Sparse finite element approximation of high-dimensional transport-dominated diffusion problems

C. Schwab, E. Süli[2] and R.A. Todor

Research Report No. 2007-04
March 2007

Seminar für Angewandte Mathematik
Eidgenössische Technische Hochschule
CH-8092 Zürich
Switzerland

[2]University of Oxford, Computing Laboratory, Wolfson Building, Parks Road, Oxford OX1 3QD, UK

Sparse finite element approximation of high-dimensional transport-dominated diffusion problems

C. Schwab, E. Süli[2] and R.A. Todor

Seminar für Angewandte Mathematik
Eidgenössische Technische Hochschule
CH-8092 Zürich
Switzerland

Research Report No. 2007-04 March 2007

Dedicated to Henryk Woźniakowski, on the occasion of his 60th birthday

Abstract

Partial differential equations with nonnegative characteristic form arise in numerous mathematical models in science. In problems of this kind, the exponential growth of computational complexity as a function of the dimension d of the problem domain, the so-called "curse of dimension", is exacerbated by the fact that the problem may be transport-dominated. We develop the numerical analysis of stabilized sparse tensor product finite element methods for such high-dimensional, non-self-adjoint and possibly degenerate second-order partial differential equations, using piecewise polynomials of degree $p \geq 1$. Our convergence analysis is based on new high-dimensional approximation results in sparse tensor-product spaces. By tracking the dependence of the various constants on the dimension d and the polynomial degree p, we show in the case of elliptic transport-dominated diffusion problems that for $p \geq 1$ the error-constant exhibits exponential decay as $d \to \infty$. In the general case when the characteristic form of the partial differential equation is non-negative, under a mild condition relating p to d, the error constant is shown to grow no faster than $\mathcal{O}(d^2)$. In any case, the sparse stabilized finite element method exhibits an optimal rate of convergence with respect to the mesh size h_L, up to a factor that is polylogarithmic in h_L.

Keywords and phrases: high-dimensional Fokker-Planck equations, partial differential equations with nonnegative characteristic form, sparse finite element method

Subject Classification: 65N30

[2]University of Oxford, Computing Laboratory, Wolfson Building, Parks Road, Oxford OX1 3QD, UK

1 Introduction

Suppose that $\Omega := (0,1)^d$, $d \geq 2$, and that $a = (a_{ij})_{i,j=1}^d$ is a symmetric positive semidefinite matrix with entries $a_{ij} \in \mathbb{R}$, $i,j = 1,\ldots,d$. In other words,

$$a^\top = a \qquad \text{and} \qquad \xi^\top a\,\xi \geq 0 \quad \forall \xi \in \mathbb{R}^d.$$

Suppose further that $b \in \mathbb{R}^d$ and $c \in \mathbb{R}$, and let $f \in \mathrm{L}^2(\Omega)$. We shall consider the partial differential equation

$$-a : \nabla\nabla u + b \cdot \nabla u + cu = f(x), \quad x \in \Omega, \tag{1.1}$$

subject to suitable boundary conditions on $\partial\Omega$ which will be stated below. Here $\nabla\nabla u$ is the $d \times d$ Hessian matrix of u whose (i,j) entry is $\partial^2 u/\partial x_i\,\partial x_j$, $i,j = 1,\ldots,d$. For two $d \times d$ matrices A and B, we define their scalar product $A : B := \sum_{i,j=1}^d A_{ij}B_{ij}$. The induced norm, called the Frobenius norm, is defined by $|A| = (A : A)^{1/2}$.

The real-valued polynomial $\alpha \in \mathcal{P}^2(\mathbb{R}^d;\mathbb{R})$ of degree ≤ 2 defined by

$$\xi \in \mathbb{R}^d \;\mapsto\; \alpha(\xi) = \xi^\top a\,\xi \in \mathbb{R}$$

is called the *characteristic polynomial* or *characteristic form* of the differential operator

$$u \;\mapsto\; \mathcal{L}u := -a : \nabla\nabla u + b \cdot \nabla u + cu$$

featuring in (1.1) and, under our hypotheses on the matrix a, the equation (1.1) is referred to as a *partial differential equation with nonnegative characteristic form* (cf. Oleĭnik & Radkevič [21]).

For the sake of simplicity of presentation we shall confine ourselves to differential operators $\mathcal{L}$ with constant coefficients. In this case,

$$a : \nabla\nabla u = \nabla \cdot (a\nabla u) = \nabla\nabla : (au) \qquad \text{and} \qquad b \cdot \nabla u = \nabla \cdot (bu).$$

At the expense of added technical difficulties most of our results can be extended to the case of variable coefficients, where $a = a(x)$, $b = b(x)$ and $c = c(x)$ for $x \in \Omega$.

Partial differential equations with nonnegative characteristic form frequently arise as mathematical models in physics and chemistry [15] (e.g. in the kinetic theory of polymers [22]; see also [3], [4], [18]) and coagulation-fragmentation problems [17]), molecular biology [9], and mathematical finance. Important special cases of these equations include the following:

(a) when the diffusion matrix $a = a^\top$ is positive definite, (1.1) is an elliptic partial differential equation;

(b) when $a \equiv 0$ and the transport direction $b \neq 0$, the partial differential equation (1.1) is a first-order hyperbolic equation;

(c) when

$$a = \begin{pmatrix} \alpha & 0 \\ 0 & 0 \end{pmatrix},$$

where α is a $(d-1) \times (d-1)$ symmetric positive definite matrix and $b = (0,\ldots,0,1)^\top \in \mathbb{R}^d$, (1.1) is a parabolic partial differential equation, with time-like direction b.

In addition to these classical types, the family of partial differential equations with nonnegative characteristic form encompasses a range of other linear second-order partial differential equations, such as degenerate elliptic equations and ultra-parabolic equations. According to a result of Hörmander [12] (cf. Theorem 11.1.10 on p.67), second-order hypoelliptic operators have non-negative characteristic form, after possible multiplication by -1, so they too fall into this category.

For classical types of partial differential equations, such as those listed under (a), (b) and (c) above, rich families of reliable, stable and highly accurate numerical techniques have been developed. Yet, only isolated attempts have been made to explore computational aspects of the class of partial differential equations with nonnegative characteristic form as a whole (cf. [13] and [14]). In particular, there has been only a limited amount of research to date on the numerical analysis of high-dimensional partial differential equations with nonnegative characteristic form (cf. Süli [27], [28]).

The field of stochastic analysis is a particularly fertile source of equations of this kind (see, for example, [5]): the progressive Kolmogorov equation satisfied by the probability density function $\psi(x_1, \ldots, x_d, t)$ of a d-component vectorial stochastic process $X(t) = (X_1(t), \ldots, X_d(t))^\top$ which is the solution of a system of stochastic differential equations including Brownian noise is a partial differential equation with nonnegative characteristic form in the $d+1$ variables $(x, t) = (x_1, \ldots, x_d, t)$. To be more precise, consider the stochastic differential equation:

$$\mathrm{d}X(t) = b(X(t))\,\mathrm{d}t + \sigma(X(t))\,\mathrm{d}W(t), \qquad X(0) = X,$$

where $W = (W_1, \ldots, W_p)^\top$ is a p-dimensional Wiener process adapted to a filtration $\{\mathcal{F}_t\,,\, t \geq 0\}$, $b \in \mathrm{C}^1_b(\mathbb{R}^d; \mathbb{R}^d)$ is the drift vector, and $\sigma \in \mathrm{C}^2_b(\mathbb{R}^d, \mathbb{R}^{d\times p})$ is the diffusion matrix. Here $\mathrm{C}^k_b(\mathbb{R}^n, \mathbb{R}^m)$ denotes the space of bounded and continuous mappings from $\mathbb{R}^n$ into $\mathbb{R}^m$, $m, n \geq 1$, all of whose partial derivatives of order k or less are bounded and continuous on $\mathbb{R}^n$. When the subscript $_b$ is absent, boundedness is not enforced.

Assuming that the random variable $X(t) = (X_1(t), \ldots, X_d(t))^\top$ has a probability density function $\psi \in \mathrm{C}^{2,1}(\mathbb{R}^d \times [0, \infty), \mathbb{R})$, then ψ is the solution of the initial-value problem

$$\begin{aligned} \frac{\partial \psi}{\partial t}(x,t) &= (A\psi)(x,t), \qquad && x \in \mathbb{R}^d,\, t > 0, \\ \psi(x,0) &= \psi_0(x), && x \in \mathbb{R}^d, \end{aligned}$$

where the differential operator $A:\ \mathrm{C}^2(\mathbb{R}^d; \mathbb{R}) \to \mathrm{C}^0(\mathbb{R}^d; \mathbb{R})$ is defined by

$$A\psi := -\sum_{j=1}^{d} \frac{\partial}{\partial x_j}\left(b_j(x)\psi\right) + \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2}{\partial x_i\,\partial x_j}\left(a_{ij}(x)\psi\right),$$

with $a(x) = \sigma(x)\,\sigma^\top(x) \geq 0$ (see Corollary 5.2.10 on p.135 in [16]). Thus, ψ is the solution of the initial-value problem

$$\begin{aligned} \frac{\partial \psi}{\partial t} + \sum_{j=1}^{d} \frac{\partial}{\partial x_j}(b_j(x)\psi) &= \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2}{\partial x_i\,\partial x_j}\left(a_{ij}(x)\psi\right), \qquad && x \in \mathbb{R}^d,\, t \geq 0, \\ \psi(x,0) &= \psi_0(x), && x \in \mathbb{R}^d, \end{aligned}$$

where, for each $x \in \mathbb{R}^d$, $a(x)$ is a $d \times d$ symmetric positive semidefinite matrix. The progressive Kolmogorov equation $\frac{\partial \psi}{\partial t} = A\psi$ is a partial differential equation with nonnegative characteristic form, called a Fokker–Planck equation.

The operator A is generally nonsymmetric (since, typically, $b \neq 0$) and degenerate (since, in general, $a(x) = \sigma(x)\sigma^\top(x)$ has nontrivial kernel). In addition, since the (possibly large) number d of equations in the system of stochastic differential equations is equal to the number of components of the independent variable x of the probability density function ψ, the Fokker–Planck equation may be high-dimensional.

The focus of the present paper is the construction and the analysis of finite element approximations to *high-dimensional* partial differential equations with non-negative characteristic form. Specifically, our aim is to extend the results from [27] and [28], developed for the case of sparse tensor-product finite element spaces consisting of piecewise multilinear functions, to polynomials of degree $p \geq 1$. The paper is structured as follows. We shall state in Section 2 the appropriate boundary conditions for the model equation (1.1), derive the weak formulation of the resulting boundary value problem, and show the existence of a unique weak solution. Section 3 is devoted to the construction of a hierarchical finite element space for univariate functions. The tensorization of this space and the subsequent sparsification of the resulting tensor-product space are described in Section 4; our chief objective is to reduce the computational complexity of the discretization without adversely effecting the approximation properties of the finite element space. In Sections 5 and 6 we build a stabilized finite element method over the sparse tensor-product space, and we explore its stability and convergence.

The convergence analysis relies on new high-dimensional approximation results in sparsified tensor-product spaces, based on continuous piecewise polynomials of degree $p \geq 1$, in the L^2 and H^1 norms. We show that the error-constants in these approximation results exhibit exponentially fast decay as functions of the dimension d. These bounds are related to those in the recent work of Griebel [10], where similar decay of the error-constant as a function of d was proved in the H^1 seminorm in the special case of $p = 1$ for "energy-norm-based" sparse-grid-spaces. Using these approximation results, we then show in the case of elliptic transport-dominated diffusion problems that if $p \geq 1$ then the error-constant exhibits exponential decay as $d \to \infty$. In the general case when the characteristic form of the partial differential equation is non-negative, and assuming that $p \geq 2$, under a mild condition relating p to d the error-constant is shown to grow no faster than $\mathcal{O}(d^2)$. In any case, the sparse stabilized finite element method exhibits an optimal rate of convergence with respect to the mesh size h_L, up to a factor that is polylogarithmic in h_L.

Our error analysis is fairly general, in the sense that only two generic structural properties of the univariate finite element space are used in the subsequent analysis: namely, (1) that the univariate finite element space can be written as a direct sum of so-called increment spaces, and (2) that there exists a projector onto the univariate finite element space which exhibits optimal approximation properties in the L^2 and H^1 norms. The specific choice of basis in the finite element space does not explicitly enter into our error analysis, as it does not affect the asymptotic rate of convergence. Of course, the implementation of the method will necessitate that a choice of basis is made; indeed, the specific choice of basis will strongly influence the sparsity structure and the conditioning of the matrix in the resulting linear system. These questions are important and we shall briefly comment on them in the concluding section, although, strictly speaking, they are beyond the scope of the present paper and will be therefore considered in detail elsewhere.

The origins of sparse tensor-product constructions and hyperbolic cross spaces can be

traced back to the works of Babenko [2] and Smolyak [26]; we refer to the papers of Temlyakov [29], DeVore, Konyagin & Temlyakov [8] for the study of high-dimensional approximation problems, to the works of Wasilkowski & Woźniakowski [30] and Novak & Ritter [19] for high-dimensional integration problems and associated complexity questions, to the paper of Zenger [31] for an early contribution to sparse tensor-product finite element methods, to the articles by von Petersdorff & Schwab [23] and Hoang & Schwab [11] for the analysis of sparse-grid methods for high-dimensional parabolic and elliptic multiscale problems, respectively, and to the recent Acta Numerica article of Bungartz & Griebel [7] for a detailed survey of the field of sparse-grid methods.

2 Boundary conditions and weak formulation

Before embarking on the construction of the numerical algorithm, we shall introduce the necessary boundary conditions and the weak formulation of the model boundary-value problem on $\Omega = (0,1)^d$ for the equation (1.1).

Let Γ denote the union of all $(d-1)$-dimensional open faces of the domain $\Omega = (0,1)^d$. On recalling that, by hypothesis, $a = a^\top$ and $\alpha(\xi) = \xi^\top a\, \xi \geq 0$ for all $\xi \in \mathbb{R}^d$, we define the subset Γ_0 of Γ by

$$\Gamma_0 := \{x \in \Gamma : \alpha(\nu(x)) > 0\};$$

here $\nu(x)$ denotes the unit normal vector to Γ at $x \in \Gamma$, pointing outward with respect to Ω. The set Γ_0 can be thought of as the *elliptic part* of Γ. The complement $\Gamma\backslash\Gamma_0$ of Γ_0 is referred to as the *hyperbolic part* of Γ. We note that, by definition, $\alpha = 0$ on $\Gamma \setminus \Gamma_0$.

On introducing the *Fichera function*

$$x \in \Gamma \mapsto \beta(x) := b \cdot \nu(x) \in \mathbb{R}$$

defined on Γ, we subdivide $\Gamma\backslash\Gamma_0$ as follows:

$$\Gamma_- := \{x \in \Gamma\backslash\Gamma_0 : \beta < 0\}, \qquad \Gamma_+ := \{x \in \Gamma\backslash\Gamma_0 : \beta \geq 0\};$$

the sets Γ_- and Γ_+ are referred to as the (hyperbolic) *inflow* and *outflow* boundary, respectively. Thereby, we obtain the following decomposition of Γ:

$$\Gamma = \Gamma_0 \cup \Gamma_- \cup \Gamma_+.$$

Lemma 1 *Each of the sets Γ_0, Γ_-, Γ_+ is a union of $(d-1)$-dimensional open faces of Ω. Moreover, each pair of opposite $(d-1)$-dimensional faces of Ω is contained either in the elliptic part Γ_0 of Γ or in its complement $\Gamma \setminus \Gamma_0 = \Gamma_- \cup \Gamma_+$, the hyperbolic part of Γ.*

Proof Since a is a constant matrix and ν is a face-wise constant vector, Γ_0 is a union of (disjoint) $(d-1)$-dimensional open faces of Γ. Indeed, if $x \in \Gamma_0$ and y is any point that lies on the same $(d-1)$-dimensional open face of Ω as x, then $\nu(y) = \nu(x)$ and therefore $\alpha(\nu(y)) = \alpha(\nu(x)) > 0$; hence $y \in \Gamma_0$ also.

A certain $(d-1)$-dimensional open face φ of Ω is contained in Γ_0 if, and only if, the opposite face $\hat{\varphi}$ is also contained in Γ_0. To prove this, let $\varphi \subset \Gamma_0$ and let $x = (x_1, \ldots, x_i, \ldots, x_d) \in \varphi$, with Ox_i signifying the (unique) co-ordinate direction such that $\nu(x) \| Ox_i$; here $O = (0, \ldots, 0)$. In other words, $x_i \in \{0,1\}$, and the $(d-1)$-dimensional face φ to which x belongs is orthogonal to the co-ordinate direction Ox_i. Hence, the point $\hat{x} = (x_1, \ldots, |x_i - 1|, \ldots, x_d)$ lies on the $(d-1)$-dimensional open face

$\hat{\varphi}$ of Ω that is opposite the face φ (i.e., $\hat{\varphi}\|\varphi$), and $\nu(\hat{x}) = -\nu(x)$. As α is a homogeneous function of degree 2 on Γ_0, it follows that

$$\alpha(\nu(\hat{x})) = \alpha(-\nu(x)) = (-1)^2\,\alpha(\nu(x)) = \alpha(\nu(x)) > 0,$$

which implies that $\hat{x} \in \Gamma_0$. By what we have shown before, we deduce that the entire face $\hat{\varphi}$ is contained in Γ_0.

Similarly, since b is a constant vector, each of Γ_- and Γ_+ is a union of $(d-1)$-dimensional open faces of Γ. If a certain $(d-1)$-dimensional open face φ is contained in Γ_-, then the opposite face $\hat{\varphi}$ is contained in the set Γ_+.

We note in passing, however, that if $\varphi \subset \Gamma_+$ then the opposite face $\hat{\varphi}$ need not be contained in Γ_-; indeed, if $\varphi \subset \Gamma_+$ and $\beta = 0$ on φ then $\beta = 0$ on $\hat{\varphi}$ also, so then both φ and the opposite face $\hat{\varphi}$ are contained in Γ_+. Of course, if $\beta > 0$ on $\varphi \subset \Gamma_+$, then $\beta < 0$ on the opposite face $\hat{\varphi}$, and then $\hat{\varphi} \subset \Gamma_-$. ■

Lemma 1 motivates the following definition.

Definition 1 *For $i \in \{0, \ldots, d\}$, a co-ordinate direction Ox_i that is orthogonal to a pair of faces of $\Omega = (0,1)^d$ which belong to Γ_0 will be called an elliptic co-ordinate direction. Otherwise, Ox_i will be called a hyperbolic co-ordinate direction.*

We consider the following boundary–value problem: find u such that

$$\mathcal{L}u := -a : \nabla\nabla u + b \cdot \nabla u + cu = f \qquad \text{in } \Omega, \tag{2.1}$$

$$u = 0 \qquad \text{on } \Gamma_0 \cup \Gamma_-. \tag{2.2}$$

Before stating the variational formulation of (2.1), (2.2), we note the following simple result from [14].

Lemma 2 *Suppose that $M \in \mathbb{R}^{d\times d}$ is a $d \times d$ symmetric positive semidefinite matrix. If $\xi \in \mathbb{R}^d$ satisfies $\xi^\top M \xi = 0$, then $M\xi = 0$.*

Since $\nu^\top a \nu = 0$ on $\Gamma \setminus \Gamma_0$ and $a \in \mathbb{R}^{d\times d}$ is a symmetric positive semidefinite matrix, we deduce from Lemma 2 with $M = a$ and $\xi = \nu$ that

$$a\nu = 0 \qquad \text{on } \Gamma \setminus \Gamma_0. \tag{2.3}$$

Let us suppose for a moment that (2.1), (2.2) has a solution u in $\mathrm{H}^2(\Omega)$. Thanks to our assumption that a is a constant matrix, we have that

$$a : \nabla\nabla u = \nabla \cdot (a \nabla u).$$

Furthermore, $a\nabla u \in [\mathrm{H}^1(\Omega)]^d$, which implies that the normal trace $\gamma_{\nu,\partial\Omega}(a\nabla u)$ of $a\nabla u$ on $\partial\Omega$ belongs to $\mathrm{H}^{\frac{1}{2}}(\partial\Omega)$. By virtue of (2.3),

$$\gamma_{\nu,\partial\Omega}(a\nabla u)|_{\Gamma\setminus\Gamma_0} = 0.$$

Note also that $\mathrm{meas}_{d-1}(\partial\Omega \setminus \Gamma) = 0$. Hence

$$\int_{\partial\Omega} \gamma_{\nu,\partial\Omega}(a\nabla u) \cdot \gamma_{0,\partial\Omega}(v)\,\mathrm{d}s = \int_{\Gamma} \gamma_{\nu,\partial\Omega}(a\nabla u)|_\Gamma \cdot \gamma_{0,\partial\Omega}(v)|_\Gamma\,\mathrm{d}s = 0 \tag{2.4}$$

for all $v \in \mathcal{V}$, where

$$\mathcal{V} = \left\{ v \in \mathrm{H}^1(\Omega) : \gamma_{0,\partial\Omega}(v)|_{\Gamma_0} = 0 \right\}.$$

This observation will be of key importance. On multiplying the partial differential equation (2.1) by $v \in \mathcal{V}$ and integrating by parts, we find that

$$(a\nabla u, \nabla v) - (u, \nabla \cdot (bv)) + (cu, v) + \langle u, v \rangle_{\Gamma_+} = (f, v) \qquad \forall v \in \mathcal{V}, \tag{2.5}$$

where $(\cdot, \cdot)$ denotes the L^2 inner–product over Ω and

$$\langle w, v \rangle_{\Gamma_\pm} = \int_{\Gamma_\pm} |\beta| wv \, \mathrm{d}s,$$

with β signifying the Fichera function $b \cdot \nu$, as before. We note that in the transition to (2.5) the boundary integral term on Γ which arises in the course of partial integration from the $-\nabla \cdot (a\nabla u)$ term vanishes by virtue of (2.4), while the boundary integral term on $\Gamma \setminus \Gamma_+ = \Gamma_0 \cup \Gamma_-$ resulting from the $b \cdot \nabla u$ term on partial integration disappears since $u = 0$ on this set by (2.2).

The form of (2.5) serves as motivation for the statement of the weak formulation of (2.1), (2.2) which is presented below. We consider the inner product $(\cdot, \cdot)_{\mathcal{H}}$ defined by

$$(w, v)_{\mathcal{H}} := (a\nabla w, \nabla v) + (w, v) + \langle w, v \rangle_{\Gamma_- \cup \Gamma_+},$$

and denote by $\mathcal{H}$ the closure of the space $\mathcal{V}$ in the norm $\| \cdot \|_{\mathcal{H}}$ defined by $\|w\|_{\mathcal{H}} := (w, w)_{\mathcal{H}}^{\frac{1}{2}}$. Clearly, $\mathcal{H}$ is a Hilbert space. For $w \in \mathcal{H}$ and $v \in \mathcal{V}$, we now consider the bilinear form $B(\cdot, \cdot) : \mathcal{H} \times \mathcal{V} \to \mathbb{R}$ defined by

$$B(w, v) := (a\nabla w, \nabla v) - (w, \nabla \cdot (bv)) + (cw, v) + \langle w, v \rangle_{\Gamma_+},$$

and for $v \in \mathcal{V}$ we introduce the linear functional $L : \mathcal{V} \to \mathbb{R}$ by

$$L(v) := (f, v).$$

We shall say that $u \in \mathcal{H}$ is a *weak solution* to the boundary–value problem (2.1), (2.2) if

$$B(u, v) = L(v) \qquad \forall v \in \mathcal{V}. \tag{2.6}$$

The existence of a unique weak solution is guaranteed by the following theorem (cf. also Theorem 1.4.1 on p.29 of [21]).

Theorem 3 *Suppose that $c \in \mathbb{R}_{>0}$. For each $f \in \mathrm{L}^2(\Omega)$, there exists a unique u in a Hilbert subspace $\hat{\mathcal{H}}$ of $\mathcal{H}$ such that (2.6) holds.*

Proof For $v \in \mathcal{V}$ fixed, we deduce by means of the Cauchy-Schwarz inequality that

$$B(w, v) \leq K_1 \|w\|_{\mathcal{H}} \|v\|_{\mathrm{H}^1(\Omega)} \qquad \forall w \in \mathcal{H},$$

where we have used the trace theorem for $\mathrm{H}^1(\Omega)$. Thus $B(\cdot, v)$ is a bounded linear functional on the Hilbert space $\mathcal{H}$. By the Riesz representation theorem, there exists a unique element $T(v)$ in $\mathcal{H}$ such that

$$B(w, v) = (w, T(v))_{\mathcal{H}} \qquad \forall w \in \mathcal{H}.$$

Since B is bilinear, it follows that $T : v \to T(v)$ is a linear operator from $\mathcal{V}$ into $\mathcal{H}$. Next we show that T is injective. Note that

$$B(v,v) = (a\nabla v, \nabla v) - (v, \nabla\cdot(bv)) + (cv, v) + \langle v, v\rangle_{\Gamma_+} \qquad \forall v \in \mathcal{V}.$$

On integrating by parts in the second term on the right-hand side we deduce that

$$\begin{aligned} B(v,v) &= (a\nabla v, \nabla v) + c\|v\|^2_{\mathrm{L}^2(\Omega)} + \tfrac{1}{2}\langle v, v\rangle_{\Gamma_-\cup\Gamma_+} \\ &\geq K_0\|v\|^2_{\mathcal{H}} \qquad \forall v \in \mathcal{V}, \end{aligned}$$

where $K_0 = \min(c, \frac{1}{2}) > 0$. Hence

$$(v, T(v))_{\mathcal{H}} \geq K_0\|v\|^2_{\mathcal{H}} \qquad \forall v \in \mathcal{V}. \tag{2.7}$$

Consequently, $T : v \to T(v)$ is an injection from $\mathcal{V}$ onto the range $\mathcal{R}(T)$ of T contained in $\mathcal{H}$. Thus, $T : \mathcal{V} \to \mathcal{R}(T)$ is a bijection. Let $S = T^{-1} : \mathcal{R}(T) \to \mathcal{V}$, and let $\hat{\mathcal{H}}$ denote the closure of $\mathcal{R}(T)$ in $\mathcal{H}$. Since, by (2.7), $\|S(w)\|_{\mathcal{H}} \leq (1/K_0)\|w\|_{\mathcal{H}}$ for all $w \in \mathcal{R}(T)$, it follows that $S : \mathcal{R}(T) \to \mathcal{V}$ is a continuous linear operator; therefore, it can be extended, from the dense subspace $\mathcal{R}(T)$ of $\hat{\mathcal{H}}$ to the whole of $\hat{\mathcal{H}}$, as a continuous linear operator $\hat{S} : \hat{\mathcal{H}} \to \mathcal{H}$. Furthermore, since

$$|L(v)| \leq \|f\|_{\mathrm{L}^2(\Omega)}\|v\|_{\mathcal{H}} \qquad \forall v \in \mathcal{H},$$

it follows that $L \circ \hat{S} : v \in \hat{\mathcal{H}} \mapsto L(\hat{S}(v)) \in \mathbb{R}$ is a continuous linear functional on $\hat{\mathcal{H}}$. Since $\hat{\mathcal{H}}$ is closed (by definition) in the norm of $\mathcal{H}$, it is a Hilbert subspace of $\mathcal{H}$. Hence, by the Riesz representation theorem, there exists a unique $u \in \hat{\mathcal{H}}$ such that

$$L(\hat{S}(w)) = (u, w)_{\mathcal{H}} \qquad \forall w \in \hat{\mathcal{H}}.$$

Thus, by the definition of $\hat{S}$, $\hat{S}(w) = S(w)$ for all w in $\mathcal{R}(T)$; hence,

$$L(S(w)) = (u, w)_{\mathcal{H}} \qquad \forall w \in \mathcal{R}(T).$$

Equivalently, on writing $v = S(w)$,

$$(u, T(v))_{\mathcal{H}} = L(v) \qquad \forall v \in \mathcal{V}.$$

Thus we have shown the existence of a unique $u \in \hat{\mathcal{H}}(\subset \mathcal{H})$ such that

$$B(u, v) = (u, Tv)_{\mathcal{H}} = L(v) \qquad \forall v \in \mathcal{V},$$

which completes the proof. ■

We note that the boundary condition $u|_{\Gamma_-} = 0$ on the inflow part Γ_- of the hyperbolic boundary $\Gamma \setminus \Gamma_0 = \Gamma_- \cup \Gamma_+$ is imposed weakly, through the definition of the bilinear form $B(\cdot,\cdot)$, while the boundary condition $u|_{\Gamma_0} = 0$ on the elliptic part Γ_0 of Γ is imposed strongly, through the choice of the function space $\mathcal{H}$. Hence, we deduce from Lemma 1 that

$$\bigotimes_{i=1}^{d} \mathrm{H}^1_{(0)}(0,1) := \mathrm{H}^1_{(0)}(0,1) \otimes \cdots \otimes \mathrm{H}^1_{(0)}(0,1) \subset \mathcal{H}, \tag{2.8}$$

where the i^{th} component $\mathrm{H}^1_{(0)}(0,1)$ in the d-fold tensor-product on the left-hand side of the inclusion is taken to be equal to $\mathrm{H}^1_0(0,1)$ if Ox_i is an elliptic co-ordinate direction; otherwise (i.e. when Ox_i is a hyperbolic co-ordinate direction), it is chosen to be equal to $\mathrm{H}^1(0,1)$.

