

# Design and analysis of finite volume methods for elliptic equations with oblique derivatives; application to Earth gravity field modelling

Jérôme Droniou\*

Matej Medla<sup>†</sup>

Karol Mikula<sup>‡</sup>

August 12, 2019

## Abstract

We develop and analyse finite volume methods for the Poisson problem with boundary conditions involving oblique derivatives. We design a generic framework, for finite volume discretisations of such models, in which internal fluxes are not assumed to have a specific form, but only to satisfy some (usual) coercivity and consistency properties. The oblique boundary conditions are split into a normal component, which directly appears in the flux balance on control volumes touching the domain boundary, and a tangential component which is managed as an advection term on the boundary. This advection term is discretised using a finite volume method based on a centred discretisation (to ensure optimal rates of convergence) and stabilised using a vanishing boundary viscosity. A convergence analysis, based on the 3rd Strang Lemma [9], is conducted in this generic finite volume framework, and yields the expected  $\mathcal{O}(h)$  optimal convergence rate in discrete energy norm.

We then describe a specific choice of numerical fluxes, based on a generalised hexahedral meshing of the computational domain. These fluxes are a corrected version of fluxes originally introduced in [29]. We identify mesh regularity parameters that ensure that these fluxes satisfy the required coercivity and consistency properties. The theoretical rates of convergence are illustrated by an extensive set of 3D numerical tests, including some conducted with two variants of the proposed scheme. A test involving real-world data measuring the disturbing potential in Earth gravity modelling over Slovakia is also presented.

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Generic finite volume scheme</b>	<b>5</b>
2.1	Mesh, space of unknowns and interpolant . . . . .	5
2.2	Prolegomena to the scheme . . . . .	6
2.3	Scheme . . . . .	7
2.4	Error estimate . . . . .	8

---

\*School of Mathematics, Monash University, Melbourne (Australia), [jerome.droniou@monash.edu](mailto:jerome.droniou@monash.edu)

<sup>†</sup>Department of Mathematics, Faculty of Civil Engineering, Slovak University of Technology, Radlinskeho 11, 810 05 Bratislava, Slovak Republic, [medla@math.sk](mailto:medla@math.sk)

<sup>‡</sup>Department of Mathematics, Faculty of Civil Engineering, Slovak University of Technology, Radlinskeho 11, 810 05 Bratislava, Slovak Republic; Algoritmy:SK s.r.o., Sulekova 6, 81106 Bratislava, Slovak Republic, [mikula@math.sk](mailto:mikula@math.sk)

<b>3</b>	<b>Numerical tests</b>	<b>9</b>
3.1	Description of the scheme . . . . .	10
3.1.1	Inner fluxes . . . . .	10
3.1.2	Surface fluxes on $\Gamma$ . . . . .	12
3.1.3	Properties of the fluxes . . . . .	13
3.2	Alternative schemes . . . . .	15
3.3	Results . . . . .	16
3.3.1	Cubic domain and non-uniform mesh . . . . .	16
3.3.2	Tesseractoid domain with non-planar $\Gamma$ . . . . .	17
3.3.3	Spherical section domain with perturbed bottom $\Gamma$ . . . . .	21
3.3.4	Cubic domain, almost tangential vector field $\mathbf{V}$ . . . . .	21
3.4	Local gravity field modelling . . . . .	26
<b>4</b>	<b>Conclusion</b>	<b>27</b>
<b>A</b>	<b>Proof of Theorem 6</b>	<b>28</b>
A.1	Coercivity . . . . .	29
A.2	Consistency . . . . .	29
<b>B</b>	<b>Proof of Proposition 12</b>	<b>32</b>
B.1	Boundary fluxes . . . . .	32
B.2	Inner fluxes . . . . .	32
B.2.1	Coercivity . . . . .	32
B.2.2	Consistency . . . . .	36

# 1 Introduction

We consider in this work a Laplace equation with oblique boundary conditions:

$$-\Delta \bar{T}(\mathbf{x}) = 0, \quad \mathbf{x} \in \Omega \quad (1a)$$

$$\nabla \bar{T}(\mathbf{x}) \cdot \mathbf{V}(\mathbf{x}) = g(\mathbf{x}), \quad \mathbf{x} \in \Gamma \quad (1b)$$

$$\bar{T}(\mathbf{x}) = 0, \quad \mathbf{x} \in \partial\Omega \setminus \Gamma, \quad (1c)$$

where  $\Omega$  is a bounded domain in  $\mathbb{R}^3$  with piecewise  $C^2$  boundary,  $\Gamma$  is a relatively open subset of  $\partial\Omega$  that is fully contained in a smooth component of this boundary,  $g \in L^2(\partial\Omega)$  and  $\mathbf{V}$  is a  $C^1$  vector field such that  $\mathbf{V}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) \neq 0$  for all  $\mathbf{x} \in \Gamma$ . Here,  $\mathbf{n}$  denotes the outer normal to  $\partial\Omega$ . We also assume that the  $(d-1)$ -dimensional measure of  $\partial\Omega \setminus \Gamma$  is non-zero. On  $\Gamma$ ,  $\mathbf{V}$  can be decomposed into a normal and a tangential component to  $\Gamma$ . After renormalising  $g$  we can assume that the normal component is  $\mathbf{n}$ , and thus that

$$\mathbf{V}(\mathbf{x}) = \mathbf{n}(\mathbf{x}) + \mathbf{W}(\mathbf{x}), \quad \forall \mathbf{x} \in \Gamma. \quad (2)$$

The properties of  $\Gamma$  ensure that  $\mathbf{W}$  is a  $C^1$  tangential vector field on  $\Gamma$ .

A motivation to study boundary value problem (BVP) (1) comes from Earth gravity field modelling. The Earth gravity potential  $G$  fulfils outside the Earth a non-homogeneous elliptic equation

$$\Delta G(\mathbf{x}) = 2\omega^2, \quad (3)$$

where  $\omega$  is the spin velocity of the Earth [21]. The magnitude of the total gravity vector  $\nabla G$  is called gravity. If the measured gravity is prescribed on the Earth surface, i.e.

$$|\nabla G(\mathbf{x})| = \bar{g}(\mathbf{x}), \quad (4)$$

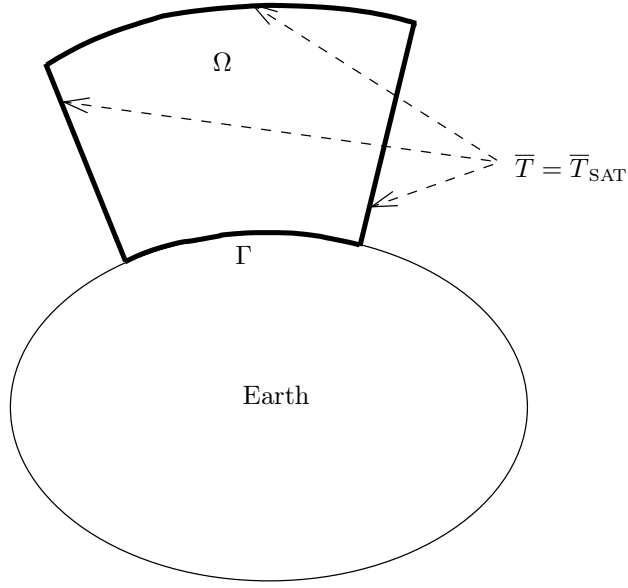


Figure 1: Illustration of the type of domains and boundary conditions that can be considered, using satellite data, in Earth gravity field simulation.

then Eq. (3) with BC (4) represents the so-called nonlinear geodetic BVP for the actual gravity potential  $G$ . The existence, uniqueness and other properties to the solution of this problem, and its variants, were studied extensively in physical geodesy community, see e.g. [1, 18, 24, 3, 17, 30, 20, 10, 11].

In Earth gravity field modelling, the actual gravity field  $G$  is usually expressed as a sum of the selected model field  $U$  and the remainder  $\bar{T}$ , i.e.

$$G(\mathbf{x}) = U(\mathbf{x}) + \bar{T}(\mathbf{x}). \quad (5)$$

If the model field  $U$  is generated by an ellipsoid with the same mass as the Earth, rotating with the same spin velocity  $\omega$  and with the constant surface potential equal to the geopotential  $W_0$  (see [31] for a definition of  $W_0$ , the potential  $G$  on the mean sea surface level), then  $U$  is called the normal gravity potential and  $\bar{T}$  is called the disturbing potential. This potential  $\bar{T}$  has no centrifugal component and it is generally accepted that the disturbing potential satisfies the Laplace equation  $\Delta \bar{T} = 0$  outside the Earth, see e.g. [21, 22]. In the satellite era, people have been able to consider a bounded domain  $\Omega$  outside the Earth where an upper part of the boundary is given as a sphere at altitude of a chosen satellite mission, and the bottom part  $\Gamma \subset \partial\Omega$  is given by a subset of the Earth surface [16, 29]. On this bottom part  $\Gamma$  the nonlinear BC (4) is given and, on the upper part, as well as on the side boundaries if one focuses on a tesseroid above the Earth, the Dirichlet-type BC obtained from satellite gravity missions can be prescribed. This allows us to fix a solution to the satellite data  $\bar{T}_{\text{SAT}}$ . See Figure 1 for an illustration.

Such nonlinear satellite-fixed geodetic BVP [28] for the disturbing potential  $\bar{T}$  can be formulated as follows

$$\Delta \bar{T}(\mathbf{x}) = 0, \quad \mathbf{x} \in \Omega, \quad (6)$$

$$|\nabla(\bar{T} + U)(\mathbf{x})| = \bar{g}(\mathbf{x}), \quad \mathbf{x} \in \Gamma, \quad (7)$$

$$\bar{T}(\mathbf{x}) = \bar{T}_{\text{SAT}}(\mathbf{x}), \quad \mathbf{x} \in \partial\Omega \setminus \Gamma. \quad (8)$$

Using the relation  $|\boldsymbol{\xi}| = \frac{\boldsymbol{\xi}}{|\boldsymbol{\xi}|} \cdot \boldsymbol{\xi}$  for  $\boldsymbol{\xi} \in \mathbb{R}^3 \setminus \{0\}$  we get (7) in the form

$$\frac{\nabla(\bar{T} + U)(\mathbf{x})}{|\nabla(\bar{T} + U)(\mathbf{x})|} \cdot \nabla(\bar{T} + U)(\mathbf{x}) = \bar{g}(\mathbf{x}).$$

Letting

$$\mathbf{v}(\mathbf{x}) = \frac{\nabla(\bar{T} + U)(\mathbf{x})}{|\nabla(\bar{T} + U)(\mathbf{x})|} \quad (9)$$

be the actual gravity vector  $\nabla G(\mathbf{x})$  unit direction, we can rewrite the nonlinear boundary condition (7) as follows

$$\mathbf{v}(\mathbf{x}) \cdot \nabla \bar{T}(\mathbf{x}) = \bar{g}(\mathbf{x}) - \mathbf{v}(\mathbf{x}) \cdot \nabla U(\mathbf{x}), \quad \mathbf{x} \in \Gamma. \quad (10)$$

Since the unit vector  $\mathbf{v}(\mathbf{x})$  is unknown and depends on  $\bar{T}(\mathbf{x})$ , boundary condition (10) is still nonlinear. However, if we set  $\bar{T}(\mathbf{x}) = 0$  in (9), which means that we approximate the unit direction  $\mathbf{v}(\mathbf{x})$  of the actual gravity vector by the unit direction of the normal gravity vector equal to  $\frac{\nabla U(\mathbf{x})}{|\nabla U(\mathbf{x})|}$ , we get a linear(ized) boundary condition

$$\mathbf{V}(\mathbf{x}) \cdot \nabla \bar{T}(\mathbf{x}) = \bar{g}(\mathbf{x}) - \bar{\gamma}(\mathbf{x}), \quad (11)$$

where  $\bar{\gamma}(\mathbf{x}) = \mathbf{V}(\mathbf{x}) \cdot \nabla U(\mathbf{x}) = \frac{\nabla U(\mathbf{x})}{|\nabla U(\mathbf{x})|} \cdot \nabla U(\mathbf{x}) = |\nabla U(\mathbf{x})|$  is the so-called normal gravity. Since all quantities depending on  $U$  are given analytically, the equation (11) represents a linear oblique derivative boundary condition. Together with equations (1a) and (1c), they are called the fixed gravimetric boundary value problem in the geodetic community [1, 24, 22, 23, 8, 16, 29] and give a basis for determining the Earth gravity field when gravity measurements are known on the Earth surface. When we denote  $\bar{g} - \bar{\gamma} = g$  and consider the problem on a bounded domain outside the Earth, we end up with the oblique derivative BVP (1) (here,  $\mathbf{V}$  has unit length and, as previously mentioned, the decomposition (2) is obtained after rescaling  $g$ ).

Let us briefly mention some results that can be found in the literature regarding the numerical approximation of second order equations with oblique derivative boundary conditions. In [5, 4] authors deal with the finite volume method for the oblique derivative boundary value problem in 2D case. In [5] they consider the oblique BC in the form

$$\bar{T}_n(\mathbf{x}) + (\alpha \bar{T})_t(\mathbf{x}) = g(\mathbf{x}), \quad (12)$$

where  $\alpha$  is a smooth function,  $\bar{T}_n$  is a derivative in the normal direction and  $(\alpha \bar{T})_t$  is a derivative in tangential direction. They develop a finite volume scheme based on the upwind principle, prove its convergence and obtain an error estimate of order  $\sqrt{h}$ . In [4], convergence results are established for the Poisson and a parabolic equation with oblique derivative boundary condition in which  $\alpha$  is constant. The convergence results are not only obtained for the approximate finite volume solutions, but also for their discrete gradients. The error estimate of order  $\sqrt{h}$  is obtained theoretically, but numerical experiments presented in these works indicate a first order rate of convergence. In [2] the authors present and analyse a 2D finite element method for the oblique derivative boundary problem, where the oblique derivative boundary is given by a graph of a real function. Finite volume methods for solving oblique derivative problems in 3D domains were suggested and numerically investigated in [26, 27, 25, 29]. These schemes are based either on upwind or central approximation of the oblique derivative. A numerical approximation of a nonlinear problem with eikonal-type boundary condition (4) was presented in [28]. This approximation is based on an iterative update of the oblique derivative condition.

In this paper we introduce and analyse novel numerical scheme for solving 3D oblique derivative boundary value problems for the Laplace equation. For the first time, there is presented a convergence analysis and error estimates for a finite volume scheme solving the oblique derivative problem in 3D. The model comes from the Earth gravity field modelling on real Earth topography but can be used in other applications as well. The presented numerical approach is general and covers various possible

discretisations of the Laplace equation inside the domain and it treats in a robust and stable way the oblique derivative boundary condition. We present numerical tests showing convergence properties of the novel scheme and compare them to further alternative numerical treatments of the oblique derivative. We also present a local Earth gravity field modelling for the region of Slovakia where we compare obtained numerical results with GPS-leveling measurements.

The paper is organised as follows. In Section 2 we present a generic finite volume method for solving the oblique derivative problem for the Laplace equation and its error estimates. In Section 3 we specify approximation of inner and surface fluxes and present results of numerical computations. We also give alternative schemes for oblique derivative treatment and discuss their pros and cons. In Section 4 we present concluding remarks. Appendix A contains proof of error estimate for the generic scheme and Appendix B contains proof of coercivity and consistency of the suggested inner flux approximation.

## 2 Generic finite volume scheme

We describe here a generic finite volume approximation of (1). The discretisation is based on a recasting of the model to transform the oblique derivative into a normal derivative, handled as a Neumann boundary condition, and a boundary advection–reaction term along  $\Gamma$ . The method is “generic” in the sense that we do not impose any specific expression of the numerical fluxes, only broad assumptions that enable the convergence analysis of the method. Our approach and analysis therefore cover many possible choices of Finite Volume methods for discretising the Laplacian in the domain.

### 2.1 Mesh, space of unknowns and interpolant

Let  $\mathfrak{T}$  be a partition of  $\Omega$  into “generalised” polyhedral finite volumes  $p$ , the generalisation coming from the fact that the faces of the polyhedra could be curved (especially those lying on  $\Gamma$ ). The mesh size is  $h := \max_{p \in \mathfrak{T}} \text{diam}(p)$ . We denote by  $\mathfrak{S}$  the set of faces of the mesh, and by  $\mathfrak{S}_{\text{int}}$  the faces contained in  $\Omega$ . The boundary faces are assumed to be compatible with  $\Gamma$  in the sense that each face in  $\mathfrak{S} \setminus \mathfrak{S}_{\text{int}}$  totally lies on  $\Gamma$ , or totally lies on the Dirichlet boundary  $\partial\Omega \setminus \Gamma$ . We let  $\mathfrak{S}_{\Gamma}$  be the set of faces on  $\Gamma$ , and  $\mathfrak{S}_{\text{Dir}} = \mathfrak{S} \setminus (\mathfrak{S}_{\text{int}} \cup \mathfrak{S}_{\Gamma})$  be the set of faces on  $\partial\Omega \setminus \Gamma$ .

For each cell  $p \in \mathfrak{T}$  we take a point  $\mathbf{x}_p \in p$  and we denote by  $\mathfrak{S}(p)$  the set of faces of  $p$ , so that  $\partial p = \cup_{\sigma \in \mathfrak{S}(p)} \sigma$ . If  $\sigma \in \mathfrak{S}(p)$ ,  $\mathbf{n}_{p,\sigma}$  is the unit outer normal to  $p$  on  $\sigma$ . Every face  $\sigma$  in  $\mathfrak{S}_{\Gamma}$  is a face of a unique finite volume  $p$ ; the dependency of  $p$  on  $\sigma$  is not made explicit as there is no risk of confusion in the formulas. We assume that:

$$\text{Each control volume } p \in \mathfrak{T} \text{ has at most one face } \sigma \text{ in } \mathfrak{S}_{\Gamma} \text{ and, in that case, } \mathbf{x}_p \in \sigma. \quad (13)$$

*Remark 1* (Assumption (13)). This assumption is not mandatory, and the design and analysis in the following sections could be adapted to meshes not satisfying (13) (see Remark 9); however, the method we consider in Section 3 naturally satisfies this property, which is why we assume it in our analysis.

For every  $p \in \mathfrak{T}$  and  $\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_{\text{int}}$ , we denote by  $q_p(\sigma)$  the finite volume such that  $\sigma = \overline{p} \cap \overline{q_p(\sigma)}$ ; here too, no risk of confusion arising we simply denote  $q$  for  $q_p(\sigma)$ . We then set  $d_{pq} = |\mathbf{x}_p - \mathbf{x}_q|$ . A face  $\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_{\text{Dir}}$  on the Dirichlet boundary of a cell  $p$  is sometimes considered as a “degenerate” cell, and also denoted by  $q$ ; for such faces, we pick a point  $\mathbf{x}_q \in \sigma$  and define again  $d_{pq} = |\mathbf{x}_p - \mathbf{x}_q|$ .

For each  $\sigma \in \mathfrak{S}_{\Gamma}$  we take  $\mathbf{x}_{\sigma} \in \sigma$ , and we denote by  $\mathfrak{E}(\sigma)$  the set of edges  $e$  of  $\sigma$ . The set of all such edges is  $\mathfrak{E}_{\Gamma} = \cup_{\sigma \in \mathfrak{S}_{\Gamma}} \mathfrak{E}(\sigma)$ , and the edges that lie in the relative interior of  $\Gamma$  are gathered in the set  $\mathfrak{E}_{\Gamma, \text{int}}$ . For  $e \in \mathfrak{E}(\sigma)$ ,  $\mathbf{n}_{\sigma,e}$  is the unit normal outward to  $\sigma$  on  $e$  in the tangent space of  $\Gamma$ .

If  $X$  is a control volume  $p$ , a face  $\sigma$  or an edge  $e$ ,  $|X|$  denotes the Lebesgue measure of  $X$  in the corresponding dimension of  $X$  (dimension 3 for a control volume, 2 for a face, 1 for an edge).

Our space of approximation has unknowns in the finite volumes, on the Dirichlet faces (“degenerate cells”), and on each edge on  $\Gamma$ , with zero values for Dirichlet faces, and for edges that are not in the

relative interior of  $\Gamma$ :

$$V_h := \{\varphi = ((\varphi_p)_{p \in \mathfrak{T}}, (\varphi_\sigma)_{\sigma \in \mathfrak{S}_{\text{Dir}}}, (\varphi_e)_{e \in \mathfrak{E}_\Gamma}) : \varphi_p \in \mathbb{R}, \varphi_\sigma = 0, \varphi_e \in \mathbb{R}, \varphi_e = 0 \text{ if } e \notin \mathfrak{E}_{\Gamma, \text{int}}\}.$$

*Remark 2.* Introducing the zero-valued unknowns is of course not necessary, but will be useful to simplify some expressions.

The norm on  $V_h$  is defined by

$$\|\varphi\|_{V_h} := (|\varphi|_{V_h, \Omega}^2 + h_\Gamma |\varphi|_{V_h, \Gamma}^2)^{1/2}, \quad (14a)$$

where  $h_\Gamma := \max_{\sigma \in \mathfrak{S}_\Gamma} \text{diam}(\sigma)$ ,

$$|\varphi|_{V_h, \Omega} := \left( \sum_{\sigma \in \mathfrak{S} \setminus \mathfrak{S}_\Gamma} \frac{|\sigma|}{d_{pq}} (\varphi_p - \varphi_q)^2 \right)^{1/2}, \quad (14b)$$

and

$$|\varphi|_{V_h, \Gamma} := \left( \sum_{\sigma \in \mathfrak{S}_\Gamma} \sum_{e \in \mathfrak{E}(\sigma)} \frac{|e|}{d_{pe}^\perp} (\varphi_p - \varphi_e)^2 \right)^{1/2} \quad (14c)$$

where  $d_{pe}^\perp$  is the orthogonal distance between  $\mathbf{x}_p$  (which belongs to  $\sigma$ ) and  $e$ . Remember that, in (14b),  $p$  and  $q$  are the two cells on each side of  $\sigma$  if  $\sigma \in \mathfrak{S}_{\text{int}}$ , and  $q = \sigma$  if  $\sigma \in \mathfrak{S}_{\text{Dir}}$  (so that  $\varphi_q = \varphi_\sigma = 0$  in that case). The term  $|\varphi|_{V_h, \Omega}$  can thus be viewed as a discrete  $H_0^1$ -(semi)norm in  $\Omega$  [13, Eq. (7.7f)], whilst  $|\varphi|_{V_h, \Gamma}$  plays the role of a discrete  $H_0^1$ -(semi)norm on the surface  $\Gamma$ . The presence of this boundary semi-norm, and its scaling by  $h_\Gamma$ , will be justified by the introduction of a small amount of diffusion on that surface to stabilise a centred approximation of an advective term on  $\Gamma$  stemming from the oblique boundary condition (see (19a)). Notice that, in (14c), Assumption (13) was used to identify the unknown on a face  $\sigma \in \mathfrak{S}_\Gamma$  with the value  $\varphi_p$  corresponding to  $p \in \mathfrak{T}$  such that  $\sigma \in \mathfrak{S}(p)$ .

