Report no. 05/19

Finite element approximation of high-dimensional transport-dominated diffusion problems

Endre Süli

Presented as invited lecture under the title
"Computational multiscale modelling: Fokker–Planck equations and their numerical analysis"
at the Foundations of Computational Mathematics conference
in Santander, Spain, 30 June – 9 July, 2005

High-dimensional partial differential equations with nonnegative characteristic form arise in numerous mathematical models in science. In problems of this kind, the computational challenge of beating the exponential growth of complexity as a function of dimension is exacerbated by the fact that the problem may be transport-dominated. We develop the analysis of stabilised sparse finite element methods for such high-dimensional, non-self-adjoint and possibly degenerate partial differential equations.

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

Oxford University Computing Laboratory
Numerical Analysis Group
Wolfson Building
Parks Road
Oxford, England OX1 3QD

September, 2005

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,

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

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 $\partial^2 u/\partial x_i\,\partial x_j$, $i,j = 1,\ldots,d$. Given 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 associated matrix norm $|A| := (A : A)^{1/2}$ is called the Frobenius norm of A.

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 [18]).

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

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

With additional 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 [13] (e.g. in the kinetic theory of polymers [19] and coagulation-fragmentation problems [15]), molecular biology [7], 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 well-known result of Hörmander [9] (cf. Theorem 11.1.10 on p.67), second-order hypoelliptic operators have nonnegative characteristic form at each point of the domain Ω, after possible multiplication by -1, so they too fall within 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. [10] and [11]). In particular, there has been no research to date on the numerical analysis of high-dimensional partial differential equations with nonnegative characteristic form.

The field of stochastic analysis is a particularly fertile source of equations of this kind (cf. [4]): 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 $\{\mathcal{F}_t\,,\ t \geq 0\}$, $b \in \mathrm{C}_b^1(\mathbb{R}^d; \mathbb{R}^d)$ is the drift vector, and $\sigma \in \mathrm{C}_b^2(\mathbb{R}^d, \mathbb{R}^{d\times p})$ is the diffusion matrix. Here $\mathrm{C}_b^k(\mathbb{R}^n, \mathbb{R}^m)$ 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 [14]). Thus, ψ is the solution of the initial-value problem

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

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[1].

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 nonnegative characteristic form. The paper is structured as follows. In order to provide a physical motivation for the mathematical questions considered here, we begin by presenting an example of a high-dimensional transport-dominated diffusion problem which arises from the kinetic theory of dilute polymers. We shall also explain briefly why such high-dimensional transport-dominated diffusion problems present a computational challenge. We shall then state in Section 3 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 4 is devoted to the construction of a hierarchical finite element space for univariate functions. The tensorisation of this space and the subsequent sparsification of the resulting tensor-product space are described in Section 5; our chief objective is to reduce the computational complexity of the discretisation without adversely effecting the approximation properties of the finite element space. In Sections 6 and 7 we build a stabilised finite element method over the sparse tensor product space, and we explore its stability and convergence.

The origins of sparse tensor product constructions and hyperbolic cross spaces can be traced back to the works of Babenko [1] and Smolyak [22]; we refer to the papers of Temlyakov [24], DeVore, Konyagin & Temlyakov [6] for the study of high-dimensional approximation problems, to the works of Wasilkowski & Woźniakowski [25] and Novak & Ritter [17] for high-dimensional integration problems and associated complexity questions, to the paper of Zenger [26] for an early contribution to the numerical solution of high-dimensional partial differential equations, to the articles by Hoang & Schwab [8] and von Petersdorff & Schwab [20] for the analysis of sparse-grid methods for high-dimensional elliptic multiscale problems and parabolic equations, and to the recent Acta Numerica article of Bungartz & Griebel [5] for an extensive survey of the field of sparse-grid methods.

[1] After the physicists Adriaan Daniël Fokker (1887–1972) and Max Planck (1858–1947).

2 An example from the kinetic theory of polymers

We present an example of a high-dimensional partial differential equation with nonnegative characteristic form which originates from the kinetic theory of dilute polymeric fluids. The fluid is assumed to occupy a domain $O \subset \mathbb{R}^n$; for physical reasons, $n = 2$ or $n = 3$ here.

There is a hierarchy of mathematical models that describe the evolution of the flow of a dilute polymer, the complexity of the model being dependent on the level of model-reduction (coarse-graining) that has taken place. The simplest model of this kind to account for noninteracting polymer chains is the so-called dumbbell model where each polymer chain which is suspended in the viscous incompressible Newtonian solvent whose flow-velocity is $u(x,t)$, $x \in O$, $t \in [0,T]$, is modelled by a dumbbell; a dumbbell consists of two beads connected by an elastic spring. At time $t \in [0,T]$ the dumbbell is characterised by the position of its centre of mass $X(t) \in \mathbb{R}^d$ and its elongation vector $Q(t) \in \mathbb{R}^d$. When a dumbbell is placed into the given velocity field $u(x,t)$, three forces act on each bead: the first force is the drag force proportional to the difference between the bead velocity and the velocity of the surrounding fluid particles; the second force is the elastic force F due to the spring stiffness; the third force is due to thermal agitation and is modelled as Brownian noise.

On rescaling the elongation vector, Newton's equations of motion for the beads give rise to the following system of stochastic differential equations:

$$\mathrm{d}X(t) = u(X(t),t)\,\mathrm{d}t, \tag{2.1}$$

$$\mathrm{d}Q(t) = \left(\nabla_u(X(t),t)\,Q(t) - \frac{1}{2\lambda}\,F(Q(t))\right)\mathrm{d}t + \frac{1}{\sqrt{\lambda}}\,\mathrm{d}W(t), \tag{2.2}$$

where $W = (W_1,\ldots,W_n)^\top$ is an n-dimensional Wiener process, $F(Q)$ denotes the elastic force acting on the chain due to elongation, and the positive parameter $\lambda = \xi/(4H)$ characterises the elastic property of the fluid, with $\xi \in \mathbb{R}_{>0}$ denoting the drag coefficient and $H \in \mathbb{R}_{>0}$ the spring stiffness.

Let $(x,q,t) \mapsto \psi(x,q,t)$ denote the probability density function of the vector-valued stochastic process $(X(t),Q(t))$; thus, $\psi(x,q,t)|\mathrm{d}x|\,|\mathrm{d}q|$ represents the probability, at time $t \in [0,T]$, of finding the centre of mass of a dumbbell in the volume element $x + \mathrm{d}x$ and having the endpoint of its elongation vector within the volume element $q + \mathrm{d}q$. Let us suppose that the elastic force $F : D \subseteq \mathbb{R}^d \to \mathbb{R}^d$, $d = 2,3$, of the spring is defined through a (sufficiently smooth) potential $U : \mathbb{R}_{\geq 0} \to \mathbb{R}$ via

$$F(q) := U'(\tfrac{1}{2}|q|^2)\,q. \tag{2.3}$$

Then, the probability density function $\psi(x,q,t)$ of the stochastic process $(X(t),Q(t))$ defined by (2.1), (2.2) satisfies the Fokker–Planck equation

$$\frac{\partial\psi}{\partial t} + \nabla_x\cdot(u\psi) + \nabla_q\cdot\left((\nabla_x u)q\,\psi - \frac{1}{2\lambda}F(q)\psi\right) = \frac{1}{2\,\lambda}\,\Delta_q\psi, \tag{2.4}$$

for $x \in O$, $q \in D$ and $t \in (0,T]$. The equation is supplemented by the initial condition $\psi(x,q,0) = \psi_0(x,q) \geq 0$ and appropriate boundary conditions.