Next, we shall consider the discretization of the weak formulation (2.6). Motivated by the tensor-product structure of the space on the left-hand side of the inclusion (2.8), we shall base our Galerkin discretization on a finite-dimensional subspace of $\mathcal{H}$ which is the tensor-product of finite-dimensional subspaces of $\mathrm{H}^1_{(0)}(0,1)$. Thus, we begin by setting up the necessary notation in the case of the univariate space $\mathrm{H}^1_{(0)}(0,1)$.

3 Univariate approximation results

Let $I = (0,1)$ and consider the sequence of partitions $\{\mathcal{T}^\ell\}_{\ell\geq 0}$, where $\mathcal{T}^0 = \{I\}$ and where the partition $\mathcal{T}^{\ell+1}$ is obtained from the previous partition

$$\mathcal{T}^\ell := \{I_j^\ell \,:\, j = 0, \ldots, 2^\ell - 1\}$$

by halving each of the intervals I_j^ℓ. The mesh-size in the partition $\mathcal{T}^\ell$ is $h_\ell := 2^{-\ell}$.

We consider the finite-dimensional linear subspace $\mathcal{V}^{\ell,p}$ of $\mathrm{H}^1(0,1)$ consisting of all continuous piecewise polynomials of degree $p \geq 1$ on the partition $\mathcal{T}^\ell$, $\ell \geq 0$. For $\ell \geq 0$ we also consider the subspace $\mathcal{V}_0^{\ell,p}$ of $\mathcal{V}^{\ell,p}$ defined by $\mathcal{V}_0^{\ell,p} := \mathcal{V}^{\ell,p} \cap \mathrm{C}_0[0,1] \subset \mathrm{H}_0^1(0,1)$ consisting of all continuous piecewise polynomial functions on $\mathcal{T}^\ell$ of degree p that vanish at both endpoints of the interval $[0,1]$.

Remark 1 When $p = 1$ the linear space $\mathcal{V}_0^{0,p}$ is trivial, that is $\mathcal{V}_0^{0,1} = \{0\}$. $\diamond$

Let us note that the families of spaces $\{\mathcal{V}_0^{\ell,p}\}_{\ell\geq 0}$ and $\{\mathcal{V}^{\ell,p}\}_{\ell\geq 0}$ are nested, i.e.

$$\mathcal{V}_0^{0,p} \subsetneq \mathcal{V}_0^{1,p} \subsetneq \mathcal{V}_0^{2,p} \subsetneq \cdots \subsetneq \mathcal{V}_0^{\ell,p} \subsetneq \cdots \subsetneq \mathrm{H}_0^1(0,1),$$

and

$$\mathcal{V}^{0,p} \subsetneq \mathcal{V}^{1,p} \subsetneq \mathcal{V}^{2,p} \subsetneq \cdots \subsetneq \mathcal{V}^{\ell,p} \subsetneq \cdots \subsetneq \mathrm{H}^1(0,1),$$

each space in each of the two chains being a proper subspace of the next space in the same chain. As in the previous section, we shall use $\mathrm{H}_{(0)}^1(0,1)$ to denote $\mathrm{H}_0^1(0,1)$ or $\mathrm{H}^1(0,1)$, as the case may be; analogously, we shall use $\mathcal{V}_{(0)}^{\ell,p}$ to denote $\mathcal{V}_0^{\ell,p}$ or $\mathcal{V}^{\ell,p}$. We shall adopt the following hypothesis.

Hypothesis $\mathbf{1}_{(0)}$ *Suppose that $p \geq 1$. For each integer $\ell \geq 0$ there exists a projector (i.e., a linear, idempotent, surjective mapping) $P_{(0)}^{\ell,p} : \mathrm{H}_{(0)}^1(0,1) \to \mathcal{V}_{(0)}^{\ell,p}$.*

Under this hypothesis,

$$\mathcal{V}_{(0)}^{\ell,p} = P_{(0)}^{\ell,p}\, \mathrm{H}_{(0)}^1(0,1), \qquad \ell \geq 0, \quad p \geq 1.$$

Now, let

$$Q_{(0)}^{\ell,p} = \begin{cases} P_{(0)}^{\ell,p} - P_{(0)}^{\ell-1,p}, & \ell \geq 1, \\ P_{(0)}^{0,p}, & \ell = 0. \end{cases}$$

Thus, for any integer $L \geq 0$,

$$P_{(0)}^{L,p} = \sum_{\ell=0}^{L} Q_{(0)}^{\ell,p}.$$

We define the *increment spaces* $\mathcal{W}_{(0)}^{\ell,p}$, $\ell = 0, 1, \ldots$, as follows:

$$\mathcal{W}_{(0)}^{\ell,p} := Q_{(0)}^{\ell,p}\, \mathrm{H}_{(0)}^1(0,1).$$

Hence, for any pair of integers $L \geq 0$ and $p \geq 1$,

$$\mathcal{V}_{(0)}^{L,p} = P_{(0)}^{L,p}\, \mathrm{H}_{(0)}^1(0,1) = \left(\sum_{\ell=0}^{L} Q_{(0)}^{\ell,p}\right) \mathrm{H}_{(0)}^1(0,1) = \sum_{\ell=0}^{L} \left(Q_{(0)}^{\ell,p}\, \mathrm{H}_{(0)}^1(0,1)\right) = \sum_{\ell=0}^{L} \mathcal{W}_{(0)}^{\ell,p}. \tag{3.1}$$

In fact, one can make a stronger statement: for any pair of integers $L \geq 0$ and $p \geq 1$, the vector space $\mathcal{V}_{(0)}^{L,p}$ is the *direct sum* of the increment spaces $\mathcal{W}_{(0)}^{\ell,p}$, $\ell = 0, \ldots, L$:

$$\mathcal{V}_{(0)}^{L,p} = \bigoplus_{\ell=0}^{L} \mathcal{W}_{(0)}^{\ell,p}. \tag{3.2}$$

This is a consequence of the following result (see, for example, [6], Theorem 2.5).

Proposition 4 *Let X be a vector space; then, there exist nontrivial subspaces X_ℓ, $\ell = 0, \ldots, L$, of X such that $X = \bigoplus_{\ell=0}^{L} X_\ell$ if, and only if, there are nonzero linear mappings $p_0, \ldots, p_L : X \to X$ such that*

(1) $\sum_{\ell=0}^{L} p_\ell = \mathrm{Id}_X$;

(2) $p_{\ell_1} \circ p_{\ell_2} = 0_X$ for all $\ell_1, \ell_2 \in \{0, \ldots, L\}$, $\ell_1 \neq \ell_2$.

Moreover, each p_ℓ is necessarily a projector and $X_\ell = \mathrm{Im}(p_\ell)$, $\ell = 0, \ldots, L$.

To prove (3.2), we shall first suppose that $p \geq 2$ and apply Proposition 4 with $X = \mathcal{V}_{(0)}^{L,p}$, $X_\ell = \mathcal{W}_{(0)}^{\ell,p}$ and $p_\ell = Q_{(0)}^{\ell,p}$, $\ell = 0, \ldots, L$, and note that $P_{(0)}^{L,p} = \sum_{\ell=0}^{L} Q_{(0)}^{\ell,p}$ is the identity-map in $\mathcal{V}_{(0)}^{L,p}$ and $Q_{(0)}^{\ell_1,p} \circ Q_{(0)}^{\ell_2,p}$ is the zero-map in $\mathcal{V}_{(0)}^{L,p}$ for all $\ell_1, \ell_2 \in \{0, \ldots, L\}$, $\ell_1 \neq \ell_2$; then, (3.2) follows from (3.1).

When $p = 1$ and $X = \mathcal{V}^{L,1}$ the argument is identical. When $p = 1$ and $X = \mathcal{V}_0^{L,1}$, however, a small modification is required since then $\mathcal{W}_0^{0,1} = \mathcal{V}_0^{0,1} = \{0\}$ and $Q_0^{0,1} = P_0^{0,1} = 0$, so Proposition 4 does not directly apply with $\ell = 0, \ldots, L$. Instead, we use Proposition 4 with $X = \mathcal{V}_0^{L,1}$, $X_\ell = \mathcal{W}_0^{\ell,1}$ and $p_\ell = Q_0^{\ell,1}$, $\ell = 1, \ldots, L$, and note that $P_0^{L,1} = \sum_{\ell=1}^{L} Q_0^{\ell,1}$ is the identity-map in $\mathcal{V}_0^{L,1}$ and $Q_0^{\ell_1,1} \circ Q_0^{\ell_2,1}$ is the zero-map in $\mathcal{V}_0^{L,1}$ for all $\ell_1, \ell_2 \in \{1, \ldots, L\}$, $\ell_1 \neq \ell_2$, to deduce that $\mathcal{V}_0^{L,1} = \bigoplus_{\ell=1}^{L} \mathcal{W}_0^{\ell,1}$. On taking the direct sum of each side in the last equality with $\mathcal{W}_0^{0,1}$, (3.2) follows since $\mathcal{W}_0^{0,1} \oplus \mathcal{V}_0^{L,1} = \mathcal{V}_0^{L,1}$.

Thus we have shown that

$$\mathcal{V}_{(0)}^{L,p} = \bigoplus_{\ell=0}^{L} \mathcal{W}_{(0)}^{\ell,p} = \mathcal{W}_{(0)}^{0,p} \oplus \mathcal{W}_{(0)}^{1,p} \oplus \cdots \oplus \mathcal{W}_{(0)}^{L,p}, \qquad L \geq 0; \tag{3.3}$$

in other words,

$$\mathcal{V}_{(0)}^{\ell,p} = \mathcal{V}_{(0)}^{\ell-1,p} \oplus \mathcal{W}_{(0)}^{\ell,p}, \qquad \ell \geq 1.$$

So far, the choice of the projectors $P_{(0)}^{\ell,p}$ has been fairly arbitrary: the argument above only made use of its algebraic properties stated in Hypothesis $1_{(0)}$. Below, we shall be interested in the convergence of certain tensor-products of the univariate projector. Specifically, we shall investigate the dependence of the convergence rates on the dimension d of the domain of definition $\Omega = (0,1)^d$ of the function u to be approximated as well as on the polynomial degree p. To this end, we shall make a second assumption on the projectors.

Hypothesis $\mathbf{2}_{(0)}$ *Let $k \geq 1$ and $p \geq 1$ be two integers, $s \in \{0, 1\}$ and $h_\ell = 2^{-\ell}$, where $\ell \geq 0$ is an integer, and suppose that $v \in \mathrm{H}^{k+1}(0,1) \cap \mathrm{H}_{(0)}^1(0,1)$. For any integer t such that $1 \leq t \leq \min(p, k)$, there exists a positive constant $c_p(s,t)$, independent of v, such that*

$$|v - P_{(0)}^{\ell,p} v|_{\mathrm{H}^s(0,1)} \leq c_p(s,t) h_\ell^{t+1-s} \, |v|_{\mathrm{H}^{t+1}(0,1)}. \tag{3.4}$$

In particular, Hypothesis $2_{(0)}$ implies that $v = \lim_{\ell\to\infty} P^{\ell,p}_{(0)} v$ for all $v \in \mathrm{H}^2(0,1)\cap\mathrm{H}^1_{(0)}(0,1)$ and all $p \geq 1$, where the limit is considered in the $\mathrm{H}^s(0,1)$-seminorm for $s \in \{0,1\}$, with the convention that, for $s = 0$, $|\cdot|_{\mathrm{H}^0(0,1)} = \|\cdot\|_{\mathrm{L}^2(0,1)}$.

3.1 Examples of univariate projectors

We shall present examples of projectors $P^{\ell,p}_0$ and $P^{\ell,p}$ which satisfy our two hypotheses. Consider the projector $P^{\ell,p} : \mathrm{H}^1(0,1) \to \mathcal{V}^{\ell,p}$ defined, for $x \in [0,1]$, by

$$(P^{\ell,p}u)(x) := u(0) + \int_0^x (\Pi^{\ell,p-1}u')(\xi)\,\mathrm{d}\xi, \qquad \ell \geq 0, \quad p \geq 1,$$

where $\Pi^{\ell,p-1} : \mathrm{L}^2(0,1) \to \mathcal{V}^{\ell,p-1}$ is the $\mathrm{L}^2(0,1)$-orthogonal projector onto $\mathcal{V}^{\ell,p-1}$.

Since $u \in \mathrm{H}^1(0,1) \subset \mathrm{C}[0,1]$, the projector is well-defined. If $p = 1$, the projection $\Pi^{\ell,p-1}(u')$ is a discontinuous, piecewise constant function of the elementwise mean values of u' over subintervals. Consequently, for $p = 1$ we have that $P^{\ell,p}u$ is equal to u at all nodes of $\mathcal{T}^\ell$. More generally, $(P^{\ell,p}u)(1) = u(1)$ for all $\ell \geq 0$ and all $p \geq 1$; furthermore,

$$P^{\ell,p}|_{\mathrm{H}^1_0(0,1)} = P^{\ell,p}_0, \qquad \text{where} \quad (P^{\ell,p}_0 u)(x) := \int_0^x (\Pi^{\ell,p-1}u')(\xi)\,\mathrm{d}\xi \qquad \text{for all } \ell \geq 1.$$

(cf. Theorem 3.14 on p.73 in Schwab [25]).

In addition, the projector $P^{\ell,p}$ has the following approximation property (cf. inequalities (3.3.29) and (3.3.30) in Schwab [25]): for any v in $\mathrm{H}^{k+1}(0,1)$, $k \geq 1$, we have that

$$|v - P^{\ell,p}v|_{\mathrm{H}^s(0,1)} \leq \left(\frac{h_\ell}{2}\right)^{t+1-s} \frac{1}{p^{1-s}} \sqrt{\frac{(p-t)!}{(p+t)!}}\; |v|_{\mathrm{H}^{t+1}(0,1)}, \; 1 \leq t \leq \min(p,k), \tag{3.5}$$

where $h_\ell = 2^{-\ell}$, $\ell \geq 0$, $p \geq 1$, $s \in \{0,1\}$, $t \in \mathbb{N}$, and $\mathbb{N}$ denotes the set of nonnegative integers. An identical bound holds for any v in $\mathrm{H}^{k+1}(0,1) \cap \mathrm{H}^1_{(0)}(0,1)$, $k \geq 1$, with $P^{\ell,p}v$ replaced by $P^{\ell,p}_{(0)}v$.

Thus we have shown that the family of finite element spaces $\{\mathcal{V}^{\ell,p}_{(0)}\}_{\ell\geq 0} \subseteq \mathrm{H}^1_{(0)}(0,1)$ and the associated projector $P^{\ell,p}_{(0)}$ satisfy the approximation property

$$|v - P^{\ell,p}_{(0)}v|_{\mathrm{H}^s(0,1)} \leq c_{p,s,t} 2^{-(t+1-s)(\ell+1)} |v|_{\mathrm{H}^{t+1}(0,1)}, \tag{3.6}$$

for all $v \in \mathrm{H}^{k+1}(0,1) \cap \mathrm{H}^1_{(0)}(0,1)$, $k \geq 1, \ell \geq 0, p \geq 1, t \in \mathbb{N}, 1 \leq t \leq \min(p,k)$ and $s \in \{0,1\}$, where

$$c_{p,s,t} := \frac{1}{p^{1-s}} \sqrt{\frac{(p-t)!}{(p+t)!}}. \tag{3.7}$$

Consequently, Hypotheses $1_{(0)}$ and $2_{(0)}$ hold, inequality (3.4) being satisfied with

$$c_p(s,t) := 2^{-(t+1-s)} c_{p,s,t} = \frac{1}{2^{t+1-s}p^{1-s}} \sqrt{\frac{(p-t)!}{(p+t)!}}. \tag{3.8}$$

3.2 Bounds on the incremental projectors for $p \geq 1$

Let us define, as above, the projection $Q_{(0)}^{\ell,p}$ onto the increment of the hierarchical family $\{\mathcal{V}_{(0)}^{\ell,p}\}_{\ell\geq 0}$ by

$$Q_{(0)}^{\ell,p} := \begin{cases} P_{(0)}^{\ell,p} - P_{(0)}^{\ell-1,p}, & \ell \geq 1, \\ P_{(0)}^{0,p}, & \ell = 0, \end{cases} \tag{3.9}$$

where now $P_{(0)}^{\ell,p}$ signifies the projector introduced in Section 3.1.

Suppose that $v \in \mathrm{H}^{k+1}(0,1) \cap \mathrm{H}^1_{(0)}(0,1)$, $k \geq 1, p \geq 1, t \in \mathbb{N}, 1 \leq t \leq \min(p,k)$ and $s \in \{0,1\}$.

(a) For $\ell \geq 1$, the triangle inequality and the approximation property (3.6) ensure that

$$|Q_{(0)}^{\ell,p} v|_{\mathrm{H}^s(0,1)} \leq \tilde{c}_{p,s,t} 2^{-(t+1-s)\ell} |v|_{\mathrm{H}^{t+1}(0,1)}, \qquad \ell \geq 1, \tag{3.10}$$

with

$$\tilde{c}_{p,s,t} = \left(1 + \frac{1}{2^{t+1-s}}\right) c_{p,s,t}. \tag{3.11}$$

(b) For $\ell = 0$ and $s = 0$, by Poincaré's inequality,

$$\begin{aligned} \|Q_{(0)}^{0,p} u\|_{\mathrm{L}^2(0,1)} &\leq \|u\|_{\mathrm{L}^2(0,1)} + \|u - P_{(0)}^{0,p} u\|_{\mathrm{L}^2(0,1)} \\ &\leq \|u\|_{\mathrm{L}^2(0,1)} + \frac{1}{\pi} |u - P_{(0)}^{0,p} u|_{\mathrm{H}^1(0,1)} \\ &= \|u\|_{\mathrm{L}^2(0,1)} + \frac{1}{\pi} \|u' - \Pi^{0,p-1} u'\|_{\mathrm{L}^2(0,1)} \\ &= \|u\|_{\mathrm{L}^2(0,1)} + \frac{1}{\pi} \sqrt{\|u'\|^2_{\mathrm{L}^2(0,1)} - \|\Pi^{0,p-1} u'\|^2_{\mathrm{L}^2(0,1)}} \\ &= \|u\|_{\mathrm{L}^2(0,1)} + \frac{1}{\pi} \sqrt{\|u'\|^2_{\mathrm{L}^2(0,1)} - |P_{(0)}^{0,p} u|^2_{\mathrm{H}^1(0,1)}} \\ &= \|u\|_{\mathrm{L}^2(0,1)} + \frac{1}{\pi} \sqrt{|u|^2_{\mathrm{H}^1(0,1)} - |Q_{(0)}^{0,p} u|^2_{\mathrm{H}^1(0,1)}}, \end{aligned}$$

and therefore since, for $a \geq b \geq 0$, $\frac{1}{\pi}\sqrt{a^2 - b^2} \leq a - b\sqrt{1 - \frac{1}{\pi^2}}$, we deduce that

$$\|Q_{(0)}^{0,p} u\|_{\mathrm{L}^2(0,1)} + \sqrt{1 - \frac{1}{\pi^2}} |Q_{(0)}^{0,p} u|_{\mathrm{H}^1(0,1)} \leq \|u\|_{\mathrm{L}^2(0,1)} + |u|_{\mathrm{H}^1(0,1)} =: \|u\|_{\mathrm{H}^1_*(0,1)}. \tag{3.12}$$

(c) For $\ell = 0$ and $s = 1$ on the other hand, we have that

$$|Q_{(0)}^{0,p} u|_{\mathrm{H}^1(0,1)} = |P_{(0)}^{0,p} u|_{\mathrm{H}^1(0,1)} \leq |u|_{\mathrm{H}^1(0,1)}. \tag{3.13}$$

Also, (3.12) implies that

$$\|Q_{(0)}^{0,p} u\|_{\mathrm{H}^1_*(0,1)} \leq \frac{\pi}{\sqrt{\pi^2 - 1}} \|u\|_{\mathrm{H}^1_*(0,1)}. \tag{3.14}$$

(d) We note that when $\ell = 0$, $s \in \{0,1\}$, $p = 1$ and $u \in \mathrm{H}^1_0(0,1)$, then we have that $Q_{(0)}^{0,1} u = P_{(0)}^{0,1} u = 0$, and inequalities (3.12) and (3.13) are trivially satisfied.

For $\ell = 0$ we set

$$\begin{aligned} \hat{c}_{p,0,(0)} &:= \|Q^{0,p}_{(0)}\|_{\mathcal{B}(\mathrm{H}^1_{(0)}(0,1),\mathrm{L}^2(0,1))}, \\ \hat{c}_{p,1,(0)} &:= \|Q^{0,p}_{(0)}\|_{\mathcal{B}(\mathrm{H}^1_{(0)}(0,1),\mathrm{H}^1_{(0)}(0,1))}, \end{aligned} \tag{3.15}$$

with the convention that the norm in $\mathrm{H}^1_0(0,1)$ is the seminorm $|\cdot|_{\mathrm{H}^1(0,1)}$ while the norm in $\mathrm{H}^1(0,1)$ is $\|\cdot\|_{\mathrm{H}^1_*(0,1)}$. It will be clear from the context whether we use zero or nonzero boundary conditions in the spaces.

Example 1 If $p \geq 2$ and $\ell = 0$, then $Q^{0,p}_0$ is, for $u \in \mathrm{H}^1_0(0,1)$, given by

$$(Q^{0,p}_0 u)(x) = \int_0^x (\Pi^{0,p-1} u')(\xi)\,\mathrm{d}\xi$$

where $\Pi^{0,p-1}$ denotes the $\mathrm{L}^2(0,1)$-projection onto $\mathcal{V}^{0,p-1}$. Now, $Q^{0,p}_0 \in \mathcal{B}(\mathrm{H}^1_0(0,1), \mathrm{H}^1_0(0,1))$ with $\hat{c}_{p,1,0} = 1$ and, as the embedding of $\mathrm{H}^1_0(0,1)$ into $\mathrm{L}^2(0,1)$ has norm $1/\pi$ by the Poincaré inequality, $\hat{c}_{p,0,0} \leq 1/\pi$. More generally, inequality (3.12) still implies that $\hat{c}_{0,0,(0)} \leq 1$, with $\mathrm{H}^1(0,1)$ equipped by the norm $\|\cdot\|_{\mathrm{H}^1_*(0,1)}$. $\diamond$

3.3 Refined bounds on the incremental projectors for $p = 1$

We shall now take a closer look at estimating $|Q^{\ell,p}_{(0)} v|_{\mathrm{H}^s(0,1)}$ in the case of $p = 1$ and $\ell \geq 1$. As in Section 3.2, $Q^{\ell,p}_{(0)}$ is defined by (3.9) where $P^{\ell,p}_{(0)}$ is the projector introduced in Section 3.1. Our objective is to sharpen our earlier expression (3.11) for the constant $\tilde{c}_{p,s,t}$ appearing in the detail-size estimate (3.10) in the special case of $p = 1$ and $s \in \{0,1\}$ (note that we necessarily have $t = 1$).

We use the following two simple auxiliary results, the first of which is a discrete version of the Poincaré inequality.

Lemma 5 *Suppose that $v \in \mathrm{H}^1_0(0,1)$ is piecewise linear on $\mathcal{T}^1 := \{[0, \frac{1}{2}], [\frac{1}{2}, 1]\}$; then*

$$\|v\|_{\mathrm{L}^2(0,1)} \leq \frac{1}{\sqrt{12}} \|v'\|_{\mathrm{L}^2(0,1)}. \tag{3.16}$$

Proof The result follows from a straightforward calculation with v taken to be the standard hat function $\varphi : x \mapsto \frac{1}{2}(1 - 2|x - \frac{1}{2}|)_+$ defined on $[0,1]$. ∎

Lemma 6 *Suppose that $v \in \mathrm{H}^1(0,1)$; then*

$$\left| \int_0^{1/2} v(t)\,\mathrm{d}t - \int_{1/2}^1 v(t)\,\mathrm{d}t \right| \leq \frac{1}{\sqrt{12}} \|v'\|_{\mathrm{L}^2(0,1)}. \tag{3.17}$$

Proof Denoting, as in the proof Lemma 5, by φ the hat function on $[0,1]$ with $\varphi(\frac{1}{2}) = \frac{1}{2}$, we note that $\mathbf{1}_{[0,\frac{1}{2}]} - \mathbf{1}_{[\frac{1}{2},1]} = \varphi'$. Then, we use partial integration and the Cauchy-Schwarz inequality to obtain

$$\begin{aligned} \left| \int_0^{1/2} v(t)\,\mathrm{d}t - \int_{1/2}^1 v(t)\,\mathrm{d}t \right| &= \left| \int_0^1 \varphi'(t) v(t)\,\mathrm{d}t \right| = \left| \int_0^1 \varphi(t) v'(t)\,\mathrm{d}t \right| \\ &\leq \|\varphi\|_{\mathrm{L}^2(0,1)} \|v'\|_{\mathrm{L}^2(0,1)} = \frac{1}{\sqrt{12}} \|v'\|_{\mathrm{L}^2(0,1)}. \end{aligned} \tag{3.18}$$

That completes the proof. ∎

Remark 2 Rescaling Lemmas 5 and 6 above from $[0,1]$ to $[0,h]$ with $h>0$ we obtain the following inequalities:

$$\|v\|_{\mathrm{L}^2(0,h)} \le \frac{h}{\sqrt{12}}\|v'\|_{\mathrm{L}^2(0,1)} \quad \forall v \in \mathrm{H}^1_0(0,h), \text{ piecewise linear on } [0,h/2]\cup[h/2,h]; \tag{3.19}$$

and

$$\left|\int_0^{h/2} v(t)\,\mathrm{d}t - \int_{h/2}^{h} v(t)\,\mathrm{d}t\right| \le \frac{h^{\frac{3}{2}}}{\sqrt{12}}\|v'\|_{\mathrm{L}^2(0,h)} \quad \forall v \in \mathrm{H}^1(0,h). \quad \diamond \tag{3.20}$$

Proposition 7 *Suppose that the projectors $P^{\ell,1}_{(0)}$, $\ell=0,1,\ldots,$ are given by*

$$(P^{\ell,1}_{(0)}v)(x) = v(0) + \int_0^x \Pi^{\ell,0}(v')(\xi)\,\mathrm{d}\xi \quad \forall v \in \mathrm{H}^1(0,1),\ \ell \ge 0, \tag{3.21}$$

then, for any $v \in \mathrm{H}^2(0,1)$, we have

$$\|Q^{\ell,1}_{(0)}v\|_{\mathrm{L}^2(0,1)} \le \frac{1}{3}2^{-2\ell}\|v''\|_{\mathrm{L}^2(0,1)} \quad \textit{and} \quad |Q^{\ell,1}_{(0)}v|_{\mathrm{H}^1(0,1)} \le \frac{1}{\sqrt{3}}2^{-\ell}\|v''\|_{\mathrm{L}^2(0,1)} \quad \forall \ell \ge 1. \tag{3.22}$$

Hence, for the constants $\tilde{c}_{1,0,1}$ and $\tilde{c}_{1,1,1}$ appearing in (3.10)*, we obtain the upper bounds*

$$\tilde{c}_{1,0,1} \le \frac{1}{3} \quad \textit{and} \quad \tilde{c}_{1,1,1} \le \frac{1}{\sqrt{3}}. \tag{3.23}$$

Proof First note that, since the projector $\Pi^{\ell,0}$ acts by averaging its argument function on each subinterval of the mesh $\mathcal{T}^\ell$, the interpolant $P^{\ell,1}_0$ in (3.21) is nodally exact, that is, $P^{\ell,1}_0 v$ equals v at all nodes (including 0 and 1) of $\mathcal{T}^\ell$, for all $\ell \ge 0$.