The unknowns in the control volumes  $p$  are destined to be approximations of the solution at  $\mathbf{x}_p$ , whereas those on the boundary edges approximate the average value of the solution on the corresponding edge. This leads to defining the following interpolant  $I_h : C(\overline{\Omega}) \rightarrow V_h$ : for  $\varphi \in C(\overline{\Omega})$  such that  $\varphi = 0$  on  $\partial\Omega \setminus \Gamma$ ,

$$\begin{aligned} I_h \varphi &= ((\varphi_p)_{p \in \mathfrak{T}}, (\varphi_\sigma)_{\sigma \in \mathfrak{S}_{\text{Dir}}}, (\varphi_e)_{e \in \mathfrak{E}_\Gamma}) \text{ with} \\ \varphi_p &= \varphi(\mathbf{x}_p) \quad \forall p \in \mathfrak{T}, \quad \varphi_\sigma = \varphi(\mathbf{x}_q) \quad \forall q = \sigma \in \mathfrak{S}_{\text{Dir}}, \quad \varphi_e = \frac{1}{|e|} \int_e \varphi \quad \forall e \in \mathfrak{E}_\Gamma. \end{aligned} \quad (15)$$

The boundary condition  $\varphi = 0$  on  $\partial\Omega \setminus \Gamma$  ensures that  $\varphi_\sigma = 0$  for all  $\sigma \in \mathfrak{S}_{\text{Dir}}$ , and that  $\varphi_e = 0$  whenever  $e \notin \mathfrak{E}_{\Gamma, \text{int}}$ .

## 2.2 Prolegomena to the scheme

Integrating (1b) over a control volume  $p \in \mathfrak{T}$ , using Green's theorem and introducing  $\mathbf{W}$  defined in (2), it holds

$$\begin{aligned} 0 &= \iiint_p -\Delta \bar{T} = - \iint_{\partial p} \nabla \bar{T} \cdot \mathbf{n}_p \\ &= - \sum_{\sigma \in \mathfrak{S}(p) \setminus \mathfrak{S}_\Gamma} \iint_\sigma \nabla \bar{T} \cdot \mathbf{n}_{p, \sigma} - \sum_{\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_\Gamma} \iint_\sigma \nabla \bar{T} \cdot (\mathbf{n}_{p, \sigma} + \mathbf{W} - \mathbf{W}). \end{aligned}$$

Denoting by  $\bar{F}_{p,\sigma}(\bar{T}) = -\iint_{\sigma} \nabla \bar{T} \cdot \mathbf{n}_{p,\sigma} d\mathbf{x}$  the exact fluxes, we invoke the boundary condition (1b) to write

$$0 = \sum_{\sigma \in \mathfrak{S}(p) \setminus \mathfrak{S}_{\Gamma}} \bar{F}_{p,\sigma}(\bar{T}) - \sum_{\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_{\Gamma}} \iint_{\sigma} g - \nabla \bar{T} \cdot \mathbf{W}.$$

The vector field  $\mathbf{W}$  is tangential to  $\Gamma$  and thus only the tangential gradient of  $\bar{T}$  is involved in the quantity  $\nabla \bar{T} \cdot \mathbf{W}$ . We can therefore write  $\nabla \bar{T} \cdot \mathbf{W} = \nabla_{\Gamma} \cdot (\bar{T} \mathbf{W}) - \bar{T} \nabla_{\Gamma} \cdot \mathbf{W}$ , where  $\nabla_{\Gamma}$  is the divergence operator on the manifold  $\Gamma$ . This leads, using the divergence theorem on each face  $\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_{\Gamma}$ , to

$$\begin{aligned} \sum_{\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_{\Gamma}} \iint_{\sigma} g &= \sum_{\sigma \in \mathfrak{S}(p) \setminus \mathfrak{S}_{\Gamma}} \bar{F}_{p,\sigma}(\bar{T}) + \sum_{\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_{\Gamma}} \iint_{\sigma} \nabla_{\Gamma} \cdot (\bar{T} \mathbf{W}) - \bar{T} \nabla_{\Gamma} \cdot \mathbf{W} \\ &= \sum_{\sigma \in \mathfrak{S}(p) \setminus \mathfrak{S}_{\Gamma}} \bar{F}_{p,\sigma}(\bar{T}) + \sum_{\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_{\Gamma}} \sum_{e \in \mathfrak{E}(\sigma)} \int_e \bar{T} \mathbf{W} \cdot \mathbf{n}_{\sigma,e} - \sum_{\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_{\Gamma}} \iint_{\sigma} \bar{T} \nabla_{\Gamma} \cdot \mathbf{W}. \end{aligned}$$

Let us denote by  $[\bar{T} \mathbf{W} \cdot \mathbf{n}]_{\sigma,e} = \int_e \bar{T} \mathbf{W} \cdot \mathbf{n}_{\sigma,e}$  the exact advection fluxes on the boundary, and by  $[\bar{T} \nabla_{\Gamma} \cdot \mathbf{W}]_{\sigma} = \iint_{\sigma} \bar{T} \nabla_{\Gamma} \cdot \mathbf{W} d\mathbf{x}$  the other contribution (akin to a reaction term) to the boundary term. This shows that the solution to (1) satisfies, for all  $p \in \mathfrak{T}$ ,

$$\sum_{\sigma \in \mathfrak{S}(p) \setminus \mathfrak{S}_{\Gamma}} \bar{F}_{p,\sigma}(\bar{T}) + \sum_{\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_{\Gamma}} \sum_{e \in \mathfrak{E}(\sigma)} [\bar{T} \mathbf{W} \cdot \mathbf{n}]_{\sigma,e} - \sum_{\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_{\Gamma}} [\bar{T} \nabla_{\Gamma} \cdot \mathbf{W}]_{\sigma} = \sum_{\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_{\Gamma}} \iint_{\sigma} g. \quad (16)$$

### 2.3 Scheme

The scheme for (1) is obtained discretising (16). As previously mentioned, we will assume generic properties on the diffusive numerical fluxes. The advective contribution to the boundary terms is discretised using a centred scheme, to which we add a small amount of (boundary) diffusion for stabilisation purposes. As discussed in Remark 16, the choice of a centred discretisation seems crucial to prove optimal error estimates.

Based on our choice of unknowns and interpolant (15), we make the following approximation, in which  $T = ((T_p)_{p \in \mathfrak{T}}, (T_{\sigma})_{\sigma \in \mathfrak{S}_{\text{Dir}}}, (T_e)_{e \in \mathfrak{E}_{\Gamma}})$  is the sought approximation of  $\bar{T}$ :

$$[\bar{T} \mathbf{W} \cdot \mathbf{n}]_{\sigma,e} \approx T_e [\mathbf{W} \cdot \mathbf{n}]_{\sigma,e} \quad \text{and} \quad [\bar{T} \nabla_{\Gamma} \cdot \mathbf{W}]_{\sigma} \approx T_p [\nabla_{\Gamma} \cdot \mathbf{W}]_{\sigma}, \quad (17)$$

where  $[\mathbf{W} \cdot \mathbf{n}]_{\sigma,e} = \int_e \mathbf{W} \cdot \mathbf{n}_{\sigma,e}$  and  $[\nabla_{\Gamma} \cdot \mathbf{W}]_{\sigma} = \int_{\sigma} \nabla_{\Gamma} \cdot \mathbf{W}$ . Here, we used Assumption (13) to utilise  $T_p$  as approximate value of  $\bar{T}$  on  $\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_{\Gamma}$ . The exact fluxes  $\bar{F}_{p,\sigma}(\bar{T})$ , for  $p \in \mathfrak{T}$  and  $\sigma \in \mathfrak{S}(p) \setminus \mathfrak{S}_{\Gamma}$ , are discretised into numerical fluxes  $\mathcal{F}_{p,\sigma}^{\Omega}(T)$  that satisfy the following conservativity condition: for all  $\varphi \in V_h$  and all  $\sigma \in \mathfrak{S}_{\text{int}}$ ,

$$\mathcal{F}_{p,\sigma}^{\Omega}(\varphi) + \mathcal{F}_{q,\sigma}^{\Omega}(\varphi) = 0. \quad (18)$$

We also select numerical diffusion fluxes  $\mathcal{F}_{\sigma,e}^{\Gamma}(T)$  on the boundary, approximations of  $-\int_e \nabla_{\Gamma} \bar{T} \cdot \mathbf{n}_{\sigma,e}$  for  $\sigma \in \mathfrak{S}_{\Gamma}$  and  $e \in \mathfrak{E}(\sigma)$ .

The resulting finite volume scheme has the form: find  $T = ((T_p)_{p \in \mathfrak{T}}, (T_{\sigma})_{\sigma \in \mathfrak{S}_{\text{Dir}}}, (T_e)_{e \in \mathfrak{E}_{\Gamma}}) \in V_h$  such that:

$$\begin{aligned} \sum_{\sigma \in \mathfrak{S}(p) \setminus \mathfrak{S}_{\Gamma}} \mathcal{F}_{p,\sigma}^{\Omega}(T) + \sum_{\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_{\Gamma}} \sum_{e \in \mathfrak{E}(\sigma)} T_e [\mathbf{W} \cdot \mathbf{n}]_{\sigma,e} - \sum_{\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_{\Gamma}} T_p [\nabla_{\Gamma} \cdot \mathbf{W}]_{\sigma} \\ + Rh_{\Gamma} \sum_{\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_{\Gamma}} \sum_{e \in \mathfrak{E}(\sigma)} \mathcal{F}_{\sigma,e}^{\Gamma}(T) = \sum_{\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_{\Gamma}} \iint_{\sigma} g, \quad \forall p \in \mathfrak{T}, \end{aligned} \quad (19a)$$

where  $R \in (0, +\infty)$  will be adjusted later (see Remark 7), and

$$\mathcal{F}_{\sigma,e}^{\Gamma}(T) + \mathcal{F}_{\tau,e}^{\Gamma}(T) = 0, \quad \forall e \in \mathfrak{E}_{\Gamma, \text{int}} \text{ with } \sigma, \tau \in \mathfrak{S}_{\Gamma} \text{ the two faces on each side of } e. \quad (19b)$$

*Remark 3* (Conservativity of the fluxes). Because they correspond to a cell-centred finite volume method, the inner fluxes  $\mathcal{F}_{p,\sigma}^\Omega$  must satisfy by design the conservativity condition (18) on any vector  $\varphi \in V_h$ . On the contrary, the fluxes  $\mathcal{F}_{\sigma,e}^\Gamma$  correspond to a cell- and edge-centred method and their conservativity is only imposed on the solution to the finite volume scheme (see Equation (19b)). See also [9, Sections 3.3.1 and 3.3.3] on this topic.

## 2.4 Error estimate

The following assumptions are made on the diffusive fluxes.

**Assumption 4.** The numerical fluxes satisfy:

1. *Coercivity*: there is  $\rho_\Omega > 0$  and  $\rho_\Gamma > 0$  such that, for all  $\varphi \in V_h$ ,

$$\sum_{\sigma \in \mathfrak{S} \setminus \mathfrak{S}_\Gamma} \mathcal{F}_{p,\sigma}^\Omega(\varphi) (\varphi_p - \varphi_q) \geq \rho_\Omega |\varphi|_{V_h,\Omega}^2, \quad (20)$$

and

$$\sum_{\sigma \in \mathfrak{S}_\Gamma} \sum_{e \in \mathfrak{E}(\sigma)} \mathcal{F}_{\sigma,e}^\Gamma(\varphi) (\varphi_p - \varphi_e) \geq \rho_\Gamma |\varphi|_{V_h,\Gamma}^2. \quad (21)$$

2. *Consistency*: there exist constants  $C_{\text{cons}}^\Omega$  and  $C_{\text{cons}}^\Gamma$  such that, for all  $u \in C^2(\bar{\Omega})$  with  $u = 0$  on  $\partial\Omega \setminus \Gamma$ ,

$$\left| \mathcal{F}_{p,\sigma}^\Omega(I_h u) + \iint_\sigma \nabla u \cdot \mathbf{n}_{p,\sigma} \right| \leq C_{\text{cons}}^\Omega h |\sigma| \|u\|_{C^2(\bar{\Omega})}, \quad \forall p \in \mathfrak{T}, \forall \sigma \in \mathfrak{S}(p) \setminus \mathfrak{S}_\Gamma, \quad (22)$$

and

$$\left| \mathcal{F}_{\sigma,e}^\Gamma(I_h u) + \int_e \nabla u \cdot \mathbf{n}_{\sigma,e} \right| \leq C_{\text{cons}}^\Gamma h_\Gamma |e| \|u\|_{C^2(\bar{\Omega})}, \quad \forall \sigma \in \mathfrak{S}_\Gamma, \forall e \in \mathfrak{E}(\sigma). \quad (23)$$

The error estimate will be established under the assumption that the following mesh regularity factor remains bounded above:

$$\text{reg}_{\mathfrak{T}} = \max \left\{ \frac{\text{diam}(p)}{d_{p,\sigma}^\perp} : p \in \mathfrak{T}, \sigma \in \mathfrak{S}(p) \setminus \mathfrak{S}_\Gamma \right\} + \max \left\{ \frac{\text{diam}(p)}{\text{diam}(q)} : p \in \mathfrak{T}, \sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_{\text{int}} \right\}, \quad (24)$$

where  $d_{p,\sigma}^\perp$  is the orthogonal distance between  $\mathbf{x}_p$  and  $\sigma$ .

*Remark 5* (Interpretation of  $\text{reg}_{\mathfrak{T}}$ ). Bounding  $\text{reg}_{\mathfrak{T}}$  above imposes that each  $\mathbf{x}_p$  must be “well within” its cell  $p$ , and that two neighbouring cells must have comparable diameters (which does not prevent local refinement, provided that it is done in layers of smoothly refined meshes).

Combining [13, Lemmas B.21 and B.31], we obtain the following discrete trace inequality: there is  $C_{\text{tr}} > 0$  depending only on  $\Omega$ ,  $\Gamma$  and an upper bound of  $\text{reg}_{\mathfrak{T}}$  such that, for all  $\varphi \in V_h$ ,

$$\sum_{\sigma \in \mathfrak{S}_\Gamma} |\sigma| \varphi_p^2 \leq C_{\text{tr}} |\varphi|_{V_h,\Omega}^2. \quad (25)$$

In the rest of the paper, the notation  $a \lesssim b$  means that  $a \leq Cb$  for a constant  $C$  that is independent of the quantities in  $a$  and  $b$ , and of the mesh (but that may depend on  $\Omega$ ,  $\mathbf{W}$ ,  $\Gamma$ ,  $\rho_\Omega$ ,  $\rho_\Gamma$ ,  $C_{\text{cons}}^\Omega$ ,  $C_{\text{cons}}^\Gamma$ ,  $R$  and an upper bound of  $\text{reg}_{\mathfrak{T}}$ ). We can now state our main error estimate.

**Theorem 6** (Error estimate). *Under Assumption 4, suppose that  $\mathbf{W}$  satisfies*

$$\|(\nabla_\Gamma \cdot \mathbf{W})^+\|_{C(\Gamma)} < \frac{2\rho_\Omega}{C_{\text{tr}}}, \quad (26a)$$



where  $(\nabla_\Gamma \cdot \mathbf{W})^+ = \max(0, \nabla_\Gamma \cdot \mathbf{W})$  is the positive part of  $\nabla_\Gamma \cdot \mathbf{W}$ , and that  $R$  is chosen such that

$$R\rho_\Gamma > \frac{1}{2} \|\mathbf{W}\|_{C(\Gamma)^d}. \quad (26b)$$

Assume that the solution  $\bar{T}$  to (1) belongs to  $C^2(\bar{\Omega})$ , and let  $T$  be the solution to the scheme (19). Then,

$$\|T - I_h \bar{T}\|_{V_h} \lesssim h \|\bar{T}\|_{C^2(\bar{\Omega})}. \quad (27)$$

*Proof.* See Appendix A.  $\square$

*Remark 7* (About Assumption (26)). Assumption (26a) imposes a relative smallness only of the *positive* part of  $\nabla_\Gamma \cdot \mathbf{W}$ . In particular, this assumption holds if  $\nabla_\Gamma \cdot \mathbf{W} \leq 0$ . Assumption (26b) shows how the user-defined parameter  $R$  must be chosen to ensure the stability of the method.

*Remark 8* (Regularity assumption on  $\bar{T}$ ). In most situations, the  $C^2$  regularity on  $\bar{T}$  can be weakened to an  $H^2$  regularity, upon additional technicalities that we do not address here to simplify the exposition. See, e.g., [13, Section 7.4] for lemmas useful for establishing consistency estimates under  $H^2$ -regularity of the function.

*Remark 9* (Assumption (13)). In case Assumption (13) is not satisfied, that is the points  $\mathbf{x}_p$  corresponding to cells that touch  $\Gamma$  do not lie on  $\Gamma$ , the scheme has to be slightly modified the following way:

- Additional unknowns on the faces on  $\Gamma$  are introduced, so that  $V_h$  is changed into

$$V_h := \{\varphi = ((\varphi_p)_{p \in \mathfrak{T}}, (\varphi_\sigma)_{\sigma \in \mathfrak{S}_{\text{Dir}} \cup \mathfrak{S}_\Gamma}, (\varphi_e)_{e \in \mathfrak{E}_\Gamma}) : \varphi_p \in \mathbb{R}, \varphi_\sigma \in \mathbb{R}, \varphi_\sigma = 0 \text{ if } \sigma \in \mathfrak{S}_{\text{Dir}}, \\ \varphi_e \in \mathbb{R}, \varphi_e = 0 \text{ if } e \notin \mathfrak{E}_{\Gamma, \text{int}}\}.$$

A point  $\mathbf{x}_\sigma$  is chosen on each  $\sigma \in \mathfrak{S}_\Gamma$  and the interpolant (15) is extended by setting, for these faces,  $\varphi_\sigma = \varphi(\mathbf{x}_\sigma)$ .

- The seminorms  $|\cdot|_{V_h, \Omega}$  and  $|\cdot|_{V_h, \Gamma}$  are modified in the following way: in (14b) the sum is taken over  $\sigma \in \mathfrak{S}$  with  $\varphi_q = \varphi_\sigma$  whenever  $\sigma \in \mathfrak{S}_\Gamma$ ; in (14c),  $\varphi_p$  is replaced with  $\varphi_\sigma$ .
- Fluxes  $\mathcal{F}_{p, \sigma}^\Omega$  are also considered for  $\sigma \in \mathfrak{S}_\Gamma$  and the scheme consists in finding  $T \in V_h$  solution to the conservativity equations (19b) and

$$\sum_{\sigma \in \mathfrak{S}(p)} \mathcal{F}_{p, \sigma}^\Omega(T) = 0, \quad \forall p \in \mathfrak{T}, \quad (28)$$

$$- \mathcal{F}_{p, \sigma}^\Omega(T) + \sum_{e \in \mathfrak{E}(\sigma)} T_e [\mathbf{W} \cdot \mathbf{n}]_{\sigma, e} - T_\sigma [\nabla_\Gamma \cdot \mathbf{W}]_\sigma + Rh_\Gamma \sum_{e \in \mathfrak{E}(\sigma)} \mathcal{F}_{\sigma, e}^\Gamma(T) = \iint_\sigma g, \quad \forall \sigma \in \mathfrak{S}_\Gamma \quad (29)$$

where, in (29),  $p$  is the only cell that has  $\sigma$  as face.

- The coercivity assumption (20) is changed into

$$\sum_{\sigma \in \mathfrak{S} \setminus \mathfrak{S}_\Gamma} \mathcal{F}_{p, \sigma}^\Omega(\varphi)(\varphi_p - \varphi_q) + \sum_{\sigma \in \mathfrak{S}_\Gamma} \mathcal{F}_{p, \sigma}^\Omega(\varphi)(\varphi_p - \varphi_\sigma) \geq \rho_\Omega |\varphi|_{V_h, \Omega}^2.$$

The analysis performed in Appendix A can then be adapted and leads to the same error estimate (27).

### 3 Numerical tests

The numerical tests presented here are obtained using internal fluxes  $\mathcal{F}_{p, \sigma}^\Omega$  corresponding to a corrected version of the ones introduced in [29], and variants. For boundary fluxes  $\mathcal{F}_{\sigma, e}^\Gamma$ , used only for stabilisation purposes, we utilise the ones provided by the Hybrid Mimetic Mixed method [14].

### 3.1 Description of the scheme

#### 3.1.1 Inner fluxes

We consider a structured, but not necessarily Cartesian, grid of points on  $\Omega$ . These points are called representative points, as this is where we will look for an approximation of the potential  $\bar{T}$ . The structured grid assumption means that the representative points can be denoted by  $\mathbf{x}_{i,j,k}$ , where  $i \in \{0, \dots, I+1\}$ ,  $j \in \{0, \dots, J+1\}$ ,  $k \in \{0, \dots, K+1\}$ , and we assume that the extremal points (corresponding to  $i = 0$ ,  $i = I+1$ ,  $j = 0$ ,  $j = J+1$ ,  $k = 0$  or  $k = K+1$ ) lie on  $\partial\Omega$ . We split the set of indices of these extremal points into  $\mathcal{I}_\Gamma = \{(i, j, k) : k = 0\}$  and its complement  $\mathcal{I}_D$ , and we assume that  $\mathcal{I}_\Gamma$  corresponds to the points  $\mathbf{x}_{i,j,k} \in \Gamma$ , so that  $\mathcal{I}_D$  is the set of indices for the points on the Dirichlet boundary  $\partial\Omega \setminus \Gamma$ . The points associated with two extremal indices lie on the edges of  $\Omega$ , whereas those with three extremal indices describe the corners of  $\Omega$ . Note that  $\Omega$  is not necessarily a hexahedron since its “faces” may not be planar. We refer to Figs. 2 and 5 for illustrations.

For each  $(i, j, k) \notin \mathcal{I}_D$ , a hexahedral finite volume is constructed around  $\mathbf{x}_{i,j,k}$  using the following procedure. Note that points with indices in  $\mathcal{I}_D$  are not associated with control volumes, as they lie on the Dirichlet boundary and they are therefore not associated with unknowns of the scheme.

- If  $(i, j, k) \in \mathcal{I}_{\text{int}} := [2, I-1] \times [2, J-1] \times [2, K-1]$ , then setting  $A = \{(m, n, o) \in \{-1, 0, 1\}^3 : |m| + |n| + |o| = 3\}$  we define, for  $(m, n, o) \in A$ , the vertex  $\mathbf{x}_{i,j,k}^{m,n,o}$  as an average of the eight neighbouring points in the grid, one of them being  $\mathbf{x}_{i,j,k}$ :

$$\mathbf{x}_{i,j,k}^{m,n,o} = \frac{1}{8} \sum_{(a,b,c) \in B(m,n,o)} \mathbf{x}_{i+a,j+b,k+c}, \quad (30)$$

where  $B(m, n, o) = \{(m, n, o), (m, n, 0), (m, 0, o), (m, 0, 0), (0, n, o), (0, n, 0), (0, 0, o), (0, 0, 0)\}$ . The finite volume around  $\mathbf{x}_{i,j,k}$  is then the hexahedron (with possibly non-planar faces) defined by the vertices  $\{\mathbf{x}_{i,j,k}^{m,n,o} : (m, n, o) \in A\}$ . See Fig. 2 (left) for an illustration.