Due to the fact that, unlike (2.2), the differential equation (2.1) does not involve random effects, the Fokker–Planck equation (2.4) for the associated probability density function is a degenerate parabolic equation for $\psi(x,q,t)$ with no diffusion in the x-direction.

In order to complete the definition of the dumbbell model, we note that the velocity field u appearing in (2.4) and the pressure p of the solvent are, in turn, found from the incompressible Navier–Stokes equations

$$\begin{aligned}
\frac{\partial u}{\partial t} + (u\cdot\nabla_x)u - \nu\,\Delta_x u + \nabla_x p &= \nabla_x\cdot\tau, && \text{in } O\times(0,T],\\
\nabla_x\cdot u &= 0, && \text{in } O\times(0,T],\\
u &= 0, && \text{on } \partial O\times(0,T],\\
u(x,0) &= u_0(x), && x\in O,
\end{aligned}$$

where the elastic extra-stress tensor $\tau = \tau(\psi)$ is defined in terms of the probability density function ψ as follows:

$$\tau(\psi) := k\,\mu\,(C(\psi) - \rho(\psi)\,I).$$

Here $k,\ \mu\in\mathbb{R}_{>0}$ are, respectively, the Boltzmann constant and the absolute temperature, I is the unit $n\times n$ tensor, and

$$\begin{aligned}
C(\psi)(x,t) &:= \int_D \psi(x,q,t)\,U'(\tfrac{1}{2}|q|^2)\,q\,q^{\top}\,\mathrm{d}q,\\
\rho(\psi)(x,t) &:= \int_D \psi(x,q,t)\,\mathrm{d}q.
\end{aligned}$$

We refer to the recent paper of Barrett, Schwab & Süli [2] for theoretical results concerning the existence of a global weak solution to this coupled Fokker–Planck–Navier–Stokes problem; see also the work of Le Bris & Lions [16] on related transport(-diffusion) problems with nonsmooth transport fields.

The Fokker–Planck equation (2.4) is a partial differential equation with nonnegative characteristic form in $2n+1$ independent variables $x\in O\subset\mathbb{R}^n$, $q\in D\subset\mathbb{R}^n$ and $t\in(0,T]\subset\mathbb{R}_{>0}$. In order to provide a rough estimate of the computational complexity of a classical algorithm for the numerical solution of the equation (2.4) supplemented with an initial condition and suitable boundary conditions, let us suppose that the spatial domain is $O\times D = (-1/2,1/2)^{2n}$ and[2] that a standard continuous piecewise linear Galerkin finite element method is used on each time level over a uniform axiparallel spatial mesh. Let us further suppose that the mesh has the relatively coarse spacing $h = 1/64$ in each of the $2n$ spatial co-ordinate directions and that a simple one-step method (such as the forward or backward Euler scheme, or the Crank–Nicolson scheme) is used to evolve the discrete solution in time. Ignoring degrees of freedom that lie on the

[2]Here, for simplicity, we took $D = (-1/2,1/2)^n$, — a ball in $\mathbb{R}^n$ of radius $1/2$ in the ℓ^∞-norm. In fact, D is a ball in $\mathbb{R}^n$ in the ℓ^2-norm of a certain fixed radius $q_{\max}\le\infty$, the maximum admissible length of the elongation vector Q; $q_{\max}=\infty$ in the case of the so-called Hookean dumbbell model.

boundary of $O \times D$, we see that the resulting system of linear equations involves around $63^4 = 15752962 \approx 1.5 \times 10^7$ unknowns on each time level when $n = 2$ (i.e., $2n = 4$) and around $63^6 = 62523502209 \approx 6.2 \times 10^{10}$ unknowns on each time level when $n = 3$ (i.e., $2n = 6$). Even on such coarse meshes the number of degrees of freedom in the numerical approximation to the analytical solution in 4 and 6 dimensions is very large, and grows very rapidly (exponentially fast, in fact,) as a function of $d = 2n + 1$, the number of independent variables. In general, on a uniform mesh of size $h = 1/N$ in each of the $2n$ spatial co-ordinate directions, the number of unknowns per time level (counting only those that are internal to $O \times D$) is $(N-1)^{2n}$. Over a unit time interval, and using the Crank–Nicolson scheme with time step $k = h$, this amounts to a total of approximately $N(N-1)^{2n} = \mathcal{O}(N^d)$ unknowns.

In addition to being high-dimensional, the equation (2.4) exhibits the features of a first-order hyperbolic equation with respect to $x \in \mathbb{R}^n$ (when variation with respect to q is suppressed), and those of a second-order parabolic transport-diffusion equation with respect to $q \in \mathbb{R}^n$ (when variation with respect to x is suppressed).

Our objective in this paper is to explore the algorithmic implications of this unpleasant combination of high-dimensionality and transport-dominated diffusion. In particular, our aim is to develop purely deterministic numerical algorithms based on the Galerkin method for high-dimensional transport-dominated diffusion problems of the form (1.1).

Alternative, stochastic, or mixed deterministic-stochastic computational approaches which have been pursued in the literature employ the intimate connection between the Fokker–Planck equation satisfied by the probability density function and the system of stochastic differential equations which govern the evolution of the underlying stochastic process (see, for example, the monograph of Öttinger [19] and the survey paper by Jourdain, Le Bris & Lelièvre [12]).

2.1 The curse of dominant transport

Classical Galerkin methods comprise a class of stable, reliable and accurate techniques for the numerical approximation of diffusion-dominated problems typified by symmetric elliptic equations (viz. equation (1.1) in the special case when a is a symmetric positive definite matrix and $b = 0$). In this case, a Galerkin method for the numerical solution of the equation (1.1), supplemented with a suitable boundary condition, coincides with the Ritz method based on energy minimisation over a finite-dimensional subspace of the infinite-dimensional Hilbert space $\mathcal{H}$ containing the weak solution u to the boundary value problem. The energy-norm is simply the norm induced by the symmetric and coercive bilinear form associated with the weak formulation of the problem, which acts as an inner product on $\mathcal{H}$. The Galerkin approximation to u is then the best approximation to u in the energy norm from the finite-dimensional subspace. If, on the other hand, $b \neq 0$, then a Galerkin method for the numerical solution of an elliptic equation of the form (1.1) cannot be rephrased in the language of energy minimisation over a finite-dimensional space; nevertheless, it will supply an accurate approximation to u, as long as a 'dominates' b in a certain sense.

In a Galerkin finite element method the finite-dimensional subspace from which the approximate solution u_h is sought consists of continuous piecewise polynomial functions of a fixed degree p which are defined over a partition of a certain fixed 'granularity' $h > 0$ of the computational domain $\Omega \subset \mathbb{R}^d$. Suppose, for example, that $d = 1$, $\Omega = (0, 1)$, $p = 1$, $a \in \mathbb{R}_{>0}$, $b \in \mathbb{R}$, $c = 0$, $f \in \mathrm{C}[0, 1]$, $f \geq 0$ and $h = 1/N$, where $N \in \mathbb{N}_{>1}$; let us also suppose for the sake of simplicity that homogeneous Dirichlet boundary conditions are imposed on $\partial\Omega = \{0, 1\}$. As long as $a \geq \frac{1}{2}h|b|$ (i.e., provided that the transport-diffusion problem is diffusion-dominated relative to the finite element partition), the qualitative behaviour of u_h will be correct, in the sense that u_h will obey a maximum principle analogous to the one satisfied by the analytical solution u.