Let us denote by I^ℓ_k, for $1 \le k \le 2^\ell$, the subintervals of $\mathcal{T}^\ell$, of length $h^\ell = 2^{-\ell}$. The nodal exactness of $P^{\ell,1}_0$ ensures that $Q^{\ell,1}_0 v\,|_{I^{\ell-1}_k} \in \mathrm{H}^1_0(I^{\ell-1}_k)$ is a multiple of the standard hat function, rescaled to $I^{\ell-1}_k$, for all $\ell \ge 1$ and $1 \le k \le 2^{\ell-1}$. Applying Lemma 5 (after rescaling to $I^{\ell-1}_k$), we obtain

$$\|Q^{\ell,1}_0 v\,|_{I^{\ell-1}_k}\|^2_{\mathrm{L}^2(I^{\ell-1}_k)} \le \frac{1}{12}(h^{\ell-1})^2\|(Q^{\ell,1}_0 v)'\,|_{I^{\ell-1}_k}\|^2_{\mathrm{L}^2(I^{\ell-1}_k)} \quad \forall v \in \mathrm{H}^2(0,1),\ \ell \ge 1. \tag{3.24}$$

The definition (3.21) ensures that

$$(Q^{\ell,1}_0 v)' = \Pi^{\ell,0}(v') - \Pi^{\ell-1,0}(v')$$

so that (since $I^{\ell-1}_k = I^\ell_{2k-1} \cup I^\ell_{2k}$)

$$(Q^{\ell,1}_0 v)'\,|_{I^{\ell-1}_k} = \left(\frac{1}{h^\ell}\int_{I^\ell_{2k-1}} v'(t)\,\mathrm{d}t\right)\mathbf{1}_{I^\ell_{2k-1}} + \left(\frac{1}{h^\ell}\int_{I^\ell_{2k}} v'(t)\,\mathrm{d}t\right)\mathbf{1}_{I^\ell_{2k}} - \left(\frac{1}{h^{\ell-1}}\int_{I^{\ell-1}_k} v'(t)\,\mathrm{d}t\right)\mathbf{1}_{I^{\ell-1}_k}.$$

Using $h^\ell = 2^{-\ell}$ we obtain

$$\begin{aligned}\|(Q^{\ell,1}_0 v)'\,|_{I^{\ell-1}_k}\|^2_{\mathrm{L}^2(I^{\ell-1}_k)} &= \frac{1}{4}\left(\frac{1}{h^\ell}\int_{I^\ell_{2k-1}} v'(t)\,\mathrm{d}t - \frac{1}{h^\ell}\int_{I^\ell_{2k}} v'(t)\,\mathrm{d}t\right)^2 \|\mathbf{1}_{I^{\ell-1}_k}\|^2_{\mathrm{L}^2(I^{\ell-1}_k)} \\ &= \frac{1}{h^{\ell-1}}\left(\int_{I^\ell_{2k-1}} v'(t)\,\mathrm{d}t - \int_{I^\ell_{2k}} v'(t)\,\mathrm{d}t\right)^2, \end{aligned} \tag{3.25}$$

from which we deduce using Lemma 6, rescaled to $I^{\ell-1}_k$, that

$$\|(Q^{\ell,1}_0 v)'\,|_{I^{\ell-1}_k}\|^2_{\mathrm{L}^2(I^{\ell-1}_k)} \le \frac{1}{12}(h^{\ell-1})^2\|v''\|^2_{\mathrm{L}^2(I^{\ell-1}_k)}. \tag{3.26}$$

Estimate (3.22) now follows from (3.24) and (3.26), upon summing over k from 1 to $2^{\ell-1}$. Finally, (3.23) follows by comparing (3.10) and (3.22). ■

4 Sparse finite element discretization

We shall now use the finite-dimensional spaces $\mathcal{V}^{L,p}$ and $\mathcal{V}_0^{L,p}$ of univariate functions to construct a tensor-product space of multivariate functions. We shall then sparsify the resulting tensor-product space with the aim to reduce its dimension without significantly compromising the approximation properties of the original tensor-product space.

4.1 Sparse tensor-product space

Let us first consider on $\Omega = (0,1)^d$ the finite-dimensional subspace $V_{(0)}^{L,p}$ of $\bigotimes_{i=1}^d \mathrm{H}^1_{(0)}(0,1)$ defined by

$$V_{(0)}^{L,p} := \bigotimes_{i=1}^{d} \mathcal{V}_{(0)}^{L,p} = \mathcal{V}_{(0)}^{L,p} \otimes \cdots \otimes \mathcal{V}_{(0)}^{L,p}, \tag{4.1}$$

where the i^{th} component $\mathcal{V}_{(0)}^{L,p}$ in this tensor-product is chosen to be $\mathcal{V}_0^{L,p}$ if Ox_i is an elliptic co-ordinate direction, and $\mathcal{V}_{(0)}^{L,p}$ is chosen as $\mathcal{V}^{L,p}$ otherwise.

In particular, if $a \equiv 0$ and therefore $\Gamma_0 = \emptyset$, then $\mathcal{V}_{(0)}^{L,p} = \mathcal{V}^{L,p}$ for each component in the tensor-product. Conversely, if a is positive definite, then $\Gamma_0 = \Gamma$ and therefore $\mathcal{V}_{(0)}^{L,p} = \mathcal{V}_0^{L,p}$ for each component of the tensor-product. In general, for $a \geq 0$ that is neither identically zero nor positive definite, $\mathcal{V}_{(0)}^{L,p} = \mathcal{V}_0^{L,p}$ for a certain number i of components in the tensor-product, where $0 < i < d$, and $\mathcal{V}_{(0)}^{L,p} = \mathcal{V}^{L,p}$ for the rest.

We denote by $\ell = (\ell_1, \ldots, \ell_d) \in \mathbb{N}^d$ a multi-index and by

$$|\ell|_\infty := \max\{\ell_i : 1 \leq i \leq d\} \quad \text{and} \quad |\ell|_1 := \ell_1 + \cdots + \ell_d$$

its ℓ_∞ and ℓ_1 norms, respectively.

Using the hierarchical decomposition (3.3) we have that

$$V_{(0)}^{L,p} = \bigoplus_{|\ell|_\infty \leq L} \mathcal{W}_{(0)}^{\ell_1,p} \otimes \cdots \otimes \mathcal{W}_{(0)}^{\ell_d,p}, \qquad \ell = (\ell_1, \ldots, \ell_d), \tag{4.2}$$

with the convention that $\mathcal{W}_{(0)}^{\ell,p} = \mathcal{W}_0^{\ell,p}$ whenever Ox_i is an elliptic co-ordinate direction, and $\mathcal{W}_{(0)}^{\ell,p} = \mathcal{W}^{\ell,p}$ otherwise.

For a fixed value of $p \geq 1$, the space $V_{(0)}^{L,p}$ has $\mathcal{O}(2^{Ld})$ degrees of freedom, a number that grows exponentially as a function of d.

In order to reduce the number of degrees of freedom, we shall replace $V_{(0)}^{L,p}$ with a lower-dimensional subspace $\hat{V}_{(0)}^{L,p}$ defined as follows:

$$\hat{V}_{(0)}^{L,p} := \bigoplus_{|\ell|_1 \leq L} \mathcal{W}_{(0)}^{\ell_1,p} \otimes \cdots \otimes \mathcal{W}_{(0)}^{\ell_d,p}, \qquad \ell = (\ell_1, \ldots, \ell_d). \tag{4.3}$$

The space $\hat{V}_{(0)}^{L,p}$ is called a *sparse tensor-product space*.

For a fixed value of $p \geq 1$, the space $\hat{V}_{(0)}^{L,p}$ has $\mathcal{O}(2^L L^{d-1})$ degrees of freedom, which is a considerable reduction compared to the $\mathcal{O}(2^{Ld})$ degrees of freedom for the space $V_{(0)}^{L,p}$.

Let us consider the d-dimensional projector

$$P^{L,p}_{(0)} \otimes \cdots \otimes P^{L,p}_{(0)} \; : \; \bigotimes_{i=1}^{d} \mathrm{H}^1_{(0)}(0,1) \to \bigotimes_{i=1}^{d} \mathcal{V}^{L,p}_{(0)} = V^{L,p}_{(0)},$$

where the i^{th} component $P^{L,p}_{(0)}$ is equal to $P^{L,p}_0$ if Ox_i is an elliptic co-ordinate direction, and equal to $P^{L,p}$ otherwise. Now, let us recall that

$$Q^{\ell,p} = \begin{cases} P^{\ell,p} - P^{\ell-1,p}, & \ell \geq 1, \\ P^{0,p}, & \ell = 0, \end{cases}$$

and

$$Q^{\ell,p}_0 = \begin{cases} P^{\ell,p}_0 - P^{\ell-1,p}_0, & \ell \geq 1, \\ P^{0,p}_0, & \ell = 0. \end{cases}$$

Thus,

$$P^{L,p}_{(0)} = \sum_{\ell=0}^{L} Q^{\ell,p}_{(0)},$$

where $Q^{\ell,p}_{(0)} = Q^{\ell,p}_0$ when $P^{\ell,p}_{(0)} = P^{\ell,p}_0$ and $Q^{\ell,p}_{(0)} = Q^{\ell,p}$ when $P^{\ell,p}_{(0)} = P^{\ell,p}$. Hence,

$$P^{L,p}_{(0)} \otimes \cdots \otimes P^{L,p}_{(0)} = \sum_{|\ell|_\infty \leq L} Q^{\ell_1,p}_{(0)} \otimes \cdots \otimes Q^{\ell_d,p}_{(0)}, \qquad \ell = (\ell_1, \ldots, \ell_d),$$

where $Q^{\ell_i,p}_{(0)}$ is equal to $Q^{\ell_i,p}_0$ when Ox_i is an elliptic co-ordinate direction, and equal to $Q^{\ell_i,p}$ otherwise.

The sparse counterpart $\hat{P}^{L,p}_{(0)}$ of the tensor-product projector $P^{L,p}_{(0)} \otimes \cdots \otimes P^{L,p}_{(0)}$ is defined by truncating the index set $\{\ell \, : \, |\ell|_\infty \leq L\}$ of the sum to $\{\ell \, : \, |\ell|_1 \leq L\}$:

$$\hat{P}^{L,p}_{(0)} := \sum_{|\ell|_1 \leq L} Q^{\ell_1,p}_{(0)} \otimes \cdots \otimes Q^{\ell_d,p}_{(0)} \; : \; \bigotimes_{i=1}^{d} \mathrm{H}^1_{(0)}(0,1) \to \hat{V}^{L,p}_{(0)}, \qquad \ell = (\ell_1, \ldots, \ell_d),$$

where $Q^{\ell_i,p}_{(0)}$ is equal to $Q^{\ell_i,p}_0$ when Ox_i is an elliptic co-ordinate direction Ox_i, and equal to $Q^{\ell_i,p}$ otherwise.

4.2 Sparse stabilized finite element method

Having defined the finite-dimensional space $\hat{V}^{L,p}_{(0)}$ in which the approximate solution will be sought, we now introduce a stabilized Galerkin finite element method over this finite-dimensional space. The main ingredients of the method are a bilinear form $b_\delta(\cdot,\cdot)$ which approximates the bilinear form $B(\cdot,\cdot)$ from the weak formulation of the boundary value problem and a linear functional $l_\delta(\cdot)$ which approximates the linear functional $L(\cdot)$.

Let us consider the bilinear form

$$b_\delta(w,v) := B(w,v) + \delta_L \sum_{\kappa \in \mathcal{T}^L} (\mathcal{L}w, b \cdot \nabla v)_\kappa.$$

Here, in the light of the fact that in the transport-dominated case $|a| \ll |b|$, the second term in the bilinear form $b_\delta(\cdot,\cdot)$ can be thought of as least-square stabilization in the direction of subcharacteristics ('streamlines').

We also define the linear functional

$$l_\delta(v) := L(v) + \delta_L \sum_{\kappa\in\mathcal{T}^L} (f, b\cdot\nabla v)_\kappa \quad (= L(v) + \delta_L(f, b\cdot\nabla v))\,.$$

Here $\delta_L \in [0, 1/c]$ is a ('streamline-diffusion') parameter to be chosen below, and $\kappa \in \mathcal{T}^L$ are d-dimensional axiparallel cubic elements of edge-length h_L in the partition of the computational domain $\Omega = (0,1)^d$. As there are 2^{Ld} such elements κ in $\mathcal{T}^L$, the computation of the *stabilization term* $\delta_L \sum_{\kappa\in\mathcal{T}^L}(\mathcal{L}w, b\cdot\nabla v)_\kappa$ in the definition of $b_\delta(w,v)$ may seem intractable for $d \gg 1$; however, it turns out that this is not so: the sum over the 2^{Ld} terms in the stabilization term can be rewritten as a sum over $Ld + \frac{1}{2}d(d-1) + 1$ terms only; see Remark 13(c).

We consider the finite-dimensional problem: find $u_h \in \hat{V}_{(0)}^{L,p}$ such that

$$b_\delta(u_h, v_h) = l_\delta(v_h) \qquad \forall v_h \in \hat{V}_{(0)}^{L,p}. \tag{4.4}$$

The idea behind the method (4.4) is to introduce mesh-dependent numerical diffusion into the standard Galerkin finite element method along subcharacteristic directions, with the aim to suppress maximum-principle-violating oscillations on the scale of the mesh, and let $\delta_L \to 0$ with $h_L \to 0$. For an analysis of the method in the case of standard finite element spaces and (low-dimensional) elliptic transport-dominated diffusion equations we refer to the monograph [24].

It would have been more accurate to write u_{h_L} and v_{h_L} instead of u_h and v_h in (4.4). However, to avoid notational clutter, we shall refrain from doing so. Instead, we adopt the convention that the dependence of $h = h_L$ on the index L will be implied, even when not explicitly noted.

We begin with the stability-analysis of the method. First, we shall show that, with an appropriate choice of the streamline-diffusion parameter δ_L, the bilinear form $b_\delta(\cdot,\cdot)$ is coercive on $V_{(0)}^{L,p} \times V_{(0)}^{L,p}$. To this end, we begin by noting that

$$\begin{aligned}
b_\delta(v_h, v_h) &= (a\nabla v_h, \nabla v_h) - (v_h, \nabla\cdot(bv_h)) + (cv_h, v_h) + \langle v_h, v_h\rangle_{\Gamma_+} + \delta_L \sum_{\kappa\in\mathcal{T}^L} (\mathcal{L}v_h, b\cdot\nabla v_h)_\kappa \\
&= (a\nabla v_h, v_h) + c\|v_h\|^2_{\mathrm{L}^2(\Omega)} + \delta_L\|b\cdot\nabla v_h\|^2_{\mathrm{L}^2(\Omega)} \\
&\quad + \frac{1}{2}\int_{\Gamma_-\cup\Gamma_+} |\beta|\,|v_h|^2\,\mathrm{d}s + \frac{1}{2}c\delta_L\int_{\Gamma_-\cup\Gamma_+}\beta|v_h|^2\,\mathrm{d}s \\
&\quad + \delta_L\sum_\kappa(-a:\nabla\nabla v_h, b\cdot\nabla v_h)_\kappa \\
&\geq \|\sqrt{a}\,\nabla v_h\|^2_{\mathrm{L}^2(\Omega)} + c\|v_h\|^2_{\mathrm{L}^2(\Omega)} + \frac{1}{2}\delta_L\|b\cdot\nabla v_h\|^2_{\mathrm{L}^2(\Omega)} \\
&\quad + \frac{1}{2}(1+c\delta_L)\int_{\Gamma_+}|\beta||v|^2\,\mathrm{d}s + \frac{1}{2}(1-c\delta_L)\int_{\Gamma_-}|\beta||v|^2\,\mathrm{d}s \\
&\quad - \frac{1}{2}\delta_L\sum_\kappa\|a:\nabla\nabla v_h\|^2_{\mathrm{L}^2(\kappa)} \qquad \forall v_h \in V_{(0)}^{L,p},
\end{aligned} \tag{4.5}$$

where we have made use of the facts that $\beta = -|\beta|$ on Γ_- and $v_h|_{\Gamma_0} = 0$.

In the case of $p = 1$, the last term in (4.5) is equal to zero, and therefore coercivity of $b_\delta(\cdot,\cdot)$ in the streamline-diffusion norm $|||\cdot|||_{\rm SD}$, defined by

$$|||v|||_{\rm SD}^2 := \|\sqrt{a}\,\nabla v\|_{{\rm L}^2(\Omega)}^2 + c\|v\|_{{\rm L}^2(\Omega)}^2 + \delta_L\|b\cdot\nabla v\|_{{\rm L}^2(\Omega)}^2 + \frac{1}{2}(1+c\delta_L)\int_{\Gamma_+}|\beta||v|^2\,{\rm d}s,$$

then follows immediately, provided that δ_L is chosen so that

$$0 \le \delta_L \le \frac{1}{c}.$$

Here $\sqrt{a} \in \mathbb{R}^{d\times d}$ is the symmetric positive semidefinite square-root of the symmetric positive semidefinite matrix $a \in \mathbb{R}^{d\times d}$.

In the case of $p > 1$, however, the final term in (4.5) is generally nonzero. Nevertheless, we shall show that, with a somewhat more restrictive choice of δ_L, the extra term can be absorbed into the first term on the right-hand side of (4.5). In order to avoid having to distinguish between the cases $p = 1$ and $p > 1$, we shall assume henceforth that $p \ge 1$, with the understanding that in the case of $p = 1$ our results can be slightly sharpened in a manner which we shall merely comment on.

To proceed with the case of a general $p \ge 1$, we require the following inverse inequality for a univariate function. For its proof, we refer to Schwab [25], p.148, Theorem 3.91.

Lemma 8 *Let $I = (a,b) \subset \mathbb{R}$, $h = b - a$ and $p \ge 1$. Then,*

$$\|v'\|_{{\rm L}^2(I)} \le \sqrt{12}\,\frac{p^2}{h}\|v\|_{{\rm L}^2(I)} \qquad \forall v \in \mathcal{P}^p(I),$$

where $\mathcal{P}^p(I)$ denotes the set of all polynomials of degree p or less defined on the closed interval $\bar{I} = [a,b]$.

Now, letting $w_i := (a\nabla v_h)_i$, we have that

$$\begin{aligned}
\|a : \nabla\nabla v_h\|_{{\rm L}^2(\kappa)}^2 &= \|\nabla\cdot(a\nabla v_h)\|_{{\rm L}^2(\kappa)}^2 = \left\|\sum_{i=1}^d \frac{\partial}{\partial x_i} w_i\right\|_{{\rm L}^2(\kappa)}^2 \\
&\le \left(\sum_{i=1}^d \left\|\frac{\partial w_i}{\partial x_i}\right\|_{{\rm L}^2(\kappa)}\right)^2 \le d\sum_{i=1}^d \left\|\frac{\partial w_i}{\partial x_i}\right\|_{{\rm L}^2(\kappa)}^2 \le \frac{12dp^4}{h_L^2}\sum_{i=1}^d \|w_i\|_{{\rm L}^2(\kappa)}^2 \\
&= \frac{12dp^4}{h_L^2}\int_\kappa \sum_{i=1}^d |w_i|^2\,{\rm d}x = \frac{12dp^4}{h_L^2}\int_\kappa |w|^2\,{\rm d}x = \frac{12dp^4}{h_L^2}\int_\kappa |a\nabla v_h|^2\,{\rm d}x \\
&= \frac{12dp^4}{h_L^2}\int_\kappa |\sqrt{a}\,\sqrt{a}\,\nabla v_h|^2\,{\rm d}x \le \frac{12dp^4}{h_L^2}\int_\kappa |\sqrt{a}\,|^2\,|\sqrt{a}\,\nabla v_h|^2\,{\rm d}x \\
&= |\sqrt{a}\,|^2\frac{12dp^4}{h_L^2}\,\|\sqrt{a}\,\nabla v_h\|_{{\rm L}^2(\kappa)}^2,
\end{aligned}$$

where $|w|$ denotes the ℓ^2 norm of $w \in \mathbb{R}^d$ and $|\sqrt{a}\,|$ again denotes the Frobenius norm of the symmetric positive semidefinite matrix $\sqrt{a} \in \mathbb{R}^{d\times d}$. Hence, after summation over all $\kappa \in \mathcal{T}^L$,

$$\sum_{\kappa\in\mathcal{T}^L} \|a : \nabla\nabla v_h\|_{{\rm L}^2(\kappa)}^2 \le |\sqrt{a}\,|^2\frac{12dp^4}{h_L^2}\,\|\sqrt{a}\,\nabla v_h\|_{{\rm L}^2(\Omega)}^2.$$

Using this bound in (4.5) we deduce that

$$\begin{aligned} b_\delta(v_h,v_h) \;\geq\; & \left(1-\delta_L|\sqrt{a}\,|^2\frac{6dp^4}{h_L^2}\right)\|\sqrt{a}\,\nabla v_h\|^2_{\mathrm{L}^2(\Omega)}+c\|v_h\|^2_{\mathrm{L}^2(\Omega)}+\frac{1}{2}\delta_L\|b\cdot\nabla v_h\|^2_{\mathrm{L}^2(\Omega)} \\ & +\frac{1}{2}(1+c\delta_L)\int_{\Gamma_+}|\beta|\,|v_h|^2\,\mathrm{d}s+\frac{1}{2}(1-c\delta_L)\int_{\Gamma_-}|\beta|\,|v_h|^2\,\mathrm{d}s. \end{aligned}$$

Let us suppose that

$$0\leq\delta_L\leq\min\left(\frac{h_L^2}{12dp^4|\sqrt{a}\,|^2}\,,\,\frac{1}{c}\right).$$

Then,

$$\begin{aligned} b_\delta(v_h,v_h) \;\geq\; & \frac{1}{2}\|\sqrt{a}\,\nabla v_h\|^2_{\mathrm{L}^2(\Omega)}+c\|v_h\|^2_{\mathrm{L}^2(\Omega)}+\frac{1}{2}\delta_L\|b\cdot\nabla v_h\|^2_{\mathrm{L}^2(\Omega)} \\ & +\frac{1}{2}(1+c\delta_L)\int_{\Gamma_+}|\beta|\,|v_h|^2\,\mathrm{d}s+\frac{1}{2}(1-c\delta_L)\int_{\Gamma_-}|\beta|\,|v_h|^2\,\mathrm{d}s \\ \geq\; & \frac{1}{2}|||v_h|||^2_{\mathrm{SD}}\qquad\forall v_h\in\hat{V}^{L,p}_{(0)}. \end{aligned} \tag{4.6}$$

Since (4.4) is a linear problem in a finite-dimensional linear space, the coercivity (4.6) of the bilinear form $b_\delta(\cdot,\cdot)$ implies the existence and uniqueness of a solution u_h to (4.4) in $\hat{V}^{L,p}_{(0)}$. Furthermore,

$$\frac{1}{2}|||u_h|||^2_{\mathrm{SD}}\leq\left(\frac{1}{c}+\delta_L\right)^{\frac{1}{2}}\|f\|_{\mathrm{L}^2(\Omega)}\,|||u_h|||_{\mathrm{SD}},$$

which, in turn, implies that

$$|||u_h|||_{\mathrm{SD}}\leq(8/c)^{\frac{1}{2}}\|f\|_{\mathrm{L}^2(\Omega)}, \tag{4.7}$$

and hence the stability of the method for all δ_L such that

$$0\leq\delta_L\leq\min\left(\frac{h_L^2}{12dp^4|\sqrt{a}\,|^2},\frac{1}{c}\right).$$

We note here that in the case of $p=1$ the constant $\frac{1}{2}$ in the coercivity result $b_\delta(v_h,v_h)\geq\frac{1}{2}|||v_h|||^2_{\mathrm{SD}}$ stated in (4.6) can be replaced by 1, under the simpler condition $0\leq\delta_L\leq 1/c$ which does not involve the matrix norm $|\sqrt{a}\,|$ or the dimension d. Consequently, the constant $(8/c)^{\frac{1}{2}}$ in the stability inequality (4.7) can then be improved to $(2/c)^{\frac{1}{2}}$, under this same condition on δ_L.

In Section 6 we shall consider the convergence analysis of the method (4.4); we shall require there the following multiplicative trace inequality.