- If  $(i, j, k) \in \mathcal{I}_\Gamma$ , so that  $k = 0$ , and  $(i, j) \in [2, I-1] \times [2, J-1]$ , we construct four vertices  $\mathbf{x}_{i,j,k}^{m,n,1}$ , for  $(m, n, 1) \in A$ , as in (30). Four more vertices  $\mathbf{x}_{i,j,k}^{m,n,0}$  are constructed by averaging the four neighbouring vertices on  $\Gamma$ :

$$\mathbf{x}_{i,j,k}^{m,n,0} = \frac{1}{4} \sum_{(a,b) \in C(m,n)} \mathbf{x}_{i+a,j+b,0}, \quad (31)$$

where  $C(m, n) = \{(m, n), (m, 0), (0, n), (0, 0)\}$ . The control volume associated with  $\mathbf{x}_{i,j,0}$  is defined by the eight vertices thus constructed, and we notice that  $\mathbf{x}_{i,j,0}$  lies on one of its faces (the one on  $\Gamma$ ), so that (13) is satisfied.

- If  $(i, j, k) \notin (\mathcal{I}_{\text{int}} \cup \mathcal{I}_\Gamma)$ ,  $\mathbf{x}_{i,j,k}$  is associated with a control volume touching the Dirichlet boundary and built from four vertices constructed as in (30) and four other vertices constructed in a similar way as in (31), using representative points on the Dirichlet boundary  $\partial\Omega \setminus \Gamma$ . See Fig. 2 (right).
- A similar construction is made for the remaining indices  $(i, j, k)$ , corresponding to control volumes with an edge along an edge of  $\Omega$ , or a vertex at one of the corners of  $\Omega$ ; for example, the vertices of the control volumes lying on an edge of  $\Omega$  are constructed as the average of two representative points  $\mathbf{x}_{a,b,c}$  with two extremal indices. See Fig. 2 (right).

A generic finite volume is therefore identified by a triplet  $(i, j, k) \notin \mathcal{I}_D$ . For simplicity and to relate more to the unstructured notations used in Section 2, we denote  $(i, j, k)$  by  $p$ . The point  $\mathbf{x}_{i,j,k}$  associated with  $p$  is therefore denoted by  $\mathbf{x}_p$ . Any face  $\sigma \in \mathfrak{S}(p) \setminus \mathfrak{S}_\Gamma$  can be associated with two representative points on each side:  $\mathbf{x}_p$  itself, and  $\mathbf{x}_q$  which might either be associated with a genuine control volume if

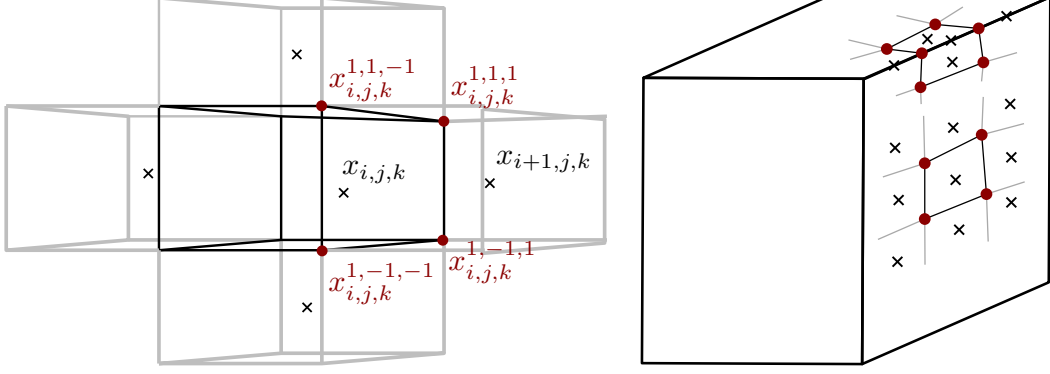


Figure 2: Illustration of the construction of an internal control volume (left), and of the faces and edges of boundary control volumes (right).

$\sigma \in \mathfrak{S}_{\text{int}}$ , or with  $\sigma$  itself (as a degenerate cell  $q$ ) if  $\sigma \in \mathfrak{S}_{\text{Dir}}$ . We write  $\sigma = \sigma_{pq}$  and notice that  $\sigma_{pq}$  may not be planar.

The four vertices  $(\mathbf{x}_{i,j,k}^{m,n,o})_{(m,n,o)}$  of  $\sigma_{pq}$  are ordered in a counterclockwise way, respective to the orientation compatible with the outer normal to  $p$ , and we denote them by  $\mathbf{x}_{pq}^{\oplus}$ ,  $\mathbf{x}_{pq}^{\boxplus}$ ,  $\mathbf{x}_{pq}^{\ominus}$ ,  $\mathbf{x}_{pq}^{\boxminus}$ . For  $\mathbf{x}_{pq}^*$  one of these vertices, we let  $\mathcal{R}(\mathbf{x}_{pq}^*)$  be the set of representative points involved in the construction of  $\mathbf{x}_{pq}^*$ ; hence, if  $\text{Card}(\mathcal{R}(\mathbf{x}_{pq}^*)) \in \{8, 4, 2, 1\}$  is the cardinality of  $\mathcal{R}(\mathbf{x}_{pq}^*)$ , we have

$$\mathbf{x}_{pq}^* = \frac{1}{\text{Card}(\mathcal{R}(\mathbf{x}_{pq}^*))} \sum_{\mathbf{y} \in \mathcal{R}(\mathbf{x}_{pq}^*)} \mathbf{y}. \quad (32)$$

We define four vectors related to the face  $\sigma_{pq}$ : the unit vector  $\mathbf{s}_{pq}$  which points from  $\mathbf{x}_p$  to  $\mathbf{x}_q$

$$\mathbf{s}_{pq} = \frac{\mathbf{x}_q - \mathbf{x}_p}{|\mathbf{x}_q - \mathbf{x}_p|}, \quad (33)$$

two tangent vectors to the face

$$\mathbf{t}_{pq}^{\circ} = \frac{\mathbf{x}_{pq}^{\oplus} - \mathbf{x}_{pq}^{\ominus}}{|\mathbf{x}_{pq}^{\oplus} - \mathbf{x}_{pq}^{\ominus}|}, \quad \mathbf{t}_{pq}^{\square} = \frac{\mathbf{x}_{pq}^{\boxplus} - \mathbf{x}_{pq}^{\boxminus}}{|\mathbf{x}_{pq}^{\boxplus} - \mathbf{x}_{pq}^{\boxminus}|},$$

and

$$\tilde{\mathbf{n}}_{pq} = \frac{1}{2}(\mathbf{x}_{pq}^{\oplus} - \mathbf{x}_{pq}^{\ominus}) \times (\mathbf{x}_{pq}^{\boxplus} - \mathbf{x}_{pq}^{\boxminus}).$$

Due to the orientation chosen on  $\sigma_{pq}$  and the ordering of the vertices of this face, if  $\mathbf{n}_{pq} : \sigma_{pq} \rightarrow \mathbb{R}^3$  is the pointwise outer unit normal to  $p$  on  $\sigma_{pq}$  we have

$$\tilde{\mathbf{n}}_{pq} = \iint_{\sigma_{pq}} \mathbf{n}_{pq}. \quad (34)$$

Since  $(\mathbf{s}_{pq}, \mathbf{t}_{pq}^{\circ}, \mathbf{t}_{pq}^{\square})$  form a basis of  $\mathbb{R}^3$ , there are  $\beta_{pq} > 0$  and  $(\alpha_{pq}^{\circ}, \alpha_{pq}^{\square}) \in \mathbb{R}^2$  such that

$$\tilde{\mathbf{n}}_{pq} = |\sigma_{pq}| \left( \frac{1}{\beta_{pq}} \mathbf{s}_{pq} - \frac{\alpha_{pq}^{\circ}}{\beta_{pq}} \mathbf{t}_{pq}^{\circ} - \frac{\alpha_{pq}^{\square}}{\beta_{pq}} \mathbf{t}_{pq}^{\square} \right). \quad (35)$$

The numerical fluxes are then given by: for  $p \in \mathfrak{T}$ ,  $\sigma \in \mathfrak{S}(p) \setminus \mathfrak{S}_\Gamma$ , and  $\varphi \in V_h$ ,

$$\mathcal{F}_{p,\sigma_{pq}}^\Omega(\varphi) = |\sigma_{pq}| \left( \frac{1}{\beta_{pq}} \frac{\varphi_p - \varphi_q}{d_{pq}} - \frac{\alpha_{pq}^\ominus}{\beta_{pq}} \frac{\varphi_{pq}^\oplus - \varphi_{pq}^\ominus}{d_{pq}^\ominus} - \frac{\alpha_{pq}^\square}{\beta_{pq}} \frac{\varphi_{pq}^\boxplus - \varphi_{pq}^\boxminus}{d_{pq}^\square} \right), \quad (36)$$

where

- $d_{pq} = |\mathbf{x}_p - \mathbf{x}_q|$  (as in Section 2),  $d_{pq}^\ominus = |\mathbf{x}_{pq}^\oplus - \mathbf{x}_{pq}^\ominus|$  and  $d_{pq}^\square = |\mathbf{x}_{pq}^\boxplus - \mathbf{x}_{pq}^\boxminus|$ , and
- for  $* \in \{\oplus, \ominus, \boxplus, \boxminus\}$ , each point  $\mathbf{y} \in \mathcal{R}(\mathbf{x}_{pq}^*)$  is associated with a genuine or degenerate cell  $r$  (possibly an edge or corner of  $\Omega$ ); we then let  $\varphi_{\mathbf{y}} = \varphi_r$  (with  $\varphi_r = 0$  if  $r$  is an edge or corner on  $\partial\Omega$ ) and, following (32), the secondary unknown  $\varphi_{pq}^*$  located at the vertex  $\mathbf{x}_{pq}^*$  is defined by

$$\varphi_{pq}^* = \frac{1}{\text{Card}(\mathcal{R}(\mathbf{x}_{pq}^*))} \sum_{\mathbf{y} \in \mathcal{R}(\mathbf{x}_{pq}^*)} \varphi_{\mathbf{y}}. \quad (37)$$

*Remark 10* (Correction of the flux in [29]). In [29], a similar flux is defined with right-hand side multiplied by  $\frac{|\tilde{\mathbf{n}}_{pq}|}{|\sigma_{pq}|}$  in (36) – that is,  $\beta_{pq}$  is multiplied by  $\frac{|\sigma_{pq}|}{|\tilde{\mathbf{n}}_{pq}|}$ . The consistency analysis in Section B.2.2 shows that this choice of flux is only consistent if the faces are asymptotically flat (that is,  $\frac{|\tilde{\mathbf{n}}_{pq}|}{|\sigma_{pq}|} \rightarrow 1$  as the mesh size tends to zero). The flux we define above is consistent even if some faces remain non-flat as the mesh size tends to zero.

### 3.1.2 Surface fluxes on $\Gamma$

We use the fluxes of the Mixed Finite Volumes [12], which is the finite volume presentation of the Hybrid Mimetic Mixed method (see [14] and [13, Section 13.2.2]). Let  $\varphi \in V_h$  and define, for  $\sigma \in \mathfrak{S}_\Gamma$  and denoting as usual by  $p \in \mathfrak{T}$  the unique control volume that contains  $\sigma$  in its boundary,

$$\nabla_{\Gamma,\sigma}\varphi = \frac{1}{|\sigma|} \sum_{e \in \mathfrak{E}(\sigma)} |e| \varphi_e \mathbf{n}_{\sigma,e}, \quad (38)$$

$$S_{p,e}(\varphi) = \varphi_e - \varphi_p - \nabla_{\Gamma,\sigma}\varphi \cdot (\bar{\mathbf{x}}_e - \mathbf{x}_p), \quad \forall e \in \mathfrak{E}(\sigma), \quad (39)$$

where  $\bar{\mathbf{x}}_e = \frac{1}{|e|} \int_e \mathbf{x}$  is the centre of mass of  $e$ . Assuming that  $\mathbf{n}_{\sigma,e}$  is constant along  $e$  (but see Remark 11 below), the Stokes formula and the definition (15) of  $I_h$  easily show that  $\nabla_{\Gamma,\sigma}$  is a consistent approximation of the tangential gradient on  $\Gamma$  in the sense that, if  $\varphi \in C^2(\Gamma)$ ,

$$\nabla_{\Gamma,\sigma} I_h \varphi = \frac{1}{|\sigma|} \int_\sigma \nabla_\Gamma \varphi. \quad (40)$$

As a consequence,  $S_{p,e}$  can be seen as the remainder of a discrete first order Taylor expansion. The HMM fluxes are then defined by: for all  $\varphi \in V_h$  and all  $\sigma \in \mathfrak{S}_\Gamma$ , the family  $(\mathcal{F}_{\sigma,e}^\Gamma(\varphi))_{e \in \mathfrak{E}(\sigma)}$  is the unique solution to

$$\sum_{e \in \mathfrak{E}(\sigma)} \mathcal{F}_{\sigma,e}^\Gamma(\varphi)(\psi_p - \psi_e) = |\sigma| \nabla_{\Gamma,\sigma}\varphi \cdot \nabla_{\Gamma,\sigma}\psi + \sum_{e \in \mathfrak{E}(\sigma)} \frac{|e|}{d_{pe}^\perp} S_{p,e}(\varphi) S_{p,e}(\psi), \quad \forall \psi \in V_h, \quad (41)$$

where we recall that  $d_{pe}^\perp$  is the orthogonal distance between  $\mathbf{x}_p$  and  $e$ .

*Remark 11* (Curved edges). The definition (38) is consistent if the only curvature of the edges is due to the curvature of  $\Gamma$ , that is, the unit normal vector  $\mathbf{n}_{\sigma,e}$  to  $\sigma$  on  $e$  along  $\Gamma$  is constant. However, the HMM remains asymptotically consistent on faces with slightly curved edges [6], in the sense that

$$\left| \mathbf{n}_{\sigma,e} - \frac{1}{|e|} \int_e \mathbf{n}_{\sigma,e} \right| = \mathcal{O}(h_\Gamma).$$

### 3.1.3 Properties of the fluxes

In this section, we show that, upon some mesh regularity assumption (that can be checked in practice during implementation), the inner and surface fluxes described in Sections 3.1.1 and 3.1.2 are coercive and consistent. As a consequence, Theorem 6 applies to the numerical scheme (19) based on these fluxes, and the error estimate (27) holds for this scheme.

We first define three mesh regularity factors. The first two are required to establish the properties of the inner fluxes (see Appendix), whereas the third one is linked to the properties of the HMM fluxes

1. The first regularity factor is related to the faces not lying on  $\Gamma$ :

$$\begin{aligned} \text{reg}_{\mathfrak{T}, \Omega} := & \max \left\{ \frac{|d_{pq}^*|}{d_{pq}^*} : \sigma_{pq} \in \mathfrak{S} \setminus \mathfrak{S}_\Gamma, * \in \{\ominus, \oplus, \boxminus, \boxplus\} \right\} \\ & + \max \left\{ \frac{1}{|\det(\mathbf{s}_{pq}, \mathbf{t}_{pq}^\ominus, \mathbf{t}_{pq}^\boxminus)|} : \sigma_{pq} \in \mathfrak{S} \setminus \mathfrak{S}_\Gamma \right\}, \end{aligned} \quad (42)$$

where  $d_{pq}^* = (|\mathbf{x}_r - \mathbf{x}_{pq}^*|)_{r \in \mathcal{R}(\mathbf{x}_{pq}^*)}$  and  $d_{pq}^* = d_{pq}^\ominus$  if  $* \in \{\ominus, \oplus\}$ ,  $d_{pq}^* = d_{pq}^\boxminus$  if  $* \in \{\boxminus, \boxplus\}$ .

2. The definition of the second regularity factor requires the introduction of a few notations associated to a pair  $(\mathbf{x}_{pq}^*, r)$ , where  $\mathbf{x}_{pq}^*$  is a vertex of a face  $\sigma_{pq} \in \mathfrak{S} \setminus \mathfrak{S}_\Gamma$  and  $r = p$  or  $q$ . We refer to Figure 3 for an illustration of these notations.

- If  $\mathbf{x}_{pq}^* \in \Omega$  is an internal vertex, we let  $F_{r,pq}^*$  be the set of the two control volumes that are neighbours of  $r$ , have  $\mathbf{x}_{pq}^*$  as vertex, but are neither  $p$  or  $q$ . The two control volumes in  $F_{r,pq}^*$  have two neighbours in common:  $r$  itself, and another control volume that we denote by  $e_{r,pq}^*$ .
- If  $\mathbf{x}_{pq}^* \in \Gamma$ , we let  $F_{r,pq}^*$  be the set made of the only control volume neighbour of  $r$ , that has  $\mathbf{x}_{pq}^*$  as vertex, but that is not  $p$  or  $q$ .
- If  $\mathbf{x}_{pq}^*$  lies on the Dirichlet boundary  $\partial\Omega \setminus \Gamma$ , it is the vertex of a face  $\sigma \in \mathfrak{S}(r) \cap \mathfrak{S}_{\text{Dir}}$ . We let  $F_{r,pq}^*$  be the set made of this face, which is identified to the degenerate control volume  $q$ .

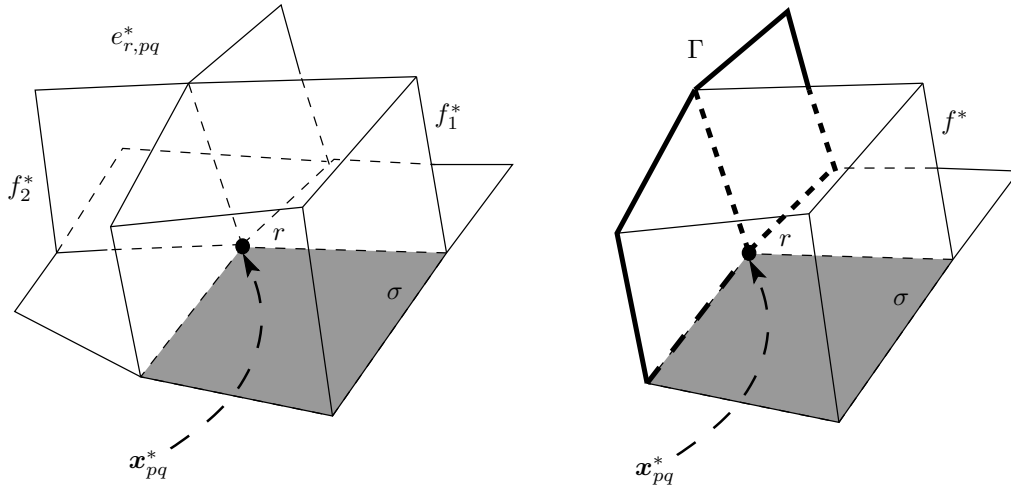


Figure 3: Illustration of the local labels around a face  $\sigma = \sigma_{pq}$  and one of its vertices  $\mathbf{x}_{pq}^*$ . Left:  $\mathbf{x}_{pq}^*$  is an internal vertex (then  $F_{r,pq}^* = \{f_1^*, f_2^*\}$ ); right:  $\mathbf{x}_{pq}^*$  lies on  $\Gamma$  (then  $F_{r,pq}^* = \{f^*\}$ ).

We then need to know, for a given face  $\sigma_{ab} \in \mathfrak{S} \setminus \mathfrak{S}_\Gamma$ , for which triplet  $(p, q, *)$  we have, for some  $r = p$  or  $q$  and  $f \in F_{r,pq}^*$ ,  $\{e_{r,pq}^*, f\} = \{a, b\}$  or  $\{r, f\} = \{a, b\}$ . These triplets are described by the following two sets

$$\begin{aligned} X_{ab} &:= \{(p, q, *) : \sigma_{pq} \in \mathfrak{S} \setminus (\mathfrak{S}_\Gamma \cup \mathfrak{S}_{\text{Dir}}) \\ &\quad \text{and, for some } r \in \{p, q\} \text{ and } f \in F_{r,pq}^*, \{a, b\} = \{e_{r,pq}^*, f\}\}, \\ Y_{ab} &:= \{(p, q, *) : \sigma_{pq} \in \mathfrak{S} \setminus (\mathfrak{S}_\Gamma \cup \mathfrak{S}_{\text{Dir}}) \\ &\quad \text{and, for some } r \in \{p, q\} \text{ and } f \in F_{r,pq}^*, \{a, b\} = \{r, f\}\}. \end{aligned} \quad (43)$$

The second regularity factor is then  $\varrho_{\mathfrak{T}, \Omega}$ , assumed to be  $> 0$ , such that

$$\begin{aligned} \varrho_{\mathfrak{T}, \Omega} &:= \min \left\{ \left[ \frac{1}{\beta_{ab}} - \epsilon_{ab} \frac{\alpha_{ab}^\circ d_{ab}}{2\beta_{ab} d_{ab}^\circ} - \epsilon_{ab} \frac{\alpha_{ab}^\square d_{ab}}{2\beta_{ab} d_{ab}^\square} \right] - \frac{1}{16} \sum_{(p,q,*) \in X_{ab}} \zeta_{X,pq}^* \frac{|\sigma_{pq}| d_{ab} \alpha_{pq}^\diamond}{|\sigma_{ab}| d_{pq}^\diamond \beta_{pq}} \right. \\ &\quad \left. - \frac{1}{16} \sum_{(p,q,*) \in Y_{ab}} \zeta_{Y,pq}^* \frac{|\sigma_{pq}| d_{ab} \alpha_{pq}^\diamond}{|\sigma_{ab}| d_{pq}^\diamond \beta_{pq}} : \sigma_{ab} \in \mathfrak{S} \setminus \mathfrak{S}_\Gamma \right\}, \end{aligned} \quad (44)$$

where

$$\epsilon_{ab} = \begin{cases} 0 & \text{if } \sigma_{ab} \in \mathfrak{S}_{\text{Dir}}, \\ 1 & \text{otherwise,} \end{cases} \quad (\zeta_{X,pq}^*, \zeta_{Y,pq}^*) = \begin{cases} (1, 3) & \text{if } \mathbf{x}_{pq}^* \in \Omega, \\ (0, 4) & \text{if } \mathbf{x}_{pq}^* \in \Gamma, \\ (0, 8) & \text{if } \mathbf{x}_{pq}^* \in \partial\Omega \setminus \Gamma, \end{cases} \quad (45)$$

and

$$\diamond = \begin{cases} \circ & \text{if } * \in \{\oplus, \ominus\}, \\ \square & \text{if } * \in \{\boxplus, \boxminus\}. \end{cases} \quad (46)$$

3. The third regularity factor is

$$\begin{aligned} \text{reg}_{\mathfrak{T}, \Gamma} &:= \max \left\{ \frac{\text{diam}(\sigma)}{d_{pe}^\perp} : \sigma \in \mathfrak{S}_\Gamma, e \in \mathfrak{E}(\sigma) \right\} \\ &\quad + \max \left\{ \frac{\text{diam}(\sigma)}{\text{diam}(\tau)} : e \in \mathfrak{E}_{\Gamma, \text{int}}, (\sigma, \tau) \text{ faces on each side of } e \right\}. \end{aligned} \quad (47)$$

**Proposition 12** (Properties of the fluxes). *The fluxes defined in Sections 3.1.1 and 3.1.2 satisfy Assumption 4 with  $\varrho_\Omega = \varrho_{\mathfrak{T}, \Omega}$ ,  $C_{\text{cons}}^\Omega$  depending only on an upper bound of  $\text{reg}_{\mathfrak{T}} + \text{reg}_{\mathfrak{T}, \Omega}$ , and  $\varrho_\Gamma > 0$  and  $C_{\text{cons}}^\Gamma$  depending only on an upper bound of  $\text{reg}_{\mathfrak{T}, \Gamma}$ .*

*Proof.* See Appendix B. □

*Remark 13* (About the regularity factors). Bounding  $\text{reg}_{\mathfrak{T}, \Omega}$  above imposes the proximity of a vertex  $\mathbf{x}_{pq}^*$  and the representative points  $\mathcal{R}(\mathbf{x}_{pq}^*)$  involved in its definition, as well as the non-degeneracy of the faces (whose diagonals  $\mathbf{t}_{pq}^\circ$  and  $\mathbf{t}_{pq}^\square$  must have a minimal angle) and the transversality of the vector  $\mathbf{x}_p \mathbf{x}_q$  and the face  $\sigma_{pq}$ . All these properties are natural given our construction of the control volumes.