This favourable behaviour of the approximate solution u_h is completely lost in the transport-dominated regime, when $a < \frac{1}{2}h|b|$; for such h, u_h exhibits maximum-principle-violating oscillations on the scale of the mesh. The oscillations will be particularly prominent in the boundary layer located in the vicinity of one of the endpoints of the interval $[0, 1]$, i.e., at $x = 0$ when $b < 0$ and $x = 1$ when $b > 0$.

An analogous situation is observed in the multidimensional case. Suppose, for example, that $\Omega = (0, 1)^d$ with $d > 1$, $p = 1$ (i.e., continuous piecewise linear polynomials in d variables are used on a simplicial partition of $\overline{\Omega}$), $a = a^\top \in \mathbb{R}^{d\times d}$ is a positive definite matrix, $b \in \mathbb{R}^d$, $c = 0$, $f \in \mathrm{C}(\overline{\Omega})$ and $h \in \mathbb{R}_{>0}$ is a mesh-parameter measuring the granularity of the finite element mesh; again, we assume that a homogeneous Dirichlet boundary condition is imposed on $\partial\Omega$. When $|a| \ll h|b|$, maximum-principle-violating oscillations will be observed in the vicinity of boundary layers; the oscillations will extend into the interior of the computational domain along subcharacteristic curves (i.e., along the transport direction b). Of course, if the mesh parameter h is sufficiently reduced so that $h|b| \ll |a|$, then the numerical approximation u_h will recover its accuracy and will appear qualitatively correct. Unfortunately the reduction of the mesh-parameter h to this level may place unachievable demands on limited computational resources.

2.2 The curse of dimensionality

The computational complexity of a numerical algorithm for the approximate solution of a transport-dominated diffusion equation is particularly unfavourable when the problem is high-dimensional. If, for example, continuous piecewise polynomial finite element basis functions of degree p are used in d dimensions on a mesh of size h and u is sufficiently smooth, in the limit of $h \to 0$ and $p \to \infty$ the error $E = \|u - u_h\|_{\mathrm{L}^2(\Omega)}$ will exhibit the optimal asymptotic convergence rate: $E \asymp C_p(u)\,(h/(p+1))^{p+1}$, where $C_p(u) = \text{Const.}|u|_{\mathrm{H}^{p+1}(\Omega)}$. Now, when $|b|/|a| \gg 1$, $C_p(u) \asymp \text{Const.}(|b|/|a|)^{p+1/2}$. Hence, for a preset tolerance $\mathtt{TOL}$, the requirement that $E = \mathtt{TOL}$ translates into requiring that

$$\frac{h}{p+1} \asymp \text{Const.}\left((|a|/|b|)^{1-1/(2(p+1))}\mathtt{TOL}^{1/(p+1)}\right).$$

At the same time, the computational complexity of the numerical method will scale as $\text{Const.}((p+1)/h)^d$. In terms of $\mathtt{TOL}$ this then gives

$$\mathtt{Complexity} \asymp \text{Const.}\left((|b|/|a|)^{d(1-1/(2(p+1)))}\mathtt{TOL}^{-d/(p+1)}\right). \tag{2.5}$$

Exponential growth of computational complexity as a function of the dimension of the problem is referred to as the *curse of dimensionality*. It is clear from (2.5) that for a transport-dominated diffusion problem, where $|b|/|a| \gg 1$, the curse of dimensionality may be particularly harmful. The focus of the paper is precisely this unfavourable situation, when the curse of dimensionality is exacerbated by dominant transport.

3 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:

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

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 mutually opposite $(d-1)$-dimensional open 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 $\hat{\varphi} \subset \Gamma_-$. ■

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

$$\mathcal{L}u \equiv -a : \nabla\nabla u + b \cdot \nabla u + cu = f \quad \text{in } \Omega, \tag{3.1}$$

$$u = 0 \quad \text{on } \Gamma_0 \cup \Gamma_-. \tag{3.2}$$

Before stating the variational formulation of (3.1), (3.2), we note the following simple result.

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 $M\xi = 0$.*

Proof First suppose that $M = (m_{ij})$ is a diagonal matrix. Since M is positive semidefinite, it follows that $m_{ii} \geq 0$ for all $i = 1, \ldots, d$. Now since $0 = \xi^\top M \xi = m_{11}\xi_1^2 + \cdots + m_{dd}\xi_d^2$, we deduce that $m_{ii}\xi_i^2 = 0$ for all $i = 1, \ldots, d$; hence also $m_{ii}\xi_i = 0$ for all $i = 1, \ldots, d$, that is, $M\xi = 0$.

Now consider the general case. Since M is a symmetric matrix, it can be diagonalised: $M = S^\top \Lambda S$, where Λ is a positive semidefinite diagonal matrix. Further, $0 = \xi^\top M \xi = (S\xi)^\top \Lambda (S\xi) = \zeta^\top \Lambda \zeta$, with $\zeta = S\xi$. From the first part of the proof we deduce that $\Lambda\zeta = 0$, and therefore $M\xi = S^\top(\Lambda S\xi) = S^\top \Lambda \zeta = S^\top 0 = 0$. ■

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

$$a\nu = 0 \qquad \text{on } \Gamma \setminus \Gamma_0. \tag{3.3}$$

Let us suppose for a moment that (3.1), (3.2) 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 $\mathrm{H}^{1/2}(\partial\Omega)$. By virtue of (3.3),

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

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

$$\int_{\partial\Omega} \gamma_{\nu,\partial\Omega}(a\nabla u)\cdot\gamma_{0,\partial\Omega}(v)\mathrm{d}s = \int_{\Gamma} \gamma_{\nu,\partial\Omega}(a\nabla u)|_{\Gamma}\cdot\gamma_{0,\partial\Omega}(v)|_{\Gamma}\mathrm{d}s = 0 \qquad \forall v\in\mathcal{V}, \tag{3.4}$$

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 (3.1) by $v\in\mathcal{V}$ and integrating by parts, we find that

$$(a\nabla u,\nabla v) - (u,\nabla\cdot(bv)) + (cu,v) + \langle u,v\rangle_{+} = (f,v) \qquad \forall v\in\mathcal{V}, \tag{3.5}$$

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

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

with β signifying the Fichera function $b\cdot\nu$, as before. We note that in the transition to (3.5) 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 (3.4), 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 (3.2).

The form of (3.5) serves as motivation for the statement of the weak formulation of (3.1), (3.2) 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

$$\|w\|_{\mathcal{H}} := (w,w)_{\mathcal{H}}^{1/2}.$$

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

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

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 (3.1), (3.2) if

$$B(u,v) = L(v) \qquad \forall v\in\mathcal{V}. \tag{3.6}$$

The existence of a unique weak solution is guaranteed by the following theorem (cf. also Theorem 1.4.1 on p.29 of [18], or [11]).