Lemma 9 (Multiplicative trace inequality) *Let $\Omega=(0,1)^d$ where $d\geq 2$ and suppose that Γ_+ is the hyperbolic outflow part of Γ. Then,*

$$\int_{\Gamma_+}|v|^2\,\mathrm{d}s\leq 4d\|v\|_{\mathrm{L}^2(\Omega)}\,\|v\|_{\mathrm{H}^1(\Omega)}\qquad\forall v\in\mathrm{H}^1(\Omega).$$

Proof We shall prove the inequality for $v \in \mathrm{C}^1(\bar{\Omega})$. For $v \in \mathrm{H}^1(\Omega)$ the result follows from the density of $\mathrm{C}^1(\bar{\Omega})$ in $\mathrm{H}^1(\Omega)$. As we have noted before, Γ_+ is a union of $(d-1)$-dimensional open faces of Ω. Let us suppose without loss of generality that the face $x_1 = 0$ of Ω belongs to Γ_+. Then,

$$v^2(0,x') = v^2(x_1,x') + \int_{x_1}^0 \frac{\partial}{\partial x_1} v^2(\xi,x')\,\mathrm{d}\xi, \qquad x' = (x_2,\ldots,x_n).$$

Hence, on integrating this over $x = (x_1,x') \in (0,1)\times(0,1)^{d-1} = \Omega$,

$$\begin{aligned}\int_{x'\in(0,1)^{d-1}} v^2(0,x')\,\mathrm{d}x' &= \int_0^1 \int_{x'\in(0,1)^{d-1}} v^2(x_1,x')\,\mathrm{d}x'\,\mathrm{d}x_1 \\ &\quad + 2\int_0^1 \int_{x'\in(0,1)^{d-1}} \int_{x_1}^0 v(\xi,x')\frac{\partial}{\partial x_1} v(\xi,x')\,\mathrm{d}\xi\;\mathrm{d}x'\,\mathrm{d}x_1 \\ &\le \|v\|^2_{\mathrm{L}^2(\Omega)} + 2\|v\|\,\|v_{x_1}\|.\end{aligned}$$

In the generic case when $\beta > 0$ on the whole of Γ_+, the set Γ_+ will contain at most d of the $2d$ faces of Ω, — at most one complete face of Ω orthogonal to the i^{th} co-ordinate direction, $i = 1,\ldots,d$. Otherwise, if $\beta = 0$ on certain faces that belong to Γ_+, the set Γ_+ may contain as many as $2d-1$ of the $2d$ faces of Ω. Thus, in the worst case,

$$\int_{\Gamma_+} |v|^2\;\mathrm{d}s \le (2d-1)\|v\|^2_{\mathrm{L}^2(\Omega)} + 4\|v\|_{\mathrm{L}^2(\Omega)} \sum_{i=1}^d \|v_{x_i}\|_{\mathrm{L}^2(\Omega)}. \tag{4.8}$$

Therefore,

$$\int_{\Gamma_+} |v|^2\;\mathrm{d}s \le 2d\sqrt{2}\max\left(1,\frac{2}{d^{\frac{1}{2}}}\right)\|v\|_{\mathrm{L}^2(\Omega)}\,\|v\|_{\mathrm{H}^1(\Omega)} \le 4d\|v\|_{\mathrm{L}^2(\Omega)}\,\|v\|_{\mathrm{H}^1(\Omega)}.$$

Hence the required result. ∎

Remark 3 It follows from (4.8) that by altering the definition of the $\mathrm{H}^1(\Omega)$ norm in a similar manner as in (3.12), the constant in Lemma 9 can be slightly improved:

$$\int_{\Gamma_+} |v|^2\;\mathrm{d}s \le 2d\|v\|_{\mathrm{L}^2(\Omega)}\,\|v\|_{\mathrm{H}^1_*(\Omega)}, \qquad \text{where} \qquad \|v\|_{\mathrm{H}^1_*(\Omega)} := \|v\|_{\mathrm{L}^2(\Omega)} + \sum_{i=1}^d \|v_{x_i}\|_{\mathrm{L}^2(\Omega)}. \quad \diamond$$

5 Approximation from sparse finite element spaces

Our objective in Section 6 will be to establish the convergence of the stabilized sparse finite element method. To this end, we first prove some combinatorial bounds on lattice sums which will then be used for quantifying the error between a function and its projection onto the sparsified finite element space. We shall also prove our key technical tool: a result on linear operators, which are bounded in suitable semi-norms, on tensor-products of Hilbert spaces. As before, $\mathbb{N}$ will denote the set of non-negative integers.

5.1 Combinatorial bounds on lattice sums

Lemma 10 *For $d \in \mathbb{N}_{>0}$ and $t > 1$ we have that*

$$\sup_{m\in\mathbb{N}} \sum_{\substack{\ell\in\mathbb{N}^d \\ |\ell|_1 = m}} t^{|\ell|_\infty - m} = d\left(1 + \frac{1}{t-1}\right)^{d-1}. \tag{5.1}$$

Proof The case $d = 1$ being trivial, we assume without loss of generality that $d \geq 2$. Let us denote by $A(m, t, d)$ the sum in (5.1) and rewrite it as

$$A(m,t,d) = \sum_{k=0}^{\infty} \sum_{\substack{\ell \in \mathbb{N}^d \\ |\ell|_1 = m,\, |\ell|_\infty = k}} t^{k-m} = \sum_{k=0}^{\infty} |\mathcal{S}(m,k,d)| t^{k-m}, \tag{5.2}$$

where the set $\mathcal{S}(m, k, d)$ is defined by

$$\mathcal{S}(m,k,d) := \{\ell \in \mathbb{N}^d \,:\, |\ell|_1 = m,\, |\ell|_\infty = k\}. \tag{5.3}$$

We deduce from (5.9) in Lemma 11 below that

$$d \sum_{m/2 < k \leq m} \binom{m-k+d-2}{d-2} t^{k-m} \leq A(m,t,d) \leq d \sum_{k=0}^{m} \binom{m-k+d-2}{d-2} t^{k-m}. \tag{5.4}$$

The statement of the theorem will follow once we have shown that the suprema over $m \in \mathbb{N}$ of both the lower and the upper bound in (5.4) are equal to the right-hand side of (5.1).

We start by considering the upper bound in (5.4), which can be written, after substituting k by $m - k$, as

$$d \sum_{k=0}^{m} \binom{k+d-2}{d-2} \left(\frac{1}{t}\right)^k.$$

The supremum over $m \in \mathbb{N}$ is thus attained for $m \to \infty$ and equals

$$d \left(\frac{1}{1-1/t}\right)^{d-1}. \tag{5.5}$$

Note that here we have used the identity

$$\frac{1}{(1-x)^{n+1}} = \sum_{k=0}^{\infty} \binom{k+n}{n} x^k \qquad \forall n \in \mathbb{N},\ \forall x \in (-1,1)$$

which follows by differentiating n times with respect to x the identity $(1-x)^{-1} = 1 + x + x^2 + \cdots$.

Now we use a similar argument to compute the supremum over $m \in \mathbb{N}$ of the lower bound in (5.4), which can be written, again after substituting k by $m - k$, as

$$d \sum_{0 \leq k < m/2} \binom{k+d-2}{d-2} \left(\frac{1}{t}\right)^k.$$

The supremum over $m \in \mathbb{N}$ is attained again for $m \to \infty$ and equals (5.5). ∎

In particular, it is a simple consequence of this theorem that, for any $d, m \in \mathbb{N}_{>0}$ and $t > 1$, we have that

$$d \cdot t^m \leq \sum_{\substack{\ell \in \mathbb{N}^d \\ |\ell|_1 = m}} t^{|\ell|_\infty} \leq d \left(1 + \frac{1}{t-1}\right)^{d-1} \cdot t^m, \tag{5.6}$$

the lower bound being trivial.

The next lemma summarizes some useful properties of the sets $\mathcal{S}(m, k, d)$ defined in (5.3) above.

Lemma 11 *Consider the sets $\mathcal{S}(m, k, d)$ defined, for $d \in \mathbb{N}_{>0}$ and $m, k \in \mathbb{N}$, by*

$$\mathcal{S}(m,k,d) := \{\ell \in \mathbb{N}^d \,:\, |\ell|_1 = m,\, |\ell|_\infty = k\};$$

then

$$\mathcal{S}(m,k,d) = \emptyset \qquad \forall k > m, \tag{5.7}$$

$$\sum_{k=0}^{\infty} |\mathcal{S}(m,k,d)| = \binom{m+d-1}{d-1}, \tag{5.8}$$

$$|\mathcal{S}(m,k,d)| \le d\binom{m-k+d-2}{d-2} \qquad \forall d \ge 2, \tag{5.9}$$

with equality for $k > m/2$.

Proof We note that (5.7) is obvious, whereas (5.8) follows from the fact that for fixed m, d, the sets $(\mathcal{S}(m,k,d))_{0\le k\le m}$ are disjoint and

$$\bigcup_{k=0}^{m} \mathcal{S}(m,k,d) = \{\ell \in \mathbb{N}^d \,:\, |\ell|_1 = m\}.$$

To prove (5.9), we consider for fixed k, m with $0 \le k \le m$, the mapping

$$\{1,2,\ldots,d\} \times \bigcup_{j=0}^{k} \mathcal{S}(m-k,j,d-1) \xrightarrow{\varphi} \mathcal{S}(m,k,d)$$

given by

$$\varphi(q,(l_1,l_2,\ldots,l_{d-1})) = (l_1,l_2,\ldots,l_{q-1},k,l_q,\ldots,l_{d-1}).$$

Obviously, φ is surjective, so that we obtain, using (5.8),

$$\begin{aligned} |\mathcal{S}(m,k,d)| &\le |\{1,2,\ldots,d\}| \cdot \sum_{j=0}^{k} |\mathcal{S}(m-k,j,d-1)| && (5.10)\\ &\le d\binom{m-k+d-2}{d-2}. && (5.11) \end{aligned}$$

For $k > m/2$ the mapping φ is also injective, which ensures equality in (5.10). Also (5.11) holds with equality for $k > m/2$ due to (5.7) and (5.8). ■

Remark 4 Of particular interest is the case $t = 2$, for which (5.1) becomes

$$\sum_{\substack{\ell\in\mathbb{N}^d\\ |\ell|_1=m}} 2^{|\ell|_\infty - m} \le d2^{d-1} \qquad \forall m \in \mathbb{N},\ \forall d \in \mathbb{N}_{>0}. \qquad \diamond \tag{5.12}$$

In the following we validate numerically the identity (5.1). The computation of the left-hand side will be based on a recursive (in d) formula for $|\mathcal{S}(m,k,d)|$ via (5.2), which reads in this case,

$$\sum_{\substack{\ell\in\mathbb{N}^d\\ |\ell|_1=m}} 2^{|\ell|_\infty - m} = \sum_{k=0}^{m} |\mathcal{S}(m,k,d)| 2^{k-m}. \tag{5.13}$$

Lemma 12 *Suppose that $m, k, d \in \mathbb{N}_{>0}$; then the following identity holds:*

$$|\mathcal{S}(m,k,d)| = \sum_{1\le n\le m/k} \binom{d}{n} \sum_{j=0}^{k-1} |\mathcal{S}(m-nk,j,d-n)|.$$

Proof The formula follows by noting that, for $\ell \in \mathcal{S}(m,k,d)$, the value $k = |\ell|_\infty$ can be attained n times (that is, by n of the co-ordinates $l_1, l_2, \ldots, l_d$), with $1 \le n \le m/k$. These n co-ordinates can be chosen freely from $\{1,2,\ldots,d\}$, and the multi-index consisting of the remaining $d-n$ co-ordinates belongs to $\mathcal{S}(m-nk, j, d-n)$ for some $0 \le j \le k-1$. ■

We shall also require the following lemma.

Lemma 13 *Suppose that $d \ge 2$, $t > 0$ and $L \ge 1$; then,*

$$\sum_{\ell \in \mathbb{N}^d \,:\, |\ell|_1 > L} 2^{-t|\ell|_1} \;=\; c_{d,t,L} \cdot 2^{-tL}\, L^{d-1},$$

where

$$c_{d,t,L} := \frac{1}{2^t\, L^{d-1}\,(d-1)!} \cdot \frac{(L+d)!}{(L+1)!} \cdot {}_2\mathrm{F}_1(L+d+1, 1; L+2, 2^{-t}),$$

and ${}_2\mathrm{F}_1(a,b;c;z)$ is the Gauss hypergeometric function. Furthermore:

(i) For any $t > 0$ and $d \ge 2$ fixed,

$$c_{d,t,L} \sim \frac{1}{(2^t-1)\,(d-1)!} \quad \text{as } L \to \infty.$$

(ii) For any $t > 0$ and $L \ge 1$,

$$c_{d,t,L} \sim \frac{1}{L^{d-1}} \left\{ 2^{tL} \left(\frac{2^t}{2^t-1} \right)^d - \binom{L+d}{d} \right\} \quad \text{as } d \to \infty.$$

In particular, for $t > 0$ and $L \ge \frac{2^t}{2^t-1}$ fixed, $\lim_{d\to\infty} c_{d,t,L} = 0$.

(iii) For any $t > 1$ fixed,

$$c_{d,t,L} \sim \frac{2^{L-d}}{\sqrt{\pi}(2^t-2)} \left(\frac{4}{L^{1-\frac{1}{2d}}} \right)^d \quad \text{as } L \to \infty \text{ and } d \to \infty, \text{ with } L-d \text{ bounded},$$

and hence $\lim\limits_{\substack{L\to\infty, d\to\infty \\ L-d \text{ bounded}}} c_{d,t,L} = 0$ for any $t > 1$ fixed.

Proof We begin by noting that

$$\begin{aligned}
\sum_{\ell \in \mathbb{N}^d \,:\, |\ell|_1 > L} 2^{-t|\ell|_1} &= \sum_{m=L+1}^{\infty} \sum_{\ell \in \mathbb{N}^d \,:\, |\ell|_1 = m} 2^{-t|\ell|_1} = \sum_{m=L+1}^{\infty} \sum_{\ell \in \mathbb{N}^d \,:\, |\ell|_1 = m} 2^{-tm} \\
&= \sum_{m=L+1}^{\infty} |\{\ell \in \mathbb{N}^d \,:\, |\ell|_1 = m\}| \cdot 2^{-tm} \\
&= \sum_{m=L+1}^{\infty} \sum_{k=0}^{\infty} |\{\ell \in \mathbb{N}^d \,:\, |\ell|_1 = m,\, |\ell|_\infty = k\}| \cdot 2^{-tm} \\
&= \sum_{m=L+1}^{\infty} \sum_{k=0}^{\infty} |\mathcal{S}(m,k,d)|\, 2^{-tm} = \sum_{m=L+1}^{\infty} \binom{m+d-1}{d-1} 2^{-tm} \\
&= \frac{1}{(d-1)!} \sum_{m=L+1}^{\infty} \frac{(m+d-1)!}{m!}\, 2^{-tm} = \frac{2^{-(L+1)t}}{(d-1)!} \sum_{n=0}^{\infty} \frac{(L+d+n)!}{(L+1+n)!}\, 2^{-tn} \\
&= \frac{2^{-(L+1)t}}{(d-1)!} \sum_{n=0}^{\infty} \frac{\Gamma(L+d+1+n) \cdot \Gamma(1+n)}{\Gamma(L+2+n)} \cdot \frac{z^n}{n!},
\end{aligned}$$

where $z = 2^{-t}$, $t > 0$. We shall rewrite the right-hand side in terms of the Gauss hypergeometric function

$$ {}_2\mathrm{F}_1(a,b;c;z) := \frac{\Gamma(c)}{\Gamma(a)\cdot\Gamma(b)} \sum_{n=0}^{\infty} \frac{\Gamma(a+n)\cdot\Gamma(b+n)}{\Gamma(c+n)} \cdot \frac{z^n}{n!}, $$

the series being convergent in the unit disk in $\mathbb{C}$. On taking $a = L+d+1$, $b = 1$, $c = L+2$, and noting that $\Gamma(1) = 1$, we have that

$$ \sum_{\ell\in\mathbb{N}^d\,:\,|\ell|_1>L} 2^{-t|\ell|_1} \quad = \quad \frac{2^{-(L+1)t}}{(d-1)!} \cdot \frac{\Gamma(L+d+1)}{\Gamma(L+2)} \cdot {}_2\mathrm{F}_1(L+d+1,1;L+2;2^{-t}). $$

We define

$$ c_{d,t,L} := \frac{1}{2^t\,(d-1)!} \cdot \frac{(L+d)!}{L^{d-1}\,(L+1)!} \cdot {}_2\mathrm{F}_1(L+d+1,1;L+2,2^{-t}). \tag{5.14} $$

Thereby,

$$ \sum_{\ell\in\mathbb{N}^d\,:\,|\ell|_1>L} 2^{-t|\ell|_1} \quad = \quad c_{d,t,L}\,2^{-tL}\,L^{d-1}. $$

We continue by exploring the asymptotic properties of $c_{d,t,L}$ in three crucial instances.

(i) *The case: $L\to\infty$, while $d\geq 2$ and $t>0$ are fixed.* Using the identity[1]

$$ {}_2\mathrm{F}_1(L+d+1,1;L+2;z) = (1-z)^{-1}\cdot {}_2\mathrm{F}_1(1-d,1,L+2;z/(z-1)), $$

we deduce that the first approximant of ${}_2\mathrm{F}_1(L+d+1,1;L+2;z)$ is $1/(1-z)$; i.e.,

$$ {}_2\mathrm{F}_1(L+d+1,1;L+2;2^{-t}) \sim \frac{1}{1-2^{-t}} \quad \text{as } L\to\infty. $$

According to Stirling's formula,

$$ n! \sim \sqrt{2\pi}\,n^{n+\frac{1}{2}}\,\mathrm{e}^{-n} \quad \text{as } n\to\infty. $$

Therefore,

$$ \frac{(L+d)!}{L^{d-1}\,(L+1)!} \sim 1 \quad \text{as } L\to\infty, $$

and hence

$$ c_{d,t,L} \sim \frac{1}{(2^t-1)\,(d-1)!} \quad \text{as } L\to\infty. $$

(ii) *The case: $d\to\infty$, while $L\geq 1$ and $t>0$ are fixed.* Using Kummer transformations[2], we deduce that

$$ {}_2\mathrm{F}_1(L+d+1,1;L+2;z) \quad = \quad \frac{\Gamma(L+2)\,\Gamma(d)}{\Gamma(L+d+1)}\,z^{-L-1}\,(1-z)^{-d} - \frac{L+1}{d\,z}\cdot {}_2\mathrm{F}_1(-L,1;d+1;1-\frac{1}{z}). $$

Therefore the first approximant of ${}_2\mathrm{F}_1(L+d+1,1;L+2;z)$ is

$$ \frac{\Gamma(L+2)\,\Gamma(d)}{\Gamma(L+d+1)}\,z^{-L-1}\,(1-z)^{-d} - \frac{L+1}{d\,z}. $$

[1]We use the minor symmetry ${}_2\mathrm{F}_1(a,b;c,z) = {}_2\mathrm{F}_1(b,a;c,z)$ of the hypergeometric function and the Kummer transformation formula ${}_2\mathrm{F}_1(a,b;c;z) = (1-z)^{-a}\,{}_2\mathrm{F}_1(a,c-b;c;z/(z-1))$ (cf. 15.3.4 in Abramowitz and Stegun [1]).

[2]We apply 15.3.3 and 15.3.9 in Abramowitz and Stegun [1], together with the identity ${}_2\mathrm{F}_1(a,0;c;z) = 1$, and use $\lim_{z\to d}\Gamma(-z)/\Gamma(1-z) = -1/d$.

That is,

$$ {}_2\mathrm{F}_1(L+d+1,1;L+2;2^{-t}) \sim \frac{\Gamma(L+2)\,\Gamma(d)}{\Gamma(L+d+1)}\,2^{t(L+1)}\,(1-2^{-t})^{-d} - \frac{L+1}{d}2^t \quad \text{as } d\to\infty. $$

This implies that

$$ c_{d,t,L} \sim \frac{1}{2^t\,(d-1)!}\,\frac{(L+d)!}{L^{d-1}\,(L+1)!}\cdot\left\{\frac{(L+1)!\,(d-1)!}{(L+d)!}\,2^{t(L+1)}\,(1-2^{-t})^{-d} - \frac{L+1}{d}2^t\right\}, $$

as $d\to\infty$, and hence

$$ c_{d,t,L} \sim \frac{1}{L^{d-1}}\left\{2^{tL}\left(\frac{2^t}{2^t-1}\right)^d - \left(\begin{array}{c} L+d \\ d \end{array}\right)\right\} \quad \text{as } d\to\infty. $$

By Stirling's formula,

$$ c_{d,t,L} \;\sim\; L\,2^{tL}\left\{\mathrm{e}^{d\left(\ln\frac{2^t}{2^t-1}-\ln L\right)} - \frac{d^L}{(L!\,2^{tL})L^d}\right\} \quad \text{as } d\to\infty. $$

Let us observe that if $L \geq \frac{2^t}{2^t-1} (> 1)$, with $t>0$, then

$$ \ln\frac{2^t}{2^t-1} - \ln L < 0 \quad \text{and} \quad \lim_{d\to\infty}\frac{d^L}{L^d} = 0. $$

Hence, $\lim_{d\to\infty} c_{d,t,L} = 0$ for any $L \geq \frac{2^t}{2^t-1}$ and $t>0$ fixed.

(iii) *The case: $L\to\infty$, $d\to\infty$, with $L-d$ bounded, while $t>1$ is fixed.* In this case one needs a uniform approximation in terms of parabolic cylinder functions (see, Olde Daalhuis [20]); the first approximant of ${}_2\mathrm{F}_1(L+d+1,1;L+2;z)$ is

$$ \frac{2^{d+L}}{\sqrt{\pi\,L}}\,\frac{\Gamma(L+2)\cdot\Gamma(d)}{\Gamma(L+d+1)}\,\frac{1}{1-2z}, \qquad |z|<\frac{1}{2}. $$

Hence,

$$ {}_2\mathrm{F}_1(L+d+1,1;L+2;2^{-t}) \;\sim\; \frac{2^{L+d}}{\sqrt{\pi\,L}}\,\frac{\Gamma(L+2)\cdot\Gamma(d)}{\Gamma(L+d+1)}\,\frac{2^t}{2^t-2}, $$

whereby

$$ c_{d,t,L} \sim \frac{1}{L^{d-1}}\cdot\frac{2^{L+d}}{\sqrt{\pi\,L}}\,\frac{1}{2^t-2} = \frac{2^{L-d}}{\sqrt{\pi}(2^t-2)}\left(\frac{4}{L^{1-\frac{1}{2d}}}\right)^d. $$

For $L\geq 7$ and $d\geq 2$ we have $4 < L^{1-\frac{1}{2d}}$; hence, $\lim_{\substack{L\to\infty,d\to\infty \\ L-d \text{ bounded}}} c_{d,t,L} = 0$ for any $t>1$ fixed. ■

The previous lemma shows that the asymptotic behaviour of $c_{d,t,L}$ is favourable, both when $d\geq 2$ is fixed and $L\gg 1$ as well as when $L\geq 1$ is fixed and $d\gg 1$. This observation then motivates the following lemma which provides a quantitative bound on

$$ \sum_{\ell\in\mathbb{N}^d\,:\,|\ell|_1>L} 2^{-t|\ell|_1} $$

reflecting these features.

Lemma 14 *The following inequalities hold for all $d\geq 2$ and all $L\geq 1$.*

(a) Suppose that $t \geq 1/(\ln 2)$; then,

$$\sum_{\ell \in \mathbb{N}^d \,:\, |\ell|_1 > L} 2^{-t|\ell|_1} \leq \frac{1}{\sqrt{2}} \left[\frac{1}{2t(\ln 2)\sqrt{\mathrm{e}\pi}} \left(\frac{2\mathrm{e}}{L} \right)^{d-\frac{1}{2}} + \left(\frac{4}{tL \ln 2} \right)^{d-1} + \frac{4^{d-1}}{(d-1)!} \right] \cdot 2^{-tL}\, L^{d-1}.$$

In particular, if $L \geq 6$ then the expression in the square bracket converges to 0 at an exponential rate as $d \to \infty$. On the other hand, for $d \geq 2$ fixed, in the limit of $L \to \infty$ the expression in the square bracket is bounded by $4^{d-1}/(d-1)!$ which, in turn, is further bounded above by $(8\pi\mathrm{e})^{-\frac{1}{2}}(4\mathrm{e}/d)^{d-\frac{1}{2}}$.

(b) Suppose that $t \geq 1$ and $0 \leq s < t+1$; then,

$$\sum_{\ell \in \mathbb{N}^d \,:\, |\ell|_1 > L} 2^{s|\ell|_\infty - (t+1)|\ell|_1} \leq c_{d,s,t,L} \cdot \begin{cases} 2^{-(t+1)L} L^{d-1} & \text{if } s = 0, \\ 2^{-tL} & \text{if } s = 1, \end{cases}$$

where,

$$c_{d,s,t,L} = \begin{cases} \frac{1}{\sqrt{2}} \left[\frac{1}{2(t+1)(\ln 2)\sqrt{\mathrm{e}\pi}} \left(\frac{2\mathrm{e}}{L} \right)^{d-\frac{1}{2}} + \left(\frac{4}{(t+1)L \ln 2} \right)^{d-1} + \frac{4^{d-1}}{(d-1)!} \right] & \text{if } s = 0, \\ d\, 2^{d-1}/(2^t - 1) & \text{if } s = 1. \end{cases} \tag{5.15}$$

Proof (a) As at the start of the proof of the previous lemma

$$\begin{aligned} \sum_{\ell \in \mathbb{N}^d \,:\, |\ell|_1 > L} 2^{-t|\ell|_1} &= \sum_{m=L+1}^{\infty} \binom{m+d-1}{d-1} 2^{-tm} \\ &\leq \frac{(d-1)^{d-\frac{1}{2}}}{(d-1)!} \sum_{m=L+1}^{\infty} \left(\frac{1}{m} + \frac{1}{d-1} \right)^{d-\frac{1}{2}} m^{d-1} 2^{-tm} \\ &\leq \frac{(d-1)^{d-\frac{1}{2}}}{(d-1)!}\, 2^{d-\frac{3}{2}} \sum_{m=L+1}^{\infty} \left[\left(\frac{1}{m} \right)^{d-\frac{1}{2}} + \left(\frac{1}{d-1} \right)^{d-\frac{1}{2}} \right] m^{d-1} 2^{-tm}. \end{aligned}$$

Here, in the transition to line 2 we made use of the bound

$$\begin{aligned} \binom{m+d-1}{d-1} &= \frac{(m+d-1)!}{m!(d-1)!} \leq \frac{(m+d-1)^{m+d-1+\frac{1}{2}} \mathrm{e}^{-(d-1)}}{m^{m+\frac{1}{2}}(d-1)!} \\ &\leq \frac{(d-1)^{d-\frac{1}{2}}}{(d-1)!} \left(\frac{1}{m} + \frac{1}{d-1} \right)^{d-\frac{1}{2}} m^{d-1}, \end{aligned}$$

which follows from the Stirling–Robbins inequality

$$\sqrt{2\pi}\, n^{n+\frac{1}{2}} \mathrm{e}^{-n+\frac{1}{12n+1}} < n! < \sqrt{2\pi}\, n^{n+\frac{1}{2}} \mathrm{e}^{-n+\frac{1}{12n}}$$

on noting that

$$\frac{1}{12(m+d-1)} - \frac{1}{12m+1} \leq 0,$$

and in the transition to line 3 we applied the bound

$$\left(\frac{1}{m} + \frac{1}{d-1} \right)^{d-\frac{1}{2}} \leq 2^{d-\frac{3}{2}} \left[\left(\frac{1}{m} \right)^{d-\frac{1}{2}} + \left(\frac{1}{d-1} \right)^{d-\frac{1}{2}} \right],$$

which follows from the elementary inequality $(a+b)^p \le 2^{p-1}(a^p+b^p)$, $a,b\ge 0$, $p\ge 1$.

Thereby,

$$\sum_{\ell\in\mathbb{N}^d\,:\,|\ell|_1>L} 2^{-t|\ell|_1} \le \mathrm{S}_1+\mathrm{S}_2,$$

where, on applying the lower bound in the Stirling–Robbins inequality to $(d-1)!$,

$$\mathrm{S}_1 = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{d-1}\,2^{d-\frac{3}{2}}\sum_{m=L+1}^{\infty}\frac{1}{\sqrt{m}}\,2^{-tm},$$

and

$$\mathrm{S}_2 = \frac{1}{(d-1)!}\,2^{d-\frac{3}{2}}\sum_{m=L+1}^{\infty} m^{d-1}2^{-tm}.$$

We start by bounding S_1. Clearly,

$$\mathrm{S}_1 = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{d-1}\,2^{d-\frac{3}{2}}\sum_{m=L+1}^{\infty}\frac{1}{\sqrt{m}}\,\mathrm{e}^{-(t\ln 2)m} = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{d-1}\,2^{d-\frac{3}{2}}\sum_{m=L+1}^{\infty} f(m),$$

where $x\mapsto f(x) := \frac{1}{\sqrt{x}}\mathrm{e}^{-(t\ln 2)x}$; the function f is positive, continuous and strictly monotonic decreasing on $\mathbb{R}_{>0}$.