The regularity factor  $\text{reg}_{\mathfrak{T}, \Gamma}$  plays the same role, for the mesh  $\mathfrak{S}_\Gamma$  of  $\Gamma$ , as the regularity factor  $\text{reg}_{\mathfrak{T}}$  for the mesh  $\mathfrak{T}$  of  $\Omega$ . See Remark 5 for an interpretation of these terms.

Bounding  $\varrho_{\mathfrak{T}, \Omega}$  below imposes that faces that share a common vertex must have comparable measures and diagonal lengths (the terms  $\frac{|\sigma_{pq}|}{|\sigma_{ab}|}$  and  $\frac{d_{ab}^\diamond}{d_{pq}^\diamond}$  remain bounded), and that  $\mathbf{s}_{pq}$  is “not too far” from the orthogonal direction to  $\sigma_{pq}$  (so that, recalling (35),  $\beta_{pq}$  remains close to 1 while  $\alpha_{pq}^\circ$  and  $\alpha_{pq}^\square$  remain small compared to 1).

All these regularity factors, as well as  $\text{reg}_{\mathfrak{T}}$ , are easy to numerically evaluate for a given mesh during the implementation. If, as the mesh is refined, these computed factors remain bounded above (for  $\text{reg}_{\mathfrak{T}}$ ,

$\text{reg}_{\mathfrak{T},\Omega}$  and  $\text{reg}_{\mathfrak{T},\Gamma}$ ) or below (for  $\varrho_{\mathfrak{T},\Omega}$ ), then it ensures the robustness and accuracy of the numerical output since the error estimate (27) then holds. Note however that these conditions on the regularity factors are merely sufficient, not necessary; the scheme can still, in some cases, converge even if these factors do not remain properly bounded.

### 3.2 Alternative schemes

In the numerical tests, we will also present the results using two alternative schemes to the ones described above. The first alternative scheme is similar to (16) in a way that it approximate the oblique derivative as an advection equation on the boundary. It uses an upwind discretisation, instead of a numerically stabilized centred discretisation, for the convective term on the boundary. The second alternative approach is similar to the approximation of inner fluxes (36). It approximates the fluxes through a boundary face on  $\Gamma$  by splitting the normal derivative into an oblique component, in the direction  $\mathbf{V}$ , and a tangential component to  $\Gamma$ . Similar splitting, but just on uniform rectangular, radial or spherical grids, was presented in [26, 27], where, however, additional points outside domain for treatment of normal derivative were introduced which is made possible with uniform structured grids.

#### Upwind scheme

The boundary advection term  $[\overline{T}\mathbf{W} \cdot \mathbf{n}]_{\sigma,e}$  in (16) is here discretised using an upwind approach (which, contrary to (19), does not require the introduction of numerical diffusion for stabilisation). The resulting scheme has the form

$$\begin{aligned} \sum_{\sigma \in \mathfrak{S}(p) \setminus \mathfrak{S}_\Gamma} \mathcal{F}_{p,\sigma}^\Omega(T) + \sum_{\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_\Gamma} \sum_{e \in \mathfrak{E}(\sigma)} \mathcal{F}_{\sigma,e}^{\Gamma,\text{adv}}(T) - \sum_{\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_\Gamma} T_p [\nabla_\Gamma \cdot \mathbf{W}]_\sigma \\ = \sum_{\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_\Gamma} \iint_\sigma g, \quad \forall p \in \mathfrak{T}, \end{aligned} \quad (48)$$

where the boundary advective numerical flux  $\mathcal{F}_{\sigma,e}^{\Gamma,\text{adv}}(T)$ , which approximates  $[\overline{T}\mathbf{W} \cdot \mathbf{n}]_{\sigma,e}$ , is given by

$$\mathcal{F}_{\sigma,e}^{\Gamma,\text{adv}}(T) = \begin{cases} T_q [\mathbf{W} \cdot \mathbf{n}]_{\sigma,e} & \text{if } [\mathbf{W} \cdot \mathbf{n}]_{\sigma,e} < 0, \\ T_p [\mathbf{W} \cdot \mathbf{n}]_{\sigma,e} & \text{if } [\mathbf{W} \cdot \mathbf{n}]_{\sigma,e} \geq 0. \end{cases}$$

*Remark 14* (Theoretical analysis). The theoretical analysis of this scheme can be conducted in a similar way as the scheme in Section 3.1, but leads to an  $\mathcal{O}(\sqrt{h})$  theoretical convergence rate (see in particular Remark 16).

#### Splitting scheme

Here, the oblique derivative is not recast as a normal component and a boundary advective component. Instead, it is directly used together with a tangential approximation to reconstruct the normal fluxes. The resulting scheme has the form

$$\sum_{\sigma \in \mathfrak{S}(p) \setminus \mathfrak{S}_\Gamma} \mathcal{F}_{p,\sigma}^\Omega(T) + \sum_{\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_\Gamma} \mathcal{F}_{p,\sigma}^\Gamma(T) = \sum_{\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_\Gamma} \iint_\sigma g, \quad \forall p \in \mathfrak{T} \quad (49)$$

where the numerical normal flux  $\mathcal{F}_{p,\sigma}^\Gamma(T)$ , that approximates  $\nabla \overline{T} \cdot \mathbf{n}$ , is given by

$$\begin{aligned} \mathcal{F}_{p,\sigma}^\Gamma(T) \\ = |\sigma| \left( \frac{1}{\beta_\sigma} g - \frac{\alpha_\sigma^\ominus}{\beta_\sigma} \frac{\frac{1}{4} \sum_{r \in \mathcal{R}(x_\sigma^\oplus)} T_r - \frac{1}{4} \sum_{r \in \mathcal{R}(x_\sigma^\ominus)} T_r}{d_\sigma^\ominus} - \frac{\alpha_\sigma^\boxplus}{\beta_\sigma} \frac{\frac{1}{4} \sum_{r \in \mathcal{R}(x_\sigma^\boxplus)} T_r - \frac{1}{4} \sum_{r \in \mathcal{R}(x_\sigma^\boxminus)} T_r}{d_\sigma^\boxplus} \right). \end{aligned} \quad (50)$$

$h$	$\text{reg}_{\mathcal{T}}$	$\text{reg}_{\mathcal{T},\Gamma}$	$\text{reg}_{\mathcal{T},\Omega}$	$\varrho_{\mathcal{T},\Omega}$
8.511e-01	7.311	5.536	3.322	6.768e-01
3.660e-01	7.427	6.394	3.279	4.295e-01
1.685e-01	7.920	5.949	3.402	3.508e-01
8.309e-02	7.879	5.967	3.383	2.798e-01
4.084e-02	8.267	6.870	3.510	2.089e-01
2.041e-02	8.655	6.584	3.585	1.669e-01
1.014e-02	8.356	6.854	3.589	1.426e-01

Table 1: The regularity parameters (24), (42), (44) and (47) for a non-uniform mesh of the cube.

Here, the coefficients  $\beta_\sigma$ ,  $\alpha_\sigma^\square$  and  $\alpha_\sigma^\circ$  are given by (35) with  $\mathbf{s}_{pq} = \mathbf{V}(\mathbf{x}_p)$ . They therefore correspond to the decomposition of  $\tilde{\mathbf{n}}_{pq}$  on the basis  $(\mathbf{V}(\mathbf{x}_p), \mathbf{t}_\sigma^\square, \mathbf{t}_\sigma^\circ)$ . The equation (50) approximates  $\nabla \bar{T} \cdot \mathbf{n}$  using the oblique derivative  $\nabla \bar{T} \cdot \mathbf{V} = g$  (see (1b)) and a tangential component, reconstructed from the boundary values of  $T$  using the same principles as in Section 3.1.1.

*Remark 15* (Theoretical analysis). Given the close proximity of the approximation (50) with the discretisation (36), the ideas developed in Appendix B could be adapted to yield an error estimate for this scheme. However, when establishing the coercivity of the method, additional boundary terms would have to be accounted for in the regularity factor (44). The scale of these additional negative terms is proportional to how tangential the oblique vector is, and the method fails to be coercive for problems with an oblique field that is too tangential to the boundary of the domain.

### 3.3 Results

To test the proposed methods we present three sets of numerical experiments. For all experiments, the exact solution is chosen to be  $\bar{T}(\mathbf{x}) = \frac{1}{\mathbf{x} - \mathbf{x}_0}$ , where  $\mathbf{x}_0 = (-0.3, -0.2, -0.1)$ . The regularity parameters (24), (42), (44), and (47) and the Experimental Order of Convergence (EOC) are presented.

#### 3.3.1 Cubic domain and non-uniform mesh

The computational domain  $\Omega$  for the first set of experiments is a cube with unit edge length. The boundary  $\Gamma$ , on which the oblique boundary condition is prescribed, corresponds to the bottom face of the cube. The mesh is a non-uniform one obtained constructing first a uniform grid with distance between representative points equal to  $h_u$ , and then moving each point by a random vector  $\mathbf{r}$  with components in  $(-0.15h_u, 0.15h_u)$ . Points on  $\partial\Omega$  are only moved in a direction tangential to the boundary. Experiments with different oblique vector fields are presented, and the regularity parameters are presented in Table 1. We notice that all parameters remain in a range that makes Theorem 6 and Proposition 12 applicable.

The first experiment, whose results are presented in Table 2, shows the convergence of the method for a constant vector field  $\mathbf{V}(\mathbf{x}) = (-1, -1, -1)$ . The method (19) displays a first order convergence in  $L^2$  and energy norms, which confirms the theoretical prediction of Theorem 6 and Proposition 12. The absence of super-convergence in  $L^2$  norm is not surprising, as specific Finite Volume methods are only known to super-converge under certain geometric conditions, and to fail to super-converge in some cases [15]. The rates of convergence for the upwind method (48) are around 1 in  $L^2$  norm but tend to 1/2 in  $V_{h,\Omega}$  norm, which is expected (see Remark 14). The splitting method (49) shows the best convergence rates: above second order in  $L^2$  norm, and above first order in  $V_{h,\Omega}$ .

The second experiment considers the non-constant vector field  $\mathbf{V}(x, y, z) = (x, y, -1)$ . In this case the surface divergence of  $\mathbf{W}(x, y, z) = (x, y, 0)$  (see (2)) is  $\nabla_\Gamma \cdot \mathbf{W}(\mathbf{x}) = 2$ . The tests show similar orders of convergence, albeit slightly reduced, as in the experiment with a constant vector field; see Table 3. The slight degradation could stem from the fact that the assumption (26a) is not fully satisfied on these meshes and with this vector field, or that the asymptotic rate has not been achieved at these mesh sizes.



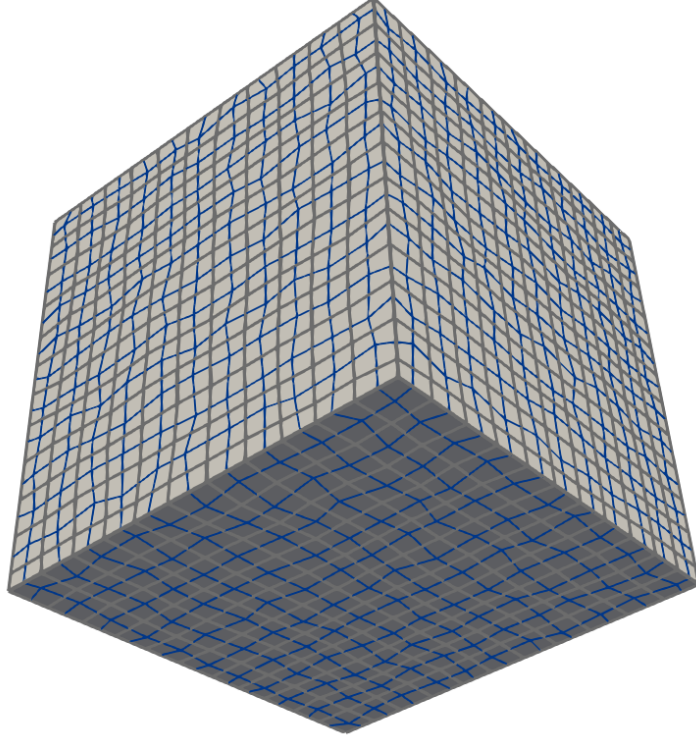


Figure 4: Non-uniform mesh of the cube

In the third experiment on the cube we consider a divergence free rotational vector field  $\mathbf{V}(x, y, z) = (-x, z, -1)$ . The results in Table 4 show that, here again, the schemes behave in a similar way as with the other two vector fields.

### 3.3.2 Tesseroid domain with non-planar $\Gamma$

The next experiments are run on a computational domain with a non-planar boundary  $\Gamma$ . The discrete computational domain then does not exactly match  $\Omega$ , and vertices of the boundary faces do not have to lie on the boundary  $\Gamma$ . Moreover, in this construction, the tangent space to  $\Gamma$  is not well defined everywhere so the co-normal  $\mathbf{n}_{\sigma,e}$  is not well defined either in the Eq. (19). In this case we approximate the normal vector  $\mathbf{n}_{\sigma,e}$  by the normalised version of  $\left(\frac{\mathbf{N}_p + \mathbf{N}_q}{2}\right) \times \mathbf{e}$ , where the vector  $\mathbf{N}_p$  is a normal to the face  $\sigma$ , the vector  $\mathbf{N}_q$  is normal to the neighbouring face on the other side of  $e$ , and the vector  $\mathbf{e}$  is a tangent vector to the edge  $e$ , chosen such that  $\mathbf{n}_{\sigma,e}$  is an outward normal to  $\sigma$ .

The experiments are performed on a non-uniform mesh of the tesseroid

$$\Omega = \left\{ (r \sin(u) \cos(v), r \sin(u) \sin(v), r \cos(u)) : r \in (1, 2), u \in \left(\frac{3\pi}{8}, \frac{5\pi}{8}\right), v \in \left(0, \frac{\pi}{4}\right) \right\}.$$

See Fig. 5 for an illustration. The oblique boundary condition is prescribed on the non-planar face corresponding to  $r = 1$ :

$$\Gamma = \left\{ (\sin(u) \cos(v), \sin(u) \sin(v), \cos(u)) : u \in \left(\frac{3\pi}{8}, \frac{5\pi}{8}\right), v \in \left(0, \frac{\pi}{4}\right) \right\}.$$

The regularity parameters of the considered meshes are presented in Table 5. As can be seen there, the regularity factor  $\varrho_{\overline{\Sigma}, \Omega}$  seems to degenerate as the mesh size is reduced, indicating that the condition that

Scheme (19)								
$h$	$L^2_\Omega$ error	EOC	$L^2_\Gamma$ error	EOC	$V_h$ error	EOC	$V_{h,\Gamma}$ error	EOC
8.511e-01	2.092e-02		4.600e-02		1.886e-01		1.844e-01	
3.660e-01	9.412e-03	0.946	2.634e-02	0.660	1.022e-01	0.726	1.462e-01	0.275
1.685e-01	3.922e-03	1.128	1.270e-02	0.940	4.767e-02	0.982	9.298e-02	0.583
8.309e-02	1.958e-03	0.982	6.756e-03	0.893	2.475e-02	0.927	6.358e-02	0.537
4.084e-02	1.003e-03	0.942	3.570e-03	0.898	1.294e-02	0.913	4.248e-02	0.567
2.041e-02	5.155e-04	0.959	1.867e-03	0.934	6.661e-03	0.957	2.678e-02	0.665
1.014e-02	2.560e-04	1.000	9.360e-04	0.986	3.287e-03	1.009	1.587e-02	0.747

Scheme (48)						
$h$	$L^2_\Omega$ error	EOC	$L^2_\Gamma$ error	EOC	$V_{h,\Omega}$ error	EOC
8.491e-01	3.224e-02		7.101e-02		1.409e-01	
3.683e-01	9.117e-03	1.512	2.530e-02	1.236	6.281e-02	0.967
1.688e-01	2.972e-03	1.436	9.999e-03	1.190	3.070e-02	0.917
8.303e-02	1.237e-03	1.236	4.555e-03	1.108	1.898e-02	0.677
4.192e-02	5.513e-04	1.182	2.147e-03	1.101	1.260e-02	0.599
2.045e-02	2.579e-04	1.058	1.036e-03	1.014	8.562e-03	0.538
1.018e-02	1.233e-04	1.058	5.075e-04	1.023	6.030e-03	0.502

Scheme (49)						
$h$	$L^2_\Omega$ error	EOC	$L^2_\Gamma$ error	EOC	$V_{h,\Omega}$ error	EOC
8.659e-01	2.154e-02		3.991e-02		8.012e-02	
3.692e-01	4.645e-03	1.800	1.186e-02	1.424	3.579e-02	0.945
1.691e-01	6.920e-04	2.437	2.228e-03	2.140	1.187e-02	1.413
8.246e-02	1.606e-04	2.035	5.954e-04	1.838	4.569e-03	1.330
4.066e-02	3.457e-05	2.172	1.353e-04	2.096	1.836e-03	1.290
2.046e-02	8.271e-06	2.083	3.127e-05	2.133	7.615e-04	1.282
1.029e-02	2.035e-06	2.039	7.885e-06	2.004	3.502e-04	1.130

Table 2: EOC for the non-uniform mesh of the cube with  $\mathbf{V}(\mathbf{x}) = (-1, -1, -1)$

Scheme (19)								
$h$	$L^2_\Omega$ error	EOC	$L^2_\Gamma$ error	EOC	$V_h$ error	EOC	$V_{h,\Gamma}$ error	EOC
8.478e-01	1.508e-02		2.908e-02		1.066e-01		1.021e-01	
3.606e-01	9.891e-03	0.493	2.639e-02	0.113	8.779e-02	0.226	1.208e-01	-0.197422
1.694e-01	5.583e-03	0.757	1.671e-02	0.605	5.182e-02	0.697	9.645e-02	0.298902
8.197e-02	3.237e-03	0.751	1.001e-02	0.705	2.929e-02	0.786	7.008e-02	0.439947
4.068e-02	1.803e-03	0.835	5.652e-03	0.815	1.571e-02	0.889	4.541e-02	0.619154
2.032e-02	9.617e-04	0.905	3.036e-03	0.895	8.098e-03	0.954	2.699e-02	0.749373
1.031e-02	5.041e-04	0.951	1.598e-03	0.945	4.148e-03	0.985	1.534e-02	0.832171

Scheme (48)						
$h$	$L^2_\Omega$ error	EOC	$L^2_\Gamma$ error	EOC	$V_{h,\Omega}$ error	EOC
8.700e-01	3.336e-02		7.199e-02		1.388e-01	
3.639e-01	8.894e-03	1.517	2.369e-02	1.275	5.440e-02	1.075
1.684e-01	3.048e-03	1.390	9.325e-03	1.210	2.413e-02	1.055
8.341e-02	1.390e-03	1.118	4.350e-03	1.085	1.256e-02	0.929
4.114e-02	6.835e-04	1.004	2.169e-03	0.984	7.078e-03	0.811
2.055e-02	3.410e-04	1.002	1.085e-03	0.997	4.233e-03	0.740
1.015e-02	1.698e-04	0.987	5.414e-04	0.985	2.700e-03	0.637

Scheme (49)						
$h$	$L^2_\Omega$ error	EOC	$L^2_\Gamma$ error	EOC	$V_{h,\Omega}$ error	EOC
8.455e-01	1.617e-02		3.542e-02		7.634e-02	
3.630e-01	2.437e-03	2.237	6.801e-03	1.952	1.956e-02	1.610
1.722e-01	7.058e-04	1.662	1.880e-03	1.725	8.495e-03	1.119
8.220e-02	1.720e-04	1.908	4.538e-04	1.921	3.432e-03	1.225
4.063e-02	3.849e-05	2.125	1.034e-04	2.099	1.455e-03	1.217
2.032e-02	9.436e-06	2.029	2.561e-05	2.014	6.904e-04	1.076
1.022e-02	2.326e-06	2.039	6.338e-06	2.033	3.302e-04	1.074

Table 3: EOC for the non-uniform mesh of the cube with  $\mathbf{V}(x, y, z) = (x, y, -1)$

Scheme (19)								
$h$	$L^2_\Omega$ error	EOC	$L^2_\Gamma$ error	EOC	$V_h$ error	EOC	$V_{h,\Gamma}$ error	EOC
8.594e-01	1.402e-02		3.084e-02		1.103e-01		1.070e-01	
3.651e-01	7.731e-03	0.695	2.187e-02	0.401	7.795e-02	0.405	1.084e-01	-0.015
1.703e-01	4.769e-03	0.633	1.455e-02	0.534	4.839e-02	0.625	9.136e-02	0.224
8.271e-02	2.620e-03	0.828	8.360e-03	0.767	2.636e-02	0.840	6.458e-02	0.480
4.070e-02	1.389e-03	0.894	4.520e-03	0.867	1.347e-02	0.947	4.045e-02	0.659
2.037e-02	7.382e-04	0.913	2.428e-03	0.897	6.985e-03	0.948	2.454e-02	0.721
1.018e-02	3.768e-04	0.969	1.248e-03	0.960	3.490e-03	1.000	1.378e-02	0.832

Scheme (48)						
$h$	$L^2_\Omega$ error	EOC	$L^2_\Gamma$ error	EOC	$V_{h,\Omega}$ error	EOC
8.524e-01	1.366e-02		1.930e-02		6.906e-02	
3.670e-01	2.301e-03	2.114	6.247e-03	1.339	2.381e-02	1.264
1.679e-01	9.552e-04	1.124	3.374e-03	0.787	1.730e-02	0.408
8.357e-02	3.550e-04	1.419	1.483e-03	1.178	8.874e-03	0.957
4.078e-02	1.550e-04	1.155	6.917e-04	1.063	5.654e-03	0.628
2.026e-02	7.428e-05	1.051	3.290e-04	1.062	3.727e-03	0.595
1.035e-02	3.648e-05	1.059	1.585e-04	1.088	2.510e-03	0.588

Scheme (49)						
$h$	$L^2_\Omega$ error	EOC	$L^2_\Gamma$ error	EOC	$V_{h,\Omega}$ error	EOC
8.668e-01	8.610e-03		1.747e-02		4.397e-02	
3.625e-01	1.853e-03	1.763	3.020e-03	2.013	1.659e-02	1.118
1.689e-01	5.089e-04	1.692	9.059e-04	1.576	7.446e-03	1.049
8.294e-02	1.174e-04	2.063	2.719e-04	1.693	3.431e-03	1.090
4.070e-02	2.681e-05	2.075	6.039e-05	2.113	1.458e-03	1.202
2.076e-02	6.741e-06	2.051	1.415e-05	2.155	6.850e-04	1.122
1.023e-02	1.644e-06	1.994	3.515e-06	1.968	3.287e-04	1.037

Table 4: EOC for the non-uniform mesh of the cube with  $\mathbf{V}(x, y, z) = (-x, z, -1)$

$h$	$\text{reg}_{\mathfrak{T}}$	$\text{reg}_{\mathfrak{T},\Omega}$	$\text{reg}_{\mathfrak{T},\Gamma}$	$\varrho_{\mathfrak{T},\Omega}$
1.092	2.538e+01	5.468	3.296	6.074e-01
4.981e-01	1.336e+01	5.970	3.468	4.051e-01
2.375e-01	1.045e+01	6.228	3.297	3.691e-01
1.158e-01	1.017e+01	6.352	3.442	2.463e-01
5.733e-02	0.998e+01	6.324	3.426	1.460e-01
2.847e-02	1.029e+01	6.582	3.598	1.392e-01
1.454e-02	1.032e+01	6.761	3.580	8.204e-02

Table 5: The regularity parameters (24), (42), (44) and (47) for a non-uniform mesh of a tesseroïd mesh.

ensure the coercivity of the inner fluxes (see Proposition 12) might not hold – which does not necessarily mean that the scheme actually fails to be coercive or to converge, since this is only a *sufficient* condition. The tests present the convergence of the methods for a non-constant vector field  $\mathbf{V}(\mathbf{x}) = (0.3, 0.2, 0.1) - \mathbf{x}$ . The results presented in Table 6 are similar to the ones obtained in Section 3.3.1, with perhaps slightly better rates of convergence across the board. In any case, the apparent decay of the regularity factor  $\varrho_{\mathfrak{T},\Omega}$  does not seem to negatively impact the convergence of the schemes.