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 (3.6) 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 continuous 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

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

Upon integrating by parts in the second term on the right–hand side we deduce that

$$\begin{aligned} B(v,v) &= (a\nabla v, \nabla v) + c\|v\|^2 + \tfrac{1}{2}\langle v, v\rangle_{\Gamma_- \cup \Gamma_+} \\ &\geq K_0 \|v\|^2_{\mathcal{H}} \qquad \forall v \in \mathcal{V}, \end{aligned} \tag{3.7}$$

where $K_0 = \min(c, \frac{1}{2}) > 0$ and $\|\cdot\| = \|\cdot\|_{\mathrm{L}^2(\Omega)}$. Hence

$$(v, T(v))_{\mathcal{H}} \geq K_0 \|v\|^2_{\mathcal{H}} \qquad \forall v \in \mathcal{V}. \tag{3.8}$$

Consequently, $T : v \mapsto T(v)$ 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 (3.8), $\|S(v)\|_{\mathcal{H}} \leq (1/K_0)\|v\|_{\mathcal{H}}$ for all $v \in \mathcal{R}(T)$, it follows that $S : \mathcal{R}(T) \to \mathcal{V}$ is a continuous linear operator; therefore, it can be extended to a continuous linear operator $\hat{S} : \hat{\mathcal{H}} \to \mathcal{H}$. Furthermore, since

$$|L(v)| \leq \|f\| \|v\|_{\mathcal{H}} \qquad \forall v \in \mathcal{H}, \tag{3.9}$$

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 $\hat{S}$,

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

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

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

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 $\mathcal{H}$. Indeed, all elements in $\mathcal{H}$ vanish on Γ_0. Hence, we deduce from Lemma 1 that

$$\bigotimes_{i=1}^{d} \mathrm{H}^1_{(0)}(0,1) \equiv \mathrm{H}^1_{(0)}(0,1) \otimes \cdots \otimes \mathrm{H}^1_{(0)}(0,1) \subset \mathcal{H}, \tag{3.10}$$

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 defined to be equal to $\mathrm{H}^1_0(0,1)$ if the co-ordinate direction Ox_i is orthogonal to a pair of $(d-1)$-dimensional open faces contained in the elliptic part Γ_0 of Γ; otherwise (i.e., when the direction Ox_i is orthogonal to a pair of $(d-1)$-dimensional open faces contained in the hyperbolic part $\Gamma \setminus \Gamma_0 = \Gamma_- \cup \Gamma_+$ of Γ), it is defined to be equal to $\mathrm{H}^1(0,1)$. Clearly, if φ and $\hat{\varphi}$ are a pair of $(d-1)$-dimensional open faces of Ω which are opposite each other (i.e., $\varphi \parallel \hat{\varphi}$), then there exists a unique $i \in \{1,\ldots,d\}$ such that the co-ordinate direction Ox_i is orthogonal to this pair of faces.

Next, we shall consider the discretisation of the problem (3.6). Motivated by the tensor product structure of the space on the left-hand side of the inclusion (3.10), we shall base our Galerkin discretisation on a finite-dimensional subspace of $\mathcal{H}$ which is the tensor product of univariate subspaces of $\mathrm{H}^1_{(0)}(0,1)$. Thus, we begin by setting up the necessary notation in the case of the univariate space $\mathrm{H}^1_{(0)}(0,1)$.

4 Univariate discretisation

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 $\mathcal{T}^\ell = \{I^\ell_j \,:\, j = 0,\ldots,2^\ell - 1\}$ by halving each of the intervals I^ℓ_j. We consider the finite-dimensional linear subspace $\mathcal{V}^\ell$ of $\mathrm{H}^1(0,1)$ consisting of all continuous piecewise polynomials of degree $p = 1$ on the partition $\mathcal{T}^\ell$. We also consider its subspace $\mathcal{V}^\ell_0 := \mathcal{V}^\ell \cap C_0[0,1] \subset \mathrm{H}^1_0(0,1)$ consisting of all continuous piecewise linear functions that vanish at both endpoints of the interval $[0,1]$.

The mesh size in the partition $\mathcal{T}^\ell$ is $h_\ell := 2^{-\ell}$ and we define $N^\ell_0 := \dim(\mathcal{V}^\ell_0)$. Clearly, $N^\ell_0 = 2^\ell - 1$ for $\ell \geq 0$. We define $M^\ell_0 := N^\ell_0 - N^{\ell-1}_0$, $\ell \geq 1$, and let $M^0_0 := N^0_0 = 0$. Analogously, we define $N^\ell := \dim(\mathcal{V}^\ell)$ and $M^\ell := N^\ell - N^{\ell-1}$ for $\ell \geq 1$, with $M^0 = N^0 = 2$. Then, $N^\ell = N^\ell_0 + 2 = 2^\ell + 1$ for all $\ell \geq 0$, and $M^\ell = M^\ell_0 = 2^{\ell-1}$, $\ell \geq 1$. In what follows, we shall not distinguish between M^ℓ_0 and M^ℓ for $\ell \geq 1$ and will simply write M^ℓ for both.

For $L \geq 1$ we consider the linearly independent set

$$\{\psi^\ell_j \,:\, j = 1,\ldots,M^\ell, \quad \ell = 1,\ldots,L\}$$

in $\mathcal{V}_0^L$, where, for $x \in [0,1]$,

$$\psi_j^\ell(x) := \left(1 - 2^\ell \left| x - \frac{2j-1}{2^\ell} \right| \right)_+, \quad j = 1, \ldots, 2^{\ell-1}, \quad \ell = 1, \ldots, L.$$

Clearly,

$$\mathcal{V}_0^L = \operatorname{span}\{\psi_j^\ell : j = 1, \ldots, M^\ell, \quad \ell = 1, \ldots, L\},$$
$$\operatorname{diam}(\operatorname{supp} \psi_j^\ell) \le 2.2^{-\ell}, \qquad j = 1, \ldots, M^\ell, \quad \ell = 1, \ldots, L.$$

Any function $v \in \mathcal{V}_0^L$ has the representation

$$v(x) = \sum_{\ell=1}^{L} \sum_{j=1}^{M^\ell} v_j^\ell \psi_j^\ell(x),$$

with a uniquely defined set of coefficients $v_j^\ell \in \mathbb{R}$.

For $L \ge 1$, we consider the $\mathrm{L}^2(0,1)$-orthogonal projector

$$P_0^L : \mathrm{L}^2(0,1) \to \mathcal{V}_0^L.$$

This has the following approximation property (cf. Brenner & Scott [3]):

$$\|v - P_0^L v\|_{\mathrm{H}^s(0,1)} \le \mathrm{Const.} 2^{-(2-s)L} \|v\|_{\mathrm{H}^2(0,1)}, \tag{4.1}$$

where $L \ge 1$, $s \in \{0,1\}$, and $v \in \mathrm{H}^2(0,1) \cap \mathrm{H}_0^1(0,1)$. In particular, $v = \lim_{L\to\infty} P_0^L v$ for all $v \in \mathrm{H}^2(0,1) \cap \mathrm{H}_0^1(0,1)$, where the limit is considered in the $\mathrm{H}^s(0,1)$-norm, $s \in \{0,1\}$.

