finite Element Analysis*, vol. 37. John Wiley & Sons, 2000.

[3] AINSWORTH, M., AND RANKIN, R. Constant free error bounds for nonuniform order discontinuous galerkin finite-element approximation on locally refined meshes with hanging nodes. *IMA journal of numerical analysis 31*, 1 (2009), 254–280.

[4] ARCHER, D. *Some collocation methods for differential equations*. PhD thesis, Rice University, 1973.

[5] ARCHER, D., AND DIAZ, J. C. A family of modified collocation methods for second order two point boundary value problems. *SIAM Journal on Numerical Analysis 15*, 2 (1978), 242–254.

[6] BANK, R. E., AND SMITH, R. K. A posteriori error estimates based on hierarchical bases. *SIAM Journal on Numerical Analysis 30*, 4 (1993), 921–935.

[7] BIALECKI, B., AND FAIRWEATHER, G. Orthogonal spline collocation methods for partial differential equations. *Journal of Computational and Applied Mathematics 128*, 1-2 (2001), 55–82.

[8] BICKLEY, W. Piecewise cubic interpolation and two-point boundary problems. *The Computer Journal 11*, 2 (1968), 206–208.

[9] CARSTENSEN, C., HU, J., AND ORLANDO, A. Framework for the a posteriori error analysis of nonconforming finite elements. *SIAM Journal on Numerical Analysis 45*, 1 (2007), 68–82.

[10] DE BOOR, C. *A Practical Guide to Splines*, vol. 27. Springer-Verlag New York, 1978.

[11] DE BOOR, C., AND SWARTZ, B. Collocation at gaussian points. *SIAM Journal on Numerical Analysis 10*, 4 (1973), 582–606.

[12] DEMARET, L., DYN, N., FLOATER, M. S., AND ISKE, A. Adaptive thinning for terrain modelling and image compression. In *Advances in multiresolution for geometric modelling*. Springer, 2005, pp. 319–338.

[13] DEMARET, L., DYN, N., AND ISKE, A. Image compression by linear splines over adaptive triangulations. *Signal Processing 86*, 7 (2006), 1604–1616.

[14] FAIRWEATHER, G., AND MEADE, D. A survey of spline collocation methods for the numerical solution of differential equations. *Mathematics for large scale computing, Lecture notes in pure and applied mathematics 120* (1989), 297–341.

[15] FISHER, N. *Orthogonal Spline Collocation Methods for Fluid Flow Problems*. PhD thesis, Colorado School of Mines. Arthur Lakes Library, 2019.

[16] FRAZER, R., JONES, W., AND SKAN, S. Approximations to functions and to the solutions of differential equations. Great Britain Aero. Res. Conf., London, reprinted in Great Britain Air Ministry Aero. *Res. Comm. Tech. Rep 1* (1937), 516–549.

[17] FYFE, D. The use of cubic splines in the solution of two-point boundary value problems. *The Computer Journal 12*, 2 (1969), 188–192.

[18] GILBARG, D., AND TRUDINGER, N. S. *Elliptic partial differential equations of second order*. springer, 2015.

[19] GOMEZ, H., AND DE LORENZIS, L. The variational collocation method. *Computer Methods in Applied Mechanics and Engineering 309* (2016), 152–181.

[20] HOUSTIS, E. N., VAVALIS, E., AND RICE, J. R. Convergence of o(h^4) cubic spline collocation methods for elliptic partial differential equations. *SIAM Journal on Numerical Analysis 25*, 1 (1988), 54–74.

[21] IRODOTOU-ELLINA, M. *Spline collocation methods for high order elliptic boundary value problems*. PhD thesis, Department of Mathematics, Aristotle University of Thessaloniki, 1987.

[22] IRODOTOU-ELLINA, M., AND HOUSTIS, E. N. An O(h^6) quintic spline collocation method for fourth order two-point boundary value problems. *BIT Numerical Mathematics 28*, 2 (1988), 288–301.

[23] ISKE, A. *Multiresolution Methods in Scattered Data Modelling*, vol. 37. Springer Science & Business Media, 2004.

[24] LAI, M., AND SCHUMAKER, L. *Spline Functions on Triangulations*. Cambridge, 2007.

[25] LAI, M.-J., AND MERSMANN, C. An adaptive triangulation method for bivariate spline solutions of PDEs. In *Approximation Theory XV: San Antonio 2016*, G. E. Fasshauer and L. L. Schumaker, Eds., Springer Proceedings in Mathematics & Statistics 201, pp. 155–175.

[26] LI, S., AND SCHUMAKER, L. L. Adaptive computation with splines on triangulations with hanging vertices. In *Approximation Theory XV: San Antonio 2016*, G. E. Fasshauer and L. L. Schumaker, Eds., Springer Proceedings in Mathematics & Statistics 201, pp. 197–218.

[27] LYCHE, T., AND SCHUMAKER, L. L. Local spline approximation methods. *Journal of Approximation Theory 15*, 4 (1975), 294–325.

[28] MARKOV, V. On functions of least deviation from zero in a given interval. *St. Petersburg 892* (1892).

[29] MONTARDINI, M., SANGALLI, G., AND TAMELLINI, L. Optimal-order isogeometric collocation at galerkin superconvergent points. *Computer Methods in Applied Mechanics and Engineering 316* (2017), 741–757.

[30] NAIMARK, M. A. *Linear Differential Operators in Hilbert Space, Part I*. F. Ungar Publishing Company, 1968. English trans. by E. R. Dawson, ed. by W. N. Everitt.

[31] PERCELL, P., AND WHEELER, M. F. A C^1 finite element collocation method for elliptic equations. *SIAM Journal on Numerical Analysis 17*, 5 (1980), 605–622.

[32] RUSSELL, R., AND SHAMPINE, L. F. A collocation method for boundary value problems. *Numerische Mathematik 19*, 1 (1972), 1–28.

[33] SABLONNIERE, P. Univariate spline quasi-interpolants and applications to numerical analysis. *arXiv preprint math/0504022* (2005).

[34] SCHOENBERG, I. J. *Cardinal Spline Interpolation*, vol. 12. Siam, 1973.

[35] SCHUMAKER, L. *Spline Functions: Basic theory*. Cambridge University Press, 2007.

[36] SCHUMAKER, L. L. Computing bivariate splines in scattered data fitting and the finite-element method. *Numerical Algorithms 48*, 1-3 (2008), 237–260.

[37] SCHUMAKER, L. L. Splines on spherical triangulations with hanging vertices. *Computer Aided Geometric Design 30*, 3 (2013), 263–275.

[38] SCHUMAKER, L. L. *Spline Functions: Computational Methods*, vol. 142. SIAM, 2015.

[39] SCHUMAKER, L. L., AND WANG, L. Spline spaces on TR-meshes with hanging vertices. *Numerische Mathematik 118*, 3 (2011), 531–548.

[40] SCHUMAKER, L. L., AND WANG, L. Splines on triangulations with hanging vertices. *Constructive Approximation 36*, 3 (2012), 487–511.

[41] ŠOLÍN, P., ČERVENỲ, J., AND DOLEŽEL, I. Arbitrary-level hanging nodes and automatic adaptivity in the hp-fem. *Mathematics and Computers in Simulation 77*, 1 (2008), 117–132.

[42] VARAH, J. M. A lower bound for the smallest singular value of a matrix. *Linear Algebra and its Applications 11*, 1 (1975), 3–5.

[43] VARGA, R. S. On diagonal dominance arguments for bounding $||A^{-1}||_\infty$. *Linear Algebra and its Applications 14*, 3 (1976), 211–217.