### 3.3.3 Spherical section domain with perturbed bottom $\Gamma$

This series of experiments is performed on a section of a spherical domain, with a perturbed bottom boundary  $\Gamma$  (see Fig. 6):

$$\begin{aligned} \Omega = \Big\{ & [(1 + 0.04(2 - r)(\sin(10\ u) + \sin(10\ v))) \sin(u) \cos(v), \\ & (1 + 0.04(2 - r)(\sin(10\ u) + \sin(10\ v))) \sin(u) \sin(v), \\ & (1 + 0.04(2 - r)(\sin(10\ u) + \sin(10\ v))) \cos(u)] : \\ & r \in (1, 2),\ u \in \left(\frac{3\pi}{8}, \frac{5\pi}{8}\right),\ v \in \left(0, \frac{\pi}{4}\right) \Big\}, \\ \Gamma = \Big\{ & [(1 + 0.04(\sin(10\ u) + \sin(10\ v))) \sin(u) \cos(v), \\ & (1 + 0.04(\sin(10\ u) + \sin(10\ v))) \sin(u) \sin(v), \\ & (1 + 0.04(\sin(10\ u) + \sin(10\ v))) \cos(u)] : \\ & u \in \left(\frac{3\pi}{8}, \frac{5\pi}{8}\right),\ v \in \left(0, \frac{\pi}{4}\right) \Big\}. \end{aligned}$$

The regularity parameters for the considered meshes are presented in Table 7. The coercivity constant  $\varrho_{\mathfrak{T},\Omega}$  is worse as in the tesseroïd case, as it becomes negative. However, once again, since a lower bound on this constant is only a sufficient condition for the theoretical analysis, this does not mean that the schemes fail to converge, as the numerical results will show. We take the non-constant vector field  $\mathbf{V}(\mathbf{x}) = (0.3, 0.2, 0.1) - \mathbf{x}$ . Table 8 shows that all three schemes behave in a similar way as in the previous tests of Sections 3.3.1 and 3.3.2. This indicates that our coercivity analysis (based on  $\varrho_{\mathfrak{T},\Omega}$ ) is actually a bit too conservative regarding the robustness range of the discretisations.

### 3.3.4 Cubic domain, almost tangential vector field $\mathbf{V}$

In this final series of numerical experiments with an analytical solution, we show the advantage of the proposed scheme (19) and of the upwind scheme (48) over the splitting scheme (49). The computational domain  $\Omega$  is the unit cube, with  $\Gamma$  being its bottom. We take the vector field  $\mathbf{V} = (11.4301, 0, -1)$ , which

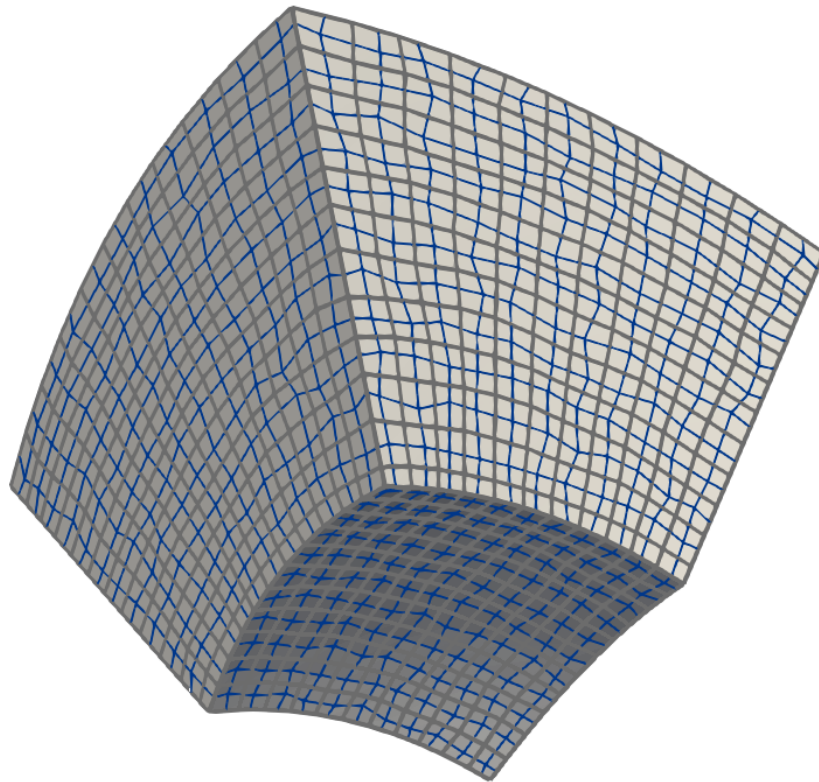


Figure 5: Non-uniform mesh of a tesseract

Scheme (19)								
$h$	$L^2_\Omega$ error	EOC	$L^2_\Gamma$ error	EOC	$V_h$ error	EOC	$V_{h,\Gamma}$ error	EOC
1.092	7.054e-03		9.644e-03		3.776e-02		3.620e-02	
4.981e-01	1.307e-03	2.148	2.189e-03	1.890	9.516e-03	1.756	1.205e-02	1.402
2.375e-01	3.656e-04	1.720	7.681e-04	1.414	3.249e-03	1.451	5.267e-03	1.117
1.158e-01	1.294e-04	1.446	3.335e-04	1.161	1.322e-03	1.252	2.698e-03	0.931
5.733e-02	5.493e-05	1.219	1.578e-04	1.065	5.788e-04	1.175	1.400e-03	0.932
2.847e-02	2.518e-05	1.114	7.580e-05	1.047	2.643e-04	1.119	7.200e-04	0.950
1.454e-02	1.248e-05	1.044	3.843e-05	1.011	1.285e-04	1.073	3.805e-04	0.949

Scheme (48)							
$h$	$L^2_\Omega$ error	EOC	$L^2_\Gamma$ error	EOC	$V_h$ error	EOC	
1.075	5.920e-03		5.346e-03		2.111e-02		
4.783e-01	1.184e-03	1.989	1.474e-03	1.592	6.483e-03	1.458	
2.326e-01	2.935e-04	1.934	5.223e-04	1.439	2.464e-03	1.342	
1.156e-01	9.156e-05	1.665	2.141e-04	1.275	1.055e-03	1.213	
5.766e-02	3.433e-05	1.411	9.374e-05	1.188	4.987e-04	1.077	
2.826e-02	1.458e-05	1.201	4.358e-05	1.074	2.661e-04	0.880	
1.435e-02	6.738e-06	1.139	2.099e-05	1.078	1.569e-04	0.779	

Scheme (49)							
$h$	$L^2_\Omega$ error	EOC	$L^2_\Gamma$ error	EOC	$V_h$ error	EOC	
1.073	5.384e-03		5.121e-03		1.800e-02		
4.950e-01	1.073e-03	2.084	1.114e-03	1.970	6.262e-03	1.364	
2.362e-01	2.165e-04	2.162	2.155e-04	2.220	2.093e-03	1.481	
1.155e-01	5.076e-05	2.028	4.964e-05	2.053	8.340e-04	1.286	
5.767e-02	1.247e-05	2.022	1.277e-05	1.954	3.506e-04	1.248	
2.845e-02	3.046e-06	1.994	3.107e-06	2.001	1.574e-04	1.133	
1.430e-02	7.544e-07	2.029	7.650e-07	2.038	7.386e-05	1.100	

Table 6: EOC for a non-uniform mesh of a tesseroid with  $\mathbf{V}(\mathbf{x}) = (0.3, 0.2, 0.1) - \mathbf{x}$

$h$	$\text{reg}_{\mathfrak{T}}$	$\text{reg}_{\mathfrak{T},\Omega}$	$\text{reg}_{\mathfrak{T},\Gamma}$	$\varrho_{\mathfrak{T},\Omega}$
1.089	3.228e+01	5.017	3.571	3.068e-01
4.928e-01	1.896e+01	6.270	3.551	-6.303e-01
2.299e-01	1.207e+01	6.259	3.777	-9.516e-01
1.159e-01	1.120e+01	6.819	3.838	-1.138
5.813e-02	1.083e+01	7.246	3.904	-1.524
2.867e-02	1.114e+01	6.890	4.253	-1.403
1.462e-02	1.118e+01	6.914	4.232	-1.642

Table 7: The regularity parameters (24), (42), (44) and (47) for a non-uniform mesh of a section of a perturbed ball.

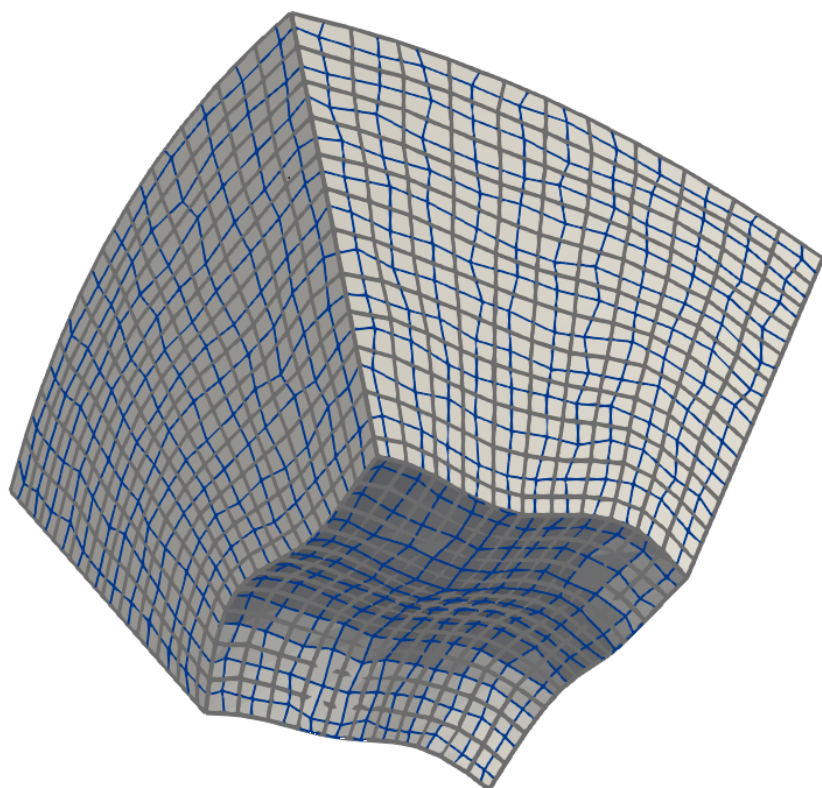


Figure 6: Section of a perturbed ball



Scheme (19)								
$h$	$L^2_\Omega$ error	EOC	$L^2_\Gamma$ error	EOC	$V_h$ error	EOC	$V_{h,\Gamma}$ error	EOC
1.089	7.577e-03		1.212e-02		5.851e-02		6.169e-02	
4.928e-01	2.595e-03	1.351	6.735e-03	0.740	4.232e-02	0.408	7.120e-02	-0.180
2.299e-01	1.519e-03	0.7021	5.123e-03	0.358	2.766e-02	0.557	6.351e-02	0.150
1.159e-01	8.678e-04	0.817	3.163e-03	0.704	1.465e-02	0.928	4.295e-02	0.571
5.813e-02	4.763e-04	0.868	1.776e-03	0.835	7.250e-03	1.019	2.550e-02	0.754
2.867e-02	2.548e-04	0.885	9.569e-04	0.875	3.594e-03	0.993	1.433e-02	0.815
1.462e-02	1.335e-04	0.959	5.028e-04	0.955	1.793e-03	1.032	7.777e-03	0.907

Scheme (48)						
$h$	$L^2_\Omega$ error	EOC	$L^2_\Gamma$ error	EOC	$V_h$ error	EOC
1.120	6.715e-03		6.435e-03		2.482e-02	
4.881e-01	1.643e-03	1.695	3.335e-03	0.791	1.269e-02	0.807
2.329e-01	7.427e-04	1.073	2.294e-03	0.505	8.068e-03	0.612
1.142e-01	3.565e-04	1.030	1.256e-03	0.845	4.444e-03	0.836
5.749e-02	1.717e-04	1.065	6.353e-04	0.993	2.336e-03	0.936
2.890e-02	8.426e-05	1.035	3.175e-04	1.008	1.245e-03	0.915
1.437e-02	4.168e-05	1.008	1.582e-04	0.997	6.884e-04	0.848

Scheme (49)						
$h$	$L^2_\Omega$ error	EOC	$L^2_\Gamma$ error	EOC	$V_h$ error	EOC
1.086	6.387e-03		5.968e-03		2.235e-02	
4.943e-01	1.229e-03	2.093	1.479e-03	1.772	7.613e-03	1.368
2.332e-01	2.756e-04	1.991	3.978e-04	1.749	2.536e-03	1.464
1.165e-01	6.656e-05	2.047	1.067e-04	1.895	9.496e-04	1.415
5.864e-02	1.597e-05	2.079	2.622e-05	2.045	3.870e-04	1.307
2.925e-02	3.894e-06	2.028	6.502e-06	2.005	1.739e-04	1.150
1.441e-02	9.641e-07	1.972	1.616e-06	1.967	8.209e-05	1.060

Table 8: EOC for  $\Omega$  the section of a perturbed ball.  $\mathbf{V}(\mathbf{x}) = (0.3, 0.2, 0.1) - \mathbf{x}$

Scheme (19)								
$h$	$L_\Omega^2$ error	EOC	$L_\Gamma^2$ error	EOC	$V_h$ error	EOC	$V_{h,\Gamma}$ error	EOC
8.605e-01	2.247e-02		4.945e-02		1.686e-01		1.610e-01	
3.606e-01	1.537e-02	0.436	4.232e-02	0.179	1.437e-01	0.184	2.014e-01	-0.257
1.692e-01	1.088e-02	0.455	3.301e-02	0.328	1.060e-01	0.401	2.012e-01	0.001
8.347e-02	6.582e-03	0.712	2.109e-02	0.634	6.669e-02	0.655	1.647e-01	0.283
4.054e-02	3.756e-03	0.776	1.255e-02	0.718	4.010e-02	0.704	1.248e-01	0.384
2.043e-02	2.046e-03	0.886	7.078e-03	0.836	2.333e-02	0.790	9.148e-02	0.453
1.017e-02	1.100e-03	0.889	3.913e-03	0.848	1.342e-02	0.792	6.596e-02	0.468

Scheme (48)						
$h$	$L_\Omega^2$ error	EOC	$L_\Gamma^2$ error	EOC	$V_h$ error	EOC
8.638e-01	5.093e-02		1.084e-01		2.195e-01	
3.664e-01	3.020e-02	0.609	8.669e-02	0.260	2.056e-01	0.760
1.685e-01	1.156e-02	1.236	4.137e-02	0.952	1.319e-01	0.571
8.343e-02	5.254e-03	1.122	2.077e-02	0.980	8.042e-02	0.703
4.077e-02	2.504e-03	1.035	1.045e-02	0.959	5.019e-02	0.658
2.044e-02	1.209e-03	1.055	5.121e-03	1.033	3.015e-02	0.738
1.022e-02	5.952e-04	1.022	2.529e-03	1.018	1.849e-02	0.705

Table 9: EOC for  $\Omega$  a cube and  $\mathbf{V}(\mathbf{x}) = (11.4301, 0, -1)$

corresponds to the outer normal on  $\Gamma$  rotated by  $85^\circ$ ; this vector field is therefore almost tangential to the boundary. The tests are run on uniform meshes.

Table 9 presents the EOC for the proposed scheme and upwind scheme. The rates are sometimes degraded compared to the previous tests, but there is a clear convergence.

For the splitting scheme (49), the fact that  $\mathbf{V}$  is almost tangential leads to very large values of  $\frac{\alpha_\sigma^*}{\beta_\sigma}$  in (50). As a consequence, the negative coefficients in the coercivity factor are too large to be controlled by the positive coefficients; the scheme really becomes non-coercive and unstable, and the BiCGStab algorithm used to solve the system fails. This breakdown of a numerical method is probably the worst situation that one wants to avoid in practice, which indicates that in severely oblique situations the proposed new methods (19) and (48) should be preferred, despite yielding sometimes reduced rates of convergence.

### 3.4 Local gravity field modelling

In this section we present local gravity field modelling over Slovakia using terrestrial gravity data. The goal of this experiment is to compute a disturbing potential using presented FVM schemes with oblique BC from terrestrial measurements and Dirichlet BCs obtained from satellite based model. Then we transform obtained potential to quasi-geoidal heights and compare them with real measurements. On the upper and side boundaries, the GO\_CONS\_GCF\_2\_DIR\_R5 model [7] was used and on the bottom boundary we used the surface gravity disturbances obtained from the available regular grid of gravity anomalies, with the resolution  $20'' \times 30''$ , that was compiled from original gravimetric measurements [19]. The gravity anomalies were transformed into the gravity disturbances by official digital vertical reference model DVRM ([www.geoportal.sk](http://www.geoportal.sk)).

The domain was bounded by  $\langle 16^\circ, 23^\circ \rangle$  meridians and  $\langle 47^\circ, 50.5^\circ \rangle$  parallels. The side boundaries were chosen sufficiently far from the area of Slovakia in order to mitigate an influence of the prescribed Dirichlet BC generated from the satellite-only geopotential model. For more details about this influence see [16]. The heights were interpolated from SRTM30 PLUS model and the upper boundary is in the

	method 19	method 48	method 49
Min	0.229	0.237	0.239
Mean	0.326	0.330	0.331
Max	0.449	0.458	0.459
Range	0.22	0.221	0.220
STD	0.052	0.050	0.050

Table 10: The GNSS-leveling test  $[m]$  at 58 points in area of Slovakia.

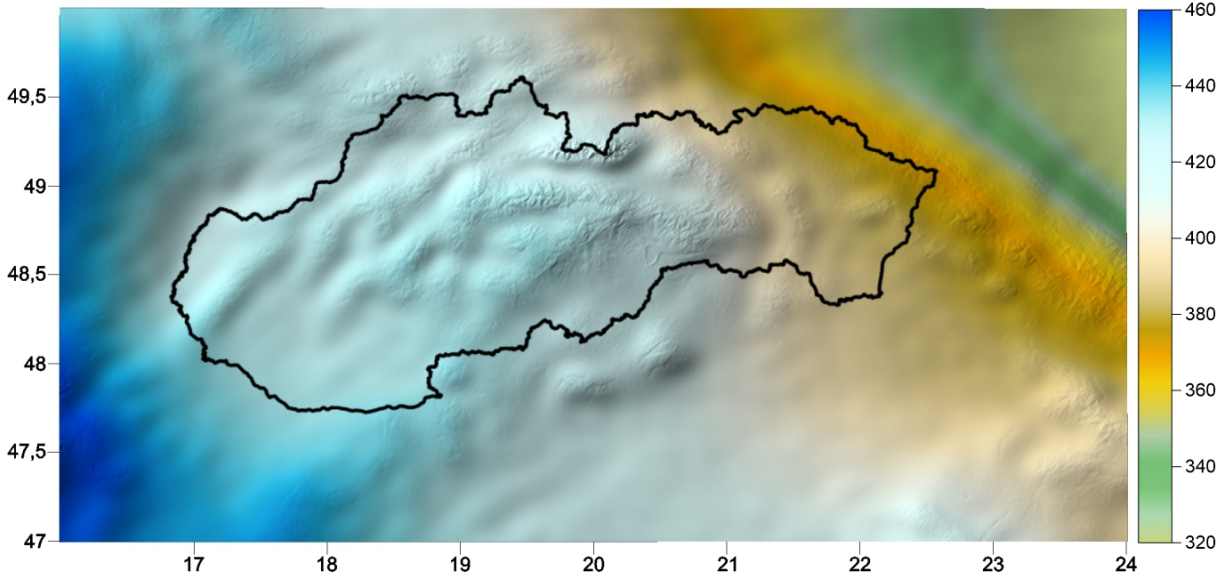


Figure 7: Disturbing potential in the area of Slovakia.

height of 240km above the reference ellipsoid.