In order to extend the construction to the multidimensional case, it is helpful to define the *increment spaces* $\mathcal{W}_0^\ell$, $\ell \ge 0$, as follows:

$$\begin{aligned} \mathcal{W}_0^0 &:= \mathcal{V}_0^0 = \{0\}, \\ \mathcal{W}^\ell &:= \operatorname{span}\{\psi_j^\ell : 1 \le j \le M^\ell\}, \qquad \ell \ge 1. \end{aligned}$$

With this notation, we can write

$$\mathcal{V}_0^\ell = \mathcal{V}_0^{\ell-1} \oplus \mathcal{W}^\ell, \qquad \ell \ge 1.$$

Therefore,

$$\mathcal{V}_0^\ell = \mathcal{W}_0^0 \oplus \mathcal{W}^1 \oplus \cdots \oplus \mathcal{W}^\ell = \mathcal{W}^1 \oplus \cdots \oplus \mathcal{W}^\ell, \qquad \ell \ge 1. \tag{4.2}$$

We proceed similarly for functions v which do not vanish at the endpoints of the interval $[0,1]$. Any $v \in \mathcal{V}^L$, $L \ge 1$, has the representation

$$v(x) = (1-x)v(0) + xv(1) + \sum_{\ell=1}^{L} \sum_{j=1}^{M^\ell} v_j^\ell \psi_j^\ell(x),$$

with a uniquely defined set of coefficients $v_j^\ell \in \mathbb{R}$. For $L \ge 0$ we shall write this expansion in compact form as

$$v(x) = \sum_{\ell=0}^{L} \sum_{j=1}^{M^\ell} v_j^\ell \psi_j^\ell(x),$$

where $\psi_1^0(x) = 1 - x$, $\psi_2^0(x) = x$, $v_1^0 = v(0)$ and $v_2^0 = v(1)$. Thus,

$$\mathcal{V}^L = \mathrm{span}\{\psi_j^\ell \,:\, j = 1, \ldots, M^\ell, \quad \ell = 0, \ldots, L\}, \qquad L \geq 0.$$

For $L \geq 0$ we consider the $\mathrm{L}^2(0,1)$-orthogonal projector[3]

$$P^L \,:\, \mathrm{L}^2(0,1) \to \mathcal{V}^L.$$

This has the following approximation property (cf. Brenner & Scott [3]):

$$\|v - P^L v\|_{\mathrm{H}^s(0,1)} \leq \mathrm{Const.} 2^{-(2-s)L} \|v\|_{\mathrm{H}^2(0,1)}, \tag{4.3}$$

where $L \geq 0$, $s \in \{0, 1\}$ and $v \in \mathrm{H}^2(0,1)$. In particular, $v = \lim_{L\to\infty} P^L v$ for all $v \in \mathrm{H}^2(0,1)$, where the limit is considered in the $\mathrm{H}^s(0,1)$-norm for $s \in \{0,1\}$.

This time, we define the increment spaces $\mathcal{W}^\ell$, $\ell \geq 0$, as follows:

$$\begin{aligned} \mathcal{W}^0 &:= \mathcal{V}^0 = \mathrm{span}\{1 - x, x\}, \\ \mathcal{W}^\ell &:= \mathrm{span}\{\psi_j^\ell \,:\, 1 \leq j \leq M^\ell\}, \qquad \ell \geq 1. \end{aligned}$$

Hence, we can write

$$\mathcal{V}^\ell = \mathcal{V}^{\ell-1} \oplus \mathcal{W}^\ell, \qquad \ell \geq 1.$$

Therefore,

$$\mathcal{V}^\ell = \mathcal{W}^0 \oplus \mathcal{W}^1 \oplus \cdots \oplus \mathcal{W}^\ell, \qquad \ell \geq 1. \tag{4.4}$$

5 Sparse tensor-product spaces

Now we return to the original multidimensional setting on $\Omega = (0,1)^d$ and consider the finite-dimensional subspace V_0^L of $\bigotimes_{i=1}^d \mathrm{H}^1_{(0)}(0,1)$ defined by

$$V_0^L := \bigotimes_{i=1}^d \mathcal{V}_{(0)}^L = \mathcal{V}_{(0)}^L \otimes \cdots \otimes \mathcal{V}_{(0)}^L, \tag{5.1}$$

where the i^{th} component $\mathcal{V}_{(0)}^L$ in this tensor product is chosen to be $\mathcal{V}_0^L$ if the co-ordinate axis Ox_i is orthogonal to a pair of $(d-1)$-dimensional open faces of Ω which belong to Γ_0, and $\mathcal{V}_{(0)}^L$ is chosen as $\mathcal{V}^L$ otherwise. In particular, if $a = 0$ and therefore $\Gamma_0 = \emptyset$, then $\mathcal{V}_{(0)}^L = \mathcal{V}^L$ for each component in the tensor product. Conversely, if a is positive definite, then $\Gamma_0 = \Gamma$ and therefore $\mathcal{V}_{(0)}^L = \mathcal{V}_0^L$ for each component of the tensor product.

[3]The choice of P^L is somewhat arbitrary; e.g., we could have defined $P^L \,:\, \mathrm{H}^1(0,1) \mapsto \mathcal{V}^L$ by $P^L v := I^0 v + P_0^L(v - I^0 v)$, where $(I^0 v)(x) = (1-x)v(0) + xv(1)$, and arrived at identical conclusions to those below. For example, (4.3) will follow from (4.1) on noting that $\|v - I^0 v\|_{\mathrm{H}^s(0,1)} \leq \mathrm{Const.}|v|_{\mathrm{H}^2(0,1)}$ for $s \in \{0,1\}$ and $v \in \mathrm{H}^2(0,1)$. In addition, this alternative projector has the appealing property:

$$P^L|_{\mathrm{H}^1_0(0,1)} = P_0^L \qquad \text{for all } L \geq 1.$$

In general, for $a \geq 0$ that is neither identically zero nor positive definite, $\mathcal{V}^L_{(0)} = \mathcal{V}^L_0$ for a certain number i of components in the tensor product, where $0 < i < d$, and $\mathcal{V}^L_{(0)} = \mathcal{V}^L$ for the rest.

Using the hierarchical decompositions (4.2) and (4.4), we have that

$$V^L_0 = \bigoplus_{|\ell|_\infty \leq L} \mathcal{W}^{\ell_1} \otimes \cdots \otimes \mathcal{W}^{\ell_d}, \qquad \ell = (\ell_1, \ldots, \ell_d), \tag{5.2}$$

with the convention that $\mathcal{W}^{\ell_i=0} = \{0\}$ whenever Ox_i is a co-ordinate direction that is orthogonal to a pair of $(d-1)$-dimensional open faces contained in Γ_0; otherwise, $\mathcal{W}^{\ell_i=0} = \mathrm{span}\{1 - x_i, x_i\}$.

The space V^L_0 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^L_0 with a lower-dimensional subspace $\hat{V}^L_0$ defined as follows:

$$\hat{V}^L_0 \;:=\; \bigoplus_{|\ell|_1 \leq L} \mathcal{W}^{\ell_1} \otimes \cdots \otimes \mathcal{W}^{\ell_d}, \qquad \ell = (\ell_1, \ldots, \ell_d), \tag{5.3}$$

again with the convention that $\mathcal{W}^{\ell_i=0} = \{0\}$ whenever Ox_i is a co-ordinate direction that is orthogonal to a pair of $(d-1)$-dimensional open faces contained in Γ_0; otherwise, $\mathcal{W}^{\ell_i=0} = \mathrm{span}\{1 - x_i, x_i\}$.

The space $\hat{V}^L_0$ is called a *sparse tensor product space*. It has

$$\dim(\hat{V}^L_0) = \mathcal{O}(2^L\, L^{d-1}) = \mathcal{O}(2^L (\log_2 2^L)^{d-1})$$

degrees of freedom, which is a considerably smaller number than

$$\dim(V^L_0) = \mathcal{O}(2^{Ld}) = \mathcal{O}(2^L (2^L)^{d-1}).$$

Let us consider the d-dimensional projector

$$P^L_{(0)} \cdots P^L_{(0)} \;:\; \bigotimes_{i=1}^{d} \mathrm{H}^1_{(0)}(0,1) \to \bigotimes_{i=1}^{d} \mathcal{V}^L_{(0)} = V^L_0,$$

where the i^{th} component $P^L_{(0)}$ is equal to P^L_0 if the co-ordinate direction Ox_i is orthogonal to a pair of $(d-1)$-dimensional open faces contained in Γ_0, and is equal to P^L otherwise.