Now, on performing the change of variable $(t\ln 2)x = y^2$,

$$\begin{aligned}\sum_{m=L+1}^{\infty} f(m) &\le \int_L^{\infty} f(x)\,\mathrm{d}x = \int_L^{\infty}\frac{1}{\sqrt{x}}\,\mathrm{e}^{-(t\ln 2)x}\,\mathrm{d}x = \frac{2}{\sqrt{t\ln 2}}\int_{\sqrt{(t\ln 2)L}}^{\infty}\mathrm{e}^{-y^2}\,\mathrm{d}y\\ &= \sqrt{\frac{\pi}{t\ln 2}}\,\frac{2}{\sqrt{\pi}}\int_{\sqrt{(t\ln 2)L}}^{\infty}\mathrm{e}^{-y^2}\,\mathrm{d}y = \sqrt{\frac{\pi}{t\ln 2}}\cdot\mathrm{erfc}\left(\sqrt{(t\ln 2)L}\right).\end{aligned}$$

We recall that,

$$\frac{2}{\sqrt{\pi}}\cdot\frac{\mathrm{e}^{-x^2}}{x+\sqrt{x^2+2}} < \mathrm{erfc}(x) \le \frac{2}{\sqrt{\pi}}\cdot\frac{\mathrm{e}^{-x^2}}{x+\sqrt{x^2+\frac{4}{\pi}}}, \qquad x>0.$$

Hence,

$$\mathrm{S}_1 \le \frac{(2e)^{d-1}}{\sqrt{2t(\ln 2)}}\sqrt{\frac{2}{\pi}}\,\frac{\mathrm{e}^{-(t\ln 2)L}}{\sqrt{(t\ln 2)L}+\sqrt{(t\ln 2)L+\frac{4}{\pi}}} \le \frac{(2e)^{d-1}}{2(t\ln 2)\sqrt{\pi}}\,L^{-\frac{1}{2}}\,2^{-tL}.$$

Next we bound S_2. Let us consider the function $x\mapsto g(x) := x^{d-1}\mathrm{e}^{-(t\ln 2)x}$; clearly, g is positive on $\mathbb{R}_{>0}$ with global maximum at $x_0 = \frac{d-1}{t\ln 2}$ and turning points at $x_\pm = \frac{d-1\pm\sqrt{d-1}}{t\ln 2}$. In particular, g is monotonic decreasing for $x\ge x_0$, and therefore

$$\begin{aligned}\sum_{m=L+1}^{\infty} m^{d-1}2^{-tm} &= \sum_{m=L+1}^{\infty} m^{d-1}\mathrm{e}^{-(t\ln 2)m} = \sum_{m=L+1}^{\infty} g(m)\\ &\le \int_L^{\infty} g(x)\,\mathrm{d}x = \int_L^{\infty} x^{d-1}\mathrm{e}^{-(t\ln 2)x}\,\mathrm{d}x, \qquad \text{for } L\ge\lceil x_0\rceil\ (\ge x_0);\end{aligned}$$

and similarly,

$$\begin{aligned}\sum_{m=L+1}^{\infty} m^{d-1}2^{-tm} &= \sum_{m=L+1}^{\infty} m^{d-1}\mathrm{e}^{-(t\ln 2)m} = \sum_{m=L+1}^{\infty} g(m)\\ &\le K(t,d)\int_L^{\infty} g(x)\,\mathrm{d}x = K(t,d)\int_L^{\infty} x^{d-1}\mathrm{e}^{-(t\ln 2)x}\,\mathrm{d}x, \qquad \text{for } L\le\lceil x_0\rceil-1,\end{aligned}$$

with

$$K(t,d) := \frac{2}{\min(\lceil x_0\rceil, x_+) - \max(\lceil x_0\rceil - 1, x_-)}, \qquad x_0 = \frac{d-1}{t\ln 2},$$

where $x \mapsto \lceil x\rceil$ denotes the ceiling function, — the smallest integer $\geq x$. We note that since $K(t,d) \geq 2$, and since $\max(\lceil x_0\rceil - 1, x_-) < x_0 \leq \min(\lceil x_0\rceil, x_+)$, the real number $K(t,d)$ is finite; in fact, plotting $d \mapsto K(t,d)$ for $t \geq 1/\ln 2$ reveals that $0 < K(t,d)/(t\ln 2) \leq 2$ for all such t, with $K(1/\ln 2, d)/(t\ln 2) = 2$ for all $d \geq 2$.

The inequality for $L \leq \lceil x_0\rceil - 1$ above follows by first observing that

$$\begin{aligned} g(\lceil x_0\rceil) &\leq g(\lceil x_0\rceil) + g(\lceil x_0\rceil - 1) \leq g(\min(\lceil x_0\rceil, x_+)) + g(\max(\lceil x_0\rceil - 1, x_-)) \\ &= \frac{2}{\min(\lceil x_0\rceil, x_+) - \max(\lceil x_0\rceil - 1, x_-)} \\ &\qquad \times \frac{\min(\lceil x_0\rceil, x_+) - \max(\lceil x_0\rceil - 1, x_-)}{2} \left(g(\min(\lceil x_0\rceil, x_+)) + g(\max(\lceil x_0\rceil - 1, x_-))\right) \\ &\leq K(t,d) \int_{\max(\lceil x_0\rceil - 1, x_-)}^{\min(\lceil x_0\rceil, x_+)} g(x)\,\mathrm{d}x \leq K(t,d) \int_{\lceil x_0\rceil - 1}^{\lceil x_0\rceil} g(x)\,\mathrm{d}x. \end{aligned}$$

Therefore, if $L = \lceil x_0\rceil - 1$, then

$$\sum_{m=L+1}^{\infty} g(m) = g(\lceil x_0\rceil) + \sum_{m=\lceil x_0\rceil + 1}^{\infty} g(m) \leq K(t,d) \int_L^{L+1} g(x)\,\mathrm{d}x + \int_{L+1}^{\infty} g(x)\,\mathrm{d}x \leq K(t,d) \int_L^{\infty} g(x)\,\mathrm{d}x.$$

If $L = \lceil x_0\rceil - 2$, then

$$\begin{aligned} \sum_{m=L+1}^{\infty} g(m) &= g(\lceil x_0\rceil - 1) + g(\lceil x_0\rceil) + \sum_{m=\lceil x_0\rceil + 1}^{\infty} g(m) \leq K(t,d) \int_{L+1}^{L+2} g(x)\,\mathrm{d}x + \int_{L+2}^{\infty} g(x)\,\mathrm{d}x \\ &\leq K(t,d) \int_{L+1}^{\infty} g(x)\,\mathrm{d}x \leq K(t,d) \int_L^{\infty} g(x)\,\mathrm{d}x. \end{aligned}$$

Finally, if $L \leq \lceil x_0\rceil - 3$, then

$$\begin{aligned} \sum_{m=L+1}^{\infty} g(m) &= \sum_{m=L+1}^{\lceil x_0\rceil - 2} g(m) + g(\lceil x_0\rceil - 1) + g(\lceil x_0\rceil) + \sum_{m=\lceil x_0\rceil + 1}^{\infty} g(m) \\ &\leq \int_L^{\lceil x_0\rceil - 1} g(x)\,\mathrm{d}x + K(t,d) \int_{\lceil x_0\rceil - 1}^{\lceil x_0\rceil} g(x)\,\mathrm{d}x + \int_{\lceil x_0\rceil}^{\infty} g(x)\,\mathrm{d}x \\ &\leq K(t,d) \int_L^{\infty} g(x)\,\mathrm{d}x. \end{aligned}$$

Thus, in any case,

$$\sum_{m=L+1}^{\infty} m^{d-1} 2^{-tm} \leq K(t,d) \int_L^{\infty} x^{d-1} \mathrm{e}^{-(t\ln 2)x}\,\mathrm{d}x, \qquad \text{for all } L \geq 1.$$

Let us perform the change of variable $x = y + L$. Then,

$$\int_L^{\infty} x^{d-1} \mathrm{e}^{-(t\ln 2)x}\,\mathrm{d}x = \mathrm{e}^{-tL(\ln 2)} \int_0^{\infty} (y+L)^{d-1} \mathrm{e}^{-(t\ln 2)y}\,\mathrm{d}y.$$

Now, $(y+L)^{d-1} \le 2^{d-2}(y^{d-1}+L^{d-1})$, and therefore,

$$\begin{aligned}\int_L^\infty x^{d-1}\mathrm{e}^{-(t\ln 2)x}\,\mathrm{d}x &\le 2^{-tL}2^{d-2}\left(\int_0^\infty y^{d-1}\mathrm{e}^{-(t\ln 2)y}\,\mathrm{d}y + L^{d-1}\int_0^\infty \mathrm{e}^{-(t\ln 2)y}\,\mathrm{d}y\right)\\ &= 2^{-tL}2^{d-2}\left(\frac{1}{(t\ln 2)^d}\int_0^\infty z^{d-1}\mathrm{e}^{-z}\,\mathrm{d}z + \frac{L^{d-1}}{t\ln 2}\int_0^\infty \mathrm{e}^{-z}\,\mathrm{d}z\right)\\ &= 2^{-tL}2^{d-2}\left(\frac{(d-1)!}{(t\ln 2)^d} + \frac{L^{d-1}}{t\ln 2}\right).\end{aligned}$$

Hence,

$$\sum_{m=L+1}^\infty m^{d-1}2^{-tm} \le 2^{-tL}2^{d-2}\frac{K(t,d)}{t\ln 2}\left(\frac{(d-1)!}{(t\ln 2)^{d-1}} + L^{d-1}\right).$$

Substituting this bound into the definition of S_2 and noting that $\frac{K(t,d)}{t\ln 2} \le 2$ yields

$$\begin{aligned}\mathrm{S}_2 &\le \frac{1}{(d-1)!}\,2^{d-\frac{3}{2}}\,2^{-tL}\,2^{d-1}\left(\frac{(d-1)!}{(t\ln 2)^{d-1}} + L^{d-1}\right)\\ &= 2^{-tL}L^{d-1}\frac{1}{\sqrt{2}}\left(\left(\frac{4}{tL\ln 2}\right)^{d-1} + \frac{4^{d-1}}{(d-1)!}\right).\end{aligned}$$

Combining the bounds on S_1 and S_2 we deduce that

$$\sum_{\ell\in\mathbb{N}^d\,:\,|\ell|_1>L} 2^{-t|\ell|_1} \le C_0\,2^{-tL}\,L^{d-1},$$

where

$$C_0 = \frac{1}{\sqrt{2}}\left[\frac{1}{2t(\ln 2)\sqrt{\mathrm{e}\pi}}\left(\frac{2\mathrm{e}}{L}\right)^{d-\frac{1}{2}} + \left(\frac{4}{tL\ln 2}\right)^{d-1} + \frac{4^{d-1}}{(d-1)!}\right].$$

In particular, if $L \ge 6$ then the expression in the square bracket converges to 0 at an exponential rate as $d\to\infty$. On the other hand, for $d \ge 2$ fixed, in the limit of $L\to\infty$ the expression in the square bracket is bounded by $4^{d-1}/(d-1)!$ which, in turn, is further bounded above by $(8\pi\mathrm{e})^{-\frac{1}{2}}(4\mathrm{e}/d)^{d-\frac{1}{2}}$ on applying to $(d-1)!$ the lower bound in the Stirling–Robbins inequality.

(b) For $s=0$, the result follows from (a) with t replaced by $t+1$. Since we assume that $t\ge 1$, trivially, $t+1 \ge 2 > 1/(\ln 2)$, we can apply part (a). Let us therefore suppose that $0<s<t+1$. Noting that for $\ell=(\ell_1,\ldots,\ell_d)\in\mathbb{N}^d$, such that $|\ell|_1=m$,

$$2^{s|\ell|_\infty-(t+1)|\ell|_1} = 2^{(s-(t+1))L+(s-(t+1))(m-L)+s(|\ell|_\infty-m)},$$

we have that

$$\begin{aligned}\sum_{\ell\in\mathbb{N}^d\,:\,|\ell|_1>L} 2^{s|\ell|_\infty-(t+1)|\ell|_1} &= \sum_{m=L+1}^\infty \sum_{\ell\in\mathbb{N}^d\,:\,|\ell|_1=m} 2^{s|\ell|_\infty-(t+1)|\ell|_1}\\ &= 2^{(s-(t+1))L}\left(\sum_{m=L+1}^\infty 2^{(s-(t+1))(m-L)}\sigma_m\right),\end{aligned}$$

where

$$\sigma_m = \sum_{\ell\in\mathbb{N}^d\,:\,|\ell|_1=m} 2^{s(|\ell|_\infty-m)}.$$

For $s>0$, $\sigma_m \le d\left(1+\frac{1}{2^s-1}\right)^{d-1}$, independent of m, by Lemma 10. The final form of the inequality under (b) for the case of $s>0$ follows on observing that $\sum_{m=L+1}^\infty 2^{(s-(t+1))(m-L)}\sigma_m$ is bounded by $d\left(1+\frac{1}{2^s-1}\right)^{d-1}/(2^{t+1-s}-1)$, independent of L. ∎

5.2 Tensorization of seminorms

Next we develop some auxiliary results concerning tensorization of seminorms. These results will then be used in the derivation of the approximation property of the d-dimensional sparse tensor-product built using the univariate finite element space scale $\{\mathcal{V}_{(0)}^{\ell,p}\}_{\ell\geq 0}$.

Let $(\mathrm{H},\langle\cdot,\cdot\rangle_{\mathrm{H}})$ and $(\mathrm{K},\langle\cdot,\cdot\rangle_{\mathrm{K}})$ be two Hilbert spaces and $T\in\mathcal{B}(\mathrm{H},\mathrm{K})$ a bounded linear operator. Clearly,

$$|u|_T := \|Tu\|_{\mathrm{K}},\quad u\in\mathrm{H}, \tag{5.16}$$

defines a seminorm on H.

Considering now four Hilbert spaces $(\mathrm{H}_i,\langle\cdot,\cdot\rangle_{\mathrm{H}_i})$, $(\mathrm{K}_i,\langle\cdot,\cdot\rangle_{\mathrm{K}_i}), i=1,2$, as well as two bounded linear operators $T_i\in\mathcal{B}(\mathrm{H}_i,\mathrm{K}_i)$, $i=1,2$, it is natural to define via

$$|u|_{T_1\otimes T_2} := \|(T_1\otimes T_2)u\|_{\mathrm{K}_1\otimes\mathrm{K}_2},\quad u\in\mathrm{H}_1\otimes\mathrm{H}_2,$$

a seminorm on the tensor-product $\mathrm{H}_1\otimes\mathrm{H}_2$ of the spaces H_1 and H_2.

Next we define the bounded linear operators with respect to seminorms of the type (5.16) and investigate their tensor-products.

Definition 2 *Let $(\mathrm{H},\langle\cdot,\cdot\rangle_{\mathrm{H}})$, $(\mathrm{K},\langle\cdot,\cdot\rangle_{\mathrm{K}})$, $(\tilde{\mathrm{H}},\langle\cdot,\cdot\rangle_{\tilde{\mathrm{H}}})$, $(\tilde{\mathrm{K}},\langle\cdot,\cdot\rangle_{\tilde{\mathrm{K}}})$ be four Hilbert spaces and consider the bounded linear operators $T\in\mathcal{B}(\mathrm{H},\mathrm{K}),\tilde{T}\in\mathcal{B}(\tilde{\mathrm{H}},\tilde{\mathrm{K}})$ and $Q\in\mathcal{B}(\mathrm{H},\tilde{\mathrm{H}})$. We say that Q is $(T,\tilde{T})$-bounded if there exists $c\geq 0$ such that*

$$|Qu|_{\tilde{T}}\leq c|u|_T\quad\forall u\in\mathrm{H}. \tag{5.17}$$

We further denote by $|Q|_{T,\tilde{T}}$ the infimum over all constants $c\geq 0$ satisfying (5.17).

Example 2 We give some examples based on the bounds in Section 3.2.

(a) We use the terminology from Definition 2, with $\mathrm{H}:=\mathrm{H}^{t+1}(0,1)\cap\mathrm{H}^1_{(0)}(0,1)$, and let $\tilde{\mathrm{H}}:=\mathrm{H}^s(0,1)$, $\mathrm{K}=\tilde{\mathrm{K}}:=\mathrm{L}^2(0,1)$, $T:=\partial^{t+1}$, $\tilde{T}:=\partial^s$, with $t\geq 1$ and $s\in\{0,1\}$. The approximation property (3.6) shows that the linear operator $\mathrm{Id}_{\mathrm{H}}-P_{(0)}^{\ell,p}$ is $(\partial^{t+1},\partial^s)$-bounded. Thus, by (3.10), the projector $Q_{(0)}^{\ell,p}$ is also $(\partial^{t+1},\partial^s)$-bounded for all $\ell\geq 1$ and $p\geq 1$ (with $\partial^0:=\mathrm{Id}_{\mathrm{L}^2(0,1)}$).

(b) Trivially, on taking $\mathrm{H}=\tilde{\mathrm{H}}:=\mathrm{H}^1_{(0)}(0,1)$ and $\mathrm{K}=\tilde{\mathrm{K}}:=\mathrm{L}^2(0,1)$, the projector $Q_{(0)}^{\ell,p}$ is (∂^1,∂^1)-bounded for all $\ell\geq 0$ and $p\geq 1$.

(c) Finally, on taking $\mathrm{H}=\mathrm{K}:=\mathrm{H}^1_{(0)}(0,1)$ (equipped with the norm $\|\cdot\|_{\mathrm{H}^1_*(0,1)}$) and $\tilde{\mathrm{H}}=\tilde{\mathrm{K}}:=\mathrm{L}^2(0,1)$, we see that $Q_{(0)}^{0,p}$ is $(\mathrm{Id}_{\mathrm{H}^1_{(0)}(0,1)},\mathrm{Id}_{\mathrm{L}^2(0,1)})$-bounded for all $p\geq 1$.

In particular, on taking $\mathrm{H}:=\mathrm{H}^1_0(0,1)$, $\tilde{\mathrm{H}}:=\mathrm{L}^2(0,1)$ and $\mathrm{K}=\tilde{\mathrm{K}}:=\mathrm{L}^2(0,1)$ we see that $Q_0^{0,p}$ is $(\partial^1,\mathrm{Id}_{\mathrm{L}^2(0,1)})$-bounded for all $p\geq 1$. ⋄

Proposition 15 *Let $(\mathrm{H}_i,\langle\cdot,\cdot\rangle_{\mathrm{H}_i})$, $(\mathrm{K}_i,\langle\cdot,\cdot\rangle_{\mathrm{K}_i})$, $(\tilde{\mathrm{H}}_i,\langle\cdot,\cdot\rangle_{\tilde{\mathrm{H}}_i})$, $(\tilde{\mathrm{K}}_i,\langle\cdot,\cdot\rangle_{\tilde{\mathrm{K}}_i})$ for $i=1,2$ be separable Hilbert spaces. Let $T_i\in\mathcal{B}(\mathrm{H}_i,\mathrm{K}_i),\tilde{T}_i\in\mathcal{B}(\tilde{\mathrm{H}}_i,\tilde{\mathrm{K}}_i)$ and $Q_i\in\mathcal{B}(\mathrm{H}_i,\tilde{\mathrm{H}}_i)$ be bounded linear operators, and assume that Q_i is $(T_i,\tilde{T}_i)$-bounded for $i=1,2$. Then $Q_1\otimes Q_2$ is $(T_1\otimes T_2,\tilde{T}_1\otimes\tilde{T}_2)$-bounded, and*

$$|Q_1\otimes Q_2|_{T_1\otimes T_2,\tilde{T}_1\otimes\tilde{T}_2}\leq|Q_1|_{T_1,\tilde{T}_1}|Q_2|_{T_2,\tilde{T}_2}.$$

In other words, if $\|\tilde{T}_i Q_i v_i\|_{\tilde{\mathrm{K}}_i} \le c_i \|T_i v_i\|_{\mathrm{K}_i}$ *for all* $v_i \in \mathrm{H}_i$, $i = 1, 2$, *then*

$$\|(\tilde{T}_1 \otimes \tilde{T}_2)(Q_1 \otimes Q_2)u\|_{\tilde{\mathrm{K}}_1 \otimes \tilde{\mathrm{K}}_2} \le c_1 c_2 \|(T_1 \otimes T_2)u\|_{\mathrm{K}_1 \otimes \mathrm{K}_2} \qquad \forall u \in \mathrm{H}_1 \otimes \mathrm{H}_2.$$

Proof For any $u \in \mathrm{H}_1 \otimes \mathrm{H}_2$ we have

$$\begin{aligned} |(Q_1 \otimes Q_2)u|_{\tilde{T}_1 \otimes \tilde{T}_2} &= \|(\tilde{T}_1 \otimes \tilde{T}_2)(Q_1 \otimes Q_2)u\|_{\tilde{\mathrm{K}}_1 \otimes \tilde{\mathrm{K}}_2} \\ &= \|(\tilde{T}_1 Q_1 \otimes \mathrm{Id}_{\tilde{\mathrm{K}}_2})(\mathrm{Id}_{\mathrm{H}_1} \otimes \tilde{T}_2 Q_2)u\|_{\tilde{\mathrm{K}}_1 \otimes \tilde{\mathrm{K}}_2}. \end{aligned} \tag{5.18}$$

Denoting $v := (\mathrm{Id}_{\mathrm{H}_1} \otimes \tilde{T}_2 Q_2)u \in \mathrm{H}_1 \otimes \tilde{\mathrm{K}}_2$ and considering an orthonormal basis $(e_i)_{i \in I}$ in $\tilde{\mathrm{K}}_2$, where $I \subset \mathbb{N}$ is a countable index set, we expand $v = \sum_{i \in I} v_i \otimes e_i$, so that

$$\begin{aligned} \|(\tilde{T}_1 Q_1 \otimes \mathrm{Id}_{\tilde{\mathrm{K}}_2})v\|^2_{\tilde{\mathrm{K}}_1 \otimes \tilde{\mathrm{K}}_2} &= \sum_{i \in I} \|\tilde{T}_1 Q_1 v_i\|^2_{\tilde{\mathrm{K}}_1} \\ &\overset{(5.17)}{\le} c_1^2 \sum_{i \in I} \|T_1 v_i\|^2_{\mathrm{K}_1} \\ &= c_1^2 \|(T_1 \otimes \mathrm{Id}_{\tilde{\mathrm{K}}_2})v\|^2_{\mathrm{K}_1 \otimes \tilde{\mathrm{K}}_2}, \end{aligned} \tag{5.19}$$

where $c_1 = |Q_1|_{T_1, \tilde{T}_1}$. We now note that

$$(T_1 \otimes \mathrm{Id}_{\tilde{\mathrm{K}}_2})v = (T_1 \otimes \mathrm{Id}_{\tilde{\mathrm{K}}_2})(\mathrm{Id}_{\mathrm{H}_1} \otimes \tilde{T}_2 Q_2)u = (\mathrm{Id}_{\mathrm{K}_1} \otimes \tilde{T}_2 Q_2)(T_1 \otimes \mathrm{Id}_{\mathrm{H}_2})u,$$

so that defining $w := (T_1 \otimes \mathrm{Id}_{\mathrm{H}_2})u \in \mathrm{K}_1 \otimes \mathrm{H}_2$ and arguing as in (5.19) to estimate the norm of $(\mathrm{Id}_{\mathrm{K}_1} \otimes \tilde{T}_2 Q_2)w$, we obtain

$$\|(\mathrm{Id}_{\mathrm{K}_1} \otimes \tilde{T}_2 Q_2)w\|_{\mathrm{K}_1 \otimes \tilde{\mathrm{K}}_2} \le c_2 \|(\mathrm{Id}_{\mathrm{K}_1} \otimes T_2)w\|_{\mathrm{K}_1 \otimes \mathrm{K}_2} = c_2 \|(T_1 \otimes T_2)u\|_{\mathrm{K}_1 \otimes \mathrm{K}_2}, \tag{5.20}$$

where $c_2 = |Q_2|_{T_2, \tilde{T}_2}$. From (5.18), (5.19), (5.20) we obtain

$$|(Q_1 \otimes Q_2)u|_{\tilde{T}_1 \otimes \tilde{T}_2} \le c_1 c_2 \|(T_1 \otimes T_2)u\|_{\mathrm{K}_1 \otimes \mathrm{K}_2} = c_1 c_2 |u|_{T_1 \otimes T_2},$$

and the desired result follows by recalling the definitions of the constants c_1, c_2. ∎

5.3 Approximation from sparse tensor-product spaces

We are now ready to embark on the study of the approximation properties of the sparse tensor-product spaces. In order to track the dependence of the constants in the error bound on the polynomial degree p, the Sobolev regularity t and the dimension d, we consider

$$\Omega := (0, 1)^d.$$

This domain has, for any d, Lebesgue measure 1.

To characterize the regularity of the function u to be approximated, we introduce, for $I \subset \{1, 2, \ldots, d\}$ with $|I| = k \ge 1$, $I = \{i_1, i_2, \ldots, i_k\}$, the notation $\mathrm{H}^{\alpha,\beta,I}(\Omega)$ for the tensor-product space consisting of d factors, each of them being either $\mathrm{H}^{\alpha}_{(0)}(0, 1)$ (in the j-th co-ordinate, if $j \in I$), or $\mathrm{H}^{\beta}_{(0)}(0, 1)$ (in the j-th co-ordinate, if $j \notin I$).

Given $I = \{i_1, i_2, \ldots, i_k\} \subset \{1, 2, \ldots, d\}$, let $I^c = \{j_1, j_2, \ldots, j_{d-k}\}$ denote the (possibly empty) complement of I with respect to $\{1, 2, \ldots, d\}$; for non-negative integers α and β we then denote by $|u|_{\mathrm{H}^{\alpha,\beta,I}(\Omega)}$ the seminorm

$$\sum_{(\alpha)_1 \le \alpha_1 \le \alpha} \cdots \sum_{(\alpha)_k \le \alpha_k \le \alpha} \sum_{(\beta)_1 \le \beta_1 \le \beta} \cdots \sum_{(\beta)_{d-k} \le \beta_{d-k} \le \beta} \left\| \left(\frac{\partial^{\alpha_1}}{\partial x_{i_1}^{\alpha_1}} \cdots \frac{\partial^{\alpha_k}}{\partial x_{i_k}^{\alpha_k}} \right) \left(\frac{\partial^{\beta_1}}{\partial x_{j_1}^{\beta_1}} \cdots \frac{\partial^{\beta_{d-k}}}{\partial x_{j_{d-k}}^{\beta_{d-k}}} \right) u \right\|_{\mathrm{L}^2(\Omega)},$$

where, for $i = 1, \ldots, k$,

$$(\alpha)_i = \begin{cases} \alpha & \text{if } Ox_i \text{ is an elliptic co-ordinate direction,} \\ 0 & \text{if } Ox_i \text{ is a hyperbolic co-ordinate direction,} \end{cases}$$

with analogous definition of $(\beta)_j$, $j = 1, \ldots, d-k$.

Theorem 16 *Let $\Omega = (0,1)^d$, $s \in \{0,1\}$, $k \geq 1$, and let a polynomial degree $p \geq 1$ be given. Let, further, $\mathrm{C}^\infty_{(0)}(\bar\Omega)$ denote the set of all functions in $\mathrm{C}^\infty(\bar\Omega)$ that vanish on Γ_0. Then, for $1 \leq t \leq p$, there exist constants $\underline{c}_{p,t}$, $\kappa(p,t,s,L) > 0$, independent of d, and κ monotonic decreasing in $L \geq 1$, such that, for any $u \in \mathrm{C}^\infty_{(0)}(\bar\Omega)$ and for any $L \geq 1$ and any $d \geq 1$, we have*

$$|u - \hat P^{L,p}_{(0)} u|_{\mathrm{H}^s(\Omega)} \leq d^{1+\frac{s}{2}} \underline{c}_{p,t} (\kappa(p,t,s,L))^{d-1+s} L^{\nu(s)} 2^{-(t+1-s)L} \cdot \max_{1\leq k\leq d} \left(\max_{\substack{I\subseteq\{1,2,\ldots,d\}\\ |I|=k}} |u|_{\mathrm{H}^{t+1,s,I}(\Omega)} \right) \tag{5.21}$$

where, for $s = 0$, the seminorm $|\cdot|_{\mathrm{H}^s(\Omega)}$ is understood to coincide with the $\mathrm{L}^2(\Omega)$-norm and $\nu(0) = d-1$, while for $s = 1$ the seminorm $|\cdot|_{\mathrm{H}^s(\Omega)}$ is the $\mathrm{H}^1(\Omega)$-seminorm and $\nu(1) = 0$.