Three experiments with the grid density  $841 \times 631 \times 301$  were performed using the FVM schemes (19), (48) and (49). The accuracy of the simulations was tested using GNSS-leveling. From the available dataset of 61 GNSS-leveling benchmarks, three evident outliers were removed. Hence, we tested the obtained local quasi-geoid model at 58 points. The results are summarised in Table 10 and, for the method (19), they are visualized in Fig. 7. We see a comparable precision of all the methods in this experiment. With this grid resolution the standard deviation of residuals between numerical results and measurements for all schemes is around 5cm.

## 4 Conclusion

We developed a framework for designing and analysing Finite Volume schemes for the Laplace equation with oblique boundary conditions. This framework, which can easily be extended to more general second order differential equations, consists in splitting the boundary condition into a normal and a tangential component, the later being handled as an advection term along the boundary of the domain; to ensure optimal convergence rates, this advection term is discretised using a centered scheme, with added numerical diffusion for stability purposes. The convergence analysis was carried out under usual coercivity and consistency assumptions on the numerical fluxes, and therefore applies to a range of possible FV discretisations. This analysis establishes first-order rates of convergence in a discrete energy ( $H^1$ ) norm.

We then constructed specific fluxes, in the case where the computational domain is discretised using generalised hexahedra, and we identified geometrical conditions, easy to check during simulations, that ensure their coercivity and consistency. Two alternative discretisations of the oblique boundary conditions were also presented: the first one uses an upwind FV discretisation of the boundary advection, the second is not based on a FV discretisation on the boundary, but rather on splitting the outer normal to the boundary into its oblique component, and a tangential component discretised using finite differences and the specific geometry of the mesh.

We then provided extensive numerical tests, designed to assess the accuracy and robustness of the method, for various choices of the computational domain, and of the oblique vector field defining the boundary conditions. These tests confirmed, for all three schemes, the theoretical first-order rate of convergence in energy norm. In some tests, the energy rate of convergence is actually apparently higher than the theoretical one (but the asymptotic convergence rate might not have been attained at the considered mesh sizes). The second variant, based on a splitting of the outer normal, seems to present the best accuracy in our initial tests, when the velocity field is not too tangential to the boundary. For a nearly tangential velocity field, this splitting scheme breaks down as the numerical solver fails to find a solution to it. On the contrary, the other two variants remain robust and convergent in this extreme situation, albeit with a reduced accuracy. All three schemes were used to compute a quasi-geoidal height in the region of Slovakia. For this test, all methods give results with comparable quality.

## A Proof of Theorem 6

The proof hinges on the 3rd Strang lemma of [9]. Let us first recast the scheme (19) under a variational formulation. Take  $\varphi \in V_h$ , multiply (19a) by  $\varphi_p$  and summing over  $p \in \mathfrak{T}$  to get

$$\begin{aligned} \sum_{p \in \mathfrak{T}} \left( \sum_{\sigma \in \mathfrak{S}(p) \setminus \mathfrak{S}_\Gamma} \mathcal{F}_{p,\sigma}^\Omega(T) \varphi_p + \sum_{\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_\Gamma} \sum_{e \in \mathfrak{E}(\sigma)} T_e [\mathbf{W} \cdot \mathbf{n}]_{\sigma,e} \varphi_p \right. \\ \left. - \sum_{\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_\Gamma} [\nabla_\Gamma \cdot \mathbf{W}]_\sigma T_p \varphi_p + Rh_\Gamma \sum_{\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_\Gamma} \sum_{e \in \mathfrak{E}(\sigma)} \mathcal{F}_{\sigma,e}^\Gamma(T) \varphi_p \right) = \sum_{p \in \mathfrak{T}} \sum_{\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_\Gamma} \iint_\sigma g \varphi_p. \end{aligned}$$

Using the conservativity of the fluxes  $\mathcal{F}_{p,\sigma}^\Omega$  (see (18)),  $\mathcal{F}_{\sigma,e}^\Gamma(T)$  (see (19b)) and  $T_e [\mathbf{W} \cdot \mathbf{n}]_{\sigma,e}$  (by definition of  $[\mathbf{W} \cdot \mathbf{n}]_{\sigma,e}$ ), and the zero value of  $T_e$  if  $e$  is a boundary edge in  $\Gamma$ , we gather the sums in the left-hand side by faces and edges as in [9, Proofs of Theorem 27 and 33] to find

$$\begin{aligned} \sum_{\sigma \in \mathfrak{S} \setminus \mathfrak{S}_\Gamma} \mathcal{F}_{p,\sigma}^\Omega(T) (\varphi_p - \varphi_q) + \sum_{\sigma \in \mathfrak{S}_\Gamma} \sum_{e \in \mathfrak{E}(\sigma)} T_e [\mathbf{W} \cdot \mathbf{n}]_{\sigma,e} (\varphi_p - \varphi_e) \\ - \sum_{\sigma \in \mathfrak{S}_\Gamma} [\nabla_\Gamma \cdot \mathbf{W}]_\sigma T_p \varphi_p + Rh_\Gamma \sum_{\sigma \in \mathfrak{S}_\Gamma} \sum_{e \in \mathfrak{E}(\sigma)} \mathcal{F}_{\sigma,e}^\Gamma(T) (\varphi_p - \varphi_e) = \sum_{p \in \mathfrak{T}} \sum_{\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_\Gamma} \iint_\sigma g \varphi_p. \end{aligned} \quad (51)$$

The solution  $T \in V_h$  to the scheme thus satisfies  $a_h(T, \varphi) = \ell_h(\varphi)$  for all  $\varphi \in V_h$ , with  $a_h(T, \varphi)$  (resp.  $\ell_h$ ) the bilinear form (resp. linear form) in the left-hand side (resp. right-hand side) of (51). Owing to the 3rd Strang lemma [9, Theorem 10], the estimate (27) follows if we establish the coercivity and consistency properties:

$$a_h(\varphi, \varphi) \gtrsim \|\varphi\|_{V_h}^2 \quad \forall \varphi \in V_h \quad (52)$$

and, letting  $\mathcal{E}_h(\bar{T}; \varphi) := \ell_h(\varphi) - a_h(\bar{T}, \varphi)$  be the consistency error,

$$\sup_{\varphi \in V_h, \|\varphi\|_{V_h} \leq 1} \mathcal{E}_h(\bar{T}; \varphi) \lesssim h \|\bar{T}\|_{C^2(\bar{\Omega})}. \quad (53)$$

## A.1 Coercivity

The coercivity properties (20) and (21) show that

$$a_h(\varphi, \varphi) \geq \rho_\Omega |\varphi|_{V_h, \Omega}^2 + Rh_\Gamma \rho_\Gamma |\varphi|_{V_h, \Gamma}^2 + \underbrace{\sum_{\sigma \in \mathfrak{S}_\Gamma} \sum_{e \in \mathfrak{E}(\sigma)} \varphi_e [\mathbf{W} \cdot \mathbf{n}]_{\sigma, e} (\varphi_p - \varphi_e) - \sum_{\sigma \in \mathfrak{S}_\Gamma} [\nabla_\Gamma \cdot \mathbf{W}]_\sigma \varphi_p^2}_{\mathcal{T}_1}. \quad (54)$$

Simple algebraic identities show that

$$\begin{aligned} \mathcal{T}_1 &= \sum_{\sigma \in \mathfrak{S}_\Gamma} \sum_{e \in \mathfrak{E}(\sigma)} [\mathbf{W} \cdot \mathbf{n}]_{\sigma, e} \left( \frac{\varphi_e - \varphi_p}{2} + \frac{\varphi_e + \varphi_p}{2} \right) (\varphi_p - \varphi_e) - \sum_{\sigma \in \mathfrak{S}_\Gamma} [\nabla_\Gamma \cdot \mathbf{W}]_\sigma \varphi_p^2 \\ &= \sum_{\sigma \in \mathfrak{S}_\Gamma} \sum_{e \in \mathfrak{E}(\sigma)} [\mathbf{W} \cdot \mathbf{n}]_{\sigma, e} \frac{1}{2} \left( -(\varphi_e - \varphi_p)^2 + \varphi_p^2 - \varphi_e^2 \right) - \sum_{\sigma \in \mathfrak{S}_\Gamma} [\nabla_\Gamma \cdot \mathbf{W}]_\sigma \varphi_p^2 \end{aligned}$$

By conservativity of  $[\mathbf{W} \cdot \mathbf{n}]_{\sigma, e}$  and zero value of  $\varphi_e$  on boundary edges of  $\Gamma$ ,

$$\sum_{\sigma \in \mathfrak{S}_\Gamma} \sum_{e \in \mathfrak{E}(\sigma)} [\mathbf{W} \cdot \mathbf{n}]_{\sigma, e} \varphi_e^2 = \sum_{e \in \mathfrak{E}_{\Gamma, \text{int}}} \left( [\mathbf{W} \cdot \mathbf{n}]_{\sigma, e} + [\mathbf{W} \cdot \mathbf{n}]_{\sigma', e} \right) \varphi_e^2 = 0.$$

Hence, since  $\sum_{e \in \mathfrak{E}(\sigma)} [\mathbf{W} \cdot \mathbf{n}]_{\sigma, e} = [\nabla_\Gamma \cdot \mathbf{W}]_\sigma$ ,

$$\mathcal{T}_1 = - \sum_{\sigma \in \mathfrak{S}_\Gamma} \sum_{e \in \mathfrak{E}(\sigma)} \frac{1}{2} [\mathbf{W} \cdot \mathbf{n}]_{\sigma, e} (\varphi_p - \varphi_e)^2 - \frac{1}{2} \sum_{\sigma \in \mathfrak{S}_\Gamma} [\nabla_\Gamma \cdot \mathbf{W}]_\sigma \varphi_p^2. \quad (55)$$

Using  $[\nabla_\Gamma \cdot \mathbf{W}]_\sigma \leq \|(\nabla_\Gamma \cdot \mathbf{W})^+\|_{C(\Gamma)} |\sigma|$  and the trace inequality (25), we write

$$- \sum_{\sigma \in \mathfrak{S}_\Gamma} [\nabla_\Gamma \cdot \mathbf{W}]_\sigma \varphi_p^2 \geq - \|(\nabla_\Gamma \cdot \mathbf{W})^+\|_{C(\Gamma)} \sum_{\sigma \in \mathfrak{S}_\Gamma} |\sigma| \varphi_p^2 \geq - \|(\nabla_\Gamma \cdot \mathbf{W})^+\|_{C(\Gamma)} C_{\text{tr}} |\varphi|_{V_h, \Omega}^2.$$

Plugging this into (55) and noticing, since  $d_{pe}^\perp \leq h_\Gamma$ , that

$$[\mathbf{W} \cdot \mathbf{n}]_{\sigma, e} \leq \|\mathbf{W}\|_{C(\Gamma)^d} |e| \leq h_\Gamma \|\mathbf{W}\|_{C(\Gamma)^d} \frac{|e|}{d_{pe}^\perp},$$

we obtain

$$\mathcal{T}_1 \geq -\frac{1}{2} h_\Gamma \|\mathbf{W}\|_{C(\Gamma)^d} |\varphi_h|_{V_h, \Gamma}^2 - \frac{1}{2} C_{\text{tr}} \|(\nabla_\Gamma \cdot \mathbf{W})^+\|_{C(\Gamma)} |\varphi_h|_{V_h, \Omega}^2.$$

Coming back to (54), we infer that

$$a_h(\varphi, \varphi) \geq \left( \rho_\Omega - \frac{1}{2} C_{\text{tr}} \|(\nabla_\Gamma \cdot \mathbf{W})^+\|_{C(\Gamma)} \right) |\varphi|_{V_h, \Omega}^2 + h_\Gamma \left( R\rho_\Gamma - \frac{1}{2} \|\mathbf{W}\|_{C(\Gamma)^d} \right) |\varphi|_{V_h, \Gamma}^2. \quad (56)$$

Owing to Assumption 26, this proves (52).

## A.2 Consistency

Using (16) and recalling that  $\ell_h(\varphi)$  is defined by the right-hand side of (51), we write

$$\begin{aligned} \ell_h(\varphi) &= \sum_{p \in \mathfrak{T}} \left( \sum_{\sigma \in \mathfrak{S}(p) \setminus \mathfrak{S}_\Gamma} \overline{F}_{p, \sigma}(\overline{T}) + \sum_{\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_\Gamma} \sum_{e \in E(\sigma)} [\overline{T} \mathbf{W} \cdot \mathbf{n}]_{\sigma, e} - \sum_{\sigma \in \mathfrak{S}(p) \cap \mathfrak{S}_\Gamma} [\overline{T} \nabla_\Gamma \cdot \mathbf{W}]_\sigma \right) \varphi_p \\ &= \sum_{\sigma \in \mathfrak{S} \setminus \mathfrak{S}_\Gamma} \overline{F}_{p, \sigma}(\overline{T}) (\varphi_p - \varphi_q) + \sum_{\sigma \in \mathfrak{S}_\Gamma} \sum_{e \in \mathfrak{E}(\sigma)} [\overline{T} \mathbf{W} \cdot \mathbf{n}]_{\sigma, e} (\varphi_p - \varphi_e) - \sum_{\sigma \in \mathfrak{S}_\Gamma} [\overline{T} \nabla_\Gamma \cdot \mathbf{W}]_\sigma \varphi_p, \end{aligned} \quad (57)$$

where we have used the conservativity of the fluxes to gather the sums by faces in the second equality. Subtracting  $a_h(I_h \bar{T}, \varphi)$  (given by the left-hand side of (51) with  $T$  replaced by  $I_h \bar{T}$ ), we can split the consistency error into four terms:

$$\mathcal{E}_h(I_h \bar{T}; \varphi) = \mathcal{T}_{c,1} + \mathcal{T}_{c,2} + \mathcal{T}_{c,3} + \mathcal{T}_{c,4} \quad (58)$$

with, setting  $I_h \bar{T} = ((\bar{T}_p)_{p \in \mathfrak{T}}, (\bar{T}_\sigma)_{\sigma \in \mathfrak{S}_{\text{Dir}}}, (\bar{T}_e)_{e \in \mathfrak{E}_\Gamma})$ ,

$$\begin{aligned} \mathcal{T}_{c,1} &= \sum_{\sigma \in \mathfrak{S} \setminus \mathfrak{S}_\Gamma} (\bar{F}_{p,\sigma}(\bar{T}) - \mathcal{F}_{p,\sigma}^\Omega(I_h \bar{T})) (\varphi_p - \varphi_q), \\ \mathcal{T}_{c,2} &= \sum_{\sigma \in \mathfrak{S}_\Gamma} \sum_{e \in \mathfrak{E}(\sigma)} \left( [\bar{T} \mathbf{W} \cdot \mathbf{n}]_{\sigma,e} - \bar{T}_e [\mathbf{W} \cdot \mathbf{n}]_{\sigma,e} \right) (\varphi_p - \varphi_e), \\ \mathcal{T}_{c,3} &= - \sum_{\sigma \in \mathfrak{S}_\Gamma} \left( [\bar{T} \nabla_\Gamma \cdot \mathbf{W}]_\sigma - \bar{T}_p [\nabla_\Gamma \cdot \mathbf{W}]_\sigma \right) \varphi_p, \\ \mathcal{T}_{c,4} &= Rh_\Gamma \sum_{\sigma \in \mathfrak{S}_\Gamma} \sum_{e \in \mathfrak{E}(\sigma)} \mathcal{F}_{\sigma,e}^\Gamma(I_h \bar{T}) (\varphi_p - \varphi_e). \end{aligned}$$

We now estimate each of these terms.

**Term  $\mathcal{T}_{c,1}$ .** Introducing  $\sqrt{d_{pq}/|\sigma|}$  and using a Cauchy–Schwarz inequality, the consistency property (22) yields

$$\begin{aligned} |\mathcal{T}_{c,1}| &\leq \left( \sum_{\sigma \in \mathfrak{S} \setminus \mathfrak{S}_\Gamma} \frac{d_{pq}}{|\sigma|} (\bar{F}_{p,\sigma}(\bar{T}) - \mathcal{F}_{p,\sigma}^\Omega(I_h \bar{T}))^2 \right)^{\frac{1}{2}} \left( \sum_{\sigma \in \mathfrak{S} \setminus \mathfrak{S}_\Gamma} \frac{|\sigma|}{d_{pq}} (\varphi_p - \varphi_q)^2 \right)^{\frac{1}{2}} \\ &\lesssim h \|\bar{T}\|_{C^2(\bar{\Omega})} \left( \sum_{\sigma \in \mathfrak{S} \setminus \mathfrak{S}_\Gamma} d_{pq} |\sigma| \right)^{\frac{1}{2}} |\varphi|_{V_h, \Omega}. \end{aligned}$$

Let  $D_{p\sigma}$  be the convex hull of  $\mathbf{x}_p$  and  $\sigma$ . If  $\sigma$  is flat, by [13, Lemma B.2] we have  $|D_{p\sigma}| = d_{p,\sigma}^\perp |\sigma|/3$  and thus, by definition of  $\text{reg}_{\mathfrak{T}}$  (which implies  $d_{pq} \leq \text{diam}(p) + \text{diam}(q) \lesssim \text{diam}(p) \lesssim d_{p\sigma}^\perp$ ),

$$\sum_{\sigma \in \mathfrak{S} \setminus \mathfrak{S}_\Gamma} d_{pq} |\sigma| \leq \sum_{p \in \mathfrak{T}} \sum_{\sigma \in \mathfrak{S}(p) \setminus \mathfrak{S}_\Gamma} d_{pq} |\sigma| \lesssim \sum_{p \in \mathfrak{T}} \sum_{\sigma \in \mathfrak{S}(p) \setminus \mathfrak{S}_\Gamma} |D_{p\sigma}| = \sum_{p \in \mathfrak{T}} |p| = |\Omega|. \quad (59)$$

This final estimate also holds in case of non-flat  $\sigma$ , as can be seen approximating  $\sigma$  by piecewise flat surfaces. Hence,

$$|\mathcal{T}_{c,1}| \lesssim h \|\bar{T}\|_{C^2(\bar{\Omega})} \|\varphi\|_{V_h}. \quad (60)$$

**Term  $\mathcal{T}_{c,2}$ .** We first estimate the consistency of the fluxes involved in this term. Using the definition of the interpolant (15) we have  $\int_e (\bar{T} - \bar{T}_e) = 0$  and thus

$$\begin{aligned} [\bar{T} \mathbf{W} \cdot \mathbf{n}]_{\sigma,e} - \bar{T}_e [\mathbf{W} \cdot \mathbf{n}]_{\sigma,e} &= \int_e \bar{T} \mathbf{W} \cdot \mathbf{n}_{\sigma,e} - \bar{T}_e \int_e \mathbf{W} \cdot \mathbf{n}_{\sigma,e} = \int_e (\bar{T} - \bar{T}_e) \mathbf{W} \cdot \mathbf{n}_{\sigma,e} \\ &= \int_e (\bar{T} - \bar{T}_e) \left( \mathbf{W} \cdot \mathbf{n}_{\sigma,e} - [\mathbf{W} \cdot \mathbf{n}]_{\sigma,e} \right) \leq h_\Gamma^2 |e| \|\bar{T}\|_{C^2(\bar{\Omega})} \|\mathbf{W}\|_{C^1(\Gamma)^d}, \end{aligned} \quad (61)$$

where the conclusion follows from a mean value theorem on  $\bar{T}$  and  $\mathbf{W}$ . Applying a Cauchy–Schwarz inequality and using (61), we infer

$$\begin{aligned} |\mathcal{T}_{c,2}| &\lesssim \left( \sum_{\sigma \in \mathfrak{S}_\Gamma} \sum_{e \in \mathfrak{E}(\sigma)} \frac{d_{pe}^\perp}{|e|} \left( [\bar{T}\mathbf{W} \cdot \mathbf{n}]_{\sigma,e} - \bar{T}_e [\mathbf{W} \cdot \mathbf{n}]_{\sigma,e} \right)^2 \right)^{\frac{1}{2}} \left( \sum_{\sigma \in \mathfrak{S}_\Gamma} \sum_{e \in \mathfrak{E}(\sigma)} \frac{|e|}{d_{pe}^\perp} (\varphi_p - \varphi_e)^2 \right)^{\frac{1}{2}} \\ &\lesssim h_\Gamma^2 \|\bar{T}\|_{C^1(\bar{\Omega})} \left( \sum_{\sigma \in \mathfrak{S}_\Gamma} \sum_{e \in \mathfrak{E}(\sigma)} d_{pe}^\perp |e| \right)^{\frac{1}{2}} |\varphi|_{V_h, \Gamma}. \end{aligned} \quad (62)$$

In a similar way as in the last equalities in (59),  $D_{pe}$  being the convex hull of  $\mathbf{x}_p$  and  $e$  we have

$$\sum_{\sigma \in \mathfrak{S}_\Gamma} \sum_{e \in \mathfrak{E}(\sigma)} d_{pe}^\perp |e| = 2 \sum_{\sigma \in \mathfrak{S}_\Gamma} \sum_{e \in \mathfrak{E}(\sigma)} |D_{pe}| = 2|\Gamma|. \quad (63)$$

Since  $|\varphi|_{V_h, \Gamma} \leq h_\Gamma^{-\frac{1}{2}} \|\varphi\|_{V_h}$ , we conclude that

$$|\mathcal{T}_{c,2}| \lesssim h_\Gamma^{\frac{3}{2}} \|\bar{T}\|_{C^2(\bar{\Omega})} \|\varphi\|_{V_h}. \quad (64)$$

*Remark 16* (Centred discretisation of the advective term). The approximation in the first term of (17) corresponds to a centred discretisation of the advection term  $\nabla_\Gamma \cdot (\bar{T}\mathbf{W})$  on  $\Gamma$ . To stabilise this centred discretisation and ensure the coercivity of the scheme, we have to add the artificial diffusion through the terms  $Rh_\Gamma \mathcal{F}^\Gamma(T)$  in (19a). A standard option to avoid adding numerical diffusion is to directly use an upwind discretisation of the advective term, as in (48). In this case, since stability would not require to introduce artificial diffusion, we would only consider cell unknowns in  $V_h$  (and not introduce edge unknowns), and we would take  $\|\cdot\|_{V_h} = |\cdot|_{V_h, \Omega}$ . The resulting scheme would be (48). Such a choice, however, would prevent us from introducing  $[\mathbf{W} \cdot \mathbf{n}]_{\sigma,e}$  in (61) and the resulting estimate would be in  $\mathcal{O}(h_\Gamma)$  instead of  $\mathcal{O}(h_\Gamma^2)$ . Carrying on as in (62) but with  $\|\cdot\|_{V_h, \Omega}$  instead of the absent  $|\cdot|_{V_h, \Gamma}$ , with the natural assumption that  $|e|h_\Gamma \lesssim |\sigma|$ , we would arrive at (64) with  $h_\Gamma^{\frac{1}{2}}$  instead of  $h_\Gamma^{\frac{3}{2}}$ . The final consistency estimate, and thus error estimate, would then be in  $\mathcal{O}(h^{\frac{1}{2}})$  instead of  $\mathcal{O}(h)$ .