Now, let

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

We also define

$$Q^\ell_0 = \begin{cases} P^\ell_0 - P^{\ell-1}_0, & \ell \geq 1, \\ P^0_0, & \ell = 0, \end{cases}$$

with the convention that $P^0_0 = 0$. Thus,

$$P^L_{(0)} = \sum_{\ell=0}^{L} Q^\ell_{(0)},$$

where $Q^\ell_{(0)} = Q^\ell_0$ when $P^\ell_{(0)} = P^\ell_0$ and $Q^\ell_{(0)} = Q^\ell$ when $P^\ell_{(0)} = P^\ell$.

Hence,

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

where $Q^{\ell_i}_{(0)}$ is equal to Q^ℓ_0 when the co-ordinate direction Ox_i is orthogonal to a pair of $(d-1)$-dimensional open faces in Γ_0, and equal to Q^ℓ otherwise.

The sparse counterpart $\hat{P}^L_0$ of the tensor-product projector $P^L_{(0)} \cdots P^L_{(0)}$ is then defined by truncating the index set $\{\ell : |\ell|_\infty \le L\}$ of the sum to $\{\ell : |\ell|_1 \le L\}$:

$$\hat{P}^L_0 := \sum_{|\ell|_1 \le L} Q^{\ell_1}_{(0)} \cdots Q^{\ell_d}_{(0)} : \bigotimes_{i=1}^d \mathrm{H}^1_{(0)}(0,1) \to \hat{V}^L_0, \qquad \ell = (\ell_1, \ldots, \ell_d),$$

where $Q^{\ell_i}_{(0)}$ is equal to Q^ℓ_0 when the co-ordinate direction Ox_i is orthogonal to a pair of $(d-1)$-dimensional open faces contained in Γ_0, and equal to Q^ℓ otherwise. In order to formulate the approximation properties of the projector $\hat{P}^L_0$, for $k \in \mathbb{N}_{\ge 1}$ we define the space $\mathcal{H}^k(\Omega)$ of functions with *square-integrable mixed* k^{th} *derivatives*

$$\mathcal{H}^k(\Omega) = \{v \in \mathrm{L}^2(\Omega) : \mathrm{D}^\alpha v \in \mathrm{L}^2(\Omega),\ |\alpha|_\infty \le k\}$$

equipped with the norm

$$\|v\|_{\mathcal{H}^k(\Omega)} = \left(\sum_{|\alpha|_\infty \le k} \|\mathrm{D}^\alpha v\|^2_{\mathrm{L}^2(\Omega)} \right)^{1/2}.$$

Now we are ready to state our main approximation result.

Proposition 4 *Suppose that* $u \in \mathcal{H}^2(\Omega) \cap \bigotimes_{i=1}^d \mathrm{H}^1_{(0)}(0,1)$. *Then, for* $s \in \{0,1\}$,

$$\|u - \hat{P}^L_0 u\|_{\mathrm{H}^s(\Omega)} \le \begin{cases} \mathrm{Const.} h_L^2 |\log_2 h_L|^{d-1} \|u\|_{\mathcal{H}^2(\Omega)}, & \text{if } s = 0, \\ \mathrm{Const.} h_L^{2-s} \|u\|_{\mathcal{H}^2(\Omega)}, & \text{if } s = 1, \end{cases} \tag{5.4}$$

where $h_L = 2^{-L}$.

Proof We follow the line of argument in the proof of Proposition 3.2 in the paper by von Petersdorff & Schwab (2004), suitably modified to accommodate our nonstandard function space $\mathcal{H}^2(\Omega) \cap \bigotimes_{i=1}^d \mathrm{H}^1_{(0)}(0,1)$ as well as the fact that the norm-equivalence properties in the $\mathrm{L}^2(0,1)$ and $\mathrm{H}^1(0,1)$ norms, employed for the wavelet basis therein, do not apply here.

In the one-dimensional case, on writing

$$Q^\ell_{(0)} u = (P^\ell_{(0)} u - u) + (u - P^{\ell-1}_{(0)} u), \qquad \ell = 1, 2, \ldots,$$

we deduce from the approximation properties of P^ℓ_0 and P^ℓ that, for $u \in \mathrm{H}^2(0,1) \cap \mathrm{H}^1_{(0)}(0,1)$,

$$\|Q^\ell_{(0)} u\|_{\mathrm{H}^s(0,1)} \le \mathrm{Const.} 2^{(s-2)\ell} \|u\|_{\mathrm{H}^2(0,1)}, \qquad \ell = 0, 1, \ldots, \tag{5.5}$$

where $s \in \{0,1\}$. We recall that

$$u = \lim_{L\to\infty} P^L_{(0)} u = \lim_{L\to\infty} \sum_{\ell=0}^{L} Q^\ell_{(0)} u = \sum_{\ell=0}^{\infty} Q^\ell_{(0)} u$$

and hence

$$u - P^L_{(0)} u = \sum_{\ell>L} Q^\ell_{(0)} u$$

for all $u \in \mathrm{H}^2(0,1) \cap \mathrm{H}^1_{(0)}(0,1)$, where the limits of the infinite series are considered in the $\mathrm{H}^s(0,1)$-norm, $s \in \{0,1\}$.

In the multidimensional case, we deduce from (5.5) that

$$\|Q^{\ell_1}_{(0)} \otimes \cdots \otimes Q^{\ell_d}_{(0)} u\|_{\mathrm{H}^s(\Omega)} \leq \mathrm{Const.} 2^{s|\ell|_\infty - 2|\ell|_1} \|u\|_{\mathcal{H}^2(\Omega)}$$

and

$$u - \hat{P}^L_0 u = \sum_{|\ell|_1 > L} Q^{\ell_1}_{(0)} \otimes \cdots \otimes Q^{\ell_d}_{(0)} u$$

for all $u \in \mathcal{H}^2(\Omega) \cap \bigotimes_{i=1}^d \mathrm{H}^1_{(0)}(0,1)$, where the limit of the infinite sum is considered in the $\bigotimes_{i=1}^d \mathrm{H}^s(0,1)$-norm, $s \in \{0,1\}$. Noting that for $\ell = (\ell_1, \ldots, \ell_d)$, such that $|\ell|_1 = m$,

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

we have that

$$\begin{aligned}
\|u - \hat{P}^L_0 u\|_{\mathrm{H}^s(\Omega)} &\leq \mathrm{Const.} \left(\sum_{|\ell|_1 > L} 2^{s|\ell|_\infty - 2|\ell|_1} \right) \|u\|_{\mathcal{H}^2(\Omega)} \\
&= \mathrm{Const.} \left(\sum_{m=L+1}^{\infty} \sum_{|\ell|_1 = m} 2^{s|\ell|_\infty - 2|\ell|_1} \right) \|u\|_{\mathcal{H}^2(\Omega)} \\
&= \mathrm{Const.}\ 2^{(s-2)L} \left(\sum_{m=L+1}^{\infty} 2^{(s-2)(m-L)} \sigma_m \right) \|u\|_{\mathcal{H}^2(\Omega)},
\end{aligned}$$

where $\sigma_m = \sum_{|\ell|_1 = m} 2^{s(|\ell|_\infty - m)}$.