Proof For $u \in \mathrm{C}^\infty_{(0)}(\bar\Omega) \subset \mathrm{H}^1_{(0)}(\Omega)$, the following identity holds in $\mathrm{H}^1(\Omega)$:

$$u = \sum_{\ell\in\mathbb{N}^d} \left(Q^{\ell_1,p}_{(0)} \otimes \cdots \otimes Q^{\ell_d,p}_{(0)} \right) u.$$

We estimate, for $s \in \{0,1\}$, the approximation error as a sum of details, i.e.

$$\left| u - \hat P^{L,p}_{(0)} u \right|_{\mathrm{H}^s(\Omega)} \leq \sum_{|\ell|_1 > L} \left| \left(Q^{\ell_1,p}_{(0)} \otimes \cdots \otimes Q^{\ell_d,p}_{(0)} \right) u \right|_{\mathrm{H}^s(\Omega)} \tag{5.22}$$

provided that the right-hand side is finite. We discuss the two cases, $s = 0$ and $s = 1$, separately.

For $s = 1$ and any $\ell = (\ell_1, \ell_2, \ldots, \ell_d) \in \mathbb{N}^d$ with $\mathrm{supp}(\ell) = I$ (that is, $\ell_j \neq 0$ iff $j \in I$) and $|I| = k$, we have to estimate the solution details

$$\left| \left(Q^{\ell_1,p}_{(0)} \otimes \cdots \otimes Q^{\ell_d,p}_{(0)} \right) u \right|^2_{\mathrm{H}^1(\Omega)} = \sum_{j=1}^d \left| \left(Q^{\ell_1,p}_{(0)} \otimes \cdots \otimes Q^{\ell_d,p}_{(0)} \right) u \right|^2_{\mathrm{H}^{1,0,\{j\}}(\Omega)} =: (\star)$$

for $\ell \in \mathbb{N}^d$.

Using Proposition 15 and the notation ∂ for the differentiation operator in dimension 1, we obtain the following chain of inequalities:

$$\begin{aligned}
(\star) \quad &\leq \quad \sum_{j\in I} \prod_{\substack{j'\in I\\ j'\neq j}} |Q^{\ell_{j'},p}_{(0)}|^2_{(\partial^{t+1},\mathrm{Id}_{\mathrm{L}^2(0,1)})} \cdot |Q^{\ell_j,p}_{(0)}|^2_{(\partial^{t+1},\partial^1)} |Q^{0,p}_{(0)}|^{2(d-k)}_{(\mathrm{Id}_{\mathrm{H}^1_{(0)}(0,1)},\mathrm{Id}_{\mathrm{L}^2(0,1)})} |u|^2_{\mathrm{H}^{t+1,1,I}(\Omega)} \\
&\quad + \sum_{j\notin I} \prod_{j'\in I} |Q^{\ell_{j'},p}_{(0)}|^2_{(\partial^{t+1},\mathrm{Id}_{\mathrm{L}^2(0,1)})} \cdot |Q^{0,p}_{(0)}|^2_{(\partial^1,\partial^1)} |Q^{0,p}_{(0)}|^{2(d-k-1)}_{(\mathrm{Id}_{\mathrm{H}^1_{(0)}(0,1)},\mathrm{Id}_{\mathrm{L}^2(0,1)})} |u|^2_{\mathrm{H}^{t+1,1,I}(\Omega)} \\
&\leq \quad \sum_{j\in I} \tilde c^{2(k-1)}_{p,0,t} \tilde c^2_{p,1,t} 4^{-(t+1)|\ell|_1+\ell_j} \hat c^{2(d-k)}_{p,0,(0)} |u|^2_{\mathrm{H}^{t+1,1,I}(\Omega)} \\
&\quad + \sum_{j\notin I} \tilde c^{2k}_{p,0,t} 4^{-(t+1)|\ell|_1} \hat c^2_{p,1,(0)} \hat c^{2(d-k-1)}_{p,0,(0)} |u|^2_{\mathrm{H}^{t+1,1,I}(\Omega)}. \\
&\leq \quad \tilde c^{2(k-1)}_{p,0,t} 4^{-(t+1)|\ell|_1} \hat c^{2(d-k-1)}_{p,0,(0)} |u|^2_{\mathrm{H}^{t+1,1,I}(\Omega)} \left(\tilde c^2_{p,1,t} \hat c^2_{p,0,(0)} \sum_{j\in I} 4^{\ell_j} + (d-k) \tilde c^2_{p,0,t} \hat c^2_{p,1,(0)} \right) \\
&\leq \quad d \bar c_{p,t} \tilde c^{2(k-1)}_{p,0,t} 4^{|\ell|_\infty - (t+1)|\ell|_1} \hat c^{2(d-k-1)}_{p,0,(0)} |u|^2_{\mathrm{H}^{t+1,1,I}(\Omega)},
\end{aligned} \tag{5.23}$$

where

$$\bar{c}_{p,t} := \max\left(\tilde{c}^2_{p,1,t}\hat{c}^2_{p,0,(0)},\ \tilde{c}^2_{p,0,t}\hat{c}^2_{p,1,(0)}\right) \tag{5.24}$$

with $\tilde{c}_{p,s,t}$ defined in (3.7), (3.11) and $\hat{c}_{p,s,(0)}$ defined in (3.15).

We note in passing that in the (important) special case when $\Gamma = \Gamma_0$, and thereby $\mathrm{H}^1_{(0)}(0,1) = \mathrm{H}^1_0(0,1)$ in each of the d co-ordinate directions, the factor $|Q^{0,p}_{(0)}|_{(\mathrm{Id}_{\mathrm{H}^1_{(0)}(0,1)},\mathrm{Id}_{\mathrm{L}^2(0,1)})}$ in the first two lines of (5.23) above can be replaced by $|Q^{0,p}_{(0)}|_{(\partial^1,\mathrm{Id}_{\mathrm{L}^2(0,1)})}$.

We thus have,

$$\begin{aligned}
&\sum_{\substack{\ell\in\mathbb{N}^d,\,|\ell|_1>L\\ \mathrm{supp}(\ell)=I}} \left|\left(Q^{\ell_1,p}_{(0)}\otimes\cdots\otimes Q^{\ell_d,p}_{(0)}\right)u\right|_{\mathrm{H}^1(\Omega)} \\
&\qquad\le \sqrt{d\bar{c}_{p,t}}\tilde{c}^{k-1}_{p,0,t}\hat{c}^{d-k-1}_{p,0,(0)} \sum_{\substack{\ell\in\mathbb{N}^d,\,|\ell|_1>L\\ \mathrm{supp}(\ell)=I}} 2^{|\ell|_\infty-(t+1)|\ell|_1}|u|_{\mathrm{H}^{t+1,1,I}(\Omega)} \\
&\qquad\le \sqrt{d\bar{c}_{p,t}}\tilde{c}^{k-1}_{p,0,t}\hat{c}^{d-k-1}_{p,0,(0)} \sum_{\ell\in\mathbb{N}^k,\,|\ell|_1>L} 2^{|\ell|_\infty-(t+1)|\ell|_1}|u|_{\mathrm{H}^{t+1,1,I}(\Omega)}\,.
\end{aligned}$$

In passing from the second to the third line in the estimate above we have dropped all $d-k$ trivial entries from the indexing of ℓ.

We now use, with arbitrary $l > L$, the estimate $\sum_{\ell\in\mathbb{N}^k,\,|\ell|_1=l} 2^{|\ell|_\infty} \le k2^{k-1+l}$ and obtain

$$\begin{aligned}
&\sum_{\substack{\ell\in\mathbb{N}^d,|\ell|_1>L\\ \mathrm{supp}(\ell)=I}} \left|\left(Q^{\ell_1,p}_{(0)}\otimes\cdots\otimes Q^{\ell_d,p}_{(0)}\right)u\right|_{\mathrm{H}^1(\Omega)} \\
&\qquad\le k\sqrt{d\bar{c}_{p,t}}\tilde{c}^{k-1}_{p,0,t}\hat{c}^{d-k-1}_{p,0,(0)}2^{k-1}\left(\sum_{l>L}2^{-tl}\right)|u|_{\mathrm{H}^{t+1,1,I}(\Omega)} \\
&\qquad= k\sqrt{d\bar{c}_{p,t}}(1-2^{-t})^{-1}\tilde{c}^{k-1}_{p,0,t}\hat{c}^{d-k-1}_{p,0,(0)}2^{k-1}2^{-t(L+1)}|u|_{\mathrm{H}^{t+1,1,I}(\Omega)} \\
&\qquad= d^{\frac12}\underline{c}_{p,t}k(2\tilde{c}_{p,0,t})^k\hat{c}^{d-k}_{p,0,(0)}2^{-tL}|u|_{\mathrm{H}^{t+1,1,I}(\Omega)},
\end{aligned} \tag{5.25}$$

where

$$\underline{c}_{p,t} := \frac12\sqrt{\bar{c}_{p,t}}((2^t-1)\tilde{c}_{p,0,t}\hat{c}_{p,0,(0)})^{-1}. \tag{5.26}$$

Now, summing (5.25) over $I\subseteq\{1,2,\dots,d\}$ we deduce that

$$\begin{aligned}
&\sum_{k=1}^d \sum_{\substack{I\subseteq\{1,2,\dots,d\}\\ |I|=k}} \sum_{\substack{\ell\in\mathbb{N}^d,|\ell|_1>L\\ \mathrm{supp}(\ell)=I}} \left|\left(Q^{\ell_1,p}_{(0)}\otimes\cdots\otimes Q^{\ell_d,p}_{(0)}\right)u\right|_{\mathrm{H}^1(\Omega)} \\
&\qquad\le d^{\frac12}\underline{c}_{p,t}2^{-tL}\sum_{k=1}^d\binom{d}{k}(2\tilde{c}_{p,0,t})^k\hat{c}^{d-k}_{p,0,(0)}\cdot k\max_{\substack{I\subseteq\{1,2,\dots,d\}\\ |I|=k}}|u|_{\mathrm{H}^{t+1,1,I}(\Omega)} \\
&\qquad\le d^{\frac12}\underline{c}_{p,t}(\kappa(p,t,1,L))^d2^{-tL}\cdot\max_{1\le k\le d}\left(k\max_{\substack{I\subseteq\{1,2,\dots,d\}\\ |I|=k}}|u|_{\mathrm{H}^{t+1,1,I}(\Omega)}\right),
\end{aligned} \tag{5.27}$$

where

$$\kappa(p,t,1,L) := 2\tilde{c}_{p,0,t}+\hat{c}_{p,0,(0)}, \qquad p\ge1,\quad 1\le t\le p,\quad L\ge1. \tag{5.28}$$

This completes the proof in the case of $s = 1$.

For $s = 0$, we write the bound (5.22) as a sum of details as follows:

$$\begin{aligned}\|u - \hat{P}^{L,p}_{(0)}u\|_{\mathrm{L}^2(\Omega)} &\leq \sum_{|\ell|_1 > L} \left\| \left(Q^{\ell_1,p}_{(0)} \otimes \cdots \otimes Q^{\ell_d,p}_{(0)} \right) u \right\|_{\mathrm{L}^2(\Omega)} \\ &= \sum_{k=1}^{d} \sum_{\substack{I \subset \{1,2,\ldots,d\} \\ |I|=k}} \sum_{\substack{\ell \in \mathbb{N}^d \\ \mathrm{supp}(\ell) = I}} \left\| \left(Q^{\ell_1,p}_{(0)} \otimes \cdots \otimes Q^{\ell_d,p}_{(0)} \right) u \right\|_{\mathrm{L}^2(\Omega)}\end{aligned}$$

We estimate the size of the detail with multi-index $\ell \in \mathbb{N}^d$ in the above sum, i.e.

$$(\star) = \left\| \left(Q^{\ell_1,p}_{(0)} \otimes \cdots \otimes Q^{\ell_d,p}_{(0)} \right) u \right\|^2_{\mathrm{L}^2(\Omega)}.$$

Using $I = \mathrm{supp}(\ell)$ and that $|I| = k$, we get

$$\begin{aligned}(\star) &\leq \left\{ \prod_{j \in I} |Q^{\ell_j,p}_{(0)}|^2_{(\partial^{t+1}, \mathrm{Id}_{\mathrm{L}^2(0,1)})} \right\} |Q^{0,p}_{(0)}|^{2(d-k)}_{(\mathrm{Id}_{\mathrm{H}^1_{(0)}(0,1)}, \mathrm{Id}_{\mathrm{L}^2(0,1)})} |u|^2_{\mathrm{H}^{t+1,0,I}(\Omega)} \\ &= \tilde{c}^{2k}_{p,0,t} \hat{c}^{2(d-k)}_{p,0,(0)} 2^{-2(t+1)|\ell|_1} |u|^2_{\mathrm{H}^{t+1,0,I}(\Omega)}.\end{aligned}$$

Summing this bound over all $I \subset \{1, 2, \ldots, d\}$ with $|I| = k$ implies

$$\|u - \hat{P}^{L,p}_0 u\|_{\mathrm{L}^2(\Omega)} \leq \sum_{k=1}^{d} \binom{d}{k} \tilde{c}^k_{p,0,t} \hat{c}^{d-k}_{p,0,(0)} \left\{ \sum_{\substack{\ell \in \mathbb{N}^k \\ |\ell|_1 > L}} 2^{-(t+1)|\ell|_1} \right\} \max_{1 \leq k \leq d} \left(\max_{\substack{I \subset \{1,2,\ldots,d\} \\ |I|=k}} |u|_{\mathrm{H}^{t+1,0,I}(\Omega)} \right).$$

Now, according to part (b) of Lemma 14 with $s = 0$, for $k \geq 2$, any $L \geq 1$ and any $t \geq 1$, we have

$$\sum_{\substack{\ell \in \mathbb{N}^k \\ |\ell|_1 > L}} 2^{-(t+1)|\ell|_1} \leq a^k \cdot 2^{-(t+1)L} L^{k-1}, \tag{5.29}$$

where $a = a_{t,L} \in \mathbb{R}_{>0}$, is independent of k, and such that

$$\frac{1}{\sqrt{2}} \left[\frac{1}{2(t+1)(\ln 2)\sqrt{\mathrm{e}\pi}} \left(\frac{2\mathrm{e}}{L} \right)^{k-\frac{1}{2}} + \left(\frac{4}{(t+1)L \ln 2} \right)^{k-1} + \frac{4^{k-1}}{(k-1)!} \right] \leq k a^{k-1}$$

for all $k = 1, 2, \ldots$. It follows from the structure of the expression in the square bracket that such a number a always exists: we define $a = a_{t,L}$ as the supremum, over all $k \geq 1$, of the $(k-1)$st root of the left-hand side of the last inequality divided by k. Clearly, $a_{t,L}$ is a monotonic decreasing function of both $t \geq 1$ and $L \geq 1$.

The values of $a = a_{t,L}$ computed for the (most pessimistic) choice of $t = 1$, as well as for $t = 10, 100, 1000$, and the range $L = 1, \ldots, 10$, are shown in Table 1. For $t > 1$, we see that $a_{t,L}$ is bounded above by $a_{1,L}$ for all $L \geq 1$; in particular, $a_{1,1} = 5.30$, rounded (up) to two decimal digits, represents an upper bound for $a_{t,L}$ for any $t \geq 1$ and $L \geq 1$; i.e. $a_{t,L} \leq 5.30$, for all $t \geq 1$ and all $L \geq 1$. It is interesting to note that $a_{1,L} = 1.19$, rounded (up) to two decimal digits, for all $L \geq 5$; therefore $a_{t,L} \leq a_{1,L} = 1.19$ for all $L \geq 5$ and all $t \geq 1$. In particular,

$$\lim_{t \to \infty, L \to \infty} a_{t,L} = \sup_{k \geq 1} \frac{4}{(\sqrt{2}(k-1)!)^{1/(k-1)}} = 1.1718.$$

L	1	2	3	4	5	6	7	8	9	10
$a_{1,L}$	5.30	2.66	1.77	1.33	1.19	1.19	1.19	1.19	1.19	1.19
$a_{10,L}$	5.28	2.65	1.77	1.33	1.19	1.19	1.19	1.19	1.19	1.19
$a_{100,L}$	5.25	2.63	1.76	1.32	1.18	1.18	1.18	1.18	1.18	1.18
$a_{1000,L}$	5.19	2.62	1.75	1.32	1.18	1.18	1.18	1.18	1.18	1.18

Table 1: Values of $a = a_{t,L}$, rounded (up) to two decimal digits, for $t = 1, 10, 100, 1000$ and $L = 1, \ldots, 10$. The numbers in the table show insensitivity of $a_{t,L}$ with respect to the choice of t, with $a_{1,L}$ being an accurate approximation to $a_{t,L}$ for all values of t of practical interest.

Now we are ready to continue our argument in the case of $s = 0$. We see that

$$\begin{aligned}
\|u - \hat{P}^{L,p}_{(0)} u\|_{\mathrm{L}^2(\Omega)} &\le \frac{\tilde{c}_{p,0,t}}{\hat{c}_{p,0,(0)}} \left\{ \sum_{k=1}^{d} \binom{d}{k} \tilde{c}^{k-1}_{p,0,t} \hat{c}^{d-k+1}_{p,0,(0)}\, k\, a^{k-1}_{t,L} L^{k-1} \right\} 2^{-(t+1)L} \\
&\qquad \times \max_{1\le k\le d} \left(\max_{\substack{I\subset\{1,2,\ldots,d\}\\ |I|=k}} |u|_{\mathrm{H}^{t+1,0,I}(\Omega)} \right) \\
&= \frac{\tilde{c}_{p,0,t}}{\hat{c}_{p,0,(0)}} \left\{ \sum_{k=0}^{d-1} \binom{d}{k+1} \tilde{c}^{k}_{p,0,t} \hat{c}^{d-k}_{p,0,(0)}\, (k+1)\, a^{k}_{t,L} L^{k} \right\} 2^{-(t+1)L} \max_{1\le k\le d} \left(\max_{\substack{I\subset\{1,2,\ldots,d\}\\ |I|=k}} |u|_{\mathrm{H}^{t+1,0,I}(\Omega)} \right) \\
&= d\, \frac{\tilde{c}_{p,0,t}}{\hat{c}_{p,0,(0)}} \left\{ \sum_{k=0}^{d-1} \binom{d-1}{k} \tilde{c}^{k}_{p,0,t} \hat{c}^{d-k}_{p,0,(0)} a^{k}_{t,L} L^{k} \right\} 2^{-(t+1)L} \max_{1\le k\le d} \left(\max_{\substack{I\subset\{1,2,\ldots,d\}\\ |I|=k}} |u|_{\mathrm{H}^{t+1,0,I}(\Omega)} \right).
\end{aligned}$$

Therefore,

$$\|u - \hat{P}^{L,p}_{(0)} u\|_{\mathrm{L}^2(\Omega)} \le \tilde{c} d (a_{t,L} \tilde{c}_{p,0,t} L + \hat{c}_{p,0,(0)})^{d-1} 2^{-(t+1)L} \max_{1\le k\le d} \left(\max_{\substack{I\subset\{1,2,\ldots,d\}\\ |I|=k}} |u|_{\mathrm{H}^{t+1,0,I}(\Omega)} \right) \tag{5.30}$$

$$\le \tilde{c} d (a_{t,L} \tilde{c}_{p,0,t} + \hat{c}_{p,0,(0)}/L)^{d-1} L^{d-1} 2^{-(t+1)L} \max_{1\le k\le d} \left(\max_{\substack{I\subset\{1,2,\ldots,d\}\\ |I|=k}} |u|_{\mathrm{H}^{t+1,0,I}(\Omega)} \right), \tag{5.31}$$

where $\tilde{c} = \tilde{c}(p,t)$. Defining

$$\kappa(p,t,0,L) := a_{t,L} \tilde{c}_{p,0,t} + \hat{c}_{p,0,(0)}/L, \qquad p \ge 1, \quad 1 \le t \le p, \quad L \ge 1, \tag{5.32}$$

we obtain

$$\|u - \hat{P}^{L,p}_{(0)} u\|_{\mathrm{L}^2(\Omega)} \le \tilde{c} d (\kappa(p,t,0,L))^{d-1} L^{d-1} 2^{-(t+1)L} \max_{1\le k\le d} \left(\max_{\substack{I\subset\{1,2,\ldots,d\}\\ |I|=k}} |u|_{\mathrm{H}^{t+1,0,I}(\Omega)} \right). \tag{5.33}$$

That completes the proof. ∎

In the error bound (5.21), the exponential dependence of the constant on the dimension d enters through $\kappa(p,s,t,L)$ defined in (5.28), (5.32). Next we discuss sufficient conditions under which this constant is less than 1.

Remark 5 Note that the factor $\kappa(p,t,s,L)^{d-1+s}$ appearing in the bound (5.21), with $\kappa(p,t,s,L)$ defined in (5.28) and (5.32) for $s = 1$ and $s = 0$, respectively, decreases exponentially with $d \to \infty$, if

$$\frac{\hat{c}_{p,0,(0)}}{L} < 1 \quad \text{when } s = 0 \qquad \text{and} \qquad \hat{c}_{p,0,(0)} < 1 \quad \text{when } s = 1,$$

and

$$\tilde{c}_{p,0,t} \leq \frac{1-(\hat{c}_{p,0,(0)}/L)}{a_{t,L}} \quad \text{when } s=0 \qquad \text{and} \qquad \tilde{c}_{p,0,t} \leq \frac{1-\hat{c}_{p,0,(0)}}{2} \quad \text{when } s=1.$$

For the projector considered in Example 1 the latter pair of inequalities is equivalent to

$$\left(1+\frac{1}{2^{t+1-s}}\right)\frac{1}{p}\sqrt{\frac{(p-t)!}{(p+t)!}} \leq \begin{cases} (1-(\hat{c}_{p,0,0}/L))/a_{t,L} & \text{when } s=0, \\ (1-\hat{c}_{p,0,0})/2 & \text{when } s=1, \end{cases} \tag{5.34}$$

while, at least in the case of a homogeneous Dirichlet boundary condition on the whole of Γ (viz. $\Gamma=\Gamma_0$ by virtue of $a=(a_{ij})_{i,j=1}^d$ being positive definite), when $\mathrm{H}^1_{(0)}(\Omega)=\mathrm{H}^1_0(\Omega)$, the first pair of inequalities holds trivially for all $L\geq 1$ since $\hat{c}_{p,0,(0)}=\hat{c}_{p,0,0}\leq 1/\pi$ (<1).

By scanning the range of validity of (5.34), we then deduce that, with the projector defined in Example 1, in the case of a homogeneous Dirichlet boundary condition on the whole Γ (viz. $\Gamma=\Gamma_0$), we have

$$\kappa(p,p,s,L)<1 \qquad \forall p\geq 2, \quad s\in\{0,1\}, \quad L\geq 1,$$

thus ensuring exponential decay of the term $\kappa(p,p,s,L)^{d-1+s}$ in (5.21) with $d\to\infty$, for all $p\geq 2$, $s\in\{0,1\}$ and $L\geq 1$. $\diamond$

Remark 6 For $p=1$ and if $\Gamma=\Gamma_0$ (i.e. the hyperbolic part $\Gamma_-\cup\Gamma_+$ of the boundary is empty), condition (5.34) (which is sharp as $p\to\infty$) is also applicable but is overly conservative. Using in (5.32), (5.28) the bounds (3.23), we obtain, for $s=1$, that

$$\kappa(1,1,1,L)\leq \frac{2}{3}+\frac{1}{\pi}\leq 0.985 \qquad \forall L\geq 1, \tag{5.35}$$

and, for $s=0$, based on Table 1 that

$$\kappa(1,1,0,L)=a_{1,L}\tilde{c}_{1,0,1}+\hat{c}_{1,0,0}/L\leq 1.77/3+1/(\pi L)\leq 0.7, \tag{5.36}$$

for all $L\geq 3$.

Remark 7 A result analogous to that contained in (5.21), in the special case of $s=1$ and $p=1$, and with $\kappa<1$ was stated in Theorem 2 in [10]. There, however, an "energy-norm-based" sparse-grid-space was use that is strictly included in $\hat{V}_0^L$. As a matter of fact, unlike (5.21), the result contained in [10] is restricted to the case of $s=1$ and $p=1$ and does not cover either $s=0$ or $p\geq 2$. $\diamond$

Remark 8 If $\Gamma_0\subsetneq\Gamma$ (i.e. the hyperbolic part $\Gamma_-\cup\Gamma_+$ of the boundary Γ is nonempty), and therefore $\mathrm{H}^1_0(\Omega)\subsetneq\mathrm{H}^1_{(0)}(\Omega)$, then we still have $\hat{c}_{p,0,(0)}\leq 1$ by (3.12) and consequently $\hat{c}_{p,0,(0)}/L<1$ for all $p\geq 2$ and all $L\geq 2$, whereby (now, corresponding to the case $s=0$ only)

$$\kappa(p,p,0,L)<1 \qquad \forall p\geq 2, \quad L\geq 2.$$

Concerning the case of $s=1$, if

$$\left(1+\frac{1}{2^p}\right)\frac{1}{p}\sqrt{\frac{1}{(2p)!}}\leq\frac{1}{d}, \tag{5.37}$$

which is a very mild condition on the minimum size of p in terms of d, then we have that

$$(\kappa(p,p,1,L))^d \leq \left(1+\frac{2}{d}\right)^d \leq \mathrm{e}^2,$$

which, in turn, ensures that $(\kappa(p,p,1,L))^d$ remains uniformly bounded for $d \gg 1$. For example, for $d \leq 7$ the condition (5.37) requires $p = 2$, for $8 \leq d \leq 71$ taking $p = 3$ will suffice, while for for $71 \leq d \leq 755$ taking $p = 4$ will be sufficient.

The discussion in the previous paragraph presupposed that the number of hyperbolic co-ordinate directions is equal to, or is very close to, d. If, however, the number of such directions is small relative to d, and can be regarded as being bounded as $d \to \infty$, then we expect that the factor $(\kappa(p,p,1,L))^d$ will exhibit exponential decay as $d \to \infty$ without the extra hypothesis (5.37), just as in the case when $\Gamma = \Gamma_0$. The proof of this would require a selective treatment of the constant $\hat{c}_{p,0,(0)}$ in the proof of Theorem 16 when $s = 1$, to monitor whether a particular factor of $\hat{c}_{p,0,(0)}$ in lines 3 and 4 of (5.23) arises from a univariate bound on $Q^{0,p}_{(0)}$ in an elliptic or in a hyperbolic co-ordinate direction. An altogether different approach to removing the condition (5.37) in the case of $\Gamma_0 \subsetneq \Gamma$ and $s = 1$ would be to show that $\hat{c}_{p,0,(0)} < 1$, uniformly in p. These lines of investigation are, however, beyond the scope of the present paper. $\diamond$

Remark 9 In the error bound (5.21) for $s = 0$, i.e. for the error in $\mathrm{L}^2(\Omega)$, we obtain for $t = p$ the optimal convergence rate h_L^{p+1} up to the polylogarithmic term $|\log_2 h_L|^{d-1}$. It is by now well accepted by sparse-grid practitioners that in very high dimensions, where necessarily $L < d$, such polylogarithmic terms dominate the convergence behaviour. However, at least in the case of $\Gamma = \Gamma_0$, the situation for $p \geq 2$ is much more favourable in this respect than for $p = 1$. This somewhat surprising phenomenon is discussed below.