**Term  $\mathcal{T}_{c,3}$ .** Notice first that

$$|[\bar{T}\nabla_\Gamma \cdot \mathbf{W}]_\sigma - \bar{T}_p [\nabla_\Gamma \cdot \mathbf{W}]_\sigma| = \left| \iint_\sigma (\bar{T} - \bar{T}_p) \nabla_\Gamma \cdot \mathbf{W} \right| \leq h_\Gamma \|\bar{T}\|_{C^1(\bar{\Omega})} |\sigma| \|\nabla_\Gamma \cdot \mathbf{W}\|_{C^0(\Gamma)}.$$

Hence, using a Cauchy–Schwarz inequality and the trace inequality (25),

$$\begin{aligned} |\mathcal{T}_{c,3}| &\leq \left( \sum_{\sigma \in \mathfrak{S}_\Gamma} \frac{1}{|\sigma|} ([\bar{T}\nabla_\Gamma \cdot \mathbf{W}]_\sigma - \bar{T}_p [\nabla_\Gamma \cdot \mathbf{W}]_\sigma)^2 \right)^{\frac{1}{2}} \left( \sum_{\sigma \in \mathfrak{S}_\Gamma} |\sigma| \varphi_p^2 \right)^{\frac{1}{2}} \\ &\lesssim h_\Gamma \|\bar{T}\|_{C^1(\bar{\Omega})} \left( \sum_{\sigma \in \mathfrak{S}_\Gamma} |\sigma| \right)^{\frac{1}{2}} C_{\text{tr}} |\varphi|_{V_h, \Omega} \leq h_\Gamma \|\bar{T}\|_{C^2(\bar{\Omega})} |\Gamma|^{\frac{1}{2}} C_{\text{tr}} \|\varphi\|_{V_h}. \end{aligned} \quad (65)$$

**Term  $\mathcal{T}_{c,4}$ .** Introducing the exact surface fluxes  $\bar{F}_{\sigma,e}(\bar{T}) = -\int_e \nabla \bar{T} \cdot \mathbf{n}_{\sigma,e}$  on  $\Gamma$ , we write

$$\mathcal{T}_{c,4} = Rh_\Gamma \underbrace{\sum_{\sigma \in \mathfrak{S}_\Gamma} \sum_{e \in \mathfrak{E}(\sigma)} (\mathcal{F}_{\sigma,e}^\Gamma(I_h \bar{T}) - \bar{F}_{\sigma,e}(\bar{T})) (\varphi_p - \varphi_e)}_{\mathcal{T}_{c,4,1}} + Rh_\Gamma \underbrace{\sum_{\sigma \in \mathfrak{S}_\Gamma} \sum_{e \in \mathfrak{E}(\sigma)} \bar{F}_{\sigma,e}(\bar{T}) (\varphi_p - \varphi_e)}_{\mathcal{T}_{c,4,2}}.$$

A Cauchy–Schwarz inequality, the consistency property (23), and (63) show that

$$\begin{aligned} |\mathcal{T}_{c,4,1}| &\leq \left( \sum_{\sigma \in \mathfrak{S}_\Gamma} \sum_{e \in \mathfrak{E}(\sigma)} \frac{d_{pe}^\perp}{|e|} (\mathcal{F}_{\sigma,e}^\Gamma(I_h \bar{T}) - \bar{F}_{\sigma,e}(\bar{T}))^2 \right)^{\frac{1}{2}} \left( \sum_{\sigma \in \mathfrak{S}_\Gamma} \sum_{e \in \mathfrak{E}(\sigma)} \frac{|e|}{d_{pe}^\perp} (\varphi_p - \varphi_e)^2 \right)^{\frac{1}{2}} \\ &\lesssim h_\Gamma \|\bar{T}\|_{C^2(\bar{\Omega})} |\varphi|_{V_h, \Gamma}. \end{aligned} \quad (66)$$

For  $\mathcal{T}_{c,4,2}$ , we use the conservativity of the fluxes  $\bar{F}_{\sigma,e}(\bar{T})$ , the fact that  $\varphi_e = 0$  for edges on the boundary of  $\Gamma$ ,  $\sum_{e \in \mathfrak{E}(\sigma)} \bar{F}_{\sigma,e}(\bar{T}) = -\iint_\sigma \Delta_\Gamma \bar{T}$ , and the trace inequality (25) to write

$$\begin{aligned} |\mathcal{T}_{c,4,2}| &= \left| \sum_{\sigma \in \mathfrak{S}_\Gamma} \sum_{e \in \mathfrak{E}(\sigma)} \bar{F}_{\sigma,e}(\bar{T}) \varphi_p \right| = \left| \sum_{\sigma \in \mathfrak{S}_\Gamma} -\iint_\sigma \Delta_\Gamma \bar{T} \varphi_p \right| \leq \left( \sum_{\sigma \in \mathfrak{S}_\Gamma} \frac{1}{|\sigma|} \left( \iint_\sigma \Delta_\Gamma \bar{T} \right)^2 \right)^{\frac{1}{2}} \left( \sum_{\sigma \in \mathfrak{S}_\Gamma} |\sigma| \varphi_p^2 \right)^{\frac{1}{2}} \\ &\leq \|\bar{T}\|_{C^2(\bar{\Omega})} |\Gamma|^{\frac{1}{2}} C_{\text{tr}} |\varphi|_{V_h, \Omega} \lesssim \|\bar{T}\|_{C^2(\bar{\Omega})} \|\varphi\|_{V_h}. \end{aligned}$$

Combined with (66) and recalling that  $\mathcal{T}_{4,c} = Rh_\Gamma \mathcal{T}_{4,c,1} + Rh_\Gamma \mathcal{T}_{4,c,2}$  this shows that

$$|\mathcal{T}_{c,4}| \lesssim h_\Gamma \|\bar{T}\|_{C^2(\bar{\Omega})} \|\varphi\|_{V_h}. \quad (67)$$

Gathering (60), (64), (65) and (67) in (58), we infer that (53) holds, which concludes the proof of Theorem 6.

## B Proof of Proposition 12

### B.1 Boundary fluxes

The coercivity of the HMM fluxes result from the construction of the method as a Gradient Discretisation Method, see [13, Chapter 13] – we note that this coercivity is purely algebraic, and not impacted by the curvature of the faces  $\sigma \in \mathfrak{S}_\Gamma$  or of their edges. In the case of flat faces and edges, the consistency of the fluxes (41) is a consequence of (40), see [13, Chapter 13] or [9, Example 31].

Consider now a curved face  $\sigma \in \mathfrak{S}_\Gamma$ , and assume that all edges of  $\sigma$  are “only curved along  $\sigma$ ” in the sense that  $\mathbf{n}_{\sigma,e}$  is constant over  $e$  for all  $e \in \mathfrak{E}(\sigma)$  (see Remark 11 otherwise). Because  $\Gamma$  is a smooth surface,  $\bar{\mathbf{x}}_e - \mathbf{x}_p$  is asymptotically close to the tangent plane to  $\Gamma$  at any point of  $\sigma$  (that is, the projection of  $\bar{\mathbf{x}}_e - \mathbf{x}_p$  in any normal direction to  $\sigma$  has length  $\mathcal{O}(h_\Gamma^2)$ ). Taylor expansions at any point of  $\sigma$  and the consistency property (40) thus give a constant  $C$  independent of the mesh such that, for  $\varphi \in C^2(\Gamma)$ ,

$$|S_{p,e}(I_h \varphi)| \leq C \|\varphi\|_{C^2(\Gamma)} h_\Gamma^2. \quad (68)$$

Using this estimate and (40), the arguments developed in [13, Chapter 13] can then easily be adapted and yield (23).

### B.2 Inner fluxes

#### B.2.1 Coercivity

Without any loss of generality we can select the vertex labels  $\mathbf{x}_{pq}^\oplus$  and  $\mathbf{x}_{pq}^\ominus$  (resp.  $\mathbf{x}_{pq}^\boxplus$  and  $\mathbf{x}_{pq}^\boxminus$ ) such that  $\alpha_{pq}^\ominus \geq 0$  (resp.  $\alpha_{pq}^\boxminus \geq 0$ ). Recalling the definition (36) of the fluxes, we use the zero value on the Dirichlet



boundary and the Young inequality  $xy \geq -\frac{1}{2}x^2 - \frac{1}{2}y^2$  to write

$$\begin{aligned}
\sum_{\sigma \in \mathfrak{S} \setminus \mathfrak{S}_\Gamma} \mathcal{F}_{p,\sigma}^\Omega(\varphi) (\varphi_p - \varphi_q) &= \sum_{\sigma_{pq} \in \mathfrak{S} \setminus \mathfrak{S}_\Gamma} |\sigma_{pq}| \left( \frac{1}{\beta_{pq}} \frac{\varphi_p - \varphi_q}{d_{pq}} + \frac{\alpha_{pq}^\circ}{\beta_{pq}} \frac{\varphi_{pq}^\oplus - \varphi_{pq}^\ominus}{d_{pq}^\circ} + \frac{\alpha_{pq}^\square}{\beta_{pq}} \frac{\varphi_{pq}^\boxplus - \varphi_{pq}^\boxminus}{d_{pq}^\square} \right) (\varphi_p - \varphi_q) \\
&= \sum_{\sigma_{pq} \in \mathfrak{S} \setminus (\mathfrak{S}_\Gamma \cup \mathfrak{S}_{\text{Dir}})} |\sigma_{pq}| \left( \frac{1}{\beta_{pq}} \frac{\varphi_p - \varphi_q}{d_{pq}} + \frac{\alpha_{pq}^\circ}{\beta_{pq}} \frac{\varphi_{pq}^\oplus - \varphi_{pq}^\ominus}{d_{pq}^\circ} + \frac{\alpha_{pq}^\square}{\beta_{pq}} \frac{\varphi_{pq}^\boxplus - \varphi_{pq}^\boxminus}{d_{pq}^\square} \right) (\varphi_p - \varphi_q) \\
&\quad + \sum_{\sigma_{pq} \in \mathfrak{S}_{\text{Dir}}} \frac{|\sigma_{pq}|}{d_{pq}} \frac{1}{\beta_{pq}} (\varphi_p - \varphi_q)^2 \\
&\geq \sum_{\sigma_{pq} \in \mathfrak{S} \setminus (\mathfrak{S}_\Gamma \cup \mathfrak{S}_{\text{Dir}})} |\sigma_{pq}| \left( \frac{1}{\beta_{pq} d_{pq}} (\varphi_p - \varphi_q)^2 - \frac{\alpha_{pq}^\circ}{2\beta_{pq} d_{pq}^\circ} (\varphi_p - \varphi_q)^2 - \frac{\alpha_{pq}^\square}{2\beta_{pq} d_{pq}^\square} (\varphi_p - \varphi_q)^2 \right. \\
&\quad \left. - \frac{\alpha_{pq}^\circ}{2\beta_{pq} d_{pq}^\circ} (\varphi_{pq}^\oplus - \varphi_{pq}^\ominus)^2 - \frac{\alpha_{pq}^\square}{2\beta_{pq} d_{pq}^\square} (\varphi_{pq}^\boxplus - \varphi_{pq}^\boxminus)^2 \right) \\
&\quad + \sum_{\sigma_{pq} \in \mathfrak{S}_{\text{Dir}}} \frac{|\sigma_{pq}|}{d_{pq}} \frac{1}{\beta_{pq}} (\varphi_p - \varphi_q)^2. \tag{69}
\end{aligned}$$

In order to establish (20), we now need to find a lower bound of this quantity in terms of sums of  $(\varphi_a - \varphi_b)^2$  for  $(a, b)$  pairs of neighbouring control volumes. The first stage is to recast  $\varphi_{pq}^\oplus - \varphi_{pq}^\ominus$  and  $\varphi_{pq}^\boxplus - \varphi_{pq}^\boxminus$  as combinations of differences of  $\varphi$  on neighbouring control volumes. Without loss of generality, we consider  $\varphi_{pq}^\oplus - \varphi_{pq}^\ominus$ . We have to deal with three cases, depending if the corresponding vertices are both internal, if one lies on  $\Gamma$ , or if one lies on the Dirichlet boundary  $\partial\Omega \setminus \Gamma$ .

**Case 1: Internal vertices.** We assume here that  $\mathbf{x}_{pq}^\oplus$  and  $\mathbf{x}_{pq}^\ominus$  are both in  $\Omega$ . Let  $\mathbf{x}_{pq}^*$  denote any one of these two vertices. Recalling the definitions in Section 3.1.3 (see also Figure 3) of  $F_{r,pq}^*$  and  $e_{r,pq}^*$ , we see that the set  $\cup_{r=p,q} (F_{r,pq}^* \cup \{r, e_{r,pq}^*\})$  is made of the eight control volumes around  $\mathbf{x}_{pq}^*$  whose unknowns are involved in the definition (37) of  $\varphi_{pq}^*$ . Hence, we can decompose  $\varphi_{pq}^\oplus - \varphi_{pq}^\ominus$  as

$$\varphi_{pq}^\oplus - \varphi_{pq}^\ominus = \frac{1}{8} \sum_{r=p,q} \left( \varphi_{e_{r,pq}^\oplus} + \sum_{f \in F_{r,pq}^\oplus} \varphi_f + \varphi_r - \varphi_r - \sum_{f \in F_{r,pq}^\ominus} \varphi_f - \varphi_{e_{r,pq}^\ominus} \right).$$

Inside the sum in the right-hand side, each cell unknown appears with the coefficients represented in Figure 8 (left). Our goal is to gather these terms together in order to write  $\varphi_{pq}^\oplus - \varphi_{pq}^\ominus$  as a combination of terms  $\varphi_b - \varphi_a$  with  $a$  and  $b$  neighbouring control volumes. This is done by splitting the coefficients in order to associate (parts of) each cell unknown with a neighbouring cell unknown, as in Figure 8 (right).

This consists in writing

$$\begin{aligned}
\varphi_{pq}^\oplus - \varphi_{pq}^\ominus &= \frac{1}{8} \sum_{r=p,q} \left( \frac{\varphi_{e_{r,pq}^\oplus}}{2} + \frac{\varphi_{e_{r,pq}^\oplus}}{2} + \sum_{f \in F_{r,pq}^\oplus} \left( 3\frac{\varphi_f}{2} - \frac{\varphi_f}{2} \right) + 3\frac{\varphi_r}{2} + 3\frac{\varphi_r}{2} \right. \\
&\quad \left. - 3\frac{\varphi_r}{2} - 3\frac{\varphi_r}{2} - \sum_{f \in F_{r,pq}^\ominus} \left( 3\frac{\varphi_f}{2} - \frac{\varphi_f}{2} \right) - \frac{\varphi_{e_{r,pq}^\ominus}}{2} - \frac{\varphi_{e_{r,pq}^\ominus}}{2} \right) \\
&= \frac{1}{16} \sum_{r=p,q} \left( \sum_{f \in F_{r,pq}^\oplus} \left( (\varphi_{e_{r,pq}^\oplus} - \varphi_f) + 3(\varphi_f - \varphi_r) \right) + \sum_{f \in F_{r,pq}^\ominus} \left( 3(\varphi_r - \varphi_f) + (\varphi_f - \varphi_{e_{r,pq}^\ominus}) \right) \right).
\end{aligned}$$

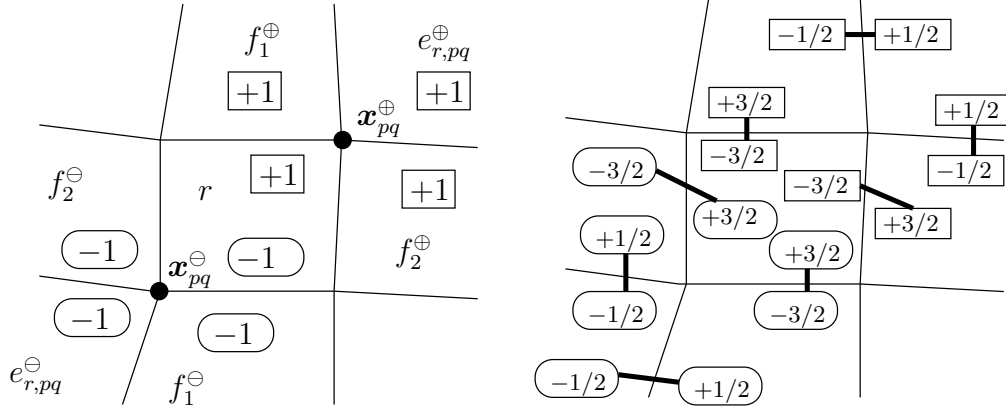


Figure 8: Internal vertices. Left: coefficients associated to each cell unknown in the expression of  $\varphi_{pq}^{\oplus} - \varphi_{pq}^{\ominus}$  (projection on 2D). Right: splitting of these coefficients, and associations of neighbouring control volumes. The sums of coefficients in each cell are the same in both pictures.

We simplify this expression by gathering the terms involving  $F_{r,pq}^{\oplus}$  and  $F_{r,pq}^{\ominus}$  under a sum  $\sum_{* \in \{\oplus, \ominus\}}$ : setting  $\delta_{\oplus} = +1$  and  $\delta_{\ominus} = -1$ , we have

$$\varphi_{pq}^{\oplus} - \varphi_{pq}^{\ominus} = \frac{1}{16} \sum_{r=p,q} \sum_{* \in \{\oplus, \ominus\}} \sum_{f \in F_{r,pq}^*} \delta_* \left( (\varphi_{e_{r,pq}^*} - \varphi_f) + 3(\varphi_f - \varphi_r) \right).$$

Using the definition (45), we arrive at

$$\varphi_{pq}^{\oplus} - \varphi_{pq}^{\ominus} = \frac{1}{16} \sum_{r=p,q} \sum_{* \in \{\oplus, \ominus\}} \sum_{f \in F_{r,pq}^*} \delta_* \left( \zeta_{X,pq}^* (\varphi_{e_{r,pq}^*} - \varphi_f) + \zeta_{Y,pq}^* (\varphi_f - \varphi_r) \right). \quad (70)$$

**Case 2: Vertex on  $\Gamma$ .** One of  $x_{pq}^{\oplus}$  or  $x_{pq}^{\ominus}$  lies on  $\Gamma$ . Without loss of generality we assume it is  $x_{pq}^{\ominus}$ . The boundary value  $\varphi_{pq}^{\ominus}$  is expressed as 1/4 of the sum of the unknowns in four faces lying on  $\Gamma$  or, equivalently, 1/8 of the sum of these four values associated with coefficients 2. Plugging this expression into  $\varphi_{pq}^{\oplus} - \varphi_{pq}^{\ominus}$  and reasoning as in Case 1, but using this time the splitting of coefficients represented in Figure 9, we arrive at

$$\varphi_{pq}^{\oplus} - \varphi_{pq}^{\ominus} = \frac{1}{16} \sum_{r=p,q} \sum_{f \in F_{r,pq}^{\oplus}} \left( (\varphi_{e_{r,pq}^{\oplus}} - \varphi_f) + 3(\varphi_f - \varphi_r) \right) + \frac{1}{16} \sum_{r=p,q} \sum_{f \in F_{r,pq}^{\ominus}} 4(\varphi_r - \varphi_f) \quad (71)$$

(the sum over  $f \in F_{r,pq}^{\ominus}$  actually only contains one term, but is written this way for homogeneity of notations). This sum can be written in the form (70), owing to the definition (45) of  $(\zeta_{X,pq}^*, \zeta_{Y,pq}^*)$ .

**Case 3: Vertex on the Dirichlet boundary.** Assuming  $x_{pq}^{\ominus} \in \partial\Omega \setminus \Gamma$ , we have  $\varphi_{pq}^{\ominus} = 0$ . Then  $F_{r,pq}^{\ominus}$  is made of the unique face  $f^{\ominus} = \sigma$  (degenerate cell) of  $r$  on  $\partial\Omega \setminus \Gamma$ , associated with a value  $\varphi_{f^{\ominus}} = 0$ . The splitting of coefficients described in Figure 10 leads to (71) with the last coefficient 4 replaced by 8; thus, recalling the definition of  $(\zeta_{X,pq}^*, \zeta_{Y,pq}^*)$  in (45), we can again write  $\varphi_{pq}^{\oplus} - \varphi_{pq}^{\ominus}$  in the form (70).

**Conclusion.** We established that the formula (70) always holds, no matter the positions of the vertices. Accounting for the definitions of  $F_{r,pq}^*$  and of  $(\zeta_{X,pq}^*, \zeta_{Y,pq}^*)$  (see (45)), we see that the right-hand side of this relation is made of at most 32 sums of  $\delta_*(\varphi_a - \varphi_b)$  (with  $(a, b) = (e_{r,pq}^*, f)$  or  $(a, b) = (f, r)$ ). Hence, using the convexity of the square function, we obtain

$$\left( \frac{\varphi_{pq}^{\oplus} - \varphi_{pq}^{\ominus}}{2} \right)^2 \leq \frac{1}{32} \sum_{r=p,q} \sum_{* \in \{\oplus, \ominus\}} \sum_{f \in F_{r,pq}^*} \left( \zeta_{X,pq}^* (\varphi_{e_{r,pq}^*} - \varphi_f)^2 + \zeta_{Y,pq}^* (\varphi_f - \varphi_r)^2 \right).$$

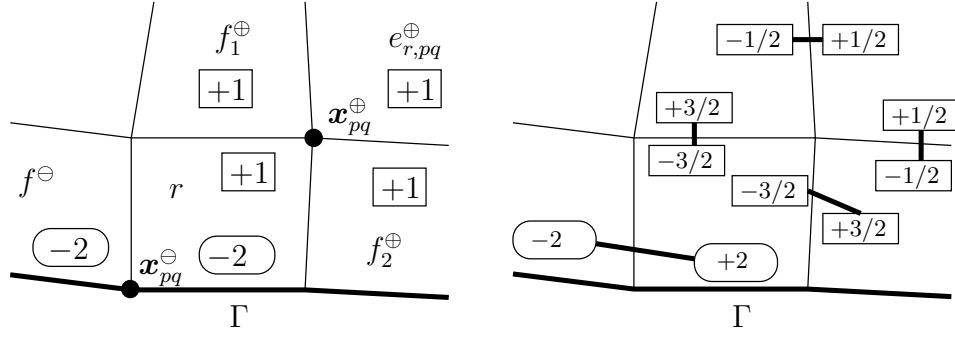


Figure 9:  $\mathbf{x}_{pq}^\ominus$  on  $\Gamma$ . Left: coefficients associated to each cell unknown in the expression of  $\varphi_{pq}^\oplus - \varphi_{pq}^\ominus$  (projection on 2D). Right: splitting of these coefficients, and associations of neighbouring control volumes. The sums of coefficients in each cell are the same in both pictures.