For $s = 0$ we have $\sigma_m \leq \mathrm{Const.} m^{d-1}$, while for $s > 0$ the bound $\sigma_m \leq \mathrm{Const.}$ holds, independent of m; we refer to [23] for a detailed proof of this fact. The final forms of the inequalities (5.4) follow, with $2^{(s-2)L} = h_L^{2-s}$ and $L = |\log_2 h_L|$, on observing that $\sum_{m=L+1}^{\infty} 2^{(s-2)(m-L)} \sigma_m$ is bounded by $\mathrm{Const.} L^{d-1}$ when $s = 0$ and by a constant independent of L when $s > 0$. ■

Since the space $\mathcal{H}^k(\Omega)$ of functions of square-integrable mixed k^{th} derivatives is a proper subspace of the classical Sobolev space $\mathrm{H}^k(\Omega) = \{v \in \mathrm{L}^2(\Omega) : \mathrm{D}^\alpha v \in \mathrm{L}^2(\Omega),\ |\alpha|_1 \leq k\}$, Proposition 4 indicates that preserving the optimal approximation order $\mathcal{O}(h^{2-s})$ of the full tensor-product space V_0^L in the $\mathrm{H}^s(\Omega)$-norm, $s = 0, 1$, upon sparsification (with a mild polylogarithmic loss of $|\log_2 h_L|^{d-1}$ in the case of $s = 0$) comes at the expense of increased smoothness requirements on the function u which is approximated from the sparse tensor-product space $\hat{V}_0^L$.

6 Sparse stabilised finite element method

Having defined the finite-dimensional space $\hat{V}_0^L$ from which the approximate solution will be sought, we now introduce the remaining ingredients of our Galerkin method: a bilinear form $b_\delta(\cdot,\cdot)$ which approximates the bilinear form $B(\cdot,\cdot)$ from the weak formulation (3.6) of the boundary value problem (3.1), (3.2) and a linear functional $l_\delta(\cdot)$ which approximates the linear functional $L(\cdot)$ from (3.6).

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 $\delta_L \in [0, 1/c]$ 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 $\Omega = (0,1)^d$; there are 2^{Ld} such elements κ in $\mathcal{T}^L$, a number that grows exponentially with d. 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 stabilisation in the direction of subcharacteristics ('streamlines').

We also define the linear functional

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

and consider the finite-dimensional problem: find $u_h \in \hat{V}_0^L$ such that

$$b_\delta(u_h, v_h) = l_\delta(v_h) \qquad \forall v_h \in \hat{V}_0^L. \tag{6.1}$$

The idea behind the method (6.1) 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 [21].

It would have been more accurate to write u_{h_L} and v_{h_L} instead of u_h and v_h in (6.1). 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. Since $u_h|_\kappa$ is multilinear in each $\kappa \in \mathcal{T}^L$ and a is a constant matrix, it follows that

$$\nabla\cdot(a\nabla u_h)|_\kappa = a : \nabla\nabla u_h|_\kappa = 0.$$

Therefore,

$$b_\delta(u_h, v_h) = B(u_h, v_h) + \delta_L(b\cdot\nabla u_h + cu_h, b\cdot\nabla v_h)$$

for all $u_h, v_h \in \hat{V}_0^L$. We note in passing that this simplification of $b_\delta(\cdot,\cdot)$ over $\hat{V}_0^L \times \hat{V}_0^L$, in comparison with its original definition, has useful computational consequences: it shows

that it is not necessary to sum over the 2^{Ld} elements κ comprising the mesh $\mathcal{T}^L$ in the implementation of the method.

Clearly,

$$\begin{aligned} b_\delta(v_h, v_h) &= (a\nabla v_h, \nabla v_h) + (cv_h, v_h) + \delta_L \|b\cdot\nabla v_h\|^2 \\ &\quad + \frac{1}{2}\int_{\Gamma_+\cup\Gamma_-} |\beta||v_h|^2 \,\mathrm{d}s + \frac{1}{2}c\delta_L \int_\Gamma \beta |v_h|^2 \,\mathrm{d}s \\ &\geq (a\nabla v_h, \nabla v_h) + c\|v_h\|^2 + \delta_L\|b\cdot\nabla v_h\|^2 \\ &\quad + \frac{1}{2}(1+c\delta_L)\int_{\Gamma_+} |\beta||v_h|^2 \,\mathrm{d}s + \frac{1}{2}(1-c\delta_L)\int_{\Gamma_-} |\beta||v_h|^2 \,\mathrm{d}s, \end{aligned} \tag{6.2}$$

where we have made use of the facts that $\beta \leq |\beta|$ on Γ_- and $v_h|_{\Gamma_0} = 0$. Since (6.1) is a linear problem in a finite-dimensional linear space, (6.2) implies the existence and uniqueness of a solution u_h to (6.1) in $\hat{V}_0^L$.

Let us also note that

$$|l_\delta(v_h)| \leq \left(\frac{1}{c} + \delta_L\right)^{1/2} \|f\| \left(c\|v_h\|^2 + \delta_L\|b\cdot\nabla v_h\|^2\right)^{1/2} \tag{6.3}$$

for all $v_h \in \hat{V}_0^L$. On noting that, by hypothesis, $1 - c\delta_L \geq 0$ and combining (6.2) and (6.3) we deduce that

$$\begin{aligned} |||u_h|||_{\mathrm{SD}}^2 &:= (a\nabla u_h, \nabla u_h) + c\|u_h\|^2 + \delta_L\|b\cdot\nabla u_h\|^2 \\ &\quad + \frac{1}{2}(1+c\delta_L)\int_{\Gamma_+} |\beta||u_h|^2 \,\mathrm{d}s + \frac{1}{2}(1-c\delta_L)\int_{\Gamma_-} |\beta||u_h|^2 \,\mathrm{d}s \\ &\leq \left(\frac{1}{c} + \delta_L\right)\|f\|^2. \end{aligned}$$

Hence,

$$|||u_h|||_{\mathrm{SD}} \leq (2/c)^{1/2}\|f\|, \tag{6.4}$$

which establishes the stability of the method (6.1), for all $\delta_L \in [0, 1/c]$.

The next section is devoted to the error analysis of the method. We shall require the following multiplicative trace inequality.

Lemma 5 (Multiplicative trace inequality) *Let $\Omega = (0,1)^d$ where $d \geq 2$ and suppose that Γ_+ is the hyperbolic outflow part of Γ. Then,*

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

Proof We shall prove the inequality for $v \in \mathrm{C}^1(\overline{\Omega})$. For $v \in \mathrm{H}^1(\Omega)$ the result follows by density of $\mathrm{C}^1(\overline{\Omega})$ 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 $x = (x_1, x') \in (0,1) \times (0,1)^{d-1} = \Omega$,

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

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,

$$\int_{\Gamma_+} |v|^2\,\mathrm{d}s \leq (2d-1)\|v\|^2 + 4\|v\| \sum_{i=1}^d \|u_{x_i}\|.$$

Therefore,

$$\int_{\Gamma_+} |v|^2\,\mathrm{d}s \leq 2d\sqrt{2}\max\{1, \frac{2}{d^{1/2}}\}\|v\|\,\|v\|_{\mathrm{H}^1(\Omega)} \leq 4d\|v\|\,\|v\|_{\mathrm{H}^1(\Omega)}.$$

Hence the required result. ■

7 Error analysis

Our goal in this section is to estimate the size of the error between the analytical solution $u \in \mathcal{H}$ and its approximation $u_h \in \hat{V}_0^L$. We shall assume throughout that $f \in \mathrm{L}^2(\Omega)$ and that the corresponding solution u belongs to $\mathcal{H}^2(\Omega) \cap \bigotimes_{i=1}^d \mathrm{H}^1_{(0)}(0,1) \subset \mathcal{H}$. 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 $v_h \in \hat{V}_0^L \subset \mathcal{V}$. Hence we deduce from (3.6) the following *Galerkin orthogonality* property:

$$b_\delta(u - u_h, v_h) = 0 \qquad \forall v_h \in \hat{V}_0^L. \tag{7.1}$$

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

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

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

$$|||u - u_h|||_{\mathrm{SD}} \leq |||\eta|||_{\mathrm{SD}} + |||\xi|||_{\mathrm{SD}}. \tag{7.2}$$