As is evident from (5.30), the factor L^{d-1}, which is the source of the polylogarithmic term $|\log_2 h_L|^{d-1}$, can be absorbed into the factor $(\kappa(p,t,0,L))^{d-1}$ if the definition (5.32) of (the dimension-independent constant) $\kappa(p,t,0,L)$ in (5.33) is changed to

$$a_{t,L}\tilde{c}_{p,0,t}L + \hat{c}_{p,0,0}, \tag{5.38}$$

which is, again, independent of d.

Based on the explicit expression for $\tilde{c}_{p,s,t}$, in (5.32), (5.28), we can still ensure that

$$a_{t,L}\tilde{c}_{p,0,t}L + \hat{c}_{p,0,0} < 1,$$

provided that the following mild extra condition relating h_L and p, which does not depend on the dimension d, holds: analogously to (5.34), we require for $L = |\log_2 h_L|$ that

$$L\left(1+\frac{1}{2^{t+1}}\right)\frac{1}{p}\sqrt{\frac{(p-t)!}{(p+t)!}} \leq \frac{1-\hat{c}_{p,0,0}}{a_{t,L}}. \tag{5.39}$$

This holds with $t = p$ and with $\hat{c}_{p,0,0} = 1/\pi$ for example if:

$$p = 2 \text{ and } L \leq 5, \qquad p = 3 \text{ and } L \leq 29, \qquad p \geq 4 \text{ and } L \leq 397. \quad \diamond \tag{5.40}$$

Remark 10 In stark contrast with the case of $p \geq 2$, for $t = p = 1$ there is no value of $L \geq 1$ for which (5.39) holds. Thus, in the case of $p = 1$ there is no $L \geq 1$ for which the factor L^{d-1} may be absorbed into the exponentially decreasing term $(\kappa(p,p,0,L))^{d-1}$ in a way that

would ensure that the resulting term still decreases exponentially as $d \to \infty$. For $p = 1$ we expect the impact of the polylogarithmic factor $|\log h_L|^{d-1}$ on the approximation error to be much more prominent for large d than for $p \geq 2$: as we can see from (5.40), for $p \geq 2$ the polylogarithmic factor $|\log h_L|^{d-1}$ can be completely suppressed for d large in the, practically relevant, preasymptotic range of L. $\diamond$

Remark 11 When $\Gamma_0 = \Gamma$, the function u to be approximated enters into the right-hand side of the estimate (5.21) in a nonstandard, yet favourable manner: through the L^2 norm of exactly one mixed derivative, — rather than through a *sum* of L^2 norms of mixed derivatives as would have been the case had we used a more conventional seminorm on the space of functions with square-integrable highest mixed derivatives. $\diamond$

6 Convergence of the sparse stabilized method

Our goal in this section is to estimate the size of the error between the analytical solution $u \in \mathcal{H}$ and its approximation $u_h \in \hat{V}_{(0)}^{L,p}$. We shall assume throughout that $f \in \mathrm{L}^2(\Omega)$ and the corresponding solution $u \in \mathcal{H}^{k+1}(\Omega) \cap \mathrm{H}^2(\Omega) \cap \bigotimes_{i=1}^d \mathrm{H}^1_{(0)}(0,1) \subset \mathcal{H}$, $k \geq 1$ and $1 \leq t \leq \min(p, k)$. Clearly,

$$b_\delta(u - u_h, v_h) = B(u, v_h) - L(v_h) + \delta_L \sum_{\kappa \in \mathcal{T}^L} (\mathcal{L}u - f, b \cdot \nabla v_h)_\kappa$$

for all $v_h \in \hat{V}_{(0)}^{L,p} \subset \mathcal{V}$. Hence we deduce from (2.6) the following *Galerkin orthogonality* property:

$$b_\delta(u - u_h, v_h) = 0 \qquad \forall v_h \in \hat{V}_{(0)}^{L,p}. \tag{6.1}$$

Let us decompose the error $u - u_h$ as follows:

$$u - u_h = (u - \hat{P}_{(0)}^{L,p} u) + (\hat{P}_{(0)}^{L,p} u - u_h) = \eta + \xi,$$

where $\eta := u - \hat{P}_{(0)}^{L,p} u$ and $\xi := \hat{P}_{(0)}^{L,p} u - u_h$. By the triangle inequality,

$$|||u - u_h|||_{\mathrm{SD}} \leq |||\eta|||_{\mathrm{SD}} + |||\xi|||_{\mathrm{SD}}. \tag{6.2}$$

We begin by bounding $|||\xi|||_{\mathrm{SD}}$. By (4.6) and (6.1), we have that

$$|||\xi|||_{\mathrm{SD}}^2 \leq b_\delta(\xi, \xi) = b_\delta(u - u_h, \xi) - b_\delta(\eta, \xi) = -b_\delta(\eta, \xi).$$

Therefore,

$$|||\xi|||_{\mathrm{SD}}^2 \leq |b_\delta(\eta, \xi)|. \tag{6.3}$$

Now,

$$\begin{aligned} b_\delta(\eta, \xi) &= (a\nabla\eta, \nabla\xi) - (\eta, b \cdot \nabla\xi) + (c\eta, \xi) + \int_{\Gamma_+} |\beta| \eta\xi \, \mathrm{d}s \\ &\quad + \delta_L \sum_{\kappa \in \mathcal{T}^L} (-a : \nabla\nabla\eta + b \cdot \nabla\eta + c\eta, b \cdot \nabla\xi)_\kappa \\ &= \mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV} + (\mathrm{V} + \mathrm{VI} + \mathrm{VII}) \,. \end{aligned}$$

For the terms I to VII we have:

$$\begin{aligned}
\text{I} &\leq \left(|\sqrt{a}\,|\,\|\nabla\eta\|_{\mathrm{L}^2(\Omega)}\right)|||\xi|||_{\mathrm{SD}},\\
\text{II} &\leq \left(\delta_L^{-\frac{1}{2}}\|\eta\|_{\mathrm{L}^2(\Omega)}\right)|||\xi|||_{\mathrm{SD}},\\
\text{III} &\leq \left(c^{\frac{1}{2}}\|\eta\|_{\mathrm{L}^2(\Omega)}\right)|||\xi|||_{\mathrm{SD}},\\
\text{V} &\leq \left(\delta_L^{\frac{1}{2}}|a|\left(\sum_{\kappa\in\mathcal{T}^L}|\eta|^2_{\mathrm{H}^2(\kappa)}\right)^{\frac{1}{2}}\right)|||\xi|||_{\mathrm{SD}},\\
\text{VI} &\leq \left(\delta_L^{\frac{1}{2}}|b|\,\|\nabla\eta\|_{\mathrm{L}^2(\Omega)}\right)|||\xi|||_{\mathrm{SD}},\\
\text{VII} &\leq \left(c\delta_L^{\frac{1}{2}}\|\eta\|_{\mathrm{L}^2(\Omega)}\right)|||\xi|||_{\mathrm{SD}}.
\end{aligned}$$

Here $|a|$ is the Frobenius norm of the matrix a and $|b|$ is the Euclidean norm of the vector b. It remains to estimate IV:

$$\begin{aligned}
\text{IV} &\leq \left(\frac{2|b|}{1+c\delta_L}\right)^{\frac{1}{2}}\left(\int_{\Gamma_+}|\eta|^2\,\mathrm{d}s\right)^{\frac{1}{2}}|||\xi|||_{\mathrm{SD}}\\
&\leq (2|b|)^{\frac{1}{2}}\,(4d)^{\frac{1}{2}}\|\eta\|^{\frac{1}{2}}_{\mathrm{L}^2(\Omega)}\|\eta\|^{\frac{1}{2}}_{\mathrm{H}^1(\Omega)}|||\xi|||_{\mathrm{SD}},
\end{aligned}$$

where in the transition to the last line we used the multiplicative trace inequality from Lemma 9. Hence, by (6.3),

$$\begin{aligned}
|||\xi|||_{\mathrm{SD}} \leq\ & |\sqrt{a}\,|\,\|\nabla\eta\|_{\mathrm{L}^2(\Omega)}+\delta_L^{-\frac{1}{2}}\|\eta\|_{\mathrm{L}^2(\Omega)}+\sqrt{c}\,\|\eta\|_{\mathrm{L}^2(\Omega)}+\sqrt{8d|b|}\,\|\eta\|^{\frac{1}{2}}_{\mathrm{L}^2(\Omega)}\,\|\eta\|^{\frac{1}{2}}_{\mathrm{H}^1(\Omega)}\\
&+\delta_L^{\frac{1}{2}}|\sqrt{a}\,|^2\left(\sum_{\kappa\in\mathcal{T}^L}|\eta|^2_{\mathrm{H}^2(\kappa)}\right)^{\frac{1}{2}}+\delta_L^{\frac{1}{2}}|b|\,\|\nabla\eta\|_{\mathrm{L}^2(\Omega)}+c\delta_L^{\frac{1}{2}}\|\eta\|_{\mathrm{L}^2(\Omega)}. \qquad (6.4)
\end{aligned}$$

The bounds on $\|\eta\|_{\mathrm{L}^2(\Omega)}$ and $\|\nabla\eta\|_{\mathrm{L}^2(\Omega)}$ will follow from Theorem 16. However, the fifth term in the sum on the right-hand side of (6.4) is nonstandard and needs to be bounded separately (except in the case of $p=1$ when this term is equal to $\delta_L^{1/2}|a||u|_{\mathrm{H}^2(\Omega)}$ and requires

no further estimation). Let us now suppose therefore that $p \geq 2$, and note that

$$\begin{aligned}
\sum_{\kappa\in\mathcal{T}^L} |\eta|^2_{\mathrm{H}^2(\kappa)} &= \sum_{i,j=1}^{d} \sum_{\kappa\in\mathcal{T}^L} \int_\kappa \left|\frac{\partial^2\eta}{\partial x_i \partial x_j}\right|^2 \mathrm{d}x \\
&= \sum_{i=1}^{d} \sum_{\kappa\in\mathcal{T}^L} \int_\kappa \left|\frac{\partial^2\eta}{\partial x_i^2}\right|^2 \mathrm{d}x + \sum_{\substack{i,j=1\\ i\neq j}}^{d} \sum_{\kappa\in\mathcal{T}^L} \int_\kappa \left|\frac{\partial^2\eta}{\partial x_i \partial x_j}\right|^2 \mathrm{d}x \\
&= \sum_{i=1}^{d} \sum_{\kappa\in\mathcal{T}^L} |\eta|^2_{\mathrm{H}^{2,0,\{i\}}(\kappa)} + \sum_{\substack{i,j=1\\ i\neq j}}^{d} |\eta|^2_{\mathrm{H}^{1,0,\{i,j\}}(\Omega)} \\
&=: \mathrm{A}^2 + \mathrm{B}^2,
\end{aligned}$$

Here, we made use of the fact that

$$\frac{\partial^2\eta}{\partial x_i \partial x_j} \in \mathrm{L}^2(\Omega) \qquad \forall i,j \in \{1,2,\ldots,d\}, \qquad i \neq j.$$

Let us first estimate

$$\mathrm{A}^2 = \sum_{i=1}^{d} \sum_{\kappa\in\mathcal{T}^L} |\eta|^2_{\mathrm{H}^{2,0,\{i\}}(\kappa)} = \sum_{i=1}^{d} \sum_{\kappa\in\mathcal{T}^L} \int_\kappa \left|\frac{\partial^2\eta}{\partial x_i^2}\right|^2 \mathrm{d}x = \sum_{i=1}^{d} \sum_{j=1}^{L} |\eta|^2_{\mathrm{H}^{2,0,\{i\}}(K_j^i)},$$

where K_j^i denotes the d-dimensional slab

$$K_j^i = (0,1) \times \cdots \times (0,1) \times (\xi_{j-1}, \xi_j) \times (0,1) \times \cdots \times (0,1) \tag{6.5}$$

where the interval (ξ_{j-1}, ξ_j) enters at position i. The reason for agglomerating the elements $\kappa \in \mathcal{T}^L$ into the slabs K_j^i, $j = 1, \ldots, L$, in this way is that the function $\partial^2\eta/\partial x_i^2$ involves no derivatives is the co-ordinate directions Ox_k for $k \neq i$. In other words, it only needs to be considered piecewise in the i^{th} co-ordinate direction; in the other $d-1$ co-ordinate directions it is defined on the whole of $(0,1)^{d-1}$ as an H^1 function.

Let us define the seminorms $|||\cdot|||_{2,i}$, $i = 1, \ldots, d$, and $|||\cdot|||_{2,*}$, by

$$|||v|||^2_{2,i} = \sum_{j=1}^{L} |v|^2_{\mathrm{H}^{2,0,\{i\}}(K_j^i)} \qquad \text{and} \qquad |||v|||^2_{2,*} = \sum_{i=1}^{d} |||v|||^2_{2,i}.$$

With this notation, we have that

$$\mathrm{A}^2 = |||\eta|||^2_{2,*} = \sum_{i=1}^{d} |||\eta|||^2_{2,i}.$$

In order to bound A, we first observe that, as a consequence of Lemma 8,

$$|v|_{\mathrm{H}^2(I)} \leq \sqrt{12}\,(p^2/h)|v|_{\mathrm{H}^1(I)} \qquad \forall v \in \mathcal{P}^p(I). \tag{6.6}$$

Hence, on recalling that

$$\eta \;=\; u - \hat{P}^{L,p}_{(0)} u = \sum_{\ell\in\mathbb{N}^d\,:\,|\ell|_1 > L} \left(Q^{\ell_1,p}_{(0)} \otimes \cdots \otimes Q^{\ell_d,p}_{(0)}\right) u, \tag{6.7}$$

for a fixed $i \in \{1, 2, \ldots, d\}$ we deduce from (6.6) with $h = h_L = 2^{-L}$ that

$$\left|\left(Q_{(0)}^{\ell_1,p} \otimes \cdots \otimes Q_{(0)}^{\ell_d,p}\right) u\right|_{\mathrm{H}^{2,0,\{i\}}(\kappa)} \;\leq\; \sqrt{12}\, p^2\, 2^L \left|\left(Q_{(0)}^{\ell_1,p} \otimes \cdots \otimes Q_{(0)}^{\ell_d,p}\right) u\right|_{\mathrm{H}^{1,0,\{i\}}(\kappa)}.$$

Now, we square the last bound, and sum over all elements $\kappa \in \mathcal{T}^L$ that are contained in the d-dimensional slab K_j^i defined in (6.5) to deduce that

$$\left|\left(Q_{(0)}^{\ell_1,p} \otimes \cdots \otimes Q_{(0)}^{\ell_d,p}\right) u\right|^2_{\mathrm{H}^{2,0,\{i\}}(K_j^i)} \;\leq\; 12\, p^4\, 2^{2L} \left|\left(Q_{(0)}^{\ell_1,p} \otimes \cdots \otimes Q_{(0)}^{\ell_d,p}\right) u\right|^2_{\mathrm{H}^{1,0,\{i\}}(K_j^i)}.$$

Hence,

$$\begin{aligned}\sum_{j=1}^{L} \left|\left(Q_{(0)}^{\ell_1,p} \otimes \cdots \otimes Q_{(0)}^{\ell_d,p}\right) u\right|^2_{\mathrm{H}^{2,0,\{i\}}(K_j^i)} &\leq 12\, p^4\, 2^{2L} \sum_{j=1}^{L} \left|\left(Q_{(0)}^{\ell_1,p} \otimes \cdots \otimes Q_{(0)}^{\ell_d,p}\right) u\right|^2_{\mathrm{H}^{1,0,\{i\}}(K_j^i)} \\ &= 12\, p^4\, 2^{2L} \left|\left(Q_{(0)}^{\ell_1,p} \otimes \cdots \otimes Q_{(0)}^{\ell_d,p}\right) u\right|^2_{\mathrm{H}^{1,0,\{i\}}(\Omega)}.\end{aligned}$$

This implies that

$$\left|\!\left|\!\left|\left(Q_{(0)}^{\ell_1,p} \otimes \cdots \otimes Q_{(0)}^{\ell_d,p}\right) u\right|\!\right|\!\right|^2_{2,i} \leq 12\, p^4\, 2^{2L} \left|\left(Q_{(0)}^{\ell_1,p} \otimes \cdots \otimes Q_{(0)}^{\ell_d,p}\right) u\right|^2_{\mathrm{H}^{1,0,\{i\}}(\Omega)},$$

and, on summing over $i = 1, \ldots, d$, and taking square-root,

$$\begin{aligned}\left|\!\left|\!\left|\left(Q_{(0)}^{\ell_1,p} \otimes \cdots \otimes Q_{(0)}^{\ell_d,p}\right) u\right|\!\right|\!\right|_{2,*} &\leq \sqrt{12}\, p^2\, 2^L \left(\sum_{i=1}^{d} \left|\left(Q_{(0)}^{\ell_1,p} \otimes \cdots \otimes Q_{(0)}^{\ell_d,p}\right) u\right|^2_{\mathrm{H}^{1,0,\{i\}}(\Omega)}\right)^{\frac{1}{2}} \\ &= \sqrt{12}\, p^2\, 2^L \left|\left(Q_{(0)}^{\ell_1,p} \otimes \cdots \otimes Q_{(0)}^{\ell_d,p}\right) u\right|_{\mathrm{H}^1(\Omega)}.\end{aligned}$$

Hence, by (6.7) and the proof of (5.21) in the case of $s = 1$,

$$\begin{aligned}\mathrm{A} = |||\eta|||_{2,*} &\leq \sum_{|\ell|_1 > L} \left|\!\left|\!\left|\left(Q_{(0)}^{\ell_1,p} \otimes \cdots \otimes Q_{(0)}^{\ell_d,p}\right) u\right|\!\right|\!\right|_{2,*} \\ &\leq \sqrt{12}\, p^2\, 2^L \sum_{|\ell|_1 > L} \left|\left(Q_{(0)}^{\ell_1,p} \otimes \cdots \otimes Q_{(0)}^{\ell_d,p}\right) u\right|_{\mathrm{H}^1(\Omega)} \\ &\leq \sqrt{12}\, p^2\, 2^L d^{\frac{3}{2}} \underline{c}_{p,t}\, (\kappa(p,t,1,L))^d\, 2^{-tL} \cdot \max_{1 \leq k \leq d} \left(\max_{\substack{I \subset \{1,2,\ldots,d\} \\ |I| = k}} |u|_{\mathrm{H}^{t+1,1,I}(\Omega)} \right).\end{aligned}$$

Thus we have shown that

$$\mathrm{A} \leq \sqrt{12}\, p^2\, d^{\frac{3}{2}} \underline{c}_{p,t}\, (\kappa(p,t,1,L))^d\, 2^{-(t-1)L} \cdot \max_{1 \leq k \leq d} \left(\max_{\substack{I \subset \{1,2,\ldots,d\} \\ |I| = k}} |u|_{\mathrm{H}^{t+1,1,I}(\Omega)} \right) \tag{6.8}$$

for $1 \leq t \leq \min(p, k)$.

Now, let us bound

$$\mathrm{B}^2 = \sum_{\substack{i,j=1 \\ i \neq j}}^{d} |\eta|^2_{\mathrm{H}^{1,0,\{i,j\}}(\Omega)}.$$

We define the seminorm $|||\cdot|||_{2,**}$ by

$$|||v|||_{2,**}^2 = \sum_{\substack{i,j=1\\ i\neq j}}^{d} |v|^2_{\mathrm{H}^{1,0,\{i,j\}}(\Omega)};$$

then,

$$\mathrm{B}^2 = |||\eta|||_{2,**}^2.$$

Now, since

$$\eta = u - \hat{P}_0^{L,p}u = \sum_{|\ell|_1>L} \left(Q_{(0)}^{\ell_1,p}\otimes\cdots\otimes Q_{(0)}^{\ell_d,p}\right)u,$$

it follows that

$$|||\eta|||_{2,**} \leq \sum_{|\ell|_1>L} \left|\!\left|\!\left| \left(Q_{(0)}^{\ell_1,p}\otimes\cdots\otimes Q_{(0)}^{\ell_d,p}\right)u \right|\!\right|\!\right|_{2,**}.$$

Given $\ell = (\ell_1,\ell_2,\ldots,\ell_d)\in\mathbb{N}^d$ with $\mathrm{supp}(\ell) = I$ (that is, $\ell_j\neq 0$ iff $j\in I$) and $|I| = k$, we have to estimate

$$\left|\!\left|\!\left| \left(Q_{(0)}^{\ell_1,p}\otimes\cdots\otimes Q_{(0)}^{\ell_d,p}\right)u \right|\!\right|\!\right|^2_{2,**} = \sum_{\substack{i,j=1\\ i\neq j}}^{d} \left| \left(Q_{(0)}^{\ell_1,p}\otimes\cdots\otimes Q_{(0)}^{\ell_d,p}\right)u \right|^2_{\mathrm{H}^{1,0,\{i,j\}}(\Omega)} =: (\star\star).$$

Using Proposition 15 and the notation ∂ for the univariate differentiation operator, we obtain the following inequality:

$$\begin{aligned}
(\star\star) \;\leq\; & \sum_{\substack{i,j\in I\\ i\neq j}} \prod_{\substack{j'\in I\\ j'\notin\{i,j\}}} |Q_{(0)}^{\ell_{j'},p}|^2_{(\partial^{t+1},\mathrm{Id}_{\mathrm{L}^2(0,1)})} \\
& \qquad \cdot |Q_{(0)}^{\ell_i,p}|^2_{(\partial^{t+1},\partial^1)} |Q_{(0)}^{\ell_j,p}|^2_{(\partial^{t+1},\partial^1)} |Q_{(0)}^{0,p}|^{2(d-k)}_{(\mathrm{Id}_{\mathrm{H}^1_{(0)}(0,1)},\mathrm{Id}_{\mathrm{L}^2(0,1)})} |u|^2_{\mathrm{H}^{t+1,1,I}(\Omega)} \\
& + \sum_{i\in I}\sum_{j\notin I} \prod_{\substack{j'\in I\\ j'\neq i}} |Q_{(0)}^{\ell_{j'},p}|^2_{(\partial^{t+1},\mathrm{Id}_{\mathrm{L}^2(0,1)})} \\
& \qquad \cdot |Q_{(0)}^{\ell_i,p}|^2_{(\partial^{t+1},\partial^1)} |Q_{(0)}^{0,p}|^2_{(\partial^1,\partial^1)} |Q_{(0)}^{0,p}|^{2(d-k-1)}_{(\mathrm{Id}_{\mathrm{H}^1_{(0)}(0,1)},\mathrm{Id}_{\mathrm{L}^2(0,1)})} |u|^2_{\mathrm{H}^{t+1,1,I}(\Omega)} \\
& + \sum_{i\notin I}\sum_{j\in I} \prod_{\substack{j'\in I\\ j'\neq j}} |Q_{(0)}^{\ell_{j'},p}|^2_{(\partial^{t+1},\mathrm{Id}_{\mathrm{L}^2(0,1)})} \\
& \qquad \cdot |Q_{(0)}^{0,p}|^2_{(\partial^1,\partial^1)} |Q_{(0)}^{\ell_j,p}|^2_{(\partial^{t+1},\partial^1)} |Q_{(0)}^{0,p}|^{2(d-k-1)}_{(\mathrm{Id}_{\mathrm{H}^1_{(0)}(0,1)},\mathrm{Id}_{\mathrm{L}^2(0,1)})} |u|^2_{\mathrm{H}^{t+1,1,I}(\Omega)} \\
& + \sum_{\substack{i,j\notin I\\ i\neq j}} \prod_{j'\in I} |Q_{(0)}^{\ell_{j'},p}|^2_{(\partial^{t+1},\mathrm{Id}_{\mathrm{L}^2(0,1)})} \\
& \qquad \cdot |Q_{(0)}^{0,p}|^2_{(\partial^1,\partial^1)} |Q_{(0)}^{0,p}|^2_{(\partial^1,\partial^1)} |Q_{(0)}^{0,p}|^{2(d-k-2)}_{(\mathrm{Id}_{\mathrm{H}^1_{(0)}(0,1)},\mathrm{Id}_{\mathrm{L}^2(0,1)})} |u|^2_{\mathrm{H}^{t+1,1,I}(\Omega)}. \qquad (6.9)
\end{aligned}$$

Hence,

$$\begin{aligned}(\star\star) \;\leq\; & \sum_{\substack{i,j\in I\\ i\neq j}} \tilde{c}_{p,0,t}^{2(k-2)} 4^{\ell_i+\ell_j-(t+1)|\ell|_1} \tilde{c}_{p,1,t}^2 \tilde{c}_{p,1,t}^2 \hat{c}_{p,0,(0)}^{2(d-k)} |u|^2_{\mathrm{H}^{t+1,1,I}(\Omega)} \\ & + \sum_{i\in I} \sum_{j\notin I} \tilde{c}_{p,0,t}^{2(k-1)} 4^{\ell_i-(t+1)|\ell|_1} \tilde{c}_{p,1,t}^2 \hat{c}_{p,1,(0)}^2 \hat{c}_{p,0,(0)}^{2(d-k-1)} |u|^2_{\mathrm{H}^{t+1,1,I}(\Omega)} \\ & + \sum_{i\notin I} \sum_{j\in I} \tilde{c}_{p,0,t}^{2(k-1)} 4^{\ell_j-(t+1)|\ell|_1} \hat{c}_{p,1,(0)}^2 \tilde{c}_{p,1,t}^2 \hat{c}_{p,0,(0)}^{2(d-k-1)} |u|^2_{\mathrm{H}^{t+1,1,I}(\Omega)} \\ & + \sum_{\substack{i,j\notin I\\ i\neq j}} \tilde{c}_{p,0,t}^{2k} 4^{-(t+1)|\ell|_1} \hat{c}_{p,1,(0)}^2 \hat{c}_{p,1,(0)}^2 \hat{c}_{p,0,(0)}^{2(d-k-2)} |u|^2_{\mathrm{H}^{t+1,1,I}(\Omega)}. \end{aligned} \tag{6.10}$$

Thus we deduce that

$$\begin{aligned}(\star\star) \;\leq\; & \tilde{c}_{p,0,t}^{2(k-2)} \hat{c}_{p,0,(0)}^{2(d-k-2)} \cdot 4^{-(t+1)|\ell|_1} |u|^2_{\mathrm{H}^{t+1,1,I}(\Omega)} \\ & \times \left(\tilde{c}_{p,1,t}^4 \hat{c}_{p,0,(0)}^4 \sum_{\substack{i,j\in I\\ i\neq j}} 4^{\ell_i+\ell_j} + 2\tilde{c}_{p,1,t}^2 \hat{c}_{p,0,(0)}^2 \tilde{c}_{p,0,t}^2 \hat{c}_{p,1,(0)}^2 \sum_{i\in I, j\notin I} 4^{\ell_i} + \tilde{c}_{p,0,t}^4 \hat{c}_{p,1,(0)}^4 \sum_{\substack{i,j\notin I\\ i\neq j}} 1 \right) \\ \leq\; & \tilde{c}_{p,0,t}^{2(k-2)} \hat{c}_{p,0,(0)}^{2(d-k-2)} \cdot 4^{-(t+1)|\ell|_1} |u|^2_{\mathrm{H}^{t+1,1,I}(\Omega)} \\ & \quad \times \Big(\tilde{c}_{p,1,t}^4 \hat{c}_{p,0,(0)}^4 k^2 4^{|\ell|_1} + 2\tilde{c}_{p,1,t}^2 \hat{c}_{p,0,(0)}^2 \tilde{c}_{p,0,t}^2 \hat{c}_{p,1,(0)}^2 k(d-k) 4^{|\ell|_1} \\ & \qquad + \tilde{c}_{p,0,t}^4 \hat{c}_{p,1,(0)}^4 [(d-k)^2 - (d-k)] \Big) \\ \leq\; & d^2 \bar{\bar{c}}_{p,t} \tilde{c}_{p,0,t}^{2(k-2)} \hat{c}_{p,0,(0)}^{2(d-k-2)} \cdot 4^{-t|\ell|_1} |u|^2_{\mathrm{H}^{t+1,1,I}(\Omega)}, \end{aligned}$$

where

$$\bar{\bar{c}}_{p,t} := \max\left(\tilde{c}_{p,1,t}^4 \hat{c}_{p,0,(0)}^4, \tilde{c}_{p,1,t}^2 \hat{c}_{p,0,(0)}^2 \tilde{c}_{p,0,t}^2 \hat{c}_{p,1,(0)}^2, \tilde{c}_{p,0,t}^4 \hat{c}_{p,1,(0)}^4 \right).$$

Therefore, we have that

$$\begin{aligned} & \sum_{\substack{\ell\in\mathbb{N}^d\,:\,|\ell|_1>L\\ \mathrm{supp}(\ell)=I}} \left|\!\left|\!\left| \left(Q_{(0)}^{\ell_1,p} \otimes \cdots \otimes Q_{(0)}^{\ell_d,p} \right) u \right|\!\right|\!\right|_{2,**} \\ & \qquad \leq d\sqrt{\bar{\bar{c}}_{p,t}}\, \tilde{c}_{p,0,t}^{k-2}\, \hat{c}_{p,0,(0)}^{d-k-2} \cdot \sum_{\substack{\ell\in\mathbb{N}^d\,:\,|\ell|_1>L\\ \mathrm{supp}(\ell)=I}} 2^{-t|\ell|_1} |u|_{\mathrm{H}^{t+1,1,I}(\Omega)} \\ & \qquad \leq d\sqrt{\bar{\bar{c}}_{p,t}}\, \tilde{c}_{p,0,t}^{k-2}\, \hat{c}_{p,0,(0)}^{d-k-2} \cdot \left(\sum_{\ell\in\mathbb{N}^k\,:\,|\ell|_1>L} 2^{-t|\ell|_1} \right) |u|_{\mathrm{H}^{t+1,1,I}(\Omega)}. \end{aligned}$$

Once again, we note in passing that in the (important) special case when $\Gamma = \Gamma_0$, and thereby $\mathrm{H}^1_{(0)}(0,1) = \mathrm{H}^1_0(0,1)$ in each of the d co-ordinate directions, the factor $|Q_{(0)}^{0,p}|_{(\mathrm{Id}_{\mathrm{H}^1_{(0)}(0,1)}, \mathrm{Id}_{\mathrm{L}^2(0,1)})}$ in the lines above can be replaced by $|Q_{(0)}^{0,p}|_{(\partial^1, \mathrm{Id}_{\mathrm{L}^2(0,1)})}$.