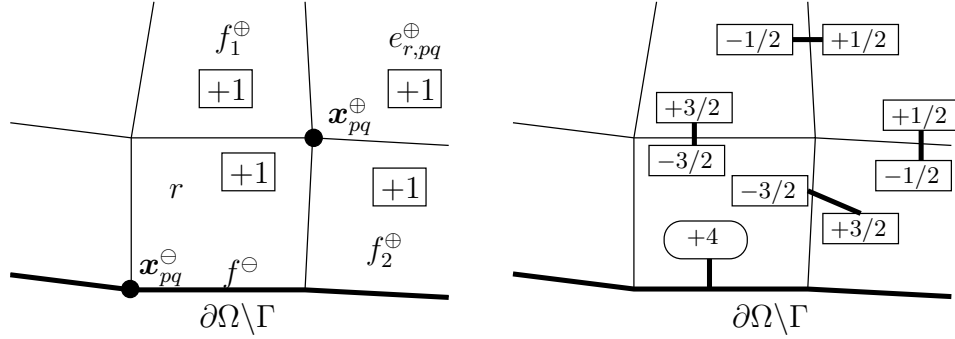


Figure 10:  $\mathbf{x}_{pq}^\ominus$  on  $\partial\Omega \setminus \Gamma$ . Left: coefficients associated to each cell unknown in the expression of  $\varphi_{pq}^\oplus - \varphi_{pq}^\ominus = \varphi_{pq}^\oplus$  (projection on 2D). Right: splitting of these coefficients, and associations of neighbouring control volumes. The sums of coefficients in each cell are the same in both pictures.

A similar estimate can be obtained for  $(\varphi_{pq}^\boxplus - \varphi_{pq}^\boxminus)^2$ , by replacing the sum range  $* \in \{\oplus, \ominus\}$  with  $* \in \{\boxplus, \boxminus\}$ . By plugging these bounds into (69) we obtain

$$\begin{aligned}
& \sum_{\sigma \in \mathfrak{S} \setminus \mathfrak{S}_\Gamma} \mathcal{F}_{p,\sigma}^\Omega(\varphi)(\varphi_p - \varphi_q) \\
& \geq \sum_{\sigma_{pq} \in \mathfrak{S} \setminus \mathfrak{S}_\Gamma} |\sigma_{pq}| \left[ \frac{1}{\beta_{pq} d_{pq}} - \epsilon_{pq} \frac{\alpha_{pq}^\circ}{2\beta_{pq} d_{pq}^\circ} - \epsilon_{pq} \frac{\alpha_{pq}^\square}{2\beta_{pq} d_{pq}^\square} \right] (\varphi_p - \varphi_q)^2 \\
& \quad - \frac{1}{16} \sum_{\sigma_{pq} \in \mathfrak{S} \setminus (\mathfrak{S}_\Gamma \cup \mathfrak{S}_{\text{Dir}})} \sum_{r=p,q} \sum_{* \in \{\oplus, \ominus, \boxplus, \boxminus\}} \frac{|\sigma_{pq}| \alpha_{pq}^\diamond}{\beta_{pq} d_{pq}^\diamond} \sum_{f \in F_{r,pq}^*} \left( \zeta_{X,pq}^* (\varphi_{e_{r,pq}^*} - \varphi_f)^2 + \zeta_{Y,pq}^* (\varphi_f - \varphi_r)^2 \right), \quad (72)
\end{aligned}$$

where  $\diamond$  is given by (46), and we have used the definition (45) of  $\epsilon_{pq}$  to integrate the last term in (69) into the first one in the right-hand side above.

All the differences of  $\varphi$  in this equation are differences  $(\varphi_a - \varphi_b)^2$  of values across a face  $\sigma_{ab} \in \mathfrak{S} \setminus \mathfrak{S}_\Gamma$ . For such a given face, the sets  $X_{ab}$  and  $Y_{ab}$  defined by (43) precisely identify the indices in the second

addend of (72) that involve the term  $(\varphi_a - \varphi_b)^2$ . Hence, (72) can be re-arranged as

$$\begin{aligned} \sum_{\sigma \in \mathfrak{S} \setminus \mathfrak{S}_\Gamma} \mathcal{F}_{p,\sigma}^\Omega(\varphi) (\varphi_p - \varphi_q) &\geq \sum_{\sigma_{ab} \in \mathfrak{S} \setminus \mathfrak{S}_\Gamma} \frac{|\sigma_{ab}|}{d_{ab}} \left\{ \left[ \frac{1}{\beta_{ab}} - \epsilon_{ab} \frac{\alpha_{ab}^\circ d_{ab}}{2\beta_{ab} d_{ab}^\circ} - \epsilon_{ab} \frac{\alpha_{ab}^\square d_{ab}}{2\beta_{ab} d_{ab}^\square} \right] \right. \\ &\quad \left. - \frac{1}{16} \sum_{(p,q,*) \in X_{ab}} \zeta_{X,pq}^* \frac{|\sigma_{pq}| d_{ab} \alpha_{pq}^\diamond}{|\sigma_{ab}| d_{pq}^\diamond \beta_{pq}} - \frac{1}{16} \sum_{(p,q,*) \in Y_{ab}} \zeta_{Y,pq}^* \frac{|\sigma_{pq}| d_{ab} \alpha_{pq}^\diamond}{|\sigma_{ab}| d_{pq}^\diamond \beta_{pq}} \right\} (\varphi_a - \varphi_b)^2. \end{aligned}$$

The definition (44) then shows that (20) holds with  $\varrho_\Omega = \varrho_{\mathfrak{T},\Omega}$ .

*Remark 17* (Alternative coercivity factor). For each  $\sigma_{pq} \in \mathfrak{S} \setminus \mathfrak{S}_\Gamma$ , take  $\varpi_{pq} > 0$ . Using before (69) the generalised Young inequality  $xy \geq -\frac{\varpi_{pq}^{-1}}{2} x^2 - \frac{\varpi_{pq}}{2} y^2$ , instead of the standard one with  $\varpi_{pq} = 1$ , the reasoning above shows that the regularity factor  $\varrho_{\mathfrak{T},\Omega}$  can be re-defined such that

$$\begin{aligned} \varrho_{\mathfrak{T},\Omega} := \max \left\{ \left[ \frac{1}{\beta_{ab}} - \epsilon_{ab} \frac{\varpi_{ab}^{-1} \alpha_{ab}^\circ d_{ab}}{2\beta_{ab} d_{ab}^\circ} - \epsilon_{ab} \frac{\varpi_{ab}^{-1} \alpha_{ab}^\square d_{ab}}{2\beta_{ab} d_{ab}^\square} \right] - \frac{1}{16} \sum_{(p,q,*) \in X_{ab}} \zeta_{X,pq}^* \frac{\varpi_{pq} |\sigma_{pq}| d_{ab} \alpha_{pq}^\diamond}{|\sigma_{ab}| d_{pq}^\diamond \beta_{pq}} \right. \\ \left. - \frac{1}{16} \sum_{(p,q,*) \in Y_{ab}} \zeta_{Y,pq}^* \frac{\varpi_{pq} |\sigma_{pq}| d_{ab} \alpha_{pq}^\diamond}{|\sigma_{ab}| d_{pq}^\diamond \beta_{pq}} : \sigma_{ab} \in \mathfrak{S} \setminus \mathfrak{S}_\Gamma \right\}. \end{aligned} \quad (73)$$

For certain choices of  $\varpi_{ab}$ , this alternative coercivity factor could remain positive and bounded above for certain meshes, for which the factor satisfying (44) is negative.

### B.2.2 Consistency

We start with a preliminary estimate. Let  $\mathbf{a} = \left( \frac{1}{\beta_{pq}}, -\frac{\alpha_{pq}^\circ}{\beta_{pq}}, -\frac{\alpha_{pq}^\square}{\beta_{pq}} \right)$ . The relation (35) yields

$$\mathbf{a} = \left[ \mathbf{s}_{pq}, \mathbf{t}_{pq}^\circ, \mathbf{t}_{pq}^\square \right]^{-1} \frac{\tilde{\mathbf{n}}_{pq}}{|\sigma_{pq}|}.$$

Since  $\left| \frac{\tilde{\mathbf{n}}_{pq}}{|\sigma_{pq}|} \right| \leq 1$  (see (34)) we infer  $|\mathbf{a}| \leq \left\| \left[ \mathbf{s}_{pq}, \mathbf{t}_{pq}^\circ, \mathbf{t}_{pq}^\square \right]^{-1} \right\|$ . The vectors  $\mathbf{s}_{pq}, \mathbf{t}_{pq}^\circ, \mathbf{t}_{pq}^\square$  having unit length, the representation of the inverse of  $\left[ \mathbf{s}_{pq}, \mathbf{t}_{pq}^\circ, \mathbf{t}_{pq}^\square \right]$  using the co-matrix and the determinant give a universal constant  $C$  such that

$$\left| \left( \frac{1}{\beta_{pq}}, -\frac{\alpha_{pq}^\circ}{\beta_{pq}}, -\frac{\alpha_{pq}^\square}{\beta_{pq}} \right) \right| \leq \frac{C}{|\det(\mathbf{s}_{pq}, \mathbf{t}_{pq}^\circ, \mathbf{t}_{pq}^\square)|}. \quad (74)$$

Let  $u \in C^2(\bar{\Omega})$  with  $u = 0$  on  $\partial\Omega \setminus \Gamma$ . In the following, we write  $\mathcal{O}(s)$  for generic functions that satisfy  $|\mathcal{O}(s)| \leq C \|u\|_{C^2(\bar{\Omega})} |s|$  with  $C$  depending only on an upper bound of the regularity factors  $\text{reg}_{\mathfrak{T}}$  and  $\text{reg}_{\mathfrak{T},\Omega}$  defined by (24) and (42). This notation naturally extends to the case where  $s$  is a vector.

Taking an arbitrary point  $\mathbf{x}_\sigma \in \sigma_{pq}$ , (34) and (35) show that

$$\begin{aligned} \bar{F}_{p,\sigma}(u) &= \iint_{\sigma_{pq}} \nabla u \cdot \mathbf{n}_{pq} = \nabla u(\mathbf{x}_\sigma) \cdot \iint_{\sigma_{pq}} \mathbf{n}_{pq} + |\sigma_{pq}| \mathcal{O}(h) \\ &= |\sigma_{pq}| \left( \frac{1}{\beta_{pq}} \nabla u(\mathbf{x}_\sigma) \cdot \mathbf{s}_{pq} - \frac{\alpha_{pq}^\circ}{\beta_{pq}} \nabla u(\mathbf{x}_\sigma) \cdot \mathbf{t}_{pq}^\circ - \frac{\alpha_{pq}^\square}{\beta_{pq}} \nabla u(\mathbf{x}_\sigma) \cdot \mathbf{t}_{pq}^\square \right) + |\sigma_{pq}| \mathcal{O}(h). \end{aligned} \quad (75)$$

Let us look at each directional derivative separately. Since  $d_{pq} \leq 2h$ , the definition (33) of  $\mathbf{s}_{pq}$  and a Taylor expansion yield

$$\nabla u(\mathbf{x}_\sigma) \cdot \mathbf{s}_{pq} = \frac{u(\mathbf{x}_p) - u(\mathbf{x}_q)}{d_{pq}} + \mathcal{O}(h). \quad (76)$$

For the derivative in the tangential direction  $\mathbf{t}_{pq}^\circ$ , Lemma 18 below shows that

$$\begin{aligned}\nabla u(\mathbf{x}_\sigma) \cdot \mathbf{t}_{pq}^\circ &= \frac{u(\mathbf{x}_{pq}^\oplus) - u(\mathbf{x}_{pq}^\ominus)}{d_{pq}^\circ} + \mathcal{O}(h) \\ &= \frac{\frac{1}{\text{Card}(\mathcal{R}(\mathbf{x}_{pq}^\oplus))} \sum_{\mathbf{y} \in \mathcal{R}(\mathbf{x}_{pq}^\oplus)} u(\mathbf{y}) + \mathcal{O}(d_{pq}^{\oplus 2}) - \frac{1}{\text{Card}(\mathcal{R}(\mathbf{x}_{pq}^\ominus))} \sum_{\mathbf{y} \in \mathcal{R}(\mathbf{x}_{pq}^\ominus)} u(\mathbf{y}) + \mathcal{O}(d_{pq}^{\ominus 2})}{d_{pq}^\circ} + \mathcal{O}(h),\end{aligned}$$

with  $\mathbf{d}_{pq}^{*2}$  the vector obtained by component-wise squaring  $\mathbf{d}_{pq}^*$ . Using the definition of  $\text{reg}_{\mathfrak{T}, \Omega}$  we infer

$$\nabla u(\mathbf{x}_\sigma) \cdot \mathbf{t}_{pq}^\circ = \frac{\frac{1}{\text{Card}(\mathcal{R}(\mathbf{x}_{pq}^\oplus))} \sum_{\mathbf{y} \in \mathcal{R}(\mathbf{x}_{pq}^\oplus)} u(\mathbf{y}) - \frac{1}{\text{Card}(\mathcal{R}(\mathbf{x}_{pq}^\ominus))} \sum_{\mathbf{y} \in \mathcal{R}(\mathbf{x}_{pq}^\ominus)} u(\mathbf{y})}{d_{pq}^\circ} + \mathcal{O}(h). \quad (77)$$

Similarly,

$$\nabla u(\mathbf{x}_\sigma) \cdot \mathbf{t}_{pq}^\square = \frac{\frac{1}{\text{Card}(\mathcal{R}(\mathbf{x}_{pq}^\boxplus))} \sum_{\mathbf{y} \in \mathcal{R}(\mathbf{x}_{pq}^\boxplus)} u(\mathbf{y}) - \frac{1}{\text{Card}(\mathcal{R}(\mathbf{x}_{pq}^\boxminus))} \sum_{\mathbf{y} \in \mathcal{R}(\mathbf{x}_{pq}^\boxminus)} u(\mathbf{y})}{d_{pq}^\square} + \mathcal{O}(h). \quad (78)$$

Plug (76)–(78) into (75) and subtract  $\mathcal{F}_{p,\sigma}^\Omega(I_h u)$  defined by (36) with  $\varphi = I_h u$ , so that  $\varphi_p = u(\mathbf{x}_p)$  for all  $p \in \mathfrak{T}$  and  $\varphi_{\mathbf{y}} = u(\mathbf{x}_q) = 0$  if  $q = \sigma \in \mathfrak{S}_{\text{Dir}}$ . This gives

$$\bar{F}_{p,\sigma}(u) - \mathcal{F}_{p,\sigma}^\Omega(I_h u) = |\sigma_{pq}| \left( \frac{\mathcal{O}(h)}{\beta_{pq}} - \frac{\alpha_{pq}^\circ \mathcal{O}(h)}{\beta_{pq}} - \frac{\alpha_{pq}^\square \mathcal{O}(h)}{\beta_{pq}} \right) + |\sigma_{pq}| \mathcal{O}(h). \quad (79)$$

Estimate (74) and the definition (42) of  $\text{reg}_{\mathfrak{T}, \Omega}$  then conclude the proof of (22), with a constant that only depends on an upper bound of this regularity factor.

**Lemma 18** (Error for smooth functions on barycentric combinations). *Let  $U$  be a convex open set of  $\mathbb{R}^3$ ,  $(\mathbf{z}_i)_{i=1,\dots,I}$  be points in  $U$ , and let  $\mathbf{z} = \sum_{i=1}^I \lambda_i \mathbf{z}_i$  for some convex coefficients  $(\lambda_i)_{i=1,\dots,I}$ . If  $\psi \in C^2(\bar{U})$  then*

$$\left| \psi(\mathbf{z}) - \sum_{i=1}^I \lambda_i \psi(\mathbf{z}_i) \right| \leq \frac{1}{2} \|\psi\|_{C^2(\bar{U})} \max_{i=1,\dots,I} |\mathbf{z}_i - \mathbf{z}|^2. \quad (80)$$

*Proof.* This lemma is classical but its (short) proof is recalled for the sake of legibility. A Taylor expansion around  $\mathbf{z}$  gives

$$\psi(\mathbf{x}) = \psi(\mathbf{z}) + \nabla \psi(\mathbf{z}) \cdot (\mathbf{x} - \mathbf{z}) + \text{Rem}(\mathbf{x}, \mathbf{z}), \quad (81)$$

where  $|\text{Rem}(\mathbf{x}, \mathbf{z})| \leq \frac{1}{2} \|\psi\|_{C^2(\bar{U})} |\mathbf{x} - \mathbf{z}|^2$ . Apply (81) to  $\mathbf{x} = \mathbf{z}_i$ , multiply by  $\lambda_i$  and sum over  $i = 1, \dots, I$ . Since  $\sum_{i=1}^I \lambda_i (\mathbf{z}_i - \mathbf{z}) = 0$  the term involving  $\nabla \psi(\mathbf{z})$  disappears and (80) follows.  $\square$

**Acknowledgement:** this research was supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (project number DP170100605), and by the Slovak Research and Development Agency (grant APVV-0522-15).

## References

- [1] G. Backus. Application of a non-linear boundary-value problem for laplace's equation to gravity and geomagnetic intensity surveys. *Q J Mech Appl Math*, 21(2):195–221, 1968.
- [2] J. W. Barrett and C. M. Elliott. Fixed mesh finite element approximations to a free boundary problem for an elliptic equation with an oblique derivative boundary condition. *Computers & Mathematics with Applications*, 11(4):335 – 345, 1985.
- [3] A. Bjerhammar and L. Svensson. On the geodetic boundary value problem for a fixed boundary surface—a satellite approach. *Bulletin géodésique*, 57(1):382–393, Mar 1983.

- [4] A. Bradji and J. Fuhrmann. On the convergence and convergence order of finite volume gradient schemes for oblique derivative boundary value problems. *Computational and Applied Mathematics*, 37(3):2533–2565, Jul 2018.
- [5] A. Bradji and T. Gallouët. Error estimate for finite volume approximate solutions of some oblique derivative boundary value problems. *International Journal on Finite Volumes*, 3(2):1–35, 2006.
- [6] F. Brezzi, K. Lipnikov, and M. Shashkov. Convergence of mimetic finite difference method for diffusion problems on polyhedral meshes with curved faces. *Math. Models Methods Appl. Sci.*, 16(2):275–297, 2006.
- [7] S. Bruinsma, C. Foerste, O. Abrikosov, J.-C. Marty, M.-H. Rio, S. Mulet, and B. Sylvain. The new esa satellite-only gravity field model via the direct approach. *Geophysical Research Letters*, 40, 07 2013.
- [8] R. Čunderlík, K. Mikula, and M. Mojzeš. Numerical solution of the linearized fixed gravimetric boundary-value problem. *Journal of Geodesy*, 82(1):15–29, Jan 2008.
- [9] D. A. Di Pietro and J. Droniou. A third Strang lemma and an Aubin–Nitsche trick for schemes in fully discrete formulation. *Calcolo*, 55(3):55:40, 2018.
- [10] G. Díaz, J. Díaz, and J. Otero. On an oblique boundary value problem related to the backus problem in geodesy. *Nonlinear Analysis: Real World Applications*, 7(2):147 – 166, 2006.
- [11] G. Díaz, J. I. Díaz, and J. Otero. Construction of the maximal solution of backus’ problem in geodesy and geomagnetism. *Studia Geophysica et Geodaetica*, 55(3):415, Aug 2011.
- [12] J. Droniou and R. Eymard. A mixed finite volume scheme for anisotropic diffusion problems on any grid. *Numer. Math.*, 105(1):35–71, 2006.
- [13] J. Droniou, R. Eymard, T. Gallouët, C. Guichard, and R. Herbin. *The gradient discretisation method*, volume 82 of *Mathematics & Applications*. Springer, 2018.
- [14] J. Droniou, R. Eymard, T. Gallouët, and R. Herbin. A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods. *Math. Models Methods Appl. Sci.*, 20(2):265–295, 2010.
- [15] J. Droniou and N. Nataraj. Improved  $L^2$  estimate for gradient schemes and super-convergence of the TPFA finite volume scheme. *IMA J. Numer. Anal.*, 38(3):1254–1293, 2018.
- [16] Z. Fašková, R. Čunderlík, and K. Mikula. Finite element method for solving geodetic boundary value problems. *Journal of Geodesy*, 84(2):135–144, Feb 2010.
- [17] E. Grafarend. The geoid and the gravimetric boundary value problem. *Report N 18*, 1989.
- [18] E. Grafarend and W. Niemeier. The free nonlinear boundary value problem of physical geodesy. *Bulletin Géodésique (1946-1975)*, 101(1):243–261, Sep 1971.
- [19] T. Grand, J. Šefara, R. Pašteka, M. Bielik, and S. Daniel. Atlas of geophysical maps and profiles, part d1: gravimetry, 2001.
- [20] B. Heck. On the non-linear geodetic boundary value problem for a fixed boundary surface. *Bulletin géodésique*, 63(1):57–67, Mar 1989.
- [21] B. Hofmann-Wellenhof and H. Moritz. Physical geodesy. 2005.
- [22] P. Holota. Coerciveness of the linear gravimetric boundary-value problem and a geometrical interpretation. *Journal of Geodesy*, 71(10):640–651, Sep 1997.
- [23] P. Holota. Neumann’s boundary-value problem in studies on earth gravity field: weak solution. *50 years of Research Institute of Geodesy, Topography and Cartography*, pages 49–69, 2005.
- [24] K. R. Koch and A. J. Pope. Uniqueness and existence for the geodetic boundary value problem using the known surface of the earth. *Bulletin Géodésique (1946-1975)*, 106(1):467–476, Dec 1972.
- [25] M. Macák, R. Cunderlik, K. Mikula, and Z. Minarechova. An upwind-based scheme for solving the oblique derivative boundary-value problem related to physical geodesy. *Journal of Geodetic Science*, 5, 01 2015.
- [26] M. Macák, K. Mikula, and Z. Minarechova. Solving the oblique derivative boundary-value problem by the finite volume method. *ALGORITMY 2012, 19th Conference on Scientific Computing, Podbanske, Slovakia, September 9-14, 2012, Proceedings of contributed papers and posters (Eds. A.Handlovicova, Z.Minarechova, D.Sevcovic)*, ISBN 978-80-227-3742-5, Publishing House of STU, pages 75–84, 2012.
- [27] M. Macák, K. Mikula, and Z. Minarechova. A novel scheme for solving oblique derivative boundary-value problem. *Studia Geophysica et Geodaetica*, 58:556–570, 2014.
- [28] M. Macák, K. Mikula, Z. Minarechová, and R. Čunderlík. On an iterative approach to solving the nonlinear satellite-fixed geodetic boundary-value problem. In N. Sneeuw, P. Novák, M. Crespi, and F. Sansò, editors, *VIII Hotine-Marussi Symposium on Mathematical Geodesy*, pages 185–192, Cham, 2016. Springer International Publishing.
- [29] M. Medla, K. Mikula, R. Čunderlík, and M. Macák. Numerical solution to the oblique derivative boundary value problem on non-uniform grids above the earth topography. *Journal of Geodesy*, 92(1):1–19, 2018.
- [30] F. Sacerdote and F. Sansò. On the analysis of the fixed-boundary gravimetric boundary-value problem. pages 507–516, Pisa, Politecnico di Milano, 06 1989.
- [31] L. Sanchez, R. Cunderlik, N. Dayoub, K. Mikula, Z. Minarechova, Z. Sima, V. Vatrť, and M. Vojtiskova. A conventional value for the geoid reference potential  $w_0$ . *Journal of Geodesy*, 90(9):815–835, 2016.