We begin by bounding $|||\xi|||_{\mathrm{SD}}$. By (6.2) and (7.1), we have that

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

Therefore,

$$|||\xi|||_{\rm SD}^2 \leq |b_\delta(\eta,\xi)|. \tag{7.3}$$

Now since $\nabla\nabla(P_L u)|_\kappa = 0$ for each $\kappa \in \mathcal{T}^L$, we have that $\nabla\nabla\eta|_\kappa = \nabla\nabla u|_\kappa$ on each $\kappa \in \mathcal{T}^L$, and therefore

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

We shall estimate each of the terms I to VII in turn:

$$\begin{aligned} \mathrm{I} &\leq \left(|a|^{1/2}\|\nabla\eta\|\right)|||\xi|||_{\rm SD}, \\ \mathrm{II} &\leq \left(\delta_L^{-1/2}\|\eta\|\right)|||\xi|||_{\rm SD}, \\ \mathrm{III} &\leq \left(c^{1/2}\|\eta\|\right)|||\xi|||_{\rm SD}, \\ \mathrm{V} &\leq \left(\delta_L^{1/2}|a|\,|u|_{\mathrm{H}^2(\Omega)}\right)|||\xi|||_{\rm SD}, \\ \mathrm{VI} &\leq \left(\delta_L^{1/2}|b|\,\|\nabla\eta\|\right)|||\xi|||_{\rm SD}, \\ \mathrm{VII} &\leq \left(c\delta_L^{1/2}\|\eta\|\right)|||\xi|||_{\rm SD}. \end{aligned}$$

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:

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

where in the transition to the last line we used the multiplicative trace inequality from Lemma 5. Hence, by (7.3),

$$\begin{aligned} |||\xi|||_{\rm SD} \leq\ & |a|^{1/2}\|\nabla\eta\| + \delta_L^{-1/2}\|\eta\| + c^{1/2}\|\eta\| + (8d)^{1/2}|b|^{1/2}\|\eta\|^{1/2}\,\|\eta\|_{\mathrm{H}^1(\Omega)}^{1/2} \\ & +\delta_L^{1/2}|a||u|_{\mathrm{H}^2(\Omega)} + \delta_L^{1/2}|b|\|\nabla\eta\| + c\delta_L^{1/2}\|\eta\|. \end{aligned} \tag{7.4}$$

On selecting

$$\delta_L := K_\delta \min\left(\frac{h_L^2}{|a|}, \frac{h_L|\log_2 h_L|^{d-1}}{d|b|}, \frac{1}{c}\right), \tag{7.5}$$

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

$$|||\xi|||_{\mathrm{SD}}^2 \leq C(u)\left(|a|h_L^2 + \frac{h_L^4|\log_2 h_L|^{2(d-1)}}{\delta_L}\right),$$

where $C(u) := \mathrm{Const.}\|u\|^2_{\mathcal{H}^2(\Omega)}$, and Const. is a positive constant independent of h_L. An identical bound holds for $|||\eta|||_{\mathrm{SD}}$. Thus we arrive at the following error bound.

Theorem 6 *Suppose that $f \in \mathrm{L}^2(\Omega)$, $c > 0$ and $u \in \mathcal{H}^2(\Omega) \cap \mathcal{H}$. Then, the following bound holds for the error $u - u_h$ between the analytical solution u of (3.6) and its sparse finite element approximation $u_h \in \hat{V}_0^L$ defined by (6.1), with $L \geq 1$ and $h = h_L = 2^{-L}$:*

$$|||u - u_h|||_{\mathrm{SD}}^2 \;\leq\; C(u)\left(|a|h_L^2 + h_L^4|\log_2 h_L|^{2(d-1)}\max\left(\frac{|a|}{h_L^2},\, \frac{d|b|}{h_L|\log_2 h_L|^{d-1}}, c\right)\right),$$

with the streamline-diffusion parameter δ_L defined by the formula (7.5) and $C(u) = \mathrm{Const.}\|u\|^2_{\mathcal{H}^2(\Omega)}$ where Const. *is a positive constant independent of the discretisation parameter h_L.*

Remark

We close with some remarks on Theorem 6 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 we used the full tensor-product finite element space V_0^L instead of the sparse tensor product space $\hat{V}_0^L$ (cf. Houston & Süli (2001)). On the other hand, as we have commented earlier, through the use of the sparse space $\hat{V}_0^L$, 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 6 that the error, in the streamline-diffusion norm $|||\cdot|||_{\mathrm{SD}}$, is $\mathcal{O}(h_L|\log_2 h_L|^{d-1})$ as h_L tends to zero, provided that the streamline-diffusion parameter is chosen as

$$\delta_L = K_\delta \frac{h_L^2}{|a|}.$$

This asymptotic convergence rate, as $h_L \to 0$, is slower, by the polylogarithmic factor $|\log_2 h_L|^{d-1}$, than the $\mathcal{O}(h_L)$ 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 $\Omega = (0,1)^d$.

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

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

so the error of the method, measured in the streamline-diffusion norm, is $\mathcal{O}(h_L^{3/2}|\log_2 h_L|^{d-1})$ when the diffusivity matrix a degenerates to zero.

c) We have confined ourselves to finite element approximations based on tensor-product piecewise polynomials of degree $p = 1$ in each of the d co-ordinate directions. Extensions of our results to the case of tensor-product piecewise polynomials of a fixed degree $p \geq 1$ in each of the co-ordinate directions are possible, although the analysis is then considerably more technical, and will be presented in a forthcoming paper [23].

d) 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.

e) For technical details concerning the efficient implementation of sparse-grid finite element methods, we refer to Zumbusch [27] and Bungartz & Griebel [5].

Acknowledgements
I am grateful to Andrea Cangiani (Università di Pavia), Kathryn Gillow (University of Oxford), Max Jensen (Humboldt Universität, Berlin), Christoph Ortner (University of Oxford) and Christoph Schwab (ETH Zürich) for helpful and constructive comments.

References

[1] 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.

[2] 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).

[3] S.C. Brenner and L.R. Scott (2002), *The Mathematical Theory of Finite Element Methods*, 2nd Edition, Volume 15 of *Texts in Applied Mathematics*, Springer–Verlag, New York.

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

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

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

[7] 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.), pp. 21–30.

[8] 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.

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

[10] 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.

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

[12] B. Jourdain, C. Le Bris and T Lelièvre (2004), 'Coupling PDEs and SDEs: the illustrative example of the multiscale simulation of viscoelastic flows. Preprint.

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

[14] 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.

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

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

[17] 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, pp. 177–188.

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

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

[20] T. von Petersdorff and C. Schwab (2004), 'Numerical solution of parabolic equations in high dimensions', M2AN Mathematical Modelling and Numerical Analysis **38**, 93–128.

[21] 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.

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

[23] E. Süli, C. Schwab, and R.-A. Todor (2005), 'Sparse finite element approximation of high-dimensional transport-dominated diffusion problems'. (In preparation).

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

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

[26] 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.

[27] 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, pp. 133–177. Springer, Berlin. (Proceedings SciTools '98).