By applying Lemma 14(a) with $t \geq 1/(\ln 2)$, we have that

$$\sum_{\ell\in\mathbb{N}^k\,:\,|\ell|_1>L} 2^{-t|\ell|_1} \leq \frac{1}{\sqrt{2}} \left[\frac{1}{2t(\ln 2)\sqrt{\mathrm{e}\pi}} \left(\frac{2\mathrm{e}}{L} \right)^{k-\frac{1}{2}} + \left(\frac{4}{tL\ln 2} \right)^{k-1} + \frac{4^{k-1}}{(k-1)!} \right] \cdot 2^{-tL}\, L^{k-1}.$$

Let $b = b_{t,L} > 0$ be a positive real number, independent of k, such that

$$\frac{1}{\sqrt{2}}\left[\frac{1}{2t(\ln 2)\sqrt{\mathrm{e}\pi}}\left(\frac{2\mathrm{e}}{L}\right)^{k-\frac{1}{2}} + \left(\frac{4}{tL\ln 2}\right)^{k-1} + \frac{4^{k-1}}{(k-1)!}\right]\cdot 2^{-tL}\,L^{k-1} \le k b^{k-1}, \quad k = 1, 2, \ldots.$$

Hence,

$$\begin{aligned}
&\sum_{k=1}^{d}\sum_{\substack{I\subset\{1,2,\ldots,d\}\\|I|=k}}\sum_{\substack{\ell\in\mathbb{N}^d\,:\,|\ell|_1>L\\ \mathrm{supp}(\ell)=I}}\left\|\left(Q_{(0)}^{\ell_1,p}\otimes\cdots\otimes Q_{(0)}^{\ell_d,p}\right)u\right\|_{2,**}\\
&\qquad\le d\sqrt{\bar{\bar{c}}_{p,t}}\sum_{k=1}^{d}\binom{d}{k}\tilde{c}_{p,0,t}^{k-2}\,\hat{c}_{p,0,(0)}^{d-k-2}\,k\,b_{t,L}^{k}\,2^{-tL}L^{k-1}\max_{1\le k\le d}\left(\max_{\substack{I\subset\{1,2,\ldots,d\}\\|I|=k}}|u|_{\mathrm{H}^{t+1,1,I}(\Omega)}\right)\\
&\qquad= \frac{d^2\sqrt{\bar{\bar{c}}_{p,t}}}{\tilde{c}_{p,0,t}\,\hat{c}_{p,0,(0)}^2}\left(b_{t,L}\tilde{c}_{p,0,t}+\frac{\hat{c}_{p,0,(0)}}{L}\right)^{d-1}L^{d-1}2^{-tL}\max_{1\le k\le d}\left(\max_{\substack{I\subset\{1,2,\ldots,d\}\\|I|=k}}|u|_{\mathrm{H}^{t+1,1,I}(\Omega)}\right).
\end{aligned}$$

Upon redefining $\kappa(p,t,0,L)$ as

$$\kappa(p,t,0,L) := \max(a_{t,L}, b_{t,L})\,\tilde{c}_{p,0,t} + \frac{\hat{c}_{p,0,(0)}}{L},$$

we deduce that

$$\mathrm{B} \le d^2\underline{\underline{c}}_{p,t}(\kappa(p,t,0,L))^{d-1}L^{d-1}2^{-tL}\max_{1\le k\le d}\left(\max_{\substack{I\subset\{1,2,\ldots,d\}\\|I|=k}}|u|_{\mathrm{H}^{t+1,1,I}(\Omega)}\right), \tag{6.11}$$

where

$$\underline{\underline{c}}_{p,t} := \frac{\sqrt{\bar{\bar{c}}_{p,t}}}{\tilde{c}_{p,0,t}\,\hat{c}_{p,0,(0)}^2}.$$

Combining the bound (6.8) on A with the bound (6.11) on B yields

$$\begin{aligned}
\left(\sum_{\kappa\in\mathcal{T}^L}|\eta|_{\mathrm{H}^2(\kappa)}^2\right)^{\frac{1}{2}} \;\le\; &\left(\sqrt{12}\,p^2\,d^{\frac{3}{2}}\underline{c}_{p,t}(\kappa(p,t,1,L))^d + d^2\underline{\underline{c}}_{p,t}(\kappa(p,t,0,L))^{d-1}|\log_2 h_L|^{d-1}h_L\right)\\
&\times h_L^{t-1}\cdot\max_{1\le k\le d}\left(\max_{\substack{I\subset\{1,2,\ldots,d\}\\|I|=k}}|u|_{\mathrm{H}^{t+1,1,I}(\Omega)}\right),
\end{aligned} \tag{6.12}$$

for $1/(\ln 2) \le t \le \min(p,k)$, $p \ge 2$, $k \ge 2$.

We also know from Theorem 16 that, for $1 \le t \le \min(p,k)$, $p \ge 1$, $k \ge 1$,

$$\|\eta\|_{\mathrm{L}^2(\Omega)} \le d\underline{c}_{p,t}(\kappa(p,t,0,L))^{d-1}h_L^{t+1}|\log_2 h_L|^{d-1}\cdot\max_{1\le k\le d}\left(\max_{\substack{I\subseteq\{1,2,\ldots,d\}\\|I|=k}}|u|_{\mathrm{H}^{t+1,0,I}(\Omega)}\right), \tag{6.13}$$

$$|\eta|_{\mathrm{H}^1(\Omega)} \le d^{\frac{3}{2}}\underline{c}_{p,t}(\kappa(p,t,1,L))^d h_L^t\cdot\max_{1\le k\le d}\left(\max_{\substack{I\subseteq\{1,2,\ldots,d\}\\|I|=k}}|u|_{\mathrm{H}^{t+1,1,I}(\Omega)}\right). \tag{6.14}$$

Let $\mathcal{H}^{t+1}(\Omega)$ denote the closure of $\mathrm{C}^\infty_{(0)}(\bar\Omega)$ in the seminorm $|\cdot|_{\mathcal{H}^{t+1}(\Omega)}$ defined by

$$|u|_{\mathcal{H}^{t+1}(\Omega)} := \max_{s\in\{0,1\}} \max_{1\le k\le d} \left(\max_{\substack{I\subseteq\{1,2,\ldots,d\}\\ |I|=k}} |u|_{\mathrm{H}^{t+1,s,I}(\Omega)} \right),$$

introduce, for ease of writing, the notation

$$\kappa_0 := \kappa(p,t,0,L) \qquad \text{and} \qquad \kappa_1 := \kappa(p,t,1,L),$$

and absorb all constants that depend on p and t only into a generic constant $C_{p,t}$. In particular, $C_{p,t}$ is independent of d and L and the coefficients a, b, c and the right-hand side f of the partial differential equation.

Remark 12 Since (6.12), (6.13), (6.14) and all of our earlier bounds are completely explicit in p and t (as well as in d and L), one could track the actual value of $C_{p,t}$ in our argument below. For clarity of presentation we shall however refrain from doing so, particularly since the emphasis here is on h-version rather than p- or hp-version finite element methods. $\diamond$

With these notational conventions, (6.13), (6.14) and (6.12) become:

$$\begin{aligned}
\|\eta\|_{\mathrm{L}^2(\Omega)} &\le C_{p,t}\, d\, \kappa_0^{d-1} h_L^{t+1} |\log_2 h_L|^{d-1} |u|_{\mathcal{H}^{t+1}(\Omega)}, && (6.15)\\
|\eta|_{\mathrm{H}^1(\Omega)} &\le C_{p,t}\, d^{\frac{3}{2}} \kappa_1^d\, h_L^t |u|_{\mathcal{H}^{t+1}(\Omega)}, && (6.16)\\
\left(\sum_{\kappa\in\mathcal{T}^L} |\eta|^2_{\mathrm{H}^2(\kappa)} \right)^{\frac{1}{2}} &\le C_{p,t} \left(d^{\frac{3}{2}} \kappa_1^d + d^2 \kappa_0^{d-1} |\log_2 h_L|^{d-1} h_L \right) h_L^{t-1} |u|_{\mathcal{H}^{t+1}(\Omega)}. && (6.17)
\end{aligned}$$

Using (6.15), (6.16) and (6.17) in (6.4) and selecting

$$\delta_L := K_\delta \min\left(\frac{h_L^2}{12dp^4 |\sqrt{a}|^2 (1+h_L|\log_2 h_L|^{d-1})^2}, \frac{h_L |\log_2 h_L|^{\frac{d-1}{2}}}{|b|}, \frac{1}{c} \right), \tag{6.18}$$

with $K_\delta \in \mathbb{R}_{>0}$ a constant, independent of h_L and d, we then deduce that

$$\begin{aligned}
|||\xi|||^2_{\mathrm{SD}} &\le C_{p,t} d^4 (\kappa_*(p,t,L))^{2(d-1)} |u|^2_{\mathcal{H}^{t+1}(\Omega)}\\
&\quad \times \left(|\sqrt{a}|^2 h_L^{2t} + |b| h_L^{2t+1} |\log_2 h_L|^{d-1} + c h_L^{2(t+1)} |\log_2 h_L|^{2(d-1)} + h_L^{2(t+1)} |\log_2 h_L|^{2(d-1)} \lambda_L \right),
\end{aligned}$$

where

$$\lambda_L := \max\left(\frac{|\sqrt{a}|^2 (1+h_L|\log_2 h_L|^{d-1})^2}{h_L^2}, \frac{|b|}{h_L |\log_2 h_L|^{d-1}}, c \right), \tag{6.19}$$

$$\kappa_*(p,t,L) = \max(\kappa(p,t,0,L), \kappa(p,t,1,L)), \tag{6.20}$$

and $1/(\ln 2) \le t \le \min(p,k)$, $p \ge 2$, $k \ge 2$. An identical bound holds for $|||\eta|||^2_{\mathrm{SD}}$.

In the case of $p = 1$ the bounds on A and B are redundant, and this simplifies the argument considerably, though ultimately we arrive at identical bounds on $|||\xi|||_{\mathrm{SD}}$ and $|||\eta|||_{\mathrm{SD}}$, only with $t = p = 1$, and with the factor $(1 + h_L|\log_2 h_L|^{d-1})^2$ replaced by 1 in the definitions of δ_L and λ_L.

Inserting the bounds on $|||\xi|||_{\mathrm{SD}}$ and $|||\eta|||_{\mathrm{SD}}$ in the right-hand side of the triangle inequality (6.2), we deduce the following theorem.

Theorem 17 *Suppose that $f \in \mathrm{L}^2(\Omega)$ in $\Omega = (0,1)^d$, that $c > 0$ and assume the regularity $u \in \mathcal{H}^{k+1}(\Omega) \cap \mathrm{H}^2(\Omega) \cap \bigotimes_{i=1}^d \mathrm{H}^1_{(0)}(0,1)$, $k \geq 1$.*

Then, for $p \geq 2$ and $1/(\ln 2) \leq t \leq \min(p,k)$, the following bound holds for the error $u - u_h$ between the analytical solution u of (2.6) and its sparse finite element approximation $u_h \in \hat{V}^{L,p}_{(0)}$ defined by (4.4), with $L \geq 1$ and $h = h_L = 2^{-L}$:

$$
\begin{aligned}
|||u - u_h|||_{\mathrm{SD}} \;&\leq\; C_{p,t} d^2 \kappa_*(p,t,L)^{d-1} |u|_{\mathcal{H}^{t+1}(\Omega)} h_L^t \\
&\times \left(|\sqrt{a}| + \sqrt{|b|} h_L^{\frac{1}{2}} |\log_2 h_L|^{\frac{d-1}{2}} + \sqrt{c} h_L |\log_2 h_L|^{d-1} + h_L |\log_2 h_L|^{d-1} \lambda_L^{\frac{1}{2}} \right), \quad (6.21)
\end{aligned}
$$

where λ_L and $\kappa_*(p,t,L)$ are defined by (6.19) and (6.20), respectively, and the stabilization parameter δ_L is given by (6.18). For $p = 1$ an identical bound holds with $k = t = p = 1$, and with the factor $(1 + h_L |\log_2 h_L|^{d-1})^2$ replaced by 1 in the definitions of δ_L and λ_L.

Remark 13 We close with some remarks on Theorem 17 and on possible extensions of the results presented here. We begin by noting that, save for the polylogarithmic factors, the definition of δ_L and the structure of the error bound in the $|||\cdot|||_{\mathrm{SD}}$ norm are exactly the same as if had we used the full tensor-product finite element space $V^{L,p}_{(0)}$ instead of the sparse tensor-product space $\hat{V}^{L,p}_{(0)}$ (cf. Houston & Süli [14]). On the other hand, as we have commented earlier, through the use of the sparse space $\hat{V}^{L,p}_{(0)}$, (discounting the effect of $p \geq 1$ on the computational cost, since we are interested in h-version methods here with p fixed at a relatively low value) computational complexity has been reduced from $\mathcal{O}(2^{Ld})$ to $\mathcal{O}(2^L (\log_2 2^L)^{d-1})$. Hence, in comparison with a streamline-diffusion method based on the full tensor-product space, a substantial computational saving has been achieved at the cost of only a marginal loss in accuracy.

a) In the diffusion-dominated case, that is when $|a| \approx 1$ and $|b| \approx 0$, we see from Theorem 17 that the error, in the streamline-diffusion norm $|||\cdot|||_{\mathrm{SD}}$, is $\mathcal{O}(h_L^p |\log_2 h_L|^{d-1})$ as h_L tends to zero, provided that the streamline-diffusion parameter is chosen as

$$
\delta_L = K_\delta \frac{h_L^2}{12 d p^4 |\sqrt{a}|^2 (1 + h_L |\log_2 h_L|^{d-1})^2} \qquad \text{when } p \geq 2,
$$

and with an analogous definition of δ_L, but with the factor $(1+h_L |\log_2 h_L|^{d-1})^2$ replaced by 1, when $p = 1$. This asymptotic convergence rate, as $h_L \to 0$, is slower, by the polylogarithmic factor $|\log_2 h_L|^{d-1}$, than the optimal $\mathcal{O}(h_L^p)$ bound on the $\|\cdot\|_{\mathrm{H}^1(\Omega)}$ norm of the error in a standard sparse Galerkin finite element approximation of Poisson's equation on $\Omega = (0,1)^d$ with continuous piecewise polynomials of degree p.

b) In the transport-dominated case, that is when $|a| \approx 0$ and $|b| \approx 1$, we select

$$
\delta_L = K_\delta \frac{h_L |\log_2 h_L|^{\frac{d-1}{2}}}{|b|},
$$

so the error of the method, measured in the streamline-diffusion norm, is $\mathcal{O}(h_L^{p+\frac{1}{2}} |\log_2 h_L|^{\frac{d-1}{2}})$ when the diffusivity matrix a degenerates to zero, — thus we see a loss of the size $\mathcal{O}(|\log_2 h_L|^{\frac{d-1}{2}})$ in comparison with the optimal $\mathcal{O}(h_L^{p+\frac{1}{2}})$ accuracy of the a classical streamline diffusion finite element approximation of a first-order scalar linear hyperbolic problem with continuous piecewise polynomials of degree p.

c) For the sake of simplicity, we have restricted ourselves to *uniform* tensor-product partitions of $[0,1]^d$. Numerical experiments indicate that, in the presence of boundary-layers, the accuracy of the proposed sparse streamline-diffusion method can be improved by using high-dimensional versions of Shishkin-type boundary-layer-fitted tensor-product nonuniform partitions.

d) When the matrix $a = (a_{ij})_{i,j=1}^d$ is positive definite, we have that $\Gamma_0 = \Gamma$ and therefore $u \in \mathrm{H}^1_{(0)}(\Omega) = \mathrm{H}^1_0(\Omega)$. Thus, in this case, $\kappa_*(p,p,L) < 1$ for all $p \geq 1$, $L \geq 1$.

The constant $(\kappa_*(p,p,L))^{d-1}$ appearing in (6.21) then converges to zero as $d \to \infty$ for all $p \geq 1$ and all $L \geq 1$. As long as the basis of the univariate space from which the sparse finite element space is constructed is a hierarchical basis on a uniform mesh, its specific choice (viz. whether it is a wavelet basis as in [23], or a standard hierarchical finite element basis) does not affect our final result. Thus we believe that the presence of the exponentially decreasing factor $(\kappa_*(p,p,L))^{d-1}$ is generic, and will be observed for error bounds in various norms. Note that the smallness of $\kappa_*(p,p,L)$ does *not* require particularly high regularity of u as expressed by the parameter $t = p$; in particular $\kappa_*(p,p,L) < 1$ for all $L \geq 1$, once $p \geq 2$ and, by (6.20), (5.35) and (5.36), also for $p = 1$ and $L \geq 3$.

e) It is important to note that the stabilization term

$$\delta_L \sum_{\kappa \in \mathcal{T}^L} (\mathcal{L}w, b \cdot \nabla v)_\kappa$$

in the definition of the bilinear form $b_\delta(w,v)$ can be rewritten as

$$\delta_L \sum_{i=1}^{d} \sum_{j=1}^{L} \left(a_{ii} \frac{\partial^2 w}{\partial x_i^2}, b \cdot \nabla v \right)_{K_j^i} + \delta_L \sum_{\substack{i=1,j=1 \\ i \neq j}}^{d} \left(a_{ij} \frac{\partial^2 w}{\partial x_i \partial x_j}, b \cdot \nabla v \right) + \delta_L (b \cdot \nabla w + cw, b \cdot \nabla v).$$

Here K_j^i, $i = 1, \ldots, d$, $j = 1, \ldots, L$, are the d-dimensional slabs defined in (6.5). Thus, instead of summing over $|\mathcal{T}^L| = 2^{Ld}$ entries we can realize the computation of the stabilization term by summing over $Ld + \frac{1}{2}d(d-1) + 1$ terms only.

f) For technical details concerning the efficient implementation of sparse-grid finite element methods, we refer to Zumbusch [32] and Bungartz & Griebel [7].

Acknowledgement. We wish to express our sincere gratitude to Adri Olde Daalhuis (University of Edinburgh). His helpful comments about the asymptotic properties of the Gauss hypergeometric function provided crucial hints for the proof of Lemma 13. We are also grateful to Christoph Ortner (University of Oxford) for suggesting a shortcut in an earlier version of a proof of Theorem 3.

References

[1] M. Abramowitz and I.A. Stegun (1972), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover Publications, New York.

[2] K. Babenko (1960), Approximation by trigonometric polynomials is a certain class of periodic functions of several variables, *Soviet Math, Dokl.* **1**, 672–675. Russian original in *Dokl. Akad. Nauk SSSR* **132**, 982–985.

[3] J.W. Barrett, C. Schwab and E. Süli (2005), Existence of global weak solutions for some polymeric flow models, *M3AS: Mathematical Models and Methods in Applied Sciences*, 6(15).

[4] J.W. Barrett and E. Süli (2006), Existence of global weak solutions to kinetic models of dilute polymers, *SIAM Multiscale Modelling and Simulation.* (Submitted for publication).

[5] R.F. Bass (1997), *Diffusion and Elliptic Operators*, Springer–Verlag, New York.

[6] T.S. Blyth and E.F. Robertson (2002), *Further Linear Algebra*, Springer–Verlag, London.

[7] H.-J. Bungartz and M. Griebel (2004), Sparse grids, *Acta Numerica*, 1–123.

[8] R. DeVore, S. Konyagin and V. Temlyakov (1998), Hyperbolic wavelet approximation, *Constr. Approx.* **14**, 1–26.

[9] J. Elf, P. Lötstedt and P. Sjöberg (2003), Problems of high dimension in molecular biology, *Proceedings of the* 19$^{\text{th}}$ *GAMM-Seminar Leipzig* (W. Hackbusch, ed.), 21–30.

[10] M. Griebel, Sparse grids and related approximation schemes for higher dimensional problems. *Foundations of Computational Mathematics 2005*, Luis-Miguel Pardo, Allan Pinkus, Endre Süli, Michael Todd, editors. Cambridge University Press 2006, pp. 106—161.

In Proceedings of the conference on Foundations of Computational Mathematics (FoCM05), Santander, Spain, 2005, London Mathematical Society, London (2006).

[11] V.H. Hoang and C. Schwab (2005), High dimensional finite elements for elliptic problems with multiple scales, *Multiscale Modeling and Simulation: A SIAM Interdisciplinary Journal* **3**(1), 168–194.

[12] L. Hörmander (2005), *The Analysis of Linear Partial Differential Operators II: Differential Operators with Constant Coefficients*, Reprint of the 1983 edition, Springer–Verlag, Berlin.

[13] P. Houston, C. Schwab and E. Süli (2002), Discontinuous *hp*-finite element methods for advection-diffusion-reaction problems, *SIAM Journal of Numerical Analysis* **39**(6), 2133–2163.

[14] P. Houston and E. Süli (2001), Stabilized *hp*-finite element approximation of partial differential equations with non-negative characteristic form, *Computing.* **66**(2), 99–119.

[15] N.G. van Kampen (1992), *Stochastic Processes in Physics and Chemistry*, Elsevier, Amsterdam.

[16] B. Lapeyre, É. Pardoux and R. Sentis (2003), *Introduction to Monte-Carlo Methods for Transport and Diffusion Equations*, Oxford Texts in Applied and Engineering Mathematics, Oxford University Press, Oxford.

[17] P. Laurençot and S. Mischler (2002), The continuous coagulation fragmentation equations with diffusion, *Arch. Rational Mech. Anal.* **162**, 45–99.

[18] C. Le Bris and P.-L. Lions (2004), Renormalized solutions of some transport equations with $\mathrm{W}^{1,1}$ velocities and applications, *Annali di Matematica* **183**, 97–130.

[19] E. Novak and K. Ritter (1998), The curse of dimension and a universal method for numerical integration, in *Multivariate Approximation and Splines* (G. Nürnberger, J. Schmidt and G. Walz, eds), International Series in Numerical Mathematics, Birkhäuser, Basel, 177–188.

[20] A.B. Olde Daalhuis (2003), Uniform asymptotic expansions for hypergeometric functions with large parameters. I, *Anal. Appl. (Singap.)* **1**, 111–120.

[21] O.A. Oleĭnik and E.V. Radkevič (1973), *Second Order Equations with Nonnegative Characteristic Form*. American Mathematical Society, Providence, RI.

[22] H.-C. Öttinger (1996), *Stochastic Processes in Polymeric Fluids*, Springer-Verlag, New York.

[23] T. von Petersdorff and C. Schwab (2004), Numerical solution of parabolic equations in high dimensions, *M2AN Mathematical Modelling and Numerical Analysis* **38**, 93–128.

[24] H.-G. Roos, M. Stynes, and L. Tobiska (1996), *Numerical Methods for Singularly Perturbed Differential Equations. Convection–Diffusion and Flow Problems*, Volume 24 of *Springer Series in Computational Mathematics*. Springer–Verlag, New York.

[25] C. Schwab (1998), *p- and hp- Finite Element Methods: Theory and Applications in Solid and Fluid Mechanics*. Numerical Metods and Scientific Computation. Clarendon Press, Oxford.

[26] S. Smolyak (1963), Quadrature and interpolation formulas for products of certain classes of functions, *Soviet Math. Dokl.* **4**, 240-243. Russian original in *Dokl. Akad. Nauk SSSR* **148**, 1042–1045.

[27] E. Süli (2005), Finite element approximation of high-dimensional transport-dominated diffusion problems, *Foundations of Computational Mathematics 2005*, Luis-Miguel Pardo, Allan Pinkus, Endre Süli, Michael Todd, editors. Cambridge University Press 2006, pp. 343-370. Available from: `http://web.comlab.ox.ac.uk/oucl/publications/natr/index.html`

[28] E. Süli (2006), Finite element algorithms for transport-diffusion problems: stability, adaptivity, tractability, Invited Lecture at the International Congress of Mathematicians, Madrid, 22-30 August 2006.
Available from: `http://web.comlab.ox.ac.uk/work/endre.suli/Suli-ICM2006.pdf`

[29] V. Temlyakov (1989), Approximation of functions with bounded mixed derivative, Volume 178 of *Proc. Steklov Inst. of Math.,* AMS, Providence, RI.

[30] G. Wasilkowski and H. Woźniakowski (1995), Explicit cost bounds of algorithms for multivariate tensor product problems, *J. Complexity* **11**, 1–56.

[31] C. Zenger (1991), Sparse grids, in *Parallel Algorithms for Partial Differential Equations* (W. Hackbusch, ed.), Vol. 31 of *Notes on Numerical Fluid Mechanics*, Vieweg, Braunschweig/Wiesbaden.

[32] G. W. Zumbusch (2000), A sparse grid PDE solver, in *Advances in Software Tools for Scientific Computing* (H. P. Langtangen, A. M. Bruaset, and E. Quak, eds.), Vol. 10 of *Lecture Notes in Computational Science and Engineering*, Ch. 4, 133–177. Springer, Berlin. (Proceedings SciTools '98).

Christoph Schwab
Seminar für Angewandte Mathematik, ETH Zürich, 8092 Zürich, Switzerland
`christoph.schwab@sam.math.ethz.ch`

Endre Süli
University of Oxford, Computing Laboratory, Wolfson Building, Parks Road, Oxford OX1 3QD, UK
`Endre.Suli@comlab.ox.ac.uk`

Radu-Alexandru Todor
Seminar für Angewandte Mathematik, ETH Zürich, 8092 Zürich, Switzerland
`todor@math.ethz.